Transforming the Emotion in Speech using a Generative Adversarial Network

Kenji Yasuda, Ryohei Orihara, Yuichi Sei, Yasuyuki Tahara and Akihiko Ohsuga
Graduate School of Information and Engineering, University of Electro-Communications, 1–5–1 Chofugaoka, Chofu-shi, Tokyo, 182–8585, Japan

Keywords: Deep Learning, Domain Transfer, Generative Adversarial Network, Unsupervised Learning, Voice Conversion.

Abstract: In recent years, natural and highly accurate outputs in domain transfer tasks have been achieved by deep learning techniques. Especially, the advent of Generative Adversarial Networks (GANs) has enabled the transfer of objects between unspecified domains. Voice conversion is a popular example of speech domain transfer, which can be paraphrased as domain transfer of speakers. However, most of the voice conversion studies have focused only on transforming the identities of speakers. Understanding other nuances in the voice is necessary for natural speech synthesis. To resolve this issue, we transform the emotions in speech by the most promising GAN model, CycleGAN. In particular, we investigate the usefulness of speech with low emotional intensity as training data. Such speeches are found to be useful when the training data contained multiple speakers.

1 INTRODUCTION

Alongside the development of deep learning in recent years, domain transfer tasks have been actively researched. Domain transfer converts an attribute of data such as a style and a form of image and speech to another without changing any other attribute. For example, in human images, domain transfer can convert the gender domain while keeping the human individual. This paper considers the domain transfer of speech, which is important for natural speech synthesis.

Domain transfer of speech is popularly performed by voice conversion (VC). The VC technique converts the speaker information while preserving the linguistic information which can be paraphrased as the speaker domain transfer. Representative VC methods are based on Gaussian mixture models (GMMs) (Toda et al., 2007) and restricted Boltzmann machines (Nakashika and Minami, 2017). Some recent VC methods use deep learning approaches such as autoencoders (Hinton and Salakhutdinov, 2006)(Sekii et al., 2017) and generative adversarial networks (GANs) (Goodfellow et al., 2014)(Miyoshi et al., 2017). Among the most attractive methods is CycleGAN (Zhu et al., 2017)(Kaneko and Kameoka, 2017)(Fang et al., 2018). CycleGAN-based methods output high-quality speech despite the unsupervised learning with non-parallel data. Here parallel data refer to a collection of the same set of utterances spoken by multiple speakers. Parallel data-based methods are expected to yield high-quality speech. However, these methods are compromised by the problematic data collection. On the other hand, non-parallel data-based methods are difficult to output high-quality speech. However, their data are easily collected. Nevertheless, CycleGAN-based methods output higher quality speech than conventional methods.

Although VC has been widely studied, few of the existing studies convert information other than the identities of speakers. Speech also includes linguistic, emotional, and non-parametric information. Understanding the emotional information is important for natural speech synthesis. Emotional speeches constructed by natural speech synthesis would benefit call centers (Sakurai and Kimura, 2013) and speech guidance (Iida et al., 1999). To this end, the present paper performs domain transfer of the emotions in speech.

Liu et al. (Liu et al., 2014), Aihara et al. (Aihara et al., 2012) and Yasuda et al. (Yasuda et al., 2018b)(Yasuda et al., 2018a)(Yasuda et al., 2018c) have investigated the issue. In (Liu et al., 2014), they focused on prosodic information and converted the emotion by replacing the fundamental frequency (F0)

Yasuda, K., Orihara, R., Sei, Y., Tahara, Y. and Ohsuga, A.
Transforming the Emotion in Speech using a Generative Adversarial Network.
DOI: 10.5220/0007258504270434
In *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART 2019)*, pages 427-434
ISBN: 978-989-758-350-6

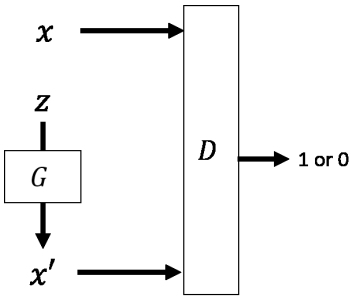

Figure 1: Schematic of a GAN.

