信号処理特論 (音楽音声信号処理特論) 2022 ゲスト講演

音声の合成と変換 その2

創造情報学専攻 猿渡研究室 特任助教
齋藤 佑樹

自己紹介

名前: 齋藤 佑樹 (SAITO Yuki) @ysato_human

🤦‍♀️ **斎**藤, 斉藤, 齊藤

略歴

2016/03: 釧路工業高等専門学校 専攻科 修了 (学士)

2018/03: 東大院 情報理工学系研究科 創造情報学専攻 修了 (修士)

2021/03: 東大院 情報理工学系研究科 システム情報学専攻 修了 (博士)

2021/04 ~ 2022/03: システム情報学専攻 特任助教

2022/04 ~ 現在: 創造情報学専攻 特任助教

研究分野

機械学習 x テキスト音声合成・変換 **(統計的音声合成)**

本講演の概要・目次

概要

統計的音声合成の基礎から, 深層学習に基づく最先端技術までを学ぶ.

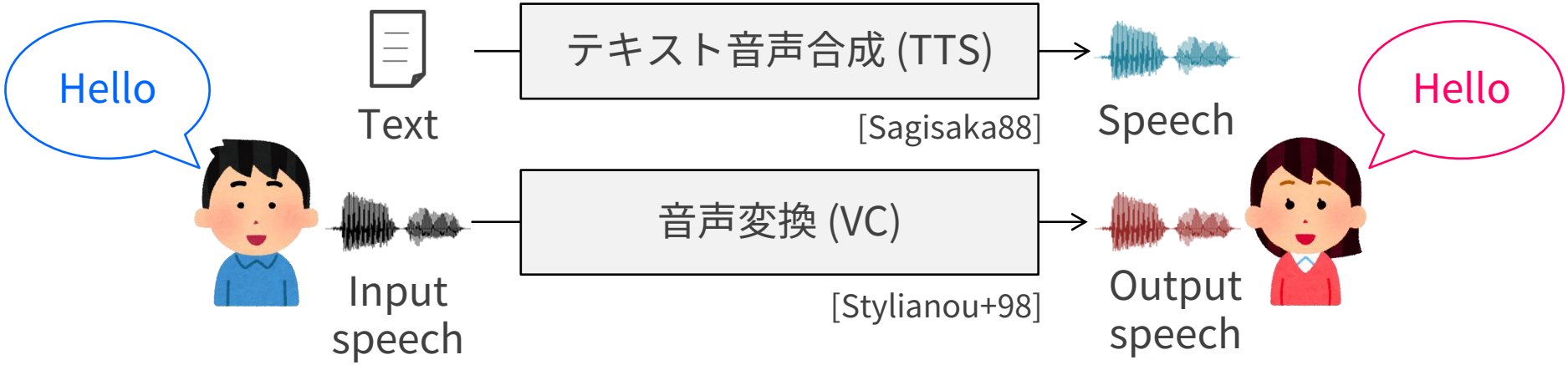
DNN 音声合成

目次

統計的音声合成とは?

音声合成 (speech synthesis)

コンピュータで人間の音声を人工的に合成・変換・加工する技術



統計的音声合成 (statistical speech synthesis) [Zen+09]

入力 → 音声 の対応関係を機械学習モデルで表現

HMM-TTS [Tokuda+13], GMM-VC [Toda+07], **DNN-TTS/VC** [Zen+13][Desai+09]

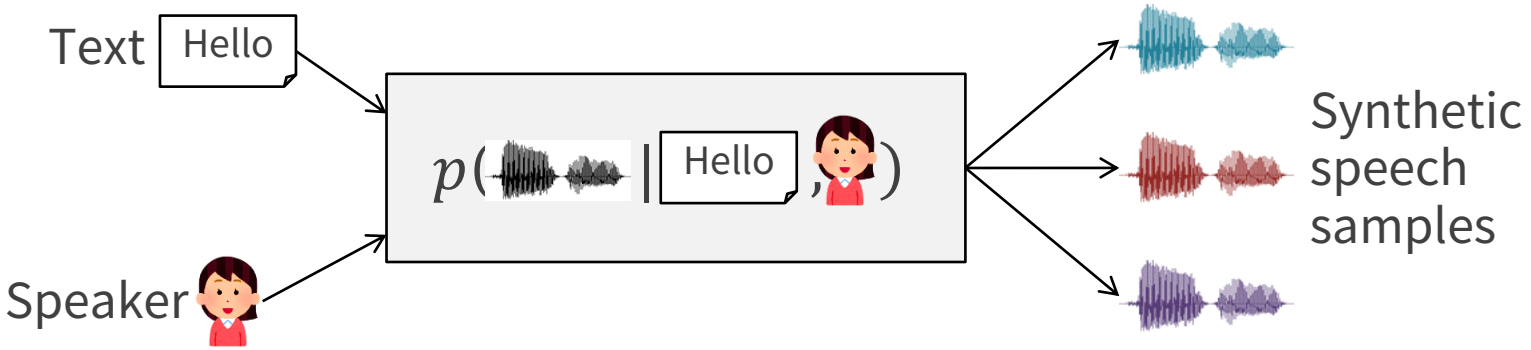
本講演のトピック: DNN ベースの統計的音声合成 (**DNN 音声合成**)

統計的音声合成研究の難しさ

統計的音声合成 = 一対多マッピングの学習

人間の発声は, 確率的にゆらぐ (その日の体調や気分など)

→ **条件付き確率モデルからのサンプリング**として解釈可能



得られた合成音声 (学習されたモデル) の評価

本質的に, 何かを作るタスクにおいて「正解」は存在しない!

c.f., 音声認識の目標 = 音声の発話内容を正確に認識すること

音声合成の究極の目標は?

ありとあらゆる音声を, 人間が望む品質で生成すること

評価基準: 音声の自然性, 話者の再現精度, etc...

Why machine learning?

比較的少ないパラメータ数で, 音声合成を効率的にモデル化可能

- c.f., 素片接続型音声合成 [Hunt+96]
  - 事前に録音された音声 (の断片) を切り貼りして音声を合成
  - 個人のあらゆる音声の録音は非現実的 (プライバシーの問題など)

ドメイン依存/非依存な音声の特徴を学習可能

- e.g., Multi-speaker TTS [Hojo+18][Jia+18][Mitsui+21]
  - 話者に非依存な発音情報と, 話者固有の特徴を同時に学習
  - 学習に用いられた既知話者だけでなく,
    未知話者の TTS も少量データで実現可能

異なる領域・分野の知見を容易に導入可能

- 音声情報処理の関連領域 (e.g., 音声認識・話者認識) だけでなく,
  自然言語処理, 画像生成などで有効な技術を用いた音声合成が可能に
  - **情報メディアを俯瞰して捉えるスキル**が問われる

本講演の概要・目次

概要

統計的音声合成の基礎から, 深層学習に基づく最先端技術までを学ぶ.

DNN 音声合成

目次

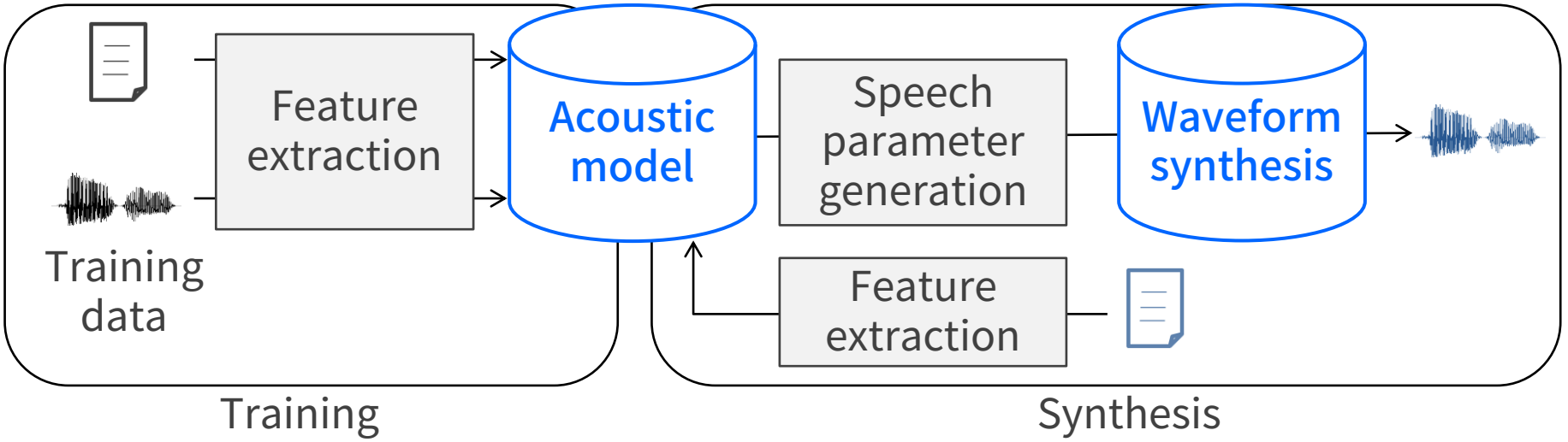
統計的音声合成の基本的な枠組み (TTS の場合)

学習時

1. 学習データ (テキスト/音声波形) から特徴量を抽出
2. テキスト特徴量から音声特徴量を生成する音響モデルを学習

生成時

1. 任意のテキストからテキスト特徴量を抽出
2. 学習後の音響モデルを用いて合成音声特徴量を生成
3. 合成音声特徴量から合成音声波形を生成



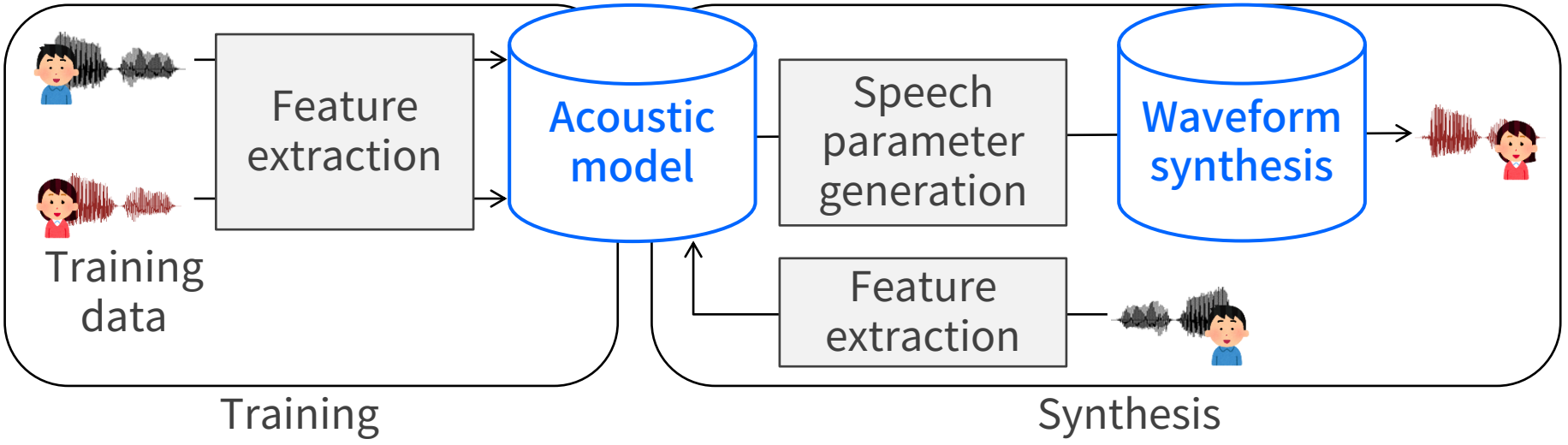
統計的音声合成の基本的な枠組み (VC の場合)

学習時

1. 学習データ (変換元/変換先話者の音声波形) から特徴量を抽出
2. 変換元話者の音声特徴量を変換する音響モデルを学習

生成時

1. 変換元話者の任意の音声から音声特徴量を抽出
2. 学習後の音響モデルを用いて合成音声特徴量を生成
3. 合成音声特徴量から合成音声波形を生成



統計的音声合成の定式化 (1/3)

Notation (青字が既知, 赤字が未知の情報)

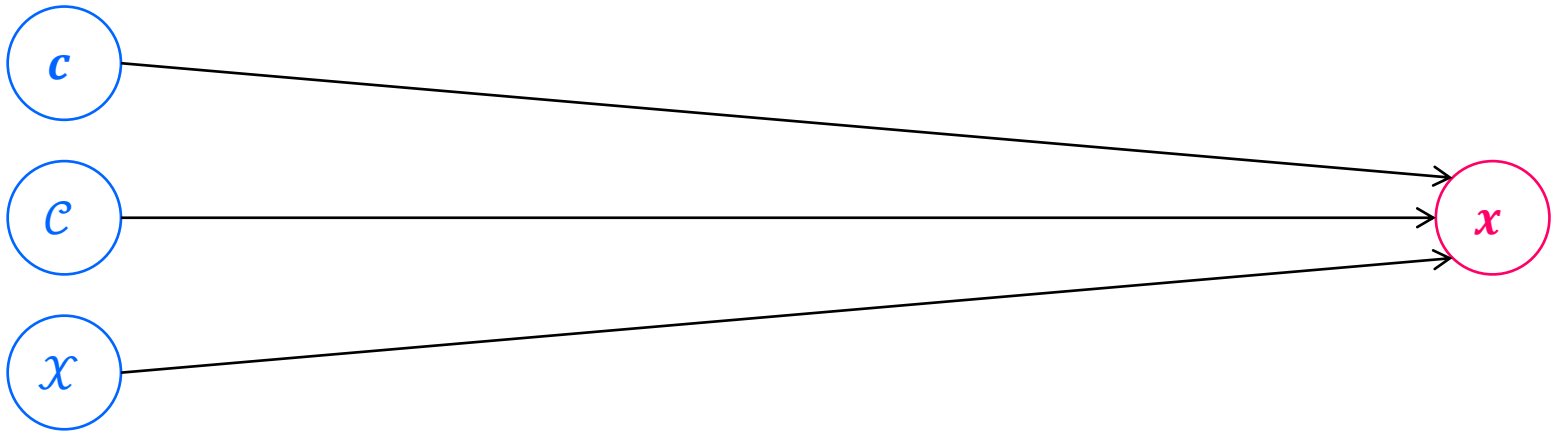
$\mathcal{X}$: 音声の学習データ, $\mathcal{C}$: 入力情報の学習データ

$\mathcal{C}$ の実体 = テキスト for TTS, 変換元話者の音声 for VC

$\boldsymbol{x}$: 合成したい音声, $\boldsymbol{c}$: $\boldsymbol{x}$ を合成するための入力情報

Formulation

学習データ $(\mathcal{X}, \mathcal{C})$ が与えられたもとで,
任意の入力 $\boldsymbol{c}$ から音声 $\boldsymbol{x}$ を生成する**予測分布** $p(\boldsymbol{x}|\boldsymbol{c}, \mathcal{C}, \mathcal{X})$ を求めたい



統計的音声合成の定式化 (2/3)

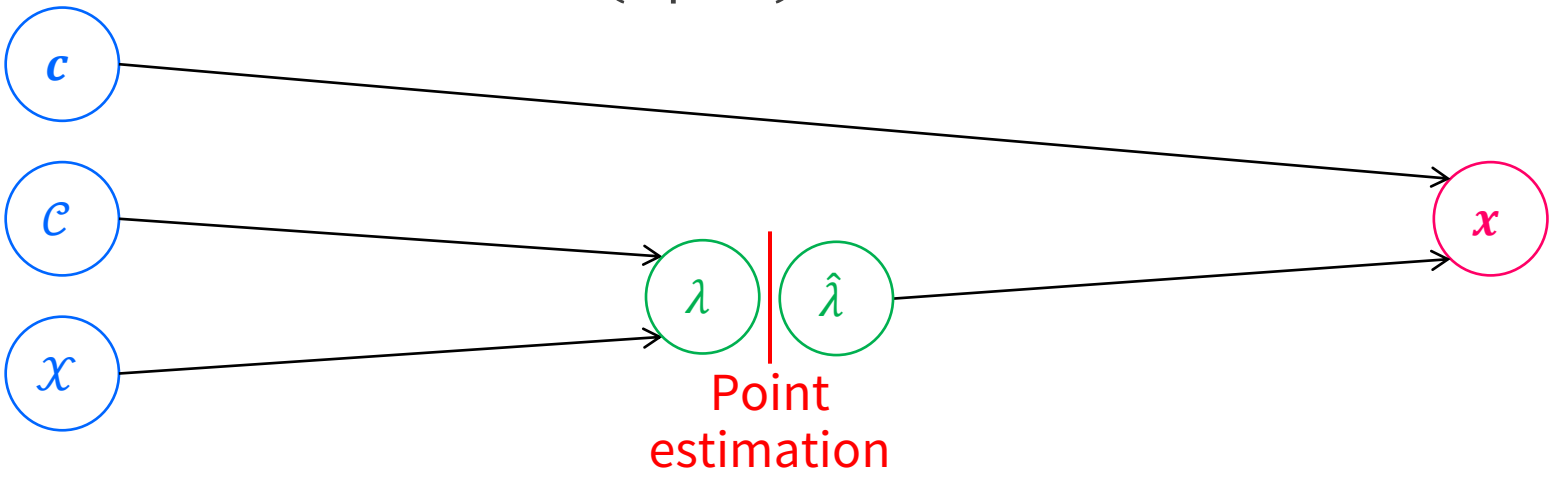
Approximation1: 統計モデル λ の導入

予測分布 $p(\boldsymbol{x}|\boldsymbol{c}, \mathcal{C}, \mathcal{X})$ を直接求めるのは困難なので,
学習データを用いて $\boldsymbol{c}$ と $\boldsymbol{x}$ の対応関係を表す**統計モデル** λ を学習

$$p(\boldsymbol{x}|\boldsymbol{c}, \mathcal{C}, \mathcal{X}) = \int p(\lambda|\mathcal{C}, \mathcal{X}) p(\boldsymbol{x}|\boldsymbol{c}, \lambda) d\lambda$$

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, p(\lambda|\mathcal{C}, \mathcal{X}) \quad \text{(Training)}$$

$$\boldsymbol{x} \sim p\left(\boldsymbol{x}\middle|\boldsymbol{c}, \hat{\lambda}\right) \quad \text{(Synthesis)}$$



統計的音声合成の定式化 (3/3)

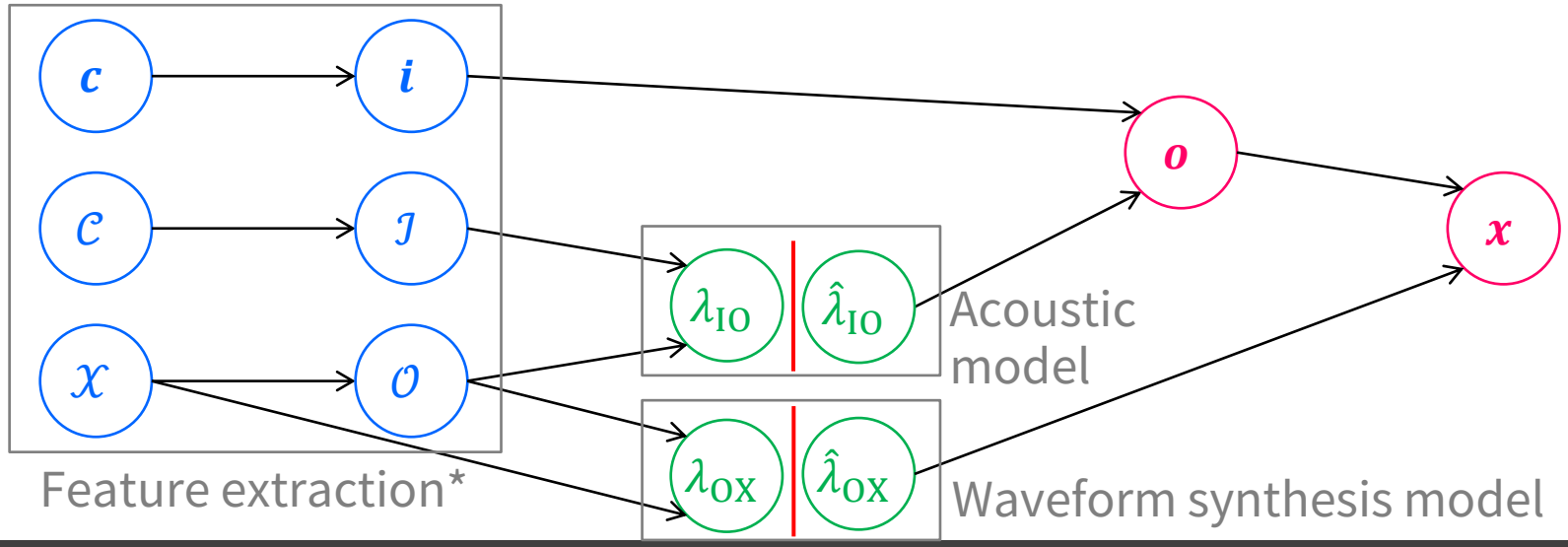
Approximation2: 中間特徴量の導入

$\boldsymbol{c}$ と $\boldsymbol{x}$ の対応関係を直接学習するのは困難なので, $\mathcal{C}$ と $\mathcal{X}$ からそれぞれ**入力特徴量** $\mathcal{I}$ と**音声特徴量** $\mathcal{O}$ を抽出し, 特徴量間の対応関係を学習

