HUMANACGAN: CONDITIONAL GENERATIVE ADVERSARIAL NETWORK WITH HUMAN-BASED AUXILIARY CLASSIFIER AND ITS EVALUATION IN PHONEME PERCEPTION

Yota Ueda[1], Kazuki Fujii[2], Yuki Saito[1], Shinnosuke Takamichi[1], Yukino Baba[3] and Hiroshi Saruwatari[1]

[1] Graduate School of Information Science and Technology, The University of Tokyo, Japan.
[2] National Institute of Technology, Tokuyama College, Japan.
[3] Faculty of Engineering, Information and Systems, University of Tsukuba, Japan.

ABSTRACT

We propose a conditional generative adversarial network (GAN) incorporating humans' perceptual evaluations. A deep neural network (DNN)-based generator of a GAN can represent a real-data distribution accurately but can never represent a human-acceptable distribution, which are ranges of data in which humans accept the naturalness regardless of whether the data are real or not. A HumanGAN was proposed to model the human-acceptable distribution. A DNN-based generator is trained using a human-based discriminator, i.e., humans' perceptual evaluations, instead of the GAN's DNN-based discriminator. However, the HumanGAN cannot represent conditional distributions. This paper proposes the HumanACGAN, a theoretical extension of the HumanGAN, to deal with conditional human-acceptable distributions. Our HumanACGAN trains a DNN-based conditional generator by regarding humans as not only a discriminator but also an auxiliary classifier. The generator is trained by deceiving the human-based discriminator that scores the unconditioned naturalness and the human-based classifier that scores the class-conditioned perceptual acceptability. The training can be executed using the backpropagation algorithm involving humans' perceptual evaluations. Our experimental results in phoneme perception demonstrate that our HumanACGAN can successfully train this conditional generator.

Index Terms— Generative adversarial network, human computation, conditional generator, auxiliary classifier, black-box optimization, speech perception

1. INTRODUCTION

Deep generative models of machine learning have contributed to media research [1, 2, 3]. A generative adversarial network (GAN) [1] is one of the strongest generative models. It has been applied in speech modeling [4, 5]. The GAN consists of a set of deep neural networks (DNNs), a generator, and a discriminator. The generator is trained to deceive the discriminator, and the discriminator is trained to distinguish between real and generated data. After iterating them, the trained generator represents a real-data distribution and can randomly generate data that follows the real-data distribution.

The GAN cannot represent an outer side of the real-data distribution. However, humans can accept out-sided media as natural. In speech perception, humans can recognize a voice as a human's even though the voice is out of the real-data distribution (i.e., an actual humans' voice). For example, that range contains synthesized voices or processed voices. In this study, we call this data range perceived by humans the *human-acceptable distribution*. HumanGAN [6] was proposed to model the human-acceptable distribution by using humans as the discriminator of the GAN. The top of **Fig. 1** shows
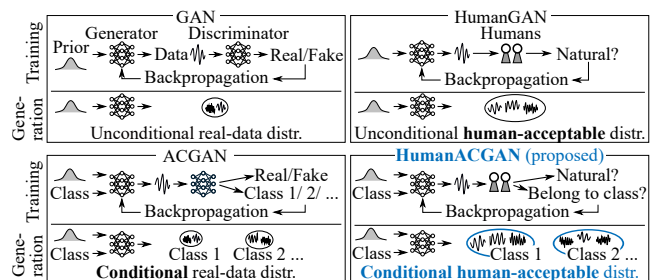
Fig. 1. Comparison of four GANs. We extend the HumanGAN to the conditional modeling in the same way the GAN was extended to the ACGAN. While an ACGAN trains a conditional generator with a DNN-based discriminator and an auxiliary classifier, the HumanACGAN trains one with a humans-based discriminator and an auxiliary classifier. The trained generator of the HumanACGAN can represent human-acceptable distributions conditioned by input class labels.

the comparison of a GAN and HumanGAN. The HumanGAN regards humans as a black-boxed system that outputs a difference in posterior probabilities given generated data. The DNN-based generator is trained using the backpropagation algorithm, including the human-based discriminator. The trained generator can represent the human-acceptable distribution.

