IMPROVING ASR ROBUSTNESS TO PERTURBED SPEECH USING CYCLE-CONSISTENT GENERATIVE ADVERSARIAL NETWORKS

Sri Harsha Dumpala, Imran Sheikh, Rupayan Chakraborty, Sunil Kumar Kopparapu

TCS Research and Innovation - Mumbai, INDIA

ABSTRACT

Naturally introduced perturbations in audio signal, caused by emotional and physical states of the speaker, can significantly degrade the performance of Automatic Speech Recognition (ASR) systems. In this paper, we propose a front-end based on Cycle-Consistent Generative Adversarial Network (CycleGAN) which transforms naturally perturbed speech into normal speech, and hence improves the robustness of an ASR system. The CycleGAN model is trained on non-parallel examples of perturbed and normal speech. Experiments on spontaneous laughter-speech and creaky voice datasets show that the performance of four different ASR systems improve by using speech obtained from CycleGAN based front-end, as compared to directly using the original perturbed speech. Visualization of the features of the laughter perturbed speech and those generated by the proposed front-end further demonstrates the effectiveness of our approach.

Index Terms— Cycle-consistent GAN, laughter speech, creaky speech, automatic speech recognition

1. INTRODUCTION

Performance of Automatic Speech Recognition (ASR) systems have seen significant jumps with the adoption of deep learning techniques. Recently, ASR systems have been shown to perform on par with human transcribers [1]. At the same time, the use of voice assistants such as Siri, Google Assistant, Amazon Alexa etc., have led to the wide use of ASR systems in various day-to-day applications. However, recent studies have shown that adversarial examples, generated by either adding a small amount of noise or by modifying a few bits of the audio signal, can be used to attack ASR systems to generate a completely different output [2, 3], even though the changes in the audio signal cannot be perceived by humans. Similar to these artificial perturbations in the audio signal, natural perturbations in human speech may also have an adverse effect on the performance of ASR systems. Natural perturbations in speech can arise due to the psychological and physical state of the speaker. Examples of naturally perturbed speech include expressive speech containing different emotions such as laughter, excitement, frustration, etc. and speech generated with different voice qualities such as creakiness, breath, etc.

In this paper, we show that the performance of the state-of-the-art deep neural network based ASR systems can significantly degrade for speech colored either by emotion or voice-quality. We show that these natural perturbations can be handled by Cycle-consistent GANs (CycleGANs) [4], a variant of Generative Adversarial Networks (GANs) [5] which can learn distributions of data across different domains even without a parallel corpus. The generator from our CycleGAN model learns to filter out the natural perturbations in speech and hence can be used as a front-end processor to improve the robustness of ASR to natural perturbations. Interestingly, in absence of perturbations, the front-end processing does not affect the ASR performance. The main contributions of this work are (a) an analysis of the performance of state-of-the-art ASR systems on naturally perturbed laughter and creaky speech, (b) an approach to train a CycleGAN model to obtain a front-end for transforming perturbed speech into normal speech, and (c) an analysis of the proposed front-end and its effectiveness in improving performance of state-of-the-art ASR systems.

The rest of the paper is organized as follows. Section 2 is the related work followed by the detailed description of our CycleGAN model in Section 3. Experiments and results are presented in Section 4 followed by an analysis on the learned transformation in Section 5 and the conclusion in Section 6.

2. RELATED WORK

Previous work have analyzed the effect of emotional speech on ASR and shown significant degradation in the performance of GMM-HMM-based ASR systems [6, 7]. They proposed adaptation of the acoustic and language models of the ASR system to capture the variations exhibited by emotive speech, in order to improve the ASR performance. As opposed to model adaptation, we propose an approach based on transforming emotional speech to normal speech. Recently, emotive-to-neutral speech conversion has been achieved by modeling prosody-based features [8]. But this approach requires a parallel corpus (i.e., same utterance spoken in neutral and with emotion), which is very difficult to collect for spontaneous speech. Similarly, GMM-HMM-based systems have been considered for synthesizing creaky speech [9], but no previous work has considered the conversion of creaky to neutral speech due to lack of a parallel corpus of creaky and neutral speech.

We propose a parallel-data-free approach to transform speech perturbed with emotions and voice quality to normal speech, based on CycleGANs [4]. CycleGAN was earlier used for voice conversion without parallel-data [10]. Compared to [10], our approach provides a front-end processor which can add robustness to ASR on utterances perturbed with emotion and voice quality. While [11] presented the initial results of our approach, this paper presents the details of our CycleGAN model, the training loss functions and additional experimental results which further validate the performance of our approach.

