SLP-P22.2

ICASSP 2024 KOREA

# Training Generative Adversarial Network-Based Vocoder with Limited Data Using Augmentation-Conditional Discriminator

*Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka*

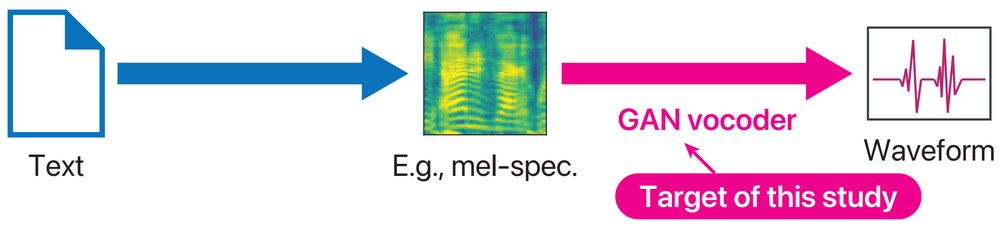
NTT Corporation, Japan

Audio samples

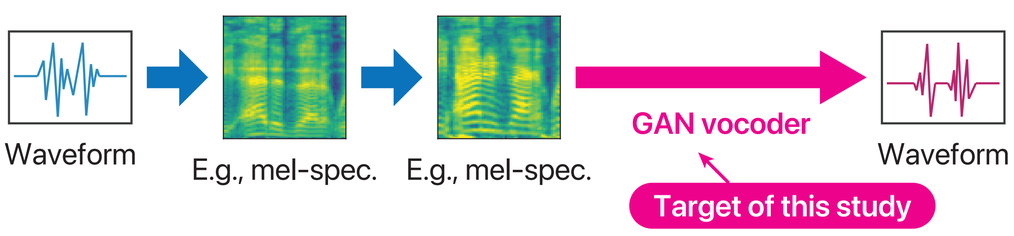
## 1 Introduction

### 1. Target of this study: GAN vocoder

Text-to-speech synthesis (Text → Waveform)




Voice conversion (Waveform → Waveform)

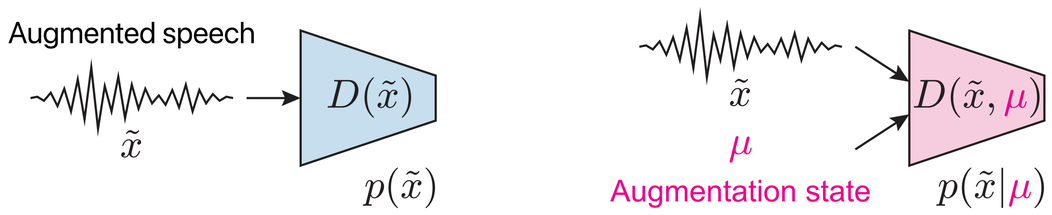



### 2. Objective: Train GAN vocoder with *limited* data

**Problem** GAN vocoder requires a large amount of training data (e.g., > 10 h)

**Objective** Reduce the amount of training data (e.g., ≈ 10 min)

### 3. Key idea: Augmentation-conditional discriminator




**(a) Standard discriminator**

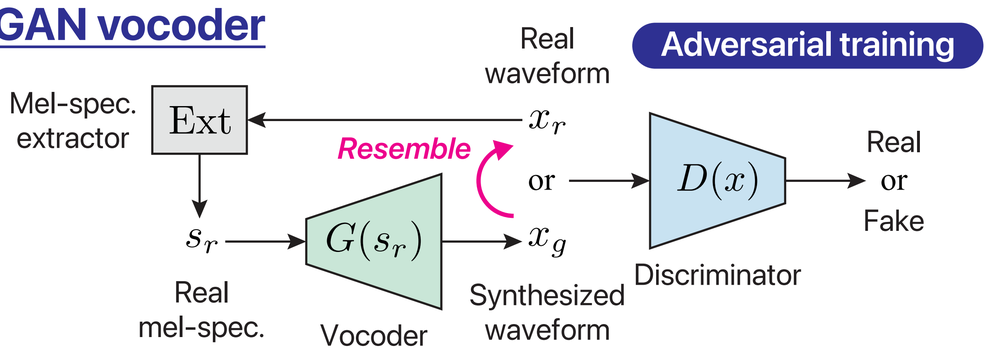
Assess speech ***without*** considering (extraordinary) augmentation state