However, the HumanGAN's generator cannot achieve more practical generative modeling such as text-to-speech synthesis [7] and voice conversion [8] because it cannot represent conditional distributions. Because the GAN was extended to the conditional GAN [9, 10], we expect that the HumanGAN can be extended to the conditional modeling. Namely, we train the HumanGAN's generator conditioned on the desired class label, and it represents the class-specific human-acceptable distribution as a result. This will contribute establishing a DNN-based framework to model the task-oriented perception by humans [11, 12].

In this paper, we propose the *HumanACGAN*, aiming to train a DNN-based conditional generator using humans' perceptual evaluations. To this end, we extend the HumanGAN by introducing an auxiliary classifier GAN (ACGAN) [10]: class-conditional expansion of a GAN. **Fig. 1** shows a comparison of four GANs: a GAN, an ACGAN, a HumanGAN, and our HumanACGAN. The ACGAN uses a DNN-based auxiliary classifier to train a conditional DNN-based generator in addition to a discriminator of the GAN. Our HumanACGAN replaces both the DNN-based discriminator and auxiliary classifier with humans. The HumanACGAN's generator is trained using human-perception-based discrimination and classification. This training is operated using an expansion of the

backpropagation-based algorithm incorporating humans' perceptual evaluations as a discriminator and an auxiliary classifier. We evaluated the HumanACGAN in phoneme perception, a task to train a DNN-based generator that represents phoneme-specific human-acceptable distributions. The experimental results show that 1) the phoneme-conditioned human-acceptable distributions are wider than the real-data ones and that 2) the HumanACGAN can successfully train a generator that represents conditional human-acceptable distributions.

2. RELATED WORKS

2.1. ACGAN

The ACGAN [10] trains a DNN-based conditional generator that represents real-data distributions conditioned on class labels. To achieve this, the ACGAN uses a DNN-based auxiliary classifier in addition to a DNN-based discriminator of the GAN. The generator $G(\cdot)$ transforms a prior noise $\boldsymbol{z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n, \cdots, \boldsymbol{z}_N]$ into data $\hat{\boldsymbol{x}} = [\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_n, \cdots, \hat{\boldsymbol{x}}_N]$, conditioned on class labels $\boldsymbol{c} = [c_1, \cdots, c_n, \cdots, c_N]$, i.e., $\hat{\boldsymbol{x}}_n = G(\boldsymbol{z}_n, \boldsymbol{c}_n)$. N denotes the number of data. The prior noise follows a known probability distribution, e.g., a uniform distribution $U(0, 1)$. Here, let real data be $\boldsymbol{x} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n, \cdots, \boldsymbol{x}_N]$. The real data also have the same corresponding class label $\boldsymbol{c}$. A discriminator $D_{\mathrm{S}}(\cdot)$ and a classifier $D_{\mathrm{C}}(\cdot)$ are used for the generator training. The $D_{\mathrm{S}}(\cdot)$ takes $\boldsymbol{x}_n$ or $\hat{\boldsymbol{x}}_n$ as an input and outputs a posterior probability that the input is real data. The $D_{\mathrm{C}}(\cdot)$ takes an $\boldsymbol{x}_n$ or $\hat{\boldsymbol{x}}_n$ as an input and outputs a posterior probability that the input source belongs to each class. Objective functions in training are formulated as

$$L_{\mathrm{S}} = \sum_{n=1}^{N} \log D_{\mathrm{S}}(\boldsymbol{x}_n) + \sum_{n=1}^{N} \log (1 - D_{\mathrm{S}}(G(\boldsymbol{z}_n, \boldsymbol{c}_n))), \quad (1)$$

$$L_{\mathrm{C}} = \sum_{n=1}^{N} \log D_{\mathrm{C}}(\boldsymbol{x}_n, \boldsymbol{c}_n) + \sum_{n=1}^{N} \log (D_{\mathrm{C}}(G(\boldsymbol{z}_n, \boldsymbol{c}_n), \boldsymbol{c}_n)). \quad (2)$$

L_{S} is the objective function of the GAN, and L_{C} enables training the conditional generator. The generator is trained to maximize $-L_{\mathrm{S}} + \lambda L_{\mathrm{C}}$, where λ is a hyperparameter.

