

ディープラーニングを利用した音声認識と音声合成

山岸順一 准教授

国立情報学研究所 コンテンツ科学研究系

英国エジンバラ大学 音声技術研究所

自己紹介：経歴

- 音声情報処理、特に音声合成の研究に17年間従事
- 2006年：博士号（東工大、工学）
- 2007～現在：エジンバラ大 Senior Research Fellow
- 2013～現在：国立情報学研究所(NII) 准教授

各年の被引用数合計



文部科学大臣表彰
若手科学者賞



日本学術振興会賞

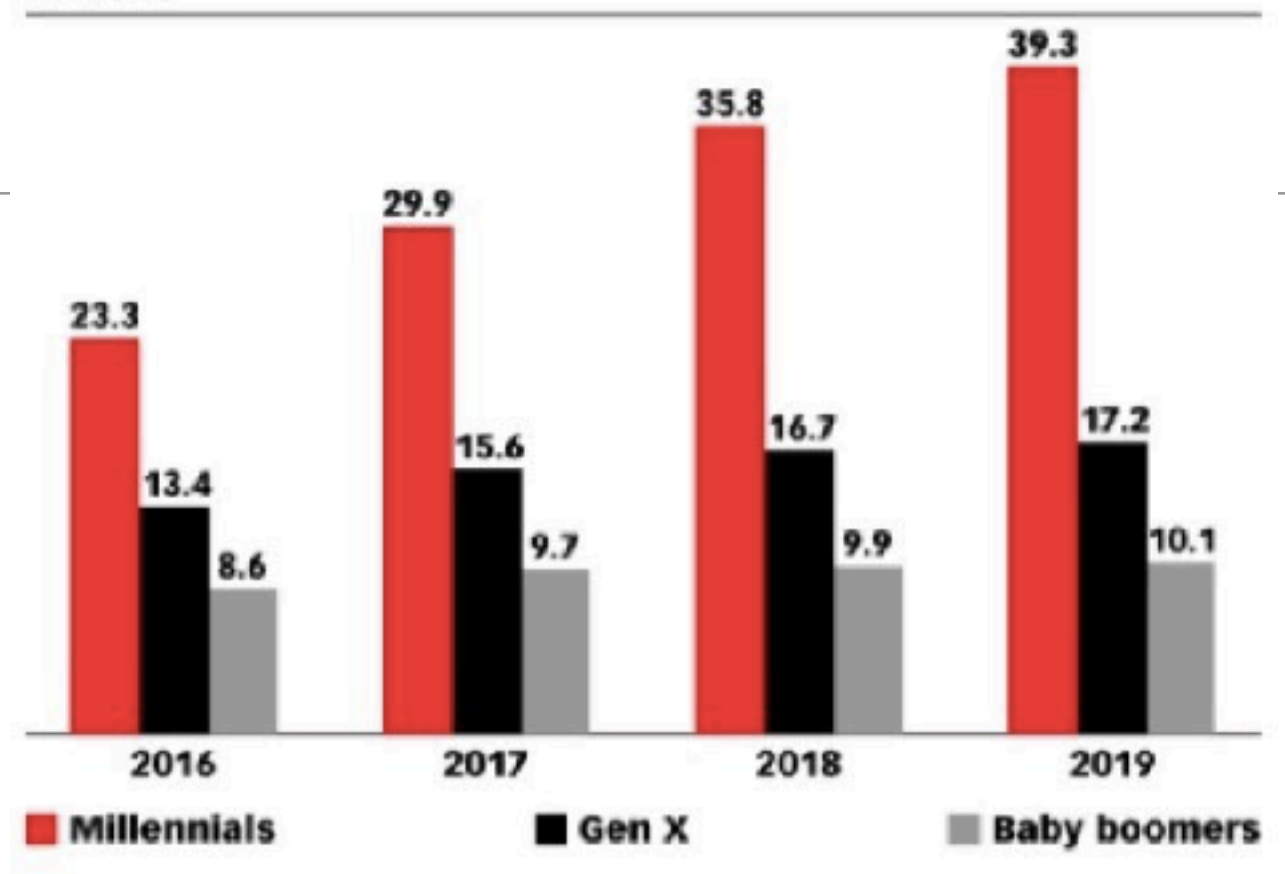
本日の構成

- 深層学習による音声合成
- 深層学習による音声認識

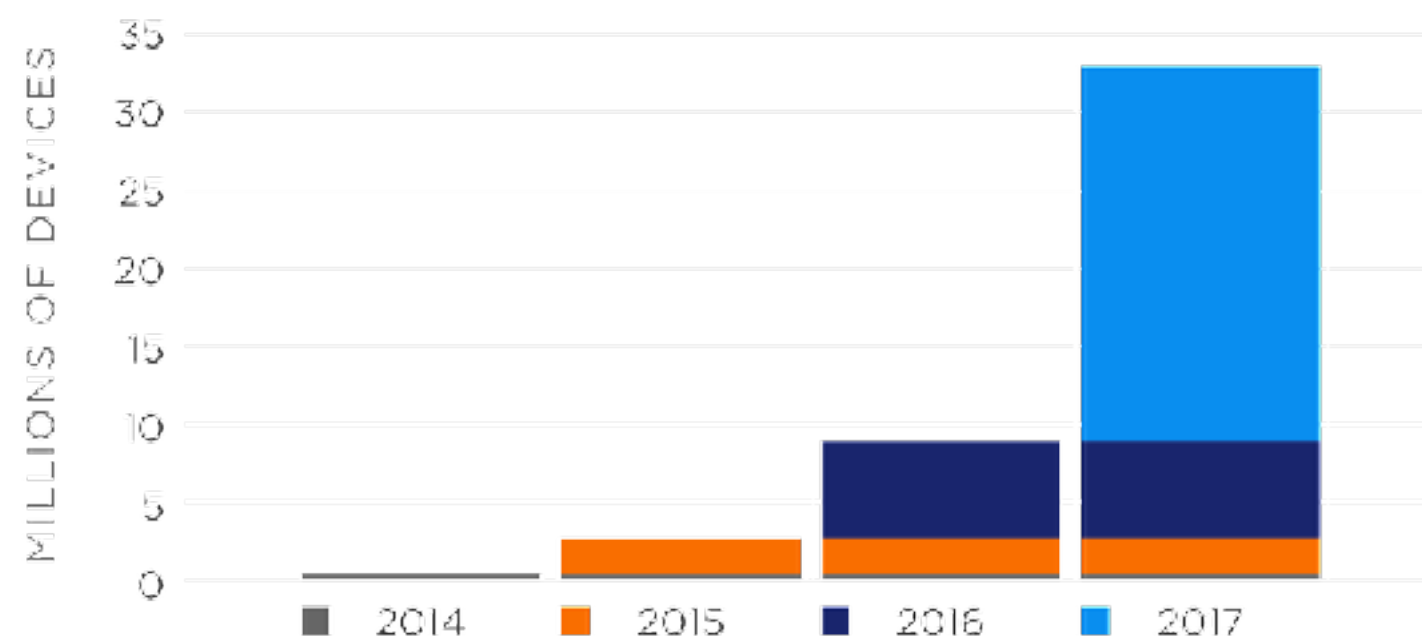
現在の音声技術

- 音声情報処理技術の普及
 - 音声検索
 - Google voice search
 - 音声翻訳
 - Google translation
 - 音声対話エージェント
 - Siri
 - Ok Google!
 - Alexa
- 若者を中心に普及が続いている
- 音声の主たる入力であるデバイスも劇的に増加

US Voice-Enabled Digital Assistant Users, by Generation, 2016-2019
millions



Voice-First Device Footprint

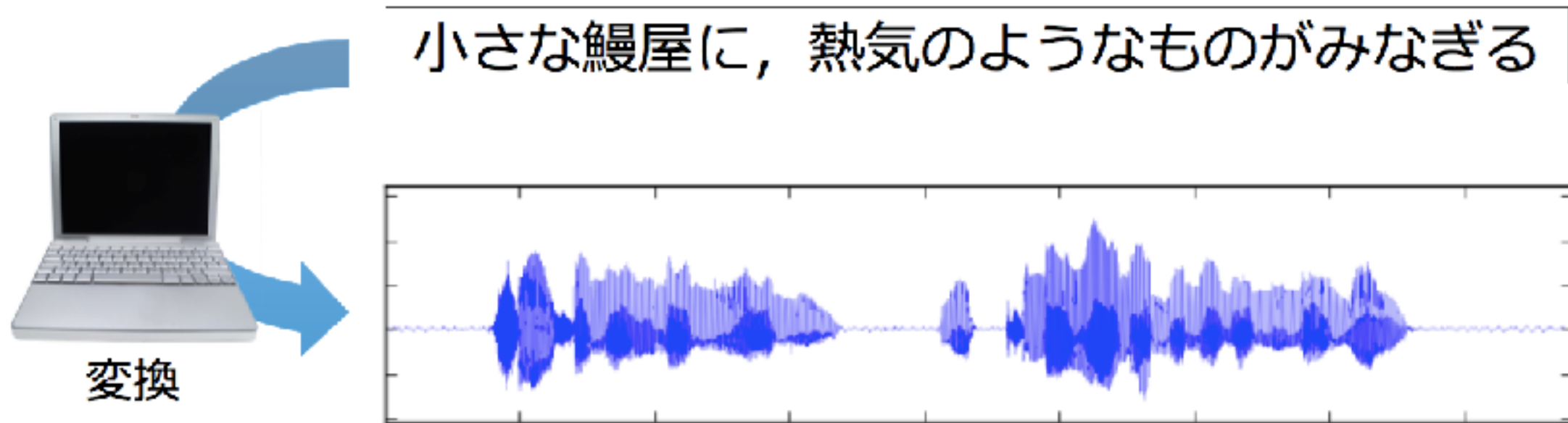


Voice Labs Analysts combined with research from CBR, KPMG, and others

<http://voicelabs.co>より引用

今日の一つ目の話題：音声合成

- テキスト音声合成：入力テキストを自然で聞き取りやすい音声へ変換



- ここ数年は企業を中心に研究・開発スピードが加速
 - Google (Google Brain, Deep Mind, Google TTS team)
 - Apple, Amazon, Microsoft Research, Facebook Research,
 - Baidu, Alibaba, iFlytek
 - エジンバラ大, CMU
 - 国内：NII, NTT, 名大, 名工大

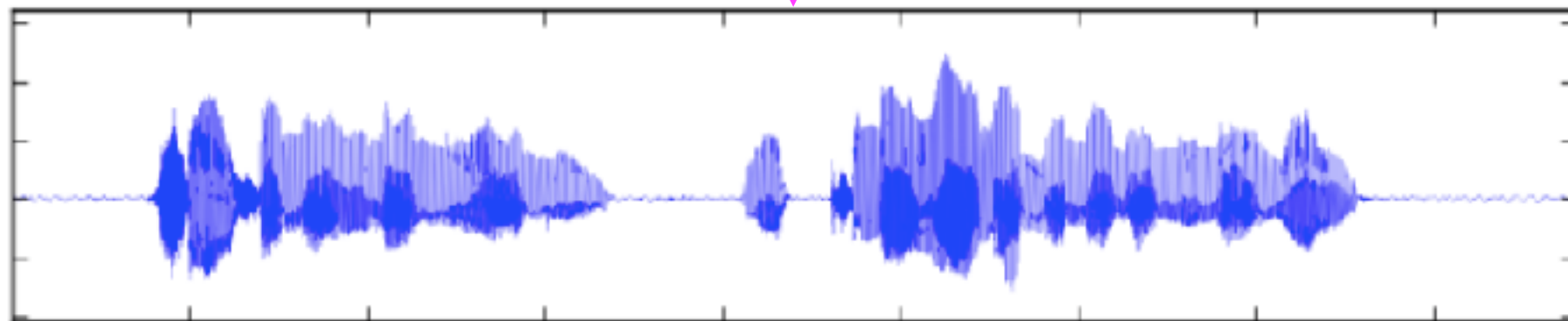
従来の音声合成：文章と音声の対応付け

小さな鰻屋に、熱気のようなものがみなぎる

文章の言語的特徴量抽出

言語情報と音響的特徴量の対応付け

音声の音響的特徴量から波形合成



従来の音声合成：文章と音声の対応付け

小さな鰻屋に、熱気のようなものがみなぎる

文章の言語的特徴量抽出

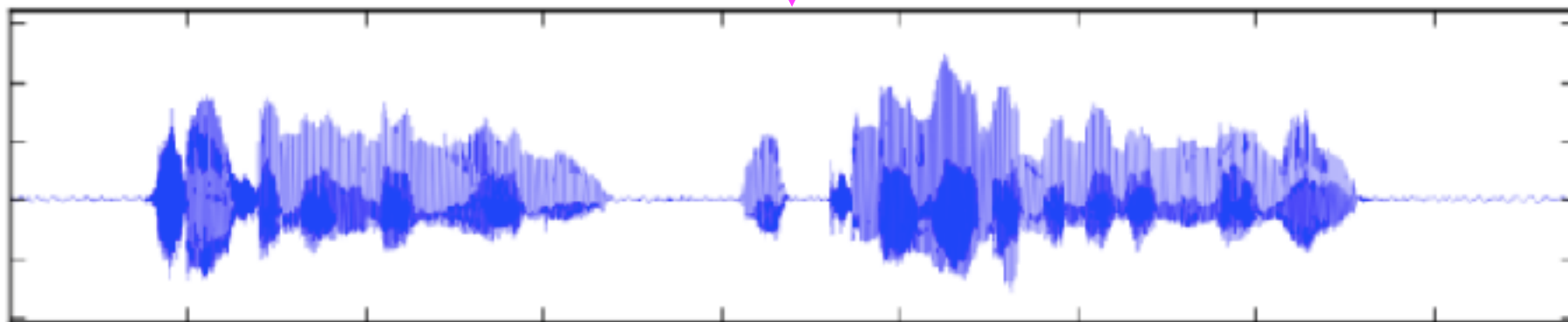
自然言語処理

言語情報と音響的特徴量の対応付け

機械学習

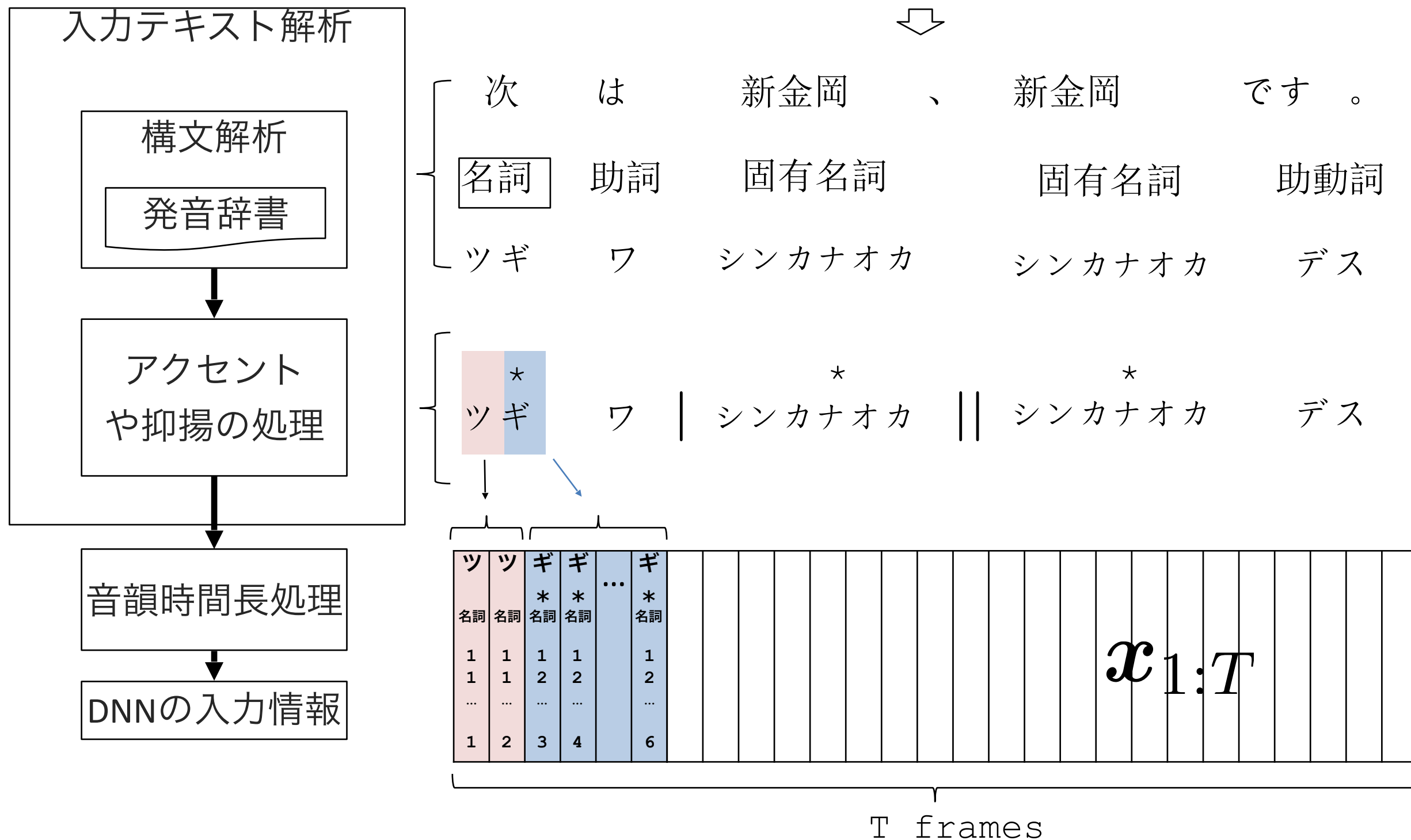
音声の音響的特徴量から波形合成

信号処理



従来の言語的特徴量抽出

次は新金岡、新金岡です。



従来の音声合成：文章と音声の対応付け

小さな鰻屋に、熱気のようなものがみなぎる

文章の言語的情報抽出

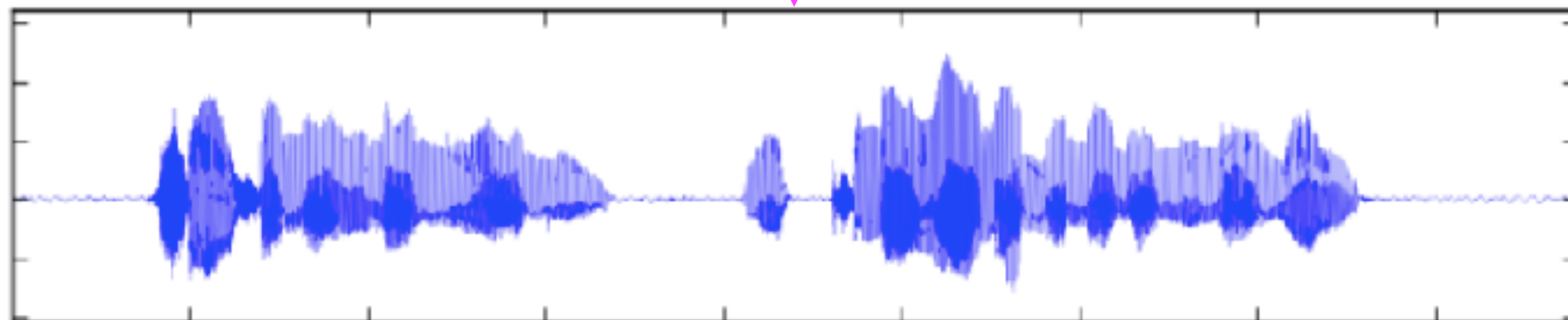
自然言語処理

言語情報と音響的特徴量の対応付け

機械学習

音声の音響的特徴量から波形合成

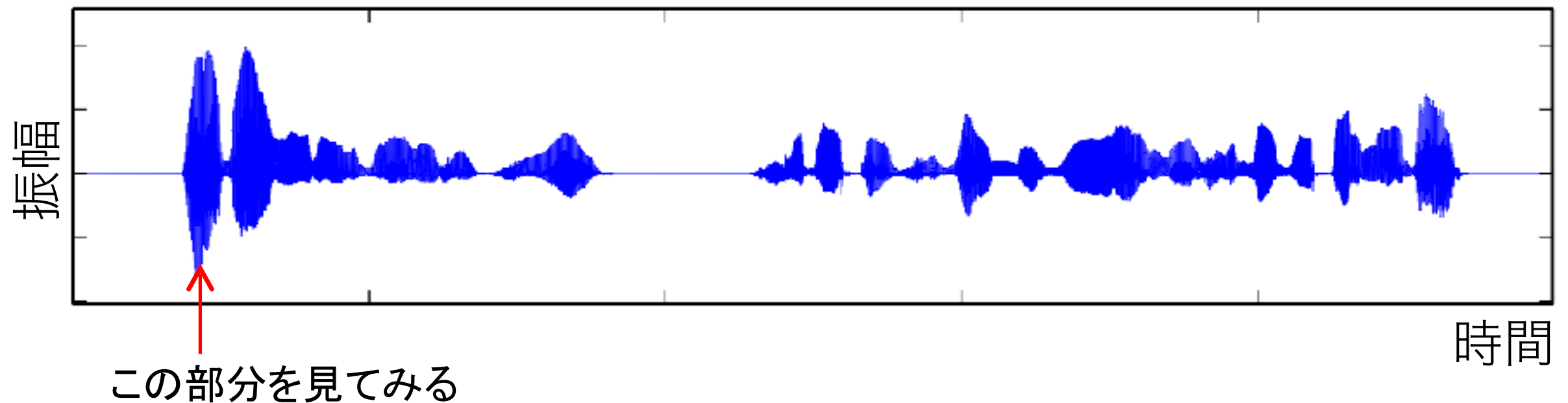
信号処理



従来の音声の音響的特徴量 (1/5)

