現代工学基礎 I

物理情報システム工学

工学部　計数工学科

猿渡 洋，妹尾 拓，近藤正章

猿渡 洋：計測工学と信号処理の基礎：

情報理解と実世界インテリジェントシステムへの応用（3）

講義資料と成績評価

- 講義資料
  - http://www.sp.ipc.i.u-tokyo.ac.jp/~saruwatari/

  ※計数工学科/情報理工学系研究科・システム情報第一研究室（猿渡研）からたどれるようにしておきます

- 成績評価
  - レポート課題**（この講義資料の最後にレポート課題があります）**
  - 出席

日程

- 9/26　ガイダンス　担当：猿渡、近藤、妹尾
- 10/3　休講
- 10/10　計測工学と信号処理の基礎：情報理解と実世界インテリジェントシステムへの応用（1）　担当：猿渡
- 10/17　計測工学と信号処理の基礎：情報理解と実世界インテリジェントシステムへの応用（2）　担当：猿渡
- 10/24　計測工学と信号処理の基礎：情報理解と実世界インテリジェントシステムへの応用（3）　担当：猿渡
- 11/ 1　制御工学の基礎：横断的制御理論とロボット制御への応用（1）　担当：妹尾
- （11/7　10/9↓へ振り替え）
- 11/ 9（金）本郷研究室見学（夕方～ 駒場5限後でも参加可能）

日程

- 11/21　制御工学の基礎：横断的制御理論とロボット制御への応用（２）
  担当：妹尾
- 11/28　制御工学の基礎：横断的制御理論とロボット制御への応用（３）
  担当：妹尾
- 12/ 5　情報の認識と計算システムの基礎：人工知能技術とそれを可能にする
  高性能計算技術（１）担当：近藤
- 12/12　先端科学技術研究センター生田・池内研究室見学会（仮）
- 12/19　情報の認識と計算システムの基礎：人工知能技術とそれを可能にする
  高性能計算技術（２）担当：近藤
- 1/ 9　情報の認識と計算システムの基礎：人工知能技術とそれを可能にする
  高性能計算技術（３）担当：近藤

目次

- データの認識・識別問題：深層学習の基礎
  - 基礎構造
  - Deep generative models
- 実波動を作り出す：統計的音声合成の基礎
  - 音声の生成過程
  - 音声合成・変換
  - 最近の研究

目次

- データの認識・識別問題：深層学習の基礎
  - 基礎構造
  - Deep generative models

- 実波動を作り出す：統計的音声合成の基礎
  - 音声の生成過程
  - 音声合成・変換
  - 最近の研究

概要

背景：その強力さが叫ばれて久しい深層学習技術

LSTM, CNN, Seq2Seq, CTC, GAN, AE, MemoryNet, SuperNN, etc.

問題：名前は聞いたことあるけど，中身をよく知らない…

本講義の目的：

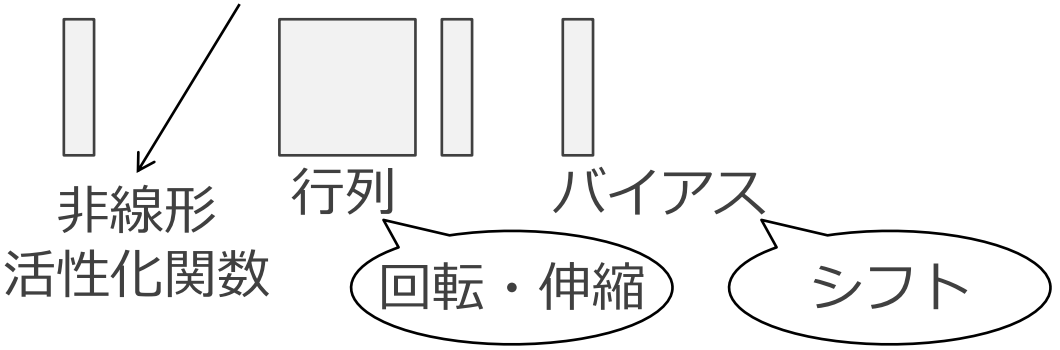
“名前は聞いたことある” から “仕組みがちょっとわかる” へ

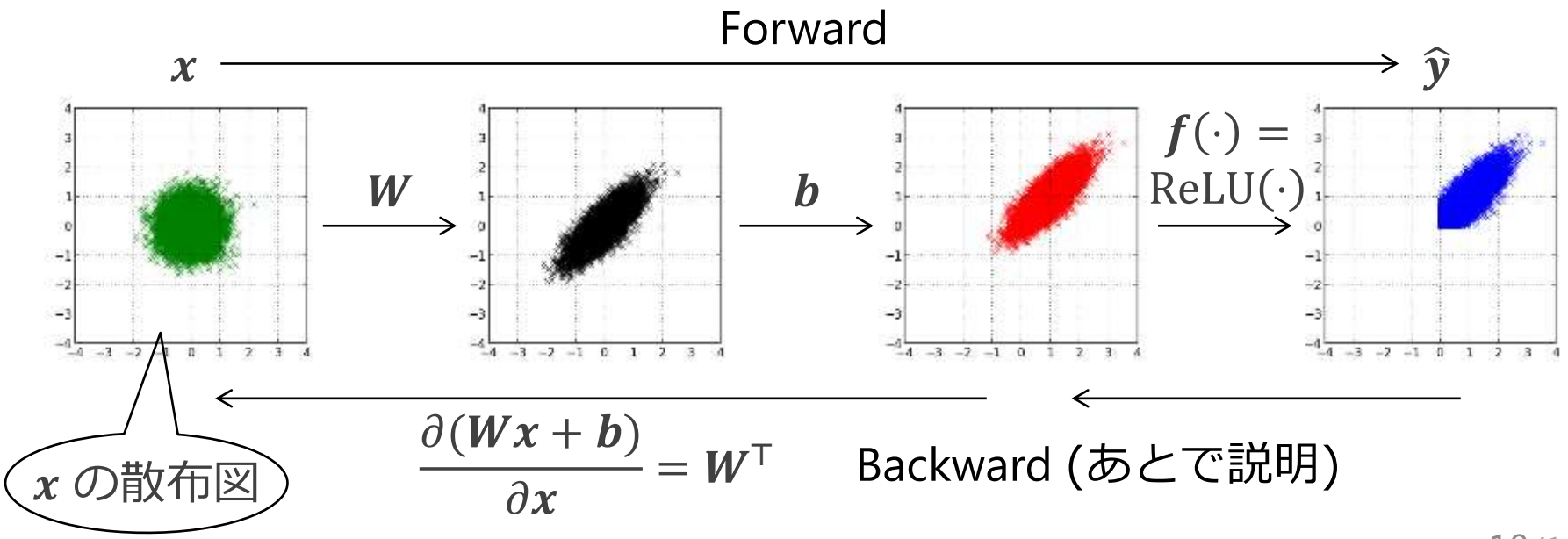
信号処理とも絡めつつ概要を紹介

Feed-Forward NN (Neural Network)

線形変換＋非線形活性化関数による変換

$$\hat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{W}\,\boldsymbol{x} + \boldsymbol{b})$$

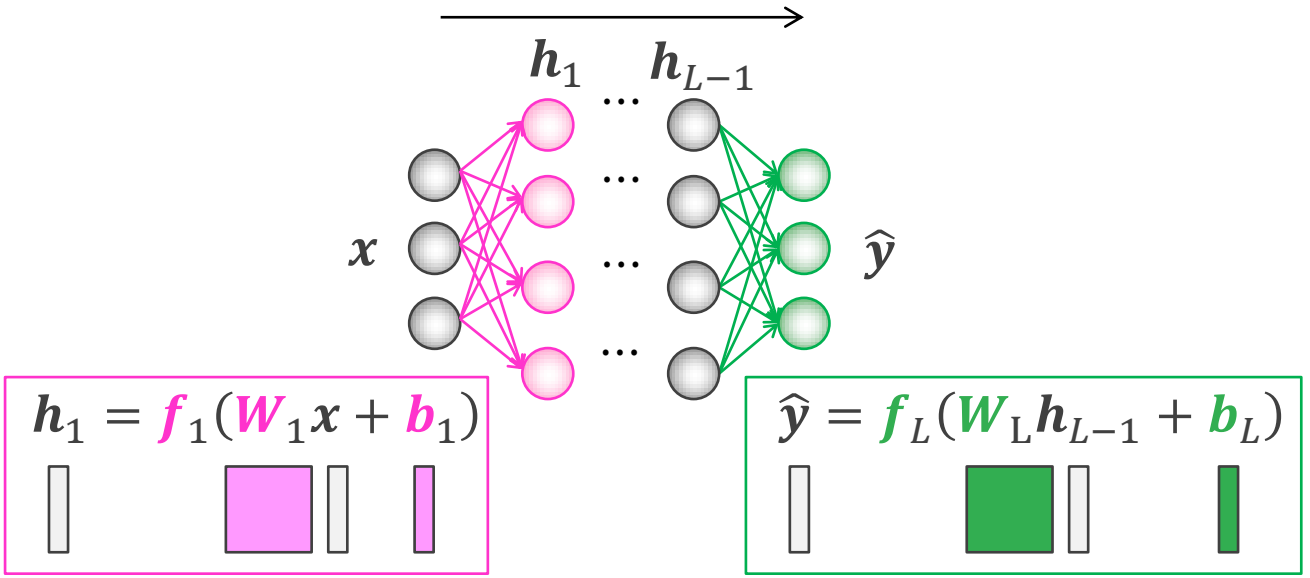




Feed-Forward NN

前のページの構造 (single-layer NN) を積み重ねる！

複数のSingle-layer NN から成る関数



Forward propagationを式で書くと…

$$\hat{y} = f_L(W_L(f_{L-1}(W_{L-1}(\dots f_1(W_1 x + b_1) \dots)) + b_{L-1}) + b_L)$$

モデルパラメータの学習

推定値 $\hat{\mathbf{y}}$ と正解値 $\mathbf{y}$ から計算される損失関数 $L(\cdot)$ を最小化

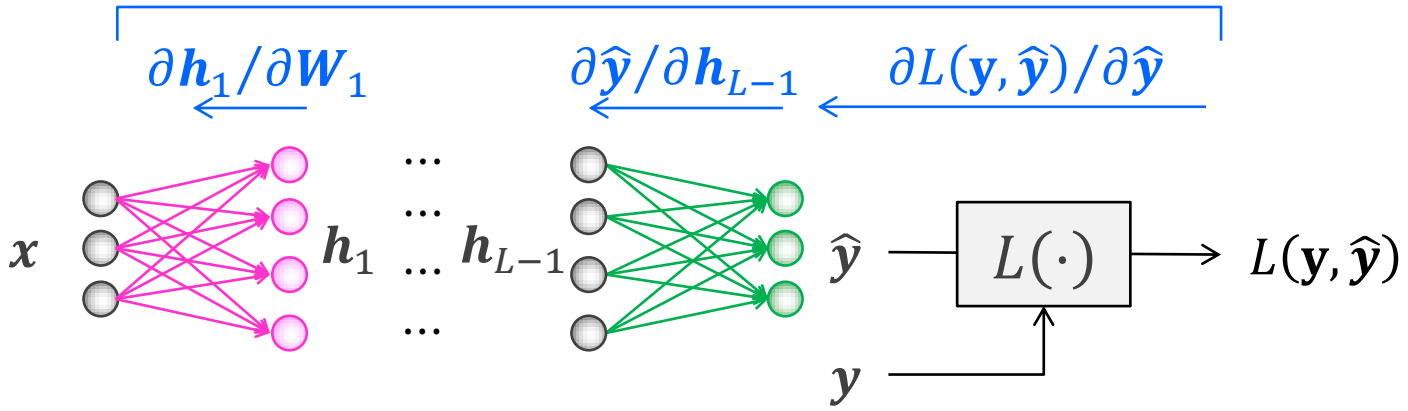
二乗誤差 $L(\mathbf{y}, \hat{\mathbf{y}}) = (\hat{\mathbf{y}} - \mathbf{y})^{\mathsf{T}}(\hat{\mathbf{y}} - \mathbf{y})$

損失関数を最小化するようにモデルパラメータ $\boldsymbol{W}, \boldsymbol{b}$ を更新

勾配法がしばしば使われる（αは学習係数 [AdaGradなどを使用]）

$$\boldsymbol{W}_1 \leftarrow \boldsymbol{W}_1 - \alpha \frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial \boldsymbol{W}_1}$$

合成関数なので，各関数の
微分の積として得られる

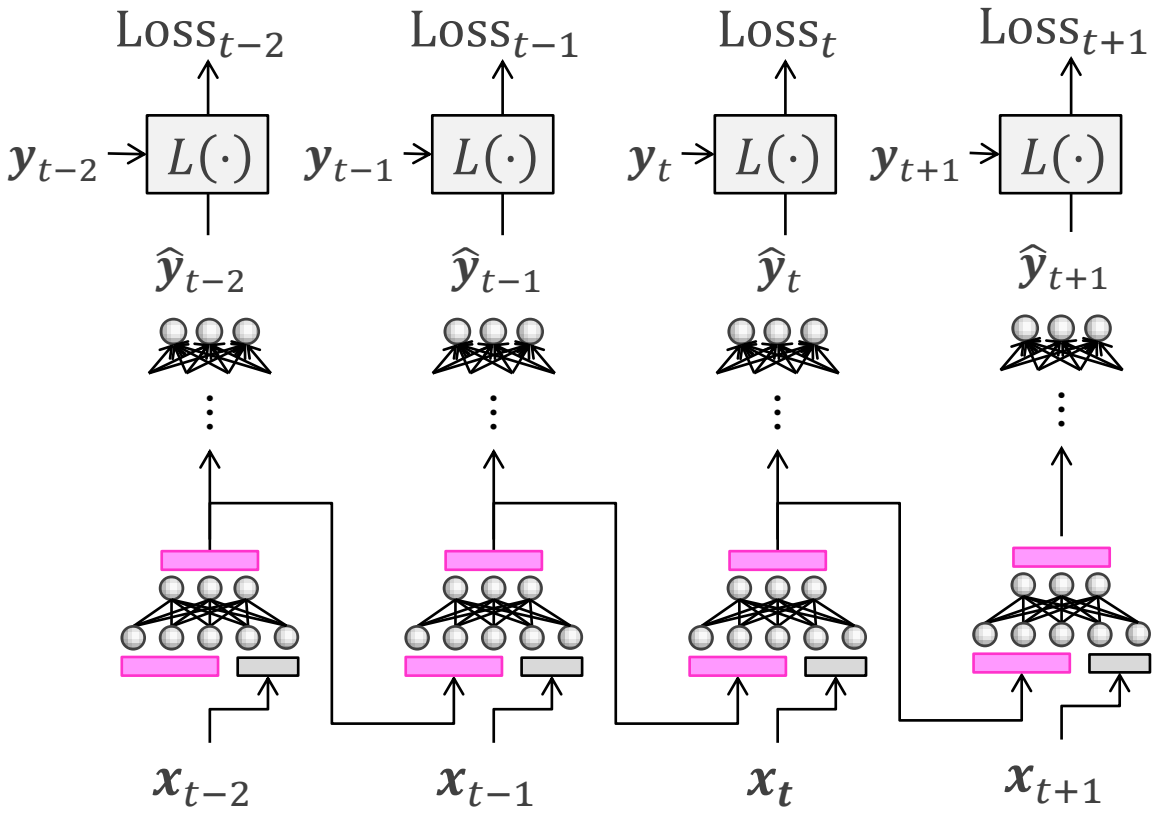


リカレント構造 & 畳み込み構造

RNN (Recurrent NN): リカレント構造を持ったNN

NNの出力の一部を入力に戻すNN (LSTMは，これの派生)

