INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Augmented CycleGANs for continuous scale normal-to-Lombard speaking style conversion

Shreyas Seshadri[1], *Lauri Juvela*[1], *Paavo Alku*[1], *Okko Räsänen*[2,1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Unit of Computing Sciences, Tampere University, Finland
firstname.surname@aalto.fi, firstname.surname@tuni.fi

Abstract

Lombard speech is a speaking style associated with increased vocal effort that is naturally used by humans to improve intelligibility in the presence of noise. It is hence desirable to have a system capable of converting speech from normal to Lombard style. Moreover, it would be useful if one could adjust the degree of Lombardness in the converted speech so that the system is more adaptable to different noise environments. In this study, we propose the use of recently developed Augmented cycle-consistent adversarial networks (Augmented CycleGANs) for conversion between normal and Lombard speaking styles. The proposed system gives a smooth control on the degree of Lombardness of the mapped utterances by traversing through different points in the latent space of the trained model. We utilize a parametric approach that uses the Pulse Model in Log domain (PML) vocoder to extract features from normal speech that are then mapped to Lombard-style features using the Augmented CycleGAN. Finally, the mapped features are converted to Lombard speech with PML. The model is trained on multi-language data recorded in different noise conditions, and we compare its effectiveness to a previously proposed CycleGAN system in experiments for intelligibility and quality of mapped speech.
Index Terms: Augmented CycleGAN, style conversion, Lombard speech, vocal effort, pulse-model in log domain vocoder

1. Introduction

Lombard speech [1] refers to a speaking style with increased vocal effort that is automatically utilized by speakers to enhance the intelligibility of speech in noisy environments. It would be beneficial to have speech technology capable of converting speech from normal style to Lombard style in a similar manner as humans do. The technology of converting speech of one style to another, while retaining the linguistic and speaker-specific information of the original speech signal, is called speaking style conversion (SSC). SSC can be regarded as a distinct area of speech technology, but it is related to other fields such as speech intelligibility enhancement in speech transmission [2]. Strict latency requirements imposed by real-time speech transmission, however, are not necessarily present in SSC, where offline processing is also possible for several potential use scenarios.

SSC has been previously studied in whisper-to-normal conversion [3–5] and in normal-to-Lombard conversion [6–8]. In addition, a parametric approach to normal-to-Lombard SSC was recently explored in [9], where a vocoder was used to extract frame level features that were then transformed from normal to Lombard style using parallel data-driven mapping models, and then synthesized as speech in the target style using the same vocoder. The proposed vocoder-based SSC methodology was extended in [10] to a non-parallel learning scheme by using cycle consistent generative adversarial networks (CycleGANs [11]), demonstrating superior quality and degree of Lombardness in comparison to the parallel-data system in [9].

Although powerful for non-parallel learning, CycleGANs have the disadvantage that they are only capable of learning a deterministic mapping from one style to another. However, increasing the level of Lombardness beyond what is required in a given environment may result in an undesired mismatch between communicative expectations and the resulting vocal expression in the given situation. Therefore, an ideal normal-to-Lombard SSC system would convert speech only to the extent that is suitable for the current level of background noise (i.e., allow conversion along the vocal effort continuum). Such a system would also require a way to control the degree of Lombardness achieved in the conversion, which is not necessarily trivial for mappings making use of machine learning.

Beyond the lack of control, the deterministic mapping in the CycleGAN has a number of other limitations. For instance, the model tends to embed information regarding the source style in the transformed signal, as the generator has to recover the details of the original sample in order to satisfy the cyclic consistency requirement [12]. The mapping also tends towards the mean of the target distribution, whereas natural speech has a rich variability of vocal efforts that may not be well captured by such a mapping. Several studies have explored non-deterministic GAN variants to counter these problems, including so-called BiCycleGANs [13] and the use of domain-specific variational information bound [14]. Among these, the Augmented CycleGANs [15] are an attractive extension to the CycleGAN framework, as they simultaneously learn the mapping and a latent space that encodes additional variability in the data distributions. In SSC, the latent space could provide a principled way to capture the natural variance in speaking styles (e.g., different degrees of Lombardness) encountered in real speech. After training, the latent space could potentially be traversed to obtain different transformations for the same input.

