Improving Speech Intelligibility through Speaker Dependent and Independent Spectral Style Conversion

Tuan Dinh[1]*, Alexander Kain*[1]*, Kris Tjaden*[2]

[1]Oregon Health & Science University
[2]University at Buffalo

dintu@ohsu.edu kaina@ohsu.edu tjaden@buffalo.edu

Abstract

Increasing speech intelligibility for hearing-impaired listeners and normal-hearing listeners in noisy environments remains a challenging problem. Spectral style conversion from habitual to clear speech is a promising approach to address the problem. Motivated by the success of generative adversarial networks (GANs) in various applications of image and speech processing, we explore the potential of conditional GANs (cGANs) to learn the mapping from habitual speech to clear speech. We evaluated the performance of cGANs in three tasks: 1) speaker-dependent one-to-one mappings, 2) speaker-independent many-to-one mappings, and 3) speaker-independent many-to-many mappings. In the first task, cGANs outperformed a traditional deep neural network mapping in terms of average keyword recall accuracy and the number of speakers with improved intelligibility. In the second task, we significantly improved intelligibility of one of three speakers, without any source speaker training data. In the third and most challenging task, we improved keyword recall accuracy for two of three speakers, but without statistical significance.

Index Terms: intelligibility, voice conversion, style conversion, conditional generative adversarial networks, dysarthria

1. Introduction

There are approximately 28 million people in the United States who have some degree of hearing loss [1]. Understanding speech can be difficult for hearing-impaired listeners, and also for normal-hearing listeners in adverse environments. In an effort to increase the intelligibility of speech, researchers have studied noise suppression and cancellation techniques [2, 3, 4, 5]. Another approach is to alter the speech signal prior to presentation in a noisy environment; these techniques can be classified into several categories, including: utilizing audio and signal properties such as amplitude compression [6], dynamic range compression [7, 8], peak-to-rms reduction [9], and formant-enhancement [10]. Other techniques exploit the knowledge of a noise masker such as optimizations based on a speech intelligibility index [11] or glimpse proportion measure [12, 13]. An extensive evaluation, which was conducted over 26 speech modification techniques [14], showed that combining spectral shaping and dynamic range compression was able to boost intelligibility in terms of equivalent intensity changes by more than 5 dB gain over unmodified natural speech in some noise conditions. However, the spectral shaping only sharpened the peaks and increased spectral energy significantly in the 1–4 kHz band, but did not involve detailed spectral modification.

This material is based upon work supported by the National Institutes of Health under Grant R01DC004689.

There are also techniques that consider the intelligibility gains due to a *clear* (CLR) speaking style [15, 16, 17], inspired by the acoustic characteristics of CLR speech such as spectral flattening and vowel space expansion [18, 19]. Typically, CLR speech is highly-articulated, with a slower speaking rate, and more frequent pauses; the exact strategy varies from speaker-to-speaker. Previously, we used hybridization experiments to establish that speech intelligibility of *habitual* (HAB) speech can be increased when certain acoustic features from parallel CLR speech are incorporated [20, 21]. This suggests that it should be possible to automatically increase the intelligibility of speech by learning a mapping between HAB and CLR features, or *style conversion*. In our previous experiments, we used a speaker dependent deep neural network (DNN) mapping the parameters of a manifold vocoder for style conversion, significantly improving the speech intelligibility of a speaker with Parkinson's disease from 24% to 46% [22]. However, DNN mappings are still limited by over-smoothing of converted spectra, leading to muffed speech [23]. Recently, the generative adversarial network (GAN) [24] has been shown to be effective in addressing the over-smoothing problem in voice conversion [23] and speech synthesis [25, 26]. We can consider the HAB-to-CLR mapping as an image-to-image translation task, in which the image is a window of the time-frequency representation of speech. In image-to-image translation, a conditional GAN (cGAN) [27] proved to be effective in generating less blurry images by combining a traditional adversarial loss and a mean absolute reconstruction loss (or L1 loss). In this paper, we leverage the cGAN architecture for HAB-to-CLR style conversion in three cases: 1) speaker-dependent one-to-one mappings, 2) speaker-independent many-to-one mappings, and 3) speaker-independent many-to-many mappings. The first case is our effort to improve our previous results [22]. In Section 3, we report on the efficacy of a cGAN-based one-to-one style conversion. However, training a speaker-dependent mapping is not possible in real-world applications, because typically CLR speech (parallel or not) is not available for a new source speaker. Therefore, in Section 4, we investigated a many-to-one mapping where we mapped HAB speech of every speaker to the CLR speech of a single speaker with the best sentence-level intelligibility. The idea of many-to-one voice conversion for intelligibility improvement was previously investigated to transform speech from any speaker regardless of accent, prosody, and background noise into a canonical target speaker [28]; however, their model required an enormous amount of training data. An issue for the many-to-one mapping is not preserving the characteristics of source speakers; therefore, in Section 5, we investigated a many-to-many mapping where a HAB-to-CLR mapping was trained on all speakers' style pairs simultaneously.

