電子情報通信学会シンポジウムセッション
もっと知りたい！ Deep Learning ～基礎から活用まで～

HITACHI
Inspire the Next

音声認識分野における ディープラーニングの基礎と最新動向

2017/3/22

神田 直之

(株)日立製作所　研究開発グループ
システムイノベーションセンタ　メディア研究部

自己紹介：神田直之

HITACHI
Inspire the Next

2006年4月～

(株)日立製作所 中央研究所 入社

- 音声認識、大規模音声データからのキーワード検出の研究開発
  - 音響モデル、言語モデル、デコーダ

2014年7月～

情報通信研究機構(NICT) 出向

- 音声認識の研究開発（音声翻訳システム）
  - 日英中韓泰緬尼越仏西
  - リカレントニューラルネット型音響モデル、CTCによる音声認識など
- IWSLT2014 音声認識評価トラック１位

2016年10月～

(株)日立製作所 研究開発グループ （兼：NICT 協力研究員）

- 音声認識の研究開発（コールセンタ、ロボット等）

講演の流れ

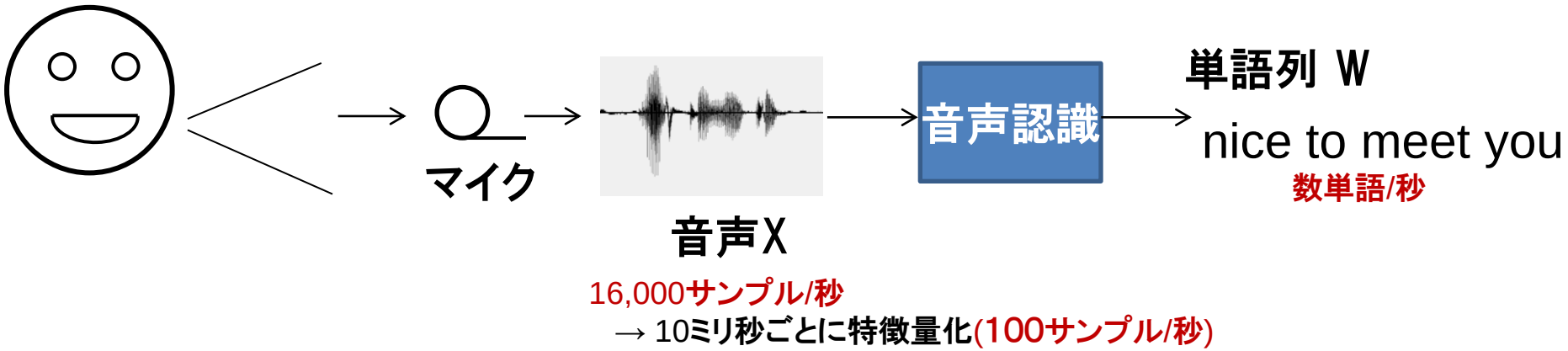
HITACHI
Inspire the Next

- Part1: 音声認識の基礎と、Deep Learningの適用方法
  - 生成モデルアプローチ
  - 識別モデルアプローチ
- Part2: 音声認識分野におけるDeep Learning研究の事例・最新動向

