Investigation of using disentangled and interpretable representations with language conditioning for cross-lingual voice conversion

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of cross-lingual voice conversion in non-parallel speech corpora and one-shot learning setting. Most prior work require either parallel speech corpora or enough amount of training data from a target speaker. However, we convert an arbitrary sentences of an arbitrary source speaker to target speaker's given only one target speaker training utterance. To achieve this, we formulate the problem as learning disentangled speaker-specific and context-specific representations and follow the idea of [1] which uses Factorized Hierarchical Variational Autoencoder (FHVAE). After training FHVAE on multi-speaker training data, given arbitrary source and target speakers' utterance, we estimate those latent representations and then reconstruct the desired utterance of converted voice to that of target speaker. We use multi-language speech corpus to learn a universal model that works for all of the languages. We investigate the use of a one-hot language embedding to condition the model on the language of the utterance being queried and show the effectiveness of the approach. We also investigate the effect of using or not using the language conditioning. Furthermore, we visualize the embeddings of the different languages and sexes. Finally, in the subjective tests, for one language and cross-lingual voice conversion, our approach achieved moderately better or comparable results compared to the baseline in speech quality and similarity.

1 Introduction

The task of Voice Conversion (VC) [2, 3] is a technique to convert source speaker's spoken sentences into those of a target speaker's voice. It requires to preserve not only the target speaker's identity, but also phonetic context spoken by the source speaker. To tackle this problem, many approaches have been proposed [4, 5, 6]. However, most prior work require parallel spoken corpus and enough amount of data to learn the target speaker's voice. Recently, there were approaches proposed for voice conversion with non-parallel corpus [7, 8, 9]. But they still require that speaker identity was known *priori*, or included in training data for the model.

Recently, Hsu et al. [1] proposed to use disentangled and interpretable representations to overcome these limitations by exploiting Factorized Hierarchical Variation Autoencoder. They achieved reasonable quality with just single utterance from a target speaker but it was still not satisfactory. Nevertheless, most prior work focus on voice conversion *within* one language. But we believe that if we can capture disentangled representations of phonetic or linguistic contexts and speaker identities, the model should be capable for more challenging *cross*-lingual setting, which means that source and target speakers are from different languages. Therefore, we focus on investigating cross-lingual voice conversion, and propose to follow the same spirit from Hsu et al. [1] and improve the performance. Our contributions are:

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

- We build a voice model which is trained on utterances from 5 different languages to let the model observe as much speaker and phonetic variations as possible.
- We conduct cross-lingual voice conversion experiments and our approach achieved moderately better or comparable results than baselines in speech quality and similarity in the subjective tests.
- We examine the effect of using additional one-hot embedding along with speaker embedding that determines the input utterance language.

2 Related Work

Voice conversion has been an important research problem for over a decade. One popular approach to tackle the problem is spectral conversion such as Gaussian mixture models (GMMs) [4] and deep neural networks (DNN) [5]. However, it requires parallel spoken corpus and dynamic time warping (DTW) is usually used to align source and target utterances. To overcome this limitation, non-parallel voice conversion approaches were proposed, for instance, *eigenvoice* [6], *i-vecotor* [10], and Variational Autoencoder [7, 9] based models. However, eigenvoice based approach [6] still requires reference speaker to train the model, and VAE based approaches [7, 9] require speaker identities to be known priori as included in training data for the model. i-vector based approach [10] looks promising which remains to be studied further. The i-vectors are converted by replacing the source latent variable by the target latent variable. The Gaussian mixture means are then reconstructed from the converted i-vector. The Gaussians with adjusted means are then applied to the source vector to perform the acoustic feature conversion. Siamese autoencoder has also been proposed for decomposing speaker identity and linguistic embeddings [11]. However, this approach requires parallel training data to learn the decomposing architecture. This decomposition is achieved by means of applying some similarity and non-similarity costs between the Siamese architectures.