Given this background, the goal of the current study is to explore the applicability of Augmented CycleGANs in a normal-to-Lombard SSC with the aim of having controllable degree of Lombardness in the conversion system. The paper builds on the data driven approach to the normal-to-Lombard parametric SSC system explored in [8–10] by using the Pulse model in log domain (PML [11]) vocoder with the Augmented CycleGANs [2]. We compare the new system to the reference CycleGAN system proposed for the same task in [10] using instrumental and subjective intelligibility tests and a quality test for the converted speech. As a result, we show that the new SSC system learns a controllable one-to-many mapping from the source style to the target style, capturing different degrees of Lombardness present in the training data.

http://dx.doi.org/10.21437/Interspeech.2019-1681

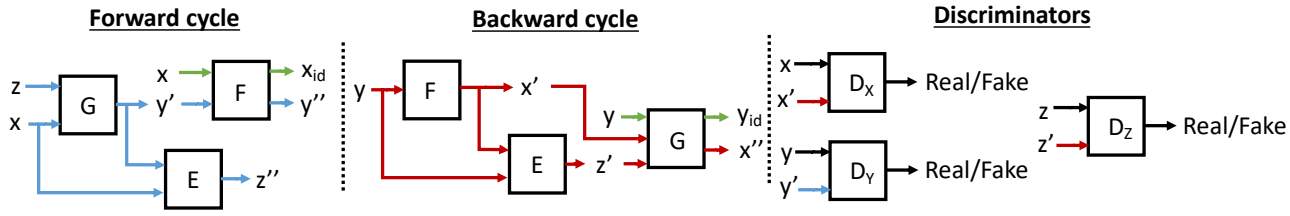

Figure 1: *Augmented CycleGAN with mapping functions G, F and E, and discriminators D_X, D_Y and D_Z. The forward cycle, backward cycle, and identity mapping are indicated with blue, red, and green respectively.*

2. Augmented CycleGAN

The Augmented CycleGAN [15] is an extension of the CycleGAN [11] that is capable of learning many-to-many bi-directional mappings between pairs of items $(x, z_x) \in X \times Z_X$ and $(y, z_y) \in Y \times Z_Y$, where, Z_X and Z_Y are trained augmented latent spaces that capture any missing information when transforming from an element in domain X to Y, and vice-versa. In our current study, we consider a one-to-many version of the Augmented CycleGAN that learns a mapping between domains $X \times Z$ and Y, where Z is the latent space with information about Y not contained in X. The basic structure of a Augmented CycleGAN is shown in Figure 1. It consists of three functions G, F and E, which map data from $X \times Z \rightarrow Y$, $Y \rightarrow X$ and $X \times Y \rightarrow Z$ respectively, and three discriminators D_X, D_Y, and D_Z, which determine whether data is from the true distributions $P(X)$, $P(Y)$, and $P(Z)$, respectively. The true data distribution for the latent space $P(Z)$ is assumed to be Gaussian.

The loss function of an augmented CycleGAN can be formulated similar to that of a standard CycleGAN used in [10]. In our implementation, we use the Wasserstein distance metric (WGAN loss) with gradient penalty [16] to determine the adverserial loss, defined as

$$\mathcal{L}_{gan}(G, D_Y, X, Y, Z) = \mathop{\mathbb{E}}_{y \sim p(Y)}[D_Y(y)] - \mathop{\mathbb{E}}_{x \sim p(X)}[D_Y(G(x, z))] + \lambda_g \mathop{\mathbb{E}}_{\hat{y} \sim p(\hat{Y})}[(||\nabla_{\hat{y}} D_Y(\hat{y})||_2 - 1)^2] \quad (1)$$

where $p(\hat{Y})$ is implicitly defined by sampling along the straight lines between pairs of points y and $G(x)$ and λ_g is the weight on the gradient penalty term of the WGAN. Similar loss terms are derived for $\mathcal{L}_{gan}(F, D_X, X, Y)$ and $\mathcal{L}_{gan}(E, D_Z, Y, Z)$ as

$$\mathcal{L}_{gan}(F, D_X, X, Y) = \mathop{\mathbb{E}}_{x \sim p(X)}[D_X(x)] - \mathop{\mathbb{E}}_{y \sim p(Y)}[D_X(F(y))] + \lambda_g \mathop{\mathbb{E}}_{\hat{x} \sim p(\hat{X})}[(||\nabla_{\hat{x}} D_X(\hat{x})||_2 - 1)^2] \quad (2)$$

$$\mathcal{L}_{gan}(E, D_Z, Y, Z) = \mathop{\mathbb{E}}_{z \sim p(Z)}[D_Z(z)] - \mathop{\mathbb{E}}_{z \sim p(Z)}[D_Z(E(y, F(y)))] + \lambda_g \mathop{\mathbb{E}}_{\hat{z} \sim p(\hat{Z})}[(||\nabla_{\hat{z}} D_Z(\hat{z})||_2 - 1)^2] \quad (3)$$