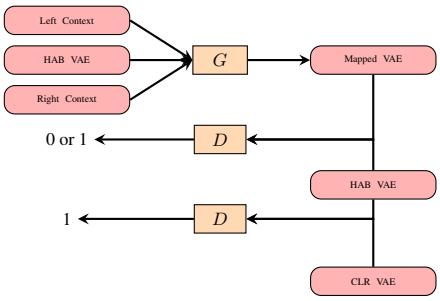

Figure 1: *cGAN framework for style conversion.*

2. Conditional Adversarial Network

Traditional GANs have a generative model or a generator (G) and a discriminative model or a discriminator (D), that together play a min-max game. Component G tries to fool component D by generating outputs close to the real data, while component D is trained to distinguish the output of component G from real data. Component G is a mapping function from random noise z to y, $G : \{z\} \rightarrow y$ [24]. In contrast, a cGAN model learns a mapping from an input x and random noise z to y, $G : \{x, z\} \rightarrow y$. The cGAN model has both G and D conditioned on input x [27], trained with the objective function $\mathcal{L}(D, G)$:

$$\min_G \max_D \mathcal{L}(D, G) = \quad (1)$$
$$\mathbb{E}_{x,y}\left[\log D(x, y)\right] + \mathbb{E}_{x,z}\left[\log(1 - D(x, G(x, z)))\right]$$

In our cGAN, we did not use random noise z, because using random noise input for generator G has proven ineffective [27, 29]. Instead, our generator mapped HAB speech features to aligned CLR speech features as shown in Figure 1. For the input vector of G, we added context by concatenating the current HAB frame in our VAE-12 representation [22] with five preceding and five following frames. We normalized the input and outputs of the network via standard scaling. The input of D consisted of either the output of G or aligned CLR feature frames, combined with the current HAB feature frame (what we wanted the output to be conditioned on). Thus, both G and D are conditioned on the current HAB feature. In addition to the adversarial loss function $\mathcal{L}(D, G)$ in Equation 1, we also minimized the L1 loss between the output of G and the ground truth; this addition was demonstrated to generate less blurry output compared to a root-mean-squared reconstruction loss [27]. We added the L1 loss with a scaling factor of 100 to $\mathcal{L}(D, G)$.

The structure of G is shown in Figure 2 [22]. By adding the input of G to the output of its last layer, we expected the network to focus on the difference between the HAB and CLR VAE-12 representations. The discriminator is a DNN with two hidden layers of 256 nodes each, and a single node output layer with sigmoidal activation function. To help stabilize the training process, we used 1) a leaky ReLU activation function with a slope of 0.2 for negative inputs for both G and D, 2) a dropout layer following each hidden layer of D with a dropout rate of 0.5, 3) the Adam optimizer with a batch size of 128, and 4) weights initialized from a zero-centered normal distribution with standard deviation 0.02 [30]. We used a momentum of 0.5, a learning rate decay of 0.00001, and learning rate of 0.0001 for D, and 0.0002 for G.