Nonetheless, cross-lingual voice conversion is also a challenging task since target language is not known in training time, and only few work has proposed, including GMMs based approach [12] and eigenvoice based approach [13], but still have inherent limitations as above.

Recently, deep generative models have been applied and successful for unsupervised learning tasks, and include Variational Autoencoder (VAE) [14], Generative Adversarial Networks (GAN) [15], and auto-regressive models [16, 17]. Among them, VAE can infer latent codes from data and generate data from them by jointly learning inference and generative networks, and VAE has been also applied for voice conversion [7, 9]. However, in their models, speaker identities are not infered from data and instead required to be known in model training time. GAN has been also exploited for non-parallel voice conversion [18] with the cycle consistency contraint [19], but it still has the limitation that it needs to know the target speaker in training time and be trained for each target.

To understand the disentangled and interpretable structure of latent codes, several work were proposed, namely, DC-IGN [20], InfoGAN [21], β-VAE [22], and FHVAE [1]. These approaches to uncover disentangled representation may help voice conversion with very limited resource from target speaker, since it might infer speaker identity information from data without supervision, as illustrated in FHVAE [1]. However, the qualities of converted voices were not good enough, therefore, we focus on the model structure of FHVAE and investigate to improve it, even with more challenging cross-lingual voice conversion setting.

3 Model

Variational autoencoder [14] (VAE) is a powerful model to uncover hidden representation and generate new data samples. Let observations be x and latent variables z. In the variational autoencoder model, the encoder (or inference network) $q_\phi(z|x)$ outputs z given input x, and decoder $p_\Phi(x|z)$ generates data x given z. The encoder and decoder are neural networks. Training is done by maximizing variational lower bound (or also called evidence lower bound):

$$\begin{aligned}\ell(\Phi, \phi) &= \mathbb{E}_q[\log p_\Phi(x, z)] - \mathbb{E}_q[\log q_\phi(z|x)] \\ &= \log p_\Phi(x) - D_{KL}(q_\phi(z|x)||p_\Phi(z|x)).\end{aligned}$$

where D_{KL} is Kullback-Leibler divergence.

However, VAE considers no structure for latent variable z. Assuming structure for z could be beneficial to exploit the inherent structures in data. Here we describe Factorized Hierarchical

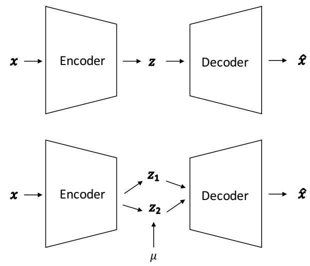

Figure 1: Structures of Variation Autoencoder (upper) and Factorized Hierarchical Variational Autoencoder (lower).

Variational Autoencoder proposed by Hsu et al [1]. Let a dataset D consist of N_{seq} i.i.d. sequences X^i. For each sequence X^i, it consists of N^i_{seg} $X^{i,j}$ observation segments. Then we define factorized latent variables of latent segment variable $Z_1^{i,j}$ and latent sequence variable $Z_2^{i,j}$. In the context of voice conversion, $Z_1^{i,j}$ is responsible for generating phonetic contexts and $Z_2^{i,j}$ is for speaker identity. When generating data $X^{i,j}$, we first sample $Z_2^{i,j}$ from isotropic Gaussian centered at μ^i shared for the entire sequence, and also $Z_1^{i,j}$ independently. Then we generate $X^{i,j}$ conditioned on $Z_1^{i,j}$ and $Z_2^{i,j}$. Thus, joint probability with a sequence X^i is:

$$p_\Phi(X^i, Z_1^i, Z_2^i, \mu^i) = p_\Phi(\mu^i) \prod_{j=1}^{N^i_{seg}} p_\Phi(X^{i,j}|Z_1^{i,j}, Z_2^{i,j}) \\ p_\Phi(Z_1^{i,j}) p_\Phi(Z_2^{i,j}|\mu^i)$$

