CYCLE-CONSISTENT ADVERSARIAL NETWORKS FOR NON-PARALLEL VOCAL EFFORT BASED SPEAKING STYLE CONVERSION

Shreyas Seshadri[⋆] *Lauri Juvela*[⋆] *Junichi Yamagishi*[†‡] *Okko Räsänen*[§⋆] *Paavo Alku*[⋆]

[⋆] Department of Signal Processing and Acoustics, Aalto University, Finland
[†] Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan
[‡] The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[§] Laboratory of Signal Processing, Tampere University of Technology, Finland

ABSTRACT

Speaking style conversion (SSC) is the technology of converting natural speech signals from one style to another. In this study, we propose the use of cycle-consistent adversarial networks (CycleGANs) for converting styles with varying vocal effort, and focus on conversion between normal and Lombard styles as a case study of this problem. We propose a parametric approach that uses the Pulse Model in Log domain (PML) vocoder to extract speech features. These features are mapped using the CycleGAN from utterances in the source style to the corresponding features of target speech. Finally, the mapped features are converted to a Lombard speech waveform with the PML. The CycleGAN was compared in subjective listening tests with 2 other standard mapping methods used in conversion, and the CycleGAN was found to have the best performance in terms of speech quality and in terms of the magnitude of the perceptual change between the two styles.

Index Terms— CycleGAN, style conversion, vocal effort, Lombard speech, pulse-model in log domain vocoder

1. INTRODUCTION

Vocal effort based speaking style conversion (SSC) is the technology of converting natural speech signals spoken in a particular style to another (e.g. whisper-to-normal or normal-to-Lombard [1]) while retaining the linguistic and speaker-specific information of the original speech signal. SSC has multiple potential applications, such as personalizing speech to the needs of the end-listener. For instance, normal speech could be converted into clear speech [2, 3] for hearing-impaired listeners. SSC can be also used for generation of context-dependent speech samples from a limited set of original recordings for recreational applications such as gaming and virtual reality. While there has already been work in whispered-to-normal speech conversion (e.g., [4–8]), SSC for other aspects of vocal effort has only been studied in a small number of previous works [9–13] that only focus on direct signal manipulation or parallel data training.

In the current study, we focus on conversion between normal and Lombard styles. Lombard speech corresponds to a speaking style that talkers naturally employ in noisy environments to improve intelligibility. We use a data driven parametric setup that uses a vocoder to extract speech features (see [14]). These features are mapped using machine learning models from the source style to the corresponding features of the target style. Finally, the mapped features are transformed to target style speech waveform with the vocoder.

Collection of a large quantity of Lombard speech data (as well as data from other styles with varying vocal effort) is laborious and potentially injurious to health of the speakers. Moreover, the amount of parallel training data, where the utterances in the source and target styles are from the same speaker speaking the same linguistic content, is limited. Our earlier work [14, 15] also suggest that the limited availability of parallel data in normal and Lombard styles causes a bottleneck in system performance. This encourages the use of non-parallel mapping models within the parametric SSC framework.

The standard approach for non-parallel training in the domain of voice conversion is the INCA algorithm [16] (see Section 3.2.2) that iteratively finds alignments between individual frames in the source and target styles. Variants of the basic INCA may use several subsequent frames [17], dynamic features [18], or custom distance metrics [19] in the INCA alignment process. As a recent alternative to INCA, Cycle-consistent adversarial networks (CycleGANs, [20]) have shown promise in the domain of voice conversion. For example [21] uses dynamic frame-level Mel-spectrum features with a standard feed-forward deep neural network (DNN) to achieve high quality voice-conversion. [22] uses a convolutional neural network (CNN) with Gated linear units (GLUs) and residual connections to map Mel features with a CycleGAN. The major advantage of the CycleGAN architecture, as compared to other non-parallel alternatives such as INCA, is that it does not rely on data alignment during the training. Instead, the method simply learns to create transformations to the source features that are statistically indistinguishable from the target domain while learning the reverse mapping at the same time. Since the alignment process is non-trivial for non-parallel data, especially in case of SSC where the source and target style features can come from the same talker but should still exhibit different signal properties (hence the need for conversion), the CycleGAN approach can potentially be especially useful.

