Adversarial Generation of Acoustic Waves with Pair Supervision

Hongyu Zhu
Department of Physics
hongyuz@andrew.cmu.edu

Katerina Fragkiadaki (advisor)
Machine Learning Department
katef@cs.cmu.edu

Abstract

Voice conversion (VC) has drawn people's attention for years however the current VC systems are far from good. In this paper, we explore VC systems based on deep neural networks. We show that most of the inter-gender generations sound muffled and do not carry speaker-specific characteristics such as stress, accents and emotions. In our experiment, we use a highway network as the baseline and add a discriminator to it to perform adversarial learning. We find the adversarial part greatly boosts the quality of generations by increasing the global variance. Further study on mel-cepstral distortions indicates that it is possible to make improvements by learning different coefficients following an appropriate weighting scheme.

1 Introduction

Speech signals contain a great amount of information such as contexts, identities and mood. In reality, speaker identities are important when it comes to leaving a voice message, media talks, concerts and so on. Under those circumstances, differentiating among speakers or even generating one person's speeches are particularly interesting problems. Voice conversion is a technique of transferring one person's voice style to another person's while preserving linguistic information. This technique is widely applied to a variety of fields including speaker verification [3, 4, 8, 35], speaker recognition [17, 19], text-to-speech systems [14] and speech enhancement [13, 33, 32].

Technically, voice conversion problems aim to find a good regression function which maps the source to the target speech. A variety of statistical methods have been explored such as Gaussian mixture models (GMMs, [26, 32, 9]), recurrent neural networks (RNNs, [23, 27]), convolutional neural networks (CNNs, [16]) and non-negative matrix factorization (NMF, [29, 37]).

From the perspective of signal processing, speech signals are presented as a sequence of numbers ranged from -1 to 1 and sampled in a specific rate (16 kHz in this paper). These data are in time domain and called waveforms. These data could be further broken up into chunks in the time domain. For each chunk, the Fourier transform is performed to calculate the magnitude of the frequency spectrum. In this way, speech signals are presented as a sequence of spectral vectors. This time-frequency representation is called spectrogram [5], in which phones and their properties are better observed. Additionally, the peak of the spectrums in each chunk, called formants, carry the information of the identity of the sound and constitute spectral envelopes. Spectral envelopes are extracted by inverse Fourier transforms, in which way Cepstrums are constructed. The cepstral coefficients from inverse Fourier transforms are referred to as Mel-Frequencies cepstral coefficients, denoted by MFCC. Also perceptual experiments indicate human ear concentrates more on lower frequency regions. MFCC is a powerful tool for speech recognition purpose.

This paper is organized as follows. In Section 2 we review related work of deep neural networks on voice conversion directly from extracted features. Based on the model that provides the best quality, in Section 3 we modify its architecture and describe the whole pipeline. In Section 4, we describe the experimental setups and evaluate the results. In Section 5 we talk about the lessons we learned and summarize our work in Section 6.

Machine Learning Master, Data Analysis Project (2018 Spring), CMU.

2 Related Work

There have been many results during past years especially in computer vision with the introduction of deep neural networks. However, there were much fewer generative models on audio signals until quite recent. The traditional models used to dominate over deep learning techniques as traditional features of temporal structures of audios contain so much information. Unlike pixels in computer vision, those features including speaker characteristics, linguistic, emotions and accents are much more difficult to disentangle. There have been some successful text-to-speech models Tacotron [36] and Wavenet [34] that are capable of generating very high quality human speeches since 2016, but style transforms directly from source waveforms or extracted features to target's have much worse performance in both parallel and non-parallel data.

17 groups submitted their own VC systems for paired data in Voice Conversion Challenge 2016 [31], using a data set with 216 utterances in each of the selected 5 male and 5 female speakers (see VCC summary for selected samples). Though those voice conversion systems are able to convert the fundamental frequencies from one source to one target while preserving the linguistic content, the speech style (stress) is clearly not learned and converted.

Besides, variational autoencoders (VAEs) are also applied in voice conversion problems ([11, 2]). As mentioned in variational autoencoding Wasserstein generative adversarial network (VAW-GAN) [12], VAEs simplify the problem by assuming the observed data is normally distributed and uncorrelated across dimension, leading to muffled converted voice. Therefore they incorporate a Wasserstein GAN into the decoder by assigning VAE's decoder as Wasserstein GAN's generator to form a VAW-GAN. Their features are `STRAIGHT` spectra, aperiodicities and pitch contours. samples are listed here. Unfortunately, from my subjective evaluations, the quality of VAW-GAN outputs are even worse than the previous submitted VCC VC systems.