This is illustrated in Figure 1. For inference, we use variational inference to approximate the true posterior and have:

$$q_\phi(Z_1^i, Z_2^i, \mu^i|X^i) = q_\phi(\mu^i) \prod_{j=1}^{N^i_{seg}} q_\phi(Z_1^{i,j}|X^{i,j}, Z_2^{i,j}) \\ q_\phi(Z_2^{i,j}|X^{i,j})$$

Since sequence variational lower bound can be decomposed to segment variational lower bound, we can use batches of segment instead of sequence level to maximize:

$$\ell(\Phi, \phi; X^{i,j}) = \ell(\Phi, \phi; X^{i,j}|\tilde{\mu}^i) + \frac{1}{N^i_{seg}} \log p_\Phi(\tilde{\mu}^i) + const$$

$$\begin{aligned} \ell(\Phi, \phi; X^{i,j}|\tilde{\mu}^i) = \mathbb{E}_{q_\phi(Z_1^{i,j}, Z_2^{i,j}|X^{i,j})}[\log p_\Phi(X^{i,j}|Z_1^{i,j}, Z_2^{i,j})] \\ -\mathbb{E}_{q_\phi(Z_2^{i,j}|X^{i,j})}[D_{KL}(q_\phi(Z_1^{i,j}|X^{i,j}, Z_2^{i,j})||p_\Phi(Z_1^{i,j}))] \\ -D_{KL}(q_\phi(Z_2^{i,j}|X^{i,j})||p_\Phi(Z_2^{i,j}|\tilde{\mu}^i)) \end{aligned}$$

where $\tilde{\mu}^i$ is the posterior mean of μ^i. Please refer to Hsu et al. [1] for more details. Additionally, Hsu et al. also proposed discriminative segment variational lower bound to encourage Z_2^i to be more sequence-specific by adding the additional term of inferring the sequence index i from $Z_2^{i,j}$. For our experiments, we exploit this FHVAE model and sequence-to-sequence model [23] as the structure of encoder-decoder for sequential data. We propose adding an input language embedding to the input of the model. This language embedding will be used to determine the input utterance language using a one-hot representation of the in-training languages.

For performing the voice conversion, we compute the average Z_2 from the training utterance(s) of source and target speakers. For a given input utterance, we compute Z_1 and Z_2 of the input utterance. There are two ways to perform voice conversion. First, we can replace Z_2 values of the source speaker with the average Z_2 from the target speaker. This approach resulted in too muffled generated result. Second, we compute a difference vector between source and target average $Z_2^{diff} = Z_2^{trg} - Z_2^{src}$.

This difference vector is added to Z_2 from the input utterance as $Z_2^{converted} = Z_2 + Z_2^{diff}$ and then decoded using FHVAE to achieve the speech features. In an informal listening test, we decided to the second approach since it resulted in significantly higher quality generated speech.

4 Experiments

4.1 Datasets

We used the TIMIT corpus [24] which is a multi-speaker speech corpus as the training data for FHVAE model. We used the training speakers as suggested by the corpus to train the model. For English test speakers, we select speakers from TIMIT testing part of the corpus. We also use a proprietary Chinese speech corpus (hereon referred to as CH) with 5200 speakers each uttering one sentence. We use Microsoft's Indian Language Speech Corpus for Indian language. We also use proprietary Korean and Japanese multi-speaker speech corpora. Additionaly. we consider using the combination of all languages corpus for training the model. For Korean, Japanese, and Indian corpora, we randomly exclude 10 percent of the speakers from each corpus for training purposes. For Chinese test speakers, we utilize speakers from the THCHS-30 speech corpus [25]. To observe the effect of having more utterances per speaker but less speakers we also train the model on VCTK corpus [26]. Finally, for objective testing (which requires availability of parallel data), we utilized four CMU-arctic voices (BDL, SLT, RMS, CLB)[27]. As speech features, we used 40th-order MCEPs (excluding the energy coefficient, dimensionality D=39), extracted using the World toolkit [28] with a 5ms frame shift. All audio files are transformed to 16kHz and 16 bit before any analysis.