Given this background, the overall goal of the current paper is to study the applicability of CycleGANs for the task of vocal effort based SSC (normal vs. Lombard), and to compare it to the standard INCA-based non-parallel approach and to our previous baseline system utilizing parallel data [14]. The systems are compared using subjective listening tests evaluating the success of style conversion and overall quality of the converted speech.

The paper is organized as follows: Section 2 describes the general structure and mathematical formulation of the CycleGAN used in the current study. Section 3 provides the basic framework for the parametric SSC system and outlines the vocoder (Pulse Model in

This study was funded by Academy of Finland grant nos. 312105, 314602, and 312490. JY was partially supported by JST CREST Grant Number JPMJCR18A6, Japan and by MEXT KAKENHI Grant Numbers (16H06302, 17H04687, 18H04120, 18H04112, 18KT0051), Japan.

Log domain, [23]) and mapping methods studied. Section 4 explains the experimental setup, including data used, system specification details and the subjective evaluation. Finally, Sections 5 and 6 describe the results and conclusions respectively.

2. CYCLE-CONSISTENT ADVERSERIAL NETWORKS

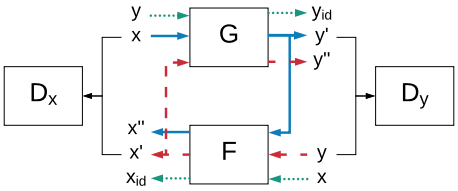

Fig. 1. CycleGAN with mapping functions G and F, and discriminators D_X and D_y. The forward cycle, backward cycle, and identity mapping are indicated with red, blue, and green respectively

A CycleGAN [20] is a non-parallel learning scheme that learns bi-directional deterministic mappings between domains X and Y, given non-aligned training samples $x_i \sim p(X)$ and $y_i \sim p(Y)$. A CycleGAN is based on the concept of adversarial learning [24], where a generative model is trained as a solution to a minmax two-player game between two neural networks called as the generator and discriminator. The basic structure of a CycleGAN is shown in Figure 1. It consists of two functions G and F, which map data from $X \rightarrow Y$ and $Y \rightarrow X$ respectively, and two discriminators D_X and D_Y, which determine whether data is from the true distributions $P(X)$ and $P(Y)$, respectively. During training, data flows in two directions: the forward cycle $x_i \xrightarrow{G} y_i' \xrightarrow{F} x_i''$ and the backward cycle $y_i \xrightarrow{F} x_i' \xrightarrow{G} y_i''$ as indicated by the blue and red arrows, respectively, in Figure 1.

The loss function of a CycleGAN has three terms. The first one, adversarial loss, measures distance of the mapped data to the true target distribution. In our implementation, we use the Wasserstein distance metric (WGAN loss) with gradient penalty [25], defined as

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

where $p(\hat{Y})$ is implicitly defined by sampling along the straight lines between pairs of points y and $G(x)$. A similar loss term is derived for $\mathcal{L}_{gan}(F, D_X, X, Y)$.

A cyclic reconstruction loss term is also defined to ensure that data passing through both G and F (or in another direction) results in an identity mapping as shown

$$\begin{aligned}\mathcal{L}_{cyc}(G, F, X, Y) = &\mathop{\mathbb{E}}_{x\sim p(X)}[||F(G(x)) - x||_1] \\ &+ \mathop{\mathbb{E}}_{y\sim p(Y)}[||G(F(y)) - y||_1]\end{aligned} \tag{2}$$

