MULTI-SPEAKER AND MULTI-DOMAIN EMOTIONAL VOICE CONVERSION USING FACTORIZED HIERARCHICAL VARIATIONAL AUTOENCODER

Mohamed Elgaar[†‡] *Jungbae Park*[†‡] *Sang Wan Lee*[†‡§]

[†]KAIST, [‡]Humelo Inc, [§]KAIST Institute for Artificial Intelligence, South Korea

{mohamed, sangwan}@kaist.ac.kr, jb@humelo.com

ABSTRACT

Due to the complexity of emotional features, there has been limited success in emotional voice conversion. One major challenge is that conversion between more than two kinds of emotions often accompanies distortion of voice signal.

The factorized hierarchical variational autoencoder (FHVAE) [1] was previously shown to have an ability, called sequence-level regularization, to generate disentangled representations of both sequence-level (such as speaker identity) and segment-level features. This study exploits the FHVAE pipeline to produce disentangled representations of emotion, making it possible to greatly facilitate emotional voice conversion.

We propose three versions of algorithms for improving the quality of disentangled representation and audio synthesis. We conducted three mean opinion score (MOS) surveys to assess the performance of our models in terms of 1) speaker's voice preservation, 2) emotion conversion, and 3) audio naturalness.

Index Terms— Emotional Voice Conversion, Variational Autoencoder, Disentangled Representation, Style Transfer

1. INTRODUCTION

Voice conversion (VC) refers to the process of modifying particular prosodic features of speech signals while preserving phonetic contents and other prosodic features. VC has been typically applied to speaker change so that the utterance of the converted voice sounds similar to that of a target speaker [2]. The VC can be applicable to various problems, including speech synthesis, low bit-rate speech coding [2], singing voice generation, video game and animation character voices generation, etc.

Emotional VC (E-VC) is another variant of the VC task. E-VC refers to the process of converting the emotion embedded in speech to a target emotion. The ability of E-VC to control synthetic speech emotion enables a human-like interaction with computer systems that use a speech interface. Virtual assistants have become a dominant application of interest due to the advancement in dialog systems, natural language processing, and speech processing. When equipped with emotion control, one would feel more comfortable with interacting with virtual assistant services, making them extremely valuable both commercially and socially.

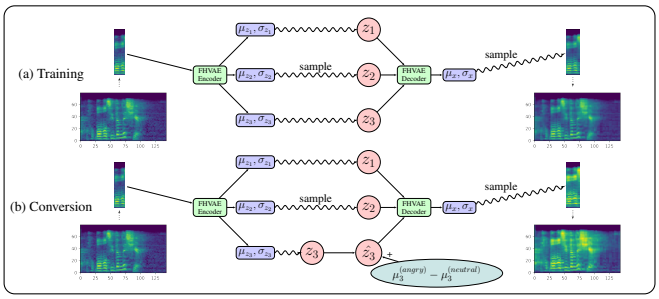


Fig. 1: Input sequence is segmented and each audio segment is used to infer the 3 latent variable distributions using variational inference [4]. z_1, z_2, z_3 refer to segment-level, sequence-level, and emotion latent variables, respectively. During conversion, z_3 is modified using vector translation in the direction of the target domain with controllable weight.

There exist a few major challenges in E-VC. The first issue is the limited amount of training data. The collection of speech with emotion annotation requires talents or professionals to act out the required emotions. That being said, people may disagree on the presented emotion, posing another challenge; the subjectivity of speech makes it difficult to evaluate the emotion control systems. [3]

Traditionally, Gaussian mixture model (GMM) based models were applied for VC [2, 5, 6]. This requires pairs of source and target speech aligned at the phoneme level, which is expensive to collect, especially for the data-hungry deep learning methods. This motivates people to use unsupervised learning methods [1, 7, 8, 9, 10, 11].

To perform emotion conversion, we employ techniques of style-transfer based on interpretable, disentangled representations [12, 13, 14]. The factorized hierarchical variational autoencoder (FHVAE) model learns a factorized representation of sequential data by making use of its multi-timescale characteristics [1]. It is able to decompose the long time scale features of speech, such as speaker identity and volume, and short timescale features, such as phonetic content. FHVAE has been evaluated for voice conversion, channel conversion (sounds recorded with different devices) and has achieved improved results [15, 16, 1].