4.2 Experimental setting

For the encoder and decoder in FHVAE model, we use Long Short Term Memory (LSTM) [29] as the first layer with 256 hidden units with a fully-connected layer on top. We use 32 dimensions for each latent variable Z_1 and Z_2. The models were trained with stochastic gradient descent. We use a mini-batch size of 256. The Adam optimizer [30] is used with β_1 = 0.95, β_2 = 0.999, ϵ = 10^{-8}, and initial learning rate of 10^{-4}. The model is trained for 500 epochs and select the model best performing on the development set.

From now on, we use the abbreviation VAE for FHVAE model. In our experiments, we consider two models: VAE-UNC (unconditioned) [31] and VAE-CND (conditioned) which mean models trained either with or without the language conditioning input. We consider four gender conversions (F: female, M: male): F2F, F2M, M2F, M2M. We also consider 25 cross-language conversions: all permutations of 5 languages considered as source and target. The voice conversion samples are available at: https://shamidreza.github.io/nips2018samples

4.3 Visualizing embeddings

In this experiment, we investigate the speaker embeddings Z_2 by visualizing them in Figure 2. For visualizing the speaker embeddings, we use the 10 test speakers (5 male, 5 female) from each language test corpus. We use VAE-UNC and VAE-CND. In Figures 2, we show the speaker embeddings computed from 1 utterance where the 2D plot of the speaker embeddings (computed using PCA) are shown. In all subplots, the female and male embedding cluster locations are clearly separated. Furthermore, the plot shows that the speaker embeddings of unique speakers fall near the same location. One phenomenon that we notice is that the speaker embeddings for different languages and gender fall to different locations for VAE-UNC, however, they fall closer to each other in VAE-CND. This might be due to the conditioning on language improving the representation ability of the model. Furthermore, we investigate the phonetic context embedding Z_1 for a sentence for four English test speakers on VAE-UNC. The phonetic context matrix over the computed utterances (compressed using PCA) is shown in Figures 3. Ideally, we want the matrices should be close to each other since the phonetic context embedding is supposed to be speaker-independent. The figure show the closeness of the embeddings at the similar time frames. There is still some minor discrepancy between the embeddings which shows room for further improvement of model architecture and/or larger speech corpus.

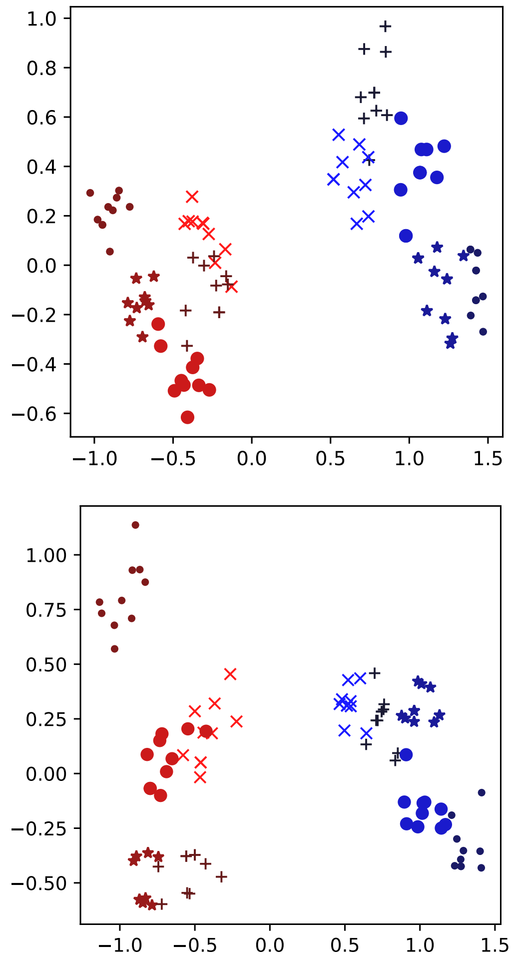

Figure 2: Visualization of speaker embeddings: unconditioned (top) versus conditioned (bottom). Red represents female speakers and blue represents male speakers. Each dot type represents a langauge