3. PERTURBED SPEECH TO NORMAL SPEECH TRANSFORMATION WITH CYCLEGANS

GANs consist of two different networks i.e., a generator G and a discriminator D. Generator is used to generate the fake samples $G(z)$, that resemble a given data distribution X, by taking random sample z from a prior distribution p_z as input, and the discriminator is used to discriminate fake samples from real samples in the data X. Both, generator and discriminator are trained using an adversarial loss function [5]. GANs have achieved impressive results in image generation [12], image-to-image translation [13] and style transfer [14]. More recently, unpaired image-to-image translation was successfully learned by adopting a variant of GAN, called cycle-consistent adversarial networks [4, 15]. We adopt the concept of CycleGAN for performing the task of non-parallel speech-to-speech emotion conversion.

We use a CycleGAN to model the transformation of perturbed speech features ($x \in X$) to normal speech features ($y \in Y$). The CycleGAN model architecture, considered in this work, is motivated from [10]. A typical GAN tries to minimize the adversarial loss $\mathfrak{L}_{adv}(G_{X\to Y}(x), y)$ which measures how far is the generated data $G_{X\to Y}(x)$ from the target data y. In case of perturbed speech to normal speech transformation without parallel utterances, a typical GAN with only the adversarial loss may not be able to preserve the context information in the speech features. The CycleGAN model can handle this using a pair of GANs with two adversarial loss functions and an additional cycle consistency loss function.

The first adversarial loss, given as:

$$\mathfrak{L}_{adv}(G_{X\to Y}(x), y) \tag{1}$$

corresponds to the forward mapping, which is the transformation from the perturbed speech to normal speech. The second adversarial loss, given as:

$$\mathfrak{L}_{adv}(G_{Y\to X}(y), x) \tag{2}$$

corresponds to the inverse mapping, which transforms the normal speech back to the perturbed speech.

The cycle consistency loss given as:

$$\begin{aligned}\mathfrak{L}_{cyc} =& E_x \left\|G_{Y\to X}(G_{X\to Y}(x)) - x\right\|_1 \\ &+ E_y \left\|G_{X\to Y}(G_{Y\to X}(y)) - y\right\|_1\end{aligned} \tag{3}$$

helps to preserve the context information, by ensuring that normal speech can be reconstructed by the cascade of the forward and inverse mapping generators and perturbed speech can be reconstructed by the cascade of the inverse and forward mapping generators, respectively.

In addition to the above mentioned losses, we also included the identity-loss function [4], given as:

$$\mathfrak{L}_{id} = E_x \left\|G_{Y\to X}(x) - x\right\|_1 + E_y \left\|G_{X\to Y}(y) - y\right\|_1 \tag{4}$$

$\mathfrak{L}_{id}$ was originally used for color preservation and we found this loss to be crucial for maintaining the linguistic information during conversion of speech.

The complete loss function ($\mathfrak{L}$) of our CycleGAN model is given as:

$$\begin{aligned}\mathfrak{L} =& \mathfrak{L}_{adv}(G_{X\to Y}(x), y) + \mathfrak{L}_{adv}(G_{Y\to X}(y), x) \\ &+ \lambda_{cyc}\, \mathfrak{L}_{cyc} + \lambda_{id}\, \mathfrak{L}_{id}\end{aligned} \tag{5}$$

The cycle consistency loss $\mathfrak{L}_{cyc}$ is scaled with a trade-of parameter λ_{cyc} whereas the identity-loss $\mathfrak{L}_{id}$ is scaled with a trade-of parameter λ_{id}.

The generator and discriminator networks in the CycleGAN model consist of convolutional blocks. The generator network consists a total of 9 convolutional blocks. These include one stride-1 convolution block, one stride-2 convolution block, 5 residual blocks [16], one $\frac{1}{2}$-stride convolution block, and one stride-1 convolution block. All convolution layers are 1-dimensional to preserve the temporal structure [17]. Similar to [18], gated linear units, which achieved state-of-the-art performance in language and speech modeling, are used as an activation function in the convolutional layers. We also used the instance normalization, proposed for style-transfer in [14]. The discriminator network consists of 4 2-dimensional convolutional blocks. Gated linear units were used as the activation function for all the convolutional blocks. For the discriminator network, we use a 6×6 patch GAN [19, 20], which classifies whether each 6×6 patch is real or fake.