The contribution of this study is threefold. First, the proposed method is based on the extension of FHVAE architecture, which introduces an additional layer of attributes using an emotion-dependent prior. For this, we explored a variety of architectures modifying encoded emotion attributes in different ways. The basic

This work was partly supported by Seoul R&BD Program(CY190019) and by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government (MSIT) (No.2019-0-01371, Development of brain-inspired AI with human-like intelligence)

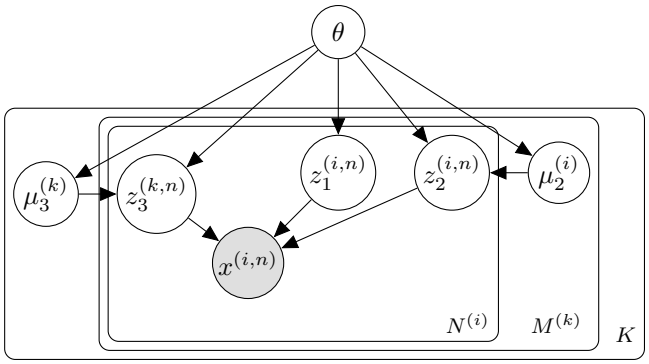


Fig. 2: Generative model of FHVAE with the addition of emotion-dependent prior μ_3.

version is shown in Fig. 1. Second, for distinct a presentation of target emotion, we introduce an additional criterion to maximize the margin between the emotion embeddings z_3 [17]. Third, to further facilitate emotion conversion, we use a loss function based on the cycle-consistency loss [18].

2. RELATED WORK

StarGAN-VC [19] uses a conditional GAN with a cycle-consistency loss to convert between many domains using a single generator. StarGAN-VC2 [11] achieves an improvement in MOS score compared to the first version, but the limitation is that the source domain should be known.

Another work [20] uses an interpretable encoder-decoder architecture that learns to generate a disentangled representation of content and style, using a GAN discriminator loss and cycle-consistency loss. They performed a subjective evaluation for audio naturalness, speaker similarity, and emotion conversion ability, comparing their model with a rule-based F0 conversion model [21] and StarGAN-VC [19]. The limitation of this model is the need for a separate encoder and decoder for every domain.

AutoVC [22] uses a simple loss function, utilizing the information bottleneck theory for zero-shot voice conversion. They presented a theorem stating that with sufficient data and bottleneck dimension setting, their model can approximate the true distribution of the target speaker's speech, and perform ideal voice conversion.

VQ-VAE [10] relies on vector quantization to encode inputs into discrete latent variables. The results indicate that the discrete encoding is a high-level description of speech and closely related to phonemes. When the discrete encoding is coupled with a one-hot speaker embedding, it is able to perform voice conversion.

3. PROPOSED METHODS

3.1. Architecture

Given a dataset of speech sequences with emotion annotation $\{(\boldsymbol{X}^{(i)}, e^{(i)})\}_{i=1}^{M}$, such that each sequence is composed of $N^{(i)}$ segments $\boldsymbol{X}^{(i)} = \{\boldsymbol{x}^{(i,n)}\}_{n=1}^{N^{(i)}}$ and each $e^{(i)}$ is one of $\{e^{(k)}\}_{k=1}^{K}$ where K is the number of emotion labels.

We extended the generative model of FHVAE (Fig. 2). The corresponding joint distribution, with parameters θ, and approximate posterior (inference model), with parameters ϕ are provided in Eq. 1 and 2, respectively. All of the distributions in Eq. 1, 2 are distributed following a multivariate isotropic Gaussian.

In Eq. 1, the conditional distribution of x is parameterized by a trainable neural network. The conditional distributions of z_2 and z_3 are centered at μ_2 and μ_3, respectively, with constant variances. The rest of the distributions in Eq. 1 are centered at 0 with constant variances.

In Eq. 2, parameters of z_1, z_2 and z_3 are inferred from X using an LSTM neural network followed by a fully connected layer. For each of μ_2 and μ_3, the parameters are a trainable lookup table for means, and a constant variances.

