Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks

Chin-Cheng Hsu[1], Hsin-Te Hwang[1], Yi-Chiao Wu[1], Yu Tsao[2], and Hsin-Min Wang[1]

[1]Institute of Information Science, Academia Sinica, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taiwan

{jeremycchsu, hwanght, tedwu, whm}@iis.sinica.edu.tw, yu.tsao@citi.sinica.edu.tw

Abstract

Building a voice conversion (VC) system from non-parallel speech corpora is challenging but highly valuable in real application scenarios. In most situations, the source and the target speakers do not repeat the same texts or they may even speak different languages. In this case, one possible, although indirect, solution is to build a generative model for speech. Generative models focus on explaining the observations with latent variables instead of learning a pairwise transformation function, thereby bypassing the requirement of speech frame alignment. In this paper, we propose a non-parallel VC framework with a variational autoencoding Wasserstein generative adversarial network (VAW-GAN) that explicitly considers a VC objective when building the speech model. Experimental results corroborate the capability of our framework for building a VC system from unaligned data, and demonstrate improved conversion quality.

Index Terms: non-parallel voice conversion, Wasserstein generative adversarial network, GAN, variational autoencoder, VAE

1. Introduction

The primary goal of voice conversion (VC) is to convert the speech from a source speaker to that of a target, without changing the linguistic or phonetic content. However, consider the case of converting one's voice into that of another who speaks a different language. Traditional VC techniques would have trouble dealing with such cases because most of them require *parallel* training data in which many pairs of speakers uttered the same texts. In this paper, we are devoted to bridging the gap between *parallel* and *non-parallel* VC systems.

We pursue a unified generative model for speech that naturally accommodates VC (Sec. 2). In this framework, we do not have to align any frames or to cluster phones or frames explicitly. The idea is skeletonized in the probabilistic graphical model (PGM) in Fig. 1. In this model, our attention is directed away from seeking alignment. Rather, what we are concerned about are 1) finding a good inference model (Sec. 2.1) for the latent variable z and 2) building a good synthesizer whose outputs match the distribution of the real speech of the target (Sec. 2.2 through 2.3). In this paper, we present a specific implementation in which a variational autoencoder (VAE [1]) assumes the inference task and a Wasserstein generative adversarial network (W-GAN [2]) undertakes speech synthesis. Our contribution is two-fold:

- We introduce the W-GAN to non-parallel voice conversion and elucidate the reason why it fits this task (Sec. 2).
- We demonstrate the ability of W-GAN to synthesize more realistic spectra (Sec. 3).
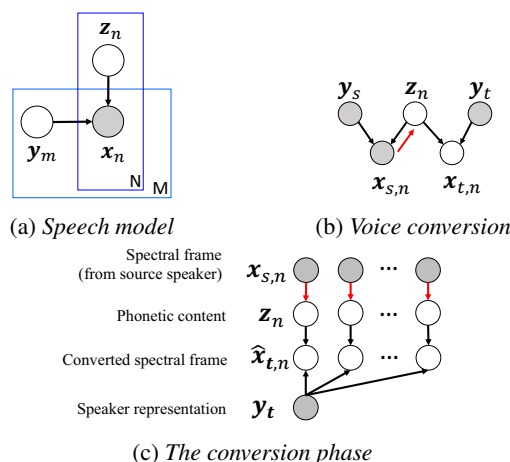

(a) *Speech model* (b) *Voice conversion*

(c) *The conversion phase*

Figure 1: *Graphical models of our voice conversion framework. Red arrows represent inference. During the conversion phase, we first infer the latent variable z and blend it with the target speaker's representation y.*

2. Non-parallel voice conversion via deep generative models

Given spectral frames $X_s = \{\boldsymbol{x}_{s,n}\}_{n=1}^{N_s}$ from the source speaker and those $X_t = \{\boldsymbol{x}_{t,n'}\}_{n'=1}^{N_t}$ from the target, assume that the real data distributions of the source and the target respectively admit a density p_s^* and p_t^*. Let f be a voice conversion function that induces a conditional distribution $p_{t|s}$. The goal of VC is to estimate f so that $p_{t|s}$ best approximates the real data distribution p_t^*:

$$p_{t|s}(f(\boldsymbol{x}_{s,n})) \approx p_t^*(\boldsymbol{x}_{t,n'}). \tag{1}$$