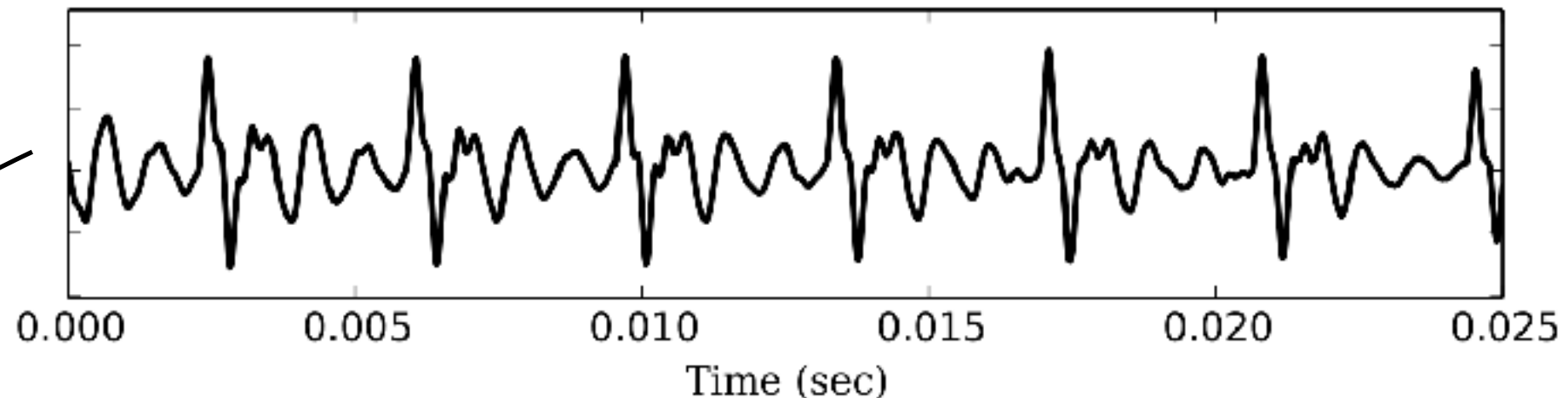
音声信号：

サンプリング周波数が48kHzの場合
1秒間に48,000の観測点がある



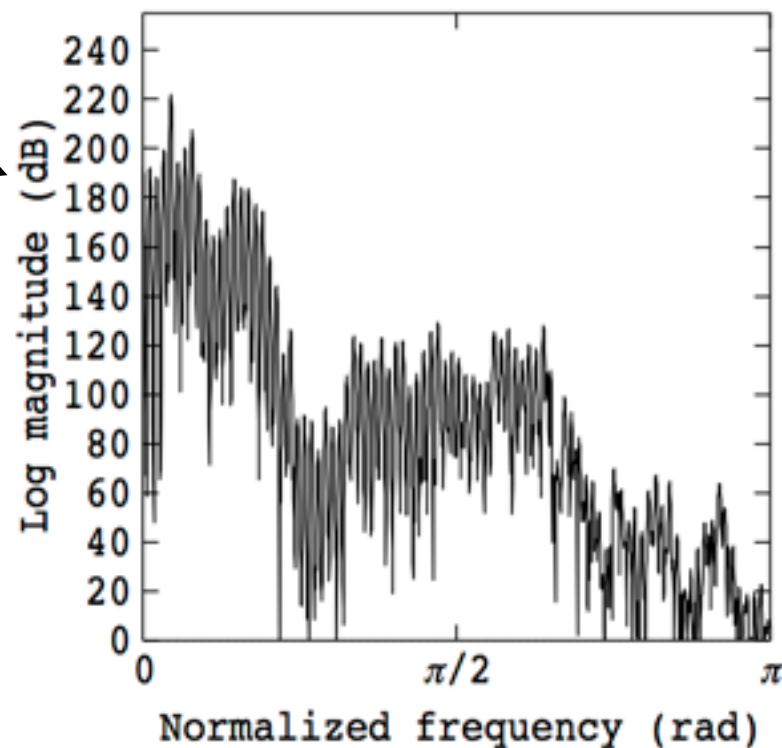
従来の音声の音響的特徴量 (2/5)

短い区間（この例では25ミリ秒）を分析



短時間

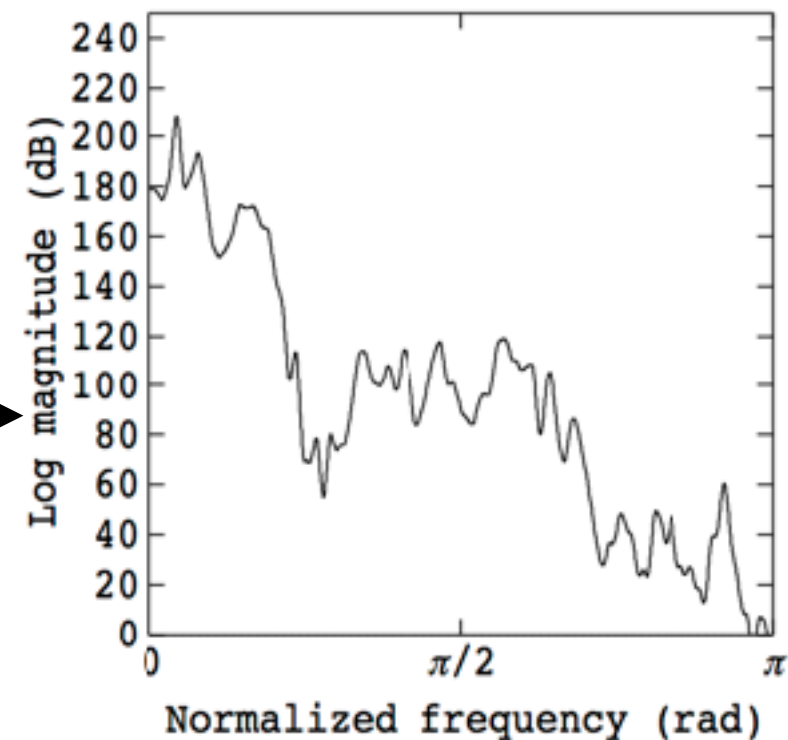
フーリエ変換



複素スペクトル：

音声に含まれる各周波数の強度および位相

DCT
変換

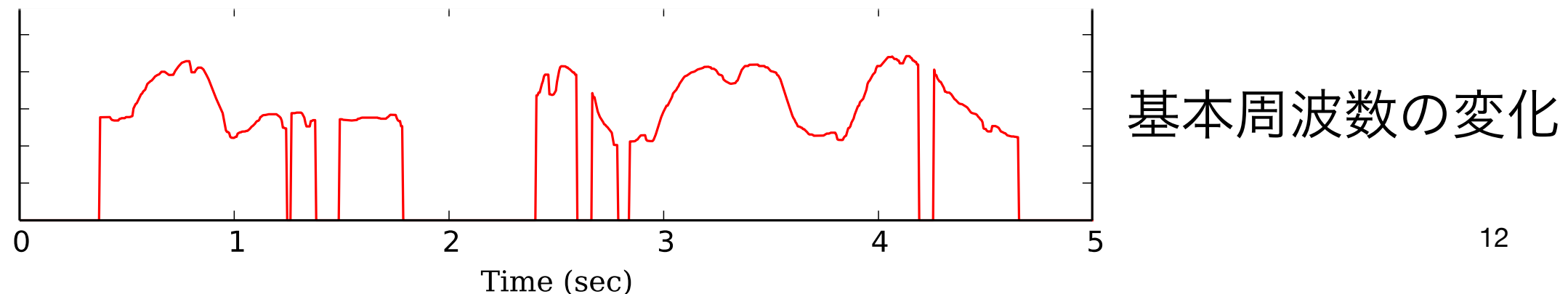
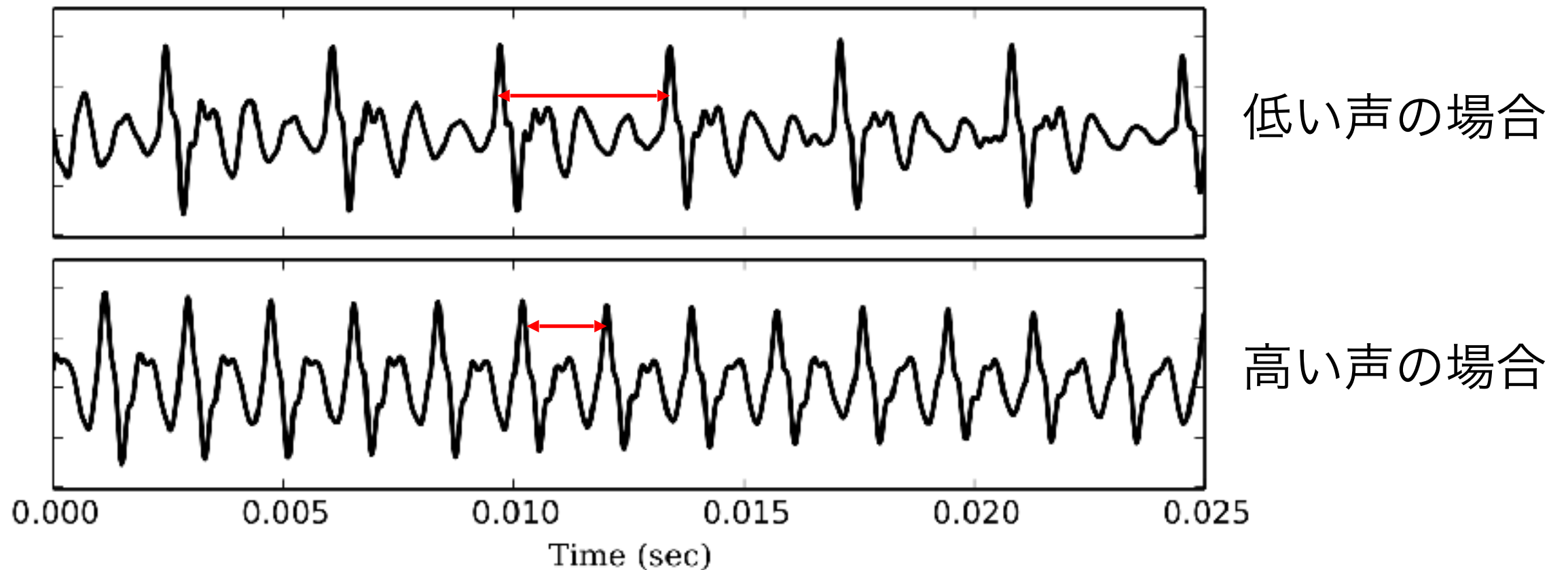


スペクトル包絡：

音色(音素)の情報

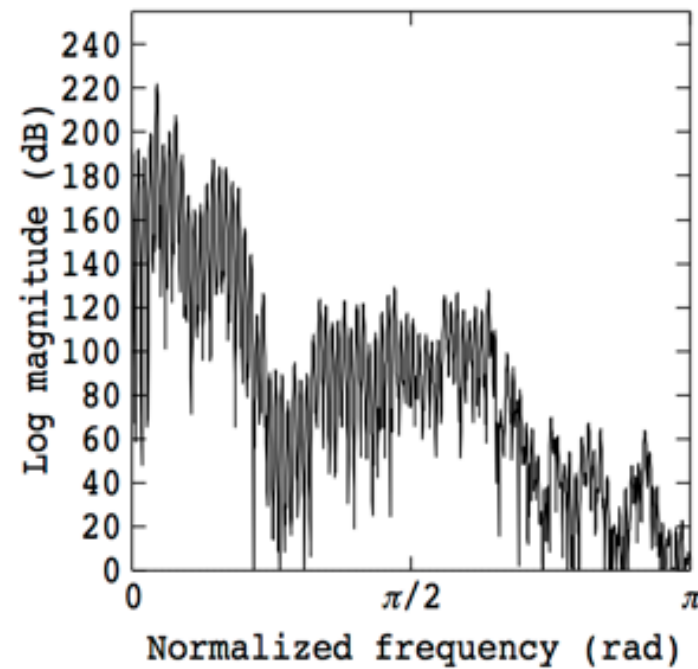
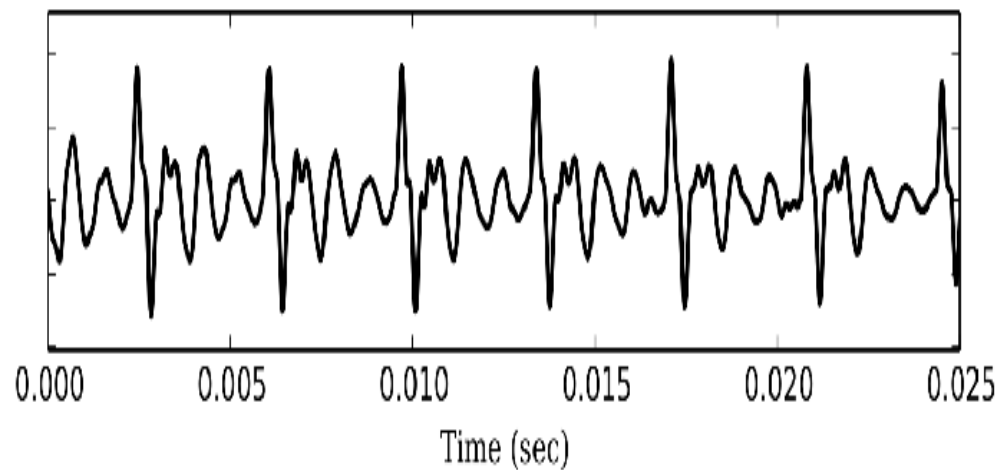
従来の音声の音響的特徴量 (3/5)

- 同じ音素でも似た波形が表れる間隔が異なる
 - 音声の基本周期 (T_0) , **基本周波数** ($F_0 = 1/T_0$)
- 高い声の方が間隔が短い

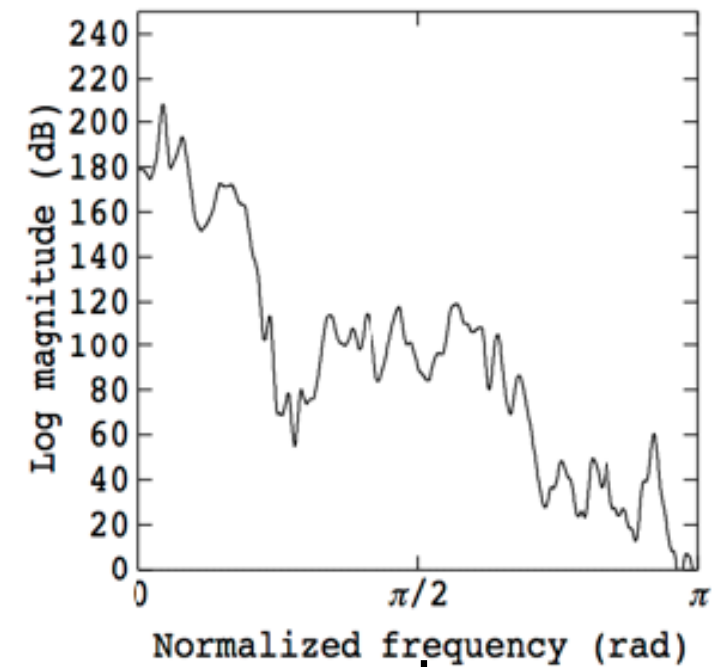


従来の音声の音響的特徴量 (4/5)

短時間波形
(1 frame)



分解(不可逆)



