HumanDiffusion: diffusion model using perceptual gradients

Yota Ueda, Shinnosuke Takamichi, Yuki Saito, Norihiro Takamune, and Hiroshi Saruwatari
The University of Tokyo, Japan

Session: Wed-P9
Paper number: 1680

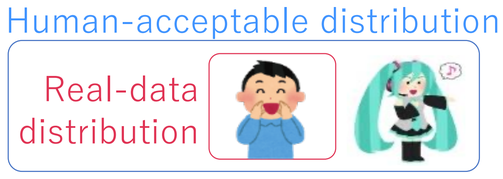
1. SYNOPSIS

1. Proposes a diffusion model that represents a human-acceptable distribution.
2. Demonstrates that the proposed algorithm successfully samples data that follow a human-acceptable distribution.

2. RESEARCH BACKGROUND

Human-acceptable distribution

- A probability distr. in which data humans can accept as natural.
- A human-acceptable distribution is typically wider than a real-data distribution.



Modeling the distribution with DNN (deep neural network)

- Can generate speech in a human-acceptable distribution.
- Can generate speech that do not exist.

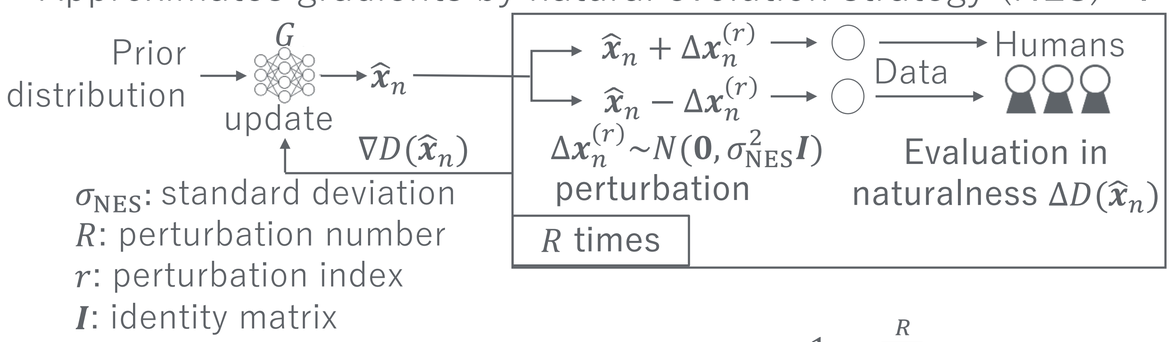
3. RELATED WORKS

GAN (generative adversarial network)[1]

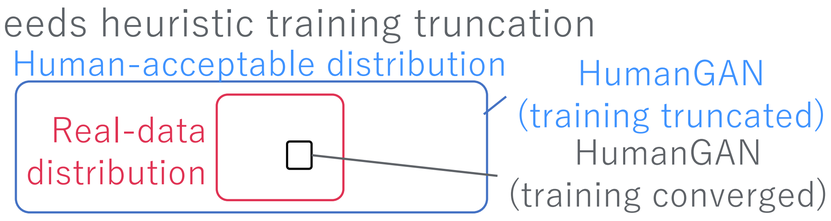
- Consists of a DNN-generator and a DNN-discriminator.
- The generator outputs data in a real-data distribution.
- Gradients ∇D for updating generator's parameters can be calculated analytically.

HumanGAN[2]

- Consists of a DNN-generator and a humans-discriminator.
- The generator outputs data in a human-acceptable distribution.
- Gradients ∇D cannot be calculated analytically.
- Approximates gradients by natural evolution strategy (NES)[4].



- Gradients are approximated as $\nabla D(\hat{\boldsymbol{x}}_n) \approx \frac{1}{2\sigma_{\mathrm{NES}}R}\sum_{r=1}^{R} \Delta D\left(\hat{\boldsymbol{x}}_n^{(r)}\right) \cdot \Delta\boldsymbol{x}_n^{(r)}$
- Problem: needs heuristic training truncation



Diffusion model[3]

- Consists of a score network S only.
- The score network models score $\nabla \log p(\boldsymbol{x})$ of the real-data distribution $p(\boldsymbol{x})$.
- Data is sampled with Langevin dynamics using estimated scores and it follows the real-data distribution.

From: https://yang-song.net/blog/2021/score/

- Sampling with Langevin dynamics

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \frac{\epsilon^2}{2}S(\boldsymbol{x}_t) + \epsilon\boldsymbol{r}$$

- If score of humans' evaluation function $D(\boldsymbol{x})$ is modeled instead of $p(\boldsymbol{x})$, sampled data follows a human-acceptable distribution.