During training, for every sequence we are given $\tilde{\mu}_2$, $\tilde{\mu}_3$. $\tilde{\mu}_2$ is unique for every sequence, and $\tilde{\mu}_3$ is common for sequences of the same emotion. The ELBO (Eq. 3) can be interpreted as follows: 1) the first term is the reconstruction loss. 2) The second term causes z_1s to be normally distributed similarly to $\mathcal{N}(0, \sigma_{z_1}^2)$. 3) The third and fourth terms of the equation cause the variables z_2 and z_3 to be normally distributed similarly to $\mathcal{N}(\tilde{\mu}_2, \sigma_{z_2}^2)$ and $\mathcal{N}(\tilde{\mu}_3, \sigma_{z_3}^2)$, respectively. 4) The fifth and sixth terms cause $\tilde{\mu}_2$, $\tilde{\mu}_3$ to be normally distributed similarly to the priors $\mathcal{N}(0, \sigma_{\mu_2}^2)$ and $\mathcal{N}(0, \sigma_{\mu_3}^2)$. Additionally, we use a discriminative loss of emotion, similarly to the discriminative loss of sequences in section 2.1 of the original model [1].

The base network architecture is identical to FHVAE with the addition of z_3 and the K-dimensional $\tilde{\mu}_3$ lookup table. During inference, z_1 is conditioned on z_2 and z_3.

3.2. Conversion Procedure

There are two approaches known to be useful for conversion: using the learned priors μ_3 and calculating the average encoded μ_3 for every emotion using the training data or a held-out validation set. From our experience, calculating the means from the data encoding generally results in a higher quality of conversion. In the experiments section, we use the learned priors μ_3 to demonstrate their quality. As illustrated in Fig. 1,

$$\hat{z_3} = z_3 + w * (\mu_3^{(target)} - \mu_3^{(source)}). \tag{4}$$

3.3. Orthogonalization for disentangled embedding

One major challenge in emotion conversion is highly correlated emotion features. For example, angry and happy emotions usually accompany an increase in pitch and volume, albeit in different patterns. Although our model generates a disentangled representation of emotion, it can further benefit from an orthogonalization procedure. We use the Gram–Schmidt orthogonalization to transform the set $\{\mu_3^{(i)}\}_{i=1}^{K}$ into the orthogonal set $\{v_3^{(i)}\}_{i=1}^{K}$. Then, the vectors $v^{(i)}$ are used for latent variable translation in Eq. 4. This turns out to be a crucial step as it results in a great improvement of conversion quality and generated audio naturalness.

3.4. Max-margin Training

We expect that if emotion embeddings are farther away from each other, then we can more robustly perform emotion conversion, and the results can be more distinctly belonging to the target class. This view is incorporated into our criterion as shown in Eq. 5, which is maximized by using gradient ascent. Note that we use $\ell 1$-norm. Because of the squared terms in the $\ell 2$-norm, more weight is given to the more distinct emotions. Therefore, it fails to induce a large margin between correlated emotions, such as laziness and sadness.

$$p_\theta(\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3) = p_\theta(\boldsymbol{\mu}_2)p_\theta(\boldsymbol{\mu}_3)\prod_{n=1}^{N} p_\theta(x^{(n)}|z_1^{(n)}, z_2^{(n)}, z_3^{(n)})p_\theta(z_1^{(n)})p_\theta(z_2^{(n)}|\boldsymbol{\mu}_2)p_\theta(z_3^{(n)}|\boldsymbol{\mu}_3). \tag{1}$$

$$q_\phi(\boldsymbol{Z}_1^{(i)}, \boldsymbol{Z}_2^{(i)}, \boldsymbol{Z}_3^{(i)}, \boldsymbol{\mu}_2^{(i)}, \boldsymbol{\mu}_3^{(i)}|\boldsymbol{X}^{(i)}) = q_\phi(\boldsymbol{\mu}_2^{(i)})q_\phi(\boldsymbol{\mu}_3^{(i)})\prod_{n=1}^{N(i)} q_\phi(z_1{}^{(i,n)}|x^{(i,n)}, z_2^{(i,n)}, z_3^{(i,n)})q_\phi(z_2^{(i,n)}|x^{(i,n)})q_\phi(z_3^{(i,n)}|x^{(i,n)}). \tag{2}$$

$$\begin{aligned}\mathcal{L}(\theta, \phi; x^{(n)}|\tilde{\mu}_2, \tilde{\mu}_3) =& \mathbb{E}_{q_\phi(z_1^{(n)}, z_2^{(n)}, z_3^{(n)}|x^{(n)})}[\log p_\theta(x^{(n)}|z_1^{(n)}, z_2^{(n)}, z_3^{(n)})] \\ &- \mathbb{E}_{q_\phi(z_2^{(n)}, z_3^{(n)}|x^{(n)})}[D_{KL}(q_\phi(z_1{}^{(n)}|x^{(n)}, z_2^{(n)}, z_3^{(n)})\|p_\theta(z_1^{(n)})] \\ &- D_{KL}(q_\phi(z_2^{(n)}|x^{(n)})\|p_\phi(z_2^{(n)}|\tilde{\mu}_2)) - D_{KL}(q_\phi(z_3^{(n)}|x^{(n)})\|p_\phi(z_3^{(n)}|\tilde{\mu}_3)) \\ &+ \frac{1}{N}\log p_\theta(\tilde{\mu}_2) + \log p_\theta(\tilde{\mu}_3).\end{aligned} \tag{3}$$