+

基本周波数

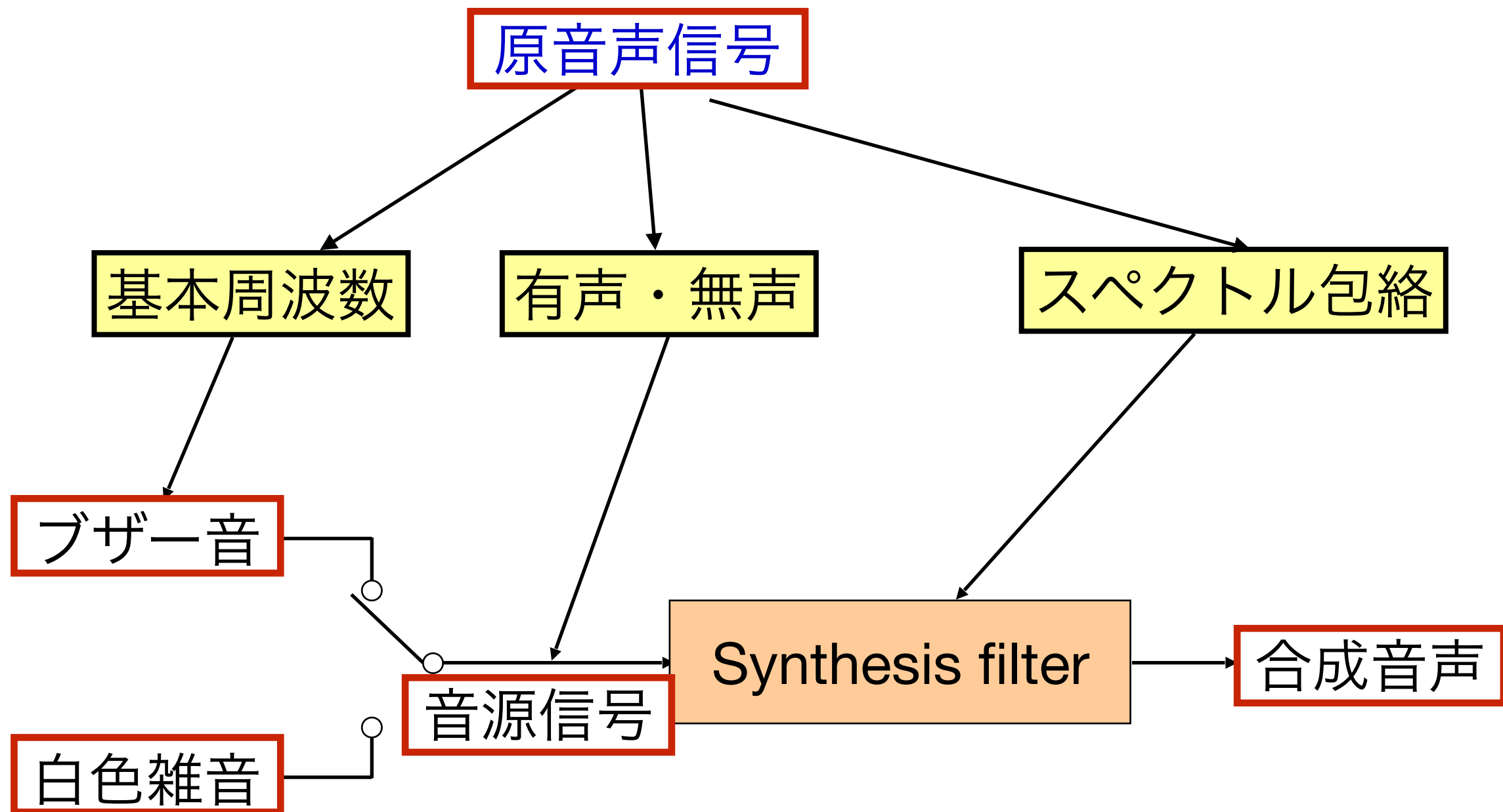
捉え難い

← 言語情報との対応関係 →

捉え易い

従来の音声の音響的特徴量 (5/5)

ボコーダによる音声波形合成



従来の音声合成：文章と音声の対応付け

小さな鰻屋に、熱気のようなものがみなぎる

文章の言語的情報抽出

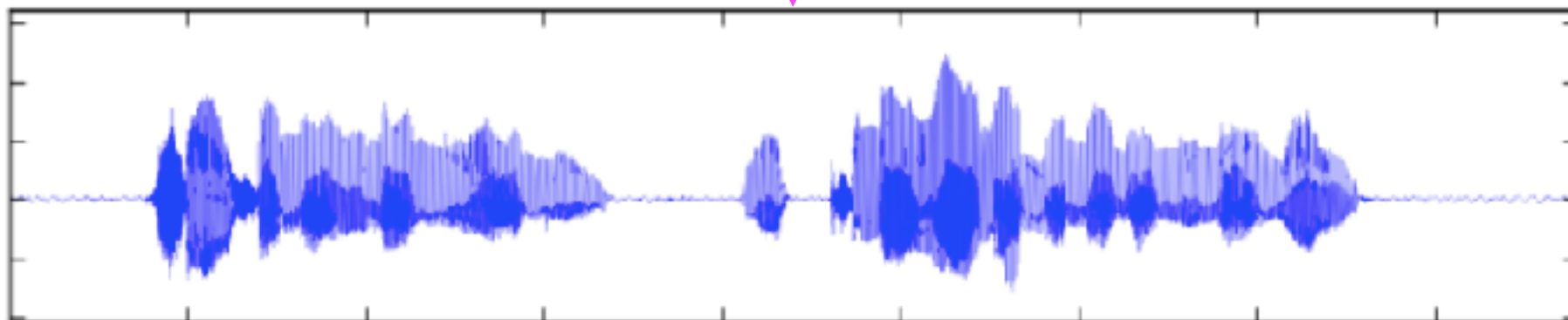
自然言語処理

言語情報と音響的特徴量の対応付け

機械学習

音声の音響的特徴量から波形合成

信号処理

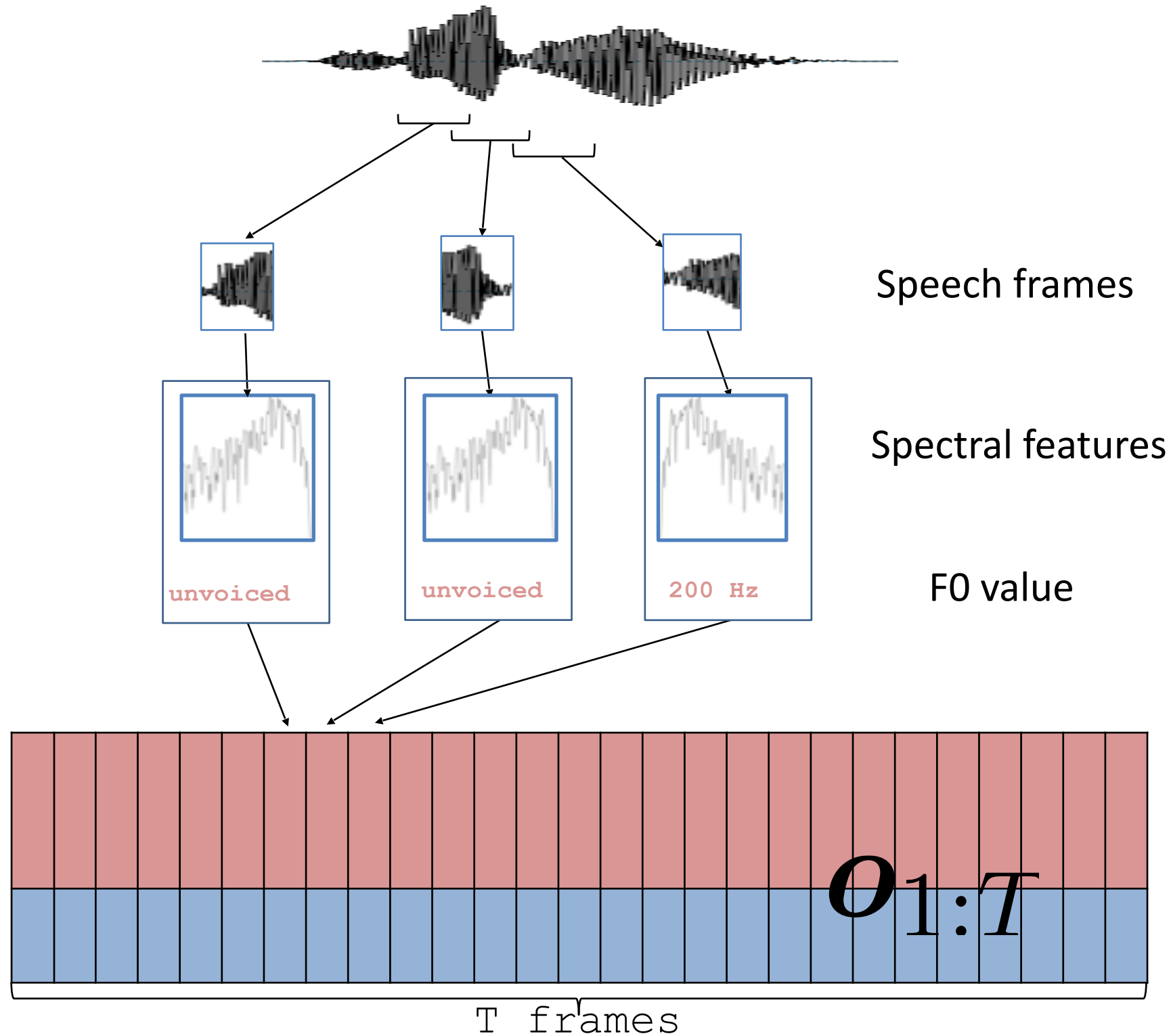


言語情報と音響特徴量の対応付け

- 言語特徴量と音響特徴量をどう対応づけるか=音声合成の歴史
 - フォルマント合成 (1970) :
 - 音素とスペクトル包絡(フォルマント)を決定論的に対応付け
 - ダイフォンの合成 (1980) :
 - ダイフォンと呼ばれる拡張単位と音声波形を対応付け
 - 波形接続合成 (1990) :
 - コンテキスト依存ダイフォンと音声波形を対応付け
 - 隠れマルコフモデル(HMM) 音声合成 (2000) :
 - コンテキスト依存音素と音響特徴量をHMMで対応付け
 - 深層学習による音声合成 (2010~)
 - 言語特徴量と音響特徴量との対応付けをDNNで学習
 - 言語特徴量と音響特徴量も明示的に利用しないend-to-end学習

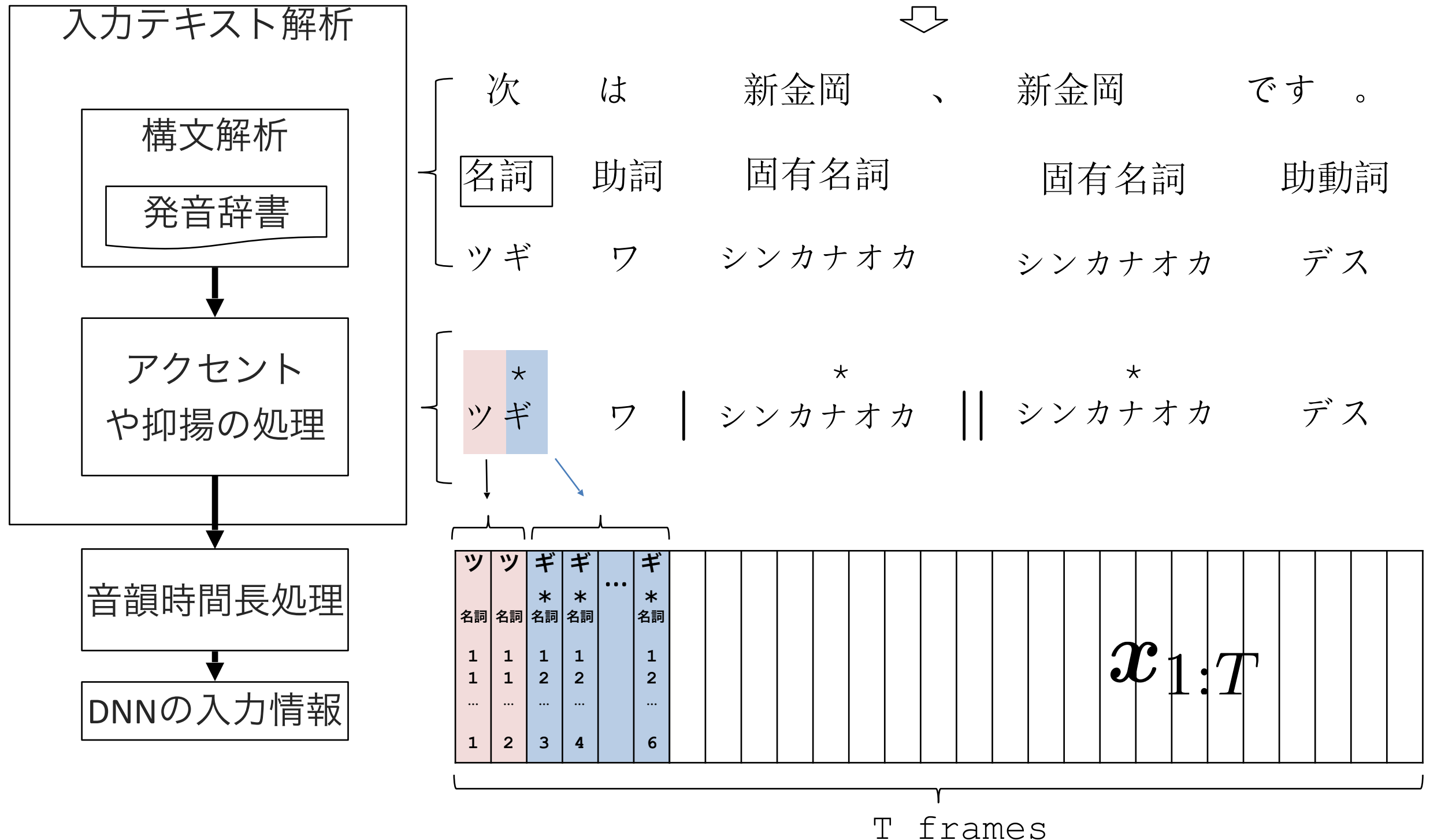
音声の音響的特徴量

音声信号全体(T frame)に処理を繰り返し、音響特徴量系列を得る

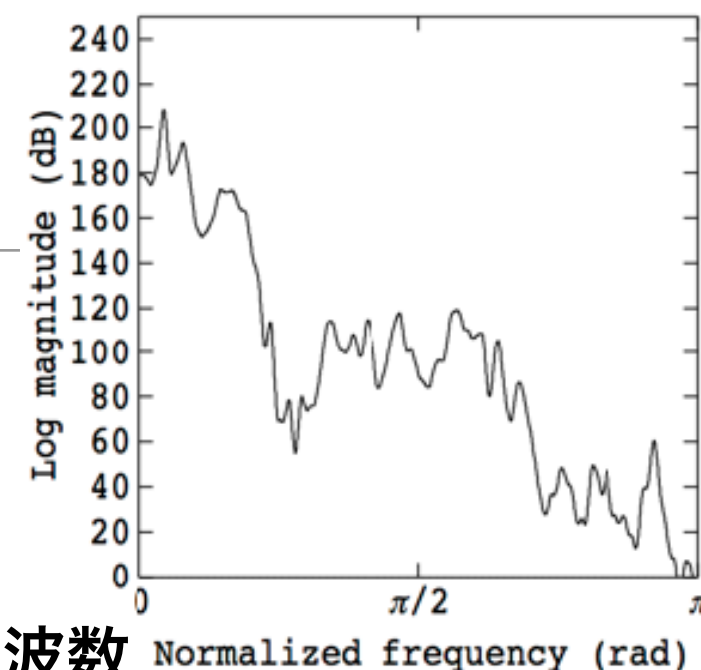
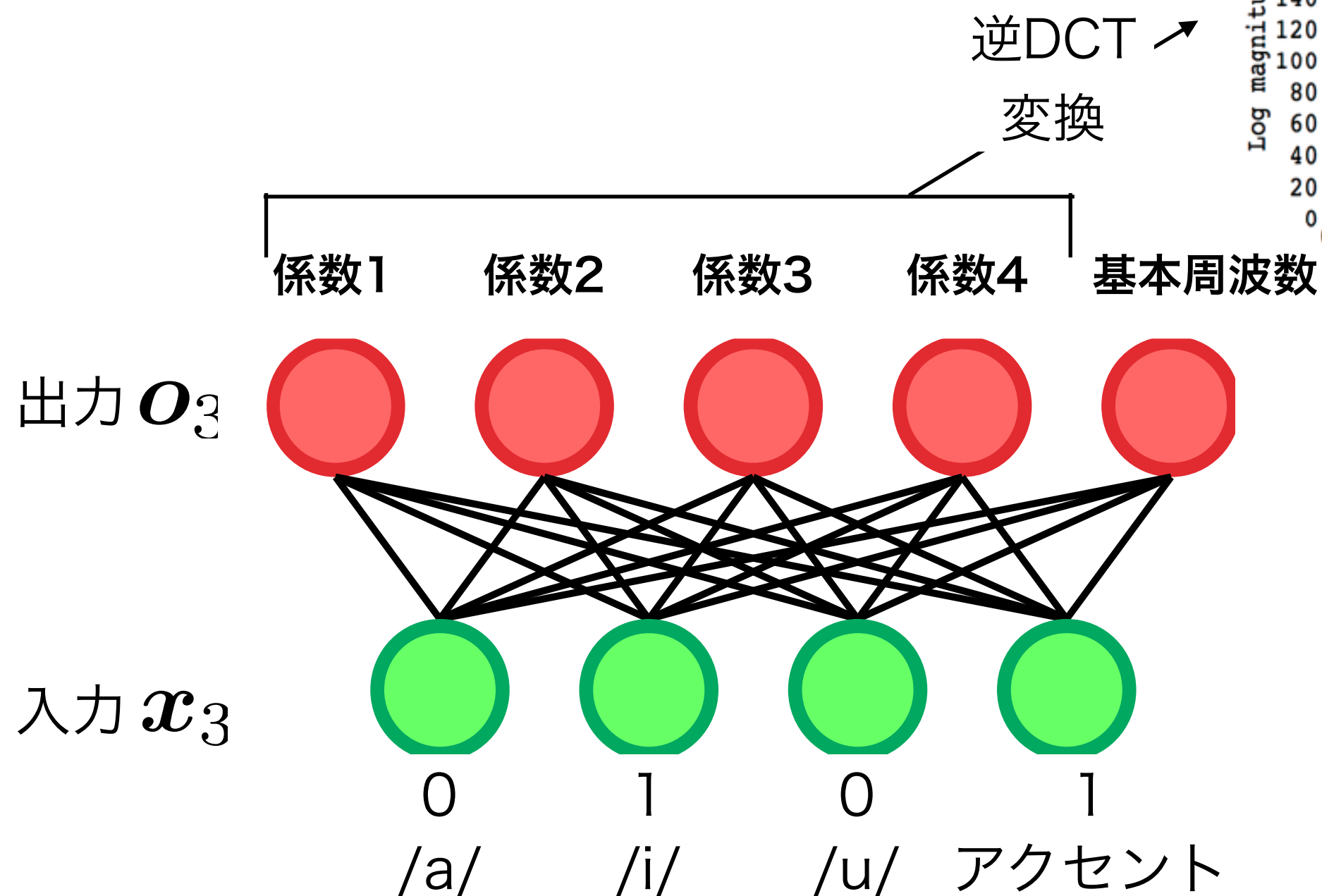