A cyclic reconstruction loss term is also defined as shown

$$\mathcal{L}_{cyc}(G, F, E, X, Y, Z) = \mathop{\mathbb{E}}_{x \sim p(X), z \sim p(Z)}[||F(G(x, z)) - x||_1] + \mathop{\mathbb{E}}_{y \sim p(Y)}[||G(F(y), E(F(y), y)) - y||_1] + \mathop{\mathbb{E}}_{x \sim p(X), z \sim p(Z)}[||E(x, G(x, z)) - z||_1] \quad (4)$$

Finally, the identity mapping loss [11, 17] is defined to ensure that input data already corresponding to target domain do not get transformed in G or F (shown in green in Figure 1).

$$\mathcal{L}_{id}(G, F, X, Y, Z) = \mathop{\mathbb{E}}_{x \sim p(X)}[||F(x) - x||_1] + \mathop{\mathbb{E}}_{y \sim p(Y), z \sim p(Z)}[||G(y, z)) - y||_1] \quad (5)$$

The mapping functions G^*, F^* and E^* are trained by alternating gradient decent on the minmax-game defined as

$$G^*, F^*, E^* = \mathop{\text{argmin}}_{G,F,E} \mathop{\max}_{D_X, D_Y, D_Z} \mathcal{L}_{all}$$

$$\text{where, } \mathcal{L}_{all} = \mathcal{L}_{gan}(G, D_Y, X, Y, Z) + \mathcal{L}_{gan}(F, D_X, X, Y) + \mathcal{L}_{gan}(E, D_Z, Y, Z) + \lambda_{cyc}\mathcal{L}_{cyc}(G, F, E, X, Y, Z) + \lambda_{id}\mathcal{L}_{id}(G, F, X, Y, Z) \quad (6)$$

where λ_{cyc} and λ_{id} control the relative importance of the cyclic reconstruction loss and the identity mapping loss respectively.

3. Normal-to-Lombard SSC

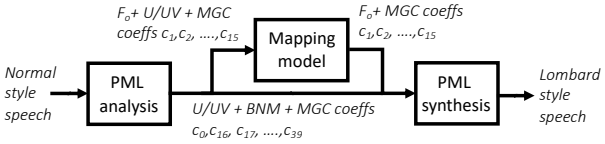

Figure 2: *Block diagram of the normal-to-Lombard speaking style conversion system.*

Following our previous works [9, 10] on SSC, the current study uses a parametric system utilizing frame level Pulse Model in Log domain (PML [18]) vocoder features for the normal-to-Lombard conversion. The basic block diagram is shown in Figure 2. First, the PML vocoder features are extracted from the input signal, followed by mapping with the Augmented CycleGAN. The modified features are then fed to the PML vocoder synthesis to generate the mapped Lombard speech utterance. Duration modification is not carried out in the current study, as we focus on differences in the degree of Lombardness between the Augmented CycleGAN system and the baseline CycleGAN system [10]. If desired, a simple system for duration modification using constant scaling of the voiced and unvoiced segments as described in [8–10] could be utilized. The sections below describe the PML vocoder and the mapping methods chosen for comparison.

3.1. PML vocoder

The PML [18] vocoder has shown good performance in two recent studies [18, 19]. PML uses a log-domain source-filter model with a sinusoidal signal analysis and a pitch synchronous pulse-based synthesis system. PML models aperiodicity via a phase distortion deviation (PDD) spectrum, which generalizes to modeling both voiced and unvoiced speech without explicit voicing decisions. PDD is thresholded to produce a binary noise mask (BNM), which is averaged in Mel-bands for parametric processing. In total, the PML vocoder features include 1) the binary noise mask (BNM), 2) fundamental frequency (F0), 3) the voicing decision (V/UV) mask, and 4) the spectral envelope. In the present study, features that are the most important for normal-to-Lombard SSC, i.e., F0 and spectral envelope encoded by the first 15 MGC coefficients c_1 ... c_{15} are used in the

mapping. The rest of the features are directly by-passed to the target domain for synthesis.