However, the use of $\ell 1$-norm gives equal weight to all pairs or emotions and produces a better separation between correlated emotions.

$$\ell_{margin} = \sum_{i=1}^{K}\sum_{j=i+1}^{K} \|\mu_3^{(i)} - \mu_3^{(j)}\|_1 \tag{5}$$

3.5. Cycle-consistency Loss

As demonstrated in section 2, cycle-consistency loss is an im- portant technique for style transfer; it helps the encoder preserve some key characteristics, such as speaker identity or phonetic content. We use a simple cycle-loss for the $\ell 2$-norm for z_2 and z_2, as presented in Algorithm 1. The full loss is shown in Eq. 6:

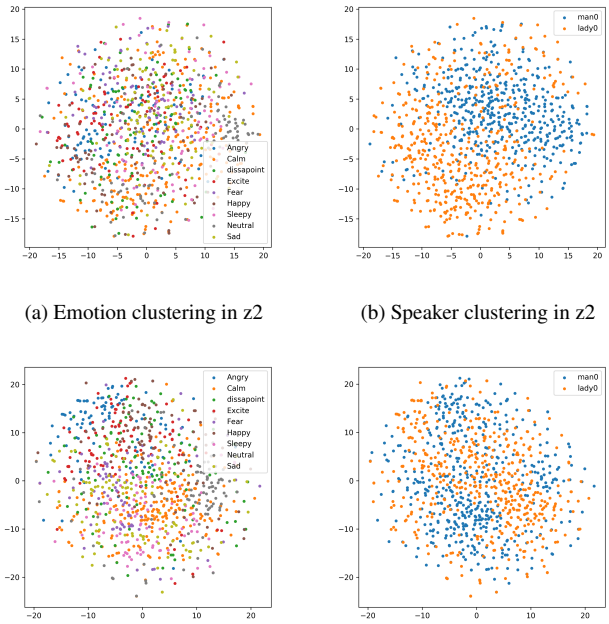
$$\begin{aligned}\mathcal{L}(\theta, \phi; x^{(n)}) = \mathcal{L}(\theta, \phi; x^{(n)}|\tilde{\mu}_2, \tilde{\mu}_3) + \alpha_1 * \ell_{disc(seq)} \\ + \alpha_2 * \ell_{disc(emo)} - \alpha_3 * \ell_{margin} \\ + \alpha_4 * \ell_{cycle} + \alpha_5 * \ell_{disc(cycle)}.\end{aligned} \tag{6}$$

Algorithm 1

procedure CYCLE(x, hyper-parameters: $\{w^{(i)}\}_{i=1}^{K}$)
  $z_1, z_2, z_3 \leftarrow encode(x)$
  $target \leftarrow random(1, K)$
  $\hat{z_3} \leftarrow z_3 + w^{(source)} * (\mu_3^{(target)} - \mu_3^{(source)})$
  $\tilde{x} \leftarrow decode(z_1, z_2, \hat{z_3})$
  $z_1^{cycle}, z_2^{cycle}, z_3^{cycle} \leftarrow encode(\tilde{x})$
  $\ell_{cycle} \leftarrow \|z_1^{cycle} - z_1\|_2^2 + \|z_2^{cycle} - z_2\|_2^2$
  $\ell_{disc(cycle)} \leftarrow \log p(k|z_3^{cycle})$

4. EXPERIMENTS

We use the same values of hyper-parameters as in the original FHVAE [1]. Additionally, we set $\sigma_{z_3} = 0.5$ and $\sigma_{\mu_3} = 0.1$. We observe that σ_{μ_3} significantly affects quality and quality is better for smaller variance. We use the following weights in Eq. 6: $\alpha_1 = 10, \alpha_2 = 10, \alpha_3 = 10, \alpha_4 = 1, \alpha_5 = 0.1$. Hyper-parameters are set according to cross-validation. Furthermore, we use the same scheme of pre and post processing as FHVAE, with 80-dimensional mel-spectrogram features.