The ACGAN trains the generator using real data, and the generator represents the real-data distribution conditioned on the data class. However, because the human-acceptable distribution is wider than the real-data distribution [6], the ACGAN cannot represent the full range of that distribution.

2.2. HumanGAN

The HumanGAN [6] was proposed to represent the human-acceptable distribution in a wider manner than a real-data distribution. A DNN-based unconditional generator of the HumanGAN is trained using humans' perceptual evaluations instead of the DNN-based discriminator of the GAN. The generator $G(\cdot)$ transforms prior noise $\boldsymbol{z}_n$ to data $\hat{\boldsymbol{x}}_n$ unconditionally. A human-based discriminator $D_{\mathrm{S}}(\cdot)$ is defined to deal with humans' perceptual evaluations. $D_{\mathrm{S}}(\cdot)$ takes $\hat{\boldsymbol{x}}_n$ as an input and outputs a posterior probability that the input is perceptually acceptable. The objective function is

$$L_{\mathrm{S}} = \sum_{n=1}^{N} D_{\mathrm{S}}(G(\boldsymbol{z}_n)). \quad (3)$$

The generator is trained to maximize L_{S}. A model parameter θ of $G(\cdot)$ is iteratively updated as $\theta \leftarrow \theta + \alpha \partial L_{\mathrm{S}} / \partial \theta$, where α is the learning coefficient, and $\partial L_{\mathrm{S}} / \partial \theta = \partial L_{\mathrm{S}} / \partial \hat{\boldsymbol{x}}_n \cdot \partial \hat{\boldsymbol{x}}_n / \partial \theta$.
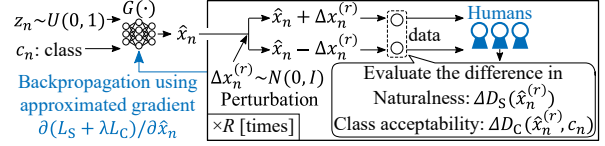
Fig. 2. Generator training process of proposed HumanACGAN. A human observes two perturbed data and evaluates their perceptual difference in two views: naturalness and class acceptability. They give global and class-specific gradients, respectively. Evaluations and perturbations are used for backpropagation to train the generator.

$\partial \hat{\boldsymbol{x}}_n / \partial \theta$ can be estimated analytically, but $\partial L_{\mathrm{S}} / \partial \hat{\boldsymbol{x}}_n$ cannot because the human-based discriminator $D_{\mathrm{S}}(\cdot)$ is not differentiable. The HumanGAN uses the natural evolution strategy (NES) [13] algorithm to approximate the gradient. A small perturbation $\Delta \boldsymbol{x}_n^{(r)}$ randomly generated from a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, and it is added to a generated datum $\hat{\boldsymbol{x}}_n$. r is the perturbation index ($1 \leq r \leq R$). $\sigma, \boldsymbol{I}$ are the standard deviation and the identity matrix, respectively. Next, a human observes two perturbed data $\{\hat{\boldsymbol{x}}_n + \Delta \boldsymbol{x}_n^{(r)}, \hat{\boldsymbol{x}}_n - \Delta \boldsymbol{x}_n^{(r)}\}$ and evaluates the difference in their posterior probabilities of naturalness:

$$\Delta D_{\mathrm{S}}(\hat{\boldsymbol{x}}_n^{(r)}) \equiv D_{\mathrm{S}}(\hat{\boldsymbol{x}}_n + \Delta \boldsymbol{x}_n^{(r)}) - D_{\mathrm{S}}(\hat{\boldsymbol{x}}_n - \Delta \boldsymbol{x}_n^{(r)}). \quad (4)$$