Finally, the identity mapping loss [20, 26] 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):

$$\begin{aligned}\mathcal{L}_{id}(G, F, X, Y) = &\mathop{\mathbb{E}}_{x\sim p(X)}[||F(x) - x||_1] \\ &+ \mathop{\mathbb{E}}_{y\sim p(Y)}[||G(y)) - y||_1]\end{aligned} \tag{3}$$

The optimal mapping functions G^* and F^* are obtained by solving the minmax-game defined as

$$\begin{aligned}&G^*, F^* = \mathop{\text{argmin}}_{G,F} \max_{D_X, D_Y} \mathcal{L}(G, F, D_X, D_Y, X, Y) \\ &\text{where, } \mathcal{L}(G, F, D_X, D_Y, X, Y) = \mathcal{L}_{gan}(G, D_Y, X, Y) \\ &+ \mathcal{L}_{gan}(F, D_X, X, Y) + \lambda_{cyc}\mathcal{L}_{cyc}(G, F, X, Y) \\ &+ \lambda_{id}\mathcal{L}_{id}(G, F, X, Y)\end{aligned} \tag{4}$$

where λ_{cyc}, λ_{id}, and λ_g from Equation 1 control the relative importance of the cyclic reconstruction loss, the identity mapping loss, and the gradient penalty term of the WGAN respectively.

3. NON-PARALLEL SPEAKING STYLE CONVERSION

The current study focuses on SSC between normal and Lombard styles. We employ a parametric approach based on the manipulation of frame level vocoder features. Pulse Model in Log domain (PML [23]) is chosen as the vocoder, as it demonstrated good performance in two recent vocoder studies [23, 27]. Figure 2 shows the block diagram of our parametric SSC system. The source style features are first extracted using the PML vocoder analysis, viz: 1) the binary noise mask (BNM), 2) fundamental frequency (F0), 3) the voicing decision (V/UV) mask, and 4) the spectral envelope. The durations of the voiced and unvoiced segments are then scaled by a constant factor respectively (see [28] for a study of the duration of phonemic classes in Lombard speech). Features that are most important for the styles in question are modified using the trained mapping model. For SSC between normal and Lombard speech, after testing various combinations of featues based on subjective quality of the converted speech, we chose the F0, voicing decisions (V/UV), and the first 10 MGC coefficients (the major properties of the spectral envelope) as the features for mapping. The modified features are then fed to the PML vocoder synthesis to generate the speech utterance in the desired target style. The sections below describe the PML vocoder and the mapping methods chosen for comparison.

3.1. PML Vocoder

PML [23] is a recent state-of-the art vocoder utilizing log-domain source-filter modeling, sinusoidal signal analysis and pitch synchronous pulse-based synthesis. The vocoder's distinctive property is its aperiodicity modeling via a phase distortion deviation (PDD) spectrum, which generalizes to modeling both voiced and unvoiced speech without explicit voicing decisions. The PDD is thresholded to produce a binary noise mask (BNM), which is averaged in Mel-bands for parametric processing.

3.2. Mapping methods

3.2.1. Parallel GMM learning

In the baseline system with parallel data, a standard GMM is used as it was shown to compare well against DNNs and non-parametric Bayesian methods in an earlier study with the present data set [14]. For the GMM training, dynamic time warping (DTW) [29] aligned source, $\mathbf{x}_s$, and target data, $\mathbf{x}_t$, are concatenated as $\mathbf{x} = [\mathbf{x}_s, \mathbf{x}_t]^T$ to obtain N samples of D-dimensional training data $\mathbf{X} = [\mathbf{x}_1, .., \mathbf{x}_N]$ for the GMM model. During mapping, the minimum mean square error (MMSE) estimate of target features $\boldsymbol{y}_t$, $\hat{\boldsymbol{y}}_t$ is calculated as