We can decompose the VC function f into two stages according to the PGM in Fig. 1b. In the first stage, a speaker-independent encoder $\mathcal{E}_\phi$ infers a latent content $\boldsymbol{z}_n$. In the second stage, a speaker dependent decoder $\mathcal{G}_\theta$ mixes $\boldsymbol{z}_n$ with a speaker-specific variable $\boldsymbol{y}$ to reconstruct the input. The problem of VC is then reformulated as:

$$\boldsymbol{x}_{t,n'} \approx f(\boldsymbol{x}_{s,n}) = \mathcal{G}_\theta(\mathcal{E}_\phi(\boldsymbol{x}_{s,n}), \boldsymbol{y}). \tag{2}$$

In short, this model explains the observation $\boldsymbol{x}$ using two latent variables $\boldsymbol{y}$ and $\boldsymbol{z}$. We will drop the frame indices whenever readability is unharmed. We refer to $\boldsymbol{y}$ as the *speaker representation* vector because it is determined solely by the speaker identity. We refer to $\boldsymbol{z}$ as the *phonetic content* vector because with a fixed $\boldsymbol{y}$, we can generate that speaker's voice by varying

z. Note that the term *phonetic content* is only valid in the context of our experimental settings where the speech is natural, noise-free, and non-emotional.

This encoder-decoder architecture facilitates VC from unaligned or non-parallel corpora. The function of the encoder is similar to a phone recognizer whereas the decoder operates as a synthesizer. The architecture enables voice conversion for the following reasons. The speaker representation y can be obtained from training. The encoder can infer the phonetic content z. The synthesizer can reconstruct any spectral frame x with the corresponding y and z. Combining these elements, we can build a non-parallel VC system via optimizing the encoder, the decoder (synthesizer), and the speaker representation. With the encoder, frame-wise alignment is no longer needed; frames that belong to the same phoneme class now hinge on a similar z. With this conditional synthesizer, VC becomes as simple as replacing the speaker representation y. (as illustrated in Fig. 1c).

We delineate our proposed method incrementally in three subsections: a conditional variational autoencoder (C-VAE) in Sec. 2.1, a generative adversarial nets (GAN) applied to improve the over-simplified C-VAE model in Sec. 2.2, and a Wasserstein GAN (W-GAN) that explicitly considers VC in the training objectives in Sec. 2.3.

2.1. Modeling speech with a C-VAE

Recent works have proven the viability of speech modeling with VAEs [3, 4]. A C-VAE that realizes the PGM in Fig. 1a maximizes a variational lower bound of the log-likelihood:

$$\log p_\theta(\boldsymbol{x}|\boldsymbol{y}) \leq -J_{vae}(\boldsymbol{x}|\boldsymbol{y}) = -(J_{obs}(\boldsymbol{x}|\boldsymbol{y}) + J_{lat}(x)), \quad (3)$$

$$J_{lat}(\phi; \boldsymbol{x}) = \mathcal{D}_{\mathrm{KL}}\big(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p_\theta(\boldsymbol{z})\big), \quad (4)$$

$$J_{obs}(\phi, \theta; \boldsymbol{x}, \boldsymbol{y}) = -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})\big], \quad (5)$$

where $\boldsymbol{x} \in X_s \cup X_t$, $\mathcal{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence, $p_\theta(\boldsymbol{z})$ is our prior distribution model of $\boldsymbol{z}$, $p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})$ is our synthesis model, and $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is our inference model. Note that the synthesis is conditioned on an extra input $\boldsymbol{y}$, thus the name conditional VAE.

In order to train the C-VAE, we have to simplify the model in several aspects. First, we choose $p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})$ to be a normal distribution whose covariance is an identity matrix. Second, we choose $p_\theta(\boldsymbol{z})$ to be a standard normal distribution. Third, the expectation over $\boldsymbol{z}$ is approximated by sampling methods. With these simplifications, we can avoid intractability and focus on modeling the statistics of the Gaussian distributions.