$\Delta D_{\mathrm{S}}(\hat{\boldsymbol{x}}_n^{(r)})$ ranges from -1 to 1. For instance, a human will answer $\Delta D_{\mathrm{S}}(\hat{\boldsymbol{x}}_n^{(r)}) = 1$ when he/she perceives that $\hat{\boldsymbol{x}}_n + \Delta \boldsymbol{x}_n^{(r)}$ is substantially more acceptable than $\hat{\boldsymbol{x}}_n - \Delta \boldsymbol{x}_n^{(r)}$. $\partial L_{\mathrm{S}} / \partial \hat{\boldsymbol{x}}$ is approximated with [13]

$$\frac{\partial L_{\mathrm{S}}}{\partial \hat{\boldsymbol{x}}_n} = \frac{1}{2\sigma^2 R} \sum_{r=1}^{R} \Delta D_{\mathrm{S}}(\hat{\boldsymbol{x}}_n^{(r)}) \cdot \Delta \boldsymbol{x}_n^{(r)}. \quad (5)$$

3. HUMANACGAN

3.1. Training

We propose the HumanACGAN. As shown in **Fig. 2**, the HumanACGAN has a conditional generator that represents class-specific human-acceptable distributions, while the HumanGAN only deals with one class. As well as the ACGAN, our HumanACGAN consists of a conditional generator, a discriminator, and an auxiliary classifier. The DNN-based generator $G(\cdot)$ is the same as that of the ACGAN; it transforms prior noise $\boldsymbol{z}_n$ into data $\hat{\boldsymbol{x}}_n$ conditioned on a class label $\boldsymbol{c}_n$. The discriminator and the auxiliary classifier are redefined to incorporate humans' perceptual evaluations. The human-based discriminator $D_{\mathrm{S}}(\cdot)$ is the same as that of the HumanGAN; it outputs a posterior probability of global (i.e., class-independent) naturalness. In addition, the human-based auxiliary classifier $D_{\mathrm{C}}(\cdot)$ evaluates class acceptability, whether or not the data can belong to the class $\boldsymbol{c}_n$. $D_{\mathrm{C}}(\cdot)$ takes $\hat{\boldsymbol{x}}_n$ generated from $G(\cdot)$ and class label $\boldsymbol{c}_n$ as inputs and outputs a posterior probability that the input is perceptually acceptable *as the class*". The objective functions are

$$L_{\mathrm{S}} = \sum_{n=1}^{N} D_{\mathrm{S}}(G(\boldsymbol{z}_n, \boldsymbol{c}_n)), \quad (6)$$

$$L_{\mathrm{C}} = \sum_{n=1}^{N} D_{\mathrm{C}}(G(\boldsymbol{z}_n, \boldsymbol{c}_n), \boldsymbol{c}_n). \quad (7)$$

A model parameter θ of $G(\cdot)$ is estimated by maximizing $L_{\mathrm{S}} + \lambda L_{\mathrm{C}}$. θ is iteratively updated as follows.

$$\theta \leftarrow \theta + \alpha \frac{\partial (L_{\mathrm{S}} + \lambda L_{\mathrm{C}})}{\partial \theta}, \quad (8)$$

$$\frac{\partial (L_S + \lambda L_C)}{\partial \theta} = \left(\frac{\partial L_S}{\partial \hat{\boldsymbol{x}}_n} + \lambda \frac{\partial L_C}{\partial \hat{\boldsymbol{x}}_n} \right) \cdot \frac{\partial \hat{\boldsymbol{x}}_n}{\partial \theta}. \tag{9}$$