言語的特徴量抽出（再掲）

次は新金岡、新金岡です。

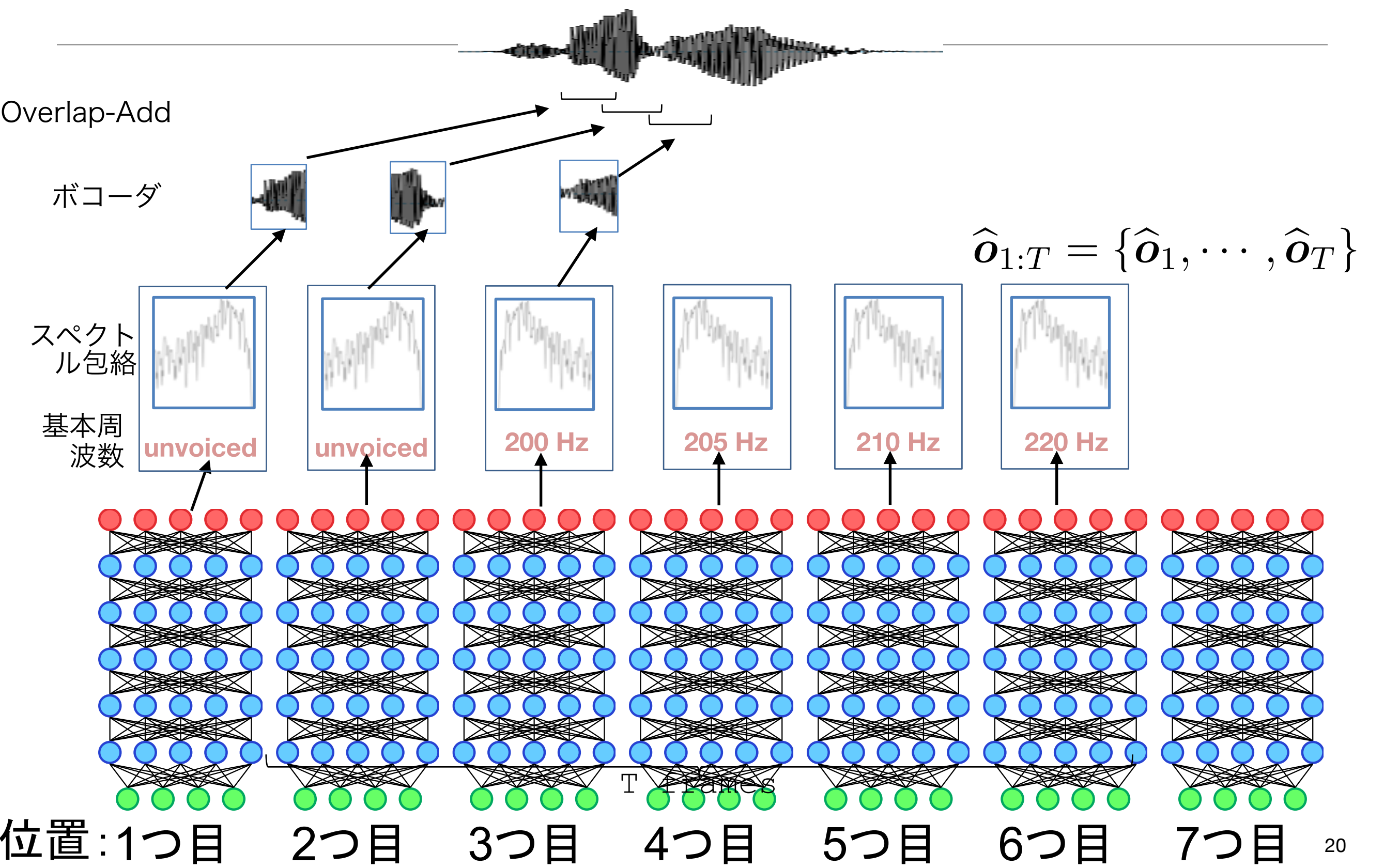


各フレームでの処理を見ると



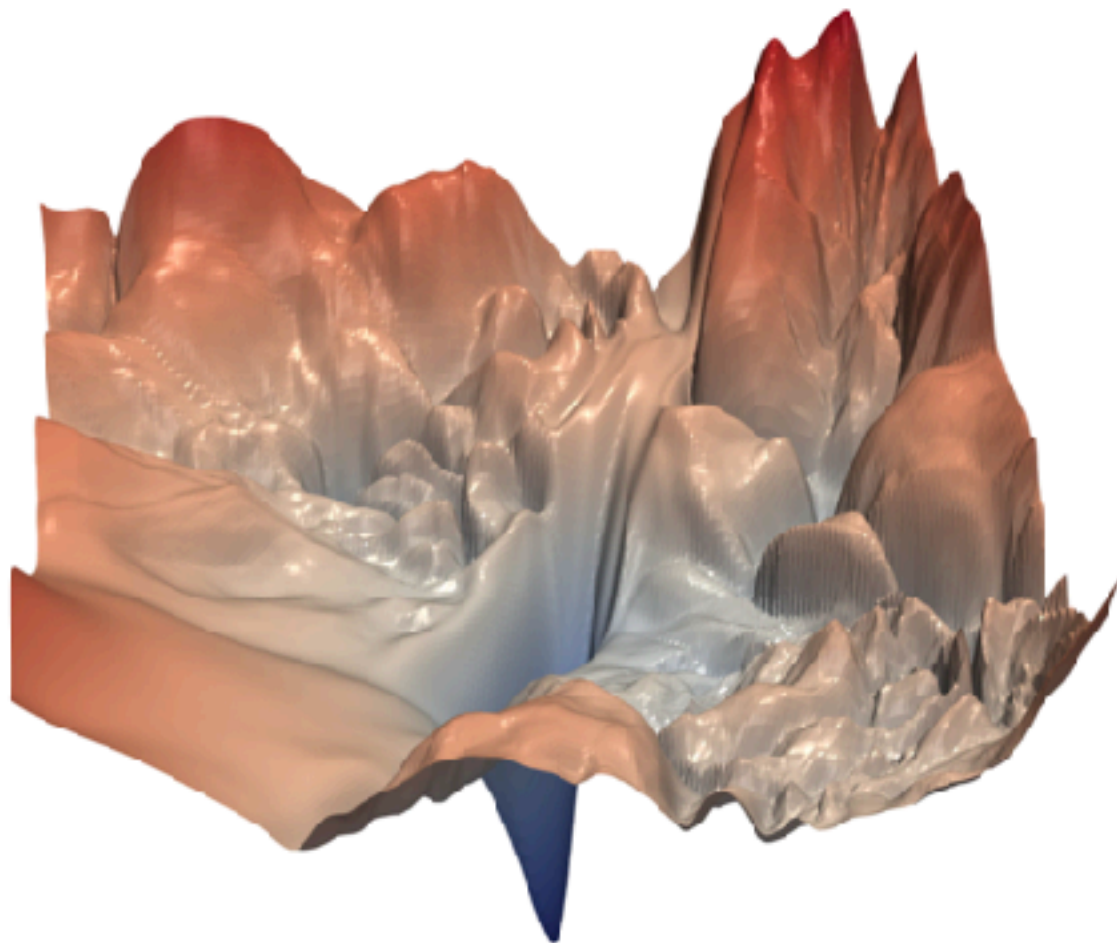
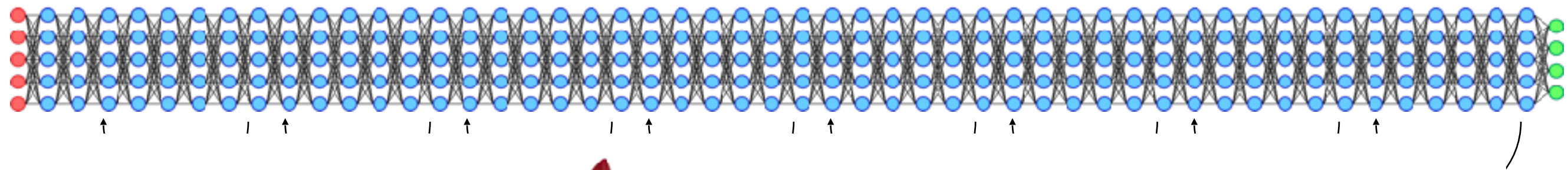
「次は新金岡、新金岡です。」のギのとある1フレーム
実際の入出力の次元数はもっと多い

全体での処理 (FeedForward型の例)

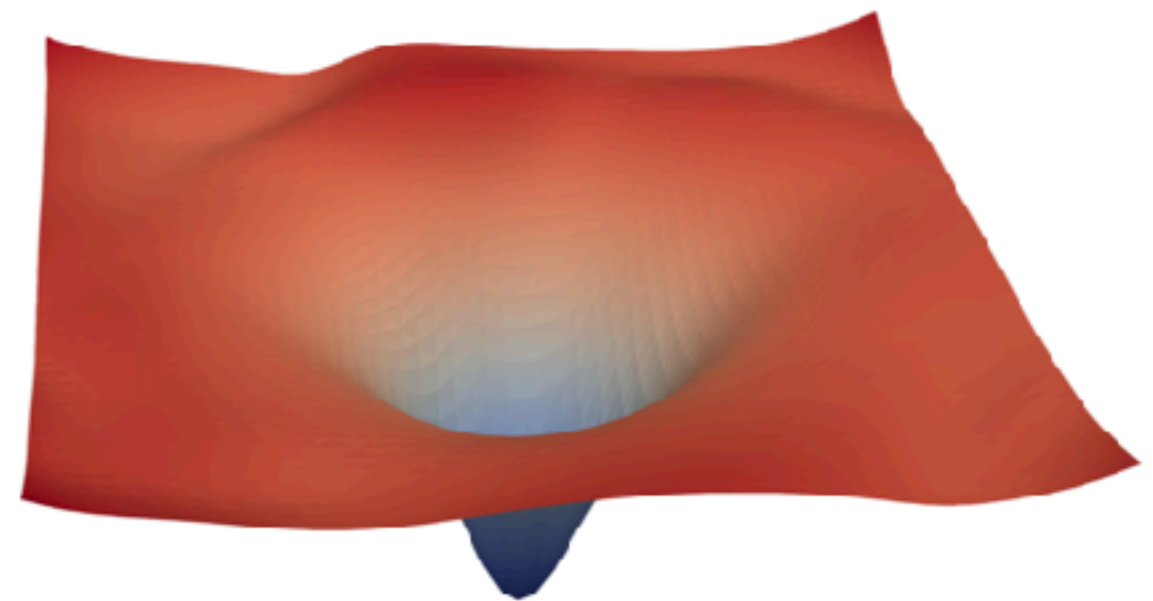


Residual/highway network

- ショートカットコネクションによる100層以上のネットワークも登場



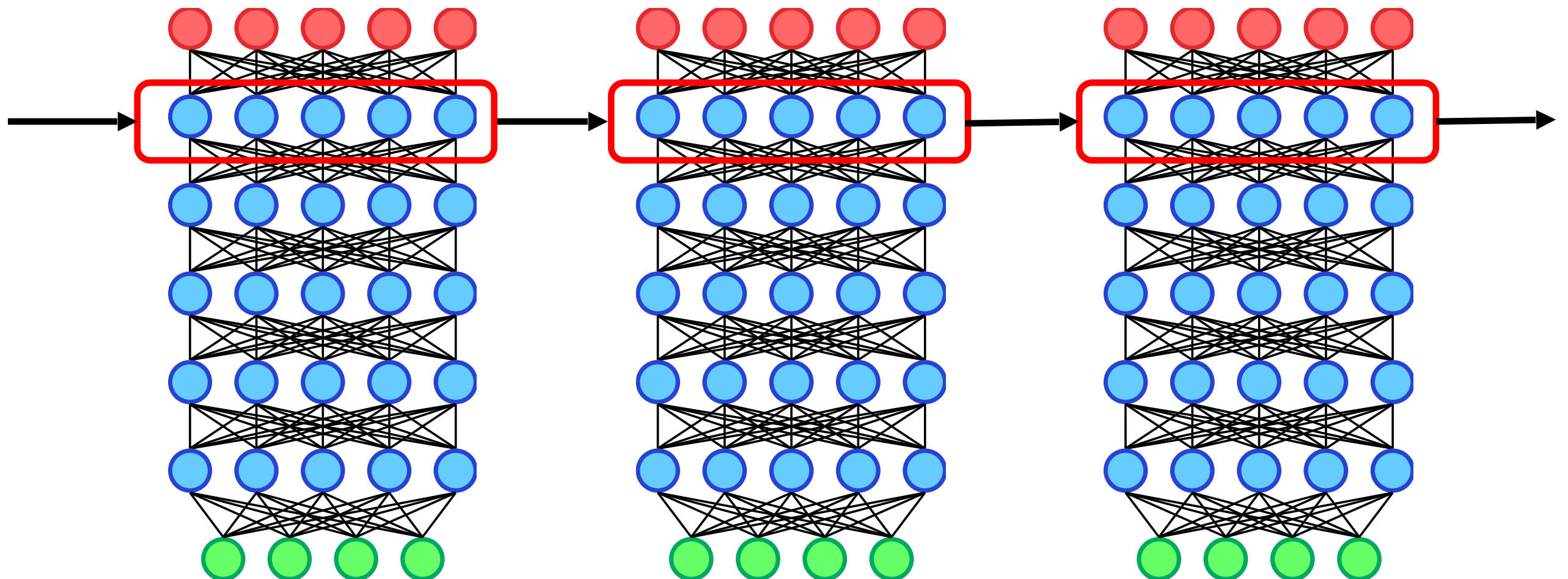
(a) without skip connections



(b) with skip connections

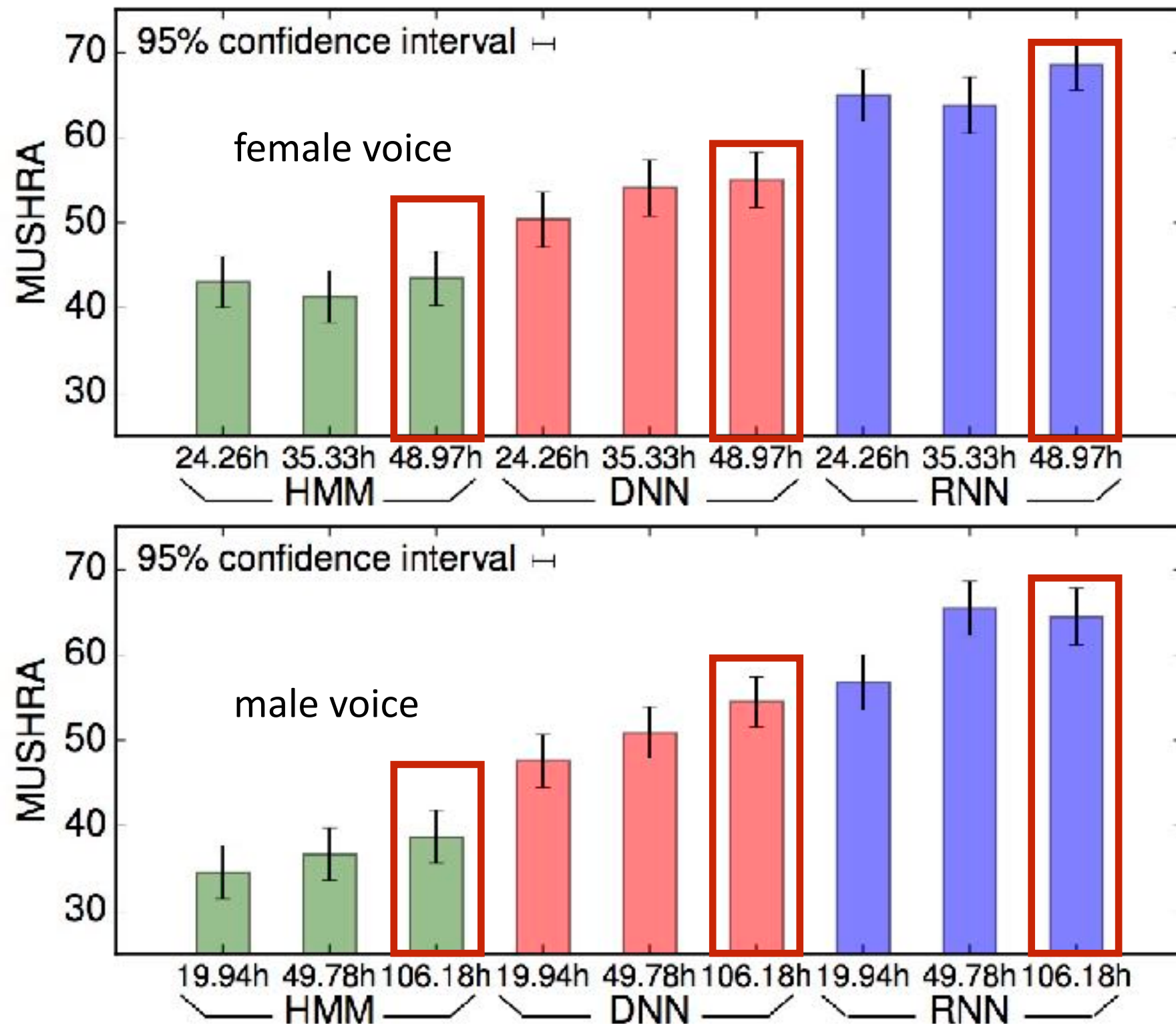
Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The vertical axis is ²¹