For the phonetic content $\boldsymbol{z}$, we have:

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z} \mid \mathcal{E}_{\phi_1}(\boldsymbol{x}), \, diag(\mathcal{E}_{\phi_2}(\boldsymbol{x}))), \quad (6)$$

where $\phi_1 \cup \phi_2 = \phi$ are the parameters of the encoder, and $\mathcal{E}_{\phi_1}$ and $\mathcal{E}_{\phi_2}$ are the inference models of mean and variance. For the reconstructed or converted spectral frames, we have:

$$p_\theta(\boldsymbol{x}_{m,n}|\boldsymbol{z}_n, \boldsymbol{y}_m) = \mathcal{N}(\boldsymbol{x}_{m,n} \mid \mathcal{G}_\theta(\boldsymbol{z}_n, \boldsymbol{y}_m), \boldsymbol{I}). \quad (7)$$

Training this C-VAE means maximizing (3). For every input $(\boldsymbol{x}_n, \boldsymbol{y}_n)$, we can sample the latent variable $\boldsymbol{z}_n$ using the re-parameterization trick described in [1]. With $\boldsymbol{z}_n$ and $\boldsymbol{y}_n$, the model can reconstruct the input, and by replacing $\boldsymbol{y}_n$, it can convert voice. This means that we are building virtually multiple models in one. Conceptually, the speaker switch lies in the speaker representation $\boldsymbol{y}$ because the synthesis is conditioned on $\boldsymbol{y}$.

2.2. Improving speech models with GANs

Despite the effectiveness of C-VAE, the simplification induces inaccuracy in the synthesis model. This defect originates from the fallible assumption that the observed data is normally distributed and uncorrelated across dimensions. This assumption gave us a defective learning objective, leading to muffled converted voices. Therefore, we are motivated to resort to models that side-step this defect.

We can improve the C-VAE by incorporating a GAN objective [5] into the decoder. A vanilla GAN [6] consists of two components: a generator (synthesizer) $\mathcal{G}_\theta$ that produces realistic spectrum and a discriminator $\mathcal{D}_\psi$ that judges whether an input is a true spectrum or a generated one. These two components seek an equilibrium in a min-max game with the Jensen-Shannon divergence $\mathcal{D}_{\mathrm{JS}}$ as the objective, which is defined as follows:

$$J_{gan}(\theta, \psi; \boldsymbol{x}) = 2\,\mathcal{D}_{\mathrm{JS}}\big(p_t^*\|p_{t|s}\big) + 2\log 2.$$

$$= \mathbb{E}_{\boldsymbol{x}\sim p_t^*}\Big[\log \frac{p_t^*}{p_t^* + p_{t|s}}\Big] + \mathbb{E}_{\boldsymbol{x}\sim p_{t|s}}\Big[\log \frac{p_{t|s}}{p_t^* + p_{t|s}}\Big]$$

$$= \mathbb{E}_{\boldsymbol{x}\sim p_t^*}\big[\log \mathcal{D}_\psi^*(\boldsymbol{x})\big] + \mathbb{E}_{\boldsymbol{z}\sim q_\phi}\big[\log \big(1 - \mathcal{D}_\psi^*(\mathcal{G}_\theta(\boldsymbol{z}))\big)\big], \quad (8)$$

where $\mathcal{D}_\psi^*$ denotes the optimal discriminator, which is the density ratio in the second equality in (8). We can view this as a density ratio estimation problem without explicit specification of distributions.

Presumably, GANs produce sharper spectra because they optimize a loss function between two distributions in a more direct fashion. We can combine the objectives of VAE and GAN by assigning VAE's decoder as GAN's generator to form a VAE-GAN [5]. However, the VAE-GAN does not consider VC explicitly. Therefore, we propose our final model: variational autoencoding Wasserstein GAN (VAW-GAN).

2.3. Direct consideration of voice conversion with W-GAN

A deficiency in the VAE-GAN formulations is that it treats VC indirectly. We simply assume that when the model is well-trained, it naturally equips itself with the ability to convert voices. In contrast, we can directly optimize a non-parallel VC loss by renovating $\mathcal{D}_{\mathrm{JS}}$ with a Wasserstein objective [2].