3.2. Mapping model

The current study compares the Augmented CycleGAN to the standard CycleGAN. Figure 3 shows the block diagram of the deep convolutional neural networks with residual connections (CNN ResNet, similar to [10]) which are used as a model for the generator G. Each of the 6 CNN layers has 256 channels, consisting of an 11-point gated convolutional unit with the last layer being a linear convolutional layer. The latent variables are modeled by a series of two fully-connected (FC) feedforward layers, each with the same number of units as the dimensionality of the latent space with leaky ReLU non-linearities (as in [20]). Adaptive instance normalization (AdaIN, [21, 22]) based on linearly transformed output of the latent variable layers (Fig. 3, bottom) is used before the application of the non-linearity on each layer of the CNN. Generators F, E in the Augmented CycleGAN and the G, F of the reference CycleGAN as well as discriminators D_X and D_Y of both are modelled using the same CNN model without the AdaIN layers, with inputs to E, representations of the source and target domains, being concatenated. Discriminator D_Z of the Augmented CycleGAN is modelled by a simple feedforward network with 2 layers with 8 gated units each. The source codes of the Augmented CycleGAN are available under an open source license[1].

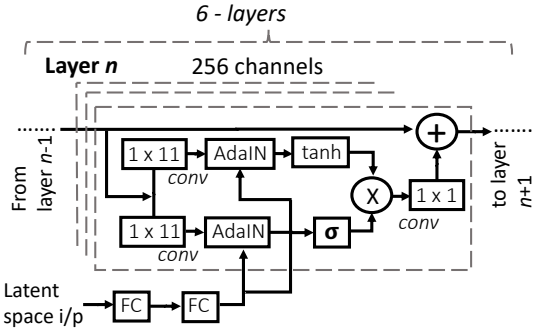

Figure 3: *Block diagram of layer n of the CNN used to model the mapping functions G, F and E and the discriminators D_x and D_y of the augmented CycleGAN.*

4. Experimental Setup

4.1. Data

Read and conversational Lombard speech corpora from two languages, Finnish and English were used to train the current model. The Finnish corpus (see [23] for details) consists of recordings from 20 Finnish speakers (10 female) in normal and Lombard style. The Lombard speech was elicited using four different background noise conditions (highly non-stationary pub noise and stationary car noise in both mild and severe conditions) played to the speakers' ears with headphones while they were being recorded [23]. The English data were from the Hurricane Challenge dataset [24] that contains both normal and Lombard speech spoken by one British male. The Finnish and English data consisted of approx. 80 and 60 minutes of speech, respectively. Data from both corpora were downsampled to 16 kHz before further processing.

[1] https://github.com/shreyas253/AugmentedCycleGAN_1dCNN

For the instrumental and subjective intelligibility tests, English Lombard grid-speech corpus from [25] was used. The dataset contained 2700 normal and Lombard utterances each respectively from 54 talkers. Each utterances contained short 6 word sentences with a color, letter and digit that are considered as keywords for a keyword spotting task [26], such as "*Set blue in B five please*". For the subjective quality experiments, two (1 male and female) speakers form the Finnish corpora were randomly chosen for evaluation, while the rest of the 18 speakers along with the English data were used for training.

4.2. System specifications

The inputs to the mapping model were z-scored to zero mean and unit variance. Analysis frames of 25 ms with a 5-ms frame shift were used by the PML vocoder. F0 was computed using the RAPT algorithm from the SPTK toolkit [27]. The binary noise mask of the PML vocoder was 25-dimensional. 40-dimensional Mel-generalized cepstrum (MGC) coefficients were used to represent the spectral envelopes which were extracted using STRAIGHT [28]. The dimension of the latent space Z was set to 2. The hyperparameters of the loss function in Equation 6, λ_g and λ_{cyc} were both set to 10. The training was run for 5000 iterations with λ_{id} set to 5 for the first 2500 iterations and linearly decreasing to 0 from then (similar to [29]). $\mathcal{L}_{all}$ in Equation 6 also included a penalty on discriminator output magnitudes (see [30]).