音声認識とは

HITACHI
Inspire the Next

- 音声Xを観測したときに、単語列Wを推定する問題



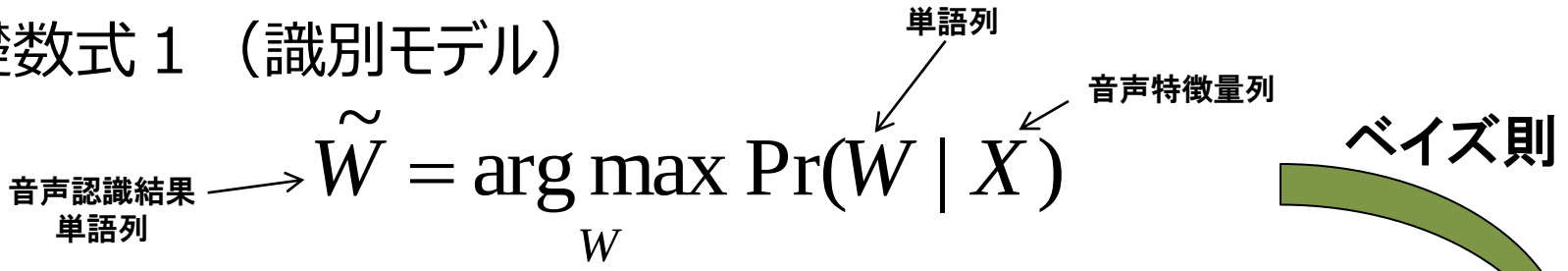
- 音声認識の問題設定
  - 連続値の列からシンボル列を推定する問題
  - 通常、入力系列長＞＞＞出力系列長
  - 外乱：話者の違い、音環境の違い（雑音、残響）、言語の違いなど

直接応用できそうな分野：動画像認識、手書き文字認識、等
その他、Deep Learningのテクニック全般は分野非依存で有効な事が多い

２つの基礎数式

HITACHI
Inspire the Next

- 基礎数式１（識別モデル）



- 基礎数式２（生成モデル）

$$\tilde{W} = \arg\max_{W} \frac{\Pr(X \mid W)\Pr(W)}{\Pr(X)} = \arg\max_{W} \Pr(X \mid W)\Pr(W)$$