2.3.1. Wasserstein distance

The Wasserstein-1 distance is defined as follows:

$$W(p_t^*, p_{t|s}) = \inf_{\gamma \in \Pi(p_t^*, p_{t|s})} \mathbb{E}_{(\boldsymbol{x}, \hat{\boldsymbol{x}})\sim\gamma}\big[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|\big] \quad (9)$$

where $\Pi(p, q)$ denotes the set of all joint distributions $\gamma(\boldsymbol{x}, \hat{\boldsymbol{x}})$ whose marginals are respectively p_t^* and $p_{t|s}$. According to the definition, the Wasserstein distance is calculated from the optimal transport, or the best frame alignment. Note that (9) is thus suitable for parallel VC.

On the other hand, the Kantorovich-Rubinstein duality [7] of (9) allows us to explicitly approach non-parallel VC:

$$W(p_t^*, p_{t|s}) = \sup_{\|\mathcal{D}\|_L \leq 1} \mathbb{E}_{\boldsymbol{x}\sim p_t^*}\big[\mathcal{D}(\boldsymbol{x})\big] - \mathbb{E}_{\boldsymbol{x}\sim p_{t|s}}\big[\mathcal{D}(\boldsymbol{x})\big], \quad (10)$$

where the supremum is over all 1-Lipschitz functions $\mathcal{D} : \mathcal{X} \to R$. If we have a parameterized family of functions $\mathcal{D}_{\psi\in\Psi}$ that are all K-Lipschitz for some K, we could consider solving the problem:

$$\max_{\psi\in\Psi} J_{wgan}, \quad (11)$$

Algorithm 1 VAE-WGAN training

function AUTOENCODE(X, y)
    $Z_\mu \leftarrow \mathcal{E}_{\phi_1}(X)$
    $Z_\sigma \leftarrow \mathcal{E}_{\phi_2}(X)$
    $Z \leftarrow$ sample from $\mathcal{N}(Z_\mu, Z_\sigma)$
    $X' \leftarrow \mathcal{G}_\theta(Z, y)$
    return X', Z

$\phi, \boldsymbol{\theta}, \boldsymbol{\psi} \leftarrow$ initialization
while not converged **do**
    $X_s \leftarrow$ mini-batch of random samples from source
    $X_t \leftarrow$ mini-batch of random samples from target
    $X'_s, Z_s \leftarrow$ AUTOENCODE(X_s, y_s)
    $X'_t, Z_t \leftarrow$ AUTOENCODE(X_t, y_t)
    $X_{t|s} \leftarrow \mathcal{G}_\theta(Z_s, y_t)$
    $J_{obs} \leftarrow J_{obs}(X_s) + J_{obs}(X_t)$
    $J_{lat} \leftarrow J_{lat}(Z_s) + J_{lat}(Z_t)$
    $J_{wgan} \leftarrow J_{wgan}(X_t, X_s)$

    // Update the encoder, generator, and discriminator
    while not converged **do**
        $\boldsymbol{\psi} \xleftarrow{update} -\nabla_\psi(-J_{wgan})$
      $\phi \xleftarrow{update} -\nabla_\phi(J_{obs} + J_{lat})$
      $\boldsymbol{\theta} \xleftarrow{update} -\nabla_\theta(J_{obs} + \alpha J_{wgan})$

where J_{wgan} is defined as

$$\mathbb{E}_{\boldsymbol{x} \sim p_t^*}\big[\mathcal{D}_\psi(\boldsymbol{x})\big] - \mathbb{E}_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\mathcal{D}_\psi(\mathcal{G}_\theta(\boldsymbol{z}), \boldsymbol{y}_t)\big] \qquad (12)$$

Alignment is not required in this formulation because of the respective expectations. What we need now is a batch of real frames from the target speaker, another batch of synthetic frames converted from the source into the target speaker, and a good discriminator $\mathcal{D}_\psi$.