CycleGAN-VC [15], inspired by CycleGAN [39] in the image domain, is proposed to convert unparalleled speech data using simultaneous forward and inverse learning procedures with adversarial loss and cycle-consistency loss Equation 1.

$$\mathcal{L} = \mathcal{L}^{\text{adv}}(G_{\text{source,target}}, D_{\text{target}}) + \mathcal{L}^{\text{adv}}(G_{\text{target,source}}, D_{\text{source}}) + \lambda_{\text{cyc}}\mathcal{L}^{\text{cyc}}(G_{\text{source,target}}, G_{\text{target,source}}), \quad (1)$$

where λ_{cyc} is a trade-off parameter to control cycle-consistency loss, which force the system to learn consistent contextual information. The generator uses gated 1D CNN to preserve temporal structure and allow the information to be selectively propagated by the states of the previous layer. The features they used are 24 Mel-cepstral coefficients, logarithmic fundamental frequencies, and aperiodicities extracted every 5 ms. They also provide there samples here. We can see the same problem as submitted VCC VC systems here, the speech style (stress) is not well converted.

To the best of our knowledge, the outputs with the best sound quality could be found here. [38, 24]. The source and target samples are selected to be so close to each other and they do not include inter-gender conversion. We have applied their model for inter-gender voice conversion, modified their architecture and compared different results.

3 The Model

There are 3 main steps in our model, preprocessing, deep neural networks and postprocessing. The preprocessing stage pairs the source and the target speech based on the contents, extracts the MFCCs and pitches from both source and target speeches and aligns both audios by applying dynamic time wrapping (DTW, [18]) on spectrogram features (see Figure 1. After features are extracted, the models try to learn the mapping from source features to target features in the training stage. For the evaluation, the features from test samples go through trained models and synthesize back to waveforms. Note that in this case, the model only maps the aligned training features, at the evaluation stage, since test source and target audios are not pre-aligned or silence-trimmed (see Section 3.1), the generated audio does not have to be the same as the corresponding test target.

3.1 Features

We use `PYWORLD` to extract features. We first read in waveforms, extract pitches f_0s and spectrograms. Then we trim silence from spectrograms, align source and target spectrograms using DTW and convert spectrogram to MFCCs to the 59th order. We drop 0th coefficient and append pitches to the rest 58 MFCCs. [30] compared the mel-cepstral and spectral differences between target and converted samples based on GMM and found many

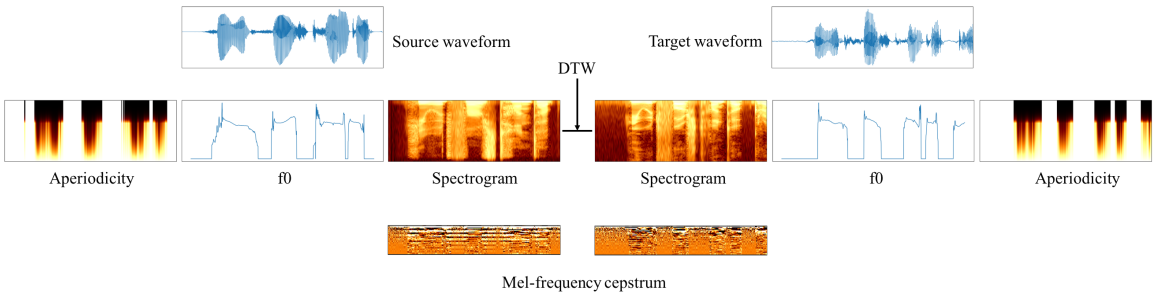

Figure 1: The preprocssing stage

mismatching local patterns. They suggested adding dynamic (delta and delta of delta) features incorporating temporal correlations among frames. Therefore we add dynamic (delta and delta of delta) features on the above 59 columns. In this way we obtain 177 features (59 static columns and 118 dynamic features) for each speech. The global minima, maxima, means and standard deviations of each column are saved separately for further re-scaling. Source and target pairs are selected based on the texts. Equation 2 is used to compute dynamic features.

$$\delta_t = \frac{\Sigma_{i=1}^{N} n\left(c_{t+i} - c_{t-i}\right)}{2\Sigma_{i=1}^{N} i^2}, \delta_t^2 = \frac{\Sigma_{i=1}^{N} n\left(\delta(c)_{t+i} - \delta(c)_{t-i}\right)}{2\Sigma_{i=1}^{N} i^2}, \tag{2}$$

where subscript t denotes time frames, c is the quantity to which the deltas are applied (in this case, ms and f_0).