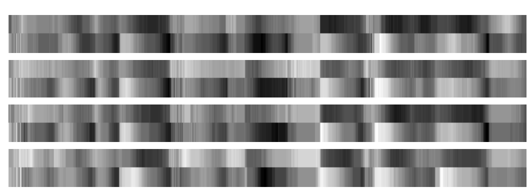

Figure 3: Visualization of phonetic context embedding sequence of a sentences aligned to each other for four English speakers. The embeddings are transformed to 2D using PCA.

4.4 Subjective evaluation

To subjectively evaluate voice conversion performance, we performed two perceptual tests. The first test measured speech quality, designed to answer the question “how natural does the converted speech sound”?, and the second test measured speaker similarity, designed to answer the question “how accurate does the converted speech mimic the target speaker"?. The listening experiments were carried out using Amazon Mechanical Turk, with participants who had approval ratings of at least 90% and were located in North America. Both perceptual tests used three trivial-to-judge trials,

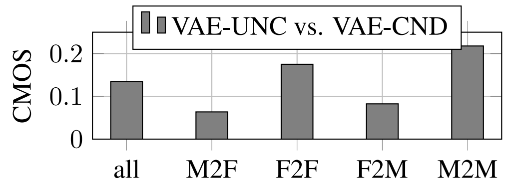

Figure 4: Speech Quality average score with gender and language break-down. Positive scores favor VAE-CND. (confidence intervals for all is close to 0.14)

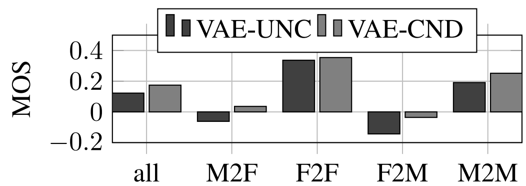

Figure 5: Speech Similarity average score with conversion break-down. Positive scores are desirable.

added to the experiment to exclude unreliable listeners from statistical analysis. No listeners were flagged as unreliable in our experiments.

4.4.1 Speech quality

To evaluate the speech quality of the converted utterances, we conducted a Comparative Mean Opinion Score (CMOS) test. In this test, listeners heard two stimuli A and B with the same content, generated using the same source speaker, but in two different processing conditions, and were then asked to indicate whether they thought B was better or worse than A, using a five-point scale comprised of +2 (much better), +1 (somewhat better), 0 (same), -1 (somewhat worse), -2 (much worse). We randomized the order of stimulus presentation, both the order of A and B, as well as the order of the comparison pairs. We utilized two processing conditions: VAE-UNC, VAE-CND. We assessed the VC approach effect by directly comparing VAE-UNC vs. VAE-CND utterances. The experiment was administered to 50 listeners with each listener judging 100 sentence pairs. We achieved $+0.15\pm0.14$ mean score towards VAE-CND. Although this is a positive difference, we did not find statistically significant difference between the quality of VAE-UNC and VAE-CND. The language-breakdown of the results are shown in Figure 4.

4.4.2 Speaker similarity

To evaluate the speaker similarity of the converted utterances, we conducted a same-different speaker similarity test [32]. In this test, listeners heard two stimuli A and B with different content, and were then asked to indicate whether they thought that A and B were spoken by the same, or by two different speakers, using a five-point scale comprised of +2 (definitely same), +1 (probably same), 0 (unsure), -1 (probably different), and -2 (definitely different). One of the stimuli in each pair was created by one of the two conversion methods, and the other stimulus was a purely MCEP-vocoded condition, used as the reference speaker. The listeners were explicitly instructed to disregard the language of the stimuli and merely judge based on the fact whether they think the utterances are from the same speaker regardless of the language. Half of all pairs were created with the reference speaker identical to the target speaker of the conversion (expecting listeners to reply “same", ideally); the other half were created with the reference speaker being the same gender, but not identical to the target speaker of the conversion (expecting listeners to reply different). We only report “same" scores. The experiment was administered to 50 listeners, with each listener judging 100 sentence pairs. The results are shown in Figure 5. We did find a consistent improvement of VAE-CND performance over VAE-UNC, however these differences where not statistically significant.