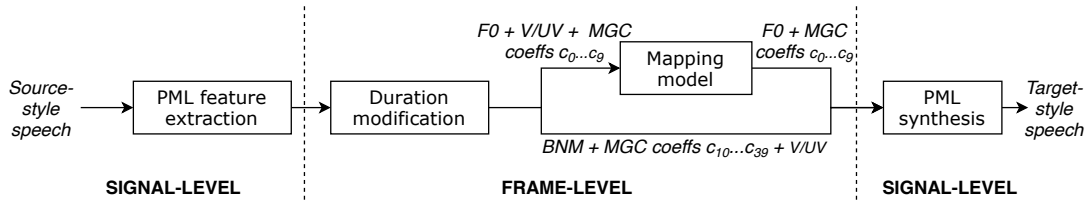

Fig. 2. Block diagram of the speaking style conversion system.

$$\hat{\mathbf{y}}_t = \sum_{k=1}^{K} p(k|\mathbf{y}_s, \mathbf{X})[\boldsymbol{\mu}_{t|k} + \boldsymbol{\Sigma}_{ts|k}\boldsymbol{\Sigma}_{ss|k}^{-1}(\mathbf{y}_s - \boldsymbol{\mu}_{s|k})]$$
$$\text{where, } p(\mathbf{X}; \boldsymbol{\theta}_k, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}([\begin{bmatrix} \boldsymbol{\mu}_{s|k} \\ \boldsymbol{\mu}_{t|k} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{ss|k} & \boldsymbol{\Sigma}_{st|k} \\ \boldsymbol{\Sigma}_{ts|k} & \boldsymbol{\Sigma}_{tt|k} \end{bmatrix}]) \quad (5)$$

where $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$, are the mean and covariance of the kth Gaussian, the weights, π_k, sum to one and $p(k|\mathbf{y}_s, \mathbf{X})$ is the probability of the kth component calculated based on π_k and marginal likelihood of the kth Gaussian. (See [30] for a detailed derivation and [30–32] for other use-cases of GMM mapping.)

3.2.2. INCA

The Iterative combination of a Nearest Neighbor search step and a Conversion step Alignment method (INCA) [16] is a commonly used non-parallel learning algorithm that iteratively looks for nearest neighbor feature pairs between the source and target speaker while also iteratively updating the conversion model to progressively improve matching to the target speaker. A GMM is used as the mapping model (see Section 3.2.1) from source to target, as the over-smoothing effect of the GMM mapping helps to minimize the negative effects of misalignments during the initial iterations (see [16]).

3.2.3. CycleGAN

Several DNN architectures were tested to model the mapping functions G and F and the discriminators D_x and D_y (section 2). In our current implementation a deep convolutional neural network with residual connections (CNN ResNet) is used (similar to [20]). Figure 3 shows the basic structure of a single layer of this network. Each layer has k-channels consisting of a w-point gated convolutional unit. The output of this unit is passed through an affine transform and added with a residual of the original input. The residual connection helps in preventing the problem of diminishing gradients in deeper networks. A CycleGAN with a feed-forward network that maps frame level features with temporal context as in [21] was also tested, but it was found to be inferior to the proposed architecture in our initial tests. The source codes of the CycleGAN are available under an open source license[1].

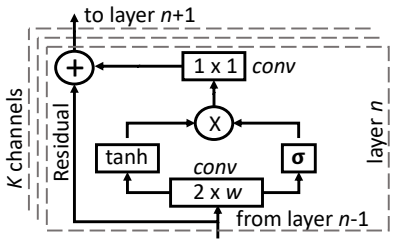

Fig. 3. Block diagram of layer n of the CNN used to model the mapping functions G and F and the discriminators D_x and D_y.

[1] https://github.com/shreyas253/CycleGAN_1dCNN/

4. EXPERIMENTAL SETUP

4.1. Data

Read and conversational Lombard speech corpus (see [33] for details) consisting of recordings from 20 Finnish speakers (10 female) was used in the present study. The read part includes a text of 90 words read by each speaker (approx. 1 min per speaker). The same text was produced in two speaking styles, normal and Lombard. The conversational part consists of realistic telephone conversations, where the subjects played the role of either a caller or a travel agent. Size of this section is approximately the same as the read section. In order to elicit Lombard speech, background noise (highly nonstationary pub noise [34]) was played to the speakers' ears with headphones while they were being recorded [33]. Data were downsampled from 48 kHz to 16 kHz for the present experiments.

