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

Aalto University School of Electrical Engineering

ACADEMY OF FINLAND

Shreyas Seshadri[1], Lauri Juvela[1], Junichi Yamagishi[2,3], Okko Räsänen[1,4] and Paavo Alku[1]

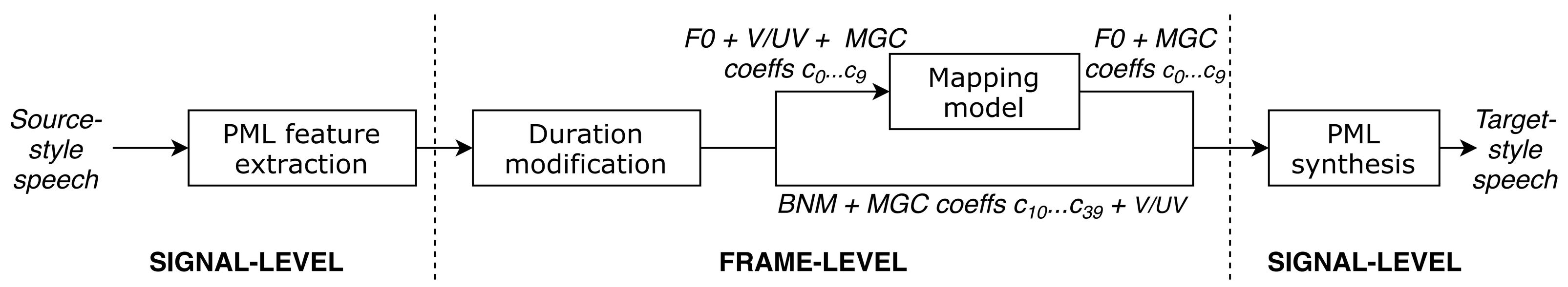
[1]Department of Signal Processing and Acoustics, Aalto University, Finland, [2]Digital Content and Media Sciences Research Division, National Institute of Informatics, Japan

[3]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom, [4]Unit of Computing Sciences, Tampere University, Finland

Goal of the study

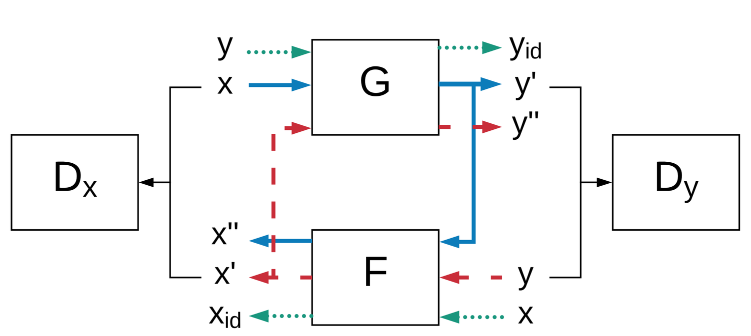
- Speaking style conversion (SSC) [1] is the technology of converting natural speech signals from one style to another.
- This study focuses on SSC for speech with varying vocal effort, focused on conversion between normal and Lombard
- We use CycleGANs [2] as a mapping model with PML vocoder features.
- The CycleGAN was compared in subjective listening tests with 2 other standard mapping methods used in conversion.

Parametric SSC system

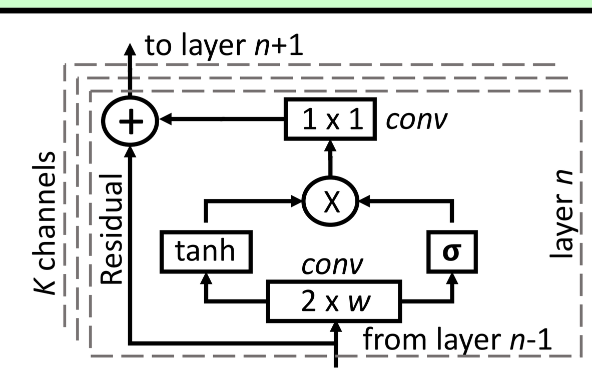


- Frame level features extracted from source signal using PML vocoder [3]: fundamental frequency, binary noise mask and spectral envelope.
- Duration modification based on characteristic voiced and unvoiced durations in each style.
- Features relevant for transformation between normal and Lombard speaking styles are transformed using a machine learning mapping model.
- Mapped features are converted to a speech waveform in the target style with the PML vocoder.

CycleGAN



Mapping functions G and F , and discriminators D_X and D_Y. The forward cycle, backward cycle, and identity mapping indicated with red, blue, and green respectively.



Block diagram of layer n of the CNN used to model G, F, D_X and D_Y.

- A CycleGAN [2] is a non-parallel learning scheme that learns bi-directional deterministic mappings.
- Trained using adversarial learning — generative models trained as a solution to a minmax two-player game between two neural networks called as the generator and discriminator.
- We use Wasserstein distance metric (WGAN loss) with gradient penalty, along with an identity mapping loss.
- The CNN shown has 8 layers and 256 channels with 11-point convolutions (similar to [4]).

Data

- Read and conversational speech recordings [5] from 20 Finnish speakers (10 female), in normal and Lombard styles.
- *Read* - each speaker read a text of 90 words (~ 1 minute).
- *Conversational* - realistic telephone conversations, where the subjects played the role of either a caller or a travel agent. Size is approximately the same as the read section.
- In order to elicit Lombard speech, background noise was played to the speakers' ears with headphones while they were being recorded.