**(b) *AugCond*$D$ (proposed)**

Assess speech ***with*** considering augmentation state

## 2 Preliminaries

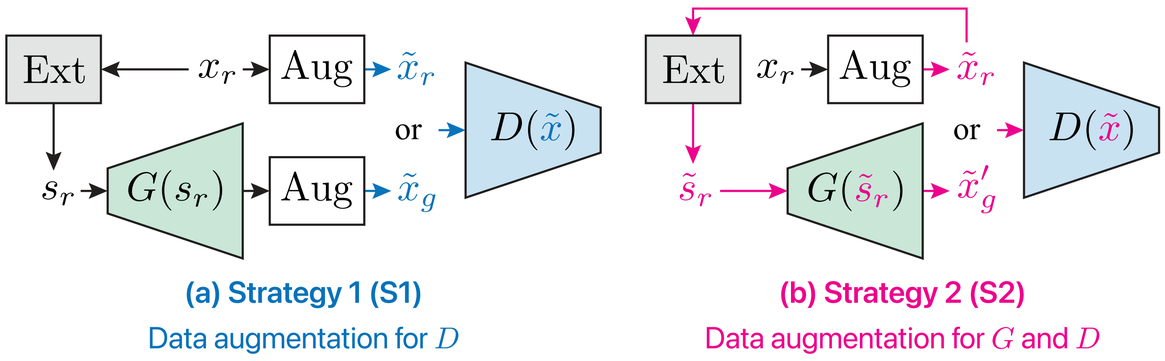
### 1. GAN vocoder




Train vocoder that can synthesize waveform resembling real waveform

### 2. Data augmentation for GAN vocoder

Two data augmentation strategies can be considered




**(a) Strategy 1 (S1)**

Data augmentation for $D$

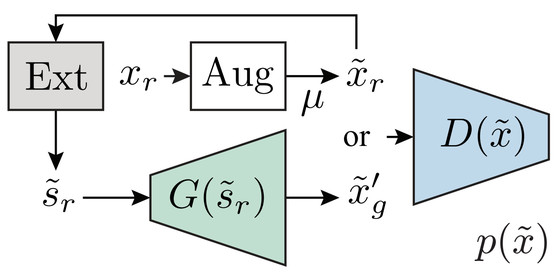
**(b) Strategy 2 (S2)**

Data augmentation for $G$ and $D$

**S2 was adopted because data augmentation for $G$ and $D$ is more effective**

## 3 Limitation of standard discriminator

### Assess speech *without* considering augmentation state



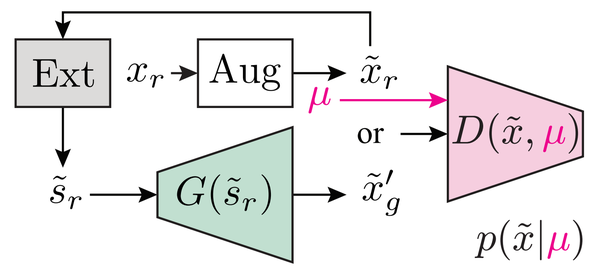

Standard discriminator is ***unconditional*** and *agnostic* to data augmentation

$$\mathcal{L}_{\mathrm{Adv}}(D) = \mathbb{E}_{(\tilde{x}_r, \tilde{s}_r, \mu)}[(D(\tilde{x}_r) - 1)^2 + (D(G(\tilde{s}_r)))^2]$$
$$\mathcal{L}_{\mathrm{Adv}}(G) = \mathbb{E}_{(\tilde{s}_r, \mu)}[(D(G(\tilde{s}_r)) - 1)^2]$$

May consider augmented (extraordinary) speech as desired real speech

## 4 Proposal: *AugCond*$D$

### Assess speech *with* considering augmentation state




***AugCond*$D$ is *conditioned on* augmentation state**

$$\mathcal{L}_{\mathrm{Adv}}(D) = \mathbb{E}_{(\tilde{x}_r, \tilde{s}_r, \mu)}[(D(\tilde{x}_r, \mu) - 1)^2 + (D(G(\tilde{s}_r), \mu))^2]$$
$$\mathcal{L}_{\mathrm{Adv}}(G) = \mathbb{E}_{(\tilde{s}_r, \mu)}[(D(G(\tilde{s}_r), \mu) - 1)^2]$$