After training the Augmented CycleGAN, each of the training utterances of the Lombard style became associated with a point in the latent space as $\hat{z} = E(F(y), y)$. Linear discriminant analysis (LDA) was used on these points to find a linear combination of latent variables that maximally separates different values of spectral tilt (represented as histogram-quantized c_1 coefficients) of the training utterances associated with the points $\hat{z}$. This line in the latent space was used as a proxy for vocal effort continuum along which there is a maximal change in the degree of Lombardness. For the experiments, three points along this line were chosen to reflect three levels of Lombardness (low, medium, and high) in the mapped utterances.

4.3. Evaluation

The mapped utterances were initially evaluated using an instrumental intelligibility test called Speech Intelligibility in Bits (SIIB, [31]) using its Gaussian variant (SIIBGauss [31]). Subjective evaluation was then carried out, including an English *Intelligibility test* and a Finnish *Quality test*. Participants were 21 Finnish native graduate students with a good command of English (e.g., [32] reports LexTale scores comparable to C1/C2 Common European Framework English proficiency in the same population). Sounds were played to the listeners in a quiet room using the Sennheiser HD598 headphones. All the sound samples being compared were normalized using sv56 standard [33]. Each listening test included a tutorial phase before the actual test. Furthermore, the listeners were asked to adjust the sound volume to a loud yet comfortable level during the tutorial session, after which the volume was kept fixed. The two tests took approximately 40 minutes for the subjects to complete. The tests were implemented using MATLAB GUI.

Objective intelligibility was measured using SIIB [31, 34] that is based on the mutual information between a clean reference and a noisy signal (as used in [9]). The test was conducted on the entire English Lombard grid-speech corpus [25] using two different noise types (unstationary factory noise and stationary Volvo noise [35]) at two signal-to-noise ratio (SNR)

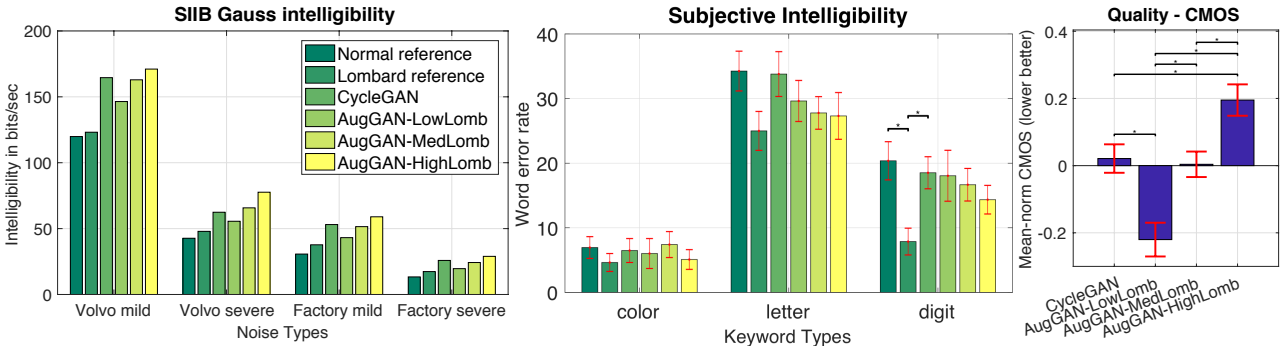

Figure 4: *From left to right, (a) the SIIB instrumental intelligibility metric over the 4 noise conditions, (b) the keyword error rates of the subjective intelligibility averaged over the 4 noise types and (c) the CMOS scores from the subjective quality test. Error bars correspond to one standard error. Bonferroni corrected significant values for (b) and (c), calculated using the Students t-test and the Mann-Whitney U-test respectively, are highlighted.*

levels here referred to as *moderate* and *severe*. These SNR levels were set to 0 dB and -5 dB for the factory noise and -15 dB and -25 dB for the Volvo noise after initial piloting (in line with [36]). SIIB was measured for the copy synthesis versions of the natural normal and natural Lombard utterances as well as for four SSC systems: the baseline CycleGAN system and three Augmented CycleGAN systems corresponding to latent space values with low, medium, and high Lombardness.