4.2. System specifications

Of the 20 speakers in the dataset, two (1 male and female) were randomly chosen for evaluation in listening tests (with the constraint that the speakers in the selection pool show clear normal and Lombard speech styles without stuttering), while the rest of the 18 speakers were used for training. The features to be mapped were normalized to zero mean and unit variance. During PML feature extraction, analysis frames of 25 ms with a 5-ms frame shift were employed. F0 was computed using the RAPT algorithm from the SPTK toolkit [35], and the range of allowed frequencies set to 50–500 Hz. The binary noise mask was 25-dimensional. The spectral envelope was extracted using STRAIGHT analysis and represented as 40-dimensional Mel-generalized cepstrum (MGC) coefficients. The scaling ratios for the duration conversion of the voiced and unvoiced segments of speech were calculated as the mean ratio of the corresponding durations in the DTW aligned segments of two speaking-styles, measured across all (un)voiced segments in the read data. These were found to be 1.08 and 0.88, i.e. the voiced and unvoiced regions were stretched and compressed, respectively (in line with [28]). The durations were modified by applying cubic spline interpolation to the resulting feature time-series.

The number of components in the GMM was set to 8 (as in [16]) for both the parallel GMM and INCA mapping methods. The INCA algorithm was allowed to run for 10 iterations (brief experiments showed that the performance doesn't improve beyond that, as noted in [16] for the use-case of voice conversion). For the CycleGAN, the number of CNN layers for the mapping functions and discriminators was set to 8 with the last layer being a linear convolutional layer. The number of channels per layer, k, was set to 256. The width of the convolutions, w, was set to 11. The hyperparameters of the loss function in Equation 4, λ_g, λ_{cyc} and λ_{ids} were set to 10, 10, and 5, respectively. The final loss function also included a penalty on discriminator output magnitudes to prevent the models from engaging in a "magnitudes race", as described in [36]. The training was run for 100 epochs with the identity mapping loss, $\mathcal{L}_{id}$ being dropped after 50 epochs (similar to [22]).

4.3. Subjective Evaluation

The subjective evaluation included a *Lombardness test* and a *Quality test*. 19 Finnish listeners participated in the listening tests. Sounds were played to the listeners in a quiet room using Sennheiser HD598 headphones. All the sound samples being compared were normalized to have the same RMS value. 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 listeners were allowed to listen to the samples as many times as they wished. Each test compared 4 unique utterances (2 speakers and 2 randomly chosen sentences) over the 3 methods (see Section 3.2) and for both normal-to-Lombard and Lombard-to-normal conversions. The tutorial sessions included utterances from speakers not included in the actual tests. The two tests took approximately one hour for the subjects to complete. A Speaking style similarity test (similar to a speaker similarity test commonly used in voice conversion [37]) was also conducted. However, the pattern of results from this was highly similar to the Lombardness-test, and the results are not therefore separately reported due to space considerations.

4.3.1. Lombardness Test

This test was set up as a MUSHRA-like (MUltiple Stimuli with Hidden Reference and Anchor, [38]) test. Each trial aimed to evaluate the 'Lombardness' of the mapped utterances. In a single trial, the listeners were given reference samples consisting of the original utterance spoken in both normal and Lombard styles and a set of unlabeled samples with the same speaker and lexical content to be rated on a Lombardness scale from 0 to 100. The utterances to be rated included a set of mapped utterances (from the 3 methods described in Section 3.2) and two hidden references of the original natural utterances in normal and Lombard styles (to be rated as 0 and 100, respectively). Each listener rated 8 trials in total. Before taking the test, the listeners were given a brief written description of Lombard speech. The listeners were also asked to focus on the style and try to ignore the speech quality. The tutorial session involved exposing the listeners to utterances in both styles. The tests were implemented using MATLAB's GUI (adapted from [39]).