3.2 Baseline

Our baseline model is a Highway feed-forward networks [25], which could be a good substitution of a much deeper network but is able to be trained efficiently using stochastic gradient descent (SGD). Highway networks regulate the information flow using transform and carry gating units.

$$\boldsymbol{y} = H(\boldsymbol{x}, \boldsymbol{W_H})T(\boldsymbol{x}, \boldsymbol{W_T}) + \boldsymbol{x}C(\boldsymbol{x}, \boldsymbol{W_C}), \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the input and output, $\boldsymbol{W_H}$ and $\boldsymbol{W_T}$ are the weights of the non-linear layer and the transform gate. Note that $\boldsymbol{y} = \boldsymbol{x}$ when carry gates are fully open $C(\boldsymbol{x}, \boldsymbol{W_C}) = 1$ transform gates are closed and $T(\boldsymbol{x}, \boldsymbol{W_T}) = 0$, $\boldsymbol{y} = H(\boldsymbol{x}, \boldsymbol{W_H})$ while carry gates are closed but transform gates are fully open.

In our baseline model, $C(\boldsymbol{x}, \boldsymbol{W_C}) = 1$, $T(\boldsymbol{x}, \boldsymbol{W_T}) = \sigma(\boldsymbol{x}_{\text{static}})$ and $H(\boldsymbol{x}, \boldsymbol{W_H})$ is a network with 4 hidden layer with hidden units 512, 256, 256, 512 and 1 last linear layer mapping last 512 hidden units to the same dimension as inputs (177, see Figure. 2). The hidden layers use ReLu (0.01) as activation functions with dropout probability 0.5. During the training process, we apply the Adagrad optimizer with learning rate 0.02 on L1 loss between the source and target feature maps. We consider L1 loss because L1 loss is generally less sensitive to outliers.

3.3 The Model

3.3.1 Generative Adversarial Network

We use the generative adversarial network (GAN) in the model. GAN is first proposed in [7]. It consists of two networks: the generator network (G) and discriminator network (D). During training the GAN, two networks are competing against each other where G is trying to generate better images to fool D and D is trying to get better in classifying whether an image is real of generated from G. During training, the two networks are optimized in an iterative manner. GANs are later extended to conditional GANs [21, 6] where the generation process can be guided by the input classes or images.

3.3.2 Multi-loss Optimization

Motivated by [20, 10], we not only apply the adversarial loss during training but also apply the L1 loss at the end of the generator, enforcing the outputs to be close to the target. The loss can be formulated as,

$$\mathcal{L}^G(\boldsymbol{y}, \boldsymbol{x}_{\text{target}}) = \lambda_{\text{static}}\mathcal{L}(\boldsymbol{y}_{\text{static}}, \boldsymbol{x}_{\text{static,target}}) + \lambda\mathcal{L}(\boldsymbol{y}, \boldsymbol{x}_{\text{target}}), \tag{4}$$

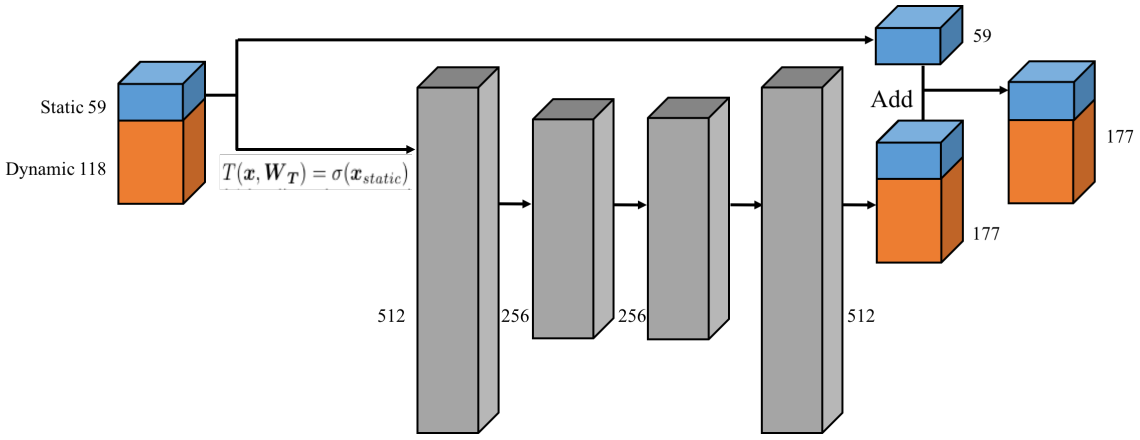

Figure 2: The baseline model

where L represents the loss in general, λ_{static} and λ are constant weights used to balance the two losses. Through our experiments, we realize that it is very important to choose λ so that the gradients back propagate from the two sides have similar norms. By combining this loss with the adversarial loss, we can represent the whole loss for the generator as,