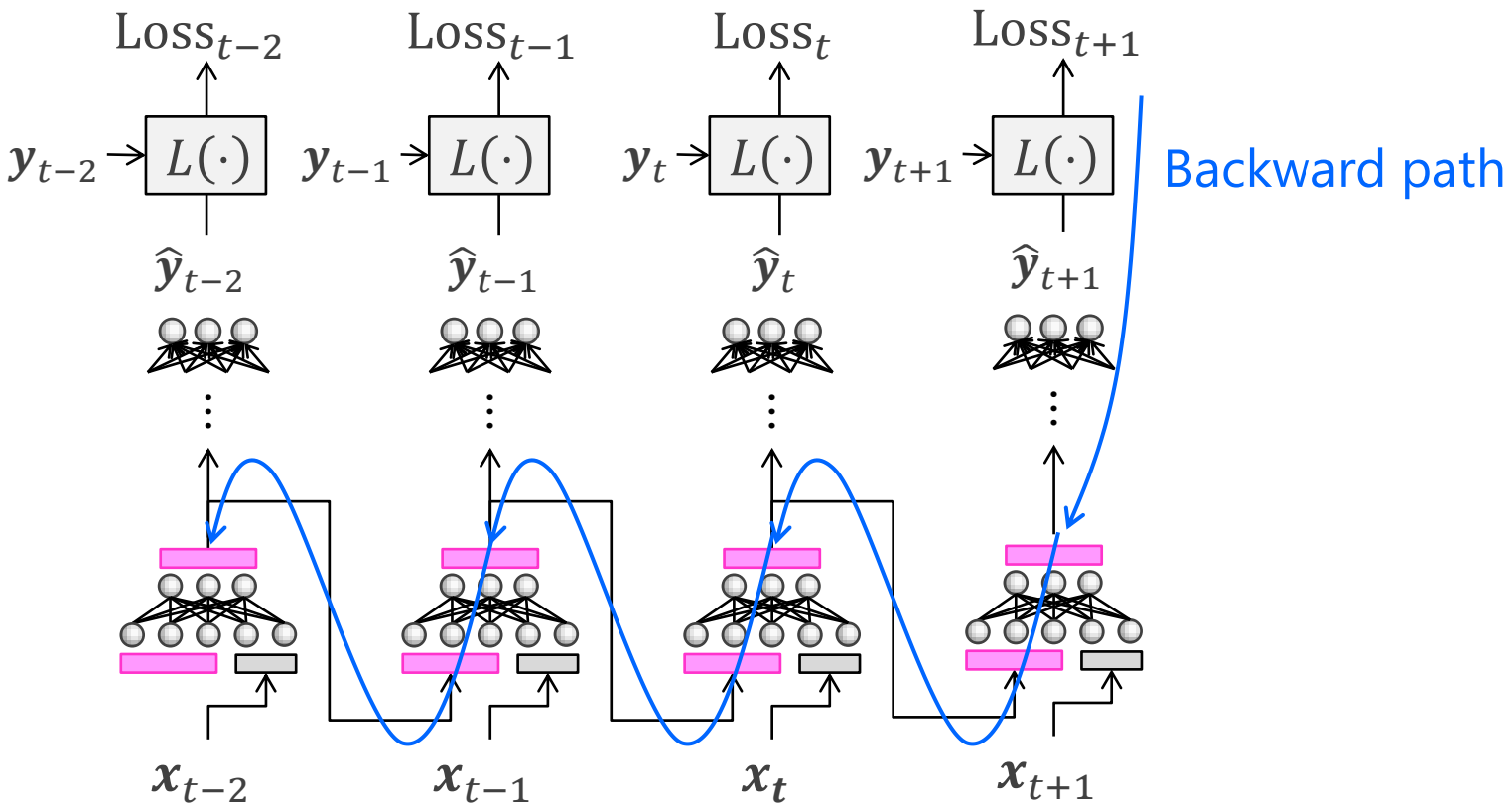
構造情報など(例えば音声の時間構造)の依存性を記憶



BPTT: Back propagation Through Time

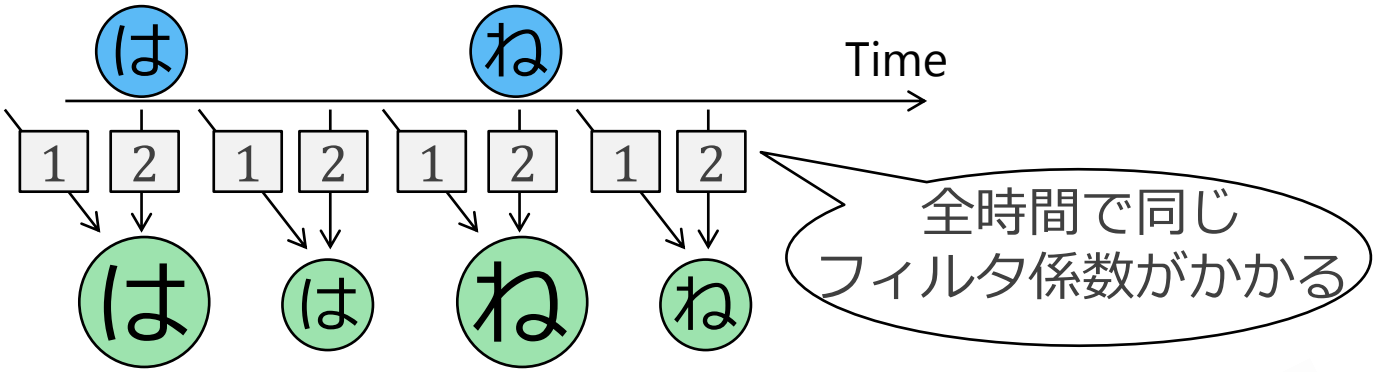
当該時間におけるbackpropagationを，過去の時間に伝播

一定時間でbackwardを打ち切る方法をTruncated BPTTという



CNN (Convolutional NN): 畳み込み構造を持ったNN

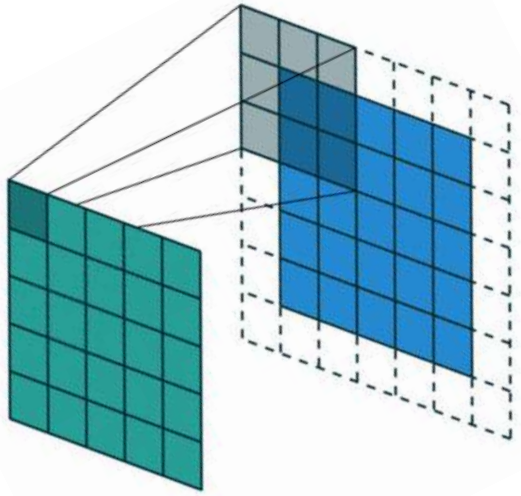
信号処理における畳み込み



畳み込み層：基本的に動作は同じ

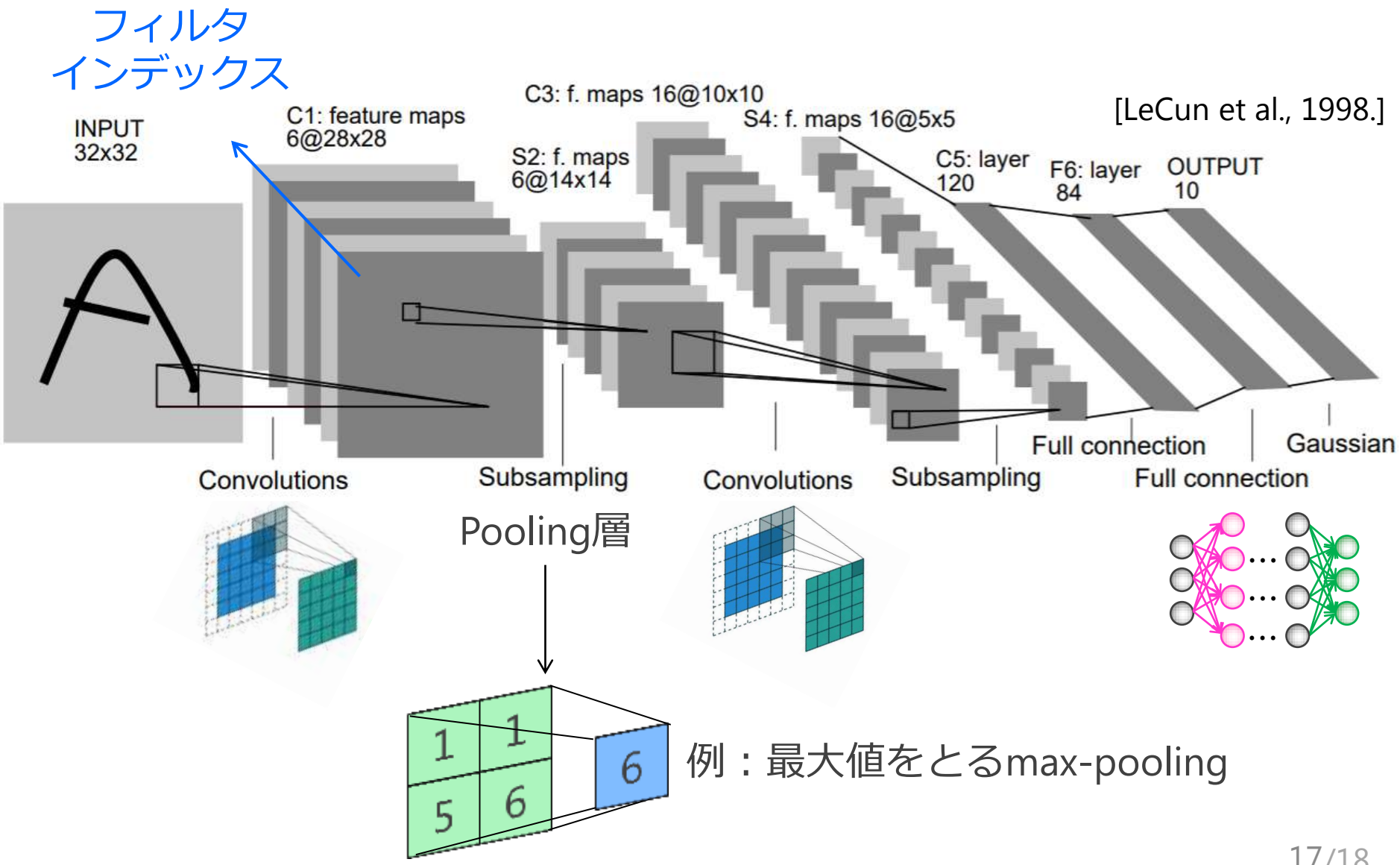
主なパラメータは

- filter size: 右図の灰色部分の形
- #stride: filterの移動幅
- #padding: 端の0埋め数
- #channel: filterの数
  - 異なるフィルタ係数を持った複数のfliterを利用



https://github.com/vdumoulin/conv_arithmetic からgifを引用

CNNの全体構造


フィルタ
インデックス
INPUT
32x32
C1: feature maps
6@28x28
S2: f. maps
6@14x14
C3: f. maps 16@10x10
S4: f. maps 16@5x5
C5: layer
120
F6: layer
84
OUTPUT
10
[LeCun et al., 1998.]
Convolutions
Subsampling
Convolutions
Subsampling
Full connection
Full connection
Gaussian
Pooling層
1
1
5
6
6
例：最大値をとるmax-pooling

自己回帰型CNN

CNNを自己回帰モデルとして扱う

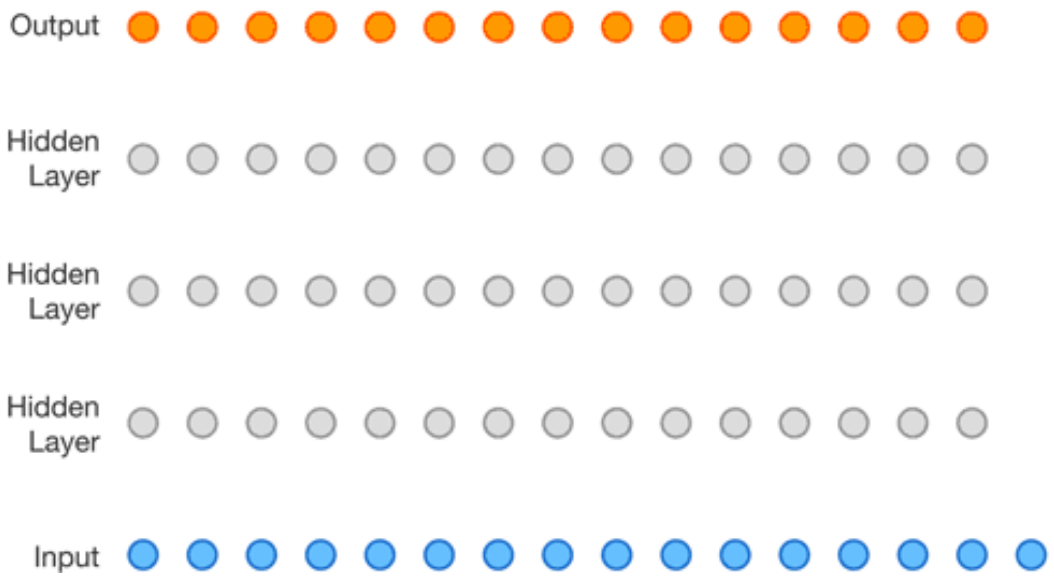
あるステップで生成した出力から，次のステップを推定

→ 信号処理の自己回帰 (エコーやハウリングなど) と同じ

系列を扱うRNNと違い，ステップごとに並列化して学習可能

WaveNet (PixelCNNの派生) [Oord et al., 2016.]

これまでに生成した波形から，次の波形を生成



https://deepmind.com/blog/wavenet-generative-model-raw-audio/ から引用

Deep Generative Model

Deep generative model (深層生成モデル)

Deep generative modelとは

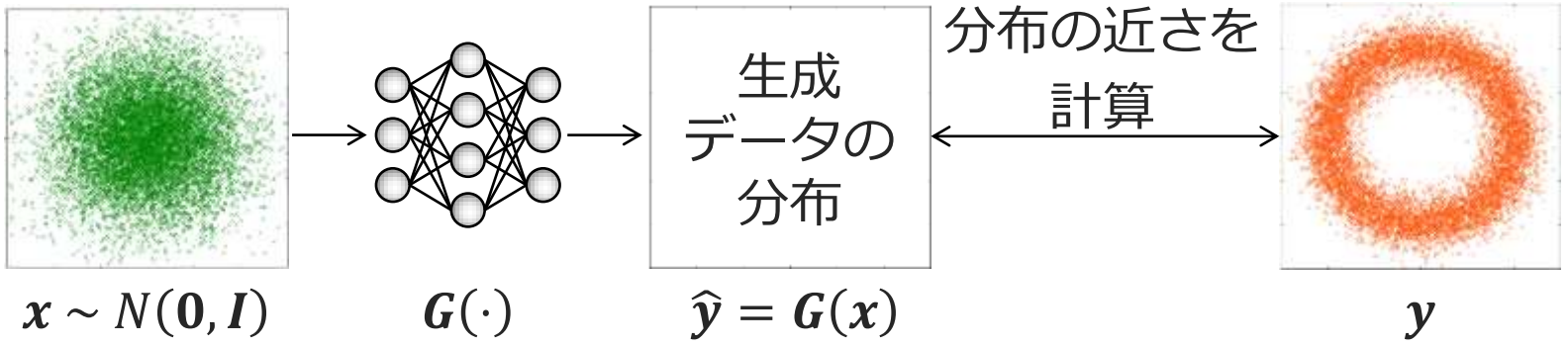
DNNを使ってデータの生成分布を表現するモデル

前述の自己回帰型CNNも，これに相当

ここでは，分布変形に基づく方法を紹介

既知の確率分布を観測データの分布に変形

生成データ $\hat{\boldsymbol{y}}$ の分布と観測データ $\boldsymbol{y}$ の分布が似るようにDNNを学習

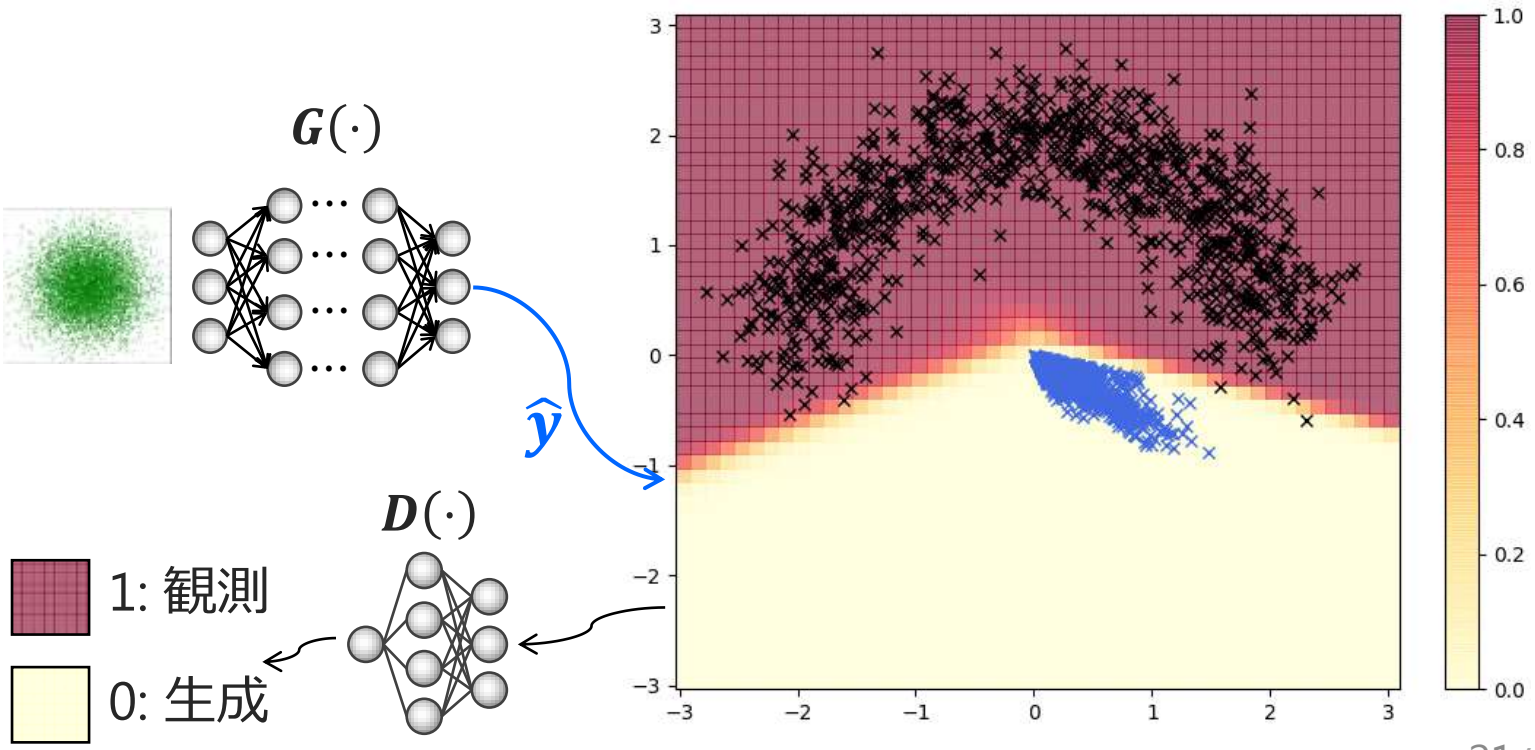


Generative Adversarial Network (GAN): 分布間距離の最小化

Generative adversarial network [Goodfellow et al., 2014.]