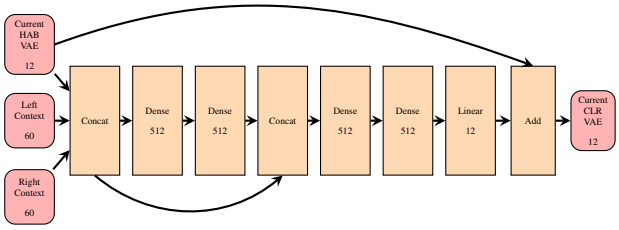

Figure 2: *Generator architecture.*

3. Experiment: One-to-One Mapping

3.1. Data

We used a database with 78 speakers consisting of control speakers (CS, N=32), speakers with multiple sclerosis (MS, N=30), and speakers with Parkinson's disease (PD, N=16) [31, 32]. All read the same 25 Harvard sentences in habitual and clear conditions (loud, slow, and fast conditions were also available). Speaker's names consisted of group, gender, and number, e. g. PDF7 was the seventh female speaker with PD.

3.2. Method

We applied our proposed cGAN conversion method to convert between two styles of a speaker; specifically, we aimed to convert the spectral aspects of a the HAB style to those of the CLR style, in an effort to improve the speech intelligibility of the former. For analysis and synthesis, we used a manifold vocoder [22]. The vocoder extracts fundamental frequency (F0), aperiodicity, and VAE-12 from each utterance.

We selected three speakers: CSM7, PDF7, and PDM6, who have been shown to benefit the most from the CLR spectrum [22]. We aligned each HAB utterance to its parallel CLR utterance of the same speaker using dynamic time warping (DTW) on 32$^{\text{nd}}$-order log filter bank features. Then, we pre-trained the generator that maps HAB VAE-12 to CLR VAE-12, minimizing a mean-squared-error loss. The pre-training stops when there is no progress in a validation set; the maximum number of epochs was 100 . Finally, we trained our proposed cGAN structure up to 300 epochs.

We created conversion stimuli using the mapped VAE-12, and F0 and aperiodicity information from the source HAB speech. To create the 25 conversion sentences, we used a leave-one-out approach, using 22 sentences for training and two for validation. Hybrid stimuli were created by replacing the HAB spectra with their aligned CLR spectra [20].

3.3. Objective Evaluation

We compared the performance of our proposed cGAN to our previous DNN [22]. Table 1 shows the average log spectral distortion (LSD) between mapped VAE-12 and CLR VAE-12. The cGAN mapping has typically smaller average LSD than its DNN counterpart. Specifically, Figure 3 shows the LSD of 25 test sentences from our two mappings. For most sentences, the LSD of the GAN mapping is lower than the LSD of the DNN mapping. Moreover, Figure 4 shows the variance ratio $\sigma^2_{\text{CLR}}/\sigma^2_{\text{MAP}}$ between CLR VAE-12 and mapped VAE-12 for each feature component. The smaller variance ratio of the cGAN mapping method suggests that the over-smoothing effect is reduced compared to the DNN-based method. Figure 5 shows a comparison of various spectrograms.

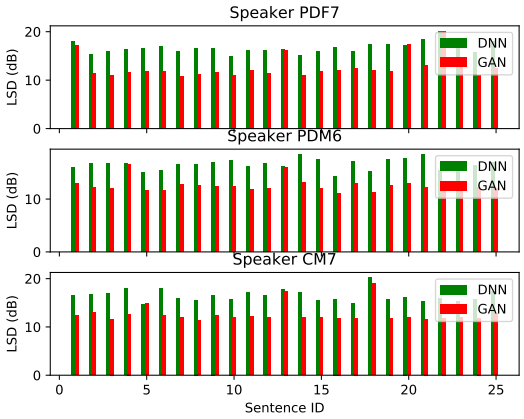

Figure 3: *Log spectral distortion (LSD) of 25 test sentences for three speakers.*