4. EXPERIMENTS AND RESULTS

We use two spontaneous speech datasets, namely, AMI meeting corpus [21] and Buckeye corpus of conversational speech [22] to analyze the effect of natural perturbations. Both these datasets consist of manual annotations and time-stamps for speech perturbed with emotions and voice-quality. From both these datasets, speech data comprising of 40 female speakers and 30 male speakers was considered for training gender-dependent CycleGAN models. We consider 210 utterances for each gender and for each class (i.e., normal speech, laughter speech and creaky speech). Out of these 210 utterances, 150 utterances are used for train and 60 utterances for test. It is to be noted that all these utterances are non-parallel. Each utterance is of 1-2 second in duration.

Table 1: ASR performance without front-end (no FE) and with front-end (FE). Numbers in parenthesis with ↓ denote reduction in the error rate.

		Google			IBM			ASpIRE		
		no FE	FE		no FE	FE		no FE	FE	
			MFBs	MFBs+APs		MFBs	MFBs+APs		MFBs	MFBs+APs
Laughter-Speech	%WER	38.4	30.9	**23.5** (14.9↓)	50.4	49.6	**42.4** (8.0↓)	53.5	45.1	**32.5** (21.0↓)
	%SER	91.8	79.6	**75.5** (16.3↓)	93.1	89.7	**89.7** (3.4↓)	93.1	91.4	**89.7** (3.4↓)
Creaky-Speech	%WER	27.4	22.9	**16.4** (11.0↓)	29.2	24.3	**21.3** (7.9↓)	32.2	30.2	**24.3** (7.9↓)
	%SER	86.1	77.8	**63.9** (22.2↓)	88.9	86.1	**86.1** (2.8↓)	94.4	91.7	**83.3** (11.1↓)

The WORLD vocoder system [23] is used to extract features from the speech signal. The speech signal is sampled at 16 kHz, and Mel filterbank (MFB) features, logarithmic fundamental frequency ($log\ F_0$) and aperiodic components (APs) are extracted within a window of length 20 msec for every 5 msec. 24-dimensional MFBs and 24-dimensional APs are modeled by the proposed CycleGAN architecture to convert the features extracted from the input perturbed speech into the features corresponding to normal speech. Previous work on speaker conversion [24, 10], have used only the spectral features (MFBs). But for perturbed speech conversion, we found that modeling both, spectral features (MFBs) and aperiodic components (APs) resulted in better conversion to normal speech than considering only spectral features (MFBs). Logarithm Gaussian normalized transformation [25] was used to convert the F_0 values from the source speech to those corresponding to the target speech.

In order to achieve a more stable training of the CycleGAN models and to generate higher quality outputs, we used the least square function to compute the adversarial loss instead of the commonly used negative log likelihood objective function [26, 4]. The CycleGAN models were trained using the Adam optimizer with a batch size of 1. The initial learning rates of the generator and the discriminator are 0.0002 and 0.0001, respectively. The learning rates were decayed by a factor of 10^5 after each epoch. In all the experiments, the cycle consistency loss trade-of parameter λ_{cyc} was set to a value of 10. The identity-loss trade-of parameter λ_{id} was set to 1 for the first 100 epochs and set to 0 after 100 epochs.

Table 1 presents the performance of Google cloud ASR[1], IBM ASR[2] and Kaldi ASR (with ASpIRE models) [27, 28] with and without our proposed front-end, when tested with laughter speech (speech perturbed with emotion) and creaky speech (speech perturbed with voice-quality). The performance is evaluated in terms of % Word Error Rate (%WER) and % Sentence Error Rate (%SER). Lower values of WER and SER indicate better performances. Table 1 shows that our proposed front-end improves the performance of each of the ASR systems. It can be observed that modeling both, spectral and aperiodic components (i.e., MFB + APs) performs better than modeling only MFBs in the proposed front-end.

Table 2: DeepSpeech model performance without front-end (no FE) and with front-end (FE) in terms of character error rate (%CER). Numbers in parenthesis with ↓ denote reduction in the error rate.

Perturbation	no FE	FE	
		MFBs	MFBs+APs
Laughter speech	56.5	53.0	**41.7** (14.8↓)
Creaky speech	33.5	29.8	**23.7** (9.8↓)

The ASR performances shown in Table 1 are influenced by the strength of the language model used by the respective ASR systems. To check ASR performance without the effect of a language model, we also present the results from the DeepSpeech model[3] which converts speech to a sequence of English characters. Table 2 shows the % Character Error Rate (%CER) performance of the DeepSpeech model with and without the proposed front-end. The DeepSpeech model was trained on 1000 hours of LibriSpeech data and did not use a language model for decoding. It can be observed from Table 2 that our proposed front-end gives significant reduction in CER of the DeepSpeech model.