分布間の近似 Jensen-Shannon divergence を最小化

$\boldsymbol{G}(\cdot)$と，観測／生成データを識別する識別モデル $\boldsymbol{D}(\cdot)$ を敵対

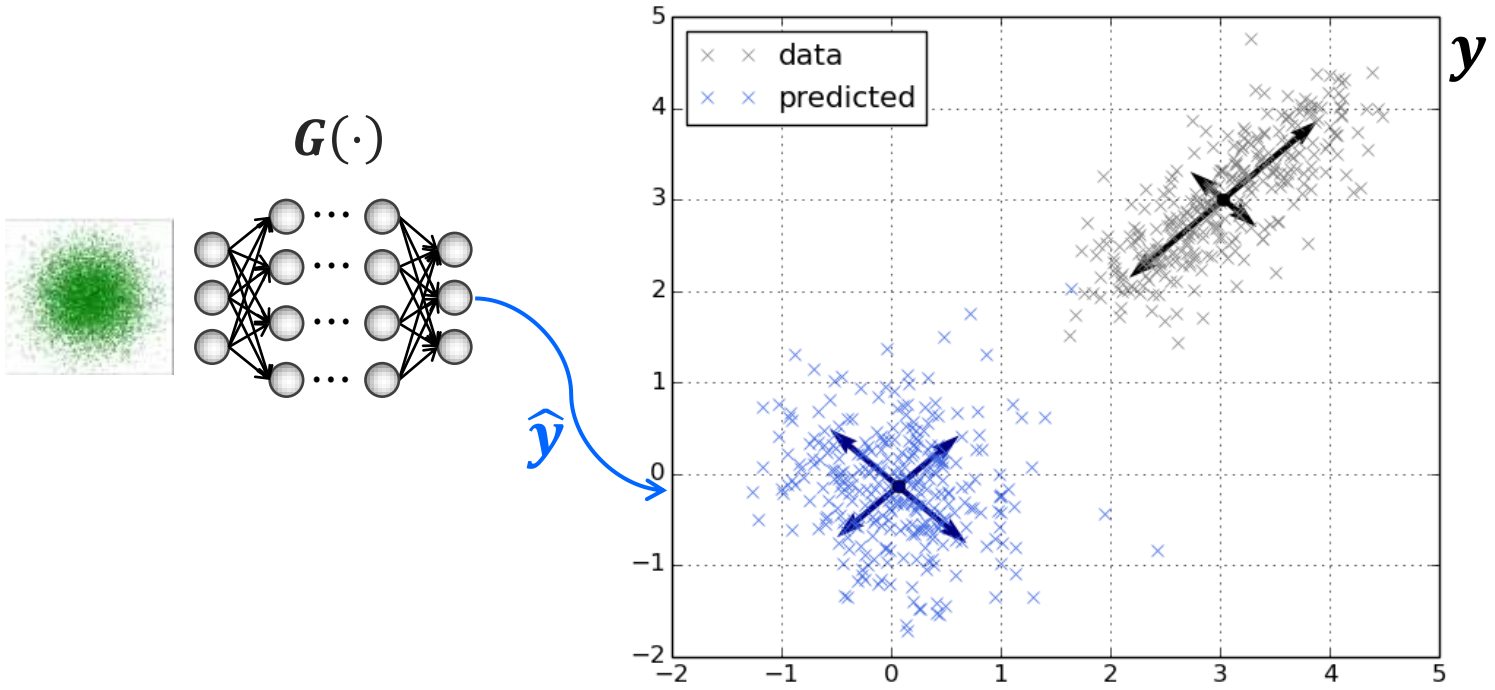


Moment Matching Network (MMN): モーメント間距離の最小化

Moment matching network [Li et al., 2015.]

分布のモーメント (平均，分散，…) 間の二乗距離を最小化

実装上は，グラム行列のノルムの差を最小化



まとめ

ここまでのまとめ

深層学習を深く学習するための基礎を紹介

基礎構造

- Feed-Forward neural networks (FFNN)
- Recurrent neural networks (RNN) … LSTMなど
- Convolutional neural networks (CNN) … WaveNetなど

Deep generative models

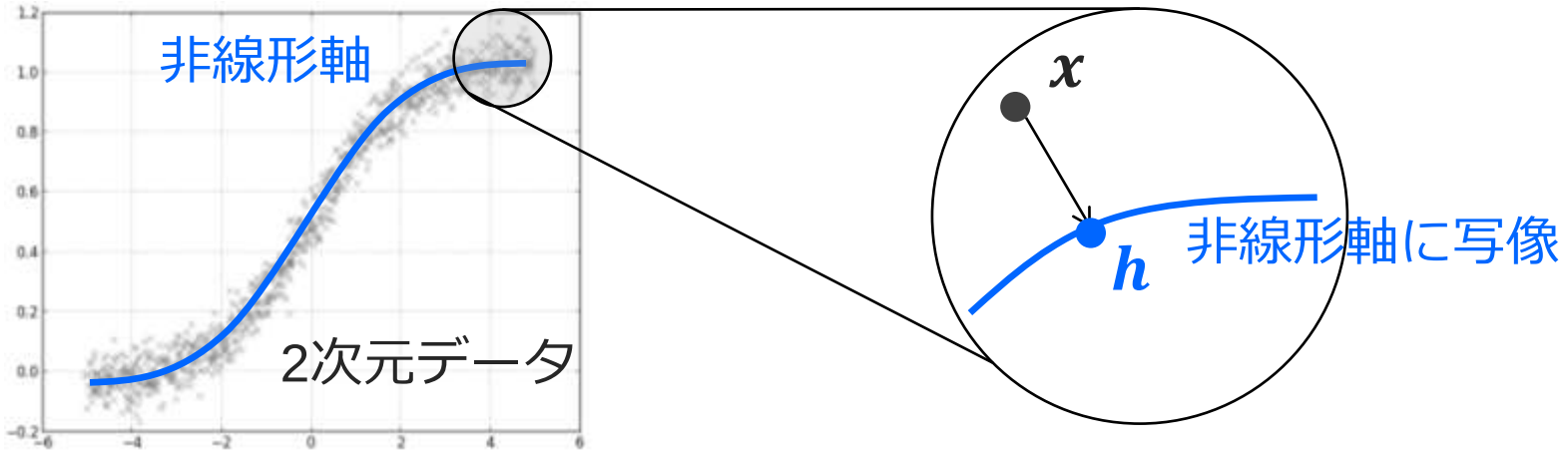
- Generative adversarial networks (GAN) … 敵対的学習
- Moment-matching networks (MMN)

付録

Auto Encoder (AE)

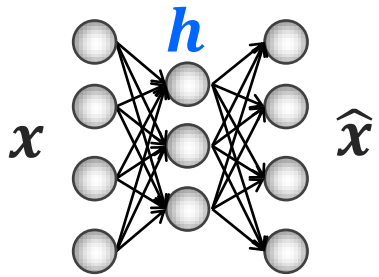
Auto-Encoder (AE): 特徴量の次元圧縮

非線形の軸を引いて，特徴量の次元を削減



Auto-Encoder: 元のデータを復元するように学習

- $\boldsymbol{h} = \boldsymbol{E}(\boldsymbol{x})$: encoder, $\hat{\boldsymbol{x}} = \boldsymbol{D}(\boldsymbol{h})$: decoder

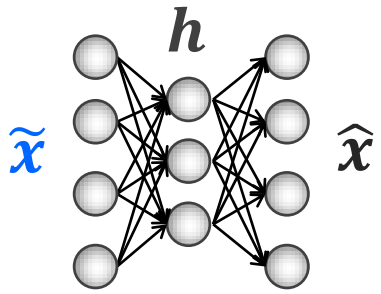


$$\boldsymbol{L}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = (\hat{\boldsymbol{x}} - \boldsymbol{x})^{\mathsf{T}}(\hat{\boldsymbol{x}} - \boldsymbol{x})$$
$$\hat{\boldsymbol{x}} = \boldsymbol{D}\big(\boldsymbol{E}(\boldsymbol{x})\big)$$

Denoising AE

より頑健な次元圧縮を行うため，入力側にノイズを付与

ノイジーな入力から，元のデータを復元する



$$\boldsymbol{L}(\boldsymbol{x},\hat{\boldsymbol{x}}) = (\hat{\boldsymbol{x}}-\boldsymbol{x})^{\mathsf{T}}(\hat{\boldsymbol{x}}-\boldsymbol{x})$$
$$\hat{\boldsymbol{x}} = \boldsymbol{D}\big(\boldsymbol{E}(\tilde{\boldsymbol{x}})\big)$$

どんなノイズを加える？

Drop: ランダムに，使用する次元を減らす

- $\tilde{\boldsymbol{x}} = [1,1,0,0,1,0,1]$^T $\circ \boldsymbol{x}$ ($\circ$ は要素積)

Gauss: ガウスノイズを付与する

- $\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{N}(\boldsymbol{0}, \lambda \boldsymbol{I})$ (λ は分散)

目次

- データの認識・識別問題：深層学習の基礎
  - 基礎構造
  - Deep generative models
- 実波動を作り出す：統計的音声合成の基礎
  - 音声の生成過程
  - 音声合成・変換
  - 最近の研究

目次

- データの認識・識別問題：深層学習の基礎
  - 基礎構造
  - Deep generative models

- 実波動を作り出す：統計的音声合成の基礎
  - 音声の生成過程
  - 音声合成・変換
  - 最近の研究

復習 (音声の生成過程・特徴分析)