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{x}_{\text{target}}) = \mathcal{L}^{G}(\boldsymbol{y}, \boldsymbol{x}_{\text{target}}) + \mathcal{L}^{\text{adv}}(D(\boldsymbol{x}_{\text{target}}, G(\boldsymbol{x}_{\text{source}}), 1). \tag{5}$$

3.3.3 The Architecture

Generator. We take the generator to be the same as the baseline model.

Discriminator. The discriminator is applied on the static features only (i.e. the first 59 dimensions). The generated feature maps go through 3 hidden layers, each with 256 hidden units. The output of the last layer is a single bit for the 2-class classification task.

4 Experiments

4.1 Data Sets

We run the experiments on `CMU ARCTIC` Databases, constructed by the Language Technologies Institute at Carnegie Mellon University. It consists of 1132 utterances from experienced speakers (US, Canadian, Indian, male or female) selected from out-of-copyright texts from Project Gutenberg. It is publicly available here. All recordings were recorded at 16 bit 32 kHz and then downsampled to 16 kHz. We mainly use US clb (US female) and US awb (Scottish male) to see if accents are learned in the model.

4.2 Feature Maps

Figure 3 is one sample extracted feature map. On the top panel, from the top to the bottom row, we show MFCC_1 – MFCC_{58} and f_0, followed by the magenta horizontal solid line, then $\delta(\text{MFCC}_1)$ – $\delta(\text{MFCC}_{58})$ and $\delta(f_0)$, another magenta line and $\delta^2(\text{MFCC}_1)$ – $\delta^2(\text{MFCC}_{58})$ and $\delta^2(f_0)$. The bottom panel is the magnification of the green region on the top panel, showing the first 4 MFCCs. We can see these 4 MFCCs correlated with each other and their magnitudes decrease with the order.

4.3 Training

Table 1 shows some training and test samples (use Adobe to open this pdf).