As well as the HumanGAN, $\partial \hat{\boldsymbol{x}}_n / \partial \theta$ can be estimated analytically, but $\partial L_S / \partial \hat{\boldsymbol{x}}_n$ and $\partial L_C / \partial \hat{\boldsymbol{x}}_n$ cannot. We formulate gradient approximation using the NES algorithm. A human observes two perturbed data $\{\hat{\boldsymbol{x}}_n + \Delta \boldsymbol{x}_n^{(r)}, \hat{\boldsymbol{x}}_n - \Delta \boldsymbol{x}_n^{(r)}\}$ and evaluates two kinds of difference in their posterior probabilities. The first is the same as the HumanGAN, i.e., the human evaluates "to what degree the inputs are perceptually different in the view of naturalness" and answers $\Delta D_S (\hat{\boldsymbol{x}}_n^{(r)})$ for approximating $\partial L_S / \partial \hat{\boldsymbol{x}}_n$ as in Eq. (4). The second is for the class-specific question, "to what degree the inputs are different in the view of class acceptability." The difference in the posterior probability $\Delta D_C (\hat{\boldsymbol{x}}_n^{(r)}, \boldsymbol{c}_n)$ is defined as

$$\Delta D_C(\hat{\boldsymbol{x}}_n^{(r)}, \boldsymbol{c}_n) \equiv D_C \left(\hat{\boldsymbol{x}}_n + \Delta \boldsymbol{x}_n^{(r)}, \boldsymbol{c}_n \right)$$
$$- D_C \left(\hat{\boldsymbol{x}}_n - \Delta \boldsymbol{x}_n^{(r)}, \boldsymbol{c}_n \right). \tag{10}$$

In the same way as the HumanGAN, $\partial L_S / \partial \hat{\boldsymbol{x}}_n$ can be approximated using $\Delta D_S (\hat{\boldsymbol{x}}_n^{(r)})$. Unlike $\Delta D_S (\cdot)$, $\Delta D_C (\cdot)$ becomes the class-specific difference. Namely, $\Delta D_C (\hat{\boldsymbol{x}}_n^{(r)}, \boldsymbol{c}_n)$ will be 1 when the human perceives that $\hat{\boldsymbol{x}}_n + \Delta \boldsymbol{x}_n^{(r)}$ is substantially more acceptable than $\hat{\boldsymbol{x}}_n - \Delta \boldsymbol{x}_n^{(r)}$ as the presented class. $\Delta D_C (\hat{\boldsymbol{x}}_n^{(r)}, \boldsymbol{c}_n)$ ranges from -1 to 1, and $\partial L_C / \partial \hat{\boldsymbol{x}}$ is approximated with

$$\frac{\partial L_C}{\partial \hat{\boldsymbol{x}}_n} = \frac{1}{2\sigma^2 R} \sum_{r=1}^{R} \Delta D_C \left(\hat{\boldsymbol{x}}_n^{(r)}, \boldsymbol{c}_n \right) \cdot \Delta \boldsymbol{x}_n^{(r)}. \tag{11}$$

Note that the HumanACGAN does not explicitly involve classification problems. In other words, it only needs to estimate the gradient of multiclass probability function (i.e., the softmax function used in an ACGAN) as the degree of class-specific acceptability.

3.2. Limitations

The HumanGAN [6] suffers from a mode collapse problem [14], a gradient vanishing, and scalability to the data size and dimensionality. These problems still remain in our HumanACGAN. Therefore, our experiments followed the initialization way and data preprocessing of the paper [6]. We first estimated histograms of the posterior probabilities and initialized θ by referring to the histograms. Also, we reduced the data dimensionality using principal component analysis (PCA).

4. EXPERIMENTAL EVALUATION

This section describes an experimental test of the effectiveness of the HumanACGAN using Japanese phonemes as classes.