音声の生成過程

音色の付与

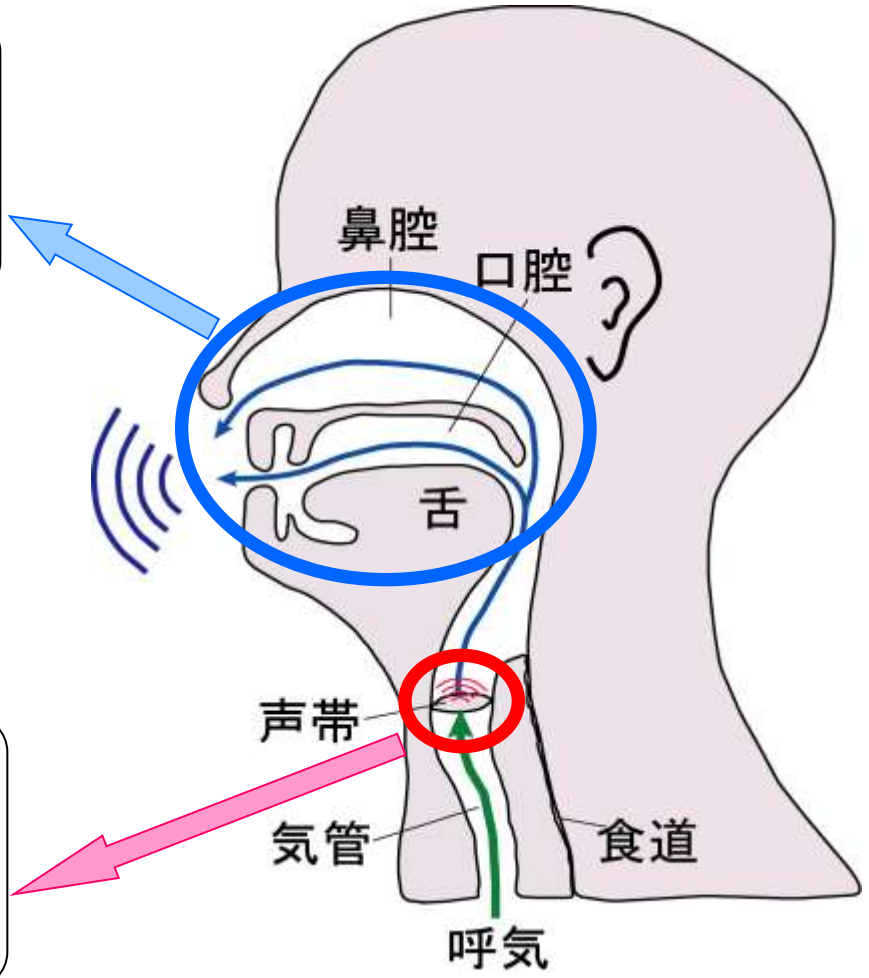
口や舌を動かして，
音色をつける！

声になる！

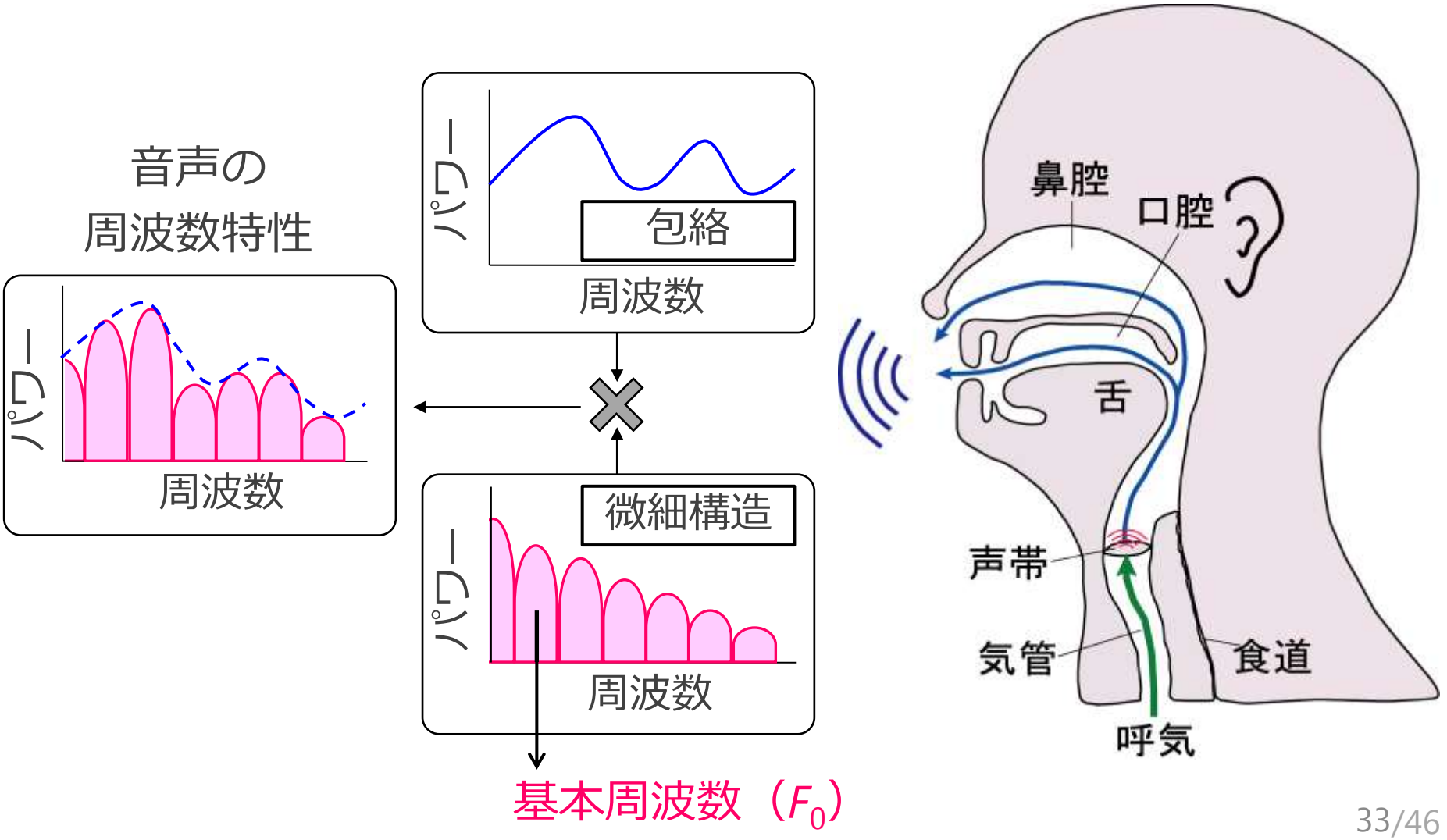
畳み込むと…

音高の生成

声帯を開閉させて，
空気を振動させる！

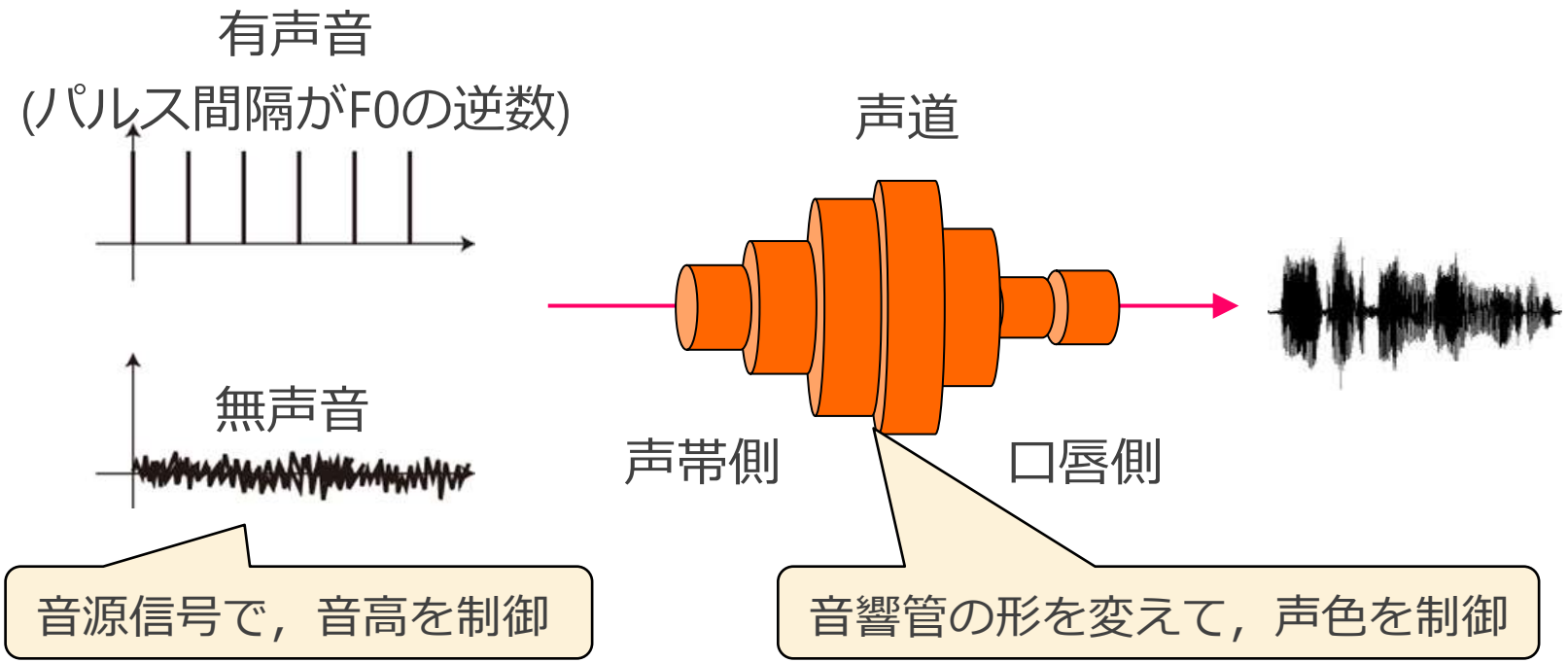


音声のスペクトル構造
（音声のスペクトル構造の2要素）



音源生成と，音響管としての声道

音源信号はインパルス列 or 白色雑音，声道は音響管連接

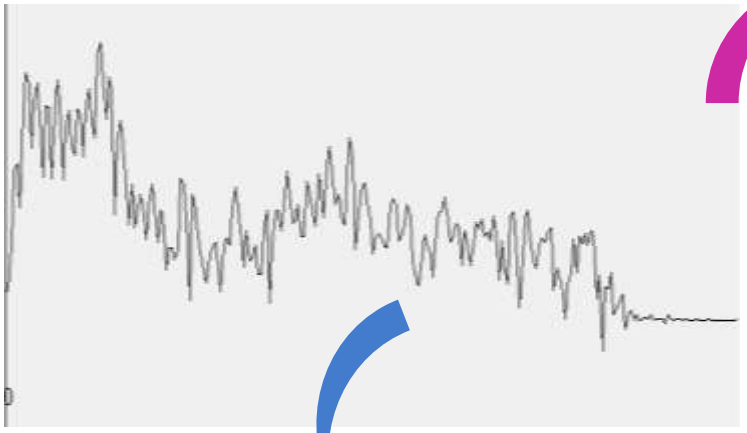


* http://ml.cs.yamanashi.ac.jp/media/20151114/1114slide.pptx から一部引用

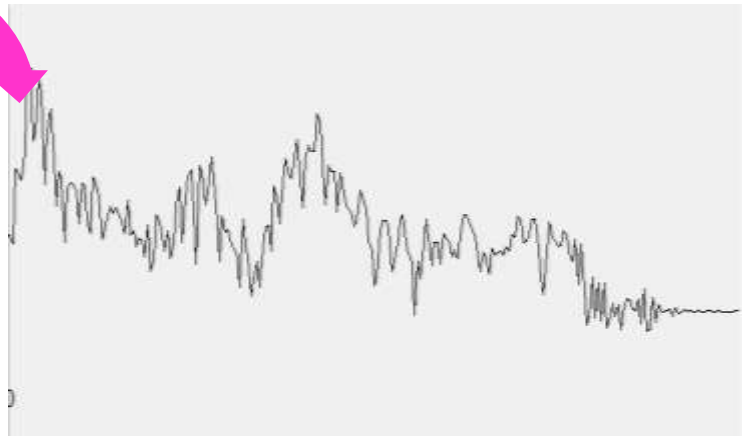
スペクトル構造の例

包絡は変わる
微細構造は変わらない

“あ”, 低いF0

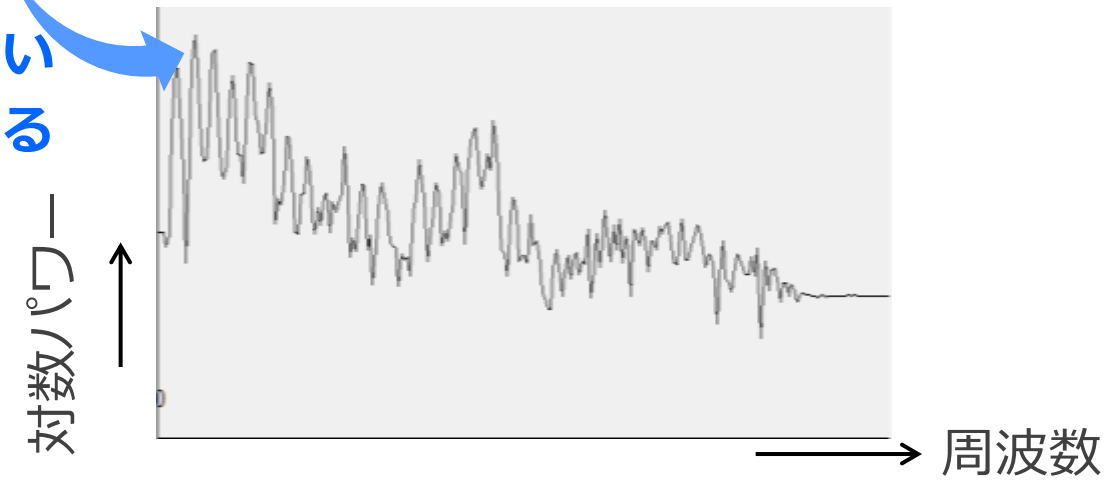


“い”, 低いF0

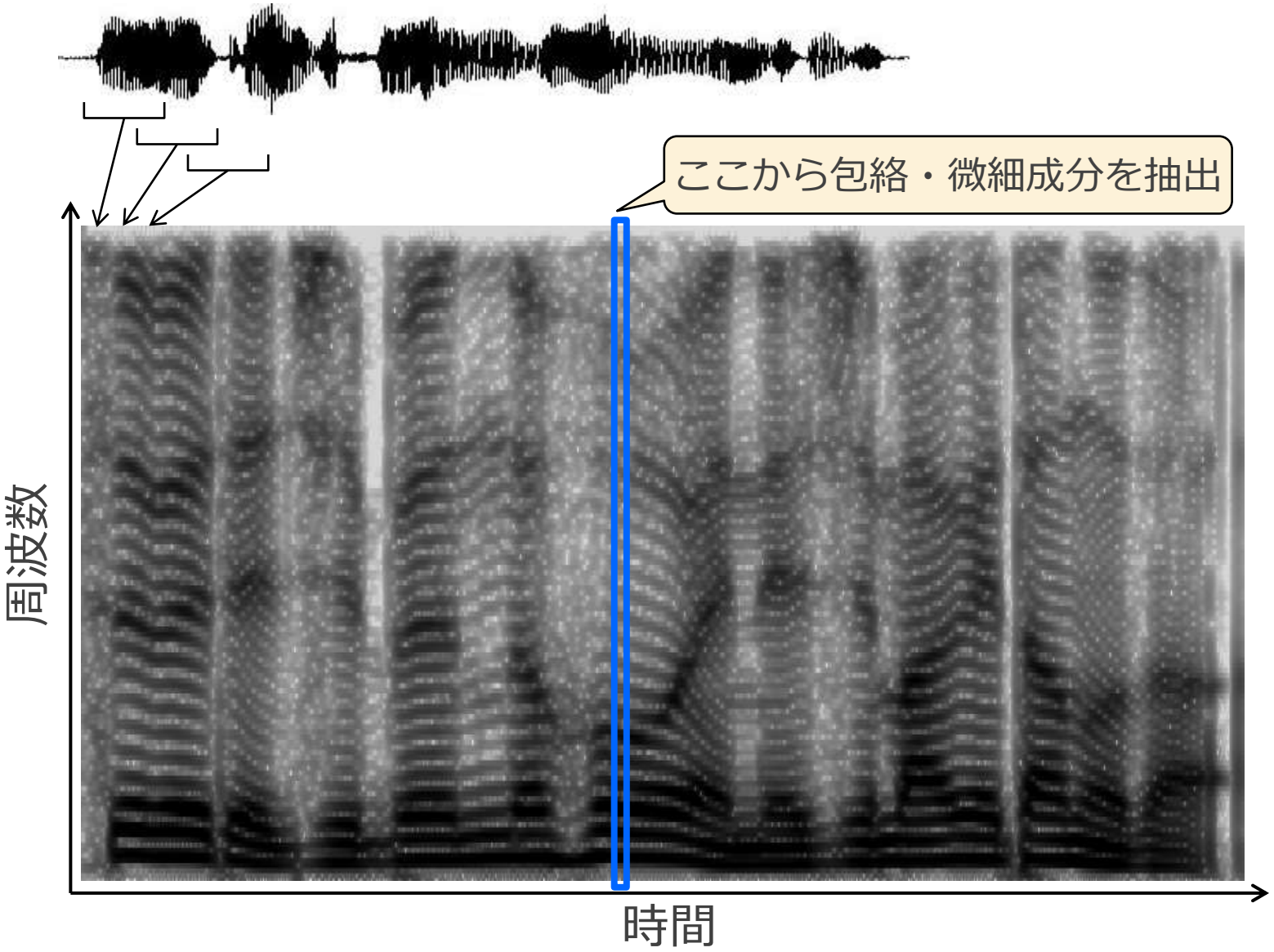


“あ”, 高いF0

包絡は変わらない
微細構造は変わる



音声の特徴分析 (フレーム分析)



音声合成・変換

音声合成：音声を人工的に作り出す技術

狭義の音声合成

テキスト音声合成 (Text-To-Speech: TTS)

広義の音声合成 (-to-speech)**

テキスト音声合成

音声変換 (Voice Conversion: VC)

概念音声合成 (Concept-To-Speech: CTS)

- 概念 → 言語生成 → 音声合成

調音・音響間マッピング

- 調音機構特性と音声の変換

マルチモーダル音声合成

- 動画像などを含む音声合成

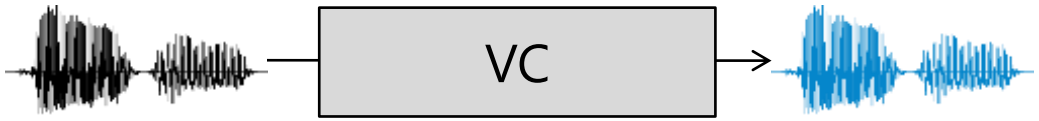
テキスト音声合成と音声変換

テキスト音声合成 (Text-To-Speech: TTS)

テキストなどから音声を合成
人以外のモノのコミュニケーションのため

音声変換 (Voice Conversion: VC)

言語情報を保持したままパラ言語・非言語情報を変換
人の発声制約を超えたコミュニケーションのため



デモ

東京大学は世界一！

昼飯はとんこつラーメンに限る！

音声合成の歴史

1939: Voder (ベル研究所)

その前身はvocoder (voice + coder)

1961: 音声合成による ‘Daisy Bell’ (ベル研究所)

～

~1990: フォルマント音声合成

専門家による音声規則設計

1990~: 波形接続型音声合成

ダイフォン音声合成，単位選択型音声合成

1995~: 統計的パラメトリック音声合成

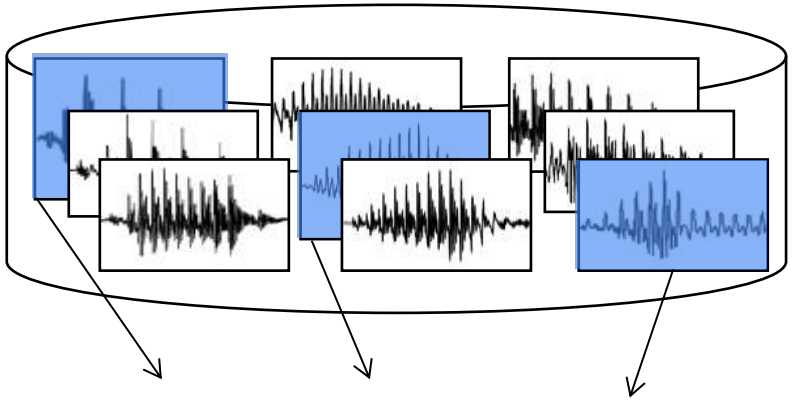
HMM音声合成・DNN音声合成

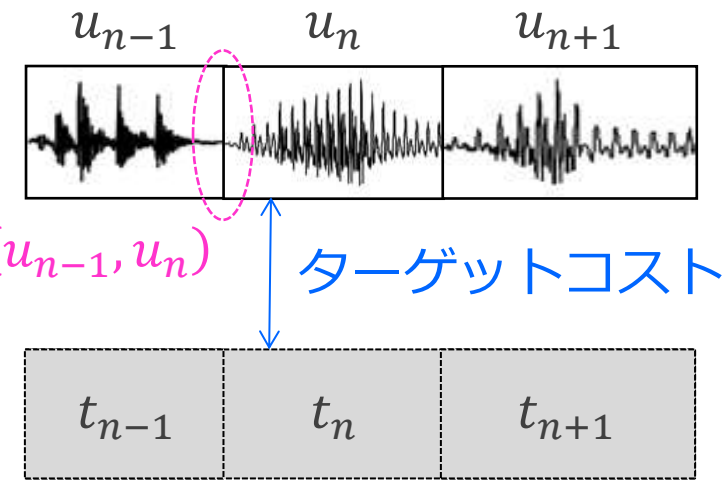
1998~: GMM音声変換・DNN音声変換

事前収録音声コーパスを用いて合成を行う
コーパスベース合成方式

サンプルベース方式 (波形接続型)