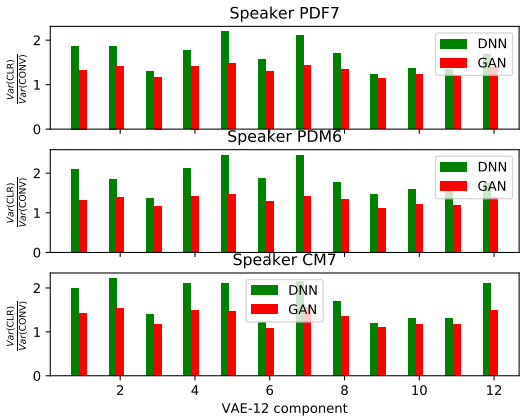

Figure 4: *Variance ratios between CLR VAE-12 (CLR) and mapped VAE-12 (MAP) features (smaller is better).*

3.4. Subjective Evaluation

LSD is not a good predictor for human perception; to evaluate speech intelligibility, we designed a test consisting of 25 sentences $\times$ 3 speakers (CSM7, PDF7, PDM6) $\times$ 5 conditions (2 purely vocoded, 1 hybrid, 2 mappings) = 375 unique trials in a Latin-square design. We performed the test on Amazon Mechanical Turk (AMT), where 60 participants listened to 25 Harvard utterances containing five keywords each. Listeners typed out each sentence as best as they could, and we calculated the average number of keywords correctly identified. The hybrid stimuli show an upper bound (or "oracle" mapping) on the intelligibility improvement. The vocoded HAB and vocoded CLR were obtained through analysis and resynthesis with unchanged parameters, using the manifold vocoder. We minimized the loudness differences between stimuli by normalizing gains in accordance with a RMSA measure. Finally, each utterance was mixed with babble noise at 0 dB SNR to avoid response saturation effects. Figure 6 shows average keyword accuracy. We observed that the cGAN mapping led to a statistically significant improvement ($p < 0.001$) for two speakers: PDM6 and CSM7, using a two-tailed t-test. In both cases, the cGAN mapping significantly outperformed the DNN-mapping, improving the intelligibility of two of three speakers, compared to our previous work where only one speaker improved [22].

mapping \ speakers	PDF7	PDM6	CSM7
DNN	16.8	16.67	16.44
GAN	12.85	12.58	12.67

Table 1: *Average LSD (in dB)*

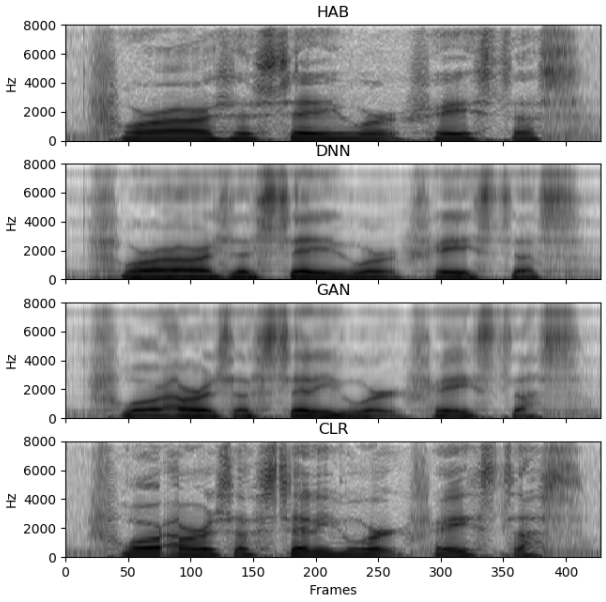

Figure 5: *Spectrogram of habitual speech (HAB), DNN mapping (DNN), cGAN mapping (GAN), and clear speech (CLR), from speaker PDF7, reading "Four hours of steady work faced us". Note the difference in formants between 2–4 kHz from the 50^{th}–100^{th} frame between the DNN and cGAN methods.*