$\mathcal{I}$ の実体 = テキスト特徴量 for TTS, 変換元話者の音声特徴量 for VC

$$\hat{\lambda}_{\mathrm{IO}} = \underset{\lambda_{\mathrm{IO}}}{\operatorname{argmax}}\, p(\lambda_{\mathrm{IO}}|\mathcal{I},\mathcal{O}), \hat{\lambda}_{\mathrm{OX}} = \underset{\lambda_{\mathrm{OX}}}{\operatorname{argmax}}\, p(\lambda_{\mathrm{OX}}|\mathcal{O},\mathcal{X}),$$

$$\boldsymbol{o} \sim p\left(\boldsymbol{o}\middle|\boldsymbol{i},\hat{\lambda}_{\mathrm{IO}}\right), \boldsymbol{x} \sim p\left(\boldsymbol{x}\middle|\boldsymbol{o},\hat{\lambda}_{\mathrm{OX}}\right)$$



*特徴量の抽出に統計モデルを用いる場合もあるが, 本講演では省略

統計的音声合成の主要な構成要素

1. 学習データ (音声コーパス) $\mathcal{C}, \mathcal{X}$
2. 特徴量 $\mathcal{I}, \mathcal{O}$
3. 音響モデル λ_{IO}
4. 波形生成モデル λ_{OX}

学習データ (音声コーパス) の用意

音声コーパス: 音声とテキスト書き起こしの集合

- 種々のアノテーションが付随する場合も
  - 音素アライメント情報: いつ, 何を話しているか
  - 話者ラベル: 誰が話しているか
  - 感情ラベル: どんな感情で話しているか

英語/日本語の主要な音声コーパス

名前	言語	話者数	ドメイン
LJSpeech [Ito+17]	英	1	オーディオブック
LibriTTS [Zen+19]	英	2,456	オーディオブック
VCTK [Veaux+12]	英	110	テキスト読み上げ
JSUT [Sonobe+17]	日	1	テキスト読み上げ
JVS [Takamichi+19]	日	100	テキスト読み上げ

猿渡・小山研でこれまでに構築した主要な音声コーパス一覧 → link

テキスト特徴量抽出のためのテキスト解析

テキスト = 言語に依存した文字の系列データ

- e.g.,「私の夢は、年収、5000000000000000＄稼ぐことです。」
- プレーンテキストからの TTS は (日本語だと特に) 困難
  - 文字・読みの多様性 (ひらがな, カタカナ, 漢字, 英数字, 句読点, etc.)
    - 同形異音語が多い (e.g.,「日本」→ にほん/にっぽん)

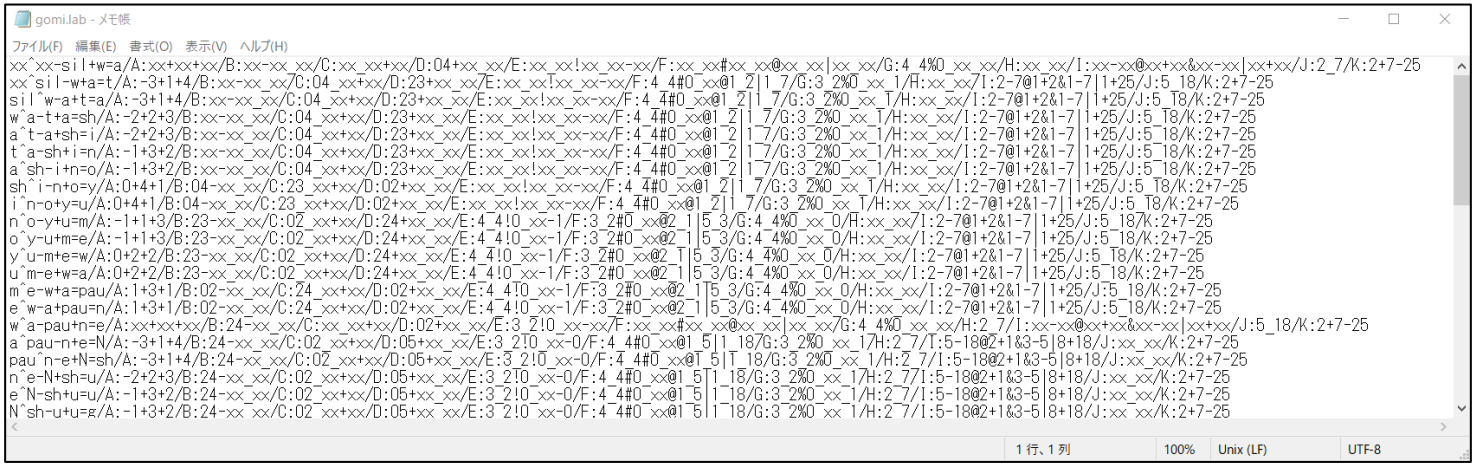
TTS におけるテキスト解析

- テキスト正規化: 読みやアクセントを推定しやすくする
  - e.g.,「 5000000000000000＄ 」→「五千兆ドル」
- 形態素解析: 単語への分割 & 品詞・読みを推定する
  - e.g.,「私」... わたし/名詞,「の」... の/助詞,「夢」...ゆめ/名詞
- 音素変換: 読み情報を機械学習で扱いやすい形式に変換する
  - e.g.,「私」... "w a t a sh i",「夢」... "y u m e"
- アクセント推定: テキストのアクセントを推定する
  - e.g.,「私」... わたし,「夢」... ゆめ

TTS で用いられる主なテキスト特徴量

フルコンテキストラベル

音素・アクセントなどの情報をひとまとめにしたラベル*



音素・韻律記号列

フルコンテキストラベルから, 読みとアクセントに関わる情報だけを簡略化して表記

直列型入力 [Kurihara+21]: "^wa[tashino#yu[me]wa..."

並列型入力 [藤井+21]: "watashinoyumewa..."

"^[_____#_[____]__..."

*厳密には, フルコンテキストラベルに対して質問した結果を数値化したものを特徴量として使用

TTS における音素継続長モデリング

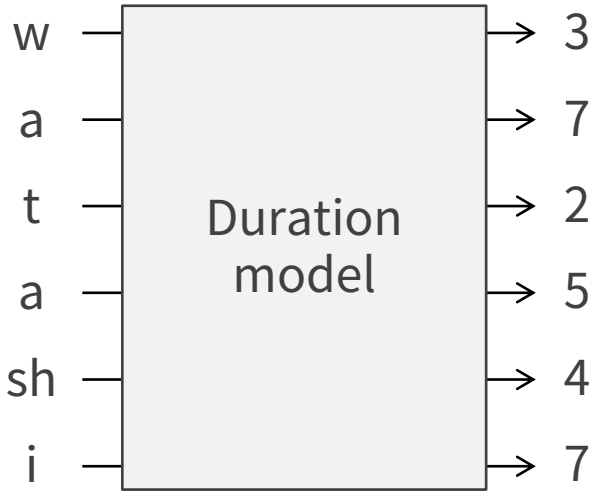
音素継続長: 当該音素がどれだけ連続するか (= 話す速さ・リズム)

プレーンテキストだけでは, 話す速さはわからない

→ 統計モデルによる予測が必要

手法1: 音素アライメント情報を用いた教師あり学習

音素単位のテキスト特徴量から, その継続長 (フレーム数) を予測



手法2: 音素アライメント情報なしの End-to-End 学習

TTS の音響モデルとアライメント情報を同時に学習 (詳細は後述)

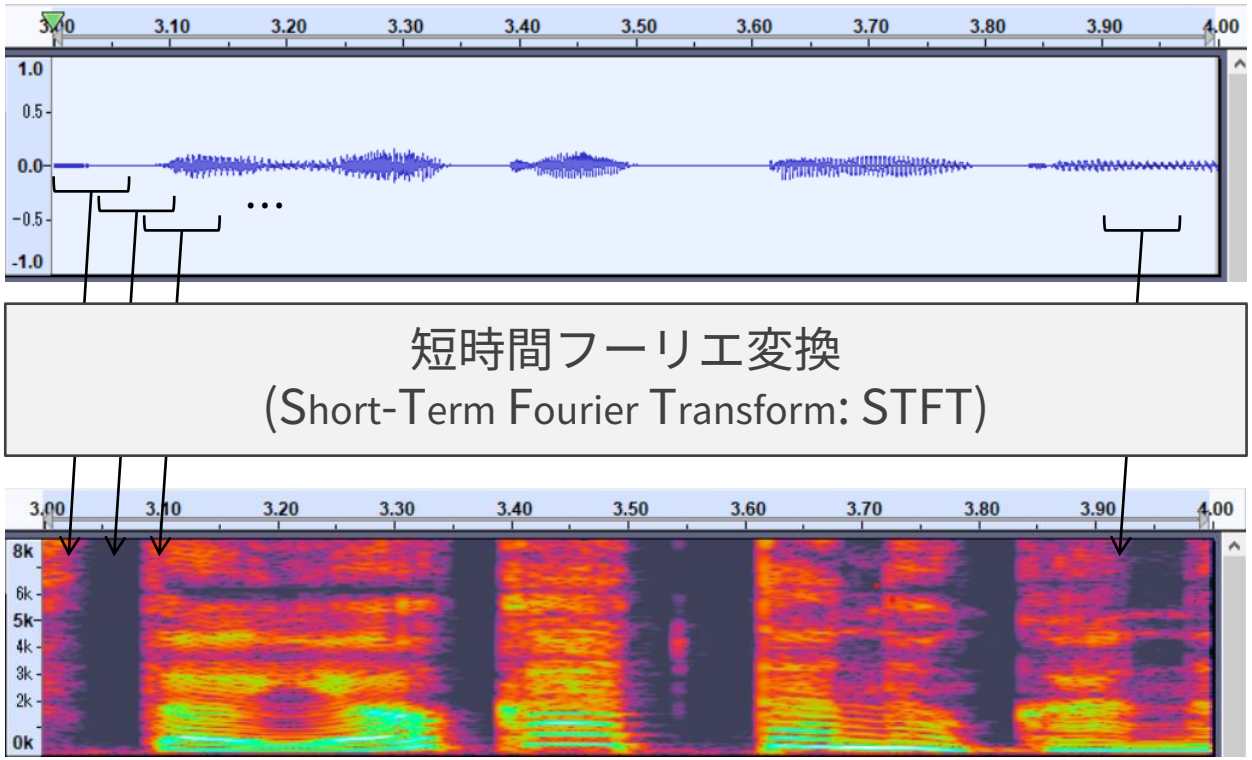
音声信号の時間・周波数表現

音声波形 = 超高密度の時系列データ

サンプリング周波数 16 kHz の音声 → **1秒間に 16,000 サンプル**

短時間フーリエ変換により, 時間・周波数表現へ変換

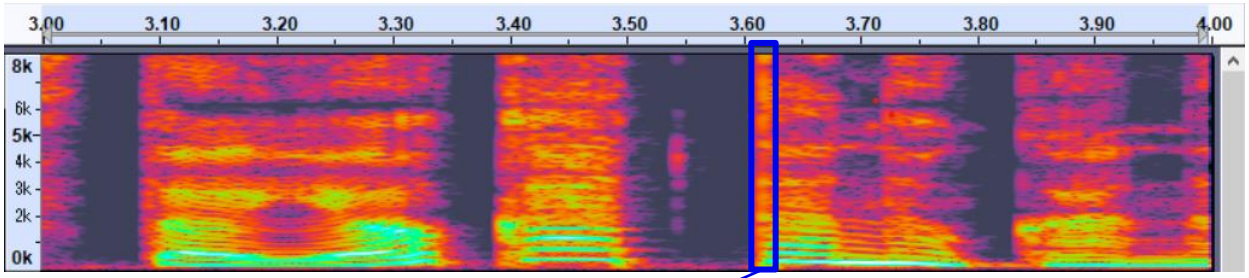
(スペクトログラム)

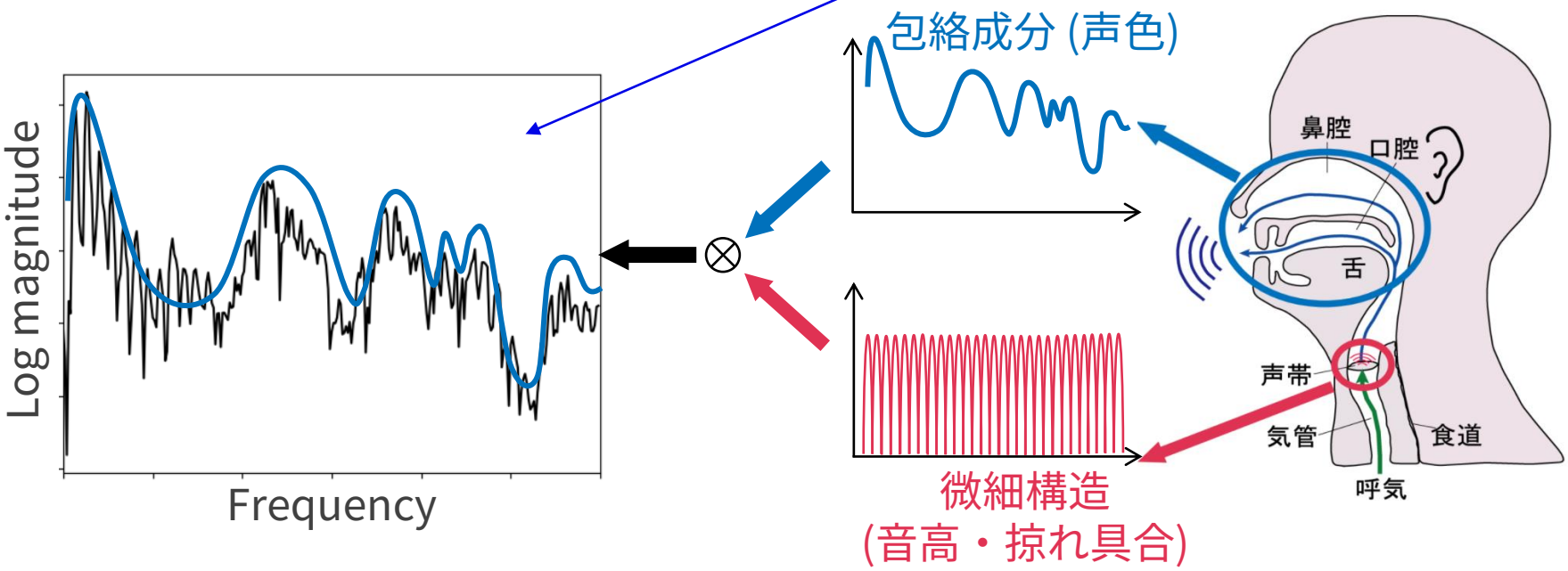


ソース・フィルタモデル: 人間による音声生成過程を近似

人間の音声 = 声帯振動の特性と声道フィルタの特性の畳み込み

前者は周波数領域での微細構造を, 後者は包絡成分を表現

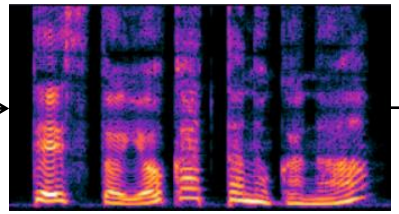


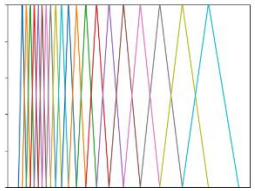


TTS/VC で用いられる主な音声特徴量

STFT analysis

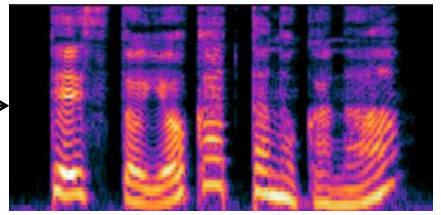
スペクトログラム
(高次元)



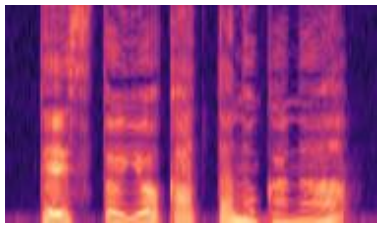


Mel-filter bank

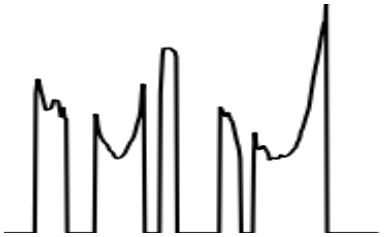
メルスペクトログラム
(低次元)



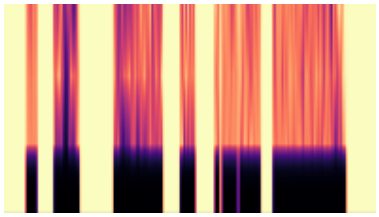
Vocoder analysis
(e.g., WORLD)
[Morise+16]



スペクトル包絡 (声色)



F0 (音高)

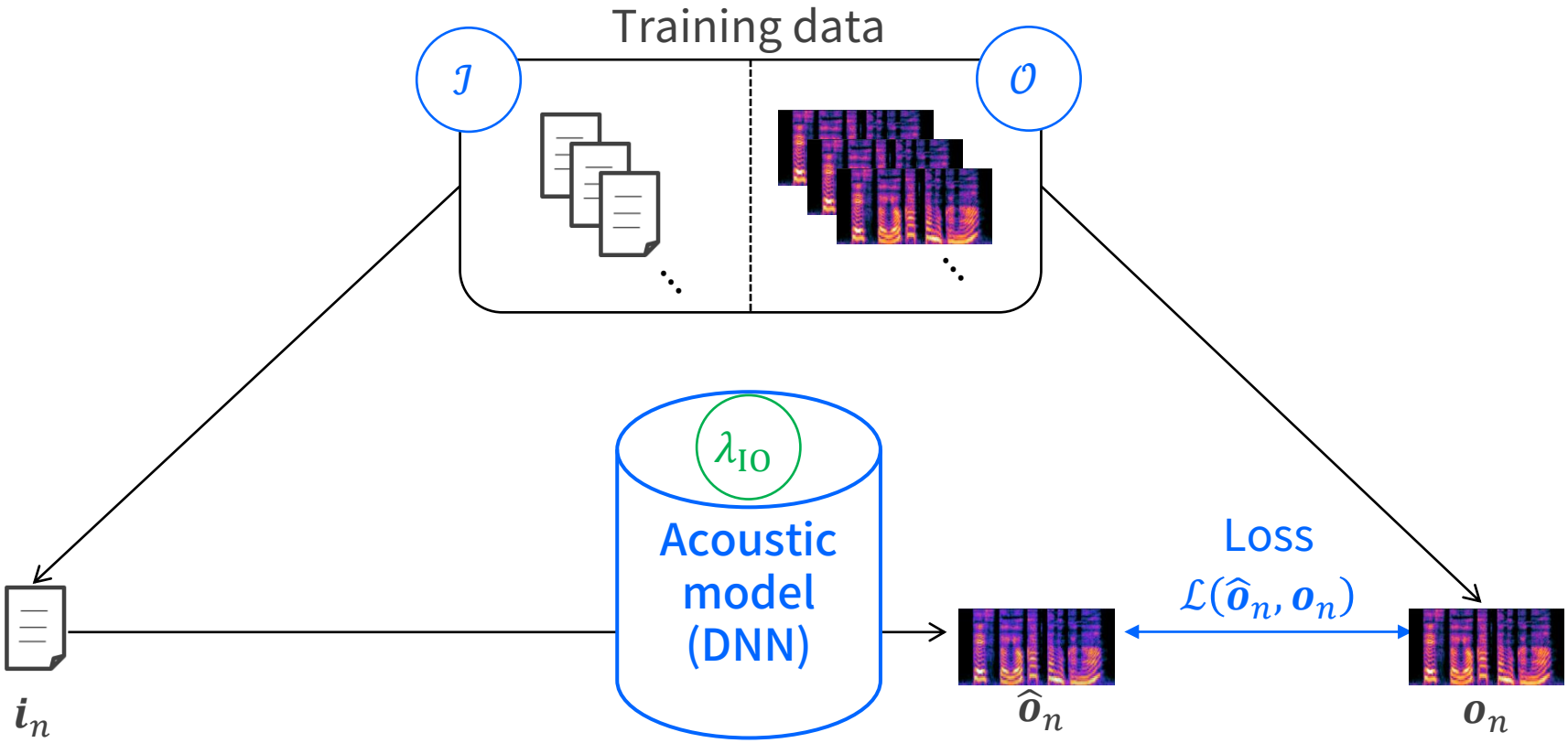


非周期性指標 (掠れ具合)