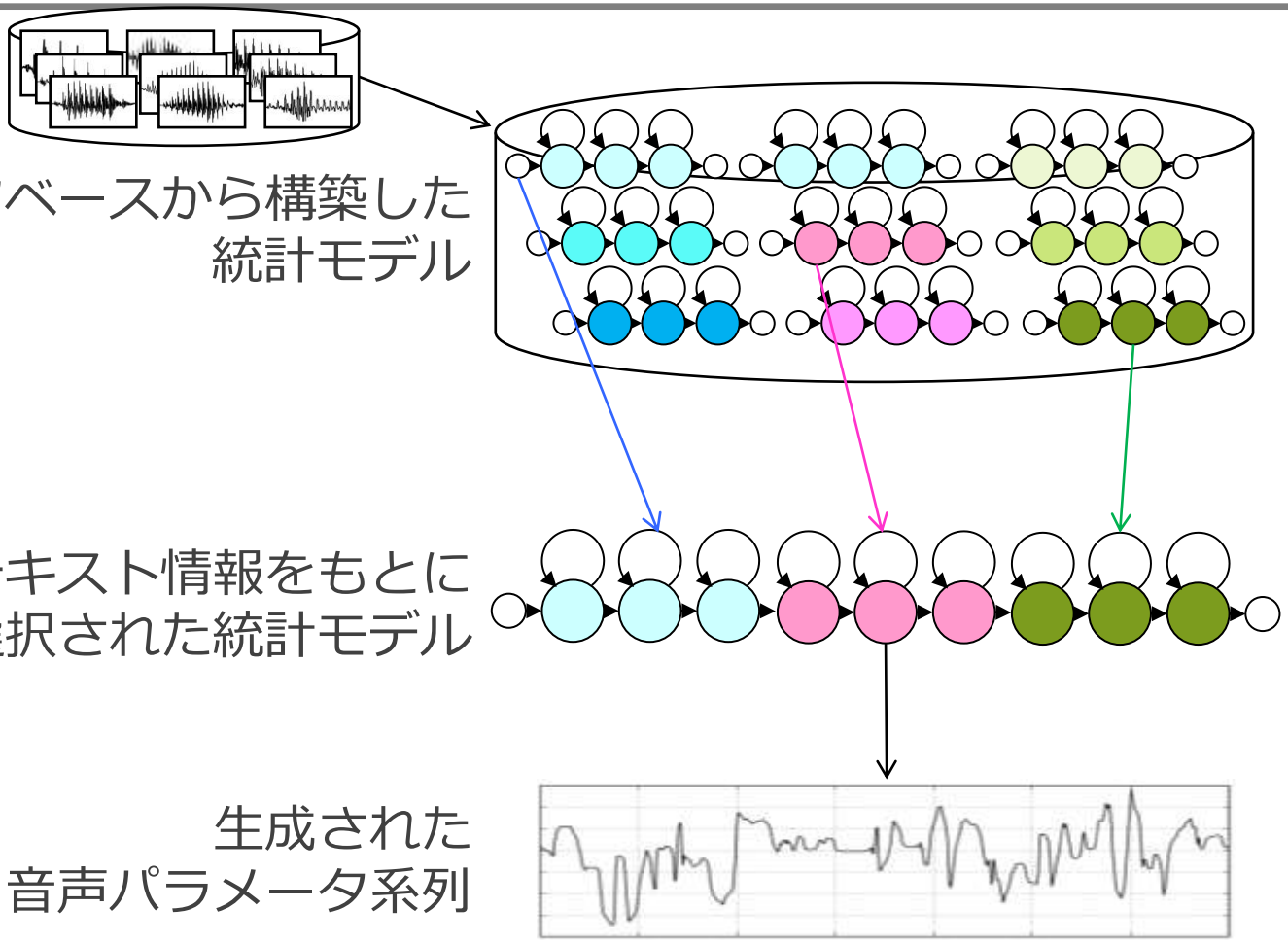
音声データベースにある
音声セグメント





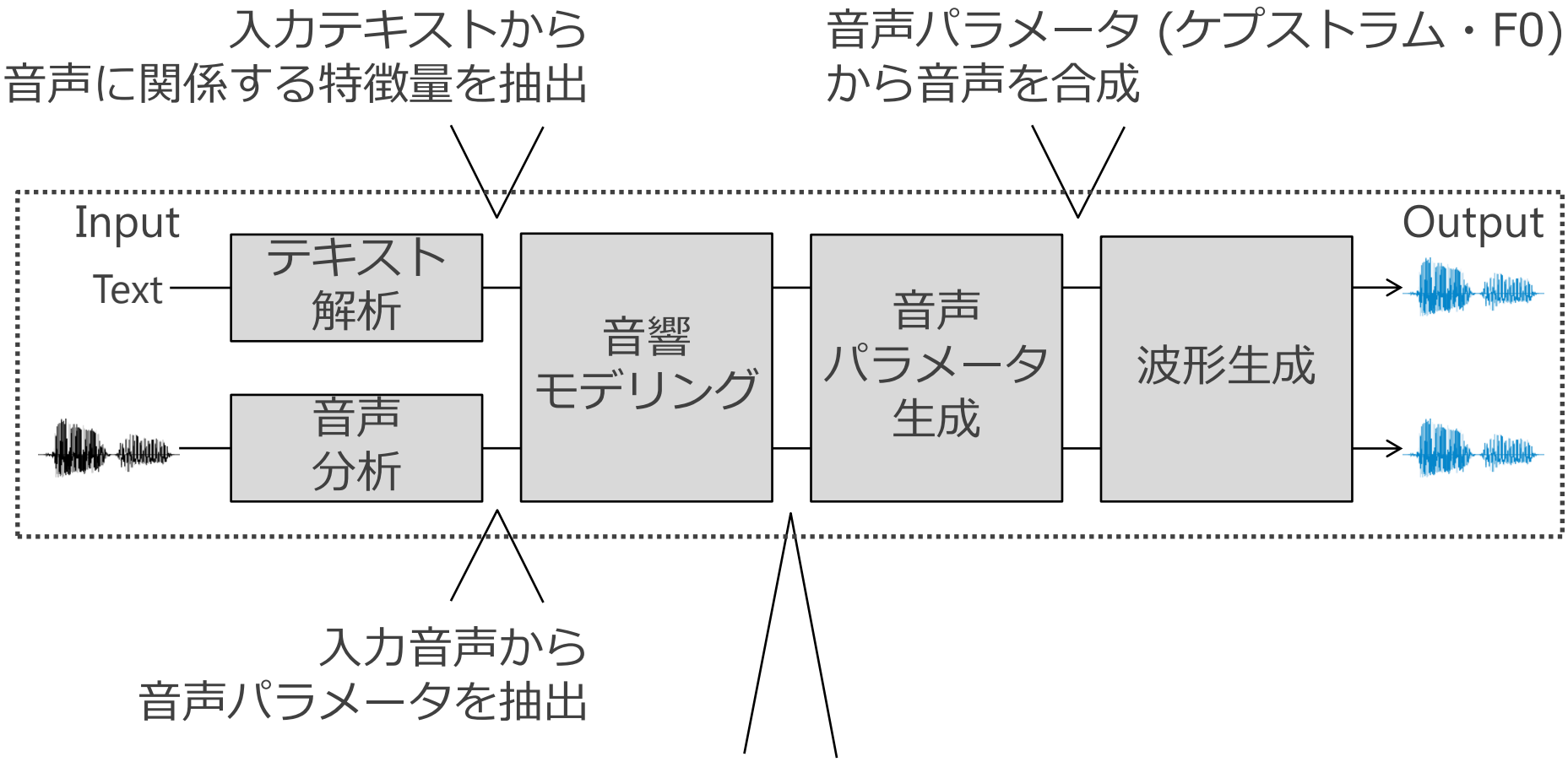
接続コストとターゲットコストの和を最小化するように
音声セグメントを選択

統計ベース方式



音声波形の代わりに統計モデルを保存して
統計モデルから音声パラメータを生成

統計ベース方式の手順


入力テキストから
音声に関係する特徴量を抽出
音声パラメータ (ケプストラム・F0)
から音声を合成
Input
Text
テキスト
解析
音声
分析
音響
モデリング
音声
パラメータ
生成
波形生成
Output
入力音声から
音声パラメータを抽出
テキスト特徴量と音声特徴量 (パラメータ) を対応付け

音声合成のためのテキスト解析

テキストを読み上げたい！

どうやって読んだらいいの？

テキストと音声を結びつける構成要素がいくつかある

①発音・音節

②アクセント・ストレス

③リズム・等時性

①発音・音節

発音

発声の最小単位である音素の違い

音節 (シラブル)

音節 … 言語依存の発声単位 (日本語ならほぼひらがな一つに対応)

- 開音節 … 母音で終わる音節。日本語の”か(k a)”など。
- 閉音節 … 子音で終わる音節。例： 英語の”it (i t)”など。

子音連結 … 同一音節中で連続する子音

- 日本語 … ほとんどCV (C: 子音、V: 母音)
- 英語 … CCCV、CCV、VCC、VCCCなどが頻出
  - straight = stra + ight

②アクセント・ストレス

音声のアクセント・ストレス

言語に依存してスペクトルとF0に現れる

例1: 日本語 (アクセント)

高いF0

わ た し は と しょ か ん へ い き ま し た。

低いF0

例2: 中国語 (アクセント: 四声)

我 去 图 书 馆

＼ ＼ ／ ー ＼ F0の変化

例3: 英語 (ストレス)

ストレス

I went to the library to study for the exam.

③リズム・等時性

音声の等時性

言語に依存した音声的単位が、時間的に等間隔に現れる

例1: 日本語 (モーラ等時性)

わ た し は と しょ か ん へ い き ま し た。

例2: 中国語 (シラブル等時性)

各点は一定時間周期で現れる

我　去　图　书　馆

例3: 英語 (ストレス等時性)

I went to the library to study for the exam.

アクセントは誰が決めてる?：NHKアクセント辞典

2016年に改定！

18年ぶり6回目。初版は1943年

前回から何が変わった？

[太田 他, 2016.]

” ついに「ク＼マ」が出た！”

”クマが出た” のアクセントは？

図２ 「熊」の NHK アナウンサー調査の結果（語別）

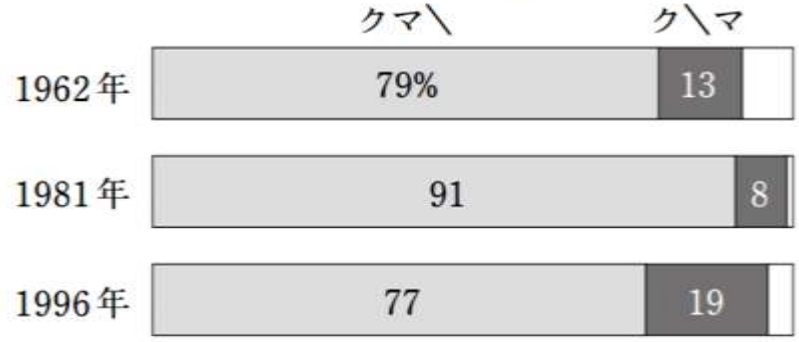
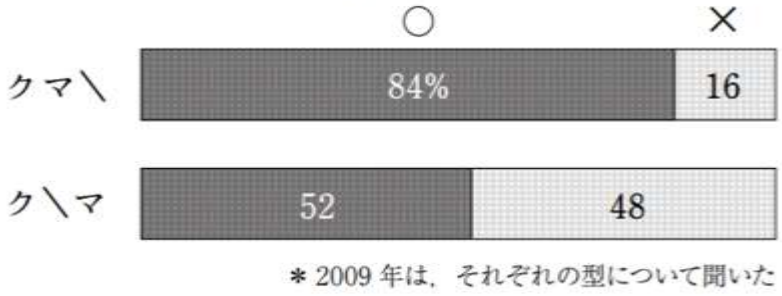


図３ 「熊」の NHK アナウンサー調査の結果（2009 年／型別）



＊2009 年は，それぞれの型について聞いた

外来語は平板化

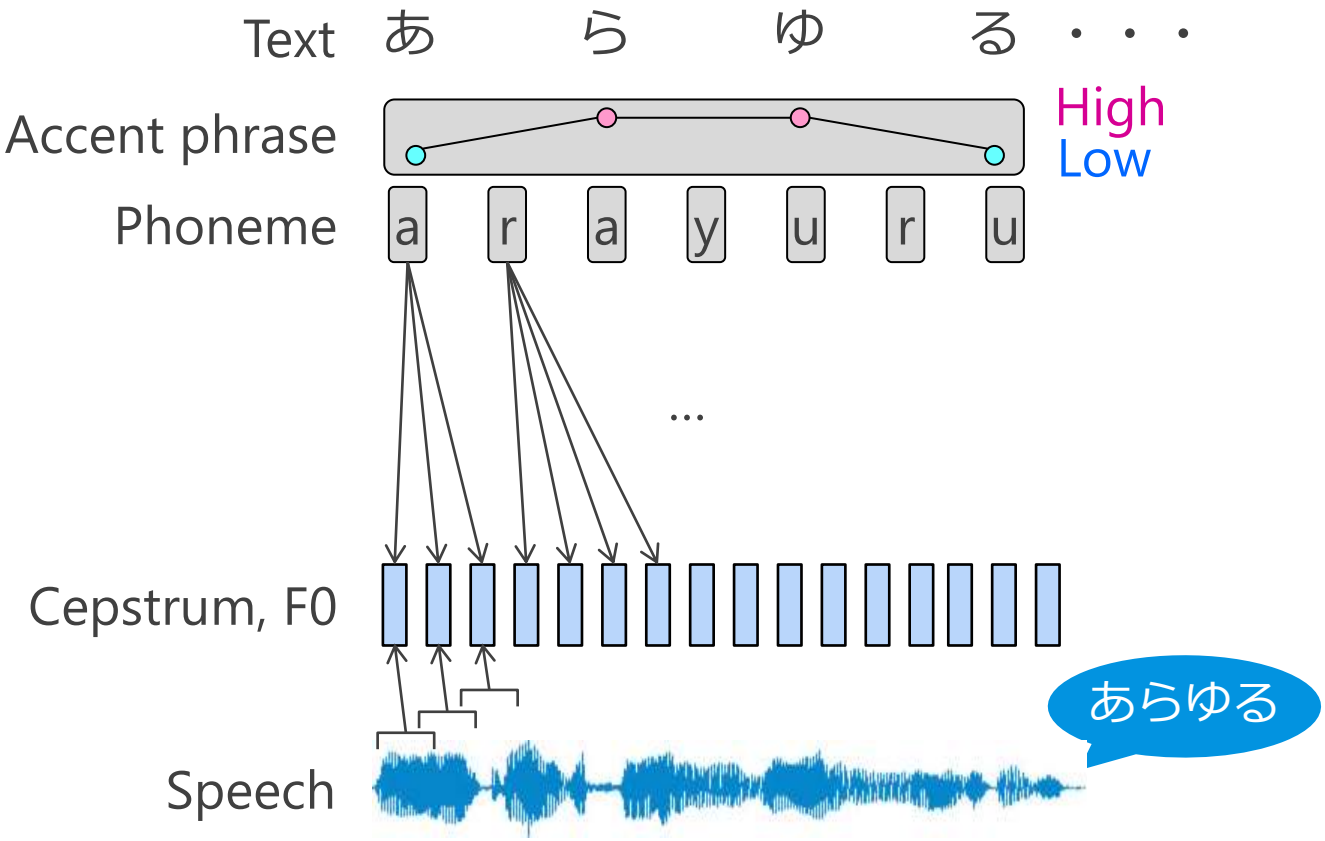
複合語 (歩み＋寄るなど) は平板から起伏化

などなど

時系列の対応付け (text-to-speech)

通常，テキスト特徴量系列と音声特徴量系列の長さは異なる

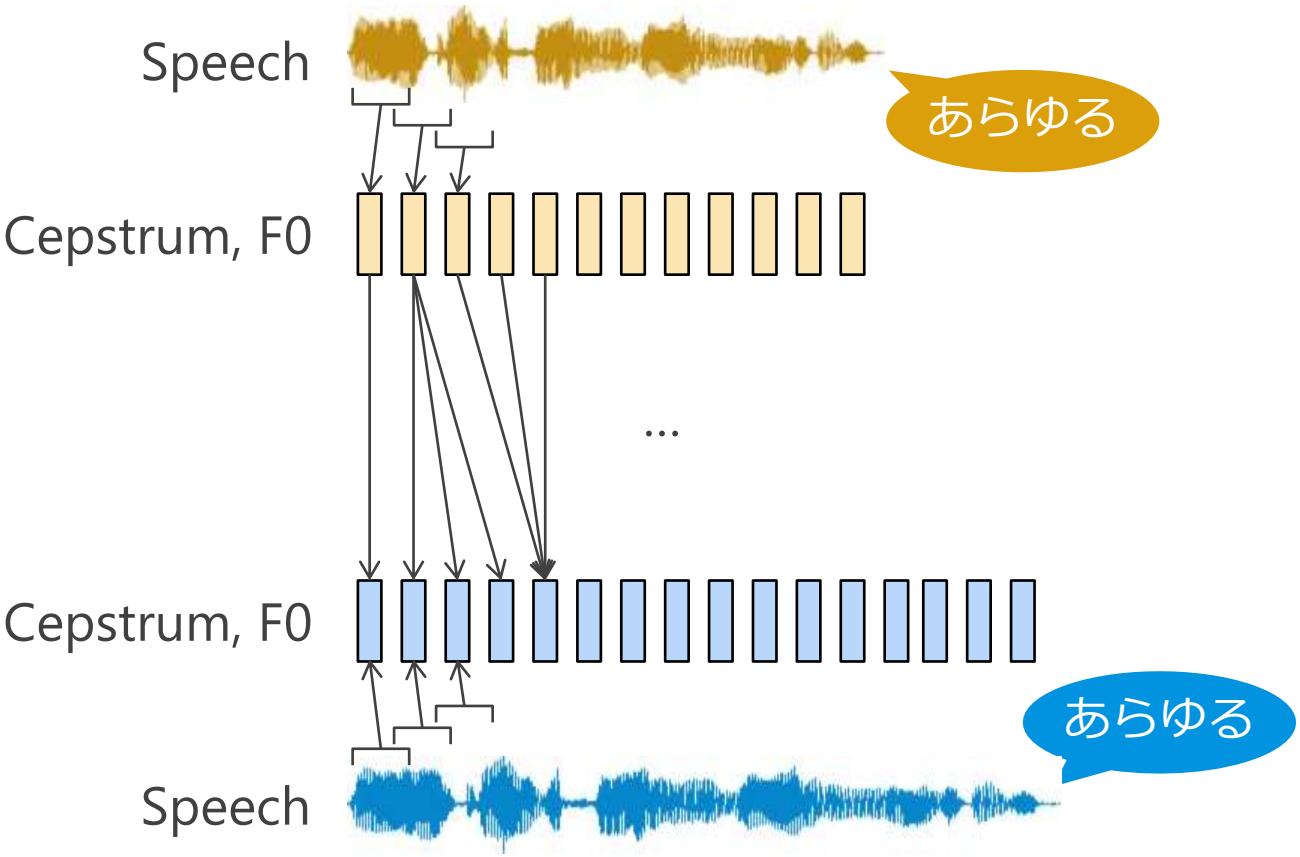
(音声認識などによる) アライメントを実施して揃える



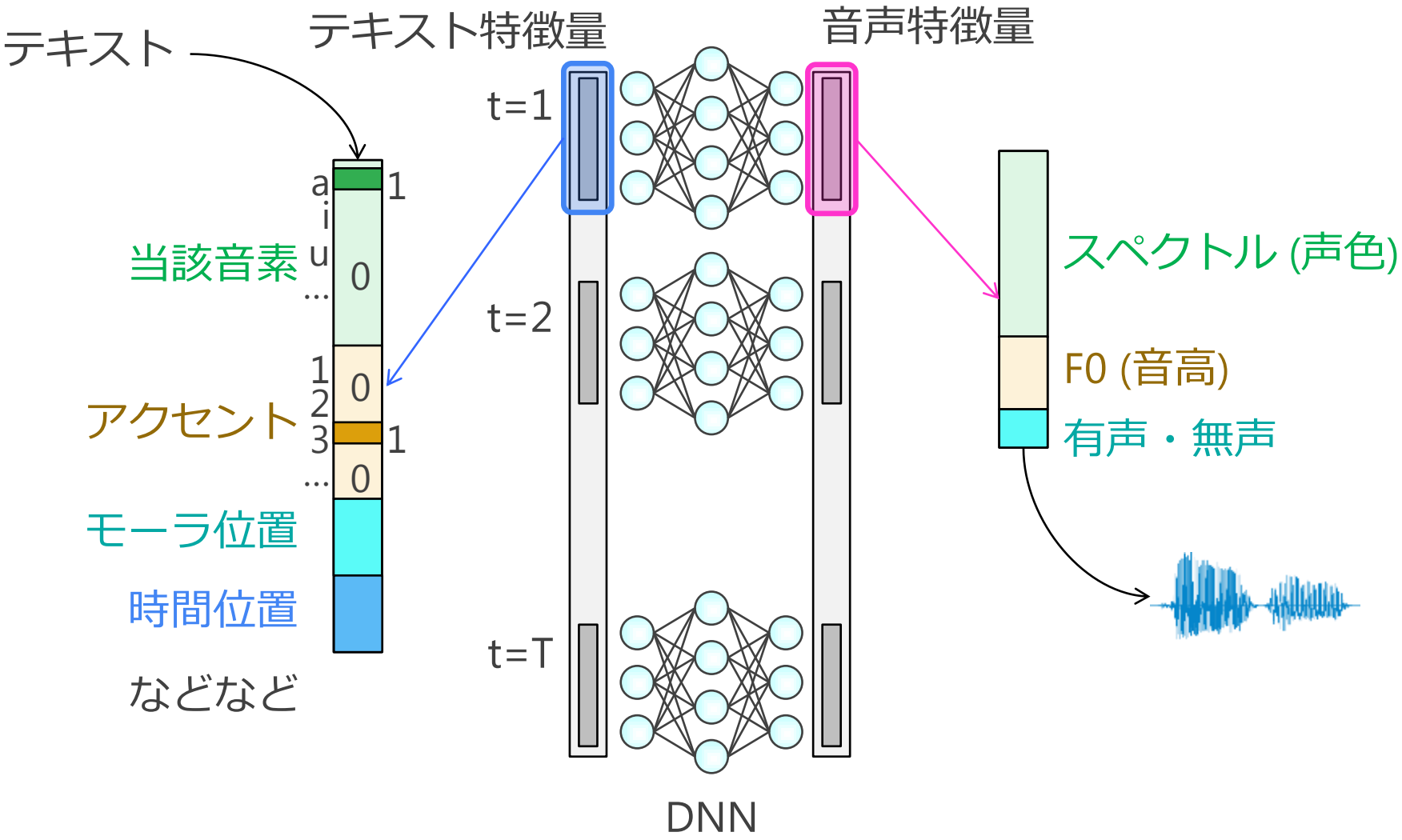
時系列の対応付け (voice conversion)

