Cycle-Consistent GAN Front-End to Improve ASR Robustness to Perturbed Speech

Sri Harsha Dumpala, Imran Sheikh, Rupayan Chakraborty, Sunil Kumar Kopparapu
TCS Research and Innovation - Mumbai, INDIA
{d.harsha, imran.as, rupayan.chakraborty, sunilkumar.kopparapu}@tcs.com

Abstract

Automatic Speech Recognition (ASR) systems, which perform well on regular speech, are found to be vulnerable to adversarial examples generated by small perturbations in the audio signal. Even naturally introduced perturbations in audio signal, caused by emotional and physical states of the speaker, can significantly degrade ASR performance. In this paper, we propose a front-end based on Cycle-Consistent Generative Adversarial Network (CycleGAN) to reduce the perturbations, and hence add robustness to ASR performance. CycleGAN is trained using non-parallel examples of perturbed and normal speech. Experiments on spontaneously generated laughter-speech and creaky voice datasets tested with Google cloud ASR show absolute improvements in WER of 14.9% and 11%, respectively, on speech converted using the CycleGAN based front-end as compared to the original perturbed speech.

1 Introduction

The advent of powerful deep learning techniques in the recent past have resulted in significant improvements in the performance of Automatic Speech Recognition (ASR) systems [1, 2, 3]. In some scenarios, ASR systems are performing at par with human-level accuracy. Further, the advent of voice assistants such as Google Home, Amazon Echo etc., have led to the wide use of ASR systems in various day-to-day applications such as voice-based search [4], home automation [5] and elderly care [6].

However, recent studies have shown that the 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 the ASR systems to generate a completely different output [7, 8, 9], even though the changes in the audio signal cannot be perceived by humans. If small artificial perturbations in the audio signal can affect the performance of ASR systems so significantly [10], natural perturbations in human speech may also have an adverse effect. Naturally perturbed speech can arise due to the psychological and physical state of the speaker. For instance, expressive speech such as excited [11, 12], frustrated etc., and speech generated with different voice qualities such as creaky, breathy, etc [13, 14, 15].

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 by either emotion or voice-quality. We show that these natural perturbations can be modeled using Cycle-consistent GANs (CycleGANs) [16, 17, 18], a variant of Generative Adversarial Networks (GANs) [19] which can learn distributions of data across different domains without a parallel corpus. Generator from our CycleGAN model has the ability to filter out the natural perturbations in speech and hence can be used as a front-end processor to improve the robustness of ASR systems. This front-end processing does not affect ASR performance in absence of the perturbations. The main contributions of this work are:

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

- Analyze the performance of state-of-the-art ASR systems tested with naturally perturbed speech, including laughter and creaky speech.
- CycleGAN based front-end to convert perturbed speech into normal speech.
- Analyze the CycleGAN front-end transformation and its effectiveness in ASR performance.

2 Related Work

Very few studies have analyzed the effect of emotional coloring of speech on ASR performance [20, 21, 22]. These studies have shown significant degradation in the performance of GMM-HMM based ASR systems, when tested with emotive speech. Moreover, most of these approaches are based on modifying the acoustic and language models of the ASR system to handle the variations exhibited by emotive speech. Recently, emotive speech utterances were converted to the neutral speech utterances by modeling prosody-based features [23]. But these approaches require a parallel data corpus (i.e., same utterance spoken in neutral and with emotion), which is very difficult to collect for spontaneous speech.

Analysis of the effect of voice qualities on ASR performance is not much explored. Only a few studies have considered the detection of creaky voice in spontaneous speech [13, 24]. Further, GMM-HMM-based systems were considered for synthesizing creaky speech [25, 26], but no previous works have considered the conversion of creaky to neutral speech. One of the main issue in the conversion of creaky to neutral speech is the lack of parallel speech corpus for creaky and neutral speech.

As compared to these works, we propose a CycleGAN-based [16] approach to transform speech perturbed with emotions and voice quality to normal speech. GANs were initially proposed for the generation of images when provided with some arbitrary random noise as input, and thereafter have achieved impressive results in image generation [27, 28], image-to-image translation [29], style transfer [30] and text-to-image synthesis [31, 32]. More recently, unpaired image-to-image translation was successfully learned by adopting a variant of GAN, called cycle-consistent adversarial networks [16, 17, 18] with an identity-mapping loss [33]. We adopt the concept of CycleGAN for performing the task of non-parallel speech-to-speech translation for emotion conversion. CycleGAN was earlier used for parallel-data-free speaker voice conversion [34] and speech enhancement [35]. To the best of our knowledge there is no prior work on converting speech perturbed with emotion or voice quality to normal speech, by considering non-parallel data. We are the first to use CycleGAN for perturbed to normal speech conversion. Moreover, our approach provides a front-end processor which can add robustness to speech recognition.