声帯振動特性と
声道フィルタ特性の
分離

TTS の音響モデル学習

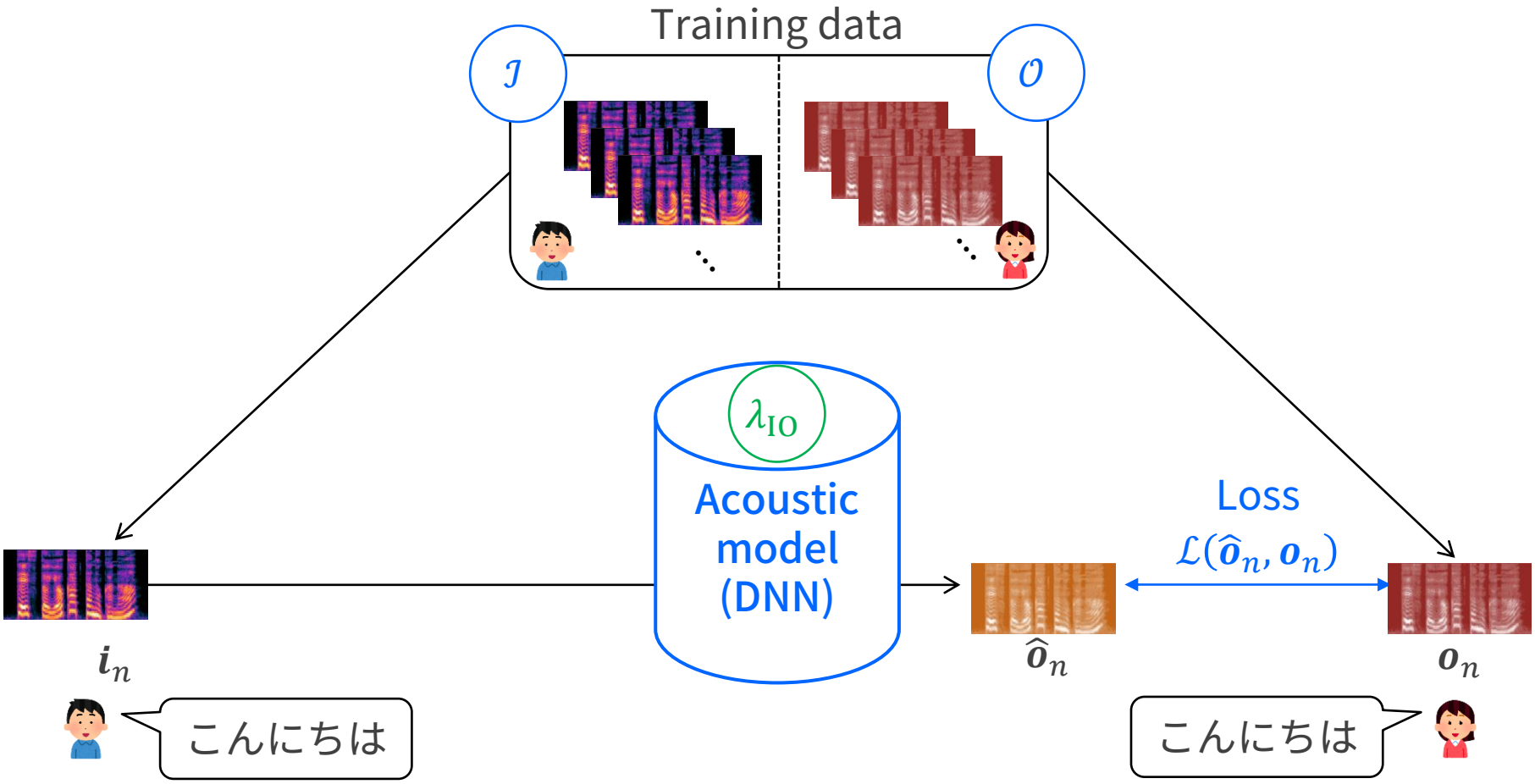
音響モデルの予測結果 $\hat{o}_n$ と学習データ o_n の誤差最小化



VC の音響モデル学習

音響モデルの予測結果 $\hat{o}_n$ と学習データ o_n の誤差最小化

同一発話内容の音声を用いた学習 = **パラレル VC**



Many many DNN-based acoustic models so far...

Table 2: A list of acoustic models and their corresponding characteristics. "Ling" stands for linguistic features, "Ch" stands for character, "Ph" stands for phoneme, "MCC" stands for mel-cepstral coefficients [82], "MGC" stands for mel-generalized coefficients [355], "BAP" stands for band aperiodicities [156, 157], "LSP" stands for line spectral pairs [135], "LinS" stands for linear-spectrograms, and "MelS" stands for mel-spectrograms. "NAR*" means the model uses autoregressive structures upon non-autoregressive structures and is not fully parallel.

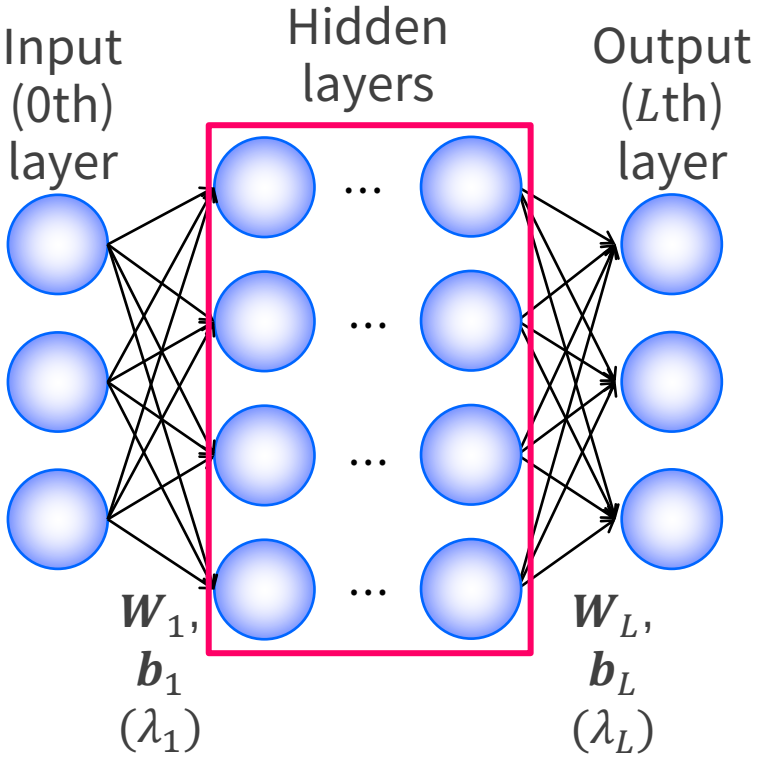
Acoustic Model	Input→Output	AR/NAR	Modeling	Structure
HMM-based [416, 356]	Ling→MCC+F0	/	/	HMM
DNN-based [426]	Ling→MCC+BAP+F0	NAR	/	DNN
LSTM-based [78]	Ling→LSP+F0	AR	/	RNN
EMPHASIS [191]	Ling→LinS+CAP+F0	AR	/	Hybrid
ARST [375]	Ph→LSP+BAP+F0	AR	Seq2Seq	RNN
VoiceLoop [333]	Ph→MGC+BAP+F0	AR	/	hybrid
Tacotron [382]	Ch→LinS	AR	Seq2Seq	Hybrid/RNN
Tacotron 2 [303]	Ch→MelS	AR	Seq2Seq	RNN
DurIAN [418]	Ph→MelS	AR	Seq2Seq	RNN
Non-Att Tacotron [304]	Ph→MelS	AR	/	Hybrid/CNN/RNN
Para. Tacotron 1/2 [74, 75]	Ph→MelS	NAR	/	Hybrid/Self-Att/CNN
MelNet [367]	Ch→MelS	AR	/	RNN
DeepVoice [8]	Ch/Ph→MelS	AR	/	CNN
DeepVoice 2 [87]	Ch/Ph→MelS	AR	/	CNN
DeepVoice 3 [270]	Ch/Ph→MelS	AR	Seq2Seq	CNN
ParaNet [268]	Ph→MelS	NAR	Seq2Seq	CNN
DCTTS [332]	Ch→MelS	AR	Seq2Seq	CNN
SpeedySpeech [361]	Ph→MelS	NAR	/	CNN
TalkNet 1/2 [19, 18]	Ch→MelS	NAR	/	CNN
TransformerTTS [192]	Ph→MelS	AR	Seq2Seq	Self-Att
MultiSpeech [39]	Ph→MelS	AR	Seq2Seq	Self-Att
FastSpeech 1/2 [290, 292]	Ph→MelS	NAR	Seq2Seq	Self-Att
AlignTTS [429]	Ch/Ph→MelS	NAR	Seq2Seq	Self-Att
JDI-T [197]	Ph→MelS	NAR	Seq2Seq	Self-Att
FastPitch [181]	Ph→MelS	NAR	Seq2Seq	Self-Att
AdaSpeech 1/2/3 [40, 403, 404]	Ph→MelS	NAR	Seq2Seq	Self-Att
DenoiSpeech [434]	Ph→MelS	NAR	Seq2Seq	Self-Att
DeviceTTS [126]	Ph→MelS	NAR	/	Hybrid/DNN/RNN
LightSpeech [220]	Ph→MelS	NAR	/	Hybrid/Self-Att/CNN
Flow-TTS [234]	Ch/Ph→MelS	NAR*	Flow	Hybrid/CNN/RNN
Glow-TTS [159]	Ph→MelS	NAR	Flow	Hybrid/Self-Att/CNN
Flowtron [366]	Ph→MelS	AR	Flow	Hybrid/RNN
EfficientTTS [235]	Ch→MelS	NAR	Flow	Hybrid/CNN
GMVAE-Tacotron [119]	Ph→MelS	AR	VAE	Hybrid/RNN
VAE-TTS [443]	Ph→MelS	AR	VAE	Hybrid/RNN
BVAE-TTS [187]	Ph→MelS	NAR	VAE	CNN
GAN exposure [99]	Ph→MelS	AR	GAN	Hybrid/RNN
TTS-Stylization [224]	Ch→MelS	AR	GAN	Hybrid/RNN
Multi-SpectroGAN [186]	Ph→MelS	NAR	GAN	Hybrid/Self-Att/CNN
Diff-TTS [141]	Ph→MelS	NAR*	Diffusion	Hybrid/CNN
Grad-TTS [276]	Ph→MelS	NAR	Diffusion	Hybrid/Self-Att/CNN
PriorGrad [185]	Ph→MelS	NAR	Diffusion	Hybrid/Self-Att/CNN

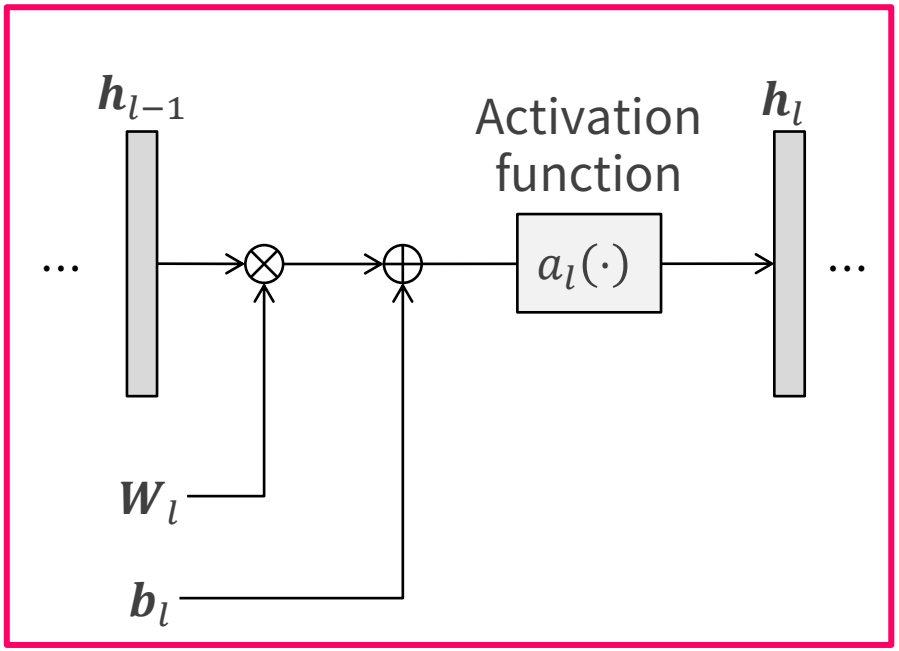
[Tan+21]

Deep Neural Network (DNN)

DNN の基本構造 = 線形変換と非線形写像の繰り返し

時系列モデリングに適した 再帰型 NN や畳み込み NN も用いられる





モデルパラメータ: $\lambda = \{\lambda_1, \cdots, \lambda_L\}$

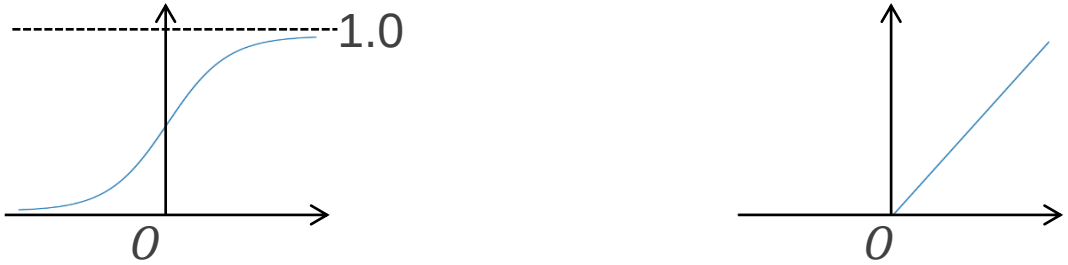
(ネットワークの結合重み・バイアスの集合)

DNN の活性化関数と学習時の誤差関数

活性化関数: 隠れ層・出力層で異なる役割

隠れ層: 非線形変換 (表現力の向上)

Sigmoid: $a(x) = 1/1 + e^{-x}$, ReLU: $a(x) = 0 \text{ if } x < 0, \text{otherwise } x$



出力層: 誤差関数を計算するドメインへの写像

Linear: $a(x) = x$, Softmax: $a(\boldsymbol{x}) = e^{\boldsymbol{x}} / \sum_m e^{x_m}$ (確率ベクトル)

誤差関数: 解きたいタスクに応じて設定

生成タスク

Mean Squared Error (MSE): $\|\hat{\boldsymbol{o}} - \boldsymbol{o}\|_2^2$ or Mean Absolute Error (MAE): $\|\hat{\boldsymbol{o}} - \boldsymbol{o}\|_1^1$

識別タスク

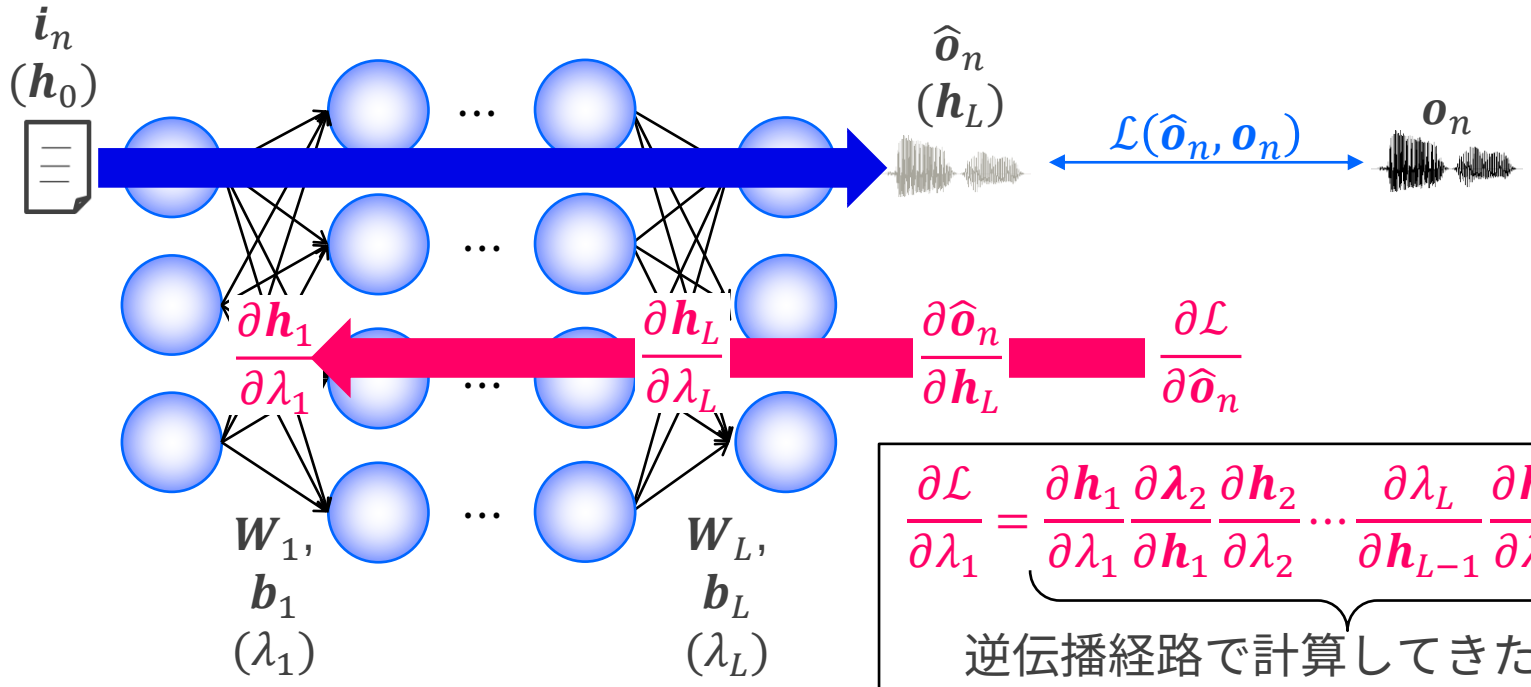
Cross-entropy: $-\sum_c o_c \log \hat{o}_c$ ($\boldsymbol{o}$ はクラスIDを示す one-hot ベクトル)

DNN の学習: 誤差逆伝播法を用いた勾配法による モデルパラメータ更新

順伝播: 入力 $\boldsymbol{i}_n$ から $\hat{\boldsymbol{o}}_n$ を出力し, 誤差関数 $\mathcal{L}(\hat{\boldsymbol{o}}_n, \boldsymbol{o}_n)$ を計算