(例えば異なる話者による) 音声も，系列長が異なる

動的時間伸縮 (DTW) などにより揃える

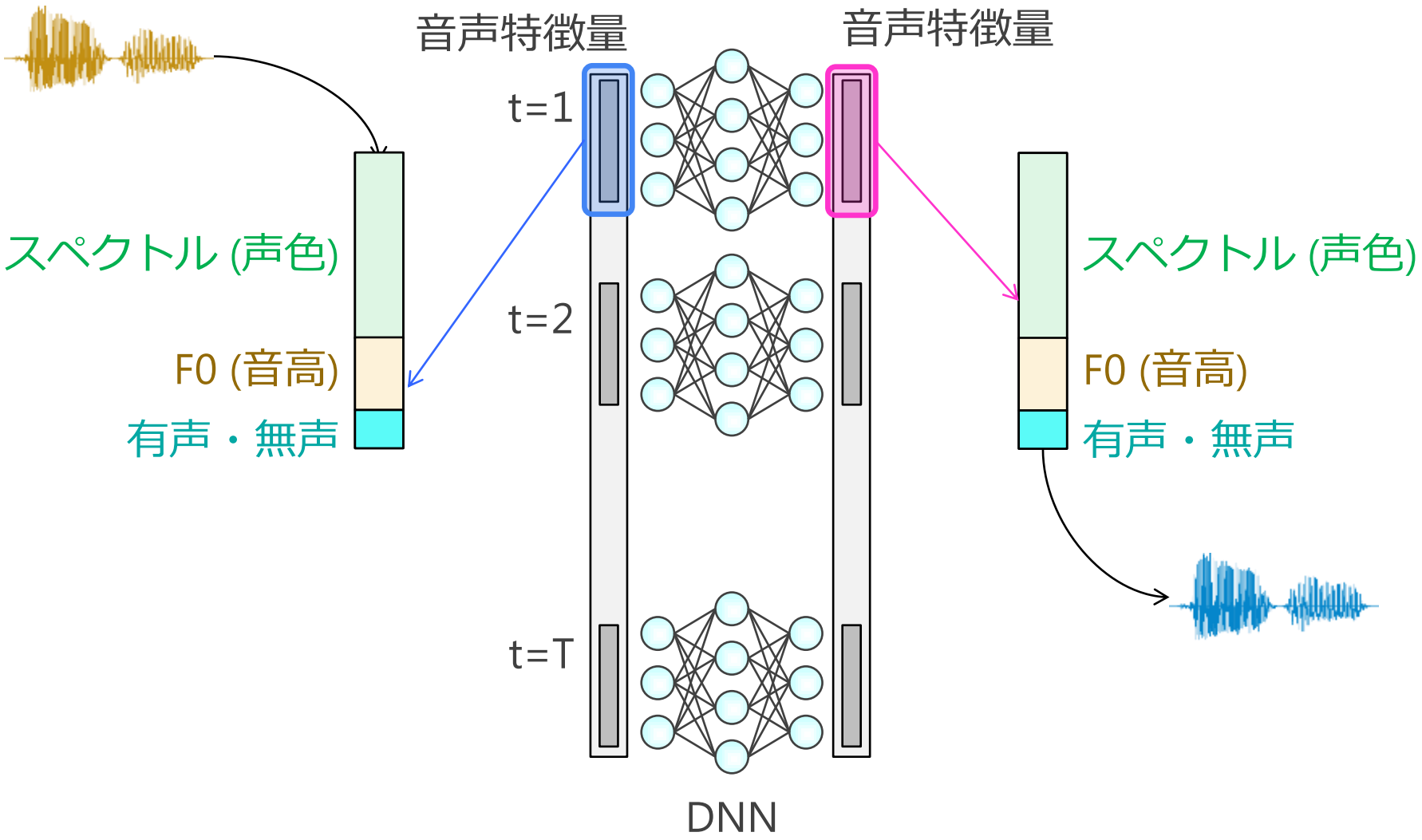


Text-to-speechでの利用

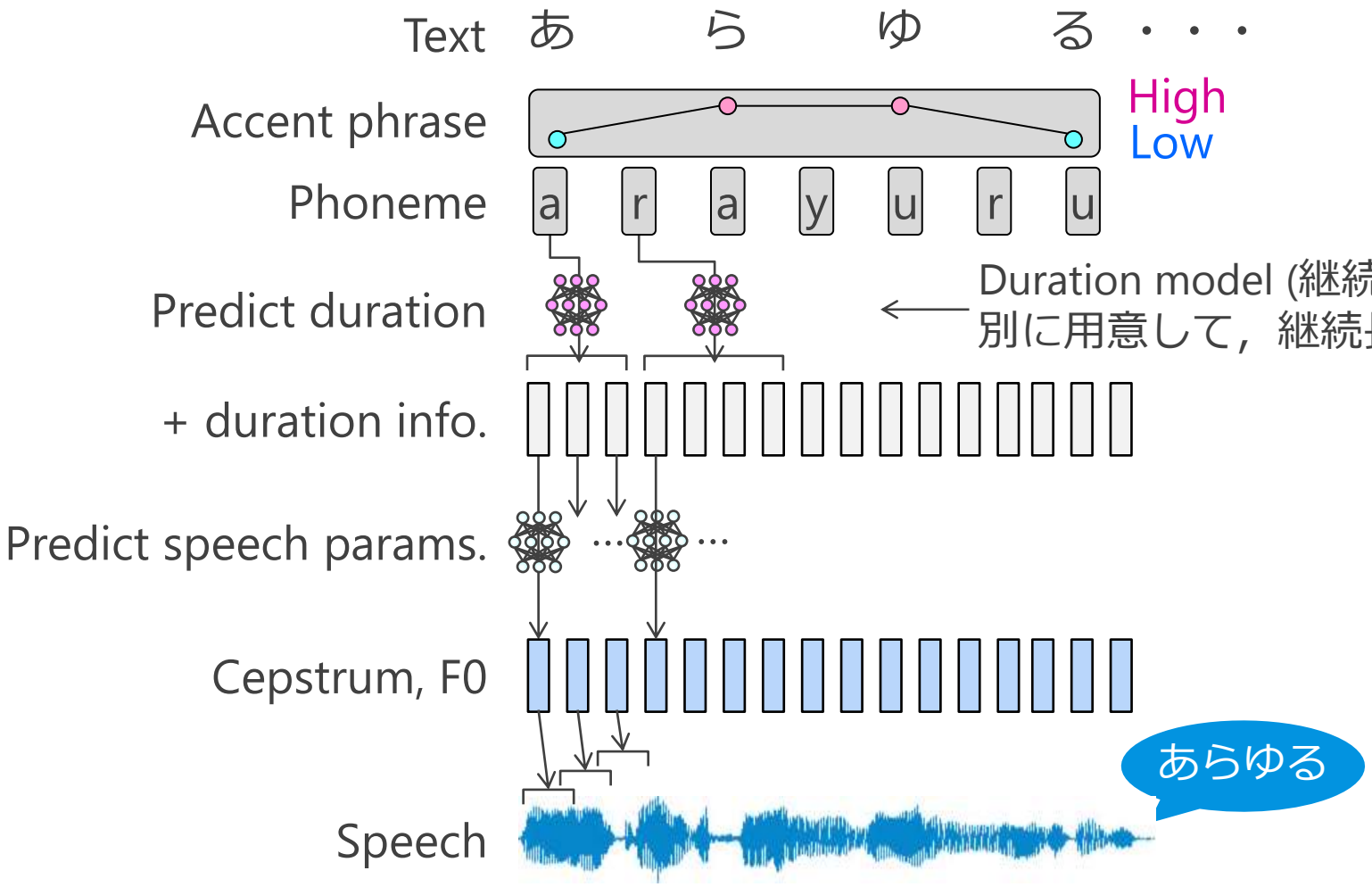


DNNは自然音声特徴量との二乗誤差を最小化するように学習

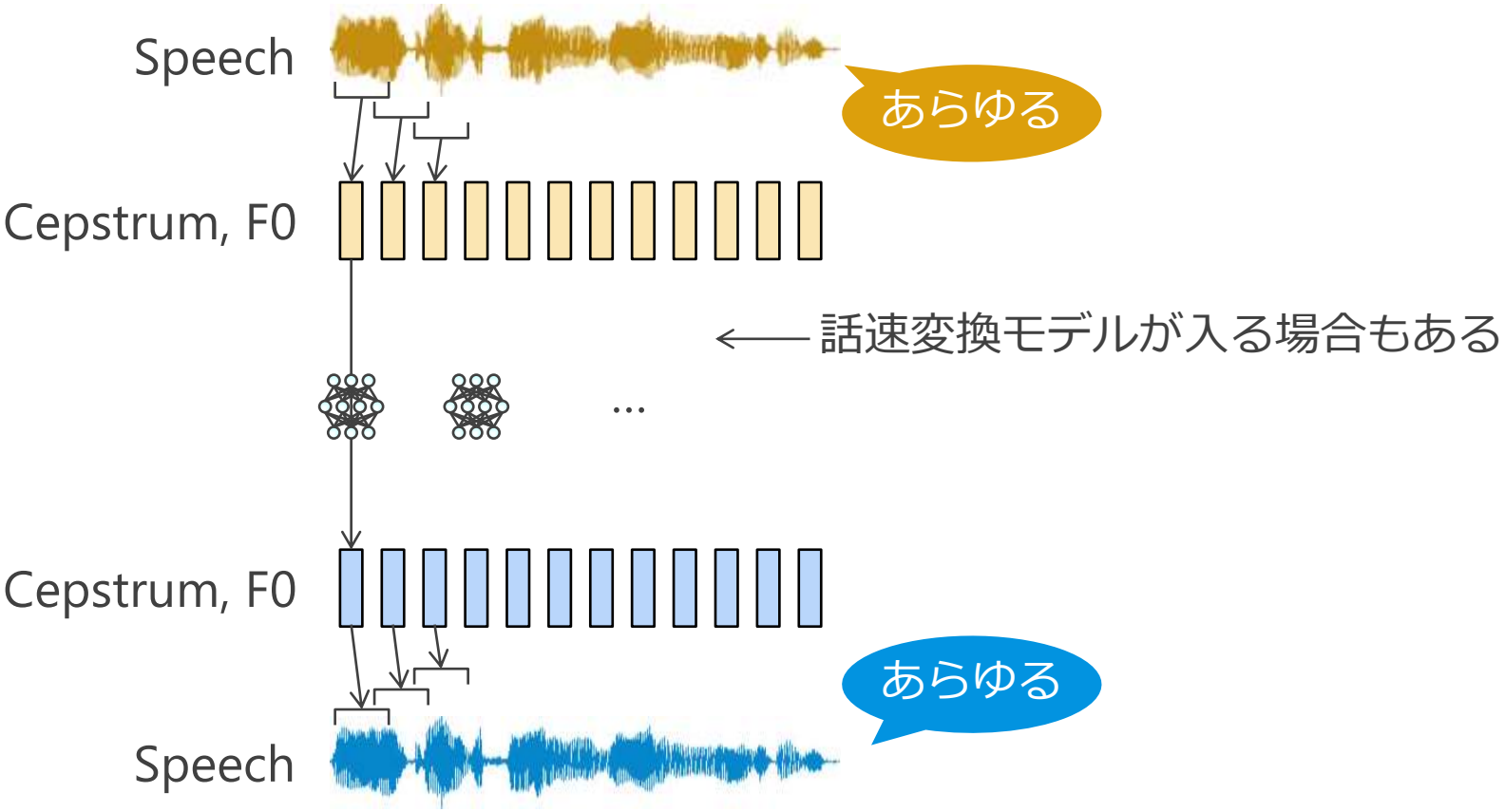
Voice conversionでの利用



Text-to-speechにおける生成手順



Voice conversionにおける生成手順



色々な発展技術

音声パラメータの時間変化を考慮したい

- 動的特徴量(各時間の特徴量差分) … $\Delta y_t = \frac{1}{2} y_{t+1} - \frac{1}{2} y_{t-1}$ (tは時間)
- リカレント構造DNN (RNN, LSTM)

DNN以外の方法

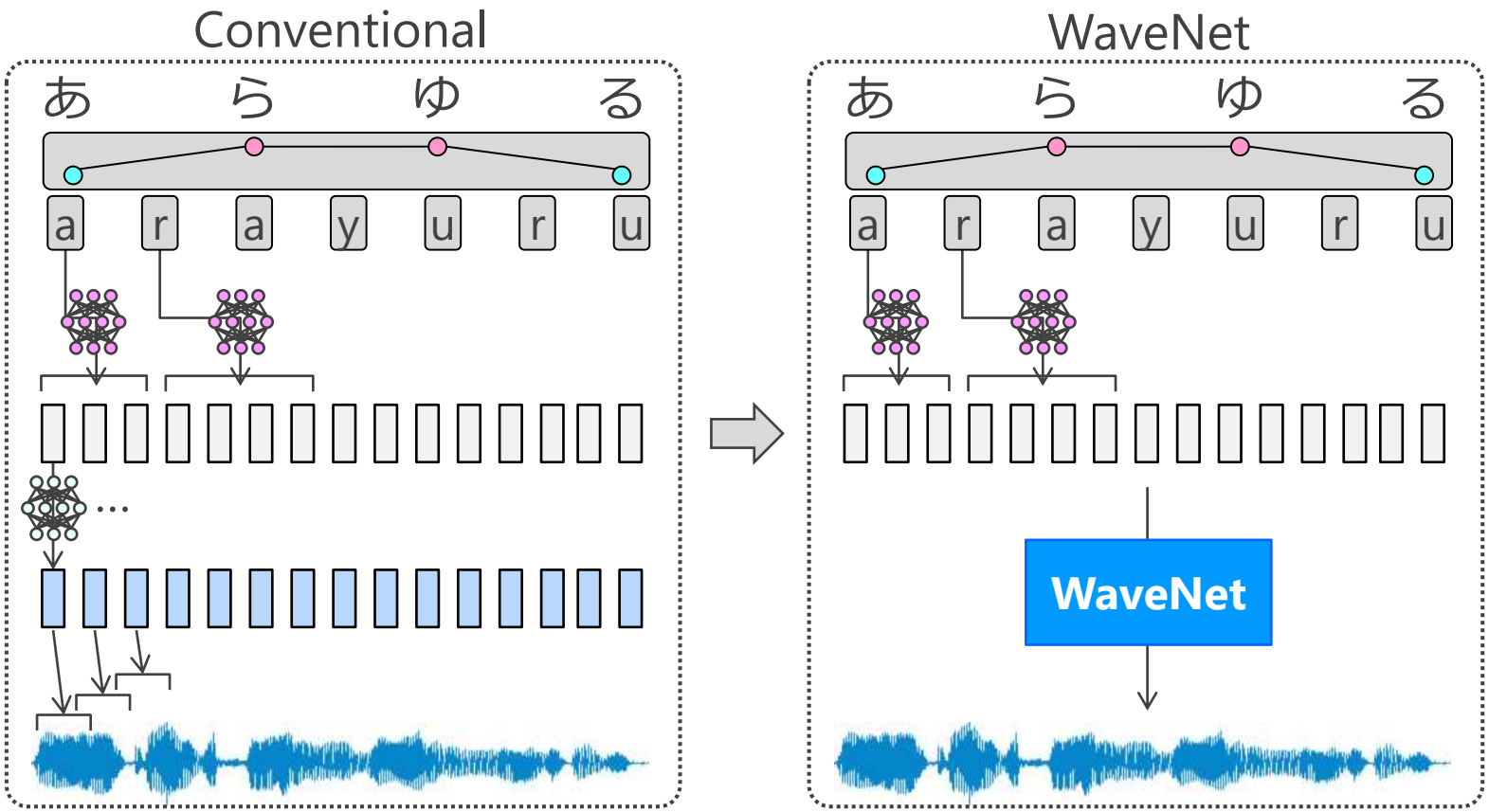
- HMM (隠れマルコフモデル)
- GMM (混合正規分布モデル)
- GPR (ガウス回帰過程)
- NMF (非負値行列因子分解)
- サンプルベースとのハイブリッド