2.3.2. VAW-GAN

Incorporating the W-GAN loss (12) with (3) yields our final objective:

$$
\begin{aligned}
J_{vawgan} = &- \mathcal{D}_{\text{KL}}\big(q_\phi(\boldsymbol{z}_n|\boldsymbol{x}_n)\|p(\boldsymbol{z}_n)\big) \\
&+ \mathbb{E}_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})\big] \\
&+ \alpha \, \mathbb{E}_{\boldsymbol{x} \sim p_t^*}\big[\mathcal{D}_\psi(\boldsymbol{x})\big] \\
&- \alpha \, \mathbb{E}_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\big[\mathcal{D}_\psi(\mathcal{G}_\theta(\boldsymbol{z}, \boldsymbol{y}_t))\big]
\end{aligned}
\qquad (13)
$$

where α is a coefficient which emphasizes the W-GAN loss. This objective is shared across all three components: the encoder, the synthesizer, and the discriminator. The synthesizer minimizes this loss whereas the discriminator maximizes it; consequently the two components have to be optimized in alternating order. For clarity, we summarize the training procedures in Alg. 1. Note that we actually use an update schedule for $\mathcal{D}_\psi$ instead of training it to real optimality.

3. Experiments

3.1. The dataset

The proposed VC system was evaluated on the Voice Conversion Challenge 2016 dataset [8]. The dataset was a parallel speech corpus; however, frame alignment was not performed in the following experiments.

We conducted experiments on a subset of 3 speakers. In the inter-gender experiment, we chose SF1 as the source and TM3 as the target. In the intra-gender experiment, we chose TF2 as the target. We used the first 150 utterances (around 10 minutes) per speaker for training, the succeeding 12 for validation, and 25 (out of 54) utterances in the official testing set for subjective evaluations.

3.2. The feature set

We used the STRAIGHT toolkit [9] to extract speech parameters, including the STRAIGHT spectra (SP for short), aperiodicity (AP), and pitch contours (F0). The rest of the experimental settings were the same as in [3], except that we rescaled log energy-normalized SP (denoted by $logSP_{en}$) to the range of $[-1, 1]$ dimension-wise. Note that our system performed frame-by-frame conversion without post-filtering and that we utilized neither contextual nor dynamic features in our experiments.

3.3. Configurations and hyper-parameters

The baseline system was the C-VAE system (denoted simply as VAE) [3] because its performance had been proven to be on par with another simple parallel baseline. In our proposed system, the encoder, the synthesizer, and the discriminator were convolutional neural networks. The phonetic space was 64-dimensional and assumed to have a standard normal distribution. The speaker representation were one-hot coded, and their embeddings were optimized as part of the generator parameters[1].

3.4. The training and conversion procedures

We first set α to 0 to exclude W-GAN, and trained the VAE till convergence to get the baseline model. Then, we proceeded on training the whole VAW-GAN via setting α to 50.

Conversion was conducted on a frame-by-frame basis as shown in Fig. 1c. First, $\mathcal{E}_\phi$ inferred the phonetic content $\boldsymbol{z}_n$ from $\boldsymbol{x}_{s,n}$ Then, we specified a speaker identity (integer, the subscript t in $\boldsymbol{y}_t$) that retrieved the speaker representation vector $\boldsymbol{y}$. The synthesizer $\mathcal{G}_\theta$ then generated a conditional output frame $\hat{\boldsymbol{x}}$ using $\boldsymbol{z}_n$ and $\boldsymbol{y}_t$.

3.5. Subjective evaluations

Five-point mean opinion score (MOS) tests were conducted in a pairwise manner. Each of the 10 listeners graded the pairs of outputs from the VAW-GAN and the VAE. Inter-gender and intra-gender VC were evaluated respectively.

The MOS results on naturalness shown in Fig. 2 demonstrate that VAW-GAN significantly outperforms the VAE baseline (p-value $\ll 0.01$ in paired t-tests). The results are in accordance with the converted spectra shown in Fig. 3, where the output spectra from VAW-GAN express richer variability across the frequency axis, hence reflecting clearer voices and enhanced intelligibility.