逆伝播: 連鎖律に基づき, 各モデルパラメータの勾配 $\partial L / \partial \lambda_l$ を計算

→ 活性化関数を含む, 各層での計算が**すべて微分可能**である必要あり

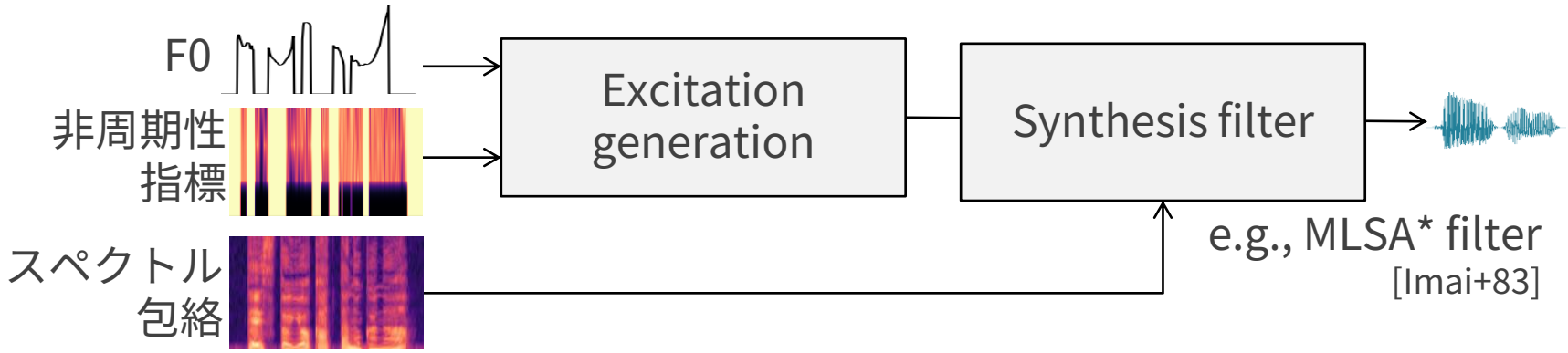


$$\frac{\partial \mathcal{L}}{\partial \lambda_1} = \underbrace{\frac{\partial \boldsymbol{h}_1}{\partial \lambda_1}\frac{\partial \boldsymbol{\lambda}_2}{\partial \boldsymbol{h}_1}\frac{\partial \boldsymbol{h}_2}{\partial \lambda_2}\cdots\frac{\partial \lambda_L}{\partial \boldsymbol{h}_{L-1}}\frac{\partial \boldsymbol{h}_L}{\partial \lambda_L}\frac{\partial \mathcal{L}}{\partial \boldsymbol{h}_L}}_{\text{逆伝播経路で計算してきた勾配の積}}$$

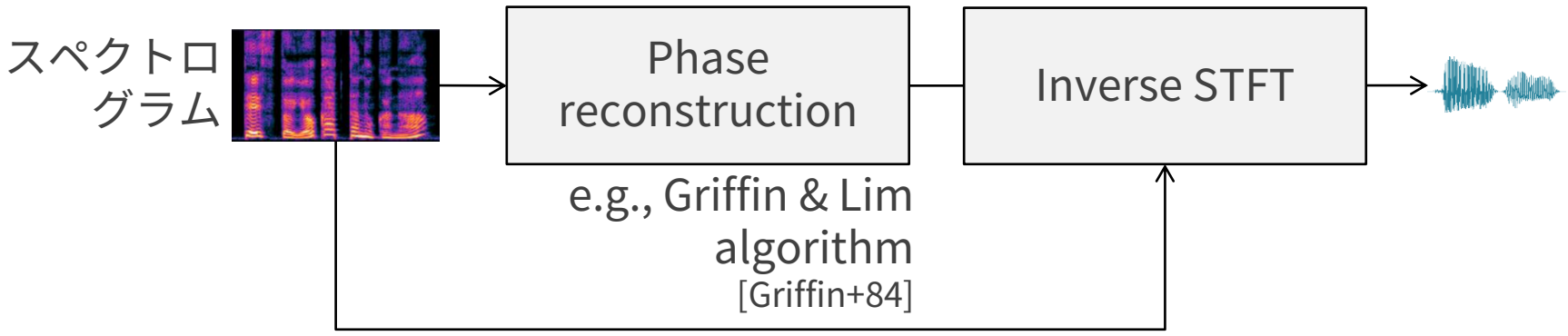
更新: $\lambda_l \leftarrow \lambda_l - \eta \dfrac{\partial \mathcal{L}}{\partial \lambda_l}$ (η は学習率)

音声波形生成

ボコーダパラメータからの生成



スペクトログラムからの生成

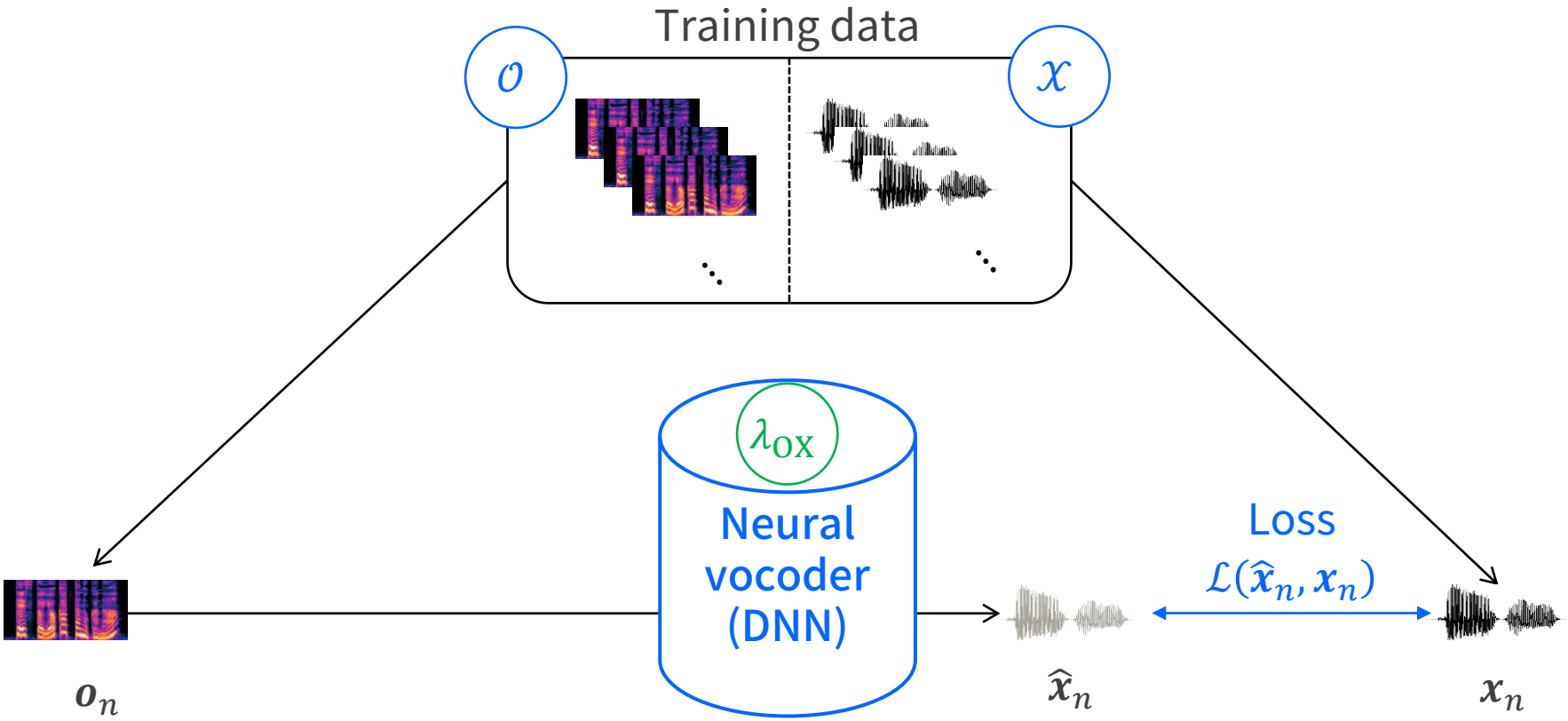


DNN を用いた生成 = ニューラルボコーダ (次スライド)

*MLSA: Mel-Log Spectrum Approximation

ニューラルボコーダの学習

ニューラルボコーダの予測結果 $\hat{x}_n$ と学習データ x_n の誤差最小化



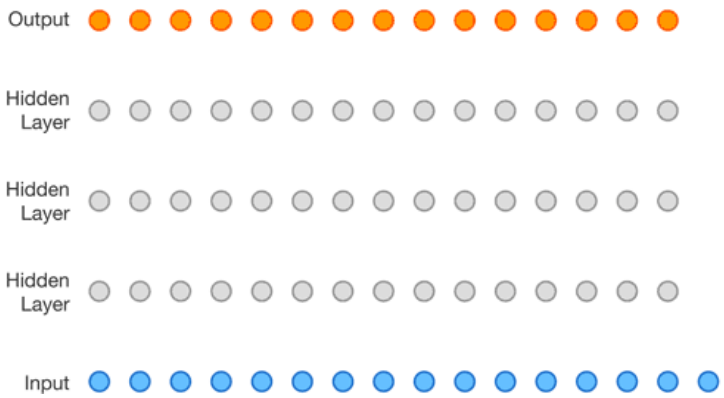
WaveNet [Oord+16]

DNN による波形生成モデリングの先駆け的研究

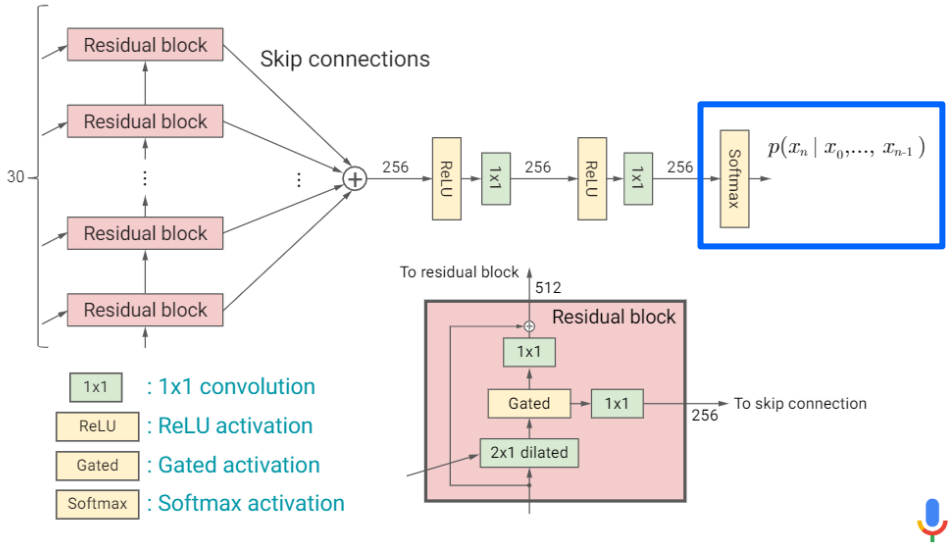
Causal dilated convolution による超長期の依存関係モデリング

Residual block による強力な非線形変換

256段階に量子化された音声信号の予測を**識別問題として定式化**



https://deepmind.com/blog/article/wavenet-generative-model-raw-audio



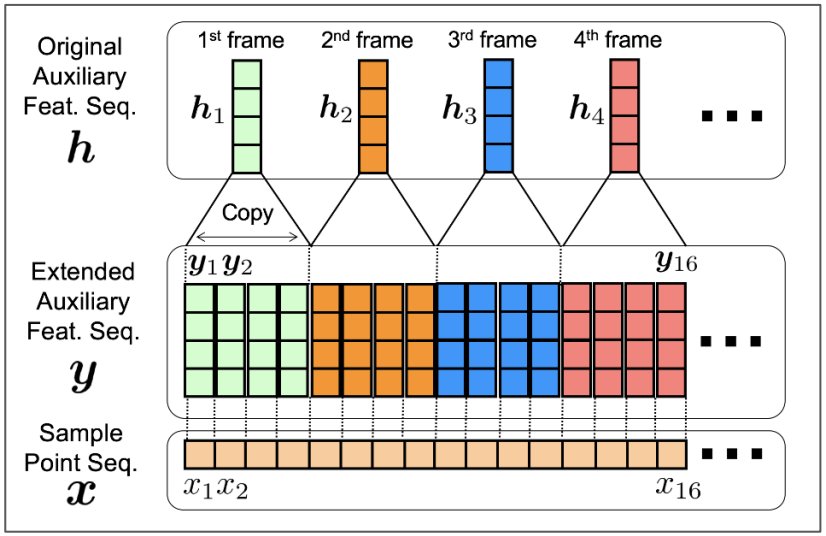
https://static.googleusercontent.com/media/research.google.com/ja//pubs/archive/45882.pdf

$$p(\boldsymbol{x}|\lambda) = \prod_{t=1}^{T} p(x_t|x_1, x_2, \cdots, x_{t-1})$$: Auto-Regressive (AR) modeling

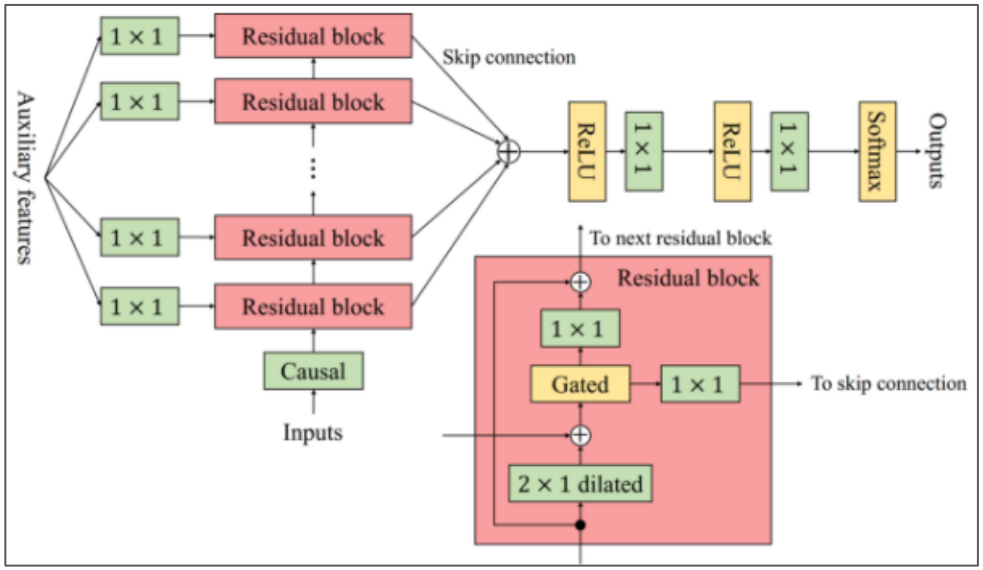
WaveNet ボコーダ [Tamamori+17]

WaveNet が表す予測分布を音声特徴量で条件付け

音声特徴量と音声波形の長さを揃えるために,
対応するサンプル数だけコピー



[Tamamori+17]



[Hayashi+17]

$$p(\boldsymbol{x}|\boldsymbol{o},\lambda) = \prod_{t=1}^{T} p(x_t|x_1, x_2, \cdots, x_{t-1}, \boldsymbol{o})$$: Conditional AR modeling

Many many neural vocoders so far...

Table 3 Small size vocoder

Vocoder	Neural network types	Characteristics
WaveNet (Oord et al., 2016)	Dilated causal gated CNN	Based on dilated CNN, the training and inference speed is slow
SampleRNN (Mehri et al., 2016)	RNN	Multi-scale RNN structure, training and inference speed is faster than Wavenet
FftNet (Jin et al., 2018)	1×1 CNN	Based on 1×1 convolution, the model structure is simple, and the training and inference speed is fast
WaveRNN (Kalchbrenner et al., 2018)	GRU	Based on single layer of GRU, the model structure is simple, and the training and inference speed is fast
Multi-Band WaveRNN (Yu et al., 2019)	GRU	Parallel generation of multiple bands, the training and inference speed is fast
LPCNet (Valin and Skoglund, 2019)	GRU	The linear prediction (LP) technology is used, the model structure is simple, and the training and inference speed is fast

Table 5 Methods of GAN-based vocoder to improve the naturalness of generated speech

Vocoder	Characteristics
MelGAN (Kumar et al., 2019)	Using multi-scale discriminant structure and feature matching loss
Parallel WaveGAN (Yamamoto et al., 2020)	Using multi-resolution STFT loss
VocGAN (Yang et al., 2020)	Using multi-resolution STFT loss, feature matching loss, multi-scale waveform generator, and JCU loss
HiFi-GAN (Kong et al., 2020)	Using multi-scale discrimination, multi-period discrimination, and MRF
Multi-Band MelGAN (Yang et al., 2021)	Using multi-resolution STFT loss

Table 4 Non-autoregressive vocoder

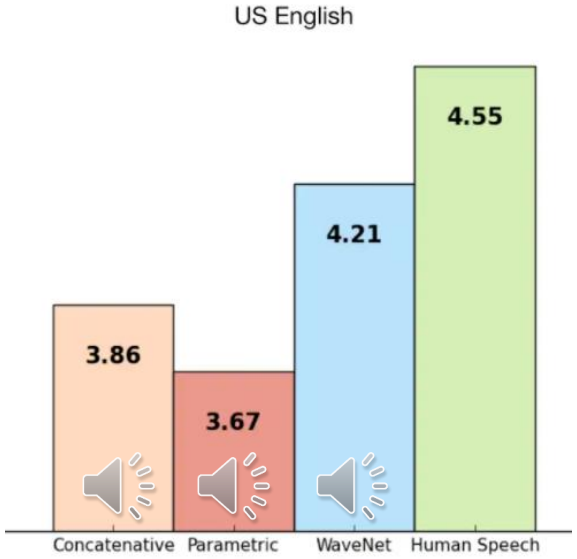
Vocoder	Neural network types	Generative model types	Characteristics
WaveNet (Oord et al., 2016)	Dilated causal gated convolution	Autoregression	Autoregressive generation, slow training and inference speed
Parallel WaveNet (Oord et al., 2018)	Dilated causal gated convolution	IAF	Based on knowledge distillation, training and inference speed is fast, Monte Carlo sampling is required to estimate KL divergence, the training process is unstable
FloWaveNet (Kim et al., 2018)	Dilated convolution	Normalizing flow	The inference speed is fast, the training convergence speed is slow, the model contains many parameters
ClariNet (Ping et al., 2018)	Dilated causal gated convolution	IAF	Based on knowledge distillation, the training and inference speed is fast, the training process is stable
WaveGlow (Prenger et al., 2019)	Non-causal dilated convolution, 1×1 convolution	Normalizing flow	The inference speed is fast, the training convergence speed is slow, the model contains many parameters
MelGAN (Kumar et al., 2019)	Dilated convolution, transposed convolution, grouped convolution	GAN	The inference speed is fast, the training convergence speed is slow
GAN-TTS (Bińkowski et al., 2019)	Dilated convolution	GAN	The training and inference speed is fast, no need for mel-spectrogram as input
Parallel WaveGAN (Yamamoto et al., 2020)	Non-causal dilated convolution	GAN	The inference speed is fast, the training convergence speed is slow, the model contains many parameters
WaveVAE (Peng et al., 2020)	Dilated causal gated convolution	IAF, VAE	The training and inference speed is fast
WaveFlow (Ping et al., 2020)	2D-dilated convolution	Autoregression	Combining the advantages of autoregressive flow and non-autoregressive flow, the training and inference speed is fast
WaveGrad (Chen et al., 2020)	Dilated convolution	Diffusion probability model	The inference speed is fast, the training convergence speed is slow
DiffWave (Kong et al., 2020)	Bidirectional dilated convolution	Diffusion probability model	The inference speed is fast, the training convergence speed is slow
Multi-Band MelGAN (Yang et al., 2021)	Dilated convolution, transposed convolution, grouped convolution	GAN	The training and inference speed is fast