3 Cyclic-GANS

GANs consist of two different networks i.e., a generator G and a discriminator D [19]. The generator is used to generate fake samples $G(z)$, that resemble a given data distribution X, by taking a random sample z from a prior distribution p_z as input. The discriminator is used to discriminate fake samples from real samples in the data X. Both, generator and discriminator are trained adversarially such that the generator learns to generate samples which resemble the original samples in the data by taking a feedback from the discriminator. The discrimintor itself gets better at discriminating the samples generated by the generator from the original samples.

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. But for applications with no parallel data, particularly in speech, a typical GAN with only the adversarial loss may not be able to preserve the context information in the speech features. In order to learn the transformation using only non-parallel data, we use Cycle-GAN architecture [16]. 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 $\mathfrak{L}adv(G_{X\to Y}(x), y)$ corresponds to the forward mapping, the second adversarial loss $\mathfrak{L}adv(G_{Y\to X}(y), x)$ corresponds to the inverse mapping and the cycle consistency loss given as:

$$\begin{aligned}\mathfrak{L}cyc =& E_x \left|G_{Y\to X}(G_{X\to Y}(x)) - x\right|_1 \\ &+ E_y \left|G_{X\to Y}(G_{Y\to X}(y)) - y\right|_1 \end{aligned} \tag{1}$$

helps to preserve the context information. The cycle consistency loss $\mathfrak{L}cyc$ is scaled with a trade-of parameter λcyc.

Table 1: ASR performance without front-end (W/O FE) and with front-end (With FE).

Perturbation		Google			ASpIRE		
		W/O FE	With FE		W/O FE	With FE	
			MFBs	MFBs+APs		MFBs	MFBs+APs
Laughter speech	%WER	38.4	30.9	**23.5**	53.5	45.1	**32.5**
	%SER	91.8	79.6	**75.5**	93.1	91.4	**89.7**
Creaky speech	%WER	27.4	22.9	**16.4**	32.2	30.2	**24.3**
	%SER	86.1	77.8	**63.9**	94.4	91.7	**83.3**

3.1 CycleGAN-based Speech-to-Speech Conversion

Our CycleGAN model architecture, considered in this work, is motivated from [34]. The block diagram of the architecture is presented as an appendix to the paper. All convolution layers are 1-dimensional to preserve the temporal structure [36]. Similar to [37], 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 instance normalization, proposed for style-transfer in [30]. For the discriminator network, we use a 6×6 patch GAN [38, 39], which classifies whether each 6×6 patch is real or fake.

Training Details: In order to achieve more stable training and to generate higher quality outputs, we used the least square loss to compute the discriminator loss in place of the negative log likelihood objective [40, 16]. We also considered the identity-loss function [16], originally used for color preservation, which we found to be crucial for maintaining the linguistic information during conversion of speech. We trained the CycleGAN models using the Adam optimizer [41] 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 for each epoch.

4 Experiments and Results

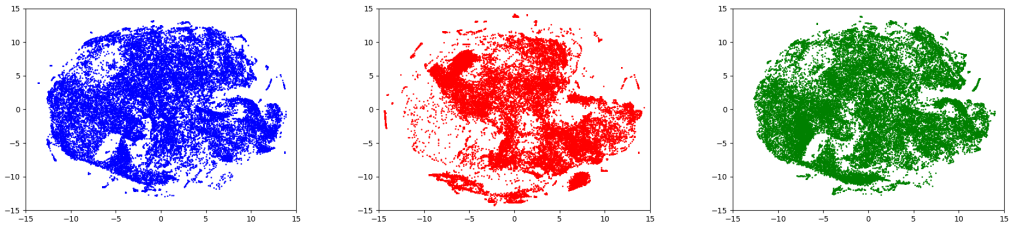
Analysis is performed by considering two spontaneous speech datasets, namely, AMI meeting corpus [42] and Buckeye corpus of conversational speech [43]. Both datasets consists of dedicated annotations along with time-stamps for speech perturbed with emotions and voice-quality. In this work, speech data collected from 40 female and 30 male speakers, in total from both datasets was considered for training gender-dependent models. For each gender and for each class (i.e., normal, laughter speech and creaky speech), 210 utterances (150 utterances for train and 60 utterances for test) were considered. It is to be noted that all these utterances are non-parallel. Each utterances is of 1-2 sec in duration.