in input with altered one suitable to express emotions for each word. In (Aihara et al., 2012), they converted F0 using a GMM. These studies used single-speaker training data. Creating converters from the training data of multiple speakers would drastically reduce the cost of collecting the training data. In (Yasuda et al., 2018b)(Yasuda et al., 2018a)(Yasuda et al., 2018c), they performed domain transfer of the emotion in speech by a CycleGAN with training data including multiple speakers, and investigated the transformation result by varying acoustic features and hyperparameters. However, they did not report on the difference between results learned from single-speaker training data and multiple-speaker training data. To improve the accuracy of domain transfer, high-quality data are generally required. However, high-quality speech data with high emotional intensity are costly to collect. If speech data with low emotional intensity could improve the accuracy of domain transfer for emotions, the speech data collection costs could be reduced. For high-quality domain transfer of speech emotions, one must investigate the training data. Therefore, we investigate whether speech with low emotional intensity provides useful training data with multiple speakers and a single speaker.

The rest of this paper is organized as follows. Section 2 describes the model used in this research and the system that transfers the emotion in the speech. In Section 3, we evaluate the system and discuss the training data. Conclusions and idea for future works are presented in Section 4.

2 METHOD FOR DOMAIN TRANSFER ON SPEECH

2.1 Generative Adversarial Network (GAN)

The GAN (see Figure 1) is a generative model of unsupervised learning proposed by Goodfellow et al.

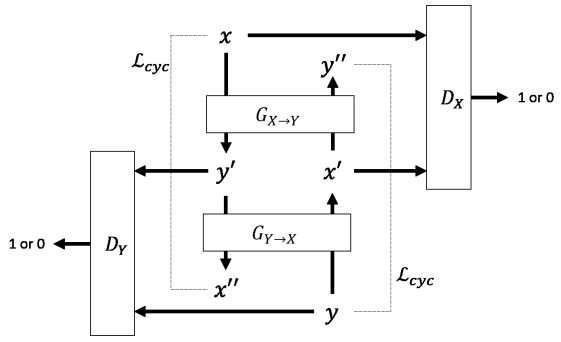

Figure 2: CycleGAN.

(Goodfellow et al., 2014)(Orihara et al., 2018). Given a dataset, the GAN generates data that are indistinguishable from the original data. The GAN is composed of two networks. The two main networks in the GAN are configured as adversarial networks that are optimized by a mini-max game. One component of the GAN is the generator G, which learns the distribution $p_{data}(x)$ of the training data x. The generator randomly samples a vector z from the prior distribution $p_z(z)$ and maps it to the output data $x' = G(z)$. The other component of the GAN is a discriminator D, which tries to discriminate between real input, that is a member of the training data, and a fake one, that is the output of the generator. The goal of D is $D(x) = 1$ for real x and $D(x) = 0$ for fake x. Whereas the optimized generator creates realistic data that fool the discriminator, the optimized discriminator precisely distinguishes the real and fake inputs:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

Many recent GANs use deep convolutional generative adversarial networks (Radford et al., 2015) which adopt convolutional neural network (CNNs) (Lecun et al., 1998) for GANs.

2.2 CycleGAN

CycleGAN is a generative model proposed by Zhu et al. (Zhu et al., 2017). The architecture of CycleGAN is shown in Figure 2. A main advantage of CycleGAN is its unsupervised learning ability from non-parallel training data. To explain the model, let us assume two domains X and Y. The generator G consists of two networks. In the first step, $G_{X\to Y}$ obtains x in X as input and generates $y' = G_{X\to Y}(x)$. In the second step, $G_{Y\to X}$ obtains y' and generates $x'' = G_{Y\to X}(G_{X\to Y}(x))$. The third step compares x'' and x. As CycleGAN assumes a one-to-one correspondence between the domains, x and x'' must be

equal. Meanwhile, the discriminator D_Y discriminates y' against Y. This process is also done vise versa. As a result, two generators are created, enabling bidirectional domain transfer.