Recurrent neural network



HMM, DNN, RNNの比較

主観評価



最近の音声合成：文章と音声波形を直接対応付け

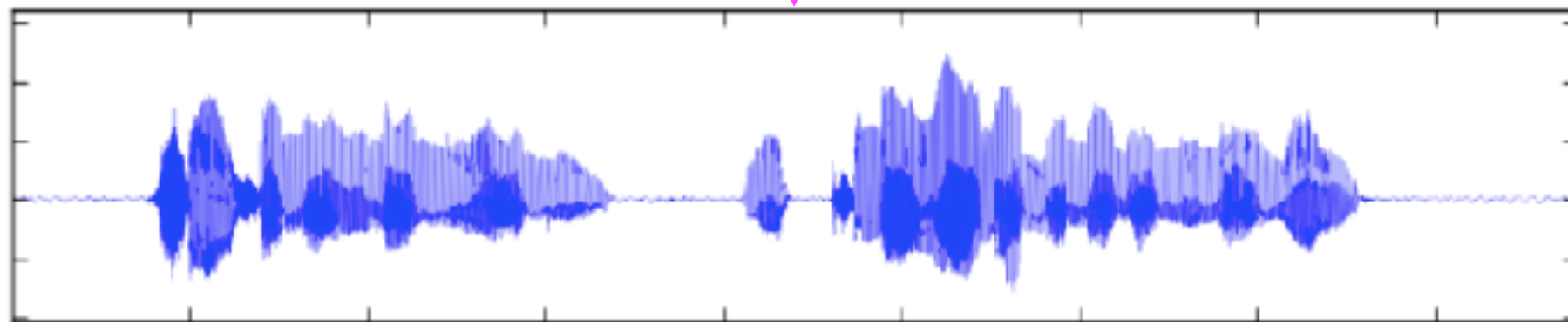
小さな鰻屋に、熱気のようなものがみなぎる

文章の言語的情報抽出

自然言語処理

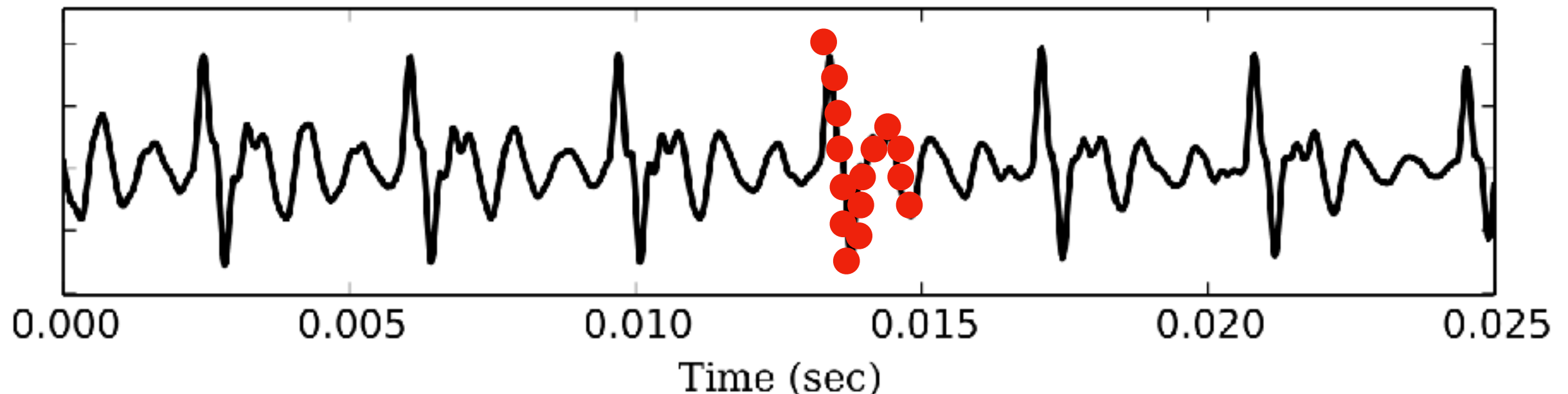
言語情報と音声波形との対応付け

機械学習

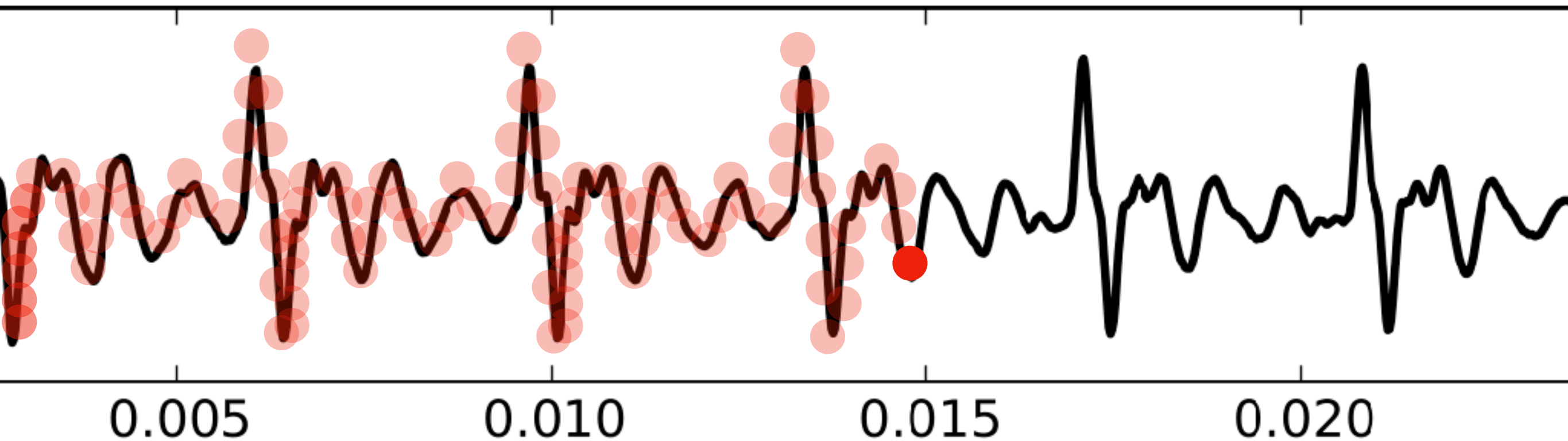


音声波形の学習と自己回帰(AR)

- 音声波形の生成は従来では信号処理が使われていた
- 音響特徴量抽出過程は不可逆であり、合成された音声の品質に上限
- 音声波形をそのまま学習・予測するニューラルネットワークも提案
- 16kHzの場合、1秒間に16,000個のデータが存在する
 - キーポイント：AR性
- 従来の信号処理でも線形のAR性が利用されて来た: LPC

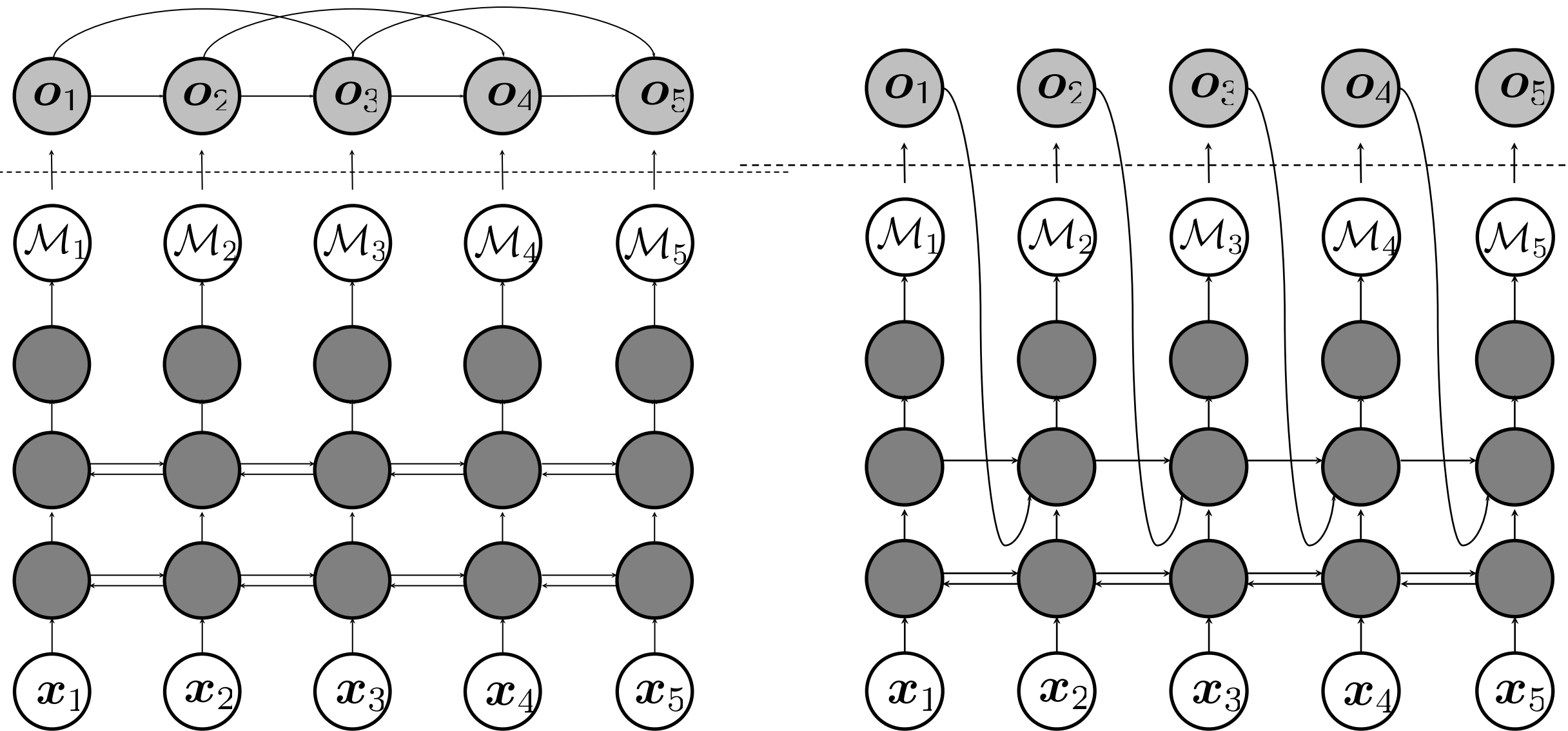


ニューラルネットワークを利用した非線形自己回帰



- 現在の波形点が、過去数百～数千点のデータポイントの非線形自己回帰であらわせると仮定
- その非線形関係をニューラルネットワークで学習

RNNを利用したARモデル

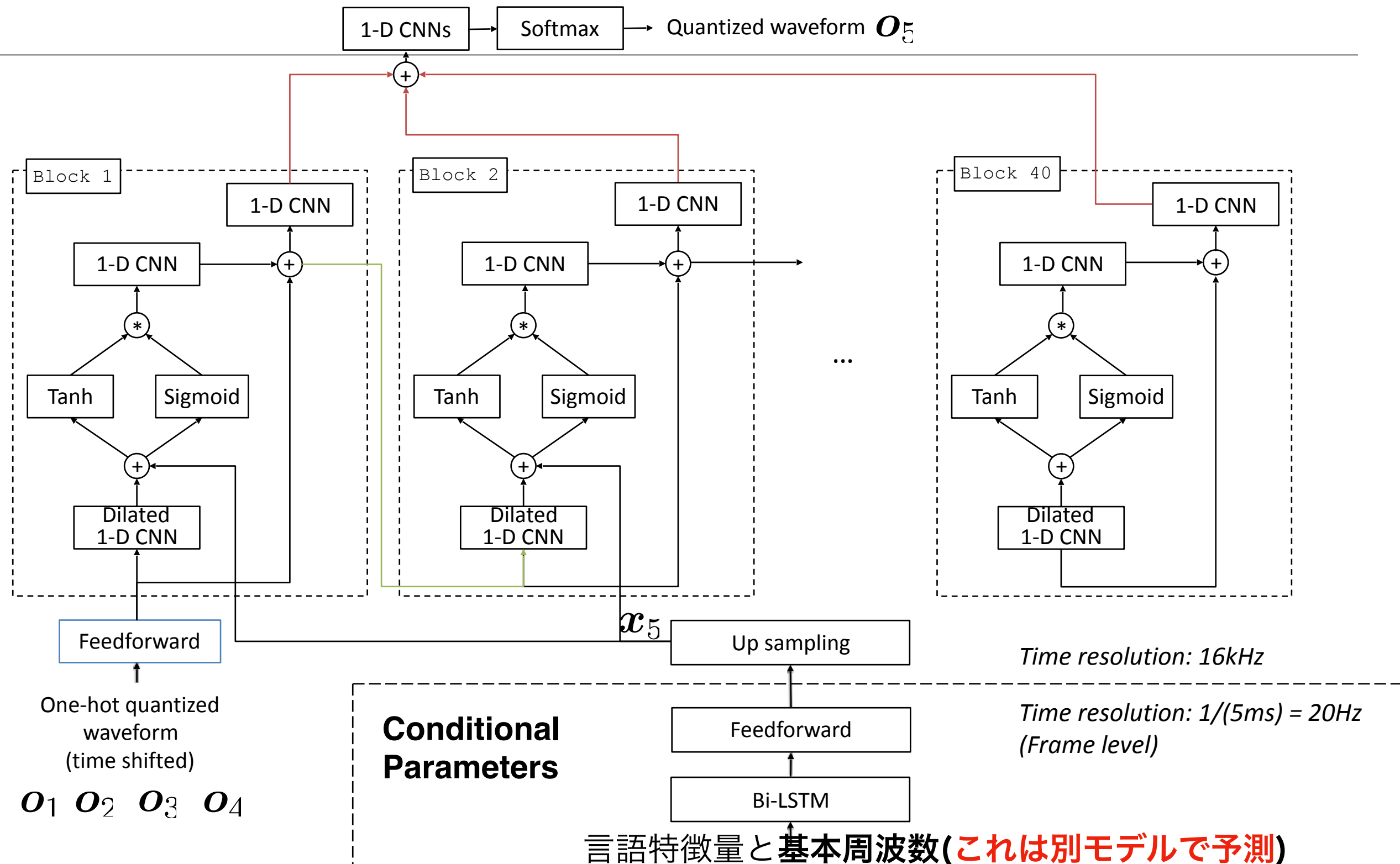


$$p(o_{1:T}; \mathcal{M}_{1:T}) = \prod_{t=1}^T p(o_t | o_{1:t-1}; \mathcal{M}_t)$$

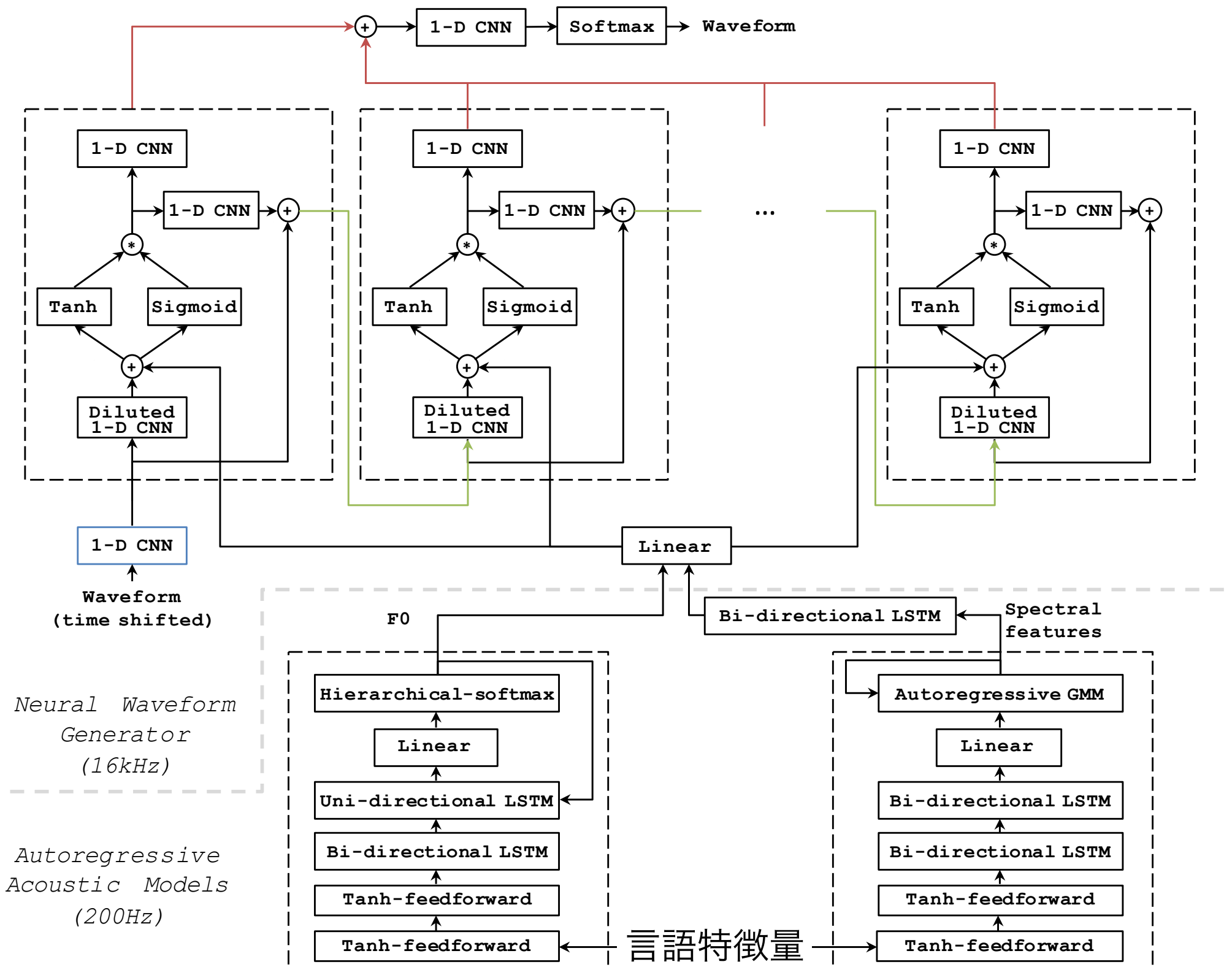
Xin Wang, Shinji Takaki, Junichi Yamagishi, "AN AUTO REGRESSIVE RECURRENT MIXTURE DENSITY NETWORK FOR PARAMETRIC SPEECH SYNTHESIS", icassp 2017 2017年3月

Xin Wang, Shinji Takaki, Junichi Yamagishi, "An RNN-based Quantized F0 Model with Multi-tier Feedback Links for Text-to-Speech Synthesis", Interspeech 2017 2017年8月

CNNを利用したARモデル:Wavenet (Google)