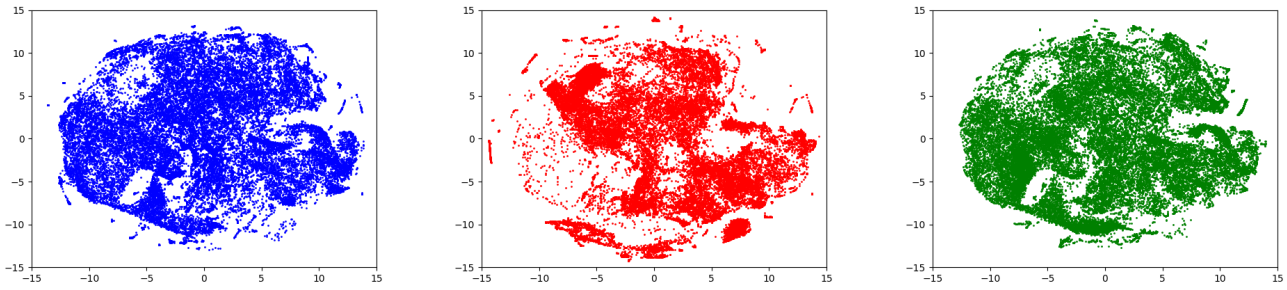
5. ANALYSIS OF THE LEARNED FRONT-END TRANSFORMATION

Figure 1 shows a 2-dimensional t-SNE projection [29] of the Mel filterbank features for (a) normal speech, (b) laughter perturbed speech [30] and (c) laughter perturbed speech transformed to normal speech by the proposed front-end. It can be observed that the filterbank features for normal speech and transformed (normal) speech are quite similar to each other and that they differ significantly from the filterbank features for laughter speech. Additionally, the spread of the filterbank features for laughter speech is reduced in the 2-dimensional t-SNE space. We hypothesize that this may be due to the reduction in vowel space for laughter speech [31].

[1]https://cloud.google.com/speech-to-text/

[2]https://www.ibm.com/watson/services/speech-to-text/

[3]https://github.com/mozilla/DeepSpeech/releases/tag/v0.4.0-alpha.3



(a) Normal Speech (b) Laughter Perturbed Speech (c) Transformed (Normal) Speech

Fig. 1: t-SNE projection of Mel filterbank output features (Best viewed in color).

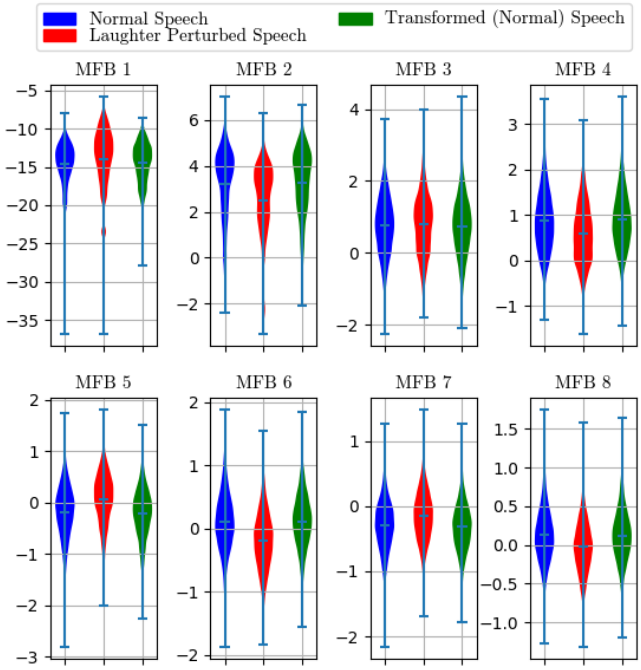


Fig. 2: Violin plot of output from filters 1 to 8 of the Mel filterbank (Best viewed in color).

For a more detailed analysis, Figure 2 shows violin plots [32] of the output of the filters 1 to 8 of the Mel filterbank, for normal speech, laughter perturbed speech and laughter perturbed speech transformed to normal speech. Output of the filters 9 to 24 do not show visible differences and hence they are not shown. It can be observed from Figure 2 that the distribution of the feature values for normal speech and transformed (normal) speech are similar and they exhibit similar variations. It implies that the front-end is able to (a) capture the distribution of the Mel filterbank outputs of both normal and laughter perturbed speech, and (b) transform laughter perturbed speech to equivalent normal speech.