The WORLD vocoder system [44] is used to extract features from the speech signal. The speech signals are sampled to 16 kHz, and then Mel filterbank (MFB) features, logarithmic fundamental frequency ($log\ F_0$) and aperiodic components (APs) [45] are extracted from the speech signal 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 features corresponding to normal speech. In previous works for speaker conversion [46, 34], only the spectral features (MFBs) were modeled. 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 [47] was used to convert the F_0 values from the source to the target speech.

Table 1 shows the performance of Google cloud ASR and Kaldi ASR (with ASpIRE models) [1, 48] with and without our proposed front-end system, 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 these values, better are the performances. It can be observed from Table 1 that modeling both, spectral and aperiodic components (i.e., MFB + APs) performs better than modeling only MFBs with the proposed

front-end. It can be observed that an absolute reduction of 14.9%, 21.0% in WER and 16.3% , 3.4% in SER, is achieved for Google ASR and ASpIRE ASR, respectively, when our proposed front-end (MFBs+APs) is used to convert laughter speech to normal speech. Similarly, an absolute reduction of 11.0%, 7.9% in WER and 22.2% , 11.1% in SER is obtained for Google ASR and ASpIRE ASR, respectively, when our proposed front-end (MFBs+APs) is used to convert creaky speech to normal speech. A demo and the source codes are made available at `https://goo.gl/dEmKcz`.

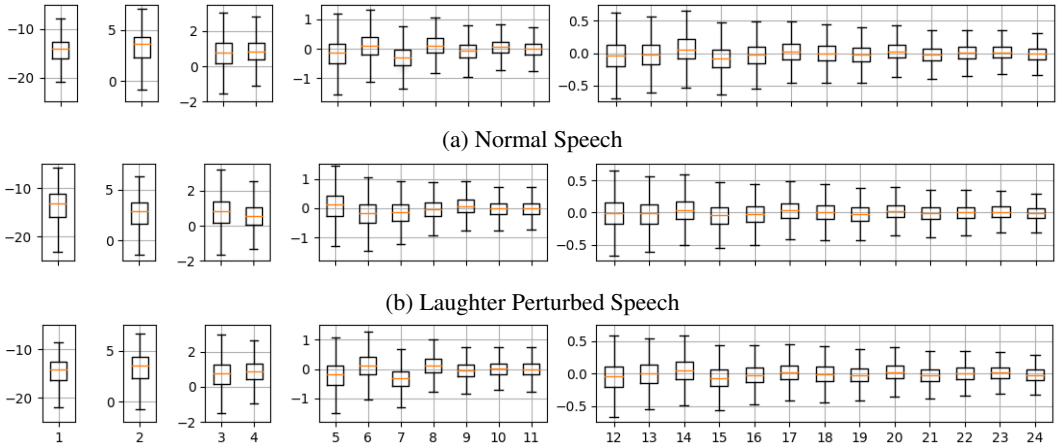
5 Analysis of the Learned Front-End Transformation



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

Figure 1: t-SNE projection of Mel filterbank output features.

Figure 1 shows a 2-dimensional t-SNE projection [49] of the Mel filterbank features for (a) normal speech, (b) laughter perturbed speech 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 normal transformed 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 of laughter speech [50].



(a) Normal Speech

(b) Laughter Perturbed Speech

(c) Normal Transformed Speech

Figure 2: Box plot of output from filters 1 to 24 of the Mel filterbank.

For a more detailed analysis, Figure 2 shows a box plot for each of the 24 Mel filterbank output features, for (a) normal speech, (b) laughter perturbed speech and (c) laughter perturbed speech transformed to normal speech. It can be observed that the feature values for normal speech and normal transformed speech are very close and they exhibit similar variations. Interestingly, for filters 1 to 8 they differ significantly from the laughter speech. We hypothesize that our front-end transformation mainly operates in this region.

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 and Kaldi ASR with ASpIRE model. We found that adding aperiodic components to spectral features gives a better performance. Visualization of the laughter speech features and the converted speech features gives insights on the transformation performed by our proposed front-end.

References

[1] 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 *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pages 539–546. IEEE, 2015.