Distinguish augmented speech with original non-augmented speech

## 5 Experiments

### 1. Investigation of benchmark performance

**Dataset: LJSpeech** (Single English speaker) [Ito&Johnson2017]

- **100%** (sufficient data): 23.7 h
- **1%** (limited data): 14.4 min

**Evaluation metrics:**

- **MOS↑:** Mean opinion score on naturalness (from 1 (bad) to 5 (excellent))
- **UTMOS↑:** Predicted mean opinion score [Saeki+2022]
- **Periodicity↓:** Distance in periodicity [Morrison+2022]
- **cFW2VD↓:** Distribution distance in wav2vec 2.0 [Kaneko+2022]

**Data augmentation:** mixup [Zhang+2018]: $\tilde{x}_r = mx_r^1 + (1-m)x_r^2 \ \ (m \sim U(0,1))$

**Compared models:**

- **HiFi** [Kong+2020]: HiFi-GAN
- **HiFi-phase** [Lee+2023]: HiFi-GAN + PhaseAug
- **HiFi-mix**: HiFi-GAN (standard $D$) + mixup
- **HiFi-ACD-mix**: HiFi-GAN + *AugCond*$D$ + mixup

| Model | Data | MOS↑ | UTMOS↑ | Periodicity↓ | cFW2VD↓ |
|---|---|---|---|---|---|
| Ground truth | – | $4.69_{\pm 0.07}$ | 4.38 | – | – |
| HiFi | 100% | $4.48_{\pm 0.08}$ | 4.23 | 0.106 | 0.022 |
| HiFi-phase | 100% | $4.49_{\pm 0.08}$ | 4.23 | 0.105 | 0.023 |
| HiFi-mix | 100% | $4.35_{\pm 0.09}$ | 4.19 | 0.108 | 0.023 |
| HiFi-ACD-mix | 100% | $4.42_{\pm 0.09}$ | 4.23 | 0.107 | 0.020 |
| HiFi | 1% | $2.89_{\pm 0.12}$ | 3.47 | 0.168 | 0.090 |
| HiFi (early stopped) | 1% | $3.53_{\pm 0.12}$ | 3.75 | 0.143 | 0.079 |
| HiFi-phase | 1% | $3.01_{\pm 0.12}$ | 3.46 | 0.166 | 0.091 |
| HiFi-phase (early stopped) | 1% | $3.62_{\pm 0.12}$ | 3.71 | 0.143 | 0.073 |
| HiFi-mix | 1% | $3.88_{\pm 0.11}$ | 3.83 | 0.125 | 0.047 |
| HiFi-ACD-mix | 1% | $4.25_{\pm 0.10}$ | 4.00 | 0.117 | 0.036 |

1. *AugCond*$D$ has no adverse effects under sufficient data conditions (100%)
2. *AugCond*$D$ improves speech quality under limited data conditions (1%)

### 2. Investigation of general utility

The same tendencies are observed in the following cases:

1. With different vocoders (HiFi-GAN V2 [Kong+2020], iSTFTNet [Kaneko+2022])

| Model | Data | UTMOS↑ | Periodicity↓ | cFW2VD↓ |
|---|---|---|---|---|
| HiFiV2-mix | 1% | 3.73 | 0.137 | 0.068 |
| HiFiV2-ACD-mix | 1% | 3.81 | 0.128 | 0.052 |
| iSTFT-mix | 1% | 3.82 | 0.121 | 0.049 |
| iSTFT-ACD-mix | 1% | 3.99 | 0.118 | 0.037 |

2. With different data augmentation (speaking rate change [Kharitonov+2021])

| Model | Data | UTMOS↑ | Periodicity↓ | cFW2VD↓ |
|---|---|---|---|---|
| HiFi-rate | 1% | 3.56 | 0.167 | 0.090 |
| HiFi-ACD-rate | 1% | 4.10 | 0.117 | 0.033 |

3. For different speaker (male speaker (9.1 min) in LibriTTS [Zen+2019])

| Model | Data | UTMOS↑ | Periodicity↓ | cFW2VD↓ |
|---|---|---|---|---|
| HiFi-mix | 9.1 min | 3.51 | 0.150 | 0.105 |
| HiFi-ACD-mix | 9.1 min | 3.66 | 0.140 | 0.074 |

## 6 Conclusions

- *AugCond*$D$ was proposed to train GAN vocoder with *limited data*
- Experimental results indicate *general utility* of *AugCond*$D$
- *Simplicity & versatility* of *AugCond*$D$ facilitates its application to various tasks