4. Experiment: Many-to-One Mapping

A real-time application for speech intelligibility enhancement ideally does not require specific training on the source's speaker's speech, and can thus be considered (source) speaker-independent. Therefore, we studied a many-to-one mapping approach where we mapped HAB speech of every speaker to the CLR speech of a single speaker with the best sentence-level intelligibility. We used the same data as the previous experiment. The CLR speech of the two speakers CSM10 and CSF15 had highest sentence-level intelligibilities [31, 32] and were thus selected as target speech for the male and female case, respectively. We trained two gender-dependent mappings that mapped all HAB VAE-12 features of all male (or female) speakers (except one of three speaker CSM7, PDF7, PDM6) to CLR VAE-12 of CSM10 or CSF15, respectively. The HAB speech of the three speakers CSM7, PDF7, and PDM6 was used for testing. The mapped VAE-12, in combination with the original F0 and aperiodicity of the source speaker, were used to create conversion stimuli. Hybrid stimuli were created by replacing HAB spectra of the source speaker with aligned CLR spectra of the target speaker using hybridization.

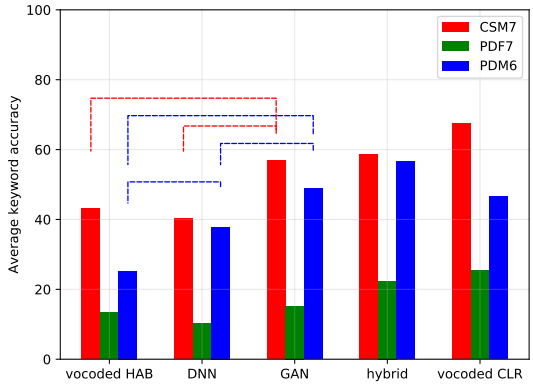

Figure 6: *Keyword recall accuracy of three speakers. The dashed lines show statistically significant differences.*

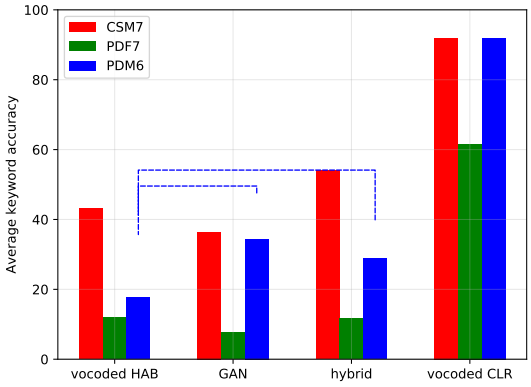

Figure 7: *Keyword recall accuracy of three speakers. The 'vocoded CLR' condition denotes clear speech of target speakers CSM10 and CSF15 for male and female cases, respectively. The dashed lines show statistically significant differences.*

4.1. Objective Evaluation

The average LSD between mapped and CLR spectrograms, with LSD between the input HAB and CLR spectrograms in parentheses, were: 17.32 (21.57) dB for CSM7, 22.7 (27.62) dB for PDF7, and 18.8 (23.28) dB for PDM6, confirming that the mapped speech is closer to CLR speech than input HAB speech.

4.2. Subjective Evaluation

To evaluate the efficacy of the method in terms of intelligibility, we designed a test consisting of 25 sentences $\times$ 3 source speakers (CSM7, PDF7, PDM6) $\times$ 3 conditions (vocoded HAB, cGAN-mapping, hybrid) + 25 sentences $\times$ 2 target speakers (CSM10, CSF15) $\times$ 1 condition (vocoded CLR) = 275 unique trials. We conducted the listening experiment similarly to the previous one, except the number of listeners was 44. Figure 7 shows the resulting keyword accuracy. We found that our many-to-one style conversion significantly improved the intelligibility of one speaker of three test speakers from 17.6% to 34.4%, using a two-tailed t-test ($p < 0.01$) , while there is no improvement in other cases.