The loss function in CycleGAN combines two objective functions, Adversarial Losses and the Cycle-Consistency Loss (Zhou et al., 2016). The total loss function is given by

$$\mathcal{L}_{full} = \mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \mathcal{L}_{adv}(G_{Y\to X}, D_X) + \lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) \quad (2)$$

where the adversarial losses are represented as follows:

$$\mathcal{L}_{adv}(G_{X\to Y}, D_Y) = E_{y\sim p_{data}(y)}[\log D_Y(y)] + E_{x\sim p_{data}(x)}[\log(1 - D_Y(G_{X\to Y}(x)))] \quad (3)$$

$$\mathcal{L}_{adv}(G_{Y\to X}, D_X) = E_{x\sim p_{data}(x)}[\log D_X(x)] + E_{y\sim p_{data}(y)}[\log(1 - D_X(G_{Y\to X}(y)))] \quad (4)$$

and the cycle-consistency loss is determined as

$$\mathcal{L}_{cyc} = E_{x\sim p_{data}}[\|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1] + E_{y\sim p_{data}(y)}[\|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1] \quad (5)$$

2.3 CycleGAN-VC

CycleGAN-VC was proposed by Kaneko and Kameoka for VC tasks. The original CycleGAN was designed for image translation applications. CycleGAN-VC is a modified version of CycleGAN that performs VC tasks using a gated CNN (Dauphin et al., 2017) and an identity mapping loss. The present research applies these ideas to the CycleGAN. Therefore, the total loss function of CycleGAN in this research given by

$$\mathcal{L}_{full} = \mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \mathcal{L}_{adv}(G_{Y\to X}, D_X) + \lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) + \lambda_{id}\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) \quad (6)$$

We will give the definition of $\mathcal{L}_{id}$ in Section 2.3.2.

2.3.1 Gated Linear Unit

Time-series data must be processed in natural language processing and speech processing. Deep learning of time-series data is usually performed by a recurrent neural network (RNN). However, because the architecture of CycleGAN is based on CNN, adopting CycleGAN to RNN is a difficult task. The activation units in a gated CNN are gated linear units (GLUs).

$$H_{l+1} = (H_l * W_l + b_l) \otimes \sigma(H_l * V_l + c_l) \quad (7)$$

where $\otimes$ means the element-wise product and σ is the sigmoid function. The GLU gate structure describes by Eq. 7 selectively propagates the time-dependent information, enabling modeling of the time-series data. CycleGAN-VC also uses GLUs as the activation units.

2.3.2 Identity Mapping Loss

When a generator designed to convert data from a specific domain obtains an input from a different domain, the foreign input should be unaltered. To this end, Kaneko and Kameoka proposed a constraint called the identity mapping loss. In the domain transfer of speech emotions, converting other domains is undesirable. The identity mapping loss is represented as follows:

$$\mathcal{L}_{id} = E_{y\sim p_{data}(y)}[\|G_{X\to Y}(y) - y\|_1] + E_{x\sim p_{data}(x)}[\|G_{Y\to X}(x) - x\|_1] \quad (8)$$

During training, $\mathcal{L}_{id}$ is added weight coefficient λ_{id} in Eq. 6.

2.4 Overall Architecture of the Proposed System

The system for transforming the emotions in speech is overviewed in Figure 3. The speech features are extracted from the speech data by TANDEM-STRAIGHT (Kawahara et al., 2008). The extracted features are the mel-frequency Cepstrum coefficients (MFCCs), F0, and the aperiodicity index (AP). Each feature is popular among voice conversion studies (Toda et al., 2007)(Kaneko and Kameoka, 2017)(Fang et al., 2018)(Sekii et al., 2017). The MFCC, F0 and AP represent the vocal tract characteristics, the pitch and intonations of the voice, and the mixed sound of random and periodic components, respectively. Therefore, MFCC and F0 are strongly dependent on the speaker and the emotion. The MFCC and F0 are converted by generators in the CycleGAN, and the AP is unchanged from those of the input AP. In this system, one set of networks is constructed for each gender because the gender greatly influences the F0 values.