6. CONCLUSION

We proposed a novel front-end based on CycleGANs to transform naturally perturbed speech to normal speech. Experiments on spontaneous laughter speech and creaky voice utterances show significant improvements in performance of the Google ASR, IBM ASR, the Kaldi ASR with ASpIRE model and that of a DeepSpeech model. We found that adding aperiodic components to spectral features gives a better performance. Visualization of the laughter speech features and the transformed speech features gives insights on the transformation performed by our proposed front-end.

7. REFERENCES

[1] Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke, "The microsoft 2017 conversational speech recognition system," in *IEEE ICASSP*, 2018.

[2] Dan Iter, Jade Huang, and Mike Jermann, "Generating adversarial examples for speech recognition," *Stanford Technical Report*, 2017.

[3] Nicholas Carlini and David Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," *arXiv:1801.01944*, 2018.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[6] Theologos Athanaselis, Stelios Bakamidis, Ioannis Dologlou, Roddy Cowie, Ellen Douglas-Cowie, and Cate Cox, "ASR for emotional speech: Clarifying the issues

and enhancing performance," *Neural Networks*, vol. 18, no. 4, 2005.

[7] Bogdan Vlasenko, Dmytro Prylipko, and Andreas Wendemuth, "Towards robust spontaneous speech recognition with emotional speech adapted acoustic models," in *KI*, 2012.

[8] V Raju, P Gangamohan, Suryakanth V Gangashetty, and Anil Kumar Vuppala, "Application of prosody modification for speech recognition in different emotion conditions," in *IEEE Tencon*, 2016.

[9] NP Narendra and K Sreenivasa Rao, "Generation of creaky voice for improving the quality of HMM-based speech synthesis," *Computer Speech & Language*, vol. 42, 2017.

[10] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv:1711.11293*, 2017.

[11] Sri Harsha Dumpala, Imran Sheikh, Rupayan Chakraborty, and Sunil Kumar Kopparapu, "Cycle-consistent gan front-end to improve asr robustness to perturbed speech," in *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language*, 2018.

[12] Emily L Denton, Soumith Chintala, Rob Fergus, et al., "Deep generative image models using a laplacian pyramid of adversarial networks," in *NIPS*, 2015.

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

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*. Springer, 2016.

[15] Yaniv Taigman, Adam Polyak, and Lior Wolf, "Unsupervised cross-domain image generation," in *ICLR*, 2016.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

[17] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *INTERSPEECH*, 2017.

[18] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," in *ICML*, 2017.

[19] Christian Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[20] Chuan Li and Michael Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *ECCV*. Springer, 2016.

[21] Jean Carletta et al., "The AMI meeting corpus: A pre-announcement," in *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, 2006.

[22] Mark A Pitt, Laura Dilley, et al., "Buckeye corpus of conversational speech," *Ohio State University*, 2007.

[23] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. on Information and Systems*, vol. 99, no. 7, 2016.

[24] Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Maximum likelihood voice conversion based on gmm with straight mixed excitation," in *Conference on Spoken Language Processing*, 2006.

[25] Kun Liu, Jianping Zhang, and Yonghong Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin," in *IEEE FSKD*, 2007.

[26] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *ICCV*. IEEE, 2017.

[27] Vijayaditya Peddinti, Guoguo Chen, Vimal Manohar, Tom Ko, Daniel Povey, and Sanjeev Khudanpur, "JHU ASpIRE system: Robust LVCSR with TDNNs, ivector adaptation and RNN-LMS," in *IEEE ASRU*, 2015.

[28] Daniel Povey, "Kaldi models," http://kaldi-asr.org/models.html, Oct 2017.

[29] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *JMLR*, vol. 9, 2008.

[30] Sri Harsha Dumpala, Karthik Venkat Sridaran, Suryakanth V Gangashetty, and B Yegnanarayana, "Analysis of laughter and speech-laugh signals using excitation source information," in *IEEE ICASSP*, 2014.

[31] Jo-Anne Bachorowski, Moria J Smoski, and Michael J Owren, "The acoustic features of human laughter," *JASA*, vol. 110, no. 3, 2001.

[32] Jerry L. Hintze and Ray D. Nelson, "Violin plots: A box plot-density trace synergism," *The American Statistician*, vol. 52, no. 2, 1998.