5. Experiment: Many-to-Many Mapping

The disadvantage of the previous many-to-one mapping is that speaker characteristics cannot be preserved. Thus, we investigate the most realistic scenario of a many-to-many mapping that aims to learn solely the style differences, while preserving the linguistic message and speaker characteristics. This task is the hardest among our three experiments because not all speakers' spectral changes across styles have been shown to benefit speech intelligibility [22]. We used the same data as the previous experiments. We aligned each HAB utterance to its parallel CLR utterance of the same speaker using DTW on 32$^{\text{nd}}$-order log filter-bank features. Then we trained all one-to-one mappings from HAB VAE-12 to CLR VAE-12 simultaneously.

	CSM7	PDF7	PDM6
vocoded HAB	36.8	10	28.8
GAN	39.6	15.6	26.8
hybrid	62	22.8	57.6
vocoded CLR	66.8	22.4	48

Table 2: *Average keyword accuracy*

5.1. Objective Evaluation

The average LSD between the cGAN mapping and CLR spectrograms for the three test speakers, with the LSD between HAB and CLR spectrograms in parentheses, are: 16.36 (17.0) dB for CSM7, 16.66 (17.53) dB for PDF7, and 16.42 (18.06) dB for PDM6, confirming that the mapped speech is closer to CLR speech than the input HAB speech.

5.2. Subjective Evaluation

We designed a test consisting of 25 sentences $\times$ 3 speakers (CSM7, PDF7, PDM6) $\times$ 4 conditions (vocoded HAB, GAN, hybrid, vocoded CLR) = 300 unique trials. We conducted the experiment similarly to the previous ones, except the number of listeners was 24. Table 2 shows average keyword recall accuracy. The cGAN resulted in improvements for two speakers: CSM7 and PDF7. However, the results were not statistically significant using a two-tailed t-test ($p > 0.05$).

6. Conclusion

We explored a cGAN architecture for spectral style conversion in increasingly challenging experiments. In the speaker-dependent one-to-one mapping case, we showed that the cGAN outperformed a DNN in terms of average keyword recall accuracy in all cases. Moreover, the cGAN significantly improved speech intelligibility of two of three speakers, compared to only one speaker when using the DNN. In the speaker-independent many-to-one mapping case, we significantly improved speech intelligibility of one of three speakers, with average keyword recall accuracy increasing from 17.6% to 34.4%. In the speaker-independent many-to-many mapping case, the cGAN improved average keyword accuracy over that of vocoded HAB speech for the two speakers CSM7 and PDF7, but without statistical significance. While these are modest results, they show promise for developing automatic speaker-independent speech-intelligibility increasing approaches, especially given the small dataset and the fact that we did not attempt to transform additional acoustic features, such as phoneme durations.

7. References

[1] K. Cruickshanks, T. Wiley, B. Tweed, B. Klein, R. Klein, J. Mares-Perlman, and D. Nondahl, "The prevalence of hearing loss in older adults," *Am. J. Epidemiol.*, vol. 148, pp. 879–885, 1998.

[2] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, September 2005.

[3] J. Chen and S. D. J. Benesty, Y. Huang, "New insights into the noise reduction wiener filter," *IEEE Trans. Audio Speech Lang. Process*, vol. 14, no. 1, pp. 1218–1234, 2006.

[4] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligbility in noise for normal hearning listerners," *Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2008.

[5] D. Williamson, Y. Wang, and D. Wang, "Estimating nonnegative matrix model activations with deep neural networks to increase perceptual speech quality," *Journal of the Acoustical Society of America*, vol. 138, no. 3, pp. 1299–1407, 2015.

[6] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. Audio Speech Lang. Process*, vol. 24, pp. 277–282, 1976.