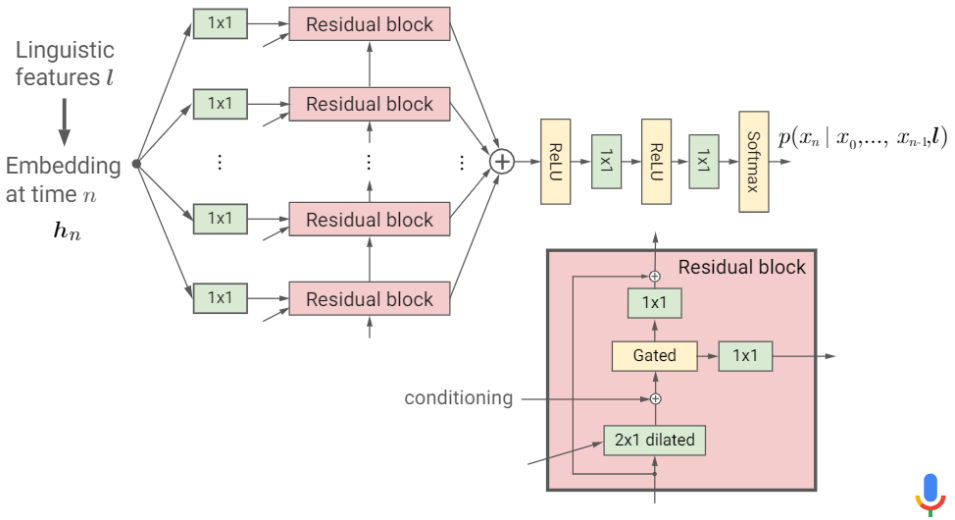
[Mu+21]

一貫学習に基づく統計的音声合成 (1/3)

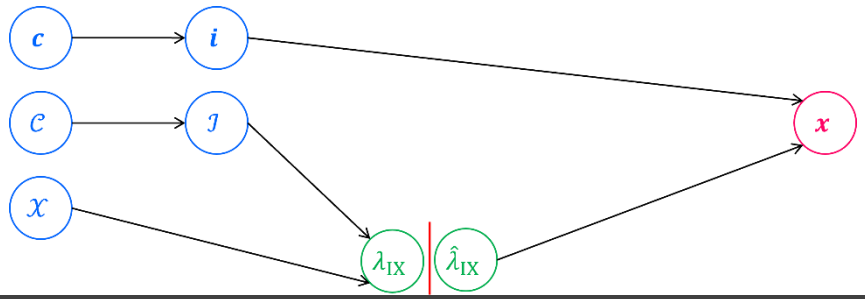
音響特徴量予測と音声波形生成の統合

e.g., テキスト特徴量で条件付けされた WaveNet [Oord+16]





https://static.googleusercontent.com/media/research.google.com/ja//pubs/archive/45882.pdf

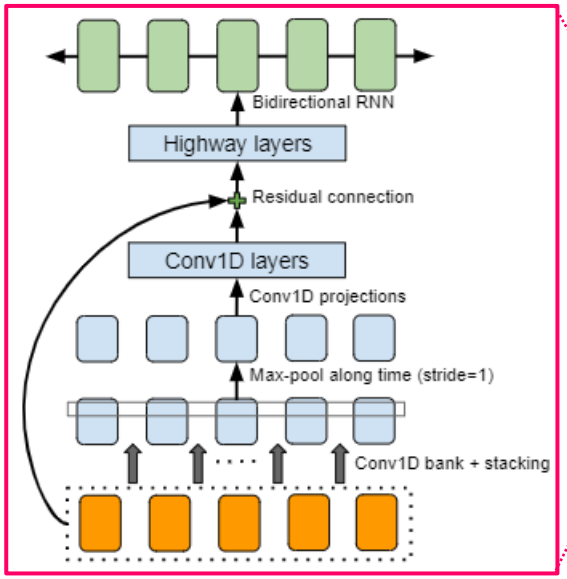


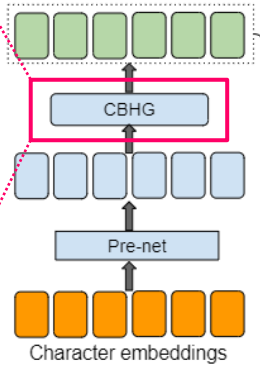
音声サンプルは ここ より引用

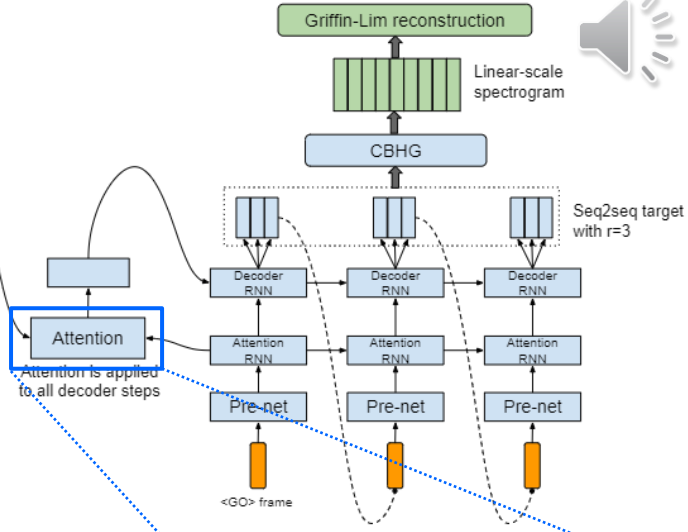
一貫学習に基づく統計的音声合成 (2/3)

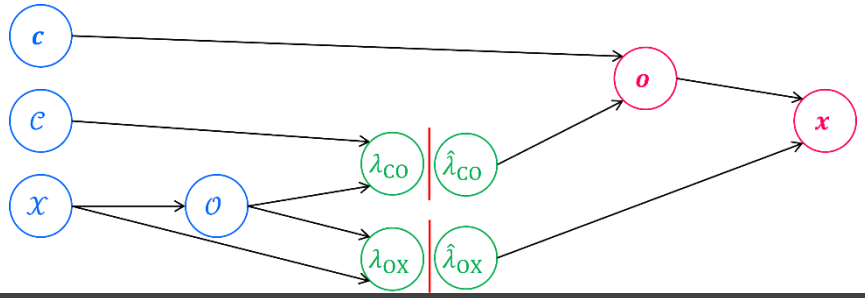
入力特徴量抽出と音響特徴量予測の統合

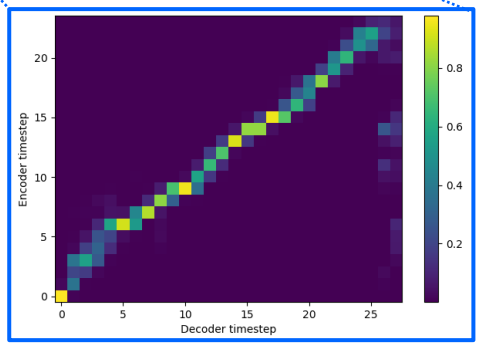
e.g., Tacotron [Wang+17]









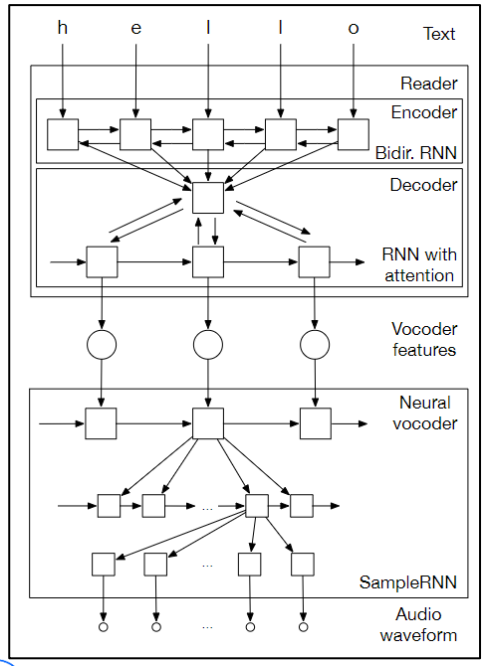


Attention の図は ここ より引用, 音声サンプルは ここ より引用

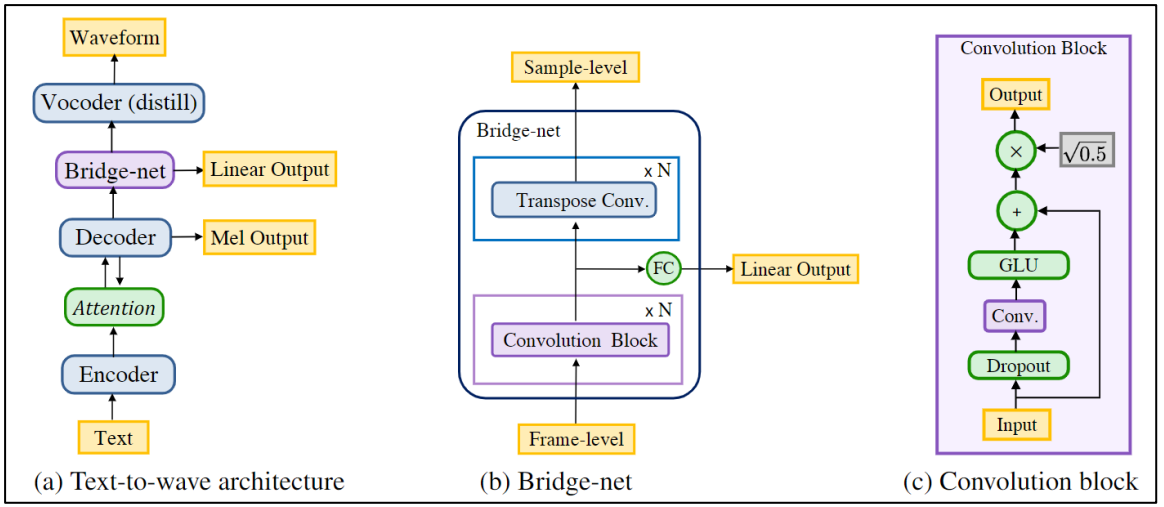
一貫学習に基づく統計的音声合成 (3/3)

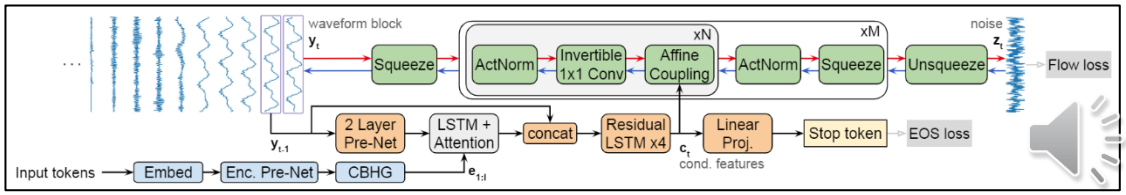
すべての統合 (真の End-to-End モデリング)

char2wav [Sotelo+17]

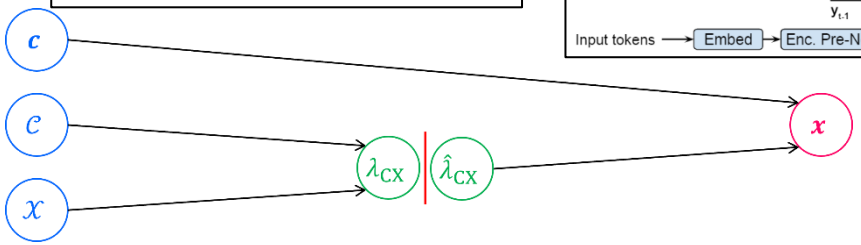


ClariNet [Ping+19]





Wave-Tacotron [Weiss+21]



音声サンプルは ここ より引用

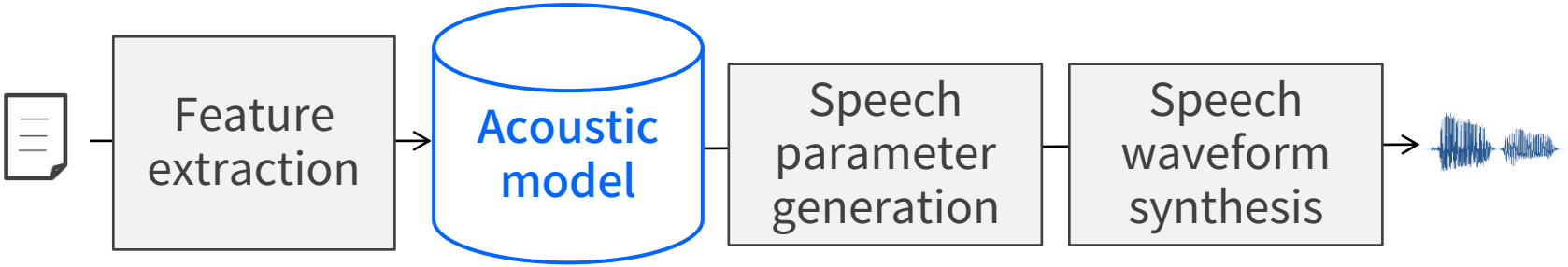
本節のまとめ

統計的音声合成 = 様々な分野の複合研究領域

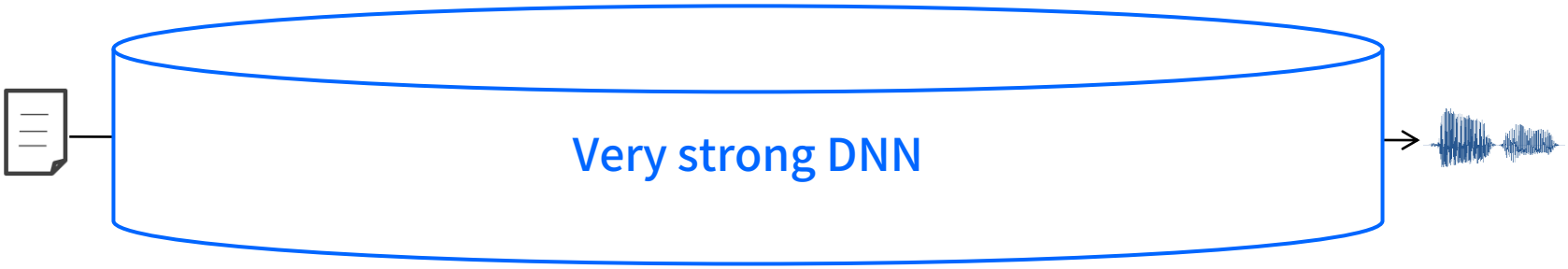
音声情報処理, 信号処理, 自然言語処理, 機械学習, etc...

部分問題への分割方式から End-to-End 方式へ

従来方式 (統計的パラメトリック音声合成 [Zen+09]):



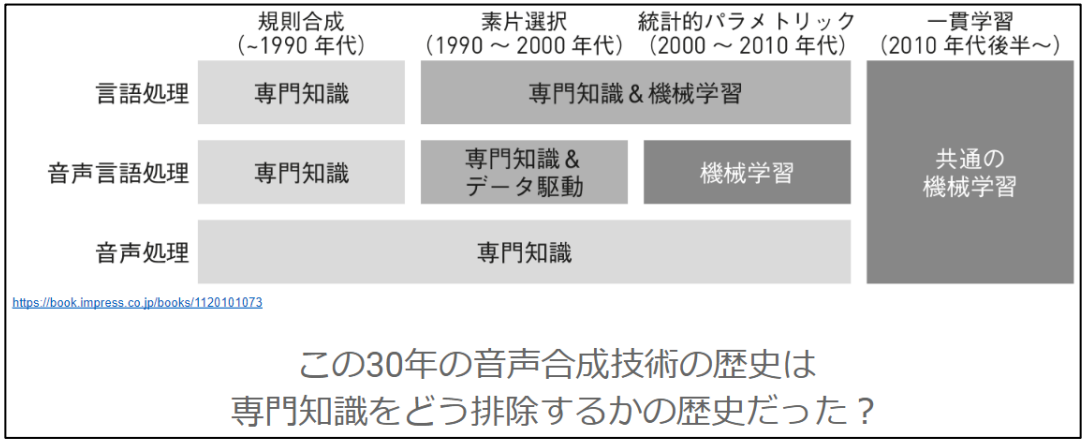
End-to-End 方式:



(余談) DNN 以前の音声合成技術は学ぶべき?

ツールとして使うだけなら... 必ずしもそうではないと思います.

高品質な音声コーパスやオープンソースな音声合成技術の公開



研究開発の対象としたいなら... 強く推奨します.

旧方式の技術は, 思わぬところで役立つことも

- HMM の構造を導入した効率的な系列モデリング [Mehta+22]
- GMM を用いた音声合成の発話スタイルモデリング [Hsu+18]

End-to-end と旧方式のいいとこ取りを採用した最新技術も

- e.g., FastSpeech 2 [Ren+21]: ボコーダパラメータを中間特徴量として利用

図は東大院「信号処理特論」の講義資料より引用 (link)

本講演の概要・目次

概要

統計的音声合成の基礎から, 深層学習に基づく最先端技術までを学ぶ.

DNN 音声合成

目次

本講演で扱うトピック

Sequence-to-sequence (seq2seq) 学習

TTS/VC における入出力系列長の違いについてどう対処するか?

深層生成モデル (Deep Generative Model: DGM)

音声の複雑な確率分布をどのように表現・学習するか?

Unsupervised/non-parallel なデータを用いた学習

TTS/VC 学習のためのペアデータ収集の難しさにどう対処するか?