The subjective intelligibility test was conducted as a keyword spotting task [26]. On each trial, the subject heard a short six-word English sentence that had a color, letter, and digit in that order, and which they were asked to identify. This experiment was conducted in the same noise conditions as the SIIB test. Each subject listened to a total of 80 stimuli which consisted of three utterances in each style category (natural normal, natural Lombard, four normal-to-Lombard-converted) and in each four noise conditions, as well as eight reference utterances in no noise condition. The listeners were only allowed to listen to the utterances once.

Finally, the subjective quality test was performed using the comparison category rating (CCR) test [37]. Each trial contained a pair of utterances without added noise. The subjects were asked to rate the perceived quality of the second utterance in comparison to the first using a continuous rating scale: -3, much worse; -2, worse; -1, slightly worse; 0, almost similar; 1, slightly better, 2, better; 3, much better. Each pair consisted of a mapped utterance and its corresponding copy synthesis version of the natural Lombard utterance. Each pair was presented in both orders and null pairs were also included. Each listener rated 36 utterances in total. The comparison mean opinion score (CMOS) [37], the average of the scores for each unique utterance pair, normalized to zero mean across each listener (as suggested in [38]) was used as the final quality measure. Lower normalized CMOS value means better speech quality.

5. Results

Results for the subjective and instrumental evaluation are shown in Figure 4. The SIIB measure shows that the CycleGAN has an intelligibility score slightly higher than that of the reference Lombard. The variations of the Augmented CycleGAN gradually increase in intelligibility from a level comparable to the reference Lombard to that of exceeding the CycleGAN. As for the subjective intelligibility test, it can be seen that the only significant differences are in the 'digit' keyword error rates between the reference normal and the reference Lombard and between the reference Lombard and the CycleGAN. The lack of statistically significant differences in the other two keywords, even between natural normal and Lombard speech, indicates that the present experimental setup with a limited number of trials per category lacks the statistical power to properly reveal more detailed differences in intelligibility (e.g., [39] used 33 times more trials per category on grid corpus data, which was infeasible for our current setup). However, the results do indicate that the intelligibility of the mapping methods lie in between the reference normal and Lombard, and the general pattern follows that of the instrumental test. Finally, the CMOS scores show that the CycleGAN is significantly worse in quality than the low Lombard variant, and significantly better than the high Lombard variant of the augmented CycleGAN. The different variations of the augmented CycleGAN are increasingly worse in quality with increasing degree of Lombardness. Example sound files are available[2].

6. Discussion and Conclusions

The results indicate that the proposed system with an Augmented CycleGAN is capable of SSC with a controllable degree of Lombardness. Moreover, the system achieves equal speech quality with the baseline CycleGAN system when the degree of Lombardness is similar in the two systems. While the pattern of increased degree of Lombardness as a function of the latent space variables is evident in the SIIB-based objective intelligibility metrics, statistical power of the listening test was not able to reveal detailed differences in intelligbility rates. However, the listening tests also indicate that the Lombardness of the Augmented CycleGAN-mapped utterances is statistically indistinguishable from Lombard speech. The reason why the system fails to reach the level of natural Lombard speech intelligibility in listening tests but exceeds that in instrumental metrics needs to investigated in future work, but may be related to the inherent degradations in signal quality due to vocoding and statistical mapping. Overall, the study shows that Augmented CycleGANs are a highly potential extension to the CycleGAN framework for speech conversion problems where non-deterministic controllable mappings are desirable.

7. Acknowledgements

This study was funded by Academy of Finland grant nos. 312105, 314602, and 312490. The authors thank the participants of the listening tests.

[2] https://shreyas253.github.io/Norm2Lomb_AugCycleGAN/

8. References

[1] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *Journal of Speech, Language, and Hearing Research*, vol. 14, pp. 677–709, Sep. 1971.

[2] ITU-T, "One-way transmission time," International Telecommunication Union, Geneva, Switzerland, Rec. G.114, May 2003.

[3] H. Konno, M. Kudo, H. Imai, and M. Sugimoto, "Whisper to normal speech conversion using pitch estimated from spectrum," *Speech Communication*, vol. 83, pp. 10–20, Oct. 2016.

[4] G. N. Meenakshi and P. K. Ghosh, "Whispered speech to neutral speech conversion using bidirectional lstms," *Proc. Interspeech 2018*, pp. 491–495, 2018.