4.3.2. Quality test

The quality test was performed using the comparison category rating (CCR) test [40]. For a given trial, the listeners were presented with pairs of speech utterances and asked to rate the perceived quality of the second utterance in comparison to the first one using a continuous rating scale that translates to English as: -3, much worse; -2, worse; -1, slightly worse; 0, almost similar; 1, slightly better, 2, better; 3, much better. In a single trial, each utterance pair consisted of a mapped utterance and its corresponding natural target style utterance. This was presented in both orders and also includes null pairs. Each listener rated 56 utterances in total. The tutorial session involved rating 3 CCR utterance pairs (including a null pair). The average of the scores for each unique utterance pair was calculated as the comparison mean opinion score (CMOS) [40], and normalized to zero mean across each listener (as suggested in [41]). As a result, lower normalized CMOS value means better speech quality.

5. RESULTS

The results of the Lombardness and Quality subjective tests are shown in Figures 4 and 5 respectively. For the normal-to-Lombard conversion, the CycleGAN produces the highest Lombardness, followed by the INCA and the parallel GMM. INCA and CycleGAN are very close in terms of quality, but are both significantly better than the parallel GMM. For the Lombard-to-normal mapping, the Lombardness of the 3 methods is almost indistinguishable. However, in terms of quality the CycleGAN is significantly better than the other two. Example sound files are available at `https://shreyas253.github.io/SpStyleConv_CycleGAN/`.

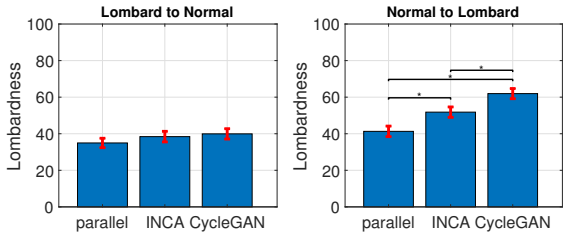

Fig. 4. Results of the Lombardness test for the Lombard-to-Normal (left, lower is better) and Normal-to-Lombard (right, higher is better) style conversions. Standard errors are shown in red. Significant differences values as measured using the Student's t-test with Bonferroni correction are highlighted.

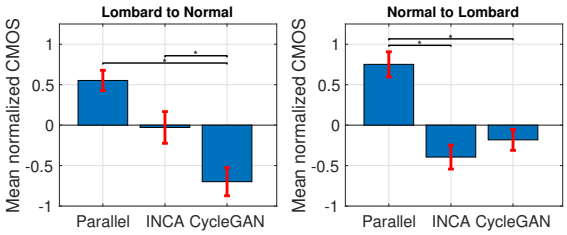

Fig. 5. Results of the Quality test for the Lombard-to-Normal (left, lower is better) and Normal-to-Lombard (right, lower is better) style conversions. Standard errors are shown in red. Significant differences values as measured using the Mann-Whitney U-test with Bonferroni correction are highlighted.

6. DISCUSSIONS AND CONCLUSION

This paper studied the use of non-parallel learning schemes to the task of vocal effort speaking style conversion, in this case between normal and Lombard speech. We compared two non-parallel methods, a new CycleGAN-based approach and INCA (a widely used method in voice conversion), to an earlier GMM-based baseline system that uses parallel data. Listening tests indicate that the CycleGAN produces encouraging results compared to the other two methods, producing the largest Lombard effect in normal-to-Lombard conversion while having indistinguishable quality from the INCA-based approach. In Lombard-to-normal conversion, the CycleGAN achieves superior speech quality to the other methods. CycleGANs should be therefore explored further in other vocal effort continuum conversion problems, as they appear to provide a strong alternative for non-parallel training on problems where parallel data scarcity is a real challenge.