We did not report objective evaluations such as mean melcepstral coefficients because we found inconsistent results with the subjective evaluations. However, similar inconsistency is common in the VC literature because it is highly likely that those evaluations are inconsistent with human auditory systems [10, 11, 12]. The performance of speaker similarity was also

---

[1]Due to space limitations, the rest of the specification of hyper-parameters and audio samples can be found on-line: https://github.com/JeremyCCHsu/vc-vawgan
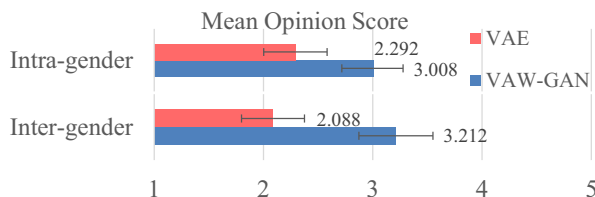
Figure 2: *MOS on naturalness. The source is SF1, and the targets are TF2 and TM3.*
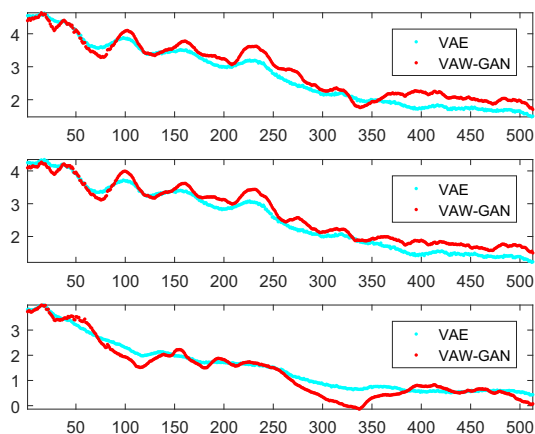

Figure 3: *Selected frames of the STRAGIHT spectra converted from SF1 to TM3. The spectral envelopes from the VAW-GAN outputs are less smooth across the frequency axis.*

unreported because we found that it remained about the same as that of [13] (System B in [14]).

4. Discussions

4.1. W-GAN improved spectrum modeling

As we can see in Fig. 3, the spectral envelopes of the synthetic speech from VAW-GAN are more structured, with more observable peaks and troughs. Spectral structures are key to the speech intelligibility, indirectly contributing to the elevated MOS. In addition, the more detailed spectral shapes in the high-frequency region reflect clearer (non-muffled) voice of the synthetic speech.

4.2. W-GAN as a variance modeling alternative

The Wasserstein objective in (13) is minimized when the distribution of the converted spectrum $p_{t|s}$ is closest to the true data distribution p_t^*. Unlike VAE that assumes a Gaussian distribution on the observation, W-GAN models the observation implicitly through a series of stochastic procedures, without prescribing any density forms. In Fig. 4, we can observe that the output spectra of the VAW-GAN system have larger variance compared to those of the VAE system. The global variance (GV) of the VAW-GAN output may not be as good as that of the data but the higher values indicate that VAW-GAN does not centralize predicted values at the mean too severely. Since speech has a highly diverse distribution, it requires more sophisticated analysis on this phenomenon.
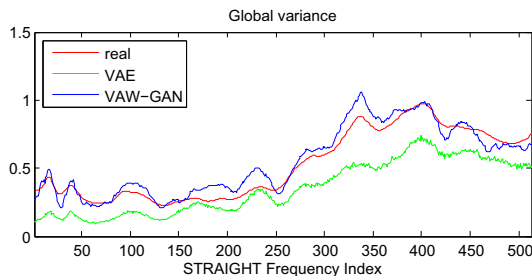

Figure 4: *The GV computed from the $logSP_{en}$ over all non-silent frames from speaker TM3.*

4.3. Imperfect speaker modeling in VAW-GAN

The reason that the speaker similarity of the converted voice is not improved reminds us of the fact that both VAE and VAW-GAN optimize the same PGM, thus the same speaker model. Therefore, modeling speaker with one global variable might be insufficient. As modeling speaker with a frame-wise variable may conflict with the phonetic vector z, we may have to resort to other PGMs. We will investigate this problem in the future.

5. Related work

To handle *non-parallel* VC, many researchers resort to frame-based, segment-based, or cluster-based alignment schemes. One of the most intuitive ways is to apply an automatic speech recognition (ASR) module to the utterances, and proceed with explicit alignment or model adaptation [15, 16]. The ASR module provides every frame with a phonetic label (usually the phonemic states). It is particularly suitable for text-to-speech (TTS) systems because they can readily utilize these labeled frames [17]. A shortcoming with these approaches is that they require an extra mapping to realize cross-lingual VC. To this end, the INCA-based algorithms [18, 19] were proposed to iteratively seek frame-wise correspondence using converted surrogate frames. Another attempt is to separately build frame clusters for the source and the target, and then set up a mapping between them [20].