(a) Emotion clustering in z2 (b) Speaker clustering in z2

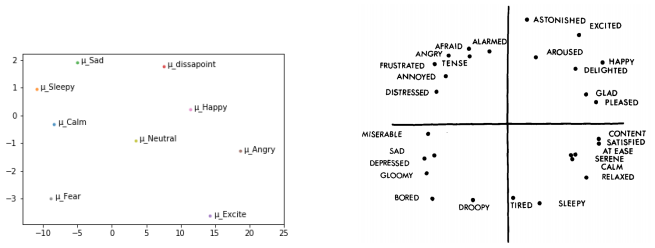
(c) Emotion clustering in z3 (d) Speaker clustering in z3

Fig. 3: Illustration of disentanglement in z2 and z3.

4.1. Visualization

We collected a dataset consisting of 25 hours of one-sentence sequences, spoken by 2 speakers and consisting of 9 emotion classes. It is recorded in a studio and the actors are asked to act out the emotion. The dataset was used for speech emotion recognition training. A model is trained without margin-loss or cycle-loss for the creation of Fig. 3 and 4. Another model was trained with margin-loss for the creation of Fig. 5. Data were divided into training and validation sets (7:3). Visualizations are done using the validation set.

A disentangled representation is known to be useful for controlling different factors of the data independently of other factors [24]. In Fig. 3 we observe that z2 exhibits clear separation for different speakers, whereas the emotions are not clearly separated. On the other hand, z3 shows a gradual change from the top-left with angry and happy (high arousal) to the bottom-right with sleepy, sad and



(a) Emotion embedding. (b) The circumplex model.

Fig. 4: The emotion embedding of our model closely resembles the valence-arousal model of emotions [23]. In (a) arousal tends to increase from top to bottom, and valence tends to increase from left to right.

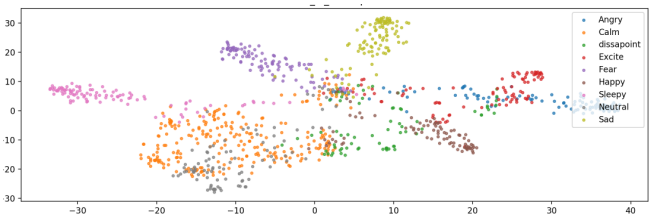


Fig. 5: Emotion embedding as a results of margin-loss.

calm (low arousal).

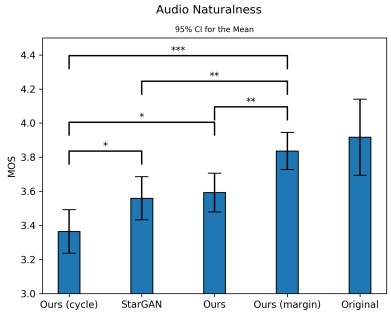
The arousal-valence dimensions of emotions appear to some extent in all of our simulations. Intriguingly, the emotion embeddings of our model bear close resemblance with the valence-arousal model [23] (Fig. 4), even though this theory is not explicitly incorporated into our training criterion.

Finally, Fig. 5, which was generated shows the effect of using margin loss on the emotion embedding. Note that the model is able to distinguish between even the highly similar emotions. Fig. 3, 4(a) and 5 are generated using the t-SNE dimensionality reduction algorithm.

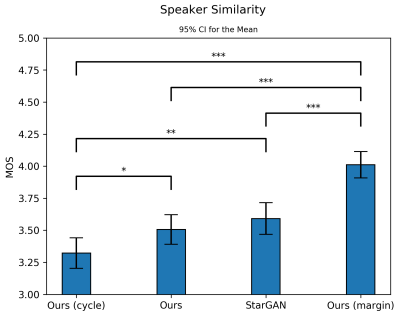
4.2. Subjective evaluation

Since there is no ground truth for unsupervised style-transfer tasks, we carried out a mean opinion score (MOS) survey to evaluate our model. We use StarGAN-VC as a baseline [19]. The dataset used is IEMOCAP [25], about 12 hours of acted emotional speech, with 5 speakers and 4 emotions (angry, happy, sad, neutral). We follow the survey setting in [20], evaluating the audio naturalness, speaker similarity before and after conversion, and the percentage of conversions perceived to be of the target emotion, by asking the worker to select one of two choices, the source and target emotion. Moreover, our model uses the Griffin-lim vocoder to synthesize speech, while StarGAN-VC uses the World vocoder.