[7] B. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 1, pp. 22–32, 1969.

[8] D. Bonardo and E. Zovato, "Speech synthesis enhancement in noisy environment," *Proceedings of INTERSPEECH*, pp. 2853–2856, 2007.

[9] T. Quatieri and R. McAulay, "Peak-to-rms reduction of speech based on a sinusoidal model," *IEEE Trans. on Audio and Electroacoustics*, vol. 39, pp. 273–288, 1991.

[10] H. Brouckxon, W. Verhelst, and B. Schuymer, "Time and frequency depedent amplification for speech intelligibility enhancement in noisy environments," *Proceedings of INTERSPEECH*, pp. 557–560, 2008.

[11] B. Sauert and P. Vary, "Near end listening enhancement: speech intelligibility improvement in noisy environments," pp. 493–496, 2006.

[12] T. Takeuchi and Y. Tatekura, "Speech intelligibility enhancement in noisy enviroment via voice conversion with glimpse proportion measure," *Proceeding of APSIPA ASC*, pp. 1713–1717, 2018.

[13] Y. Tang and M. Cooke, "Learning static spectral weightings for speech intelligibility enhancement in noise," *Computer Speech & Language*, vol. 49, pp. 1–16, 2018.

[14] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligbility-enhancing speech modificantions: the hurricane challenge," *Proceedings of INTERSPEECH*, pp. 3552–3556, 2013.

[15] M. Picheny, N. Durlach, and L. Braida, "Speaking clearly for the hard of hearing i: Intelligibility differences between clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 28, pp. 96–103, 1985.

[16] S. H. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 112, pp. 259–271, 2002.

[17] S. Ferguson, "Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 116, pp. 2365–2373, 2004.

[18] A. Amano-Kusumoto and J. Hosom, "A review of research on speech intelligibility and correlations with acoustic features," *Center for Spoken Language Understanding, Oregon Health and Science University (Technical Report CSLU-011-0001)*, 2011.

[19] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles," *Computer Speech & Lang.*, vol. 28, no. 2, pp. 629–647, 2014.

[20] A. Kain, A. Amano-Kusumoto, and J.-P. Hosom, "Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility," *Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2308–2319, 2008.

[21] K. Tjaden, A. Kain, and J. Lam, "Hybridizing conversational and clear speech to investigate the source of increased intelligibility in speakers with parkinson's disease," *Journal of Speech, language, and hearing research*, vol. 57, pp. 1191–1205, 2014.

[22] T. Dinh, A. Kain, and K. Tjaden, "Using a manifold vocoder for spectral voice and style conversion," *Proceedings of INTERSPEECH*, pp. 1388–1392, 2019.

[23] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," *Proceedings of INTERSPEECH*, 2017.

[24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proceedings of the NIPS*, vol. 2, pp. 2672–2680, 2014.

[25] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashiro, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," *Proceedings of ICASSP*, 2017.

[26] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Trans. on Audio, Speech and Lang. Process*, vol. 26, no. 1, pp. 84–96, 2017.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proceedings of CVPR*, pp. 5967–5976, 2016.

[28] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," *Proceedings of INTERSPEECH*, 2019.

[29] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *Proceedings of INTERSPEECH*, pp. 2008–2012, 2017.

[30] S. Chintala, "How to train a gan?" https://github.com/soumith/ganhacks, 2016.

[31] K. Tjaden, J. Lam, and G. E. Wilding, "Vowel acoustics in parkinson's disease and multiple sclerosis: comparison of clear, loud, and slow speaking conditions," *Journal of Speech, language, and hearing research*, vol. 56, pp. 1485–1502, 2013.

[32] K. Tjaden, J. E. Sussman, and G. E. Wilding, "Impact of clear, loud, and slow speech on scaled intelligiblity and speech serverity in parkinson's disease and multiple sclerosis," *Journal of Speech, language, and hearing research*, vol. 57, pp. 779–792, 2014.