2.5 Network Architecture and Training Details

The generators and discriminators are designed as 2D CNNs following StarGAN-VC (Kameoka et al., 2018). The features of CycleGAN-VC (Kaneko and Kameoka, 2017) used in our architecture are shown

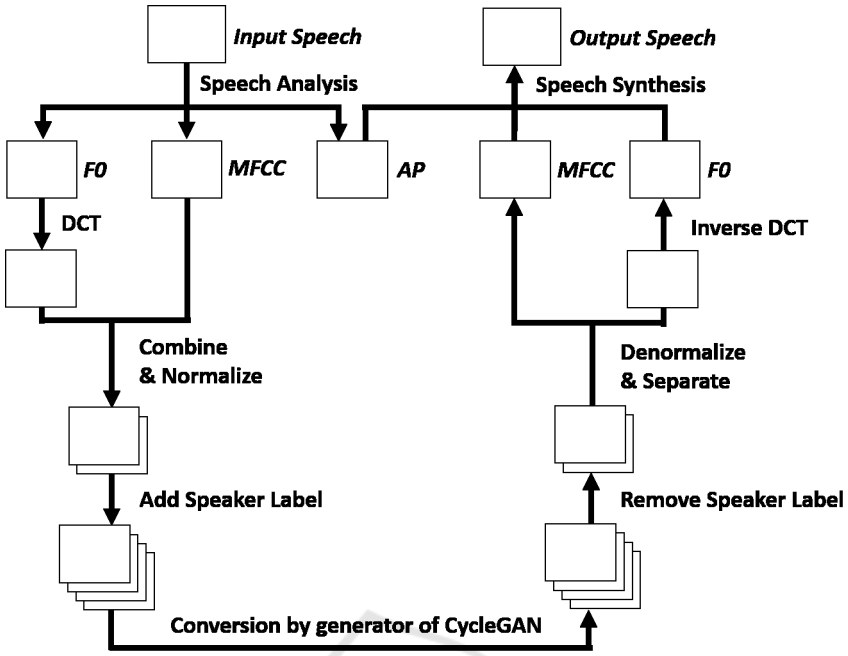

Figure 3: The system for transforming the emotions in speech.

below. Each generator comprises three downsampling layers, six residual layers (He et al., 2016), two upsampling layers, and a convolutional layer. Each discriminator is composed of a convolutional layer and four downsampling layers. The activation units employ GLUs in the hidden layers of the generators and discriminators, and sigmoid functions in the output layers. For layers other than the input and output layers in the generators and discriminators, instance normalization (Ulyanov et al., 2016) is applied. The networks are trained with the Adam optimizer (Kingma and Ba, 2015) with a batch size of 1. The training rate is 0.001 for generators and 0.00001 for discriminators. λ_{cyc} and λ_{id} are set to 10 and 15, respectively.

3 EXPERIMENTS

3.1 Datasets

Experiments were performed on the OGVC dataset (Arimoto et al., 2012), a Japanese speech dataset with emotion labels. There are eight types of emotions, namely, acceptance (ACC), anger (ANG), anticipation (ANT), disgust (DIS), fear (FEA), joy (JOY), sadness (SAD), and surprise (SUR). Speeches expressing these emotions were given by four professional voice actors: two males (MOY, MTY) and two females (FOY, FYN). Each datum in OGVC is labeled with one emotion type and the degree of its intensity on a scale of 0 to 3. In the present experiment, we use three emotions (“ANG”, “JOY”, “SAD”), following the choice made by (Aihara et al., 2012), each with an intensity level of 2 and 3. The speech data are downsampled to 16 kHz, and the acoustic features are extracted every 5 ms by TANDEM-STRAIGHT. In acoustic features, the first channel represents a 24-dimensional MFCC, the second channel represents a 24-dimensional F0 after discrete cosine transform (sliding window size: 25, frame size: 1), and the third and fourth channels employ one-hot speaker labels. The data combination is shown in Table 1. In Table 1, the column of “name” is combination name, the column of “speaker” is the name of speakers used in training and evaluation, the column of “intensity” is the emotional intensity used in training, the column of “train” is the number of the utterances used in training, and the column of “evaluation” is the number of the utterances used in evaluation.

Since the purpose of this research is conversion of emotional speeches, we are uninterested in less emotional speech. Therefore, only speech data of intensity 3 is used for evaluation data, and those that apparently failed in speech synthesis are removed from evaluation data.

Table 1: Conversion pattern.

Name	Speaker	Intesity	Train	Evaluation
MA	MOY, MTY	2, 3	68	6
MB	MOY, MTY	3	34	6
FA	FOY, FYN	2, 3	68	6
FB	FOY, FYN	3	34	6
MAX	MOY	2, 3	34	3
MAY	MTY	2, 3	34	3
FAX	FOY	2, 3	34	3
FAY	FYN	2, 3	34	3

3.2 Experimental Conditions

We conduct both objective and subjective evaluation experiments.