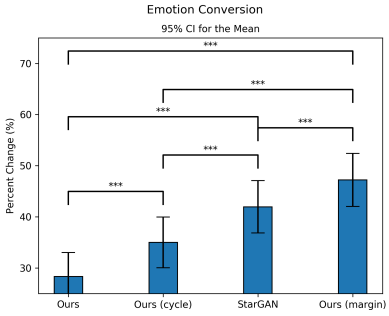
The survey was performed using Amazon Mechanical Turk. The dataset was divided into training, validation and testing sets (70:25:5). We randomly selected 24 audio samples from the test set for the survey. We evaluate the baseline model and 3 variations of our model, vanilla, only with margin-loss and only with cycle-loss. The 24 audio samples are converted into the 3 emotions other than the source, resulting in 72 samples from every model. In the audio naturalness test, the original samples are still only 24, while the model outputs are 72 samples each. We gather responses from 5



(a) Audio Naturalness



(b) Speaker Similarity



(c) Emotion Conversion (%)

Fig. 6: Subjective evaluation results. * = p-value < 0.05, ** = p-value < 0.005, *** = p-value < 0.0005

unique workers for every sample. The unique number of participants for each test: a) 113, b) 76, c) 80. The results are shown in Fig. 6. Our model superior performance in maintaining the speaker identity, following from the properties of FHVAE[1].

5. CONCLUSION

Our emotion conversion framework is based on an extension to FHVAE [1] to generate a disentangled representation of emotion and perform emotional voice conversion. We propose the addition of novel loss terms and the orthogonalization of learned embeddings to remedy the shortcomings of vanilla FHVAE in emotional voice conversion. We show that our models achieve quality comparative with

[1]Audio samples are available at `https://mohdelgaar.github.io/humelo-emoconv-icassp2020`.

advanced GAN-based methods while showing significant improvement in maintaining the speaker's voice.

6. ACKNOWLEDGEMENT

The authors would like to thank Onur Babacan for the many valuable discussions, and the Humelo data team (Seokwon Jung, Hyeonmook Park, Kiwoong Yeom) for collecting the dataset used in section 4.1.

7. REFERENCES

[1] Wei-Ning Hsu, Yu Zhang, and James Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in neural information processing systems*, 2017, pp. 1878–1889.

[2] Y. Stylianou and O. Cappe, "A system for voice conversion based on probabilistic classification and a harmonic plus noise model," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, May 1998, vol. 1, pp. 281–284 vol.1.

[3] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," 2019.

[4] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2014.

[5] Hiromichi Kawanami, Yohei Iwami, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Gmm-based voice conversion applied to emotional speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[6] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Gmm-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 134–138, 2012.

[7] Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, Patrick Nguyen, and Ruoming Pang, "Hierarchical generative modeling for controllable speech synthesis," *CoRR*, vol. abs/1810.07217, 2018.

[8] 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," *CoRR*, vol. abs/1704.00849, 2017.

[9] Li-Wei Chen, Hung-Yi Lee, and Yu Tsao, "Generative adversarial networks for unpaired voice transformation on impaired speech," *arXiv preprint arXiv:1810.12656*, 2018.

[10] Aaron van den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 6306–6315.

[11] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion," *CoRR*, vol. abs/1907.12279, 2019.

[12] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, 2016, pp. 2172–2180.

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework.," *ICLR*, vol. 2, no. 5, pp. 6, 2017.

[14] Guan, Shaobo in Insight Fellows Program, "Controlled image synthesis and editing using a novel tl-gan model," 2018, [Online; accessed 16-Oct-2019].

[15] Seyed Hamidreza Mohammadi and Taehwan Kim, "Investigation of using disentangled and interpretable representations for one-shot cross-lingual voice conversion," *CoRR*, vol. abs/1808.05294, 2018.

[16] Seongkyu Mun and Suwon Shon, "Domain mismatch robust acoustic scene classification using channel information conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 845–849.

[17] Chongxuan Li, Jun Zhu, Tianlin Shi, and Bo Zhang, "Max-margin deep generative models," in *Advances in neural information processing systems*, 2015, pp. 1837–1845.

[18] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[19] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[20] Jian Gao, Deep Chakraborty, Hamidou Tembine, and Olaitan Olaleye, "Nonparallel emotional speech conversion," *CoRR*, vol. abs/1811.01174, 2018.

[21] Jianhua Tao, Yongguo Kang, and Aijun Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1145–1154, 2006.

[22] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*, 2019, pp. 5210–5219.

[23] James A Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[24] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[25] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.