Compared Mapping Methods

Parallel GMM

- A standard GMM is used with 8 components.
- DTW aligned features used to train frame-level models.

INCA

- Non-parallel learning scheme [6] that iteratively looks for nearest neighbor feature pairs between the source and target while also iteratively updating the conversion model to progressively improve matching to the target style.
- Same 8-component GMM model as in the parallel training.
- Algorithm is run for 10 iterations.

Subjective Evaluation

Lombardness of mapped speech

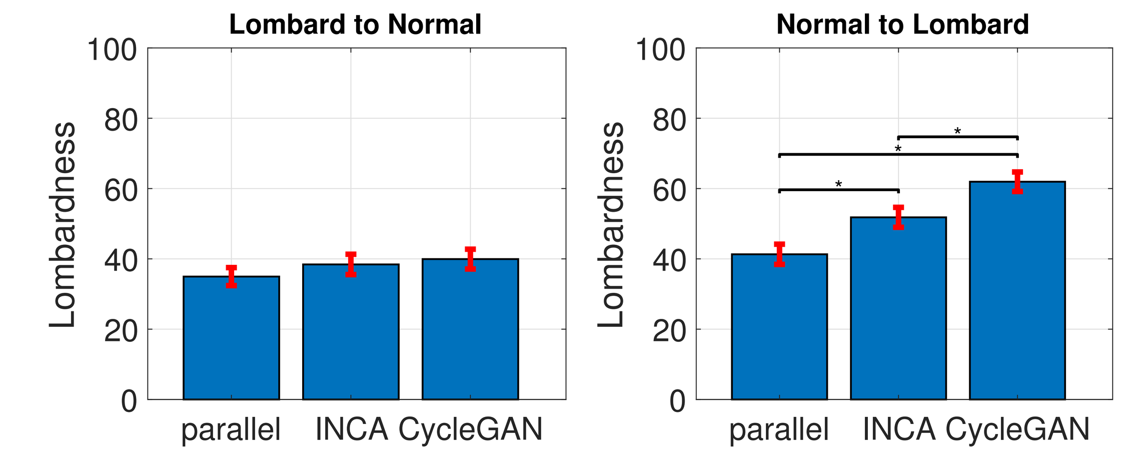
- Setup as a MUSHRA-like (MUltiple Stimuli with Hidden Reference and Anchor) test.
- Aim: evaluate the Lombardness of the mapped utterances for the normal-to-Lombard and Lombard-to-normal mappings.
- Listeners rated Lombardness of mapped samples on a scale 0–100 based on known reference samples in normal and Lombard style.

Quality of mapped speech

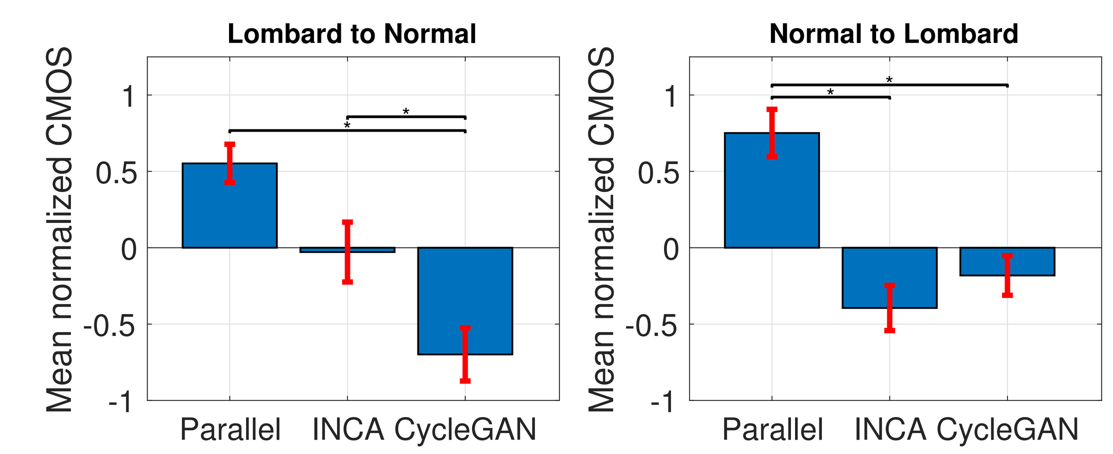
- Comparison category rating (CCR) test.
- Listeners were presented with pairs of speech utterances and asked to rate the perceived quality of the second utterance in comparison to the first one using a continuous rating scale from -3, much worse to 3, much better.
- Each utterance pair consisted of a mapped utterance and its corresponding natural Lombard utterance.
- Ratings converted to CMOS scores (smaller is better).

Results

Lombardness Test



Quality Test



Conclusions

- This work 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.
- CycleGAN produces encouraging results compared to the baseline 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, CycleGAN achieves superior speech quality to the other methods.
- CycleGANs seems like a promising candidate for SSC problems, as they appear to provide a strong alternative for non-parallel training on problems where parallel data scarcity is a real challenge.
- The implementation of the CycleGAN is available on https://github.com/shreyas253/CycleGAN_1dCNN/

References

1. 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.
2. 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.
3. 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.
4. T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," *Proc. EUSIPCO 2018*, 2018.
5. 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, Sep. 2016, pp. 2771–2775.
6. 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.