[5] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, no. 7-8, pp. 515–520, 2002.

[6] K. Nathwani, G. Richard, B. David, P. Prablanc, and V. Roussarie, "Speech intelligibility improvement in car noise environment by voice transformation," *Speech Communication*, vol. 91, pp. 17–27, Jul. 2017.

[7] D.-Y. Huang, S. Rahardja, and E. P. Ong, "Lombard effect mimicking," in *Proc. SSW*, Kyoto, Japan, Sep. 2010, pp. 258–263.

[8] A. R. López, S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Speaking style conversion from normal to Lombard speech using a glottal vocoder and Bayesian GMMs," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 1363–1367.

[9] S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Vocal effort based speaking style conversion using vocoder features and parallel learning," *IEEE Access*, vol. 7, pp. 17 230–17 246, 2019.

[10] S. Seshadri, L. Juvela, J. Yamagishi, O. Räsänen, and P. Alku, "Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion." in *Proc. ICASSP-2019*, pp. 6835–6839.

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proc. ICCV 2017*, pp. 2223–2232, 2017.

[12] C. Chu, A. Zhmoginov, and M. Sandler, "CycleGAN, a master of steganography," 2017. [Online]. Available: https://arxiv.org/abs/1712.02950

[13] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *NeurIPS 30*, 2017, pp. 465–476.

[14] H. Kazemi, S. Soleymani, F. Taherkhani, S. Iranmanesh, and N. Nasrabadi, "Unsupervised image-to-image translation using domain-specific variational information bound," in *NeurIPS 31*, 2018, pp. 10 348–10 358.

[15] A. Almahairi, S. Rajeshwar, A. Sordoni, P. Bachman, and A. Courville, "Augmented CycleGAN: Learning many-to-many mappings from unpaired data," in *Proc. ICML*, Stockholm Sweden, 2018, pp. 195–204.

[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS*, 2017, pp. 5767–5777.

[17] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation." in *Proc. ICLR-2017*, 2017.

[18] G. Degottex, P. Lanchantin, and M. Gales, "A log domain pulse model for parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 57–70, Jan. 2018.

[19] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "GlottDNN-A full-band glottal vocoder for statistical parametric speech synthesis," in *Proc. Interspeech*, San Francisco, USA, Sep. 2016, pp. 2473–2477.

[20] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR*, 2019.

[21] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. ICCV*, 2017, pp. 1501–1510.

[22] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," in *Proc. ICLR*, 2017.

[23] E. Jokinen, U. Remes, and P. Alku, "The use of read versus conversational Lombard speech in spectral tilt modeling for intelligibility enhancement in near-end noise conditions," in *Proc. Interspeech*, San Francisco, USA, Sep. 2016, pp. 2771–2775.

[24] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Hurricane natural speech corpus [sound]," 2013. [Online]. Available: https://datashare.is.ed.ac.uk/handle/10283/347

[25] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual Lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[26] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[27] SPTK Working Group, "Speech Signal Processing Toolkit (SPTK) version 3.8," http://sp-tk.sourceforge.net/, 2014.

[28] H. Kawahara, J. Estill, and O. 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 *Proc. MAVEBA*, 2001.

[29] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," *Proc. EUSIPCO 2018*, pp. –, 2018.

[30] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. ICLR-2018*, 2018.

[31] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2018.

[32] S. Kakouros and O. Räsänen, "Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features," *Cognitive Science*, vol. 40, pp. 1739–1774, 2016.

[33] ITU-R, "Objective measurement of active speech level ITU-T recommendation," International Telecommunication Union, Geneva, Switzerland, Rec. P.56, 1993.

[34] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2153–2166, 2018.

[35] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[36] E. Jokinen, U. Remes, P. Alku, E. Jokinen, U. Remes, and P. Alku, "Intelligibility enhancement of telephone speech using Gaussian process regression for normal-to-Lombard spectral tilt conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 10, pp. 1985–1996, Oct. 2017.

[37] ITU-T, "Methods for objective and subjective assessment of quality," International Telecommunication Union, Rec. ITU-R P.800, Aug. 1996.

[38] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 3976–3980.

[39] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication*, vol. 49, no. 5, pp. 402–417, 2007.