音声合成の最近の研究

WaveNet

2016年，波形を直接生成するニューラルネットが提案された

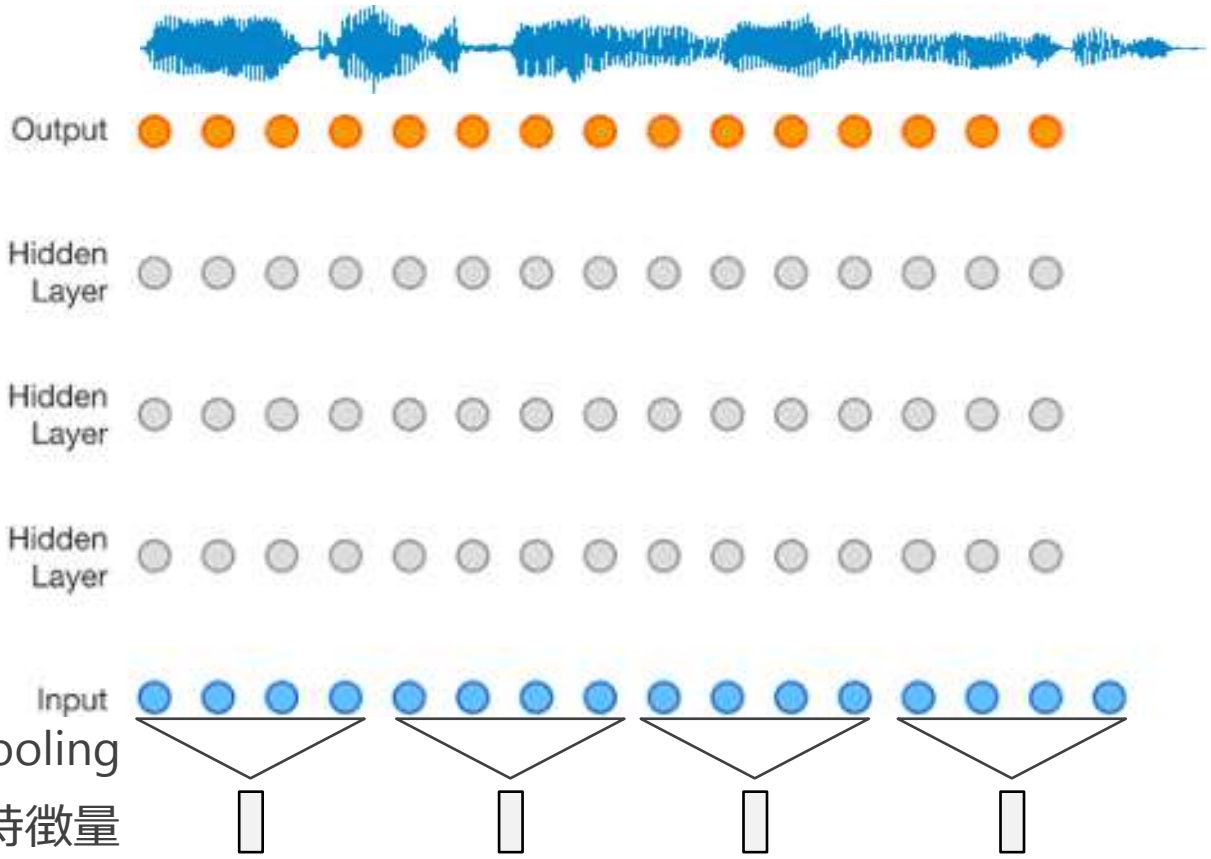
音声特徴量生成，波形生成，フレーム分析を取り除き，
サンプルごとに波形を生成する単一のニューラルネットワーク



WaveNet音声合成

[Oord et al., 2016.]

- 波形の離散化 … float型信号を256 (65536) 段階に圧縮して処理
- 自己回帰モデル … 一つ前の結果から現在の結果が分かる，LPCと類似
- Dilated convolution … ユニットを飛ばした畳み込みで受容野を大きく

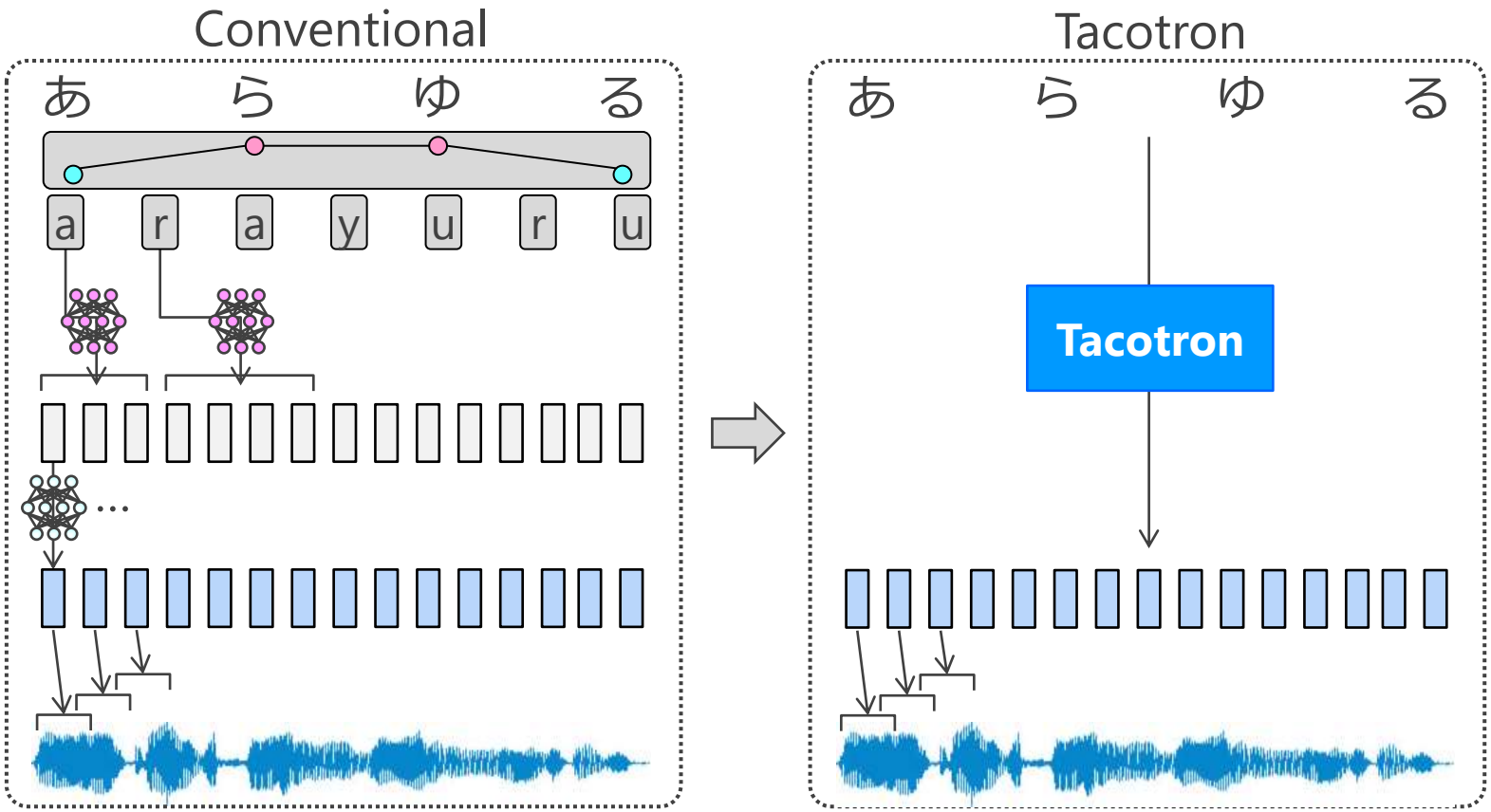


Tacotron: Sequence-to-sequence変換によるテキストー音声系列変換

2017年，言語処理部と継続長生成を置き換えるものが登場

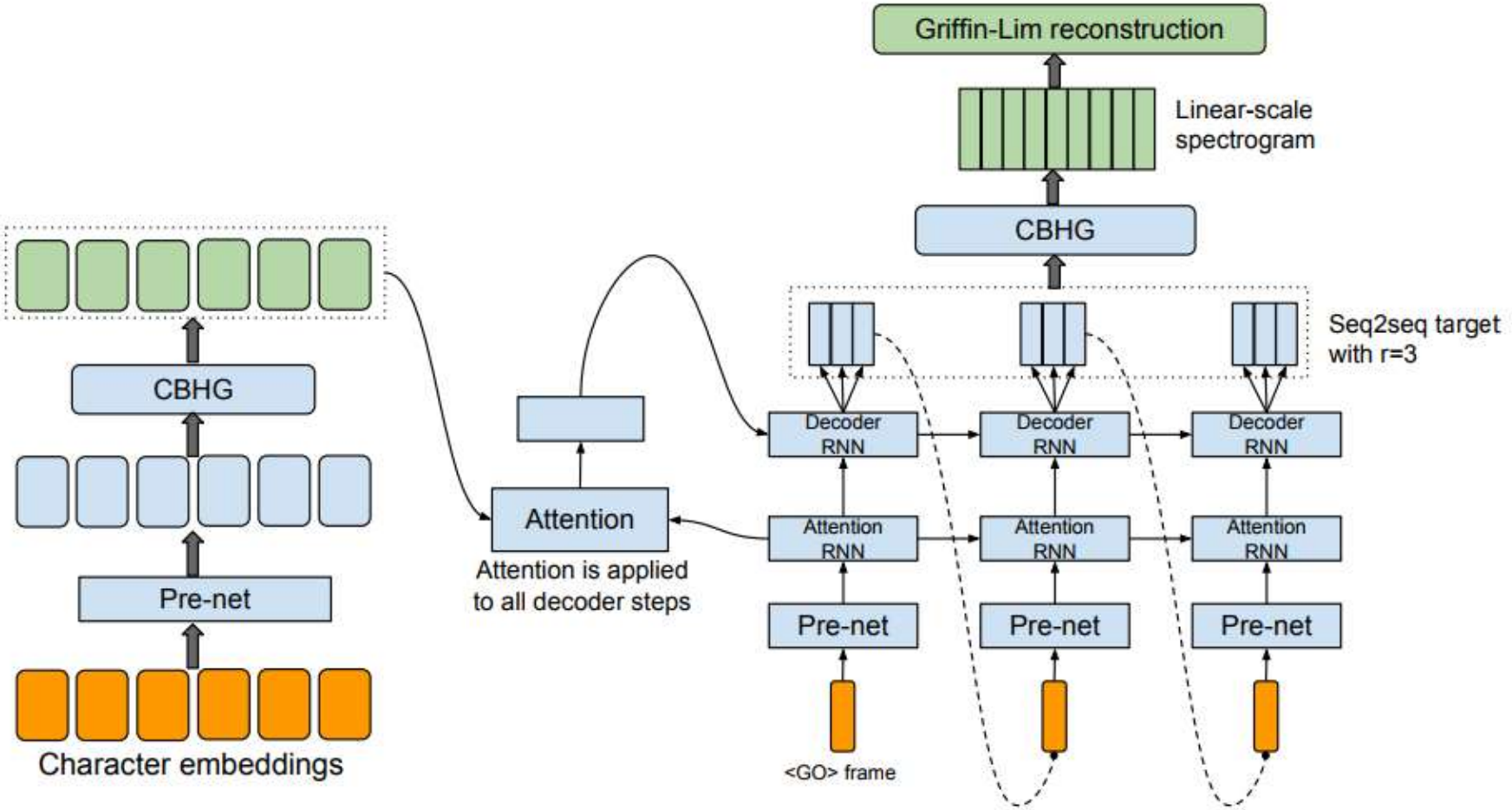
一番のキーは，機械翻訳で提案されたattention modelの利用

Attention model … 可変長変換を実現するモデルの一つ (詳細は省略)



Tacotron

[Wang et al., 2017.]



Anti-spoofingに敵対する音声合成

[Saito et al., 2017.]

課題

自然音声と合成音声の違いを，自動的に見つけてくれないか？

Anti-spoofing

音声合成による「声のなりすまし」を防ぐ技術

Anti-spoofingに敵対する学習

Anti-spoofingを騙すように音声合成器を更新

Anti-spoofingと交互に更新

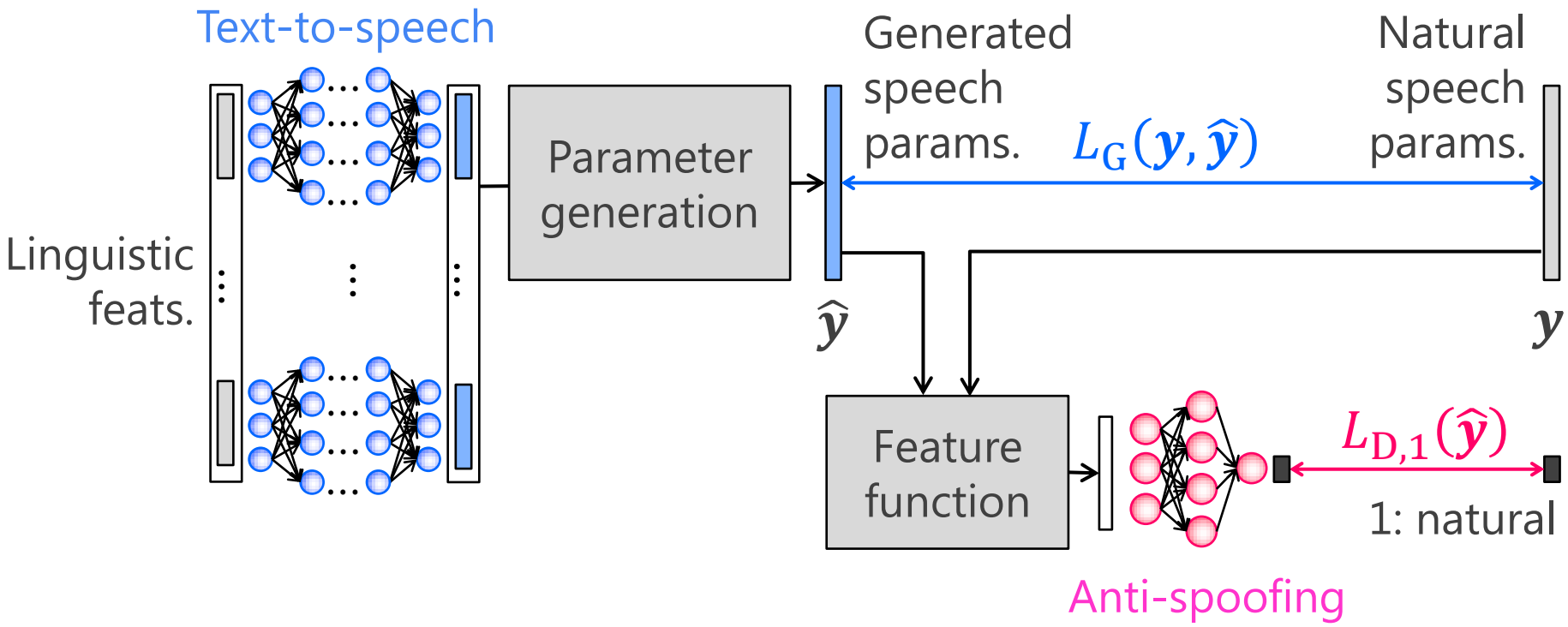
音声合成の更新

Anti-spoofingの更新

DNN音声合成のための敵対学習

[ICASSP2017 SLP Grant Award]

[Saito et al., 2017.]



$L(y, \hat{y}) = L_G(y, \hat{y}) + \omega_D L_{D,1}(\hat{y})$ を最小化

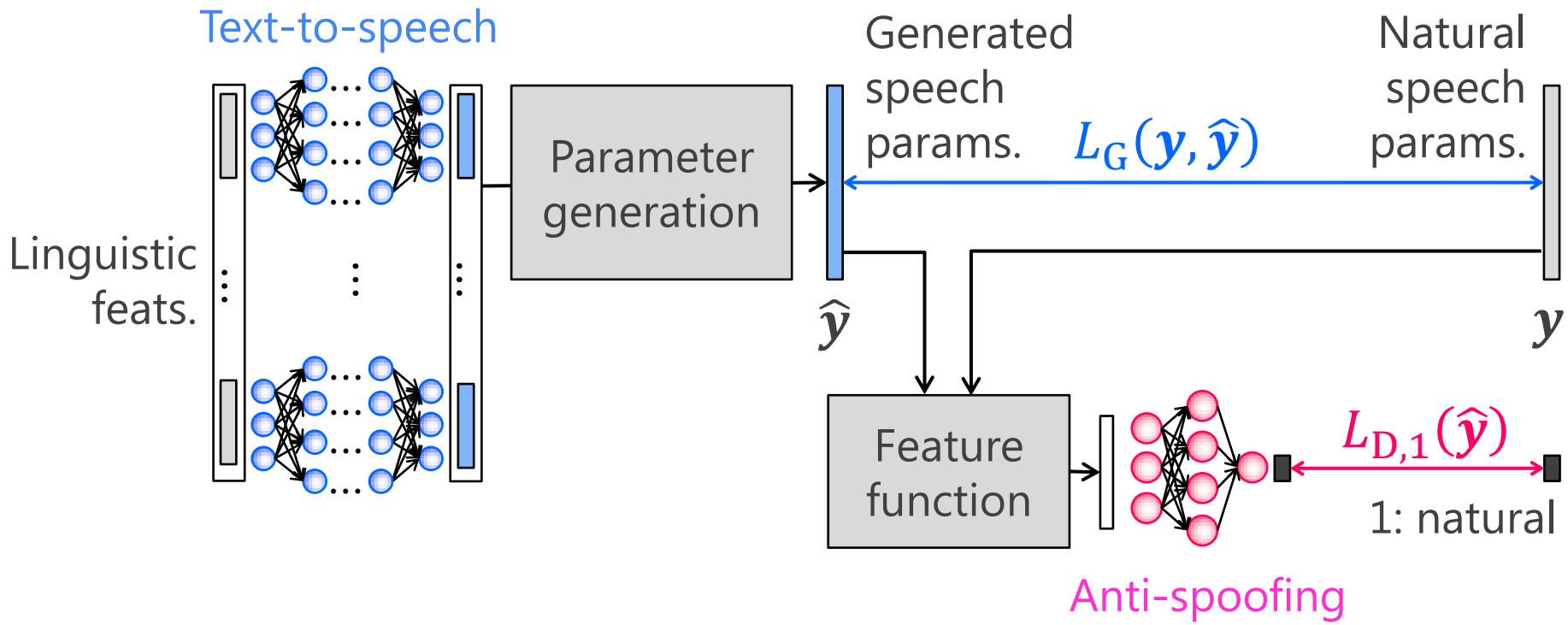
$L_G(y, \hat{y})$: 生成誤差

$L_{D,1}(\hat{y})$: Anti-spoofingを騙す損失

DNN音声合成のための敵対学習

[ICASSP2017 SLP Grant Award]

[Saito et al., 2017.]