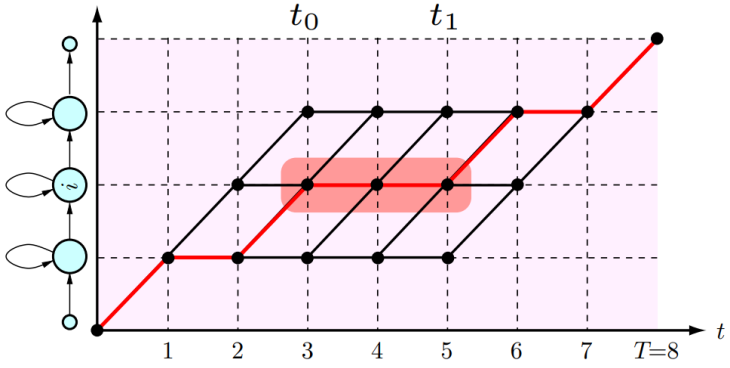
TTS/VC における系列マッピング

統計的音声合成における重要なタスクの一つ

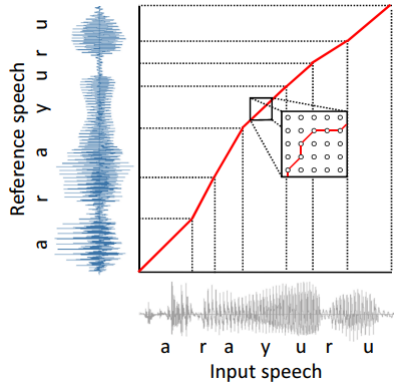
- TTS: 入力長は数十 (テキストの音素数), 出力長は数百フレーム
- VC: 発話内容が同じでも, 音声特徴量のフレーム数は異なる

旧方式でのアプローチ (基本的には動的計画法に基づく)

- TTS: Forced alignment 由来の音素継続長を用いた教師あり学習
- VC: 動的時間伸縮 (DTW) を用いた事前の特徴量アラインメント



https://www.sp.nitech.ac.jp/~tokuda/tokuda_interspeech09_tutorial.pdf



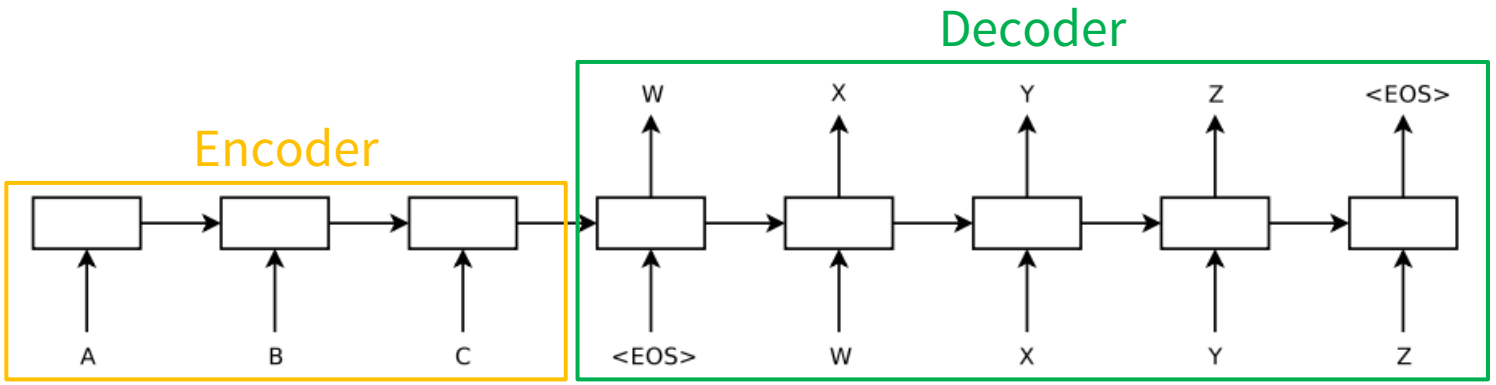
http://sython.org/papers/thesis/saito21PhD_thesis.pdf

問題点: アラインメント時のエラーが学習に伝播

Sequence-to-Sequence (seq2seq) 学習

入力系列と出力系列の対応関係を学習

- 当初は機械翻訳 (machine translation) における技術
- Encoder: 入力系列を圧縮 / Decoder: 圧縮表現から出力系列を予測
  - 圧縮表現 = context vector (入力系列の文脈情報を保持)

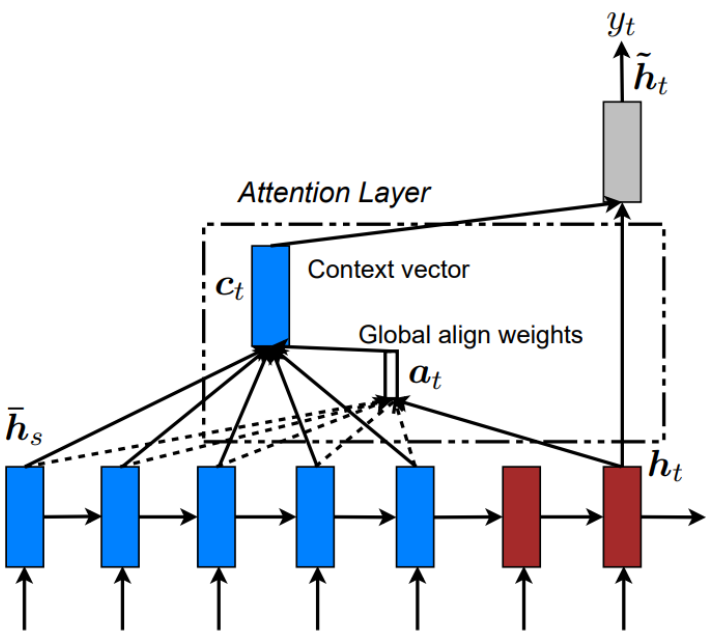


オリジナルの手法 [Sutskever+14]: **系列のアラインメントは不明**

- 系列から系列への変換は可能だが，要素同士の対応関係はわからない

注意機構 (attention) を用いた seq2seq 学習

Context vector の推定方法を改良

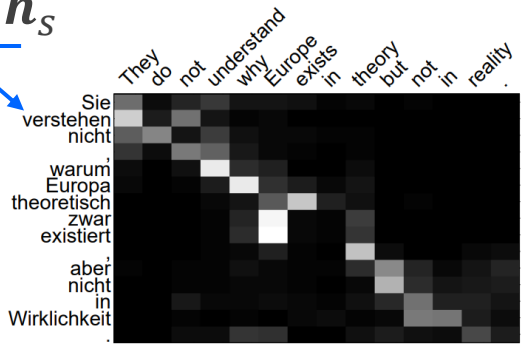


$$a_t(s) = \mathrm{align}(h_t, \bar{h}_s) \quad \text{Softmax}$$

$$= \frac{\exp\left(\mathrm{score}(h_t, \bar{h}_s)\right)}{\sum_{s'} \exp\left(\mathrm{score}(h_t, \bar{h}_{s'})\right)}$$

Hidden states の類似度 (内積)

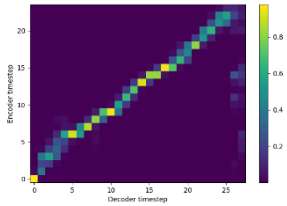
$$c_t = \sum_s a_t(s)\bar{h}_s$$



Attention について補足

TTS/VC では, 基本的に alignment は対角

対角に近づくように学習を誘導することも [Tachibana+18]



各計算の要素は, (Query, Key, Value) として解釈可能 → 次スライド

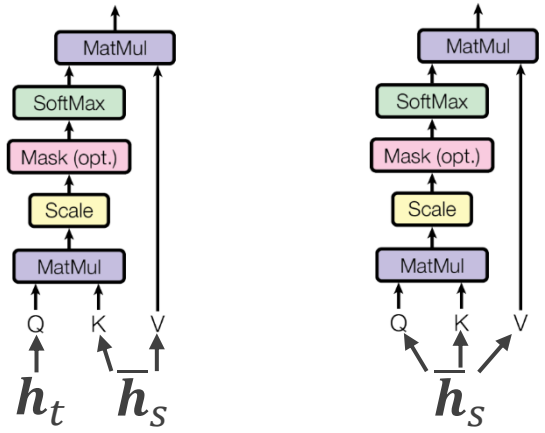
[Luong+15]

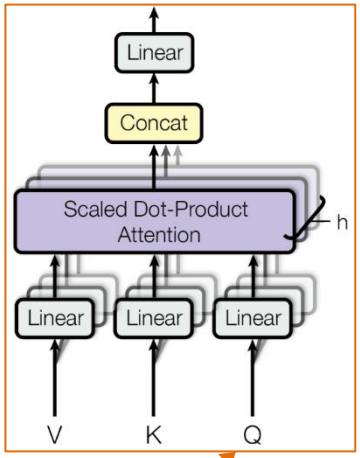
(余談?) Transformer & self-attention

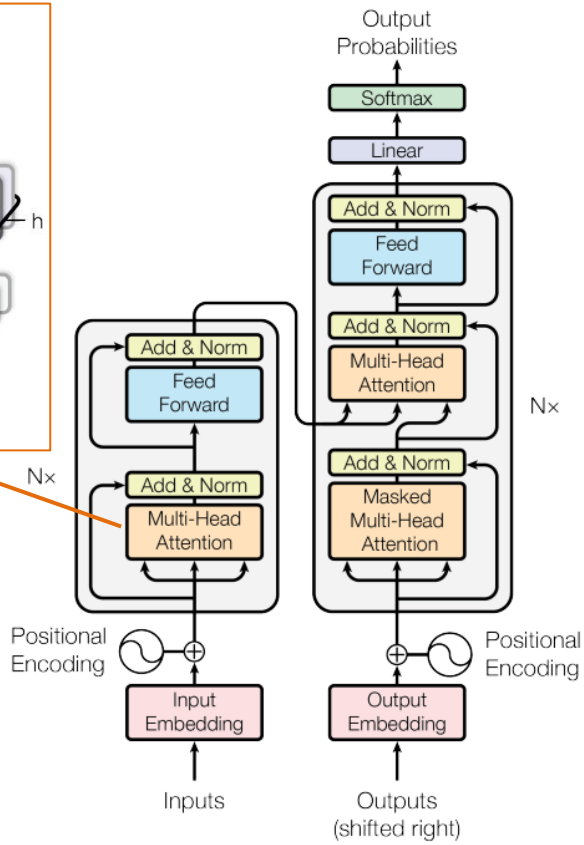
"Attention Is All You Need" 論文で登場

Self-attention により, 入力系列内での複雑な依存関係をモデル化

Attention　　Self-attention







位置情報のモデル化

Position-wise Feed-Forward Network

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Positional Encoding

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

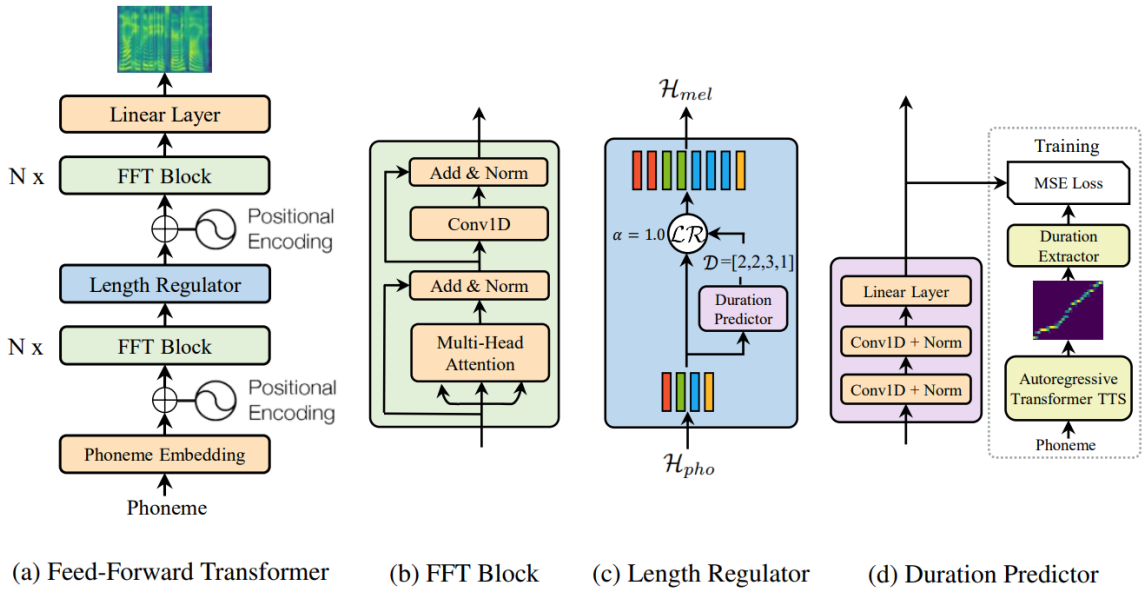
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

[Vaswani+17]

Length regulator を用いた End-to-End 学習

旧方式における継続長モデルを音響モデルに内包 & 同時に学習

代表例: FastSpeech [Ren+19a] (Non-AR なので高速に推論可能)



(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator (d) Duration Predictor

改良版の FastSpeech2 [Ren+21]

音声の {F0, energy} を予測する **variance adaptors** も内包

合成音声の品質と制御性を改善

VC にも応用可能 [Hayashi+21]

本講演で扱うトピック

Sequence-to-sequence (seq2seq) 学習

TTS/VC における入出力系列長の違いについてどう対処するか?

深層生成モデル (Deep Generative Model: DGM)

音声の複雑な確率分布をどのように表現・学習するか?

Unsupervised/non-parallel なデータを用いた学習

TTS/VC 学習のためのペアデータ収集の難しさにどう対処するか?

深層生成モデル (DGM)

深層生成モデル: データ $\boldsymbol{x}$ の分布 $p(\boldsymbol{x})$ を表現する DNN

学習: 学習データ $\mathcal{X}$ を用いて, 真のデータ分布 $p^*(\boldsymbol{x})$ を近似する
生成分布 $p(\boldsymbol{x}|\lambda)$ のモデルパラメータ λ を推定

生成: $p(\boldsymbol{x}|\lambda)$ から新しいデータをサンプリング

e.g., WaveNet = AR 深層生成モデル $p(\boldsymbol{x}|\lambda) = \prod_{t=1}^{T} p(x_t|x_1, x_2, \cdots, x_{t-1})$

応用例

新しいデータの生成 (創作活動支援, データ拡張など)

生成分布を用いた半教師あり学習 (データの事前分布として利用)

代表的な DGMs (参考: [Ruthotto+21])

Variational AutoEncoder (VAE) [Kingma+14]

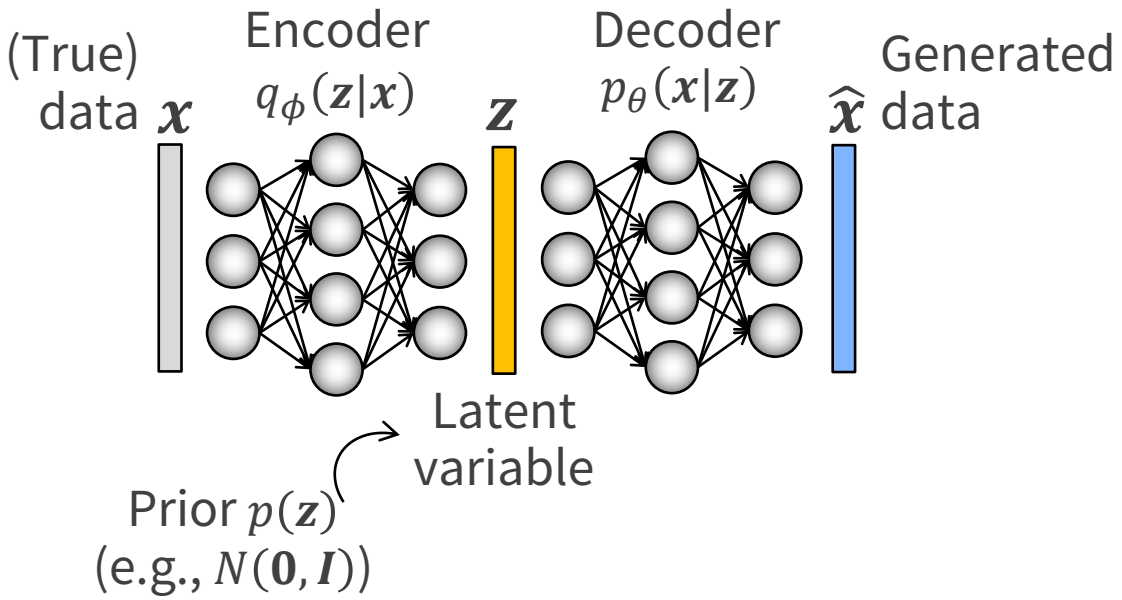
Generative Adversarial Network (GAN) [Goodfellow+14]

Flow [Rezende+15]

VAE: 潜在変数の分布に制約をつけた autoencoder

Encoder: データから潜在変数を抽出

Decoder: 潜在変数からデータを生成



$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right) - \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$$

潜在変数に対する正則化項

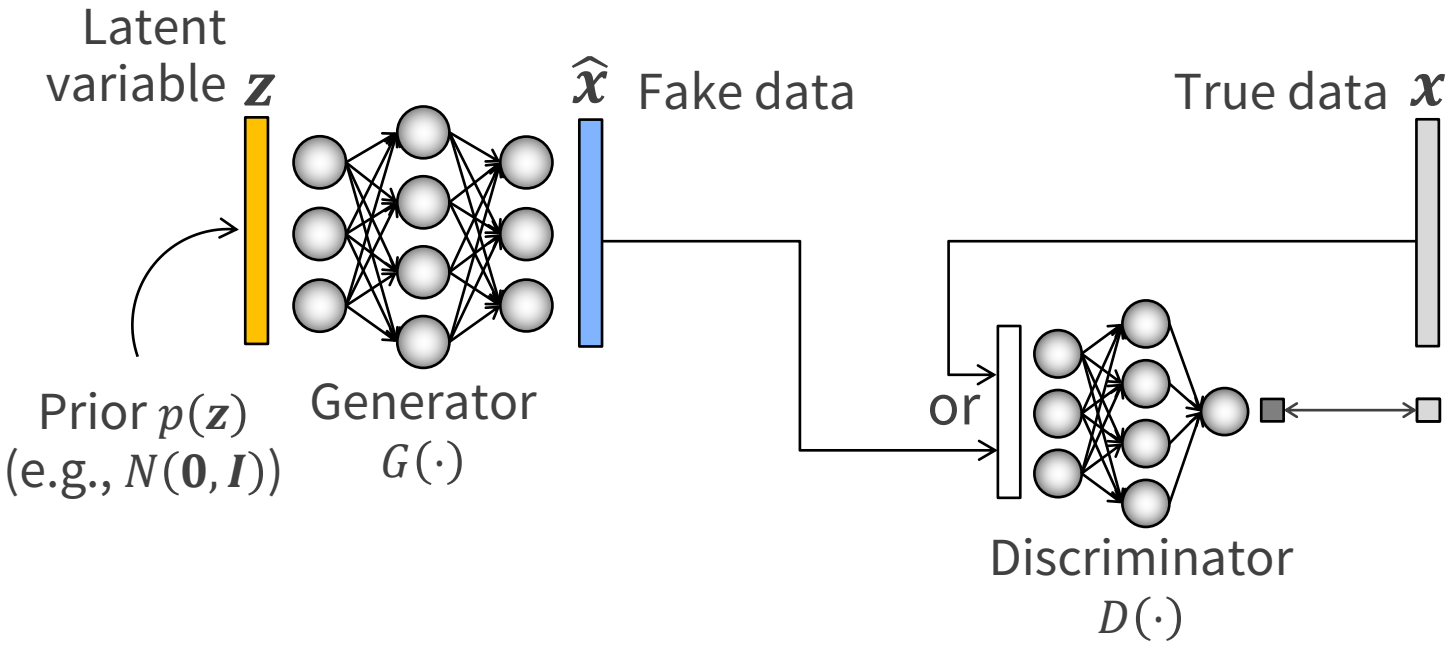
データの再構築誤差

[Kingma+14]

GAN: Discriminator と Generator のミニマックス最適化

Discriminator D: 真のデータと偽のデータを識別

Generator G: 潜在変数から D を騙せるような偽のデータを生成



$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p^*(\boldsymbol{x})}[\log D(x)] + \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}\big[\log\big(1 - D(G(z))\big)\big]$$

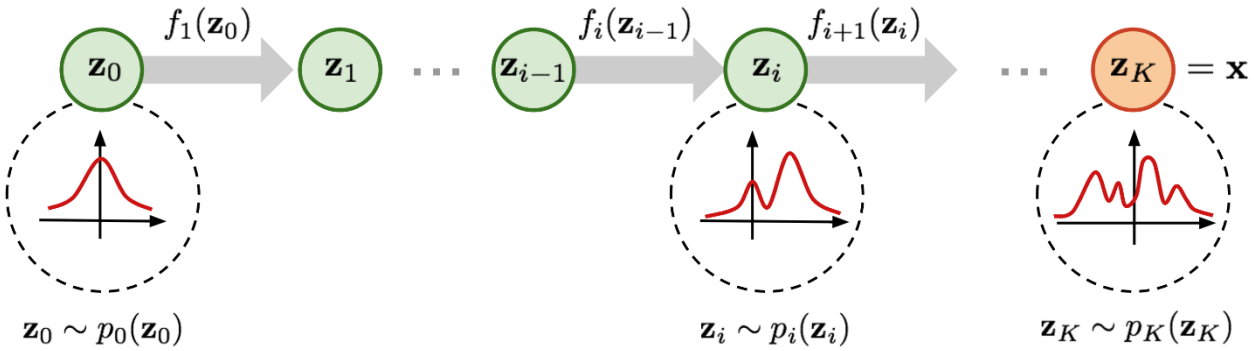
真のデータと偽のデータの分布間距離規範最小化

[Goodfellow+14]

Flow: 多層の変数変換に基づく生成分布の学習

可逆非線形変換 f により, 潜在変数の分布をデータの分布に変換

- f を NN で表現し, 変数変換によりデータの確率分布を表現
  - $x = f_K \circ f_{K-1} \circ \cdots \circ f_1(\boldsymbol{z}_0)$ ... 潜在変数からのデータ生成
- f の具体例
  - Coupling [Dinh+15], Residual [Behrmann+19], AR [Kingma+16]



https://lilianweng.github.io/posts/2018-10-13-flow-models/

$$\log p\,(\boldsymbol{x}) = \log p_0(\boldsymbol{z}_0) - \sum_{k=1}^{K} \log \left| \det \frac{d f_i}{d \boldsymbol{z}_{i-1}} \right|$$

変数変換由来の Jacobian

(おまけ) Denoising Diffusion Probabilistic Model (DDPM)

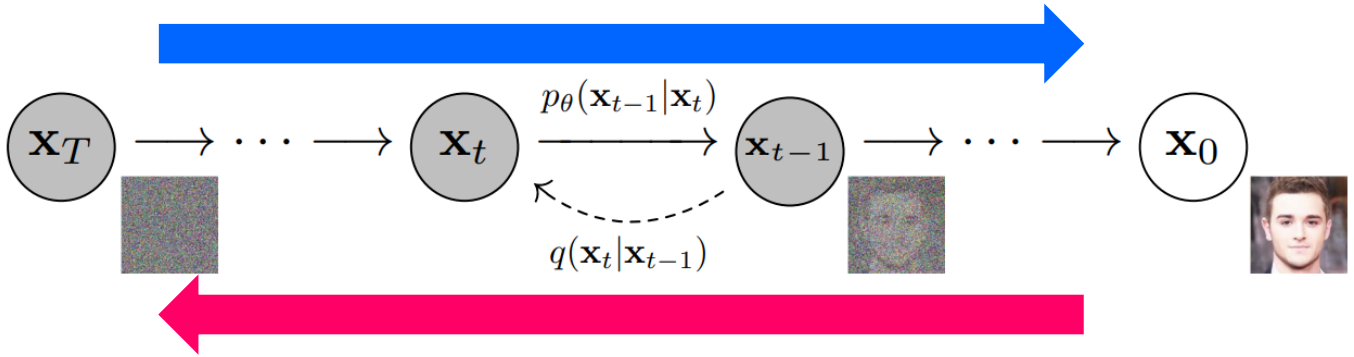
非平衡熱力学に基づく生成モデルのクラス

Forward (diffusion) process: データに Gaussian ノイズを付加

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

Reverse process: ノイズからデータを復元

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$



Training: データに付加されたノイズ ϵ を予測する

$$\mathcal{L} = \mathbb{E}_{t\sim[1,T],\, x_0\sim q(x_0),\, \epsilon\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}[\|\epsilon - \epsilon_\theta(x_t, t)\|]$$

DNN によるノイズ予測

[Ho+20]

本講演で扱うトピック

Sequence-to-sequence (seq2seq) 学習
- TTS/VC における入出力系列長の違いについてどう対処するか?