局所条件変数もARモデルで学習



局所条件変数も AR-RNNで学習

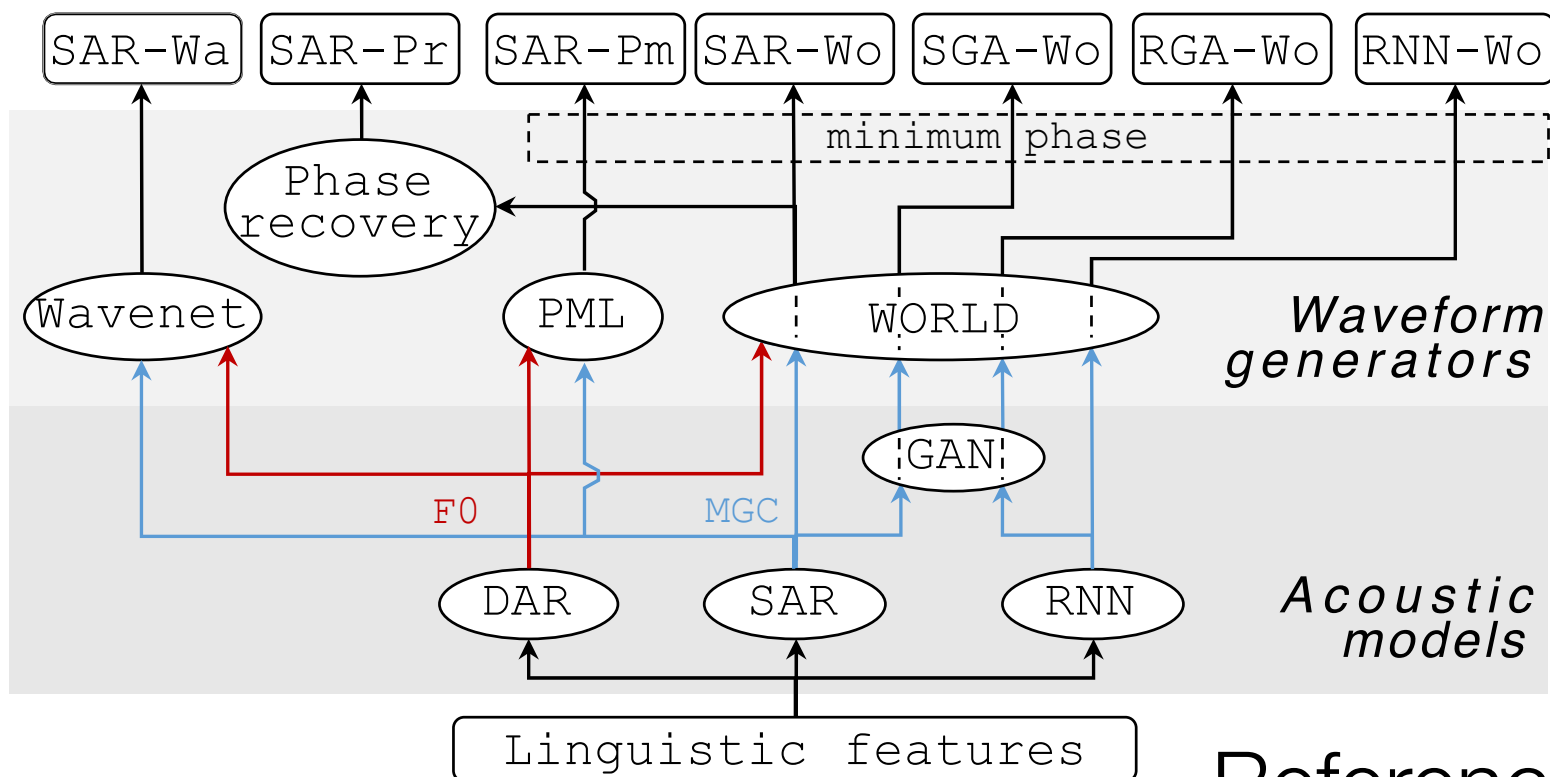
抑揚の予測精度
r=0.94

合成音声

本人

おまけ

本当にこれが信号処理やGANよりも良いのか？



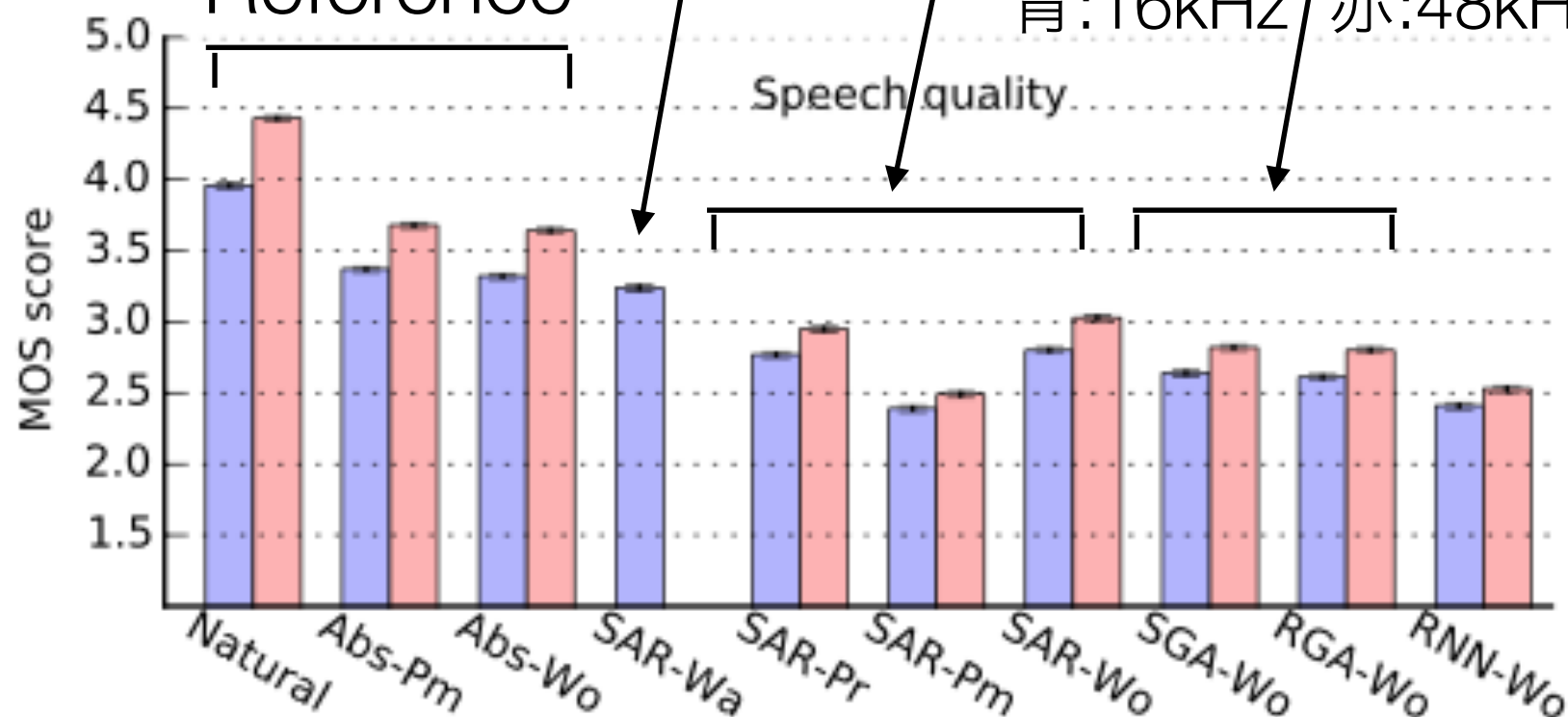
Wavenet

信号処理

GAN

Reference

青:16kHz 赤:48kHz

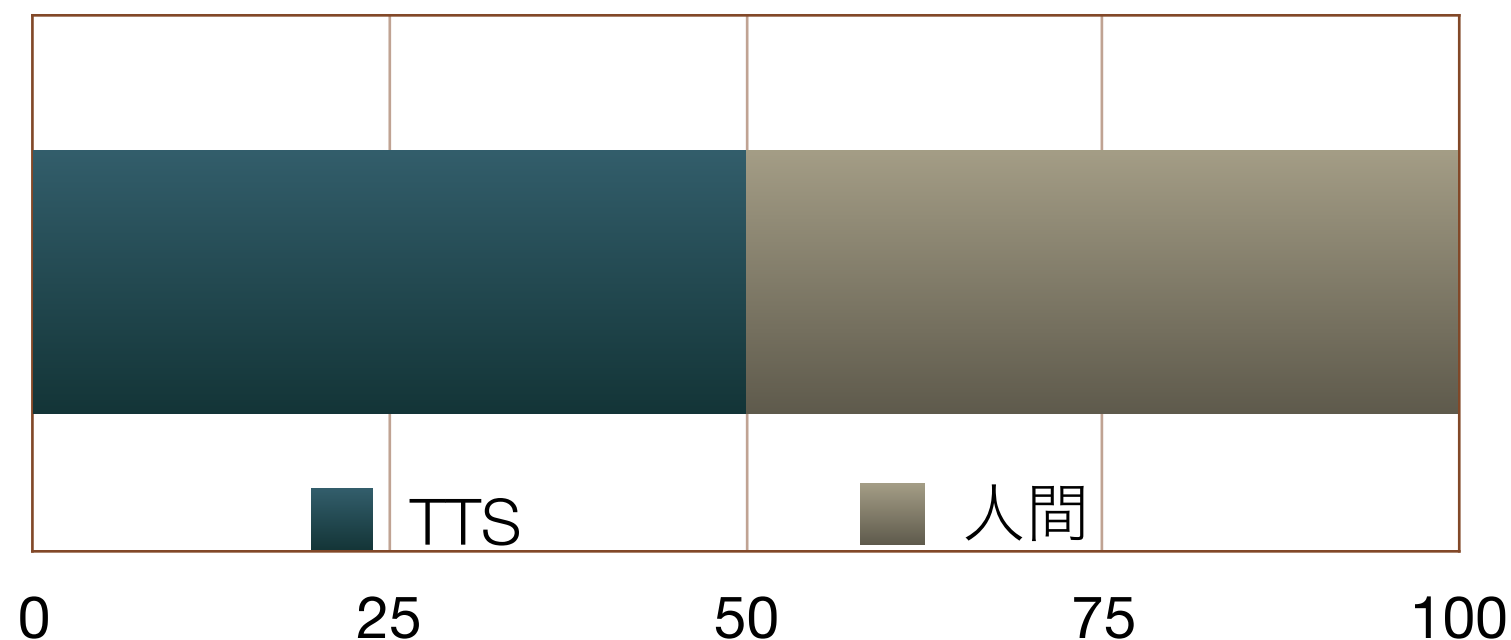


Wavenetのが合成音声の中で自然性が一番高い

簡易チューリングテスト

- 100名の被験者
- 同じ1文章を読み上げた人間の音声と合成音声の2つの音声を聞き、どちらがコンピュータが生成した音声かを回答, 正答率を%で表示
- 1文章からではどちらが合成音声でどちらかが人間か区別がつかない

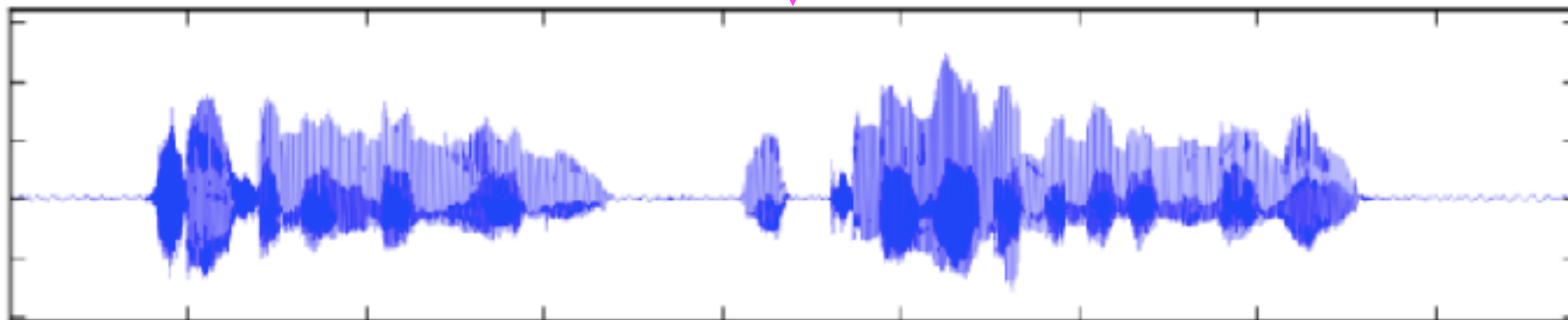
ニューラルネットワークを利用した日本語テキスト音声合成の簡易チューリングテスト



End-to-end 音声合成

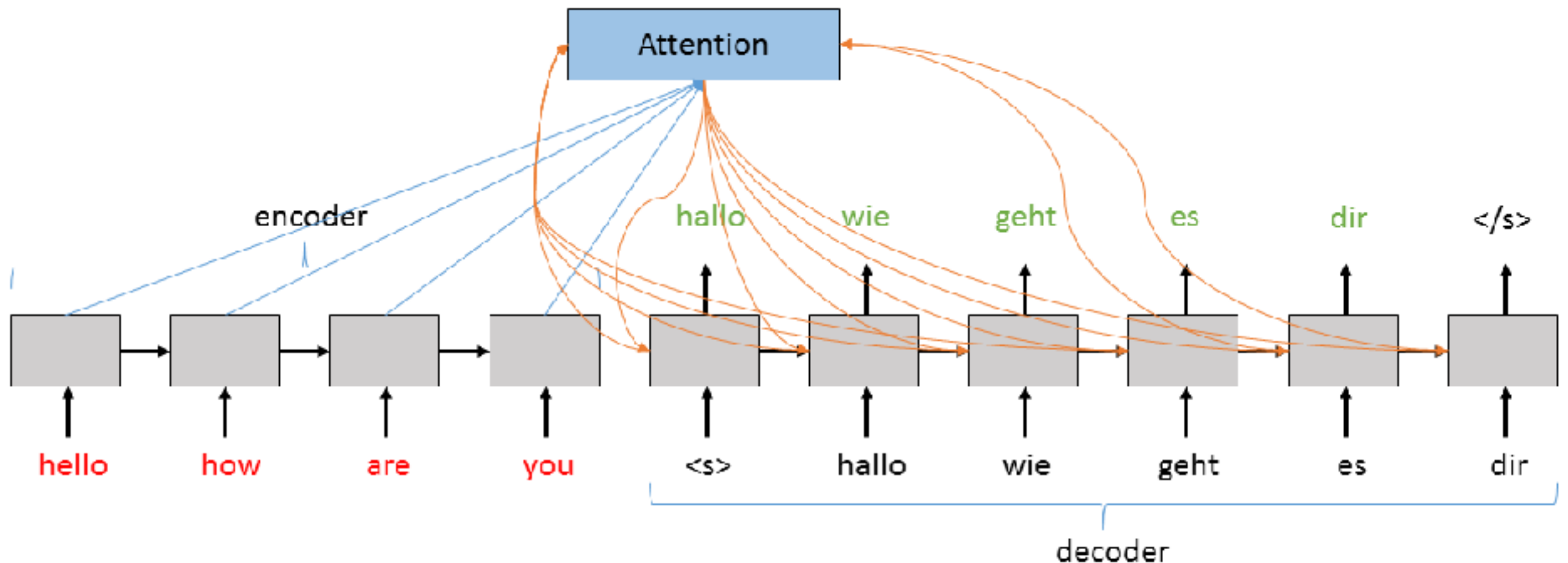
“The cat in the hat”

文章と音声波形との直接対応付け



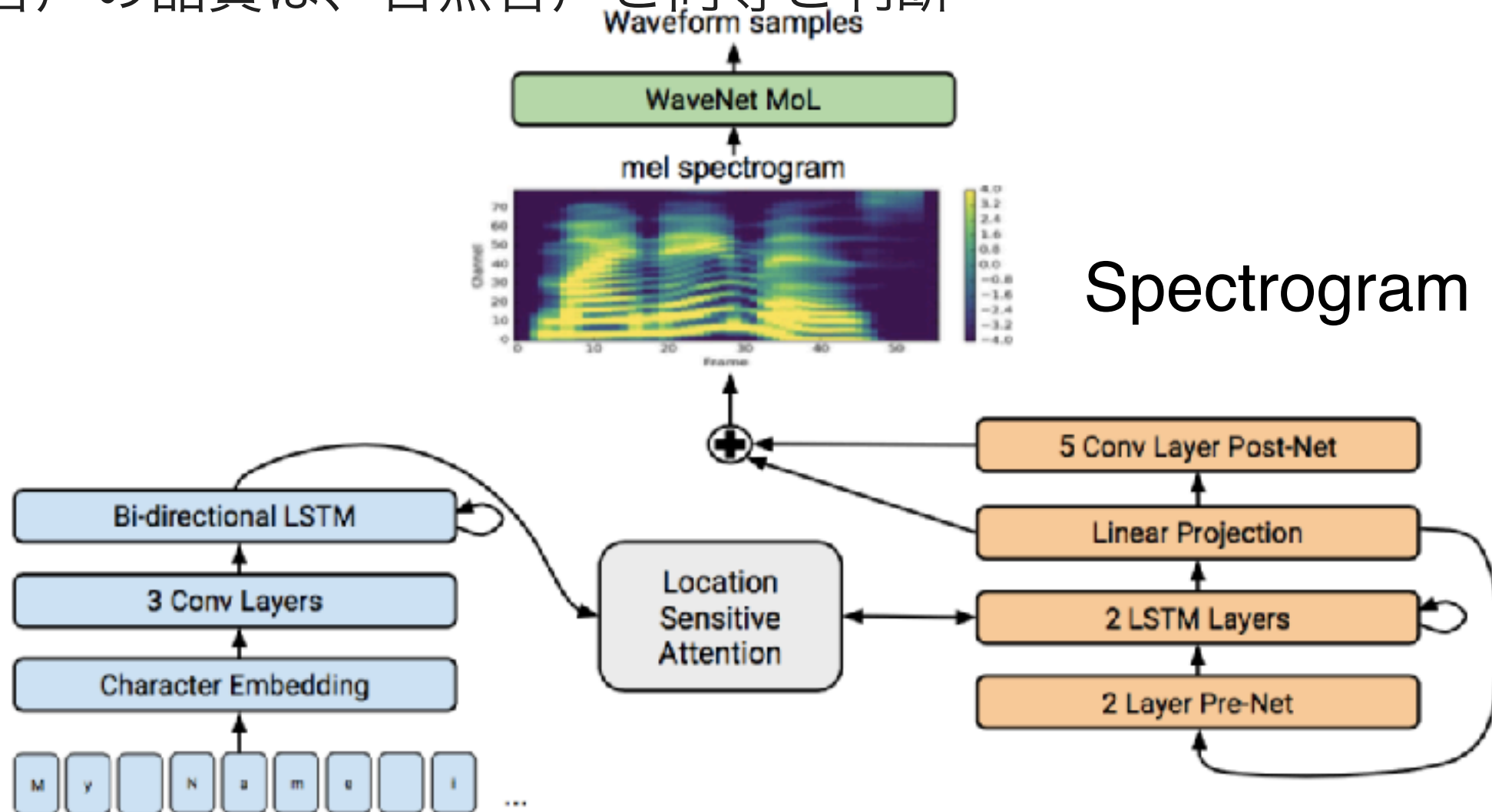
Sequence-to-sequence models

- Sequence-to-sequence neural network models
- 機械翻訳によく用いられる
- 入力言語の単語系列を、単語の数や順序のことなる言語の文字列へ変換



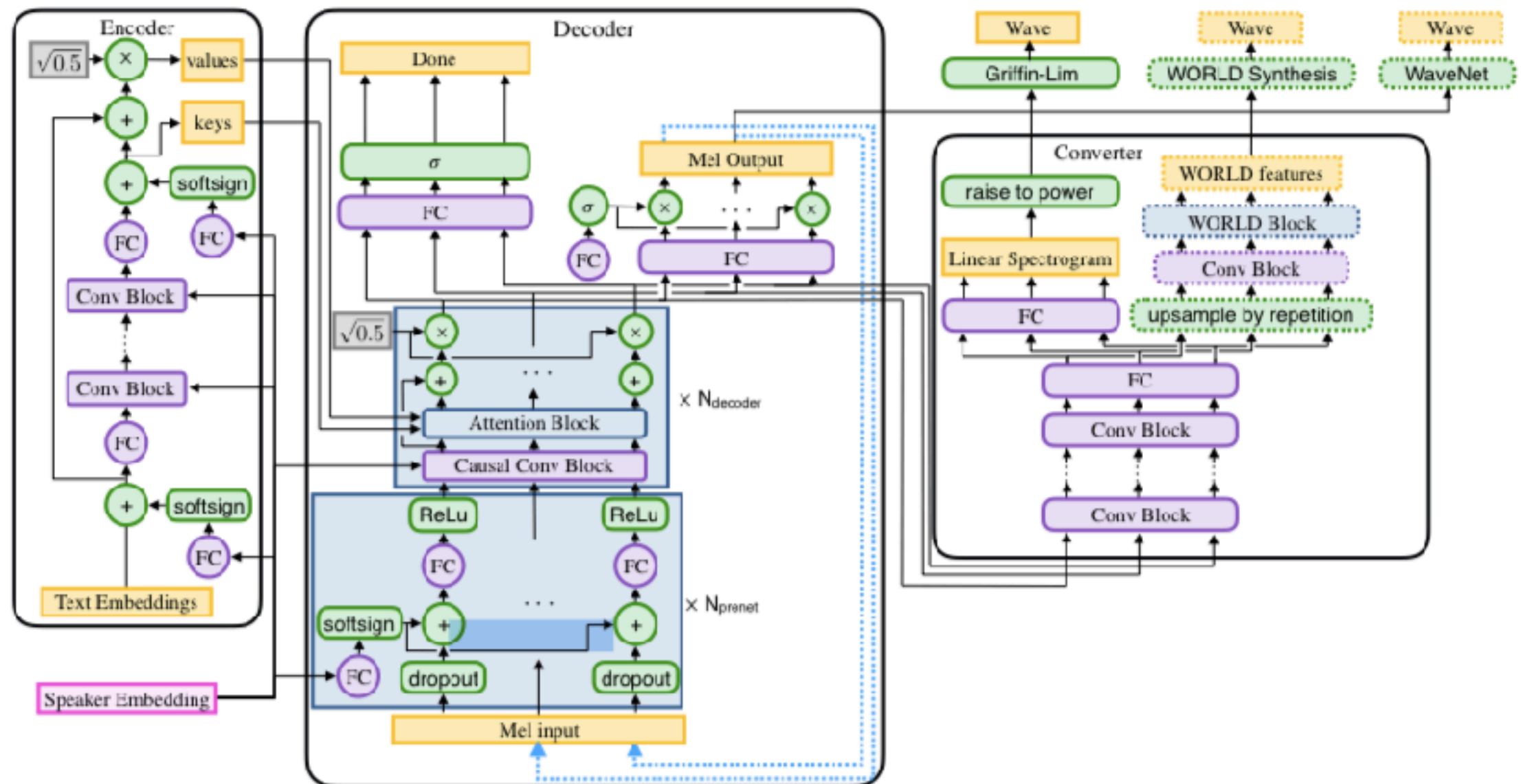
Tacotron2 (Google)

- Seq2seq modelsを利用し、文字列を直接入力、音声波形を出力
- 例えるなら、文字列と音声波形系列の”機械翻訳”
- 合成音声の品質は、自然音声と同等と判断



複数話者音声合成

- 一つのモデルで複数の人の声を合成
- Deep Voice3 from Baidu (Tacotron2 + dot-product attention + **speaker embedding**)



よく使われる音声コーパス

- CSTR VCTK Corpus:
 - 109名の英語話者のコーパス
 - 無償
 - 無響室収録
 - <http://dx.doi.org/10.7488/ds/1994>



INFORMATION SERVICES

Contact us

🏠 ホーム / College of Science & Engineering / School of Informatics / Centre for Speech Technology Research (CSTR) / VCTK / アイテム表示

CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit

No Thumbnail

Citation

Veaux, Christophe; Yamagishi, Junichi; MacDonald, Kirsten. (2017). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). <http://dx.doi.org/10.7488/ds/1994>.