[2] Gakuto Kurata, Bhuvana Ramabhadran, George Saon, and Abhinav Sethy. Language modeling with highway lstm. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 244–251. IEEE, 2017.

[3] Wayne Xiong, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, and Andreas Stolcke. The microsoft 2017 conversational speech recognition system. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5934–5938. IEEE, 2018.

[4] Yanzhang He, Rohit Prabhavalkar, Kanishka Rao, Wei Li, Anton Bakhtin, and Ian McGraw. Streaming small-footprint keyword spotting using sequence-to-sequence models. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 474–481. IEEE, 2017.

[5] 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. *Proc. INTERSPEECH. ISCA*, 2017.

[6] Kazunari Tamamizu, Seiji Sakakibara, Sachio Saiki, Masahide Nakamura, and Kiyoshi Yasuda. Capturing activities of daily living for elderly at home based on environment change and speech dialog. In *International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management*, pages 183–194. Springer, 2017.

[7] Dan Iter, Jade Huang, and Mike Jermann. Generating adversarial examples for speech recognition. Stanford technical report at `https://web.stanford.edu/class/cs224s/reports/Dan_Iter.pdf`, 2017.

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

[9] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554*, 2018.

[10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[11] Sri Harsha Dumpala, Karthik Venkat Sridaran, Suryakanth V Gangashetty, and B Yegnanarayana. Analysis of laughter and speech-laugh signals using excitation source information. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 975–979. IEEE, 2014.

[12] Sri Harsha Dumpala, Ashish Panda, and Sunil Kumar Kopparapu. Analysis of the effect of speech-laugh on speaker recognition system. *Proc. Interspeech 2018*, pages 1751–1755, 2018.

[13] Thomas Drugman, John Kane, and Christer Gobl. Data-driven detection and analysis of the patterns of creaky voice. *Computer Speech & Language*, 28(5):1233–1253, 2014.

[14] Sri Harsha Dumpala and KNRK Raju Alluri. An algorithm for detection of breath sounds in spontaneous speech with application to speaker recognition. In *International Conference on Speech and Computer*, pages 98–108. Springer, 2017.

[15] Sri Harsha Dumpala and Sunil Kumar Kopparapu. Improved speaker recognition system for stressed speech using deep neural networks. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 1257–1264. IEEE, 2017.

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

[17] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2868–2876, 2017.

[18] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865, 2017.

[19] 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*, pages 2672–2680, 2014.

[20] 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*, 18(4):437–444, 2005.

[21] Bogdan Vlasenko, Dmytro Prylipko, and Andreas Wendemuth. Towards robust spontaneous speech recognition with emotional speech adapted acoustic models. In *35th German Conference on Artificial Intelligence (KI-2012), Saarbrücken, Germany (September 2012)*, pages 103–107. Citeseer, 2012.

[22] Kohei Mukaihara, Sakriani Sakti, and Satoshi Nakamura. Recognizing emotionally coloured dialogue speech using speaker-adapted dnn-cnn bottleneck features. In *International Conference on Speech and Computer*, pages 632–641. Springer, 2017.

[23] VV Vidyadhara Raju, P Gangamohan, Suryakanth V Gangashetty, and Anil kumar Vuppala. Application of prosody modification for speech recognition in different emotion conditions. In *Region 10 Conference (TENCON), 2016 IEEE*, pages 951–954. IEEE, 2016.

[24] NP Narendra and K Sreenivasa Rao. Automatic detection of creaky voice using epoch parameters. In *Interspeech*, pages 2347–2351, 2015.

[25] Tuomo Raitio, John Kane, Thomas Drugman, and Christer Gobl. Hmm-based synthesis of creaky voice. In *Interspeech*, pages 2316–2320, 2013.

[26] NP Narendra and K Sreenivasa Rao. Generation of creaky voice for improving the quality of hmm-based speech synthesis. *Computer Speech & Language*, 42:38–58, 2017.

[27] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[28] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[31] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1060–1069. JMLR, 2016.

[32] Han Zhang, Tao Xu, and Hongsheng Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916. IEEE, 2017.

[33] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *2017 International Conference on Learning Representations (ICLR)*, 2016.

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

[35] Zhong Meng, Jinyu Li, Yifan Gong, et al. Cycle-consistent speech enhancement. *arXiv preprint arXiv:1809.02253*, 2018.

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

[37] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International Conference on Machine Learning*, pages 933–941, 2017.