Objective experiments evaluate the emotional expressions using classifiers constructed in random forest of Weka (Garner, 1995). The maximum and minimum values of the 24-dimensional MFCC and F0, and the speaker information, are available for training the classifiers. The 10-fold cross validation results of the classifiers are shown in Table 2.

The subjective experiments evaluated the naturalness, identifiability of the speakers, and emotional expressions of the conversion results using ten subjects. The naturalness and identifiability of the speakers are evaluated by the mean opinion score, and the emotional expressions are evaluated by the classification accuracy.

3.3 Results

3.3.1 Objective Evaluation

The recalls and precisions of the classifier's outputs for the created speeches are shown in Table 3. MA is relatively more accurate than MB, and there is a tradeoff between FA and FB. These results confirm that when training with multiple speakers, speech with emotional intensity 2 contributes to the accuracy of the emotional expressions. In Table 3, MAX&MAY is a combination of MAX and MAY, and FAX&FAY is a combination of FAX and FAY. We also find that MAX&MAY is relatively more accurate than MB, and that FB is relatively more accurate than FAX&FAY. Therefore, speech with emotional intensity 2 can improve the accuracy of the emotional expressions when training with a single male speaker. However, in the case of a single female speaker, it does not improve. We conclude that speech with emotional intensity 2 provides useful training data when training with multiple speakers.

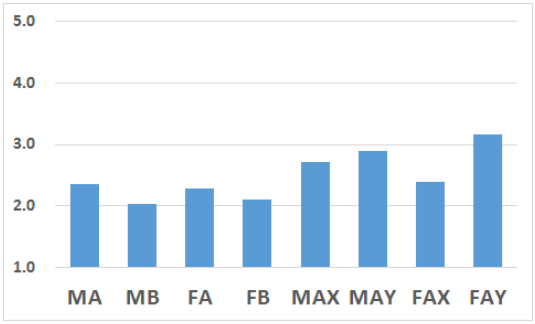

Figure 4: Naturalness results.

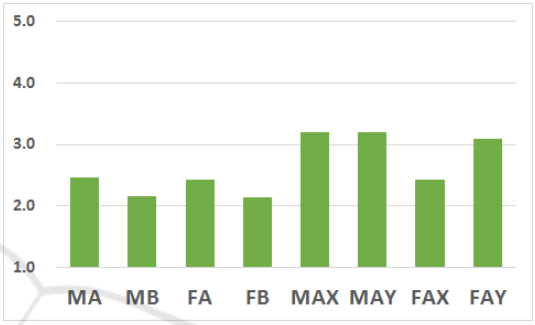

Figure 5: Similarities results.

3.3.2 Subjective Evaluation

The naturalness evaluation results are shown in Figure 4. MA is relatively more natural than MB, and FA is relatively more natural than FB. These results demonstrate that when training with multiple speakers, speech with emotional intensity 2 can contribute to the naturalness. MAX and MAY are relatively more natural than MB, and FAX and FAY are relatively more natural than FB. That is, speech with emotional intensity 2 also contributes to the naturalness when training with a single speaker. In fact, training with a single speaker achieved higher naturalness than training with multiple speakers.

The speaker identifiability evaluation results, which are evaluated by the similarity between the speaker identity found in the converted voice and the original, are shown in Figure 5. In MA, the speaker identities are more similar than in MB. In FA, the speaker identities are more similar than in FB. These results demonstrate that when training with multiple speakers, speech with emotional intensity 2 can contribute to the similarity. In MAX and MAY, the speaker identities are more similar than in MB. In FAX and FAY, the speaker identities are more similar than in FB. These results show that even when training with a single speaker, speech with emotional intensity 2 improve the similarity. Also, it is found that training

Table 2: Performance of classifiers [%].

	ANG		JOY		SAD	
	Recall	Precision	Recall	Precision	Recall	Precision
Male	52.9	50.7	56.9	62.1	68.1	65.3
Female	69.1	58.0	63.9	69.7	61.1	67.7

Table 3: Classification results of emotional expressions by classifiers [%].