検索



- ☒ サイト検索
- ☐ このコレクション

登録利用者

ログイン

登録

本日の構成

- 深層学習による音声合成
- **深層学習による音声認識**

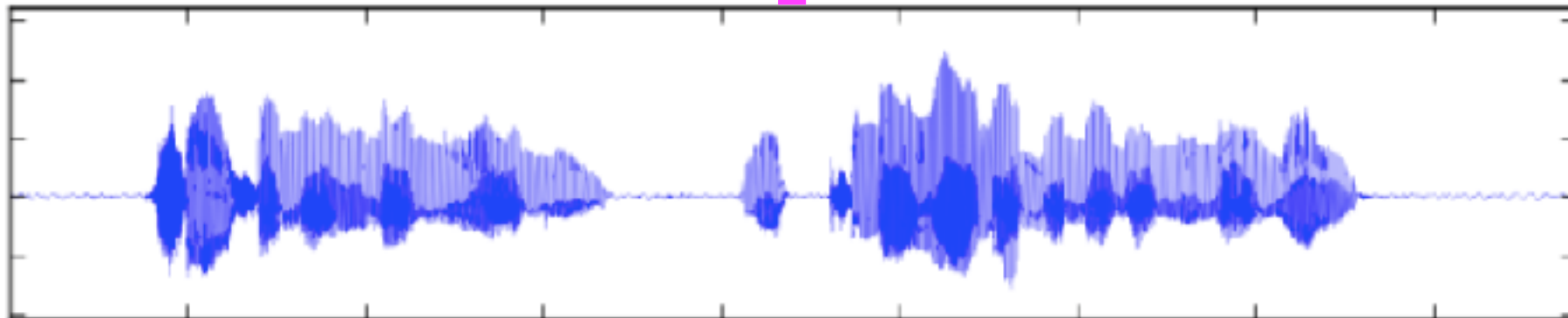
従来の音声認識のプロセス

“The cat in the hat”

言語モデルによるデコーディング

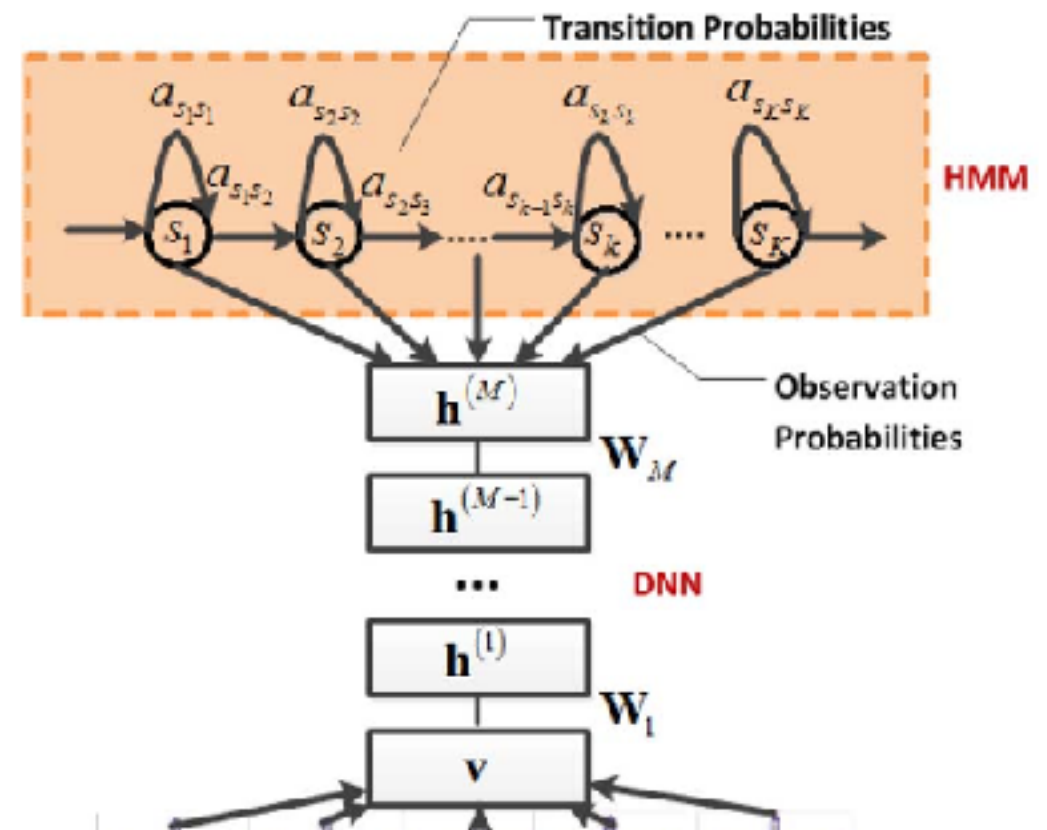
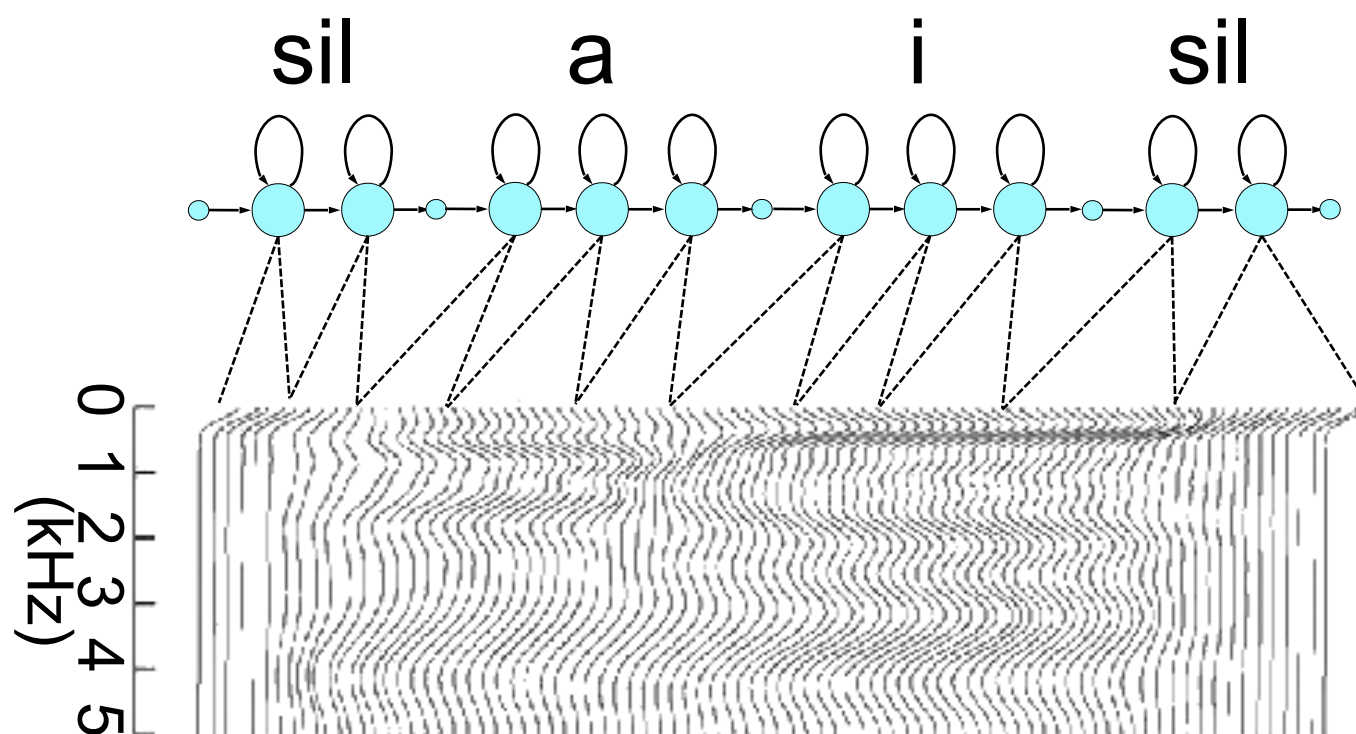
音響モデル

音響特徴量

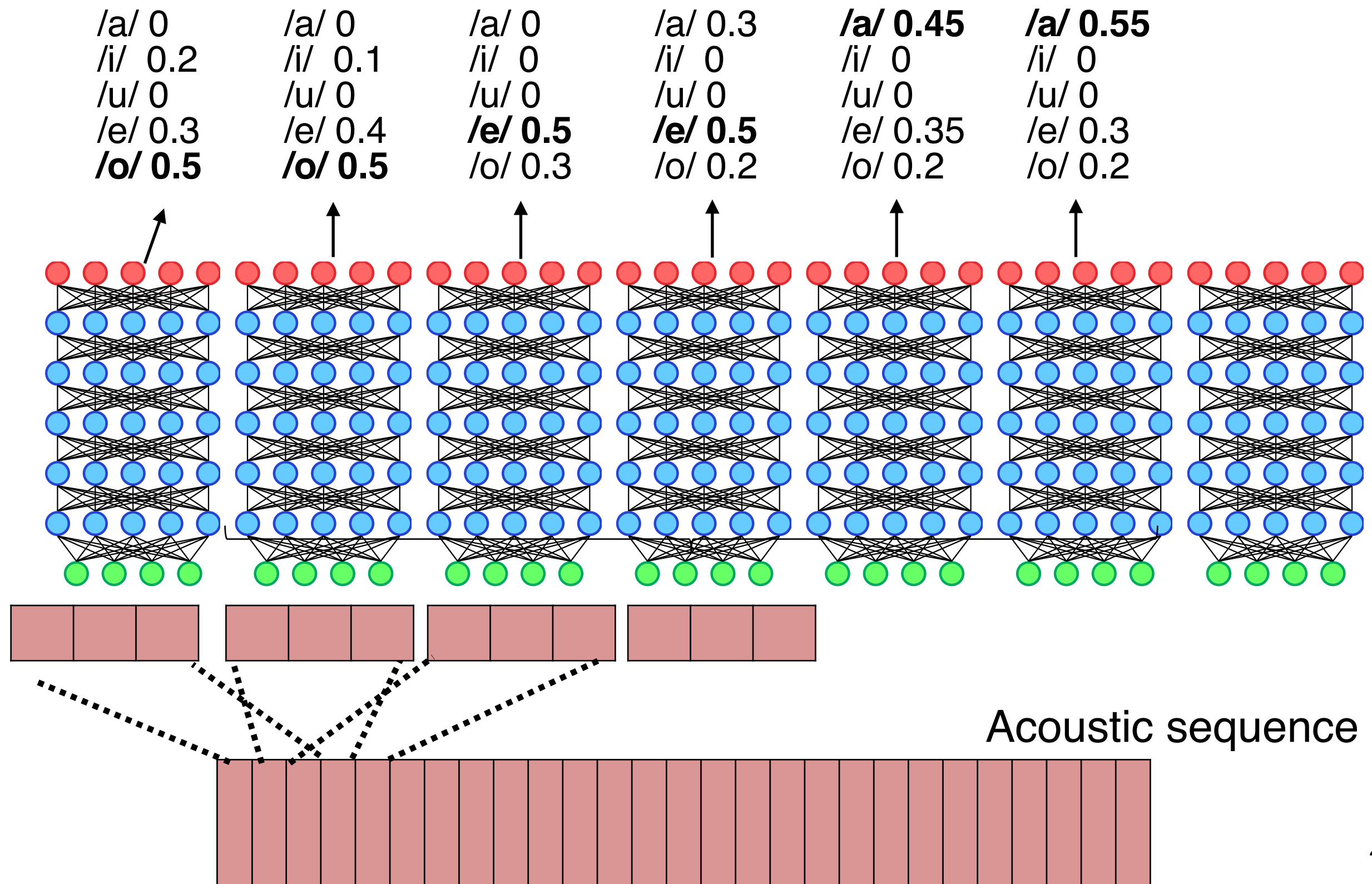


音声認識の歴史

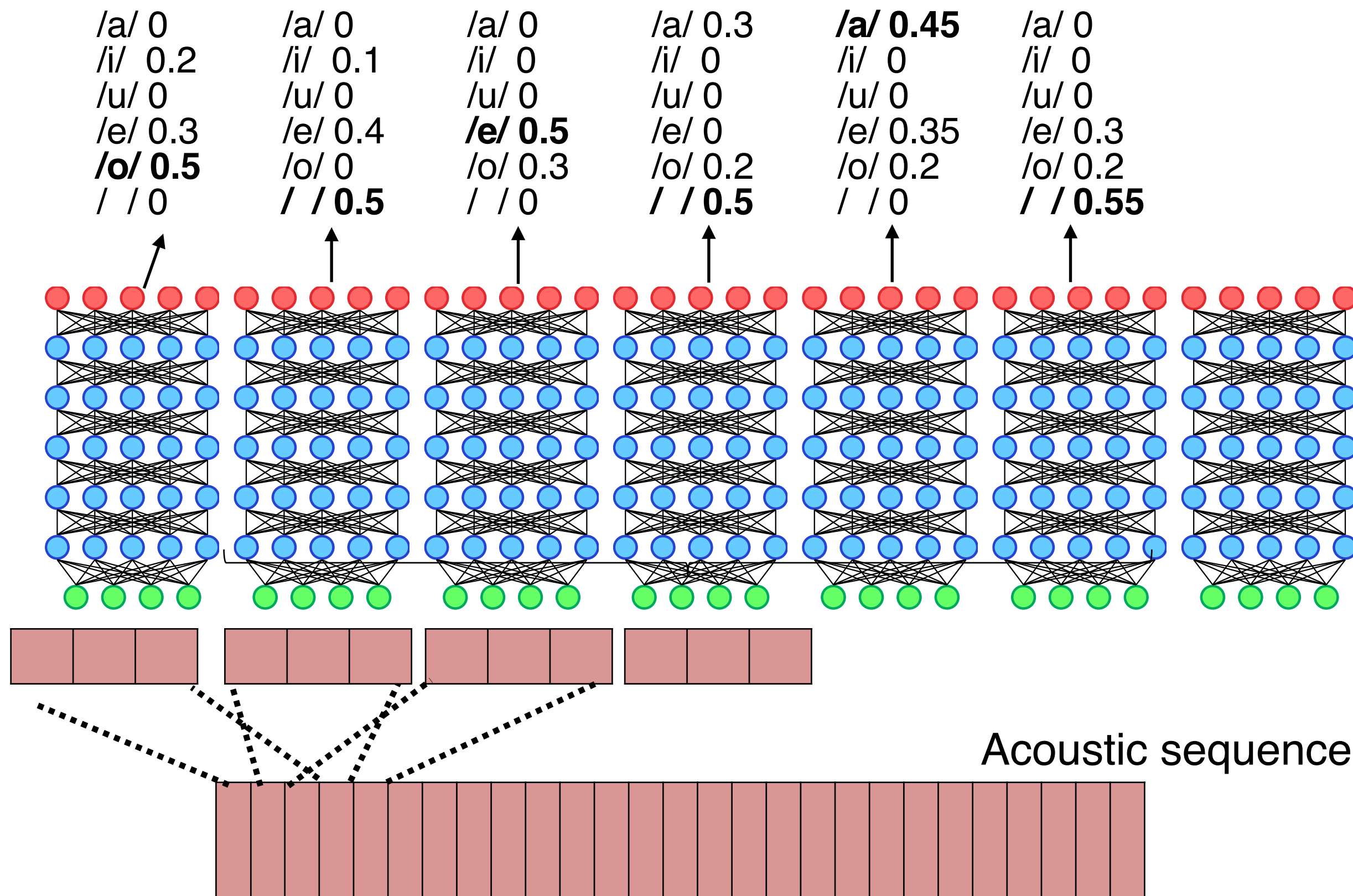
- 1970s: Dynamic Time Warping (DTW)
- 1980s: Hidden Markov models
- 1990s: Large vocabulary continuous speech recognition (LVCSR)
- 2000s: Discriminative training (minimize word/phone error rate)
- 2010s: Hybrid of HMM and deep learning (product)
- **Current: end-to-end ASR (research)**