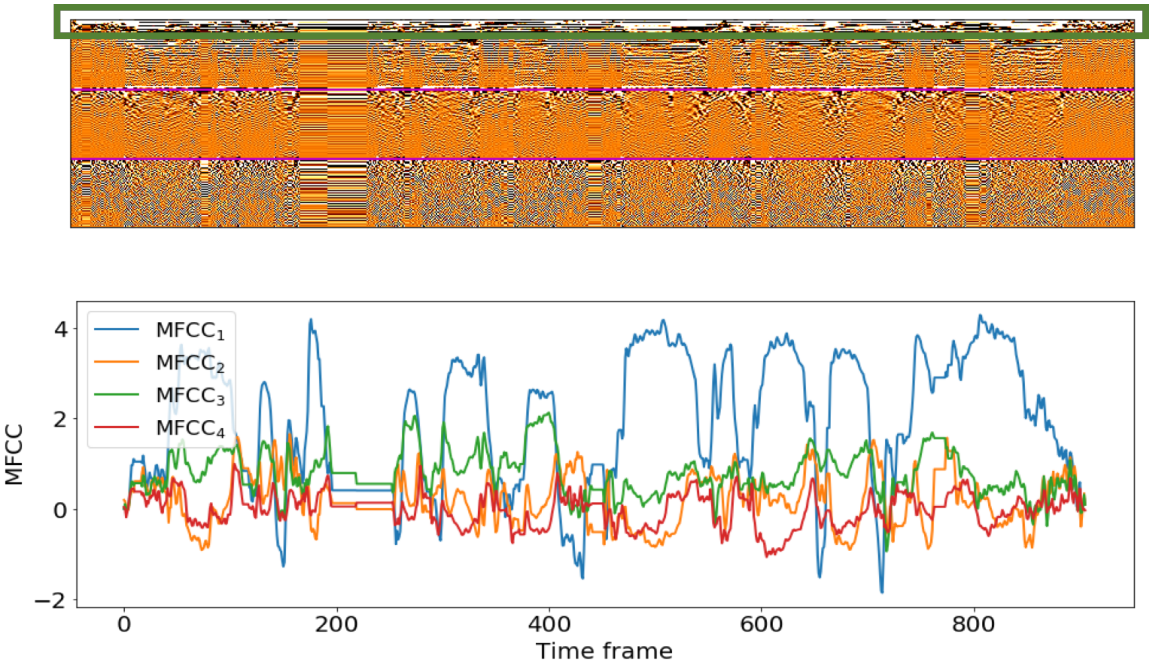

Figure 3: The feature map

Table 1: Some samples

	Source	Target	Baseline	GAN
Training				
Test				

4.4 Objective Evaluation

4.4.1 Global Variance

Many voice conversion approaches suffer from the muffling effect [1], due to over-smoothing of spectral envelopes. One approach to reduce the muffling effect is to encode the global variance (GV) in the features [30] and another one is to put a GAN [12]. Figure 4 show the effect of GAN. The top panel from left to right columns shows MFCC scatter plots of the target, VAE and VAW-GAN and the bottom panel shows MFCC scatter plots of the target, baseline and (baseline+)GAN. In both cases, adding a GAN is helpful to enhance the variances of MFCCs. The global variance could also be viewed as variances of MFCCs across different orders (see Figure 5).

4.4.2 Modulation Spectrum

Similar to GV, modulation spectrum (MS) is another feature that tends to fill the large quality gap between natural speech and synthetic speech [28]. It is defined as a the Fourier transform of the parameter sequence. It is useful in discovering the wrong behaviour across frequency bands and providing possibilities to boost performance using filtering algorithms. Here we use MS as a metric to evaluate the generations. Fig 6 shows the averaged MS over generations in the test set. GAN is very close to the target.

4.4.3 Mel-cepstral Distortion

The Mel-cepstral distortion is defined as

$$\text{MelCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(\text{mc}_d - \hat{\text{mc}}_d)^2}, \tag{6}$$