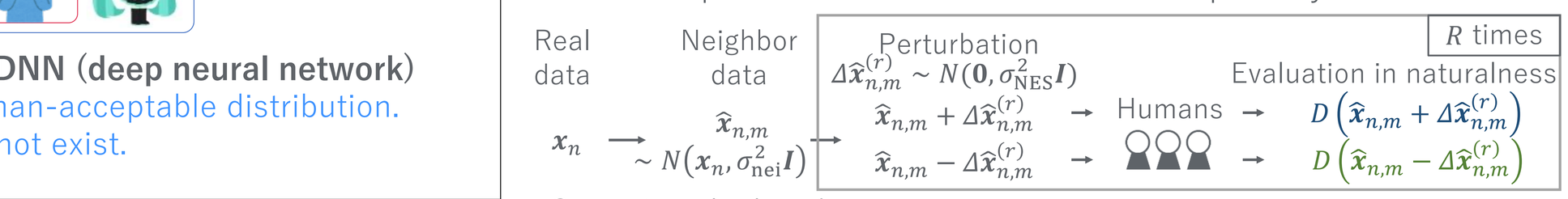
4. PROPOSED METHOD: HumanDiffusion

Method

- Sampling data $\boldsymbol{x} \sim \frac{1}{\int D} D(\boldsymbol{x})$ with score $\nabla \log D(\boldsymbol{x})$

Approximation of score

- Approximating score because it cannot be calculated analytically.
- Neighbor data is sampled from real data for gradient around them.
- Evaluates perturbed data in naturalness acceptability for NES.



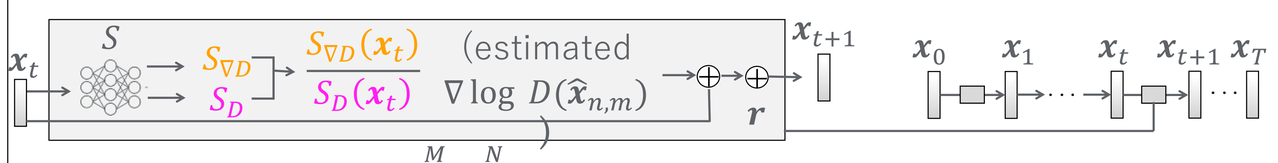
- Score is calculated as

$$\nabla \log D(\hat{\boldsymbol{x}}_{n,m}) = \frac{\nabla D(\hat{\boldsymbol{x}}_{n,m})}{D(\hat{\boldsymbol{x}}_{n,m})}$$

$$\nabla D(\hat{\boldsymbol{x}}_{n,m}) = \frac{1}{2\sigma_{\mathrm{NES}}R}\sum_{r=1}^{R}\left(D\left(\hat{\boldsymbol{x}}_{n,m} + \Delta\hat{\boldsymbol{x}}_{n,m}^{(r)}\right) - D\left(\hat{\boldsymbol{x}}_{n,m} - \Delta\hat{\boldsymbol{x}}_{n,m}^{(r)}\right)\right) \cdot \Delta\hat{\boldsymbol{x}}_{n,m}$$

$$D(\hat{\boldsymbol{x}}_{n,m}) = \mathrm{argmax}_D\, p_D\left(\left(D\left(\hat{\boldsymbol{x}}_{n,m} + \Delta\hat{\boldsymbol{x}}_{n,m}^{(r)}\right) + D\left(\hat{\boldsymbol{x}}_{n,m} - \Delta\hat{\boldsymbol{x}}_{n,m}^{(r)}\right)\right)/2\right)$$

Training score network and sampling



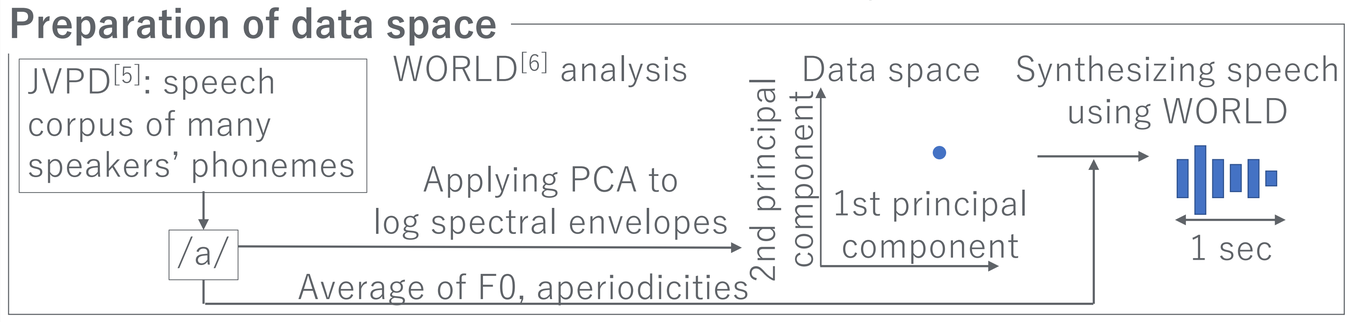
- Loss function: $L = \sum_{m=1}^{M}\sum_{n=1}^{N}\left(\left|S_{\nabla D}(\hat{\boldsymbol{x}}_{n,m}) - \nabla D(\hat{\boldsymbol{x}}_{n,m})\right|^2 + \left|S_D(\hat{\boldsymbol{x}}_{n,m}) - D(\hat{\boldsymbol{x}}_{n,m})\right|^2\right)$
- Sampling: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \frac{\epsilon^2}{2}\frac{S_{\nabla D}(\boldsymbol{x}_t)}{S_D(\boldsymbol{x}_t)} + \epsilon\boldsymbol{r}$