7. 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] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *Journal of Speech, Language, and Hearing Research*, vol. 29, no. 4, pp. 434–446, Dec. 1986.

[3] R. M. Uchanski, S. S. Choi, L. D. Braida, C. M. Reed, and N. I. Durlach, "Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate," *Journal of Speech, Language, and Hearing Research*, vol. 39, no. 3, pp. 494–509, Jun. 1996.

[4] 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.

[5] Z. Tao, X.-D. Tan, T. Han, J.-H. Gu, Y.-S. Xu, and H.-M. Zhao, "Reconstruction of normal speech from whispered speech based on RBF neural network," in *Proc. IITSI*, Jinggangshan, China, Apr. 2010, pp. 374–377.

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

[7] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad, "Fundamental frequency generation for whisper-to-audible speech conversion," in *ICASSP, 2014*. IEEE, 2014, pp. 2579–2583.

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

[9] 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.

[10] À. Calzada and J. C. Socoró, "Vocal effort modification through harmonics plus noise model representation," in *Proc. NOLISP*, Las Palmas de Gran Canaria, Spain, Nov. 2011, pp. 96–103.

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

[12] K. I. Nordstrom, G. Tzanetakis, and P. F. Driessen, "Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1087–1096, Aug. 2008.

[13] C. d'Alessandro and B. Doval, "Experiments in voice quality modification of natural speech signals: The spectral approach," in *Proc. ESCA/COCOSDA Workshop on Speech Synthesis*, Blue Mountains, Australia, Nov. 1998, pp. 277–282.

[14] 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.

[15] S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Vocal effort based speaking style conversion using vocoder features and parallel learning," *Submitted for publication*, 2018.

[16] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[17] H. Benisty, D. Malah, and K. Crammer, "Non-parallel voice conversion using joint optimization of alignment by temporal context and spectral distortion," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7909–7913.

[18] N. Shah and H. Patil, "Effectiveness of dynamic features in inca and temporal context-inca," *Proc. Interspeech 2018*, pp. 711–715, 2018.

[19] N. Shah, M. C. Madhavi, and H. Patil, "Unsupervised vocal tract length warped posterior features for non-parallel voice conversion," in *Proc. Interspeech 2018*, 2018, pp. 1968–1972. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1712

[20] 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.

[21] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," *Proc. ICASSP 2018*, pp. 5279–5283, 2018.

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

[23] 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.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[25] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5767–5777. [Online]. Available: http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf

[26] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation."

[27] 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.

[28] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 366–378, Feb. 2009.

[29] D. Ellis, "Dynamic time warp (DTW) in Matlab," http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/., 2003, [Online; accessed 20-March-2018].

[30] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, Seattle, USA, 1998, pp. 285–288.

[31] E. Jokinen, U. Remes, M. Takanen, K. Palomäki, M. Kurimo, and P. Alku, "Spectral tilt modelling with GMMs for intelligibility enhancement of narrowband telephone speech," in *Proc. Interspeech*, Singapore, Sep. 2014, pp. 2036–2040.

[32] E. Jokinen, U. Remes, and P. Alku, "Comparison of Gaussian process regression and Gaussian mixture models in spectral tilt modelling for intelligibility enhancement of telephone speech," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, pp. 85–89.

[33] ——, "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.

[34] ETSI, "Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database," ETSI, Sophia Antipolis Cedex, France, version 1.2.4, 2011.

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

[36] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[37] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the voice conversion challenge 2016 evaluation results." in *Interspeech*, 2016, pp. 1637–1641.

[38] ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, Geneva, Switzerland, Rec. ITU-R BS.1534-3, Nov. 2015.

[39] E. Vincent, "Mushram 1.0," http://c4dm.eecs.qmul.ac.uk/downloads/, 2005, [Online; accessed 20-March-2018].

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

[41] 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.