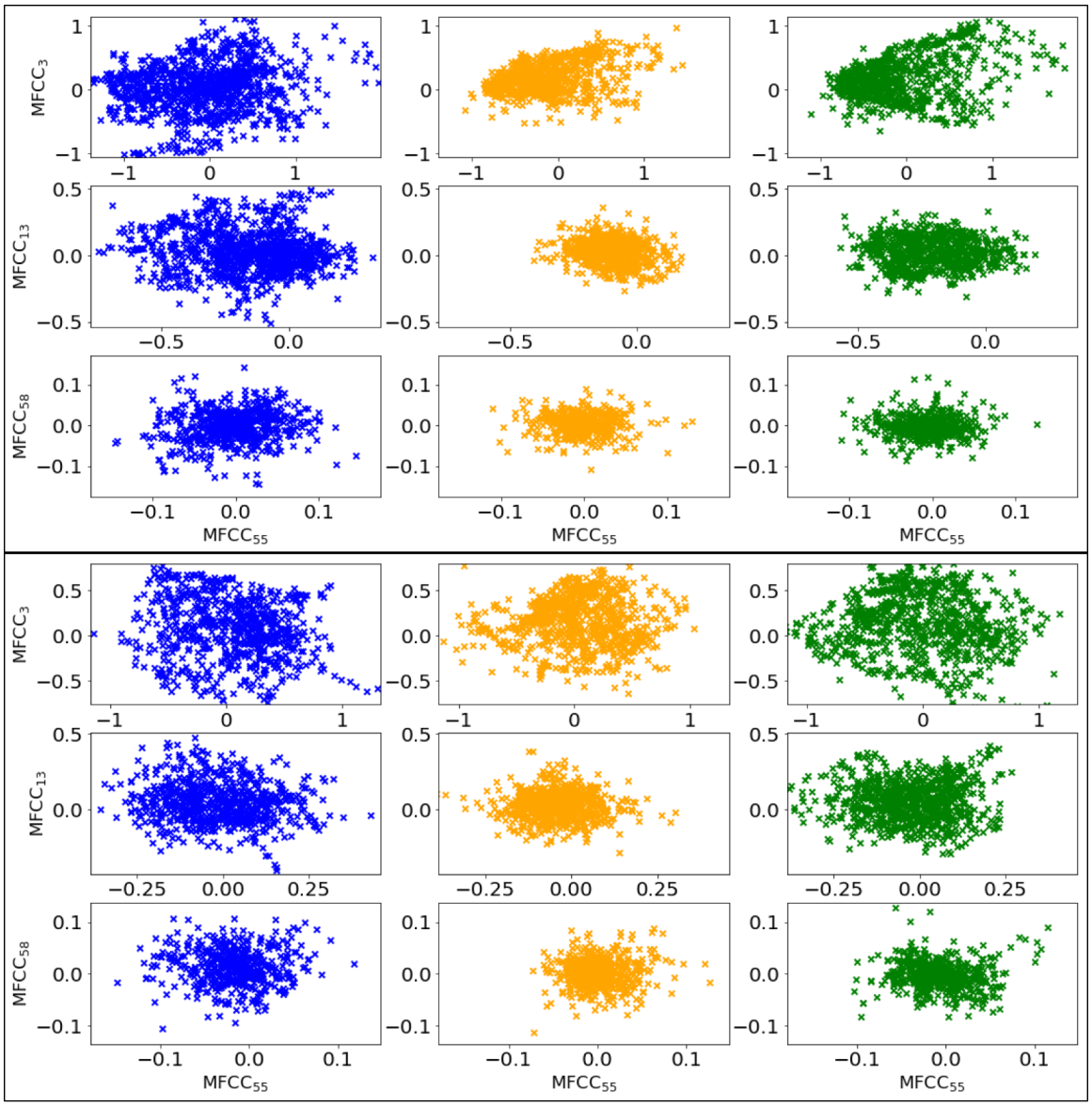

Figure 4: Scatter plots of MFCCs w/o GAN

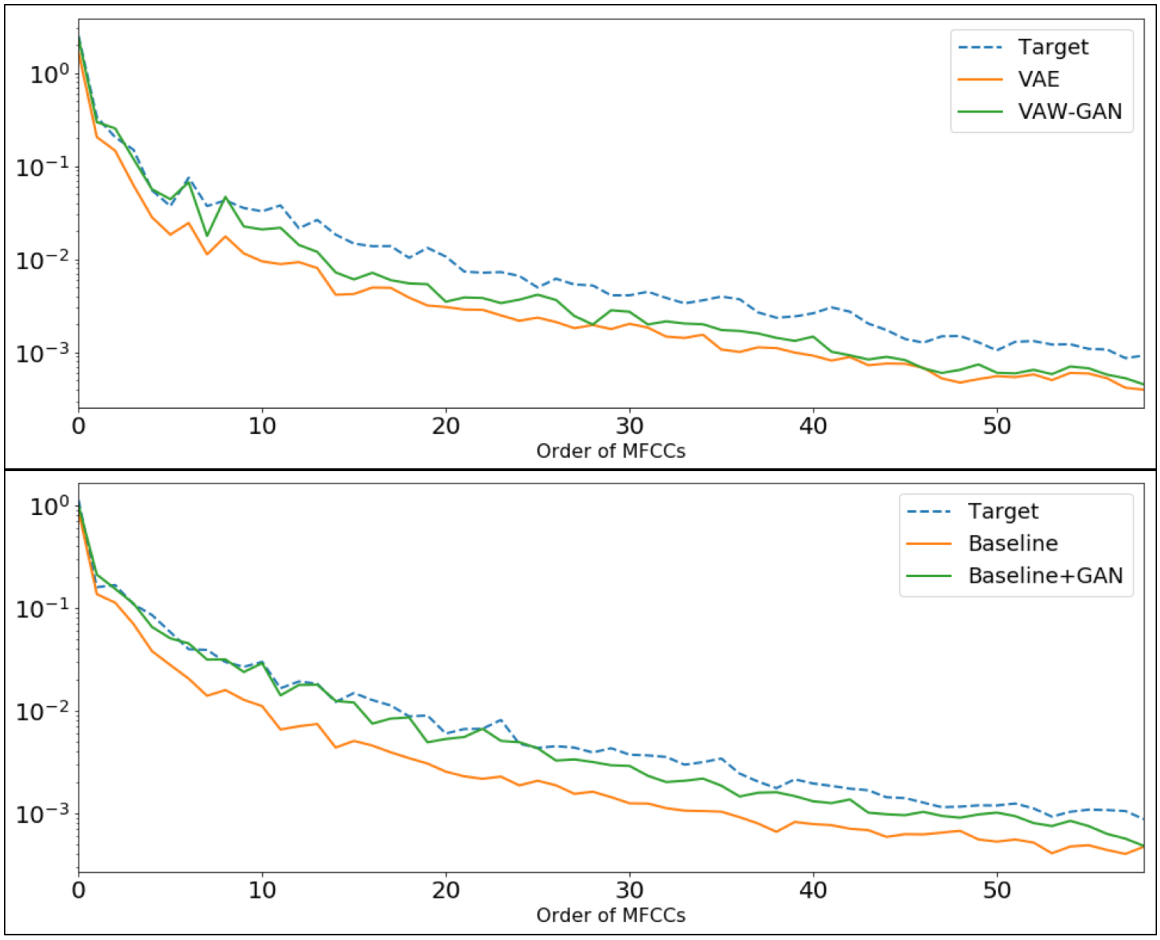

Figure 5: Global variances of MFCCs w/o GAN

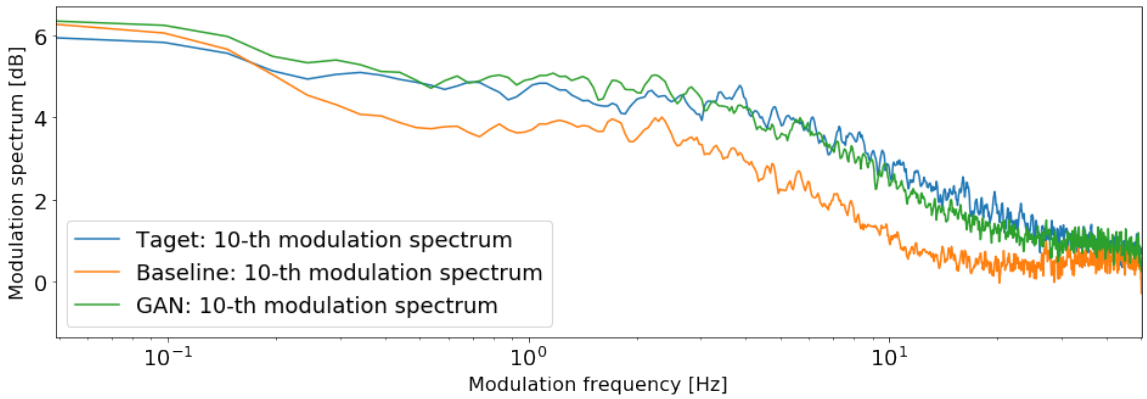

Figure 6: Modulation Spectrum

where mc_d and $\hat{\mathrm{mc}}_d$ are dth coefficients of the target and converted spectra respectively. Interestingly, though GAN gets a big performance boost in generating MFCCs, the mel-cepstral distortion is highly correlated and roughly as the same level as the baseline model (see Figure 7). This might indicate different orders of MFCCs has different significance levels to the final generation and putting different weights according to orders when learning might be a good idea. We may leave this as one of the future plans.

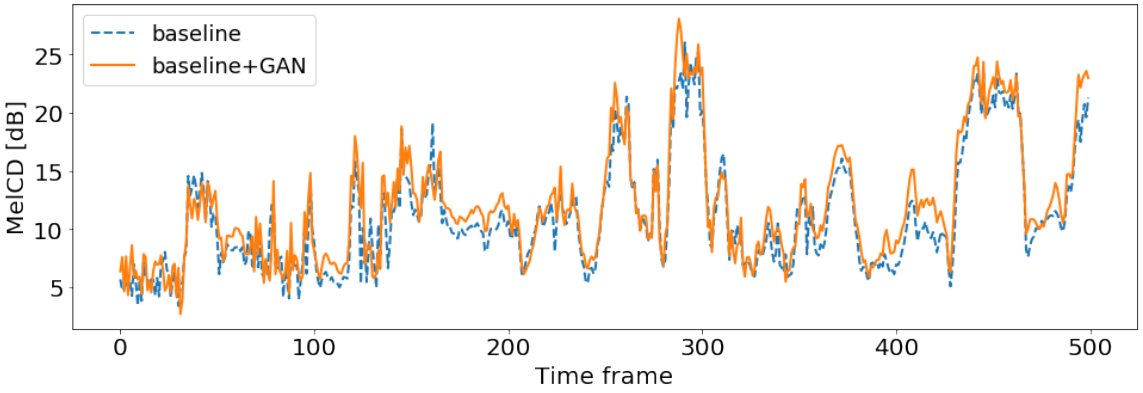

Figure 7: Mel-cepstral distortion of the audios (first 500 time frames)

5 Discussion

There are still many unsolved problems existing in VC. A major issue lies in understanding the audio domain, including waveforms, spectrograms and mel-cepstral features. As also mentioned in this blog and [22], audios should not be treated the same as images whose algorithms could make the most use of local correlations in features. Also various audio features have interpolation issues. For example if we interpolate two different vowel regions, the result may sound terribly distorted, muffled or robotic. Unlike smooth, Gaussian-like distributions of image pixels, the histograms (distributions) of audio features are spiky. Additionally, vision problems usually have high tolerance. Two images that are visually the same may have quite different RGB values but this is generally not the case with audios. Two spectrograms with slightly distorted values may lead to bad quality after synthesizing.