	ANG		JOY		SAD	
	Recall	Precision	Recall	Precision	Recall	Precision
MA	81.8	42.9	50.0	66.7	41.7	100.0
MB	72.7	47.1	30.0	37.5	41.7	62.5
FA	41.7	27.8	41.7	35.7	16.7	50.0
FB	25.0	30.0	58.3	43.8	36.4	44.4
MAX&MAY	83.3	45.5	27.3	37.5	16.7	40.0
FAX&FAY	45.5	31.3	36.4	30.8	10.0	33.3

Table 4: Classification results of emotional expressions by subjects [%].

	ANG		JOY		SAD	
	Recall	Precision	Recall	Precision	Recall	Precision
MA	23.3	27.5	25.8	27.4	39.1	31.9
MB	19.1	22.1	35.0	35.0	36.7	32.6
FA	31.7	23.0	19.2	24.7	29.2	34.3
FB	20.8	19.5	12.5	17.0	30.9	25.4
MAX	18.3	17.5	11.7	16.7	28.3	22.7
MAY	16.7	16.9	6.0	7.7	41.7	34.7
FAX	18.3	18.3	8.0	10.8	38.0	30.2
FAY	8.0	7.7	15.0	20.0	20.0	15.9

with a single speaker achieves higher similarities than training with multiple speakers.

The results of the emotional expression evaluations are shown in Table 4. Here, the accuracies of MA and MB have little difference. However, FA is relatively more accurate than FB. These results show that when training with multiple speakers, speech with emotional intensity 2 contributes slightly to the accuracy of the emotional expressions. We also find that MB is relatively more accurate than MAX and MAY, and that FB is relatively more accurate than FAX and FAY. These results show that when training with a single speaker, speech with emotional intensity 2 cannot contribute to the accuracy of the emotional expressions. Some cases show extremely low scores. The reason is unclear.

From the results of the three evaluation experiments, we conclude that when training with multiple speakers, speech with emotional intensity 2 can contribute to the accuracy of the naturalness, the similarities, and the emotional expressions in speech. Therefore, speech with emotional intensity 2 provides useful training data when training with multiple speakers in conversion of emotional intensity 3. However, when training with a single speaker, speech with emotional intensity 2 can contribute to the accuracy of the naturalness and similarities, but not to the accuracy of the emotional expressions. Therefore, including speech with emotional intensity 2 in the training data will not help the transfer domain of the speech emotions when training with a single speaker. More speech with emotional intensity 3 might improve the accuracy of the emotional expressions when training with a single speaker.

3.4 Inception Scores

We calculated the inception scores (Salimans et al., 2016) of features generated by CycleGAN. The feature targeted in this additional experiment is F0 and MFCC of emotional intensity 3 generated from the evaluation data. In this experiment, we used InceptionV3 of Keras (Chollet et al., 2015). In order to make the input size of feature same as the inception model,

Table 5: Results of inception scores.

	Number of features	Inception score
All training data (intensity 2 & 3)	1500	1.0000968
All training data (intensity 3)	792	1.0000962
MA	178	1.0000932
MB	178	1.0000933
FA	172	1.0000973
FB	172	1.0000973
MAX	90	1.0000981
MAY	88	1.0000910
FAX	82	1.0000979
FAY	90	1.0000906

it was resized and complemented by adding random value. The inception scores are shown in Table 5. In Table 5, the inception score from each experiment have little difference.

4 CONCLUSION AND FUTURE WORKS

To improve natural speech synthesis, we performed domain transfer of the emotions in speech. We convert speeches to the "ANG", "JOY", and "SAD" domains by a network based on CycleGAN, which demonstrates high performance in VC. In addition, we investigate the usefulness of speech with emotional intensity 2 as training data. The evaluation experiment confirmed that speech with emotional intensity 2 provides useful training data when training with multiple speakers in conversion with emotional intensity 3.

In future works, we will investigate many-to-many conversions of emotion and conduct investigations on networks that can convert to arbitrary attributes, such as Semi-Latent GAN (Yin et al., 2017) and StarGAN (Choi et al., 2017).

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP16K00419, JP16K12411, JP17H04705, JP18H03229, JP18H03340, JP18K19835.