5 Conclusions

We proposed to exploit FHVAE model for challenging non-parallel and cross-lingual voice conversion, even with very small number of training utterances such as only one target speaker's utterance. We use multi-language corpus to learn disentangled representations from speech. We also introduce a one-hot language embedding to the model in order to improve the model performance. We perform several visualizations to show the effect of the model. In the subjective tests, we found some improvement in the quality and similarity performance of the system, although the differences where not statistically significant.

References

[1] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in neural information processing systems*, pp. 1876–1887, 2017.

[2] Y. Stylianou, "Voice transformation: a survey," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 3585–3588, IEEE, 2009.

[3] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[4] 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*, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[6] Z. Wu, T. Kinnunen, E. S. Chng, and H. Li, "Mixture of factor analyzers using priors from non-parallel speech for voice conversion," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 914–917, 2012.

[7] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. 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*, pp. 1–6, IEEE, 2016.

[8] P. Song, W. Zheng, and L. Zhao, "Non-parallel training for voice conversion based on adaptation method," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 6905–6909, IEEE, 2013.

[9] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[10] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using i-vector plda: Towards unifying speaker verification and transformation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 5535–5539, IEEE, 2017.

[11] S. H. Mohammadi and A. Kain, "Siamese autoencoders for speech style extraction and switching applied to voice identification and conversion," *Proceedings of Interspeech*, pp. 1293–1297, 2017.

[12] B. Ramani, M. A. Jeeva, P. Vijayalakshmi, and T. Nagarajan, "A multi-level gmm-based cross-lingual voice conversion using language-specific mixture weights for polyglot synthesis," *Circuits, Systems, and Signal Processing*, vol. 35, no. 4, pp. 1283–1311, 2016.

[13] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit, "Cross-language voice conversion based on eigenvoices," *Proceedings of Interspeech*, 2009.

[14] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[15] 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*, pp. 2672–2680, 2014.

[16] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel recurrent neural networks,” *arXiv preprint arXiv:1601.06759*, 2016.

[17] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[18] T. Kaneko and H. Kameoka, “Parallel-data-free voice conversion using cycle-consistent adversarial networks,” *arXiv preprint arXiv:1711.11293*, 2017.

[19] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” *arXiv preprint arXiv:1703.10593*, 2017.

[20] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, “Deep convolutional inverse graphics network,” in *Advances in Neural Information Processing Systems*, pp. 2539–2547, 2015.

[21] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “Infogan: Interpretable representation learning by information maximizing generative adversarial nets,” in *Advances in Neural Information Processing Systems*, pp. 2172–2180, 2016.

[22] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “beta-vae: Learning basic visual concepts with a constrained variational framework,” 2016.

[23] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1,” *NASA STI/Recon technical report n*, vol. 93, 1993.

[25] D. Wang and X. Zhang, “Thchs-30: A free chinese speech corpus,” *arXiv preprint arXiv:1512.01882*, 2015.

[26] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2017.

[27] J. Kominek and A. W. Black, “The cmu arctic speech databases,” in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[28] M. Morise, F. Yokomori, and K. Ozawa, “World: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[29] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[31] S. H. Mohammadi and T. Kim, “Investigation of using disentangled and interpretable representations for one-shot cross-lingual voice conversion,” *Proc. Interspeech*, 2018.

[32] A. B. Kain, “High resolution voice transformation,” 2001.