6 Conclusion

In this paper, we modify and apply the model [38, 24] to inter-gender samples and compare the results with those from Cycle-GAN, VAE, VAW-GAN. The (baseline+)GAN model generates the most authentic and closest audios to the target, but all the models suffer from not learning the stress, accents and emotions from targets. Although there is still a long way to go in the state-of-the-art deep VC models, we still have some takeaways from this project.

- The synthetic speech has much less dependency on aperiodicities than spectrograms, mel-ceptral and pitch contours. Also mel-ceptral is more efficient than spectrograms as they compress most information of spectrograms into the first few coefficients.
- GANs are very useful in enhancement of GV.
- TTS models perform much better than VC models as they have generic features on the natural pronunciations of targets.
- Audio features are not like image pixels. Breakthroughs are needed in understanding and modelling stress, emotions and accents.

References

[1] Hadas Benisty and David Malah. Voice conversion using gmm with enhanced global variance. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[2] Merlijn Blaauw and Jordi Bonada. Modeling and transforming speech using variational autoencoders. In *INTERSPEECH*, pages 1770–1774, 2016.

[3] William M Campbell, Douglas E Sturim, Douglas A Reynolds, and Alex Solomonoff. Svm based speaker verification using a gmm supervector kernel and nap variability compensation. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.

[4] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[5] John R Deller Jr, John G Proakis, and John H Hansen. *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[6] E Denton, S Chintala, A Szlam, and R Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[8] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. End-to-end text-dependent speaker verification. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5115–5119. IEEE, 2016.

[9] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj. Voice conversion using partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):912–921, 2010.

[10] William Seto Katerina Fragkiadaki Hsiao-Yu Fish Tung, Adam W Harley. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[11] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from non-parallel corpora using variational auto-encoder. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*, pages 1–6. IEEE, 2016.

[12] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv preprint arXiv:1704.00849*, 2017.

[13] Zeynep Inanoglu and Steve Young. Data-driven emotion conversion in spoken english. *Speech Communication*, 51(3):268–283, 2009.

[14] Alexander Kain and Michael W Macon. Spectral voice conversion for text-to-speech synthesis. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 285–288. IEEE, 1998.

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

[16] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino. Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks. In *Proc. INTERSPEECH*, pages 1283–1287, 2017.

[17] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447, 2007.

[18] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005.

[19] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1695–1699. IEEE, 2014.

[20] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *CoRR*, abs/1511.05440, 2015.

[21] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[22] Seyed Hamidreza Mohammadi and Alexander Kain. An overview of voice conversion systems. *Speech Communication*, 88:65–82, 2017.

[23] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[24] Yuki Saito, Shinnosuke Takamichi, Hiroshi Saruwatari, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Statistical parametric speech synthesis incorporating generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(1):84–96, 2018.

[25] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[26] Yannis Stylianou, Olivier Cappé, and Eric Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on speech and audio processing*, 6(2):131–142, 1998.

[27] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4869–4873. IEEE, 2015.

[28] Shinnosuke Takamichi, Tomoki Toda, Graham Neubig, Sakriani Sakti, and Satoshi Nakamura. A postfilter to modify the modulation spectrum in hmm-based speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 290–294. IEEE, 2014.

[29] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Exemplar-based voice conversion using sparse representation in noisy environments. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 96(10):1946–1953, 2013.

[30] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235, 2007.

[31] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi. The voice conversion challenge 2016. In *INTERSPEECH*, pages 1632–1636, 2016.

[32] Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2505–2517, 2012.

[33] Oytun Turk and Marc Schroder. Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):965–973, 2010.

[34] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *CoRR*, abs/1609.03499, 2016.

[35] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. Deep neural networks for small footprint text-dependent speaker verification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4052–4056. IEEE, 2014.

[36] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech syn. *arXiv preprint arXiv:1703.10135*, 2017.

[37] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li. Exemplar-based sparse representation with residual compensation for voice conversion. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(10):1506–1521, 2014.

[38] Shan Yang, Lei Xie, Xiao Chen, Xiaoyan Lou, Dongyan Huang, and Haizhou Li. Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework. *arXiv preprint arXiv:1707.01670*, 2017.

[39] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.