4.1. Experimental setup

Two phonemes we used were /i/ and /e/ of Japanese vowels. We basically followed the experimental setup of the HumanGAN paper [6]. The used data consisted of 199 female speakers' utterances recorded in the JVPD [15] corpus. Before extracting speech features, the speech waveforms were downsampled at 16 kHz, and their powers were normalized. 513-dimensional log spectral envelopes, fundamental frequency (F0), and aperiodicities (APs) were extracted every 5 ms from the speech waveforms using the WORLD vocoder [16, 17]. We extracted the speech features of the vowels /i/ and /e/ using phoneme alignment obtained by Julius [18]. We applied PCA to the log spectral envelopes and used the first and second principal components. The two-dimensional principal components
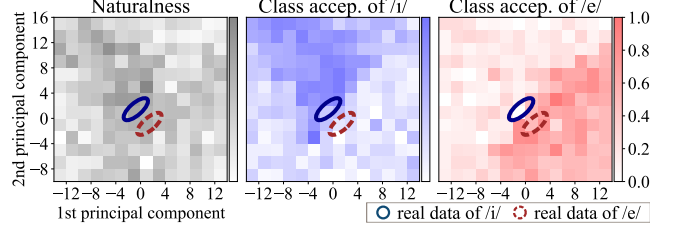
Fig. 3. Color maps representing posterior probabilities of naturalness and class acceptability ("accep.") of /i/ and /e/.
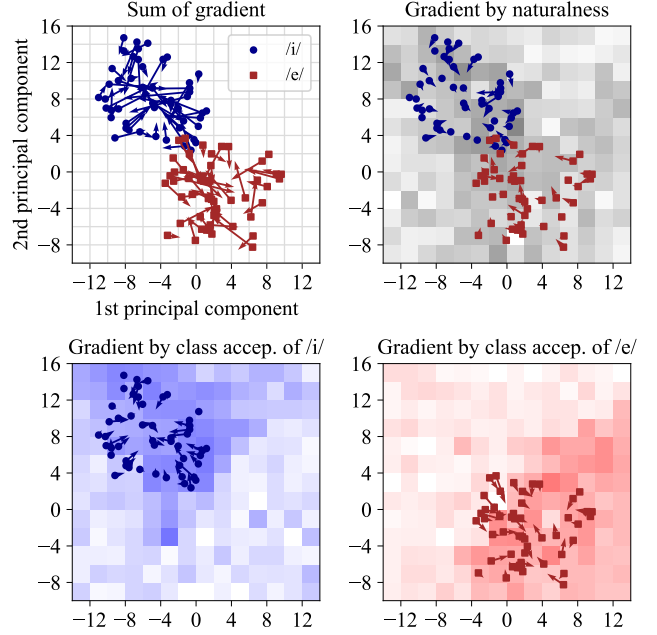

Fig. 4. Generated data (points) and gradient (arrows) estimated by our algorithm. "accep." denotes acceptability. The gradients of the upper left figure are the weighted sum of the other three gradients. We can see that the gradients for naturalness and class acceptability are respectively pointing to darker (i.e., higher posterior) zones.

were normalized to have zero-mean and unit-variance. The speech waveforms to be evaluated were synthesized using features obtained from a DNN-based conditional generator in the following way. First, the first and second principal components were generated by the generator and de-normalized. For the other features, i.e., the F0 and the APs, we used the average of all speakers. These corresponded to the speech features of one frame (5 ms). Next, we copied the features for 200 frames to make the perceptual evaluations easy and synthesized 1 second of speech waveforms using the WORLD vocoder.

4.2. Difference between real-data and human-acceptable distributions

First, we confirmed that the human-acceptable distributions were wider than the real-data distributions. We carried out two tests: evaluations of humans' tolerance of naturalness and class acceptability. In this experiment, class acceptability denotes whether the data sounded like the presented phoneme. We split the two-dimensional space into grids and generated a speech waveform for every grid. Then, we presented a speech waveform to a listener, and the listener rated the naturalness on a 5-point scale from 1 (bad) to 5 (excellent). Next, we presented a speech waveform and a phoneme label
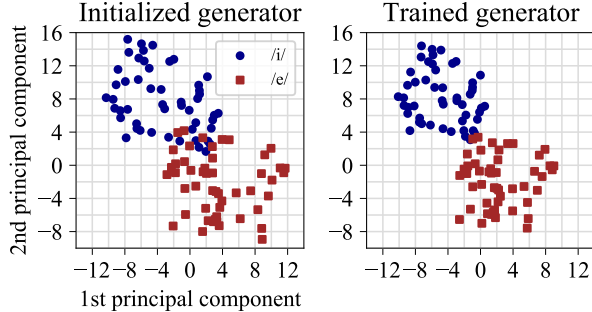
Fig. 5. Data generated from the initialized or trained generator. The colors of the data points correspond to those of the color maps for class acceptability in **Fig. 3**. Based on posterior probabilities shown in **Fig. 3**, we can say that the training makes the data move to have higher posterior probabilities.
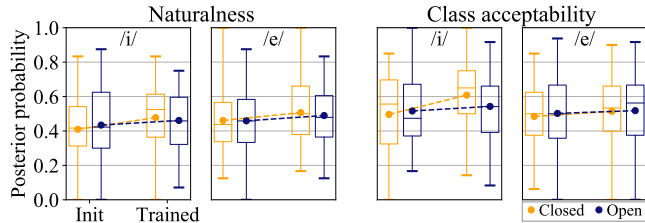
Fig. 6. Posterior probability of data generated from initialized ("Init") or trained ("Trained") generators. The boxes indicate first, second (i.e., median), and third quantiles. The line plot indicates mean value.