深層生成モデル (Deep Generative Model: DGM)
- 音声の複雑な確率分布をどのように表現・学習するか?

Unsupervised/non-parallel なデータを用いた学習
- TTS/VC 学習のためのペアデータ収集の難しさにどう対処するか?

Motivation

TTS/VC における課題: 学習データ収集の難しさ

- TTS: (テキスト, 音声) のペアが必要
  - 発話内容の書き起こしはコストが大きい (そもそも不可能な場合も)
- VC: (変換元, 変換先) 話者の同一発話内容 (パラレル) データが必要
  - あらゆる話者のパラレルデータを集めるのは非現実的

Core idea: 音声から発話内容に関する特徴を抽出・分離

- 本講演では, 以下を紹介
  - 音声認識 (Automatic Speech Recognition: ASR) の利用
  - GAN や VAE の導入

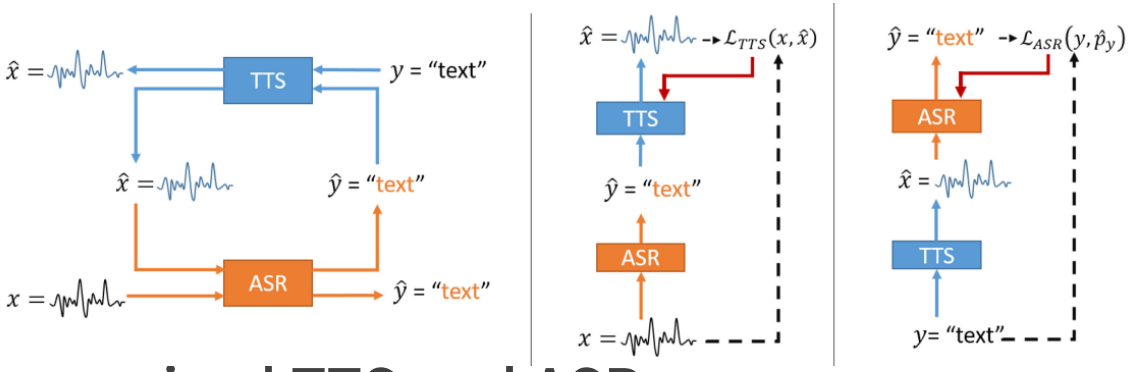
(余談) ノンパラレルなデータでの VC 学習 = アラインメント不要

- VC での seq2seq 学習があまり研究されていないのはこれが一因?

ASR の導入による TTS の (ほぼほぼ) 教師なし学習

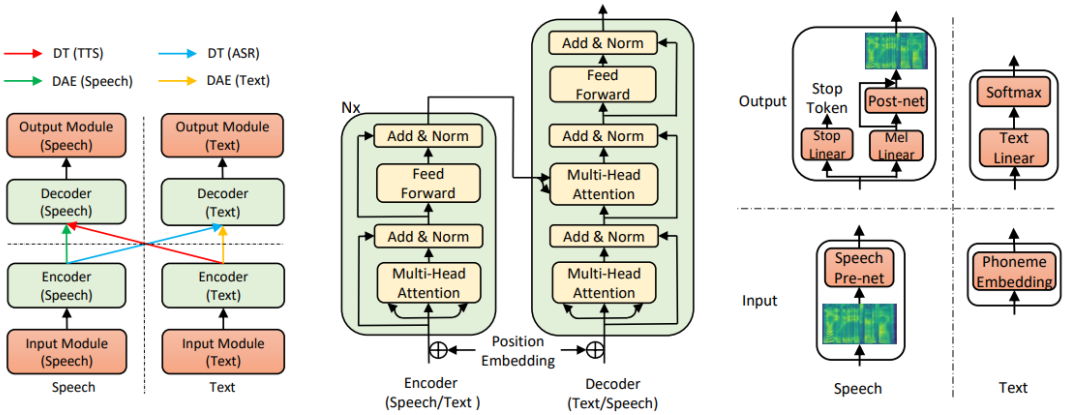
Machine speech chain [Tjandra+20]

ASR 由来の疑似テキストを用いて TTS を学習 (逆も可能)



Almost unsupervised TTS and ASR [Ren+19b]

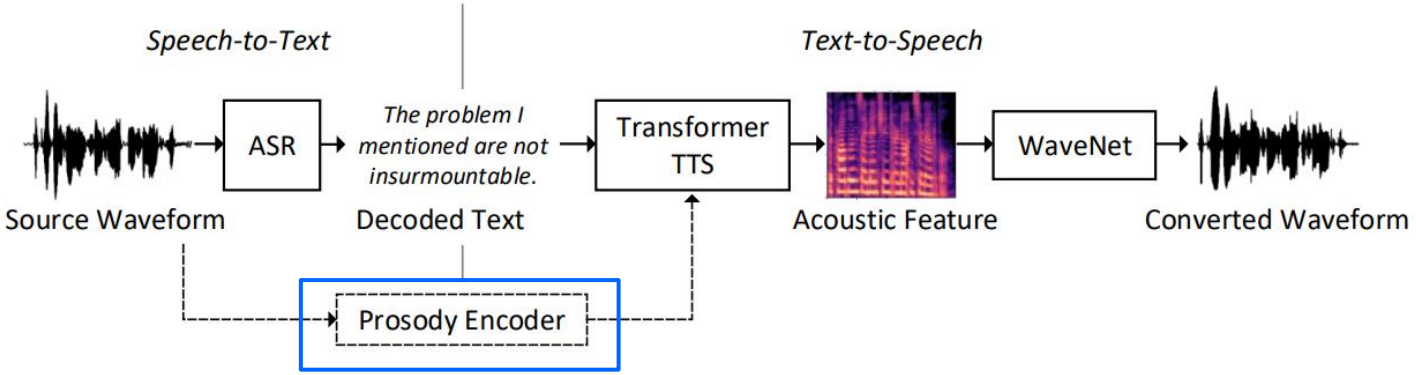
テキスト/音声の (denoising AE) を2つ用意し, TTS/ASR のパスを実現



ASR と TTS の結合によるノンパラレル VC

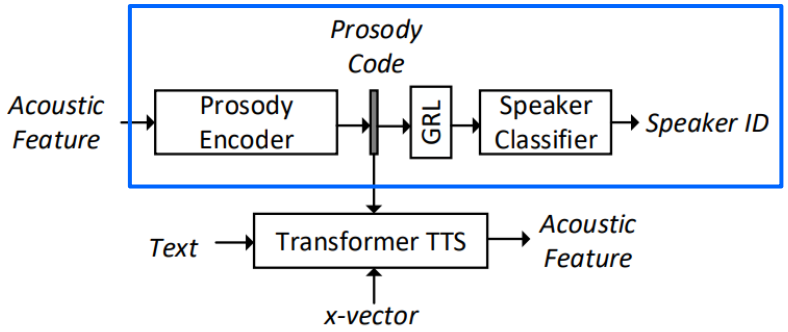
Voice Conversion Challenge 2020 のチャンピオン[Zhang+20]

ASR と 多話者 TTS と WaveNet Vocoder を結合



Speaker adversarial training により, 話者非依存な韻律情報を学習

→ 入力音声のイントネーションやアクセントが不変であることを担保



Prosody code を用いた話者識別が失敗するように学習

GRL: Gradient Reversal Layer (backprop. 時に符号を反転させる)

敵対学習に基づくノンパラレル VC

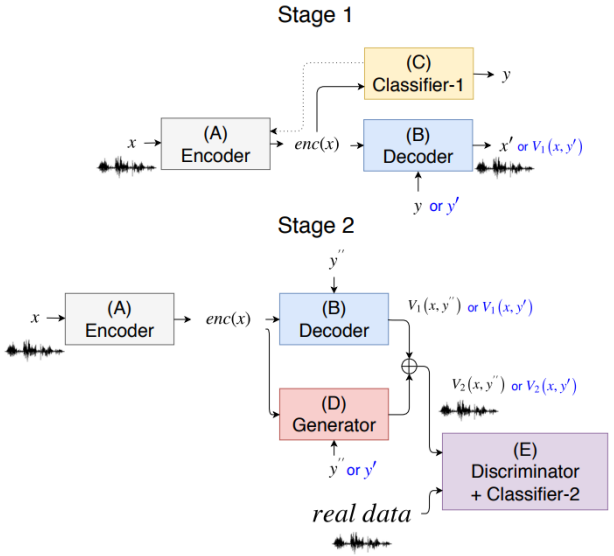
話者敵対学習 & GAN の導入 [Chou+18]

Stage 1:

話者敵対学習で話者非依存な情報を学習

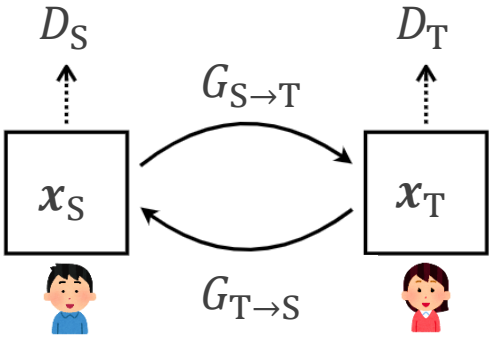
Stage 2:

GAN で当該話者の自然な音声を合成
(厳密には Auxiliary Classifier GAN が正しい?)

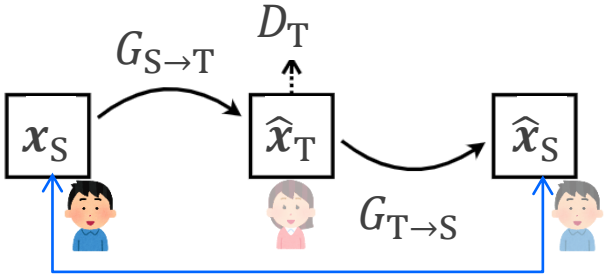


CycleGAN-VCs* [Kaneko+18]

話者変換モデル×2
& Discriminator×2

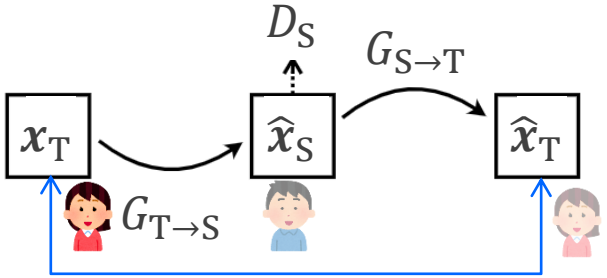


ドメイン T の敵対学習



$S \to T \to S$ の
循環無矛盾学習

ドメイン S の敵対学習



$T \to S \to T$ の
循環無矛盾学習

*多話者 VC に拡張した StarGAN-VCs [Kameoka+18] も存在

本節のまとめ

高品質な統計的音声合成を支える基盤技術を紹介

- Seq2seq 学習: 系列間のアラインメントをデータドリブンで学習
  - (Self-)Attention は数多くの TTS/VC モデルの根幹になりつつある
- 深層生成モデル: 音声の複雑な確率分布を表現する DNN
  - VAE, GAN, Flow, DDPM を紹介 (いずれも TTS/VC で広く活用)
- TTS/VCの教師なし or ノンパラレル学習: データ収集コストの削減
  - ASR の利用から敵対学習の導入までを紹介

より深く学びたい人へ

- TTS 技術のサーベイ論文 → [Mu+21][Tan+21]
- VC 技術のサーベイ論文 → [Sisman+20]
- 深層生成モデルのサーベイ論文 → [Ruthotto+21]

(余談) GAN ベース TTS/VC のルーツは日本だった!

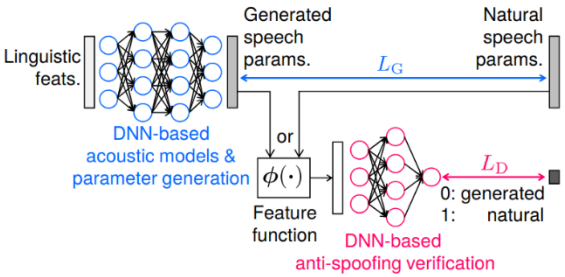
Training algorithm to deceive Anti-Spoofing Verification for DNN-based speech synthesis

Yuki Saito; Shinnosuke Takamichi; Hiroshi Saruwatari

Publication Year: 2017 , Page(s): 4900 - 4904

Cited by: Papers (8)

▸ Abstract **HTML**



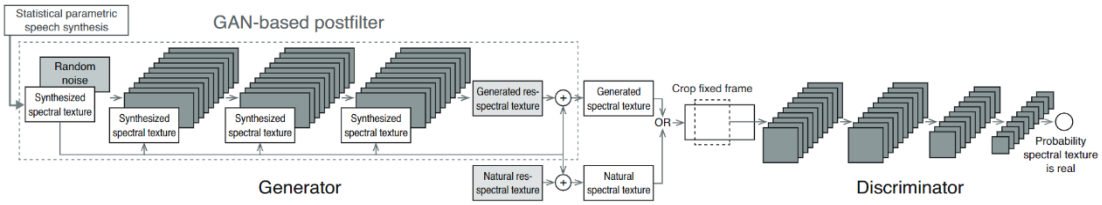
Generative adversarial network-based postfilter for statistical parametric speech synthesis

Takuhiro Kaneko; Hirokazu Kameoka; Nobukatsu Hojo; Yusuke Ijima; Kaoru Hiramatsu; Kunio Kashino

Publication Year: 2017 , Page(s): 4910 - 4914

Cited by: Papers (37)

▸ Abstract **HTML**



詳細は「2021年度 ガウス過程と機械学習特論」でのゲスト講演資料を参照

本講演の概要・目次

概要

統計的音声合成の基礎から, 深層学習に基づく最先端技術までを学ぶ.

DNN 音声合成

目次

本講演で紹介する技術

2022年6月現在における SOTA 技術を紹介

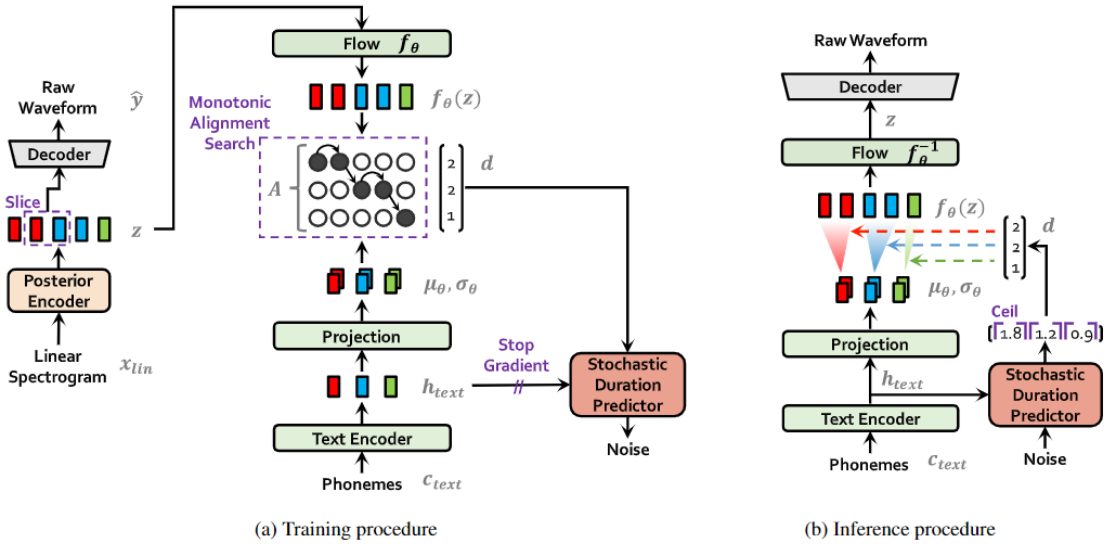
- TTS
  - VITS: Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech [Kim+21]
  - NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality [Tan+22]
- VC
  - NVC-Net: End-to-End Adversarial Voice Conversion [Nguyen+22]
- Neural vocoder
  - BigVGAN: A Universal Neural Vocoder with Large-Scale Training [Lee+22]

注: 齋藤の個人的な主観・解釈を大いに含みます.