REFERENCES

Aihara, R., Takashima, R., Takiguchi, T., and Ariki, Y. (2012). Gmm-based emotional voice conversion spectrum spectrum and prosody. *Reports of the ... spring meeting the Acoustical Society of Japan, 1-R-29*, pages 503–504.

Arimoto, Y., Kawatsu, H., Ohno, S., and Iida, H. (2012). Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment. *Acoustical Science and Technology*, 33(6):359–369.

Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., and Choo, J. (2017). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR*, abs/1711.09020.

Chollet, F. et al. (2015). Keras. https://keras.io.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. (2017). Language modeling with gated convolutional networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 933–941, International Convention Centre, Sydney, Australia. PMLR.

Fang, F., Yamagishi, J., Echizen, I., and Lorenzo-Trueba, J. (2018). High-quality nonparallel voice conversion based on cycle-consistent adversarial network. *IEEE ICASSP 2018*, pages SP–P6.2.

Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. In *In Proc. of the New Zealand Computer Science Research Students Conference*, pages 57–64.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507.

Iida, A., Campbell, N., and Yasumura, M. (1999). Design and evaluation of synthesized speech with emotion. *Trans.IPS.Japan*, 40(2):479–486.

Kameoka, H., Kaneko, T., Tanaka, K., and Hojo, N. (2018). Stargan-vc: Non-parallel many-to-many voice conver-

sion with star generative adversarial networks. *CoRR*, abs/1806.02169.

Kaneko, T. and Kameoka, H. (2017). Parallel-data-free voice conversion using cycle-consistent adversarial networks. *CoRR*, abs/1711.11293.

Kawahara, H., Morise, M., Takahashi, T., Nishimura, R., Irino, T., and Banno, H. (2008). Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3933–3936.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR*.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324.

Liu, D., Domoto, K., Inoue, Y., and Utsuro, T. (2014). Emotional voice conversion utilizing f0 contour and duration of word accent type. *IEICE Tech. Rep. Speech*, 114(52):159–164.

Miyoshi, H., Saito, Y., Takamichi, S., and Saruwatari, H. (2017). Voice conversion using sequence-to-sequence learning of context posterior probabilities. In Lacerda, F., editor, *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 1268–1272. ISCA.

Nakashika, T. and Minami, Y. (2017). Speaker-adaptive-trainable boltzmann machine and its application to non-parallel voice conversion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017(1):16.

Orihara, R., Narasaki, R., Yoshinaga, Y., Morioka, Y., and Kokojima, Y. (2018). Approximation of time-consuming simulation based on generative adversarial network. *Proc. 42nd IEEE International Conference on Computer Software and Applications*, pages 171–176.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434.

Sakurai, A. and Kimura, S. (2013). The use of speech technologies in call centers - including para- and non-linguistic information. *IPSJ SIG Technical Reports*, 2013(2):1–6.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. (2016). Improved techniques for training gans. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc.

Sekii, Y., Orihara, R., Kojima, K., Sei, Y., Tahara, Y., and Ohsuga, A. (2017). Fast many-to-one voice conversion using autoencoders. In *ICAART (2)*, pages 164–174. SciTePress.

Toda, T., Black, A. W., and Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio, Speech Language Processing*, 15(8):2222–2235.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. S. (2016). Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022.

Yasuda, K., Orihara, R., Sei, Y., Tahara, Y., and Ohsuga, A. (2018a). An experimental study on transforming the emotion in speech using cyclegan. *Joint Agent Workshops and Sympososium*, page 5B.

Yasuda, K., Orihara, R., Sei, Y., Tahara, Y., and Ohsuga, A. (2018b). An experimental study on transforming the emotion in speech using gan. *IEICE Tech. Rep. SP*, 118(198):19–22.

Yasuda, K., Orihara, R., Sei, Y., Tahara, Y., and Ohsuga, A. (2018c). Transforming the emotion in speech using cyclegan. *IEICE Tech. Rep. AI*, 118(116):61–66.

Yin, W., Fu, Y., Sigal, L., and Xue, X. (2017). Semi-latent gan: Learning to generate and modify facial images from attributes. *CoRR*, abs/1704.02166.

Zhou, T., Krähenbühl, P., Aubry, M., Huang, Q., and Efros, A. A. (2016). Learning dense correspondence via 3d-guided cycle consistency. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 117–126. IEEE Computer Society.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.