(/i/ or /e/) to a listener, and the listener rated the class acceptability on the same scale. One-through-five of the obtained scores corresponded to $0.00, 0.25, 0.50, 0.75$, and 1.00 of the posterior probability ($D_S(\hat{\boldsymbol{x}}_n)$ or $D_C(\hat{\boldsymbol{x}}_n, \boldsymbol{c}_n)$), respectively. We used the Lancers crowdsourcing platform [19] to execute the evaluations. The posterior probabilities were averaged for each grid. Each grid was scored by at least five listeners. The total number of listeners was 105.

Fig. 3 shows the results. As described in **Section 4.1**, the real data ware normalized to have zero-mean and unit-variance, and the ACGAN represents this range. However, as shown in **Fig. 3**, the human-acceptable distributions, i.e., darker zones of the color maps, were wider than the real-data distribution for both naturalness and class acceptability. Therefore, we obtained support for the HumanACGAN being able to represent this distribution, which an ACGAN cannot adequately represent.

4.3. Transition of generated data during training

Next, we executed the HumanACGAN training and qualitatively evaluated the generated data and approximated gradient. The prior noise followed a two-dimensional uniform distribution $U(0, 1)$. We randomly generated this prior noise before the training and fixed it during the iterations. Half of the data belonged to the class label /i/, and the remaining data belonged to /e/. The generator was a small feed-forward neural network consisting of a two-unit input layer, 2×4-unit sigmoid hidden layers, and a two-unit linear output layer. The output layer was conditioned using the two dimensional one-hot class label vector. We performed random initialization of the generator until these conditions were satisfied: The data generated from the initialized generator should cover ranges of the higher posterior probabilities of naturalness and class acceptability, and the data of each class should leak into the other class distribution for

determining the ability to classify. We empirically set the hyperparameter $\lambda = 2$ and used the gradient descent method with a learning rate of $\alpha = 0.0005$ for the training. We used Chainer [20] for the implementation. The number of generated data N, the number of perturbations R, the number of training iterations, and the standard deviation of NES σ were set to 50, 5, 4, and 2.0, respectively. We carried out two perceptual evaluations during the HumanACGAN training. First, we presented two speech waveforms to a listener, and the listener rated which one was more natural on a 5-point scale: 1: the first one, 3: equal, and 5: the second one. Next, we presented two speech waveforms and one phoneme label (/i/ or /e/) to a listener, and the listener rated which one sounded like the presented phoneme on the same scale. One-through-five of the obtained scores corresponded to $1.0, 0.5, 0.0, -0.5$, and -1.0 of the difference in the posterior probabilities ($\Delta D_S(\hat{\boldsymbol{x}}_n)$ or $\Delta D_C(\hat{\boldsymbol{x}}_n, \boldsymbol{c}_n)$), respectively.