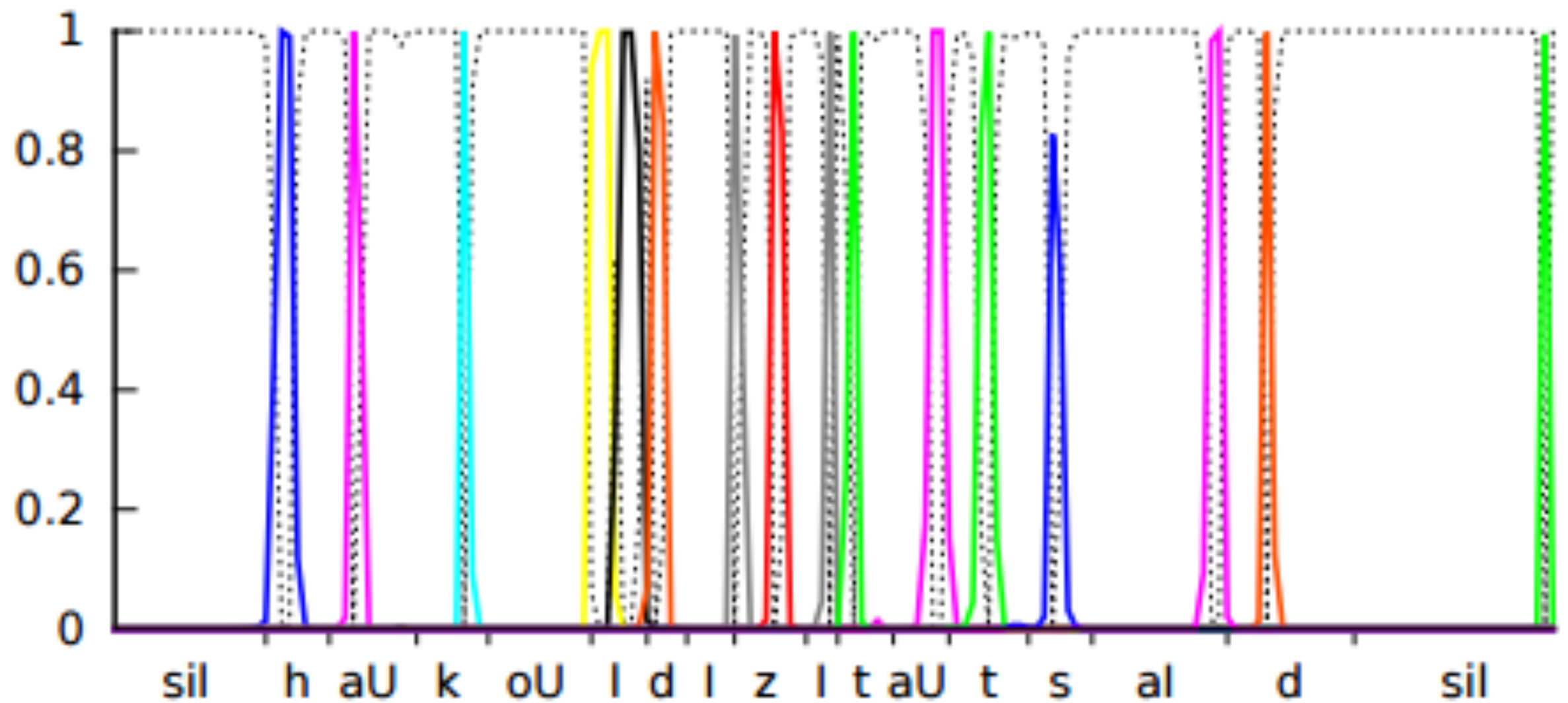
各時刻の各音素の確率をDNNで予測



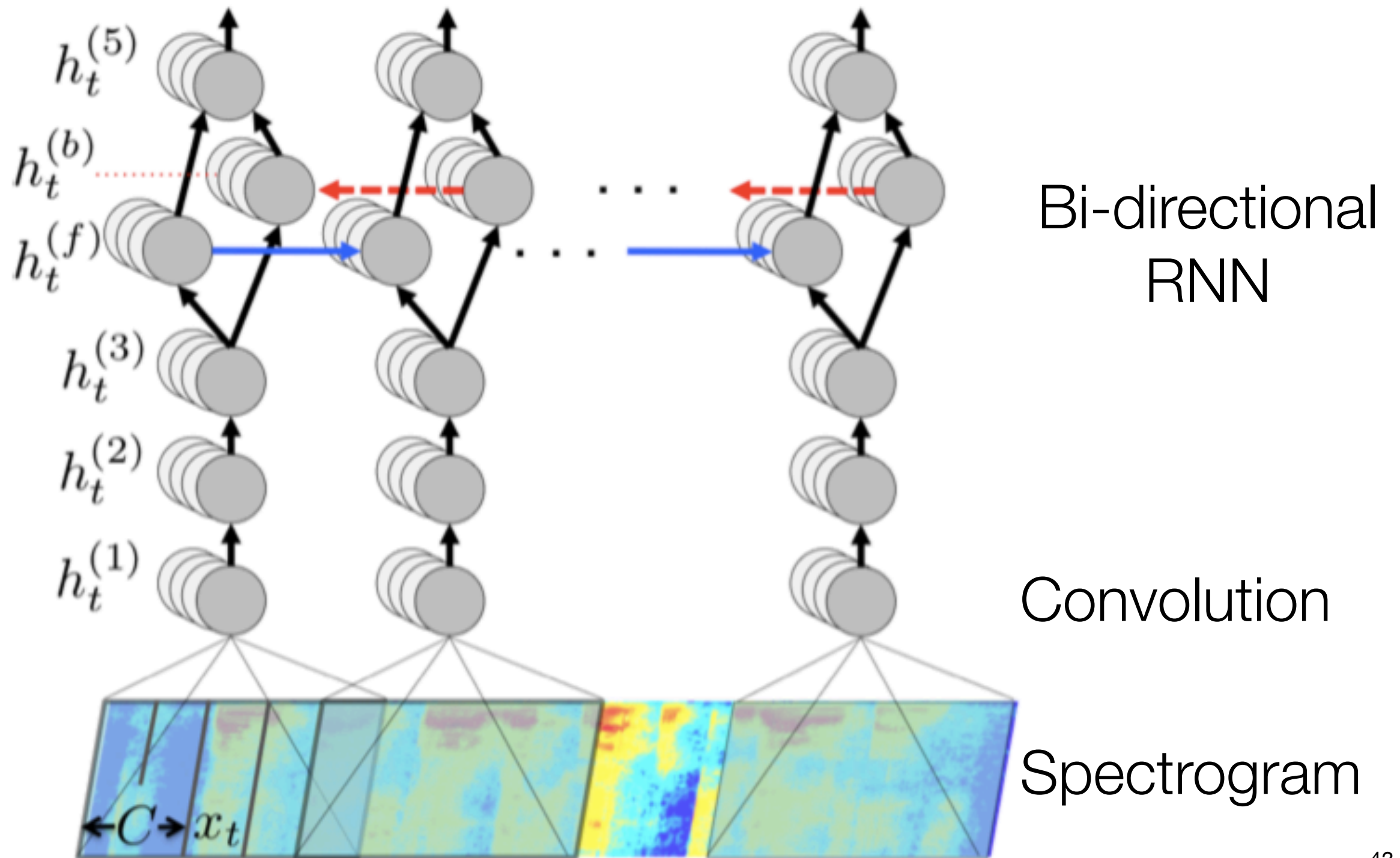
前の音素の継続を表すヌル記号 / / を導入
発話全体で学習・予測 (CTCと呼ばれる)



Output: Phone activation patterns over time



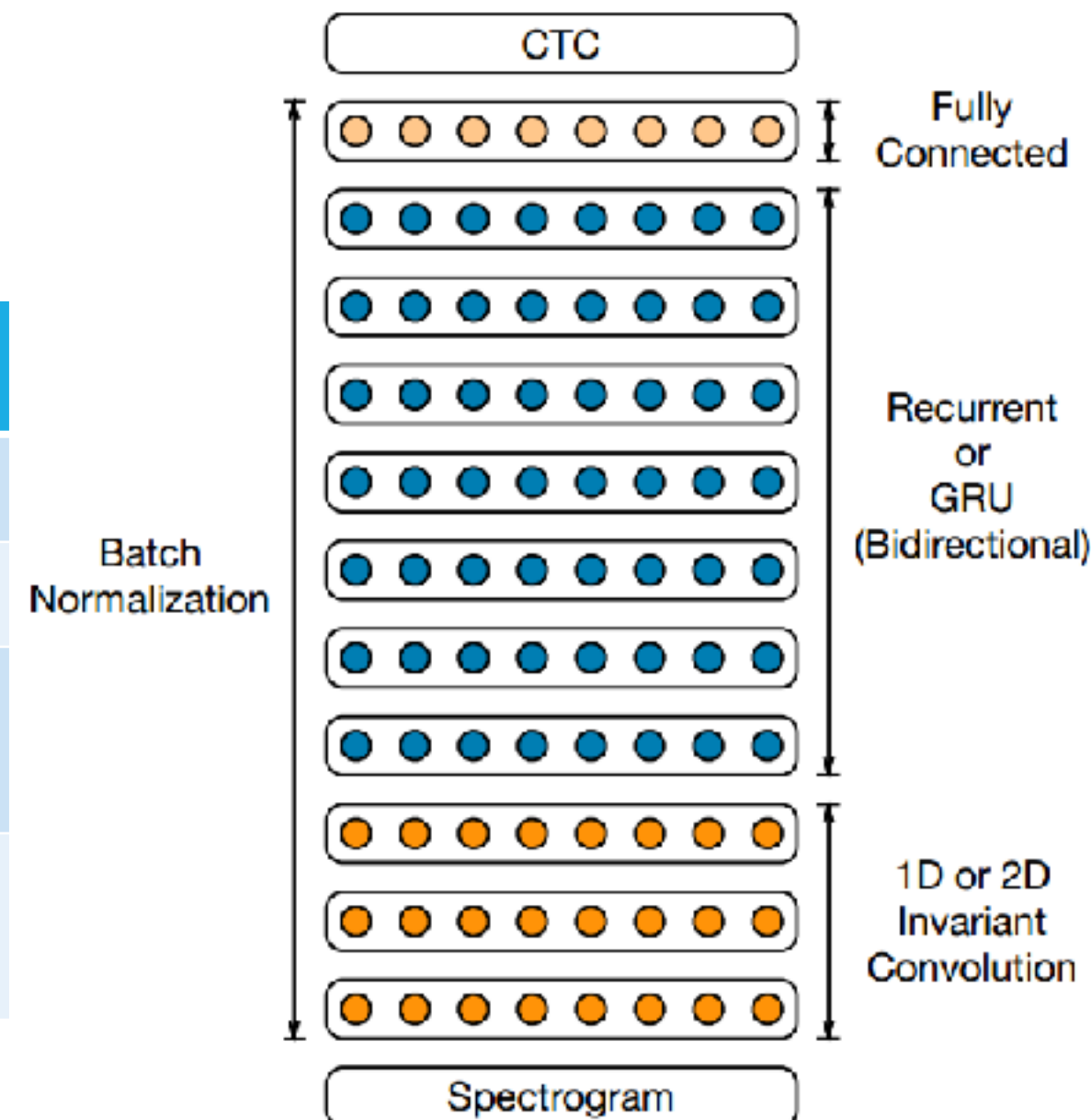
最近のASR用DNNの例: Deep Speech



Deep Speech 2 (Baidu Research)

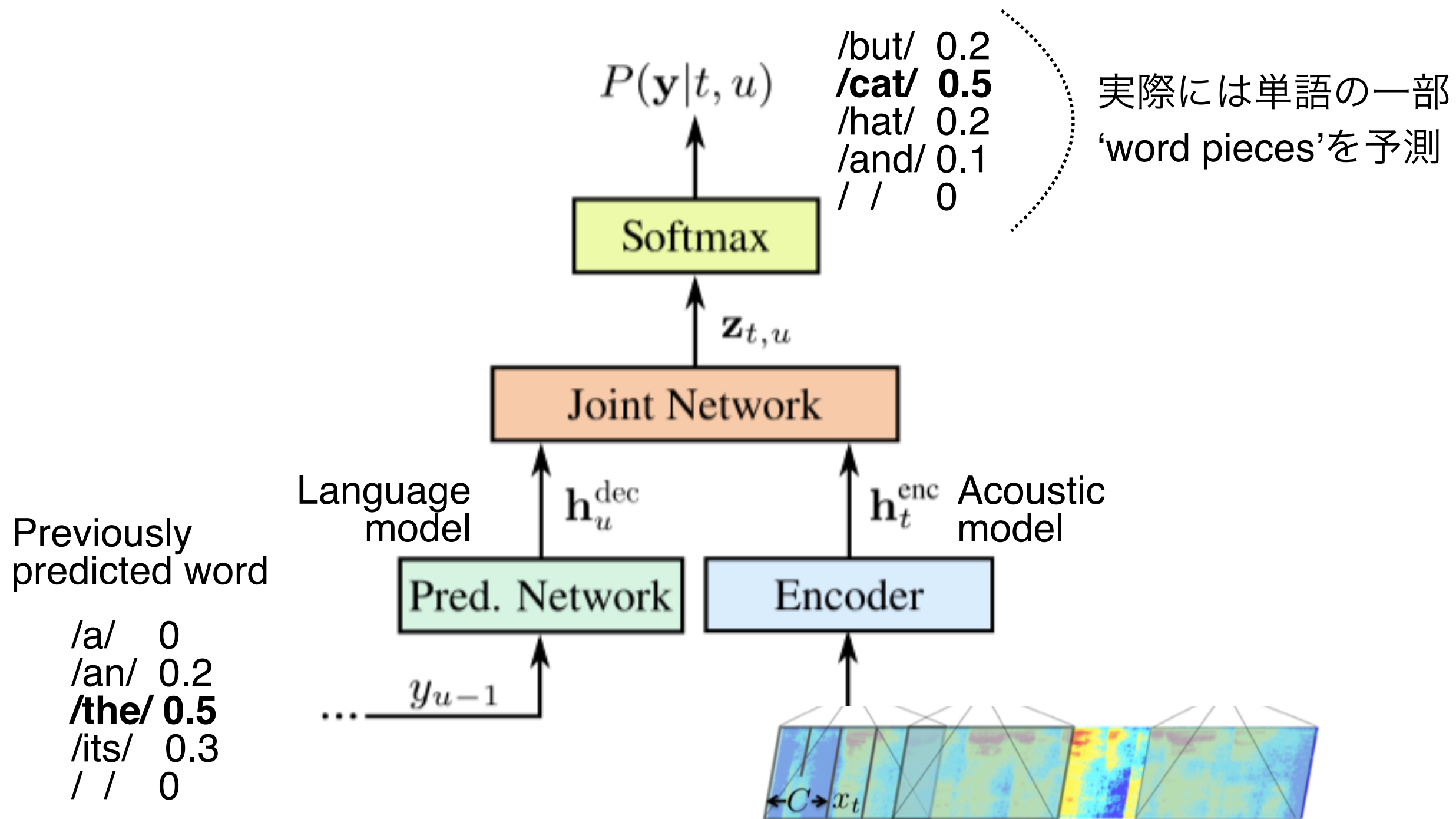
- 3 層の2D convolution
- 7 層の双方向リカレント
- 単語誤り率

Test set	Deep speech 2	Human
WSJ eval'92	3.60	5.03
WSJ eval'93	4.98	8.08
LibriSpeech test-clean	5.33	5.83
LibriSpeech test-other	13.25	12.69

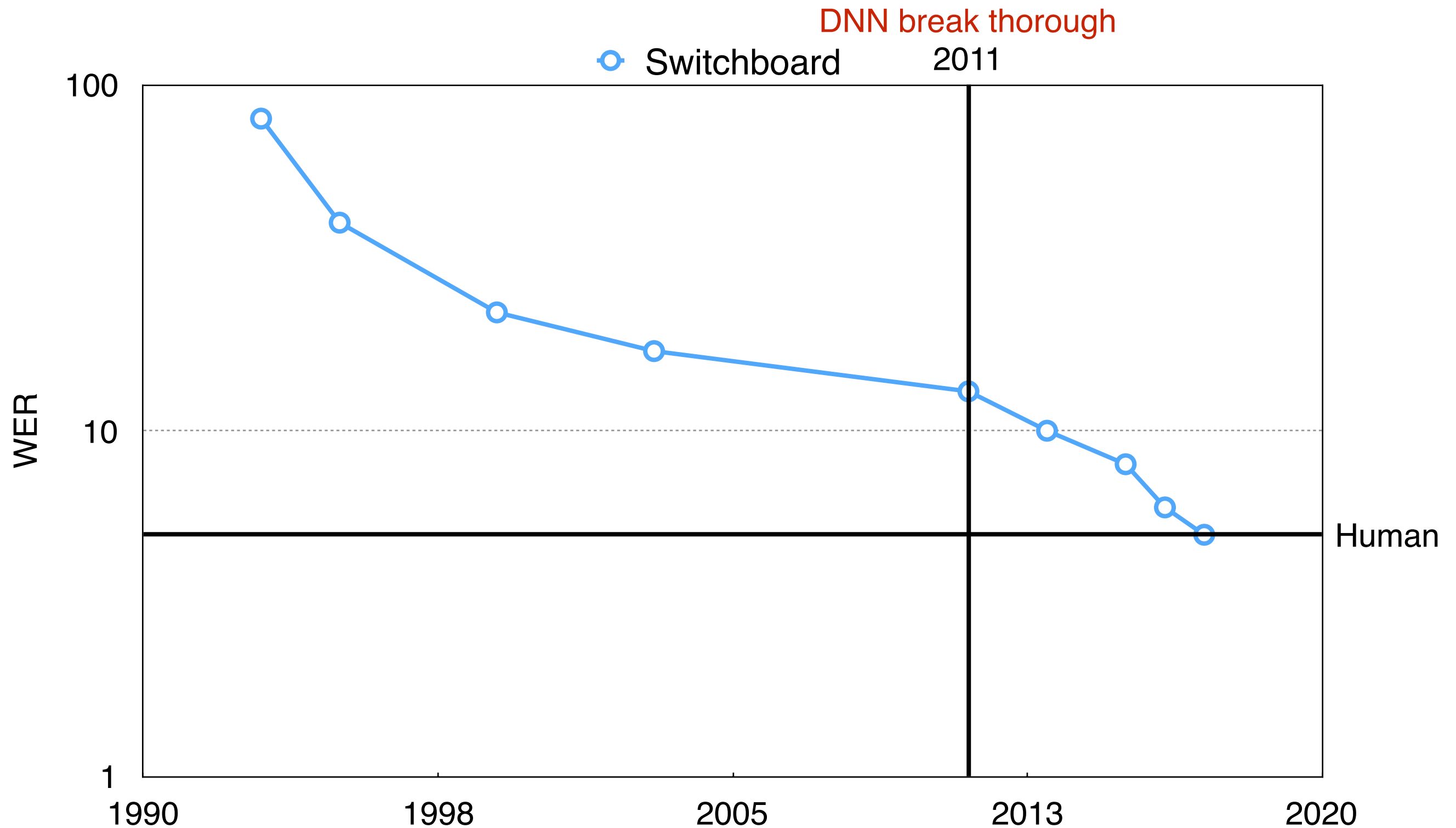


音素ではなく、単語を直接出力する

“Recurrent neural transducer”も提案



Switchboard databaseの性能向上



まとめ

- 深層学習を利用した音声合成
 - 音響特徴量系列と言語特徴量系列を対応づける音響モデルだけでなく
 - 音声波形生成
 - テキスト処理
 - も全てDNNで行うend-to-end学習が成功
 - 音声品質も劇的に向上
 - 人間の品質とほぼ同じ
- 深層学習を利用した音声合成
 - DNNにより単語誤り率は激減、人間の単語認識精度と同レベル
 - end-to-end学習によりモデル構造は非常にわかりやすくなった