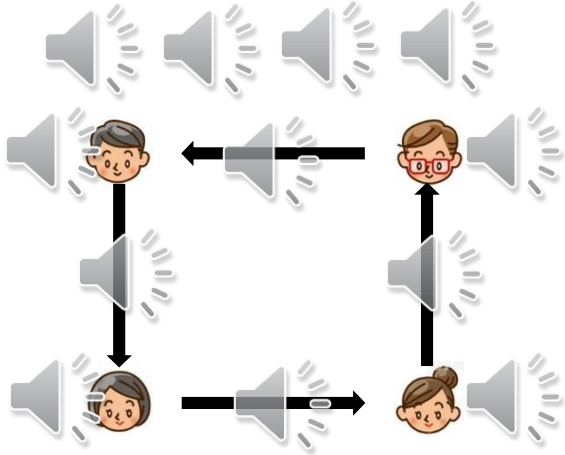
VITS [Kim+21]

これまで紹介してきた技術の全部載せ

Flow, VAE, GAN, Duration Predictor



Multi-speaker TTS も VC もできちゃう



様々な変種・改良版が存在

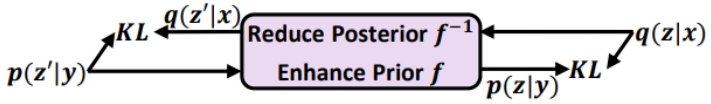
- AdaVITS [Song+22]: モデルの軽量化
- VISinger [Zhang+22]: 歌声合成版
- YourTTS [Casanova+21]: Zero-shot TTS/VC 版

音声サンプルは ここ より引用

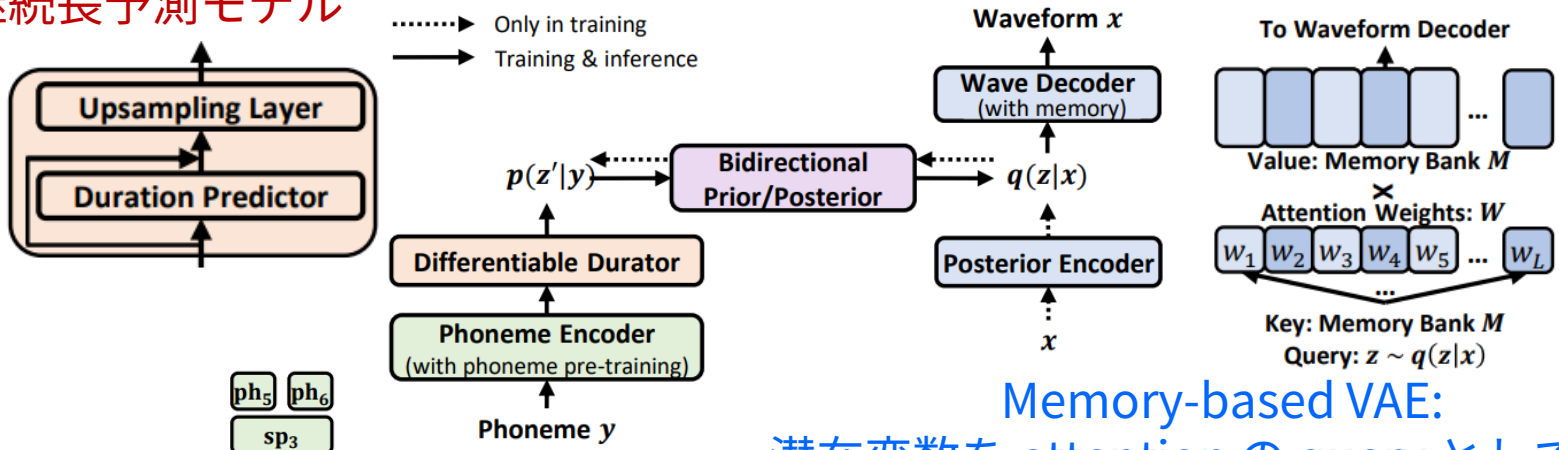
NaturalSpeech [Tan+22]

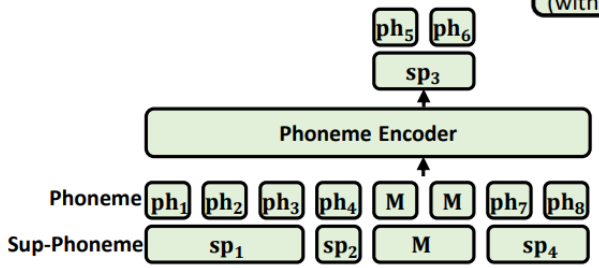
Conditional VAE ベースの TTS モデル

Flow を用いた事前分布/事後分布学習



学習可能な
継続長予測モデル





大規模テキストデータ &
sub-phonemes を用いた
事前学習

Memory-based VAE:
潜在変数を attention の query として使用

Table 2: MOS comparison between NaturalSpeech and human recordings. Wilcoxon rank sum test is used to measure the p-value in MOS evaluation.

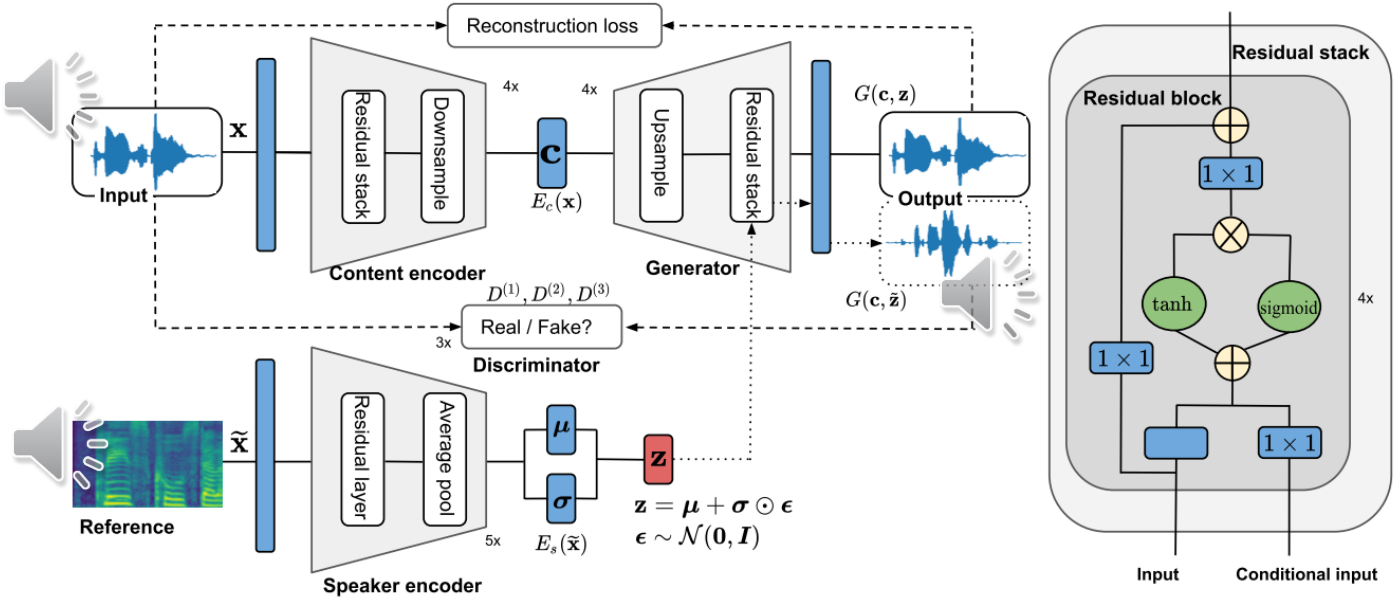
Human Recordings	NaturalSpeech	Wilcoxon p-value
4.58 ± 0.13	4.56 ± 0.13	0.7145

音声サンプルは ここ より引用

NVC-Net [Nguyen+22]

敵対学習に基づく End-to-End VC モデル

- 基本的には AE ベース → ノンパラレルデータで学習可能
- Discriminator の設計を工夫
  - **異なる時間解像度**での特徴を捉えるように設計 $(D^{(1)}, D^{(2)}, D^{(3)})$
  - **話者ごとに true/fake を分類**するように final layer を分岐



音声サンプルは ここ より引用

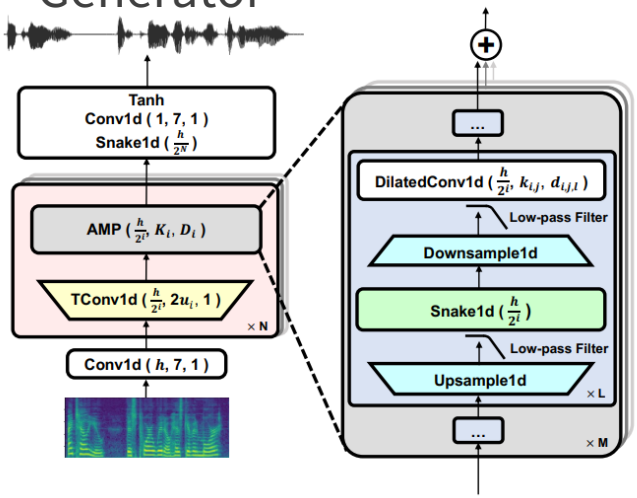
BigVGAN [Lee+22]

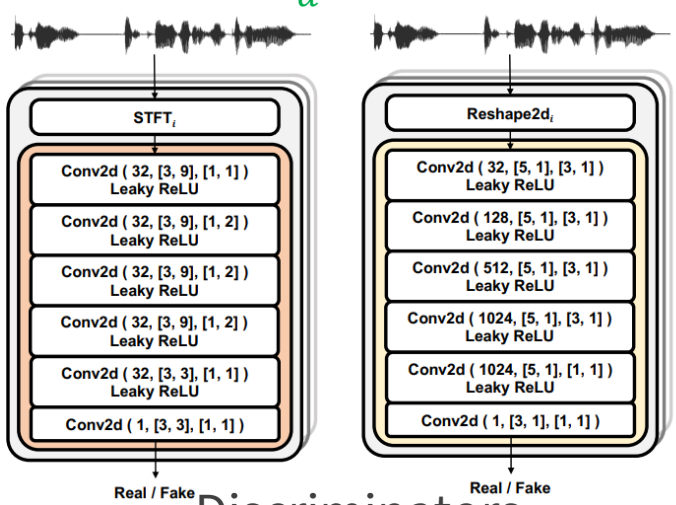
2022年6月9日に arXiv で公開された超大規模ニューラルボコーダ

- Up to **112M** parameters! (> 14M in conventional HiFi-GAN [Kong+20])
- Anti-aliased Multi-Periodicity composition (AMP): 音声の周期性をモデル化
  - 周期的な inductive bias をもたらす Snake 活性化 & anti-aliasing (LPF)

$$f_\alpha(x) = x + \frac{1}{\alpha}\sin^2(\alpha x)$$

Generator





Discriminators
(multi-resolution & multi-period)

ここがスゴイ: unseen audio data でも高品質に合成可能

笑い声/歓声 音楽 未知言語

音声サンプルは ここ より引用

本節のまとめ

統計的音声合成の最先端技術を (一部) 紹介

- 未知話者/未知環境の音声データも高品質に合成可能になりつつある
- Big Tech の機械学習研究者による貢献が大きい
  - TTS/VC が機械学習タスクの一種として捉えられるようになったため
- 一方で, 日本人研究者による貢献も非常に大きい

これ以上やれることはなくなったのか? → NO.

- 合成音声の品質が「人間と同程度」にようやく達しただけ
- 機械と人間が音声で自然にコミュニケーションするためには...
  - より緻密な韻律 (イントネーション・リズム) の制御
  - ドメイン外のデータに対する柔軟な適応機構
    - e.g., 読み上げ音声 → 自発音声 など
  - 音声発話の誤りを即座に訂正可能な機構

本講演の概要・目次

概要

統計的音声合成の基礎から, 深層学習に基づく最先端技術までを学ぶ.

DNN 音声合成

目次

まとめ & 今後の研究潮流

本講演: 深層学習に基づく統計的音声生成 (DNN 音声合成)

統計的音声合成の基礎から, 近年の手法までを紹介

(私が考える) 今後の研究潮流

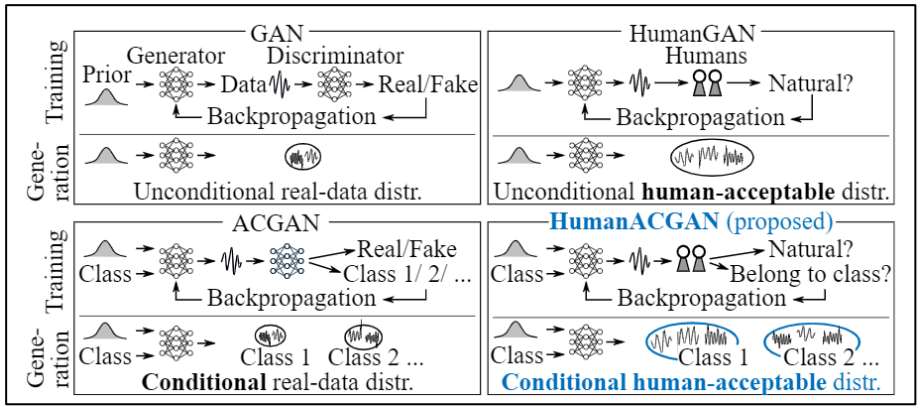
TTS/VC: (基本的には) 人間に聴かせてナンボの研究領域

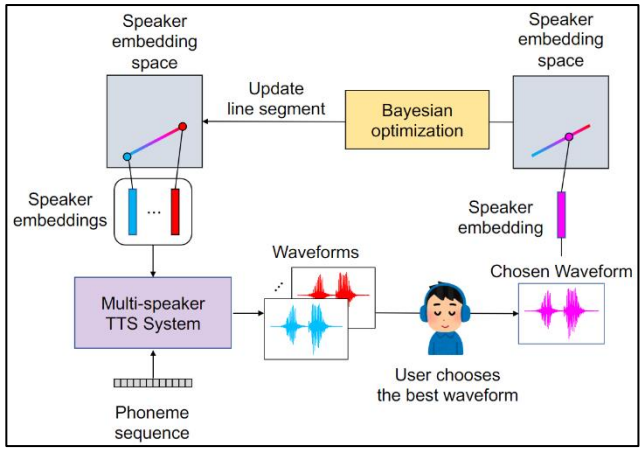
→ 人間をどうやって統計的音声生成の枠組みに取り入れるか?

最近の興味: 人間の知覚評価を取り入れた音響モデル学習・適応

e.g., HumanGAN [Fujii+20], HumanACGAN [Ueda+21],

人間の話者知覚評価に基づく音声合成の話者適応 [宇田川+21]





参考文献リスト (1/5)

[Sagisaka88] Y. Sagisaga, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," In Proc. ICASSP, 1988.

[Stylianou+98] Y. Stylianou et al., "Continuous probabilistic

transform for voice conversion," IEEE Trans. on ASLP, vol. 6, no. 9, 1998.

[Zen+09] H. Zen et al., "Statistical parametric speech synthesis," Speech Communication, vol. 51, no. 11, 2009.

[Tokuda+13] K. Tokuda et al., "Speech synthesis based on hidden Markov models," Proceedings of IEEE, vol. 101, no. 5, 2013.

[Toda+07] T. Toda et al., "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. on ASLP, vol. 15, no. 8. 2007.

[Zen+13] H. Zen et al., "Statistical parametric speech synthesis using deep neural networks," In Proc. ICASSP, 2013.

[Desai+09] S. Desai et al., "Voice conversion using artificial neural networks," In Proc. ICASSP, 2009.

[Hunt+96] A. J. Hunt et al., "Unit selection in concatenative speech synthesis system using a large speech database," In Proc. ICASSP, 1996.

[Hojo+18] N. Hojo et al., "DNN-based speech synthesis using speaker codes," IEICE Trans. on Information and Systems, vol. E101-D, no. 2, 2018.

[Jia+18] Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," In Proc. NeurIPS, 2018.

[Mitsui+21] K. Mitsui et al., "Deep Gaussian process based multi-speaker speech synthesis with latent speaker representation," Speech Communication, vol. 132, 2021.

参考文献リスト (2/5)

[Ito+17] K. Ito et al., "The LJ Speech dataset," 2017.

[Zen+19] H. Zen et al., "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," In Proc. INTERSPEECH, 2019.

[Veaux+12] C. Veaux et al., "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2012.

[Sonobe+17] R. Sonobe et al., "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," arXiv, 2017

[Takamichi+19] S. Takamichi et al., "JVS corpus: free Japanese multi-speaker voice corpus," arXiv, 2007.

[Kurihara+21] K. Kurihara et al., "Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS," IEICE Trans. on Information and Systems, vol. E104.D, no. 2, 2021.

[藤井+21] 藤井 他, "韻律情報で条件付けした非自己回帰型End-to-End日本語音声合成の検討," 情報処理学会研究報告, 2021.

[Morise+16] M. Morise et al., "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. on Information and Systems, vol. E99.D, no. 7, 2016.

[Tan+21] X. Tan et al., "A survey on neural speech synthesis," arXiv, 2021.

[Imai+83] S. Imai et al., "Mel log spectrum approximation (MLSA) filter for speech synthesis," Electronics and Communications in Japan (Part I: Communications), vol. 66, no. 2, 1983.

[Griffin+84] D. Griffin et al., "Signal estimation from modified short-time Fourier transform," IEEE Trans. on ASSP, vol. 32, no. 2, 1984.

参考文献リスト (3/5)

[Oord+16] A. v. D. Oord et al., "WaveNet: A generative model for raw audio," In Proc. SSW, 2016.

[Tamamori+17] A. Tamamori et al., "Speaker-dependent WaveNet vocoder," In Proc. INTERSPEECH, 2017.

[Hayashi+17] T. Hayashi et al., "An investigation of multi-speaker training for WaveNet vocoder," In Proc. ASRU, 2017.

[Mu+21] Z. Mu et al., "Review of end-to-end speech synthesis technology based on deep learning," arXiv: 2021.

[Wang+17] Y. Wang et al., "Tacotron: Towards end-to-end speech synthesis," In Proc. INTERSPEECH, 2017.

[Sotelo+17] J. Sotelo et al., "Char2Wav: End-to-end speech synthesis," In Proc. ICLR Workshop, 2017.

[Ping+19] W. Ping et al., "ClariNet: Parallel wave generation in end-to-end text-to-speech," In Proc. ICLR, 2019.

[Weiss+21] R. J. Weiss et al., "Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis," In Proc. ICASSP, 2021.

[Mehta+22] S. Mehta et al., "Neural HMMs are all you need (for high-quality attention-free TTS)," In Proc. ICASSP, 2022.

[Hsu+18] W.-N. Hsu et al., "Hierarchical generative modeling for controllable speech synthesis," In Proc. ICLR, 2019.

[Ren+21] Y. Ren et al., "FastSpeech 2: Fast and high-quality end-to-end text to speech," In Proc. ICLR, 2021.

[Sutskever+14] I. Sutskever et al., "Sequence to sequence learning with neural networks," In Proc. NIPS, 2014.

参考文献リスト (4/5)

[Luong+15] M.-T. Luong et al., "Effective approaches to attention-based neural machine translation," In Proc. EMNLP, 2015.

[Vaswani+17] A. Vaswani et al., "Attention is all you need," In Proc. NIPS, 2017.

[Ren+19a] Y. Ren et al., "FastSpeech: Fast, robust and controlalble text-to-speech," In Proc. NeurIPS, 2019.

[Hayashi+21] T. Hayashi et al., "Non-autoregressive sequence-to-sequence voice conversion," In Proc. ICASSP, 2021.

[Ruthotto21] L. Ruthotto et al., "An introduction to deep generative modeling," arXiv, 2021.

[Kingma+14] D. P. Kingma et al., "Auto-encoding variational Bayes," In Proc. ICLR, 2014.

[Goodfellow+14] I. J. Goodfellow et al., "Generative adversarial nets," In Proc. NIPS, 2014.

[Rezende+15] D. Rezende et al., "Variational inference with normalizing flow," In Proc. ICML, 2015.

[Ho+20], J. Ho et al., "Denoising Diffusion Probabilistic Models," In Proc. NeurIPS, 2020.

[Tjandra+20] A. Tjandra et al., "Machine Speech Chain," IEEE/ACM Trans. on ASLP, vol. 28, 2020.

[Ren+19b] Y. Ren et al., "Almost unsupervised text-to-speech and automatic speech recognition," In Proc. ICML, 2019.

[Zhang+20] J.-X. Zhang et al., "Voice conversion by cascading automatic speech recognition and text-to-speech synthesis with prosody transfer," In Proc. of Joint Workshop for Blizzard Challenge and Voice Conversion Challenge 2020, 2020.

[Chou+18] J.-c. Chou et al., "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," In Proc. INTERSPEECH, 2018.

[Kaneko+18] T. Kaneko et al., "CycleGAN-VC: parallel-data-free voice conversion using cycle-consistent adversarial networks," In Proc. EUSIPCO, 2018.

参考文献リスト (5/5)

[Kameoka+18] H. Kameoka et al., "StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks," In Proc. SLT, 2018.

[Sisman+21] B. Sisman et al., "An overview of voice conversion and its challenges: From statistical modeling to deep learning," IEEE/ACM Trans. on ASLP, vol. 29, 2021.

[Kim+21] J. Kim et al., "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," In Proc. ICML, 2021.

[Tan+22] X. Tan et al., "NaturalSpeech: End-to-end text to speech synthesis with human-level quality," arXiv, 2022:

[Nguyen+22] B. Nguyen et al., "NVC-Net: End-to-end adversarial voice conversion," In Proc. ICASSP, 2022.

[Lee+22] S.-g. Lee et al., "BigVGAN: A universal neural vocoder with large-scale training," arXiv, 2022.

[Fujii+20] K. Fujii et al., "HumanGAN: generative adversarial network with human-based discriminator and its evaluation in speech perception modeling," In Proc. ICASSP, 2020.

[Ueda+21] Y. Ueda et al., "HumanACGAN: Conditional generative adversarial network with human-based auxiliary classifier and its evaluation in phoneme perception," In Proc. ICASSP, 2021.

[宇田川+21] 宇田川 他, "人間の知覚評価フィードバックによる音声合成の話者適応," 電子情報通信学会研究報告, 2021.