Fig. 4 shows the gradient of every data point. The posterior probability of **Fig. 3** is also drawn for reference, but note that we never used it during the training. We can say that the gradient by both naturalness and class acceptability points to each darker range. This qualitatively indicates that the gradient shown in Eqs. (5), (11) was properly estimated. **Fig. 5** shows the data generated from initialized and trained generators. On the basis of posterior probabilities shown in **Fig. 3**, we can say that the training makes the data move so that they have higher posterior probabilities of both naturalness and class acceptability. This qualitatively indicates that the loss functions shown in Eqs. (6), (7) improved the generator to represent the conditional human-acceptable distributions.

4.4. Increase in posterior probabilities during training

Finally, we quantitatively verified that the training increases the posterior probabilities of naturalness $D_S(\cdot)$, and class acceptability $D_C(\cdot)$. We prepared two types of data: closed and open. The closed data were generated from the prior noise used while the training, and open data was generated from a newly sampled prior noise that was not used during the training. The posterior probabilities of the closed/open data were scored in the same manner as the ones in **Section 4.2**. The total number of listeners was 160.

Fig. 6 shows the box plots of the posterior probability. The training iteration increased the posterior probabilities of both naturalness and class acceptability with not only the closed data but also the open data. Therefore, we can say that our training can increase the objective values consisting of posterior probabilities and that the generator of the HumanACGAN can represent conditional human-acceptable distributions.

5. CONCLUSION

We proposed the HumanACGAN, which can conditionally represent humans' perceptually acceptable distributions. We reconfigured a discriminator and an auxiliary classifier of an ACGAN to utilize humans' perceptual evaluations. The DNN-based conditional generator of the HumanACGAN was trained using two kinds of human's perceptual evaluations: naturalness and class acceptability, which were used for the discriminator and the auxiliary classifier, respectively. We evaluated the effectiveness of the HumanACGAN using qualitative and quantitative experiments. We are planning to expand the scalability of the HumanACGAN in terms of scalability to the data size and dimensionality as part of our future work.

6. REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Montreal, Canada, Dec. 2014, pp. 2672–2680.

[2] D. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv*, vol. abs/1312.6114, 2013.

[3] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," in *Proc. ICLR*, San Diego, U.S.A., May 2015.

[4] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, Jan. 2018.

[5] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on generative adversarial networks," in *proc. ICASSP*, Brighton, United Kingdom, May 2019, pp. 6955–6959.

[6] K. Fujii, Y. Saito, S. Takamichi, Y. Baba, and H. Saruwatari, "HumanGAN: generative adversarial network with human-based discriminator and its evaluation in speech perception modeling," in *Proc. ICASSP*, Barcelona, Spain, May 2020, pp. 6239–6243.

[7] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[8] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[9] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[10] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. ICLR*, Vancouver, Canada, Apr. 2018.

[11] C. Chiu, Y. Koyama, Y. Lai, T. Igarashi, and Y. Yue, "Human-in-the-loop differential subspace search in high-dimensional latent space," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 85–1, 2020.

[12] J. Peterson, J. Suchow, K. Aghi, A. Ku, and T. Griffiths, "Capturing human category representations by sampling in deep feature spaces," *arXiv preprint arXiv:1805.07644*, May 2018.

[13] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proc. ICML*, Stockholm, Sweden, Jul. 2018, vol. 2, pp. 2137–2146.

[14] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," in *Proc. NIPS*, Barcelona, Spain, Dec. 2016.

[15] "Vowel database: Five Japanese vowels of males, females, and children along with relevant physical data (JVPD)," http://research.nii.ac.jp/src/en/JVPD.html.

[16] 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. E99-D, no. 7, pp. 1877–1884, Jul. 2016.

[17] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, Nov. 2016.

[18] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, Aalborg, Denmark, Sep. 2001, pp. 1691–1694.

[19] "Lancers," https://www.lancers.jp/.

[20] S. Tokui, R. Okuta, T. Akiba, Y. Niitani, T. Ogawa, S. Saito, S. Suzuki, K. Uenishi, B. Vogel, and H. Y Vincent, "Chainer: A deep learning framework for accelerating the research cycle," in *Proc. KDD*, Anchorage, U.S.A., Aug. 2019, pp. 2002–2011.