人間の声に似せようと努力

ウソ（合成音）に騙されまいと攻防

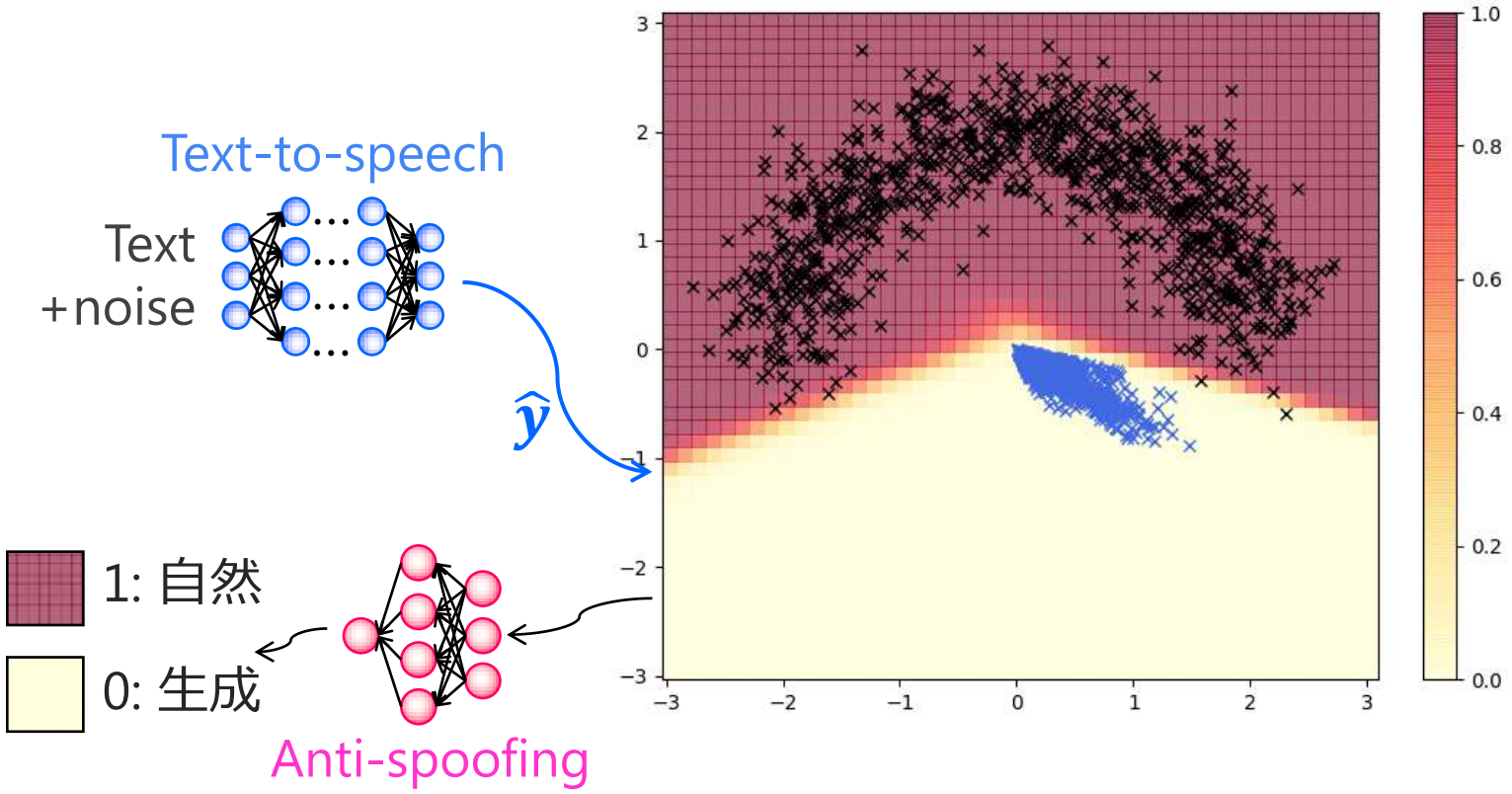
Generative Adversarial Network (GAN): 分布間距離の最小化

[Goodfellow et al., 2014.]

Generative adversarial network

分布間の近似 Jensen-Shannon divergence を最小化

合成器と，自然／合成音声を識別するanti-spoofingを敵対



一期一会な音声合成

[INTERSPEECH2017 Travel Grant Award]

[Takamichi et al., 2017.]

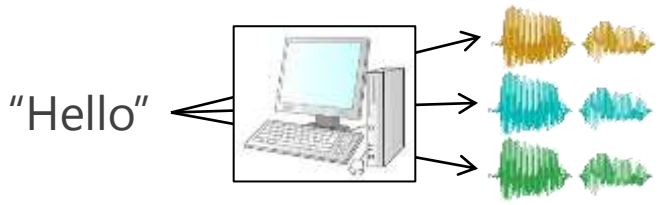
音声は一期一会．同じ音声には二度と出会えない！

でも，計算機の話す音声はいつも同じ…

一期一会音声合成

音質を保ちながら，人間の様に毎回違う声を合成する音声合成

Moment-matching network を利用



Moment-matching network

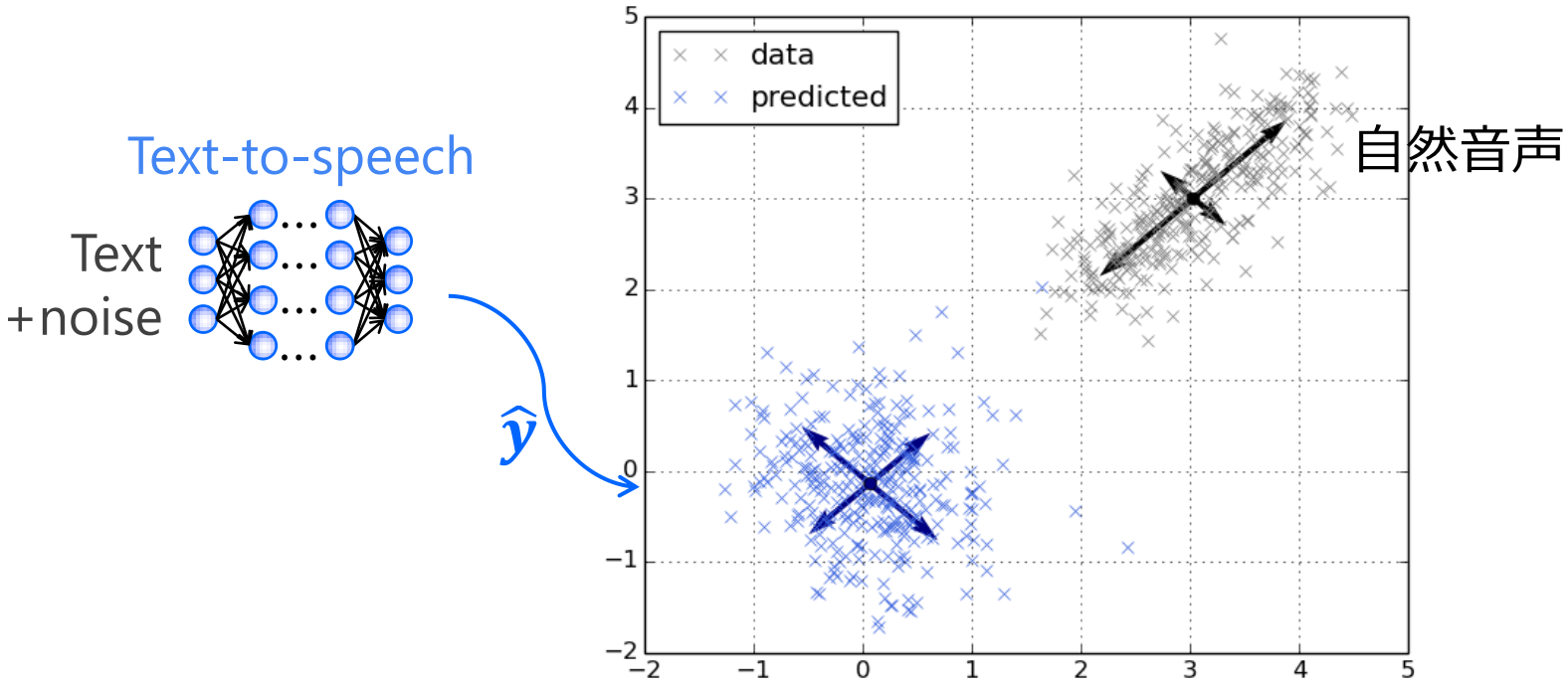
[Li et al., 2015.]

Moment matching network [Li et al., 2015.]

分布のモーメント (平均，分散，...) 間の二乗距離を最小化

- モーメントを使って，音声のばらつきを表現

実装上は，グラム行列のノルムの差を最小化

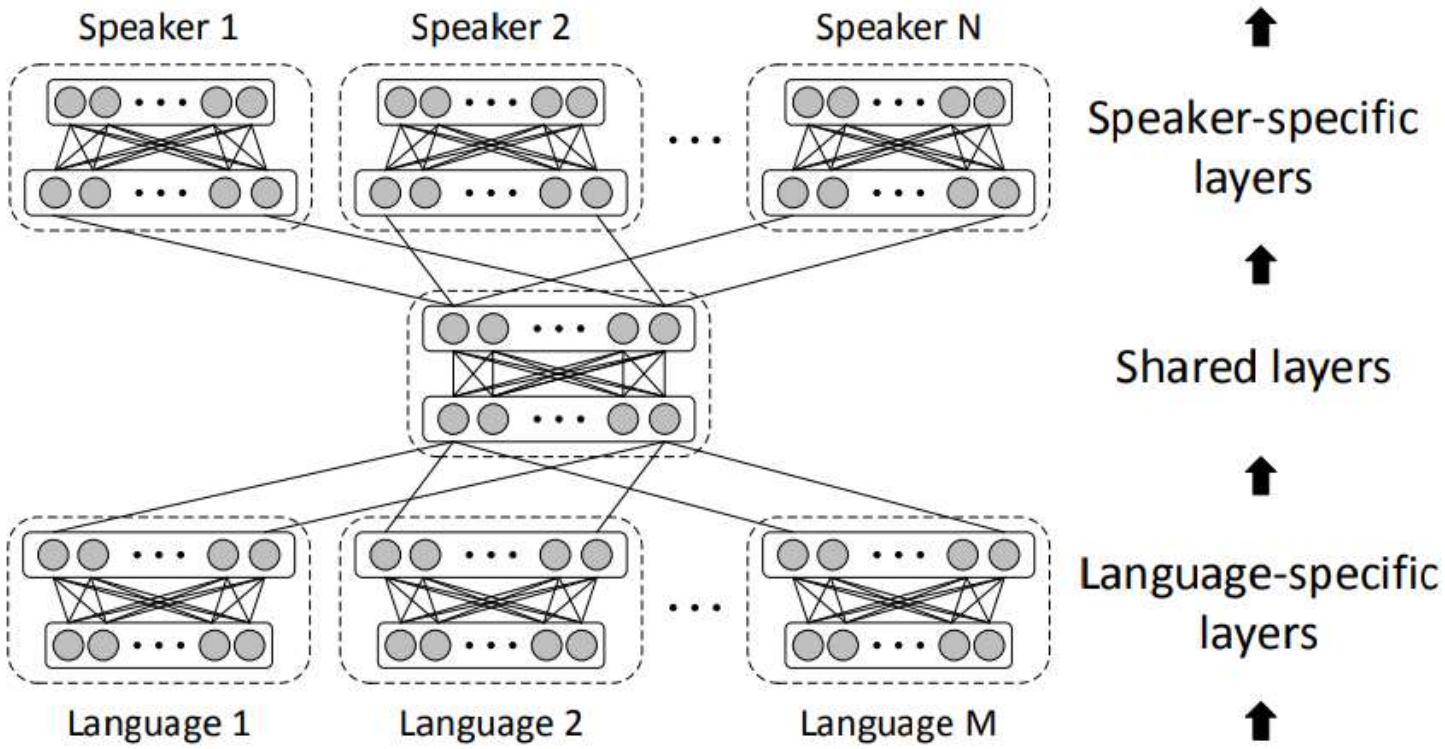


多言語・他話者モデリング

[Fan et al., 2016.]

複数言語・複数話者の音声合成を一気に作りたい！

DNNを分割して，各言語・各話者をモデル化する専用の層を作る



http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7472737

主成分分析やvariational autoencoderによる教師無し特徴量抽出

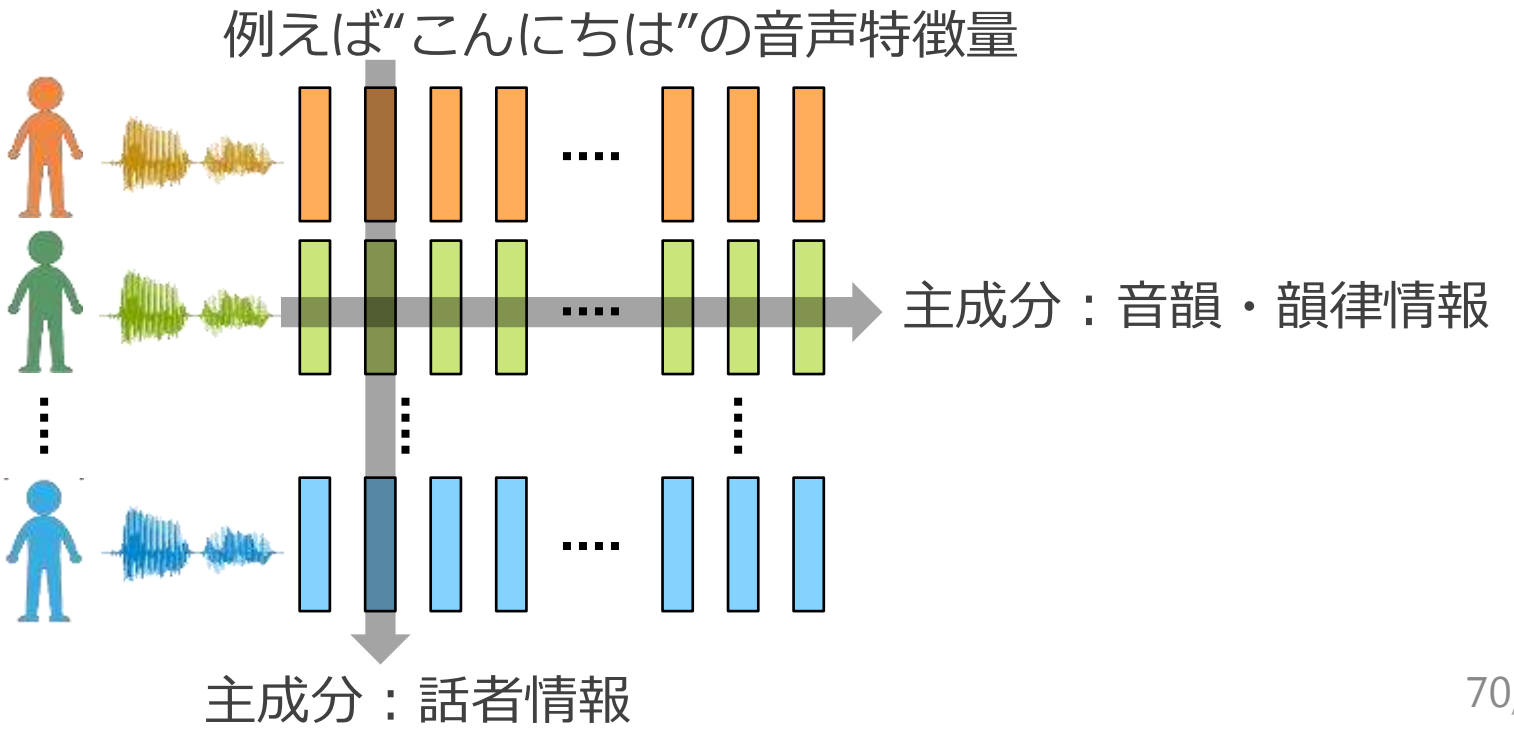
音声特徴量の潜在変数とは？

音韻・韻律情報 (何を発話しているか)：

- 同じ文を話していれば，話者間で共通の因子

話者情報 (誰が発話しているか)：

- 発話内容によらず同一話者内では共通の因子



近年の応用研究 (紹介だけ)

発話障害・聴覚障害を補助する音声合成

話せない／聞こえない音声を話せる／聞こえるように

音声翻訳・言語教育を支援する音声合成

言語の壁を越えて声を伝える

希少言語の音声合成

音声言語文化の定量化

あらゆる声からあらゆる声を合成

音環境への頑健性など

話し言葉の音声合成

ただ読みあげる，一方的に話かける音声合成を超える

ただ喋るだけの音声合成の時代は既に終わっている．
これからは，人にどう寄り添うか，人をどう拡張するか，
何に声の芸術性を見出すかが重要になる．

レポート課題（猿渡担当分）

※注意：レポートには表紙（表紙には，講義タイトル，担当回の教員名，学籍番号．科類，学年，氏名を記載）を必ず付けること。

１．「対象を統計的な現象としてモデル化する」ことの意義について述べよ。また、身の周りの物理現象に関して「最尤推定」を当てはめた例を１つ示せ。

２．表には1・裏には0と書かれたコインがある（表裏の生起確率は等しいものとする）。このコインをN個同時に投げ、「その表示面の数字の和」を確率変数Xとみなす。これがNの増加とともにガウス分布に近づくことを示して中心極限定理を説明せよ。なお、図を使ってヒストグラムの概形を示しても良い（一様分布が釣鐘型に変わる様子を示せ）。

３．独立成分分析に関して説明せよ。