Recent advances include [21], in which the authors exploited i-vectors to represent speakers. Their work differed from ours in that they adopted explicit alignment during training. In [22], the authors represented the phonetic space with senone probabilities outputted from an ASR module, and then generated voice by means of a TTS module. Despite differences in realization, our models do share some similarity ideally.

6. Conclusions

We have presented a voice conversion framework that is able to directly incorporate a non-parallel VC criterion into the objective function. The proposed VAW-GAN framework improves the outputs with more realistic spectral shapes. Experimental results demonstrate significantly improved performance over the baseline system.

7. Acknowledgements

8. References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013. [Online]. Available: http://arxiv.org/abs/1312.6114

[2] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *CoRR*, vol. abs/1701.07875, 2017. [Online]. Available: http://arxiv.org/abs/1701.07875

[3] C. Hsu, H. Hwang, Y. Wu, Y. Tsao, and H. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016, Jeju, South Korea, December 13-16, 2016*. IEEE, 2016, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/APSIPA.2016.7820786

[4] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 1770–1774. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1183

[5] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, M. Balcan and K. Q. Weinberger, Eds., vol. 48. JMLR.org, 2016, pp. 1558–1566. [Online]. Available: http://jmlr.org/proceedings/papers/v48/larsen16.html

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial networks," *CoRR*, vol. abs/1406.2661, 2014. [Online]. Available: http://arxiv.org/abs/1406.2661

[7] C. Villani, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Berlin: Springer, 2009.

[8] T. Toda, L. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 1632–1636. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1066

[9] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, no. 3-4, pp. 187–207, 1999.

[10] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[11] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, pp. 1506–1521, 2014.

[12] H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "A probabilistic interpretation for artificial neural network-based voice conversion," *Proc. APSIPA*, 2015.

[13] Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally linear embedding for exemplar-based spectral conversion," *Proc. INTERSPEECH*, in press.

[14] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the voice conversion challenge 2016 evaluation results," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 1637–1641. [Online]. Available: https://doi.org/10.21437/Interspeech.2016-1331

[15] M. Dong, C. Yang, Y. Lu, J. W. Ehnes, D. Huang, H. Ming, R. Tong, S. W. Lee, and H. Li, "Mapping frames with DNN-HMM recognizer for non-parallel voice conversion," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2015, Hong Kong, December 16-19, 2015*. IEEE, 2015, pp. 488–494. [Online]. Available: http://dx.doi.org/10.1109/APSIPA.2015.7415320

[16] M. Zhang, J. Tao, J. Tian, and X. Wang, "Text-independent voice conversion based on state mapped codebook," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*. IEEE, 2008, pp. 4605–4608. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2008.4518682

[17] P. Song, W. Zheng, and L. Zhao, "Non-parallel training for voice conversion based on adaptation method," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*. IEEE, 2013, pp. 6905–6909. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2013.6639000

[18] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. Audio, Speech & Language Processing*, vol. 18, no. 5, pp. 944–953, 2010. [Online]. Available: http://dx.doi.org/10.1109/TASL.2009.2038669

[19] Y. Agiomyrgiannakis, "The matching-minimization algorithm, the INCA algorithm and a mathematical framework for voice conversion with unaligned corpora," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 5645–5649. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2016.7472758

[20] H. Ney, D. Sündermann, A. Bonafonte, and H. Höge, "A first step towards text-independent voice conversion," in *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*. ISCA, 2004.

[21] J. Wu, Z. Wu, and L. Xie, "On the use of i-vectors and average voice model for voice conversion without parallel data," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016, Jeju, South Korea, December 13-16, 2016*. IEEE, 2016, pp. 1–6. [Online]. Available: http://dx.doi.org/10.1109/APSIPA.2016.7820901

[22] F. Xie, F. K. Soong, and H. Li, "A KL divergence and dnn-based approach to voice conversion without parallel training sentences," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, N. Morgan, Ed. ISCA, 2016, pp. 287–291. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-116

[23] *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016, Jeju, South Korea, December 13-16, 2016*. IEEE, 2016. [Online]. Available: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7803478

[24] N. Morgan, Ed., *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*. ISCA, 2016. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016