5. EXPERIMENTAL EVALUATIONS

Purpose

- Confirm HumanDiffusion models human-acceptable distribution.

Preparation of data space

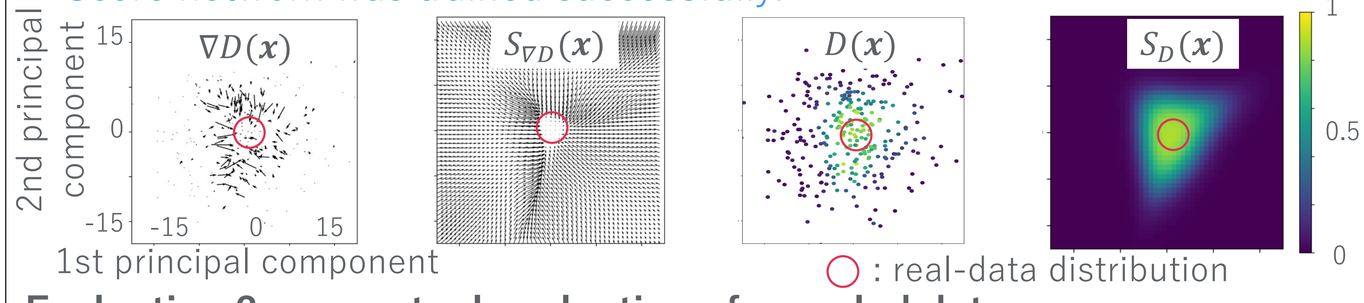


Hyper-parameters

- n: 100, m: 3, #humans: 105, σ_{nei}: 10, σ_{NES}: 1.0, ϵ: 0.003, T: 10000, R: 20
- Score network: 3 fully-connected layers, optimizer: Adam

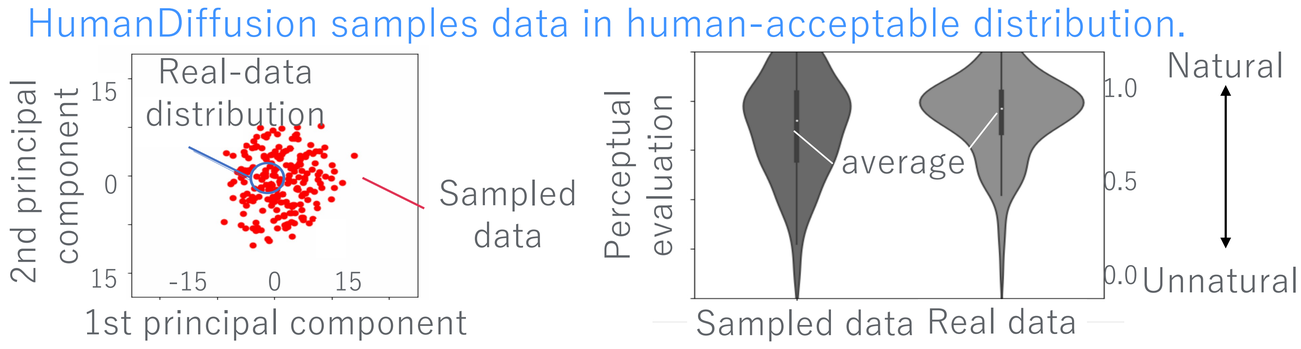
Evaluation 1: confirm score network learns $D(\boldsymbol{x})$ and $\nabla D(\boldsymbol{x})$

- $S_D(x)$ got higher values around the real-data distribution.
- $S_{\nabla D}(x)$ pointed to the real-data distribution.
- Score network was trained successfully.



Evaluation 2: perceptual evaluation of sampled data

- Humans evaluated sampled data and real data in naturalness.
- The evaluation value was almost same although the distribution of sampled data was wider than real-data distribution.
- HumanDiffusion samples data in human-acceptable distribution.



References [1] Goodfellow et al., Proc. NIPS, 2014. [2] Fujii et al., Proc. ICASSP, 2020. [3] Song, Yang, et al. arXiv:2011.13456 (2020). [4] Ilyas et al., Proc. ICML, 2018. [5] "http://research.nii.ac.jp/src/en/JVPD.html [6] Morise et al., IEICE, 2016.