[38] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, volume 2, page 4, 2017.

[39] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

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

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

[42] Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100, 2005.

[43] Mark A Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. Buckeye corpus of conversational speech (2nd release)[www. buckeyecorpus. osu. edu] columbus, oh: Department of psychology. *Ohio State University (Distributor)*, 2007.

[44] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

[45] Hideki Kawahara, Jo Estill, and Osamu Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[46] Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Maximum likelihood voice conversion based on gmm with straight mixed excitation. 2006.

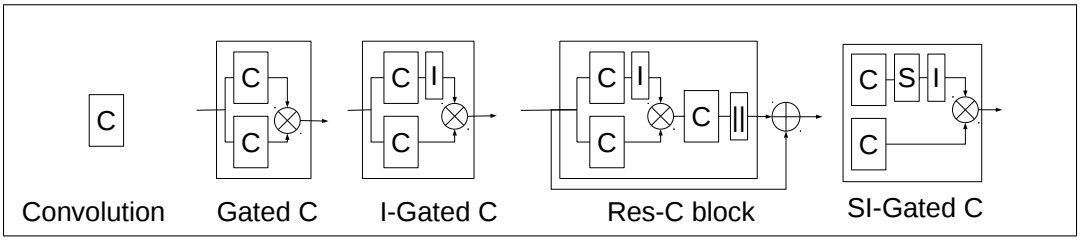
[47] Kun Liu, Jianping Zhang, and Yonghong Yan. High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin. In *Fuzzy Systems and Knowledge Discovery, FSKD. Fourth International Conference on*, volume 4, pages 410–414. IEEE, 2007.

[48] Daniel Povey. Kaldi models. http://kaldi-asr.org/models.html, last accessed: Nov. 2018.

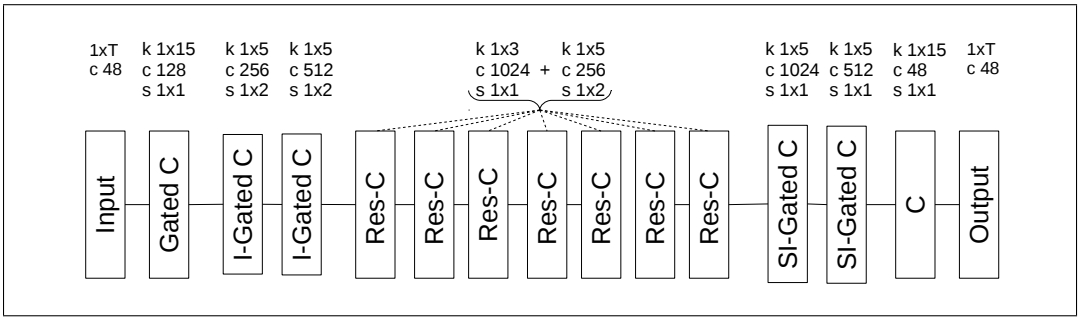
[49] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[50] Jo-Anne Bachorowski, Moria J Smoski, and Michael J Owren. The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, 110(3):1581–1597, 2001.

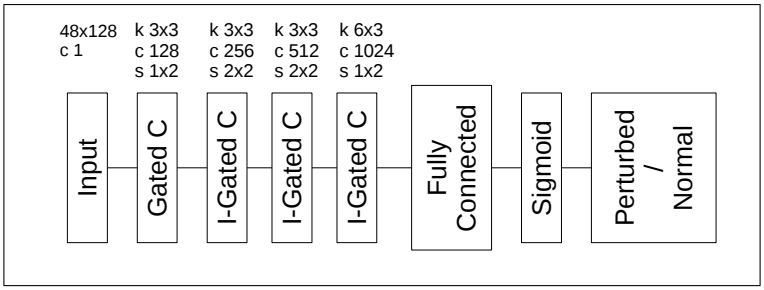
Appendix



(a) Blocks used in the generator and discriminator networks. Gated-C is gated convolution, I-Gated-C is instance normalized gated convolution, Res-C is residual convolution block and SI-Gated-C is pixel shuffled I-Gated-C.



(b) Generator block diagram. 'c' refers to channels, 'k' refers to convolution kernel size and 's' refers to stride. 'T' denotes the number of frames in the input.



(c) Discriminator block diagram. 'c' refers to channels, 'k' refers to convolution kernel size and 's' refers to stride.

Figure 3: Block diagram representation of the generator and the discriminator networks used in our CycleGAN architecture.