Wの最適化に無関係

音響モデル：単語列Wから，音声特徴量列Xが生成される確率

言語モデル：単語列Wが生成される確率

２つの基礎数式

HITACHI
Inspire the Next

- 基礎数式１（識別モデル）

単語列
音声特徴量列
音声認識結果
単語列

$$\tilde{W} = \arg\max_{W} \Pr(W \mid X)$$

ベイズ則

DL以前も以後も
主流の方式

- 基礎数式２（生成モデル）

$$\tilde{W} = \arg\max_{W} \frac{\Pr(X \mid W)\Pr(W)}{\Pr(X)}$$

$\Pr(X)$ はWの最適化に無関係

$$= \arg\max_{W} \Pr(X \mid W)\Pr(W)$$

音響モデル：単語列Wから，音声特徴量列Xが生成される確率

言語モデル：単語列Wが生成される確率

HITACHI
Inspire the Next

生成モデルによる音声認識

生成モデルに基づく音声認識

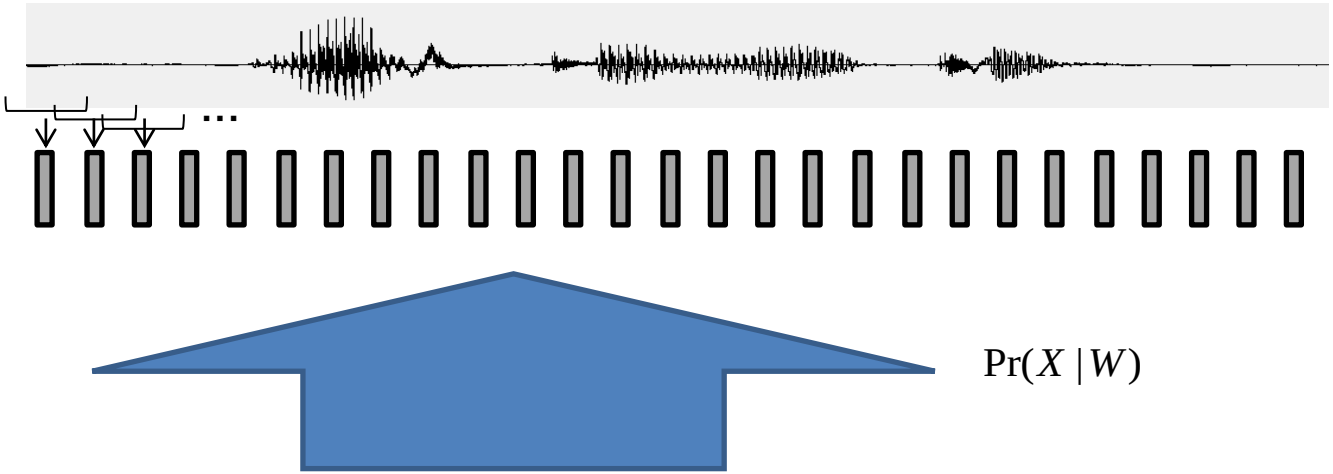
数式

$$\tilde{W} = \arg\max_{W} \Pr(X \mid W) \Pr(W)$$

音響モデル　言語モデル

$$\cong \quad ?$$

模式図



生成モデルに基づく音声認識

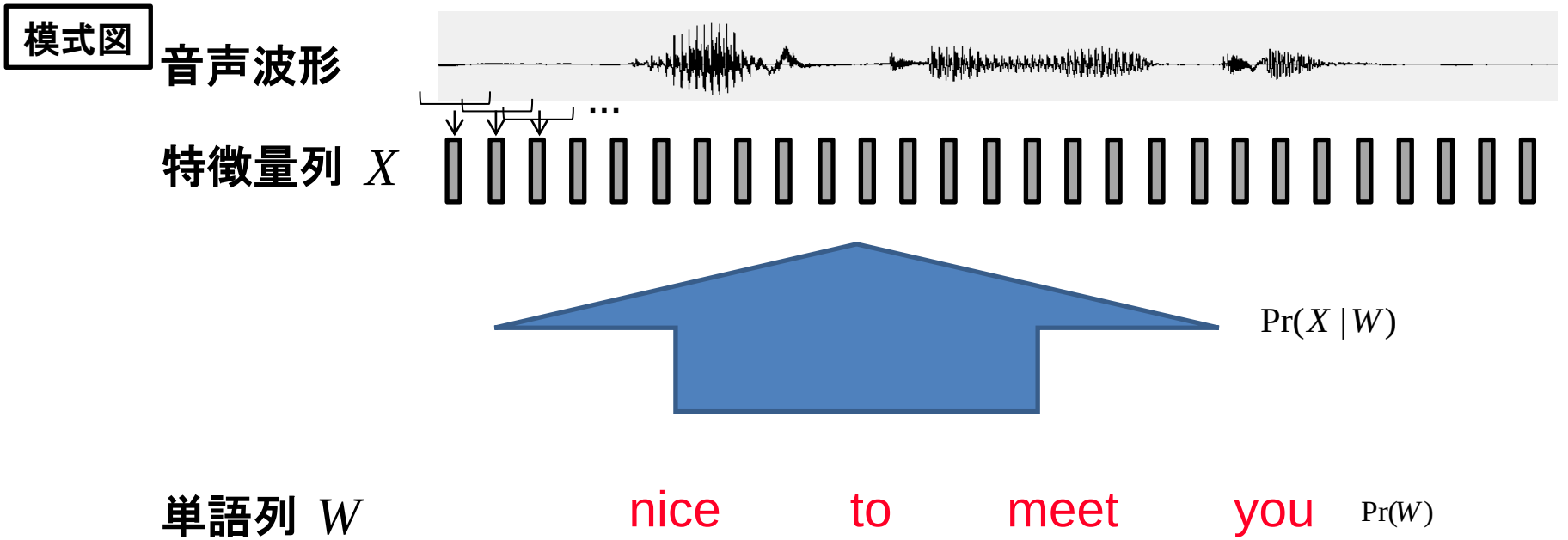
数式

$$\tilde{W} = \arg\max_{W} \Pr(X \mid W) \Pr(W)$$

$$\cong \arg\max_{W} \{\max_{P \in \Psi(W)} \Pr(X \mid P) \Pr(P \mid W)\} \Pr(W)$$

音響モデル　発音モデル　言語モデル

模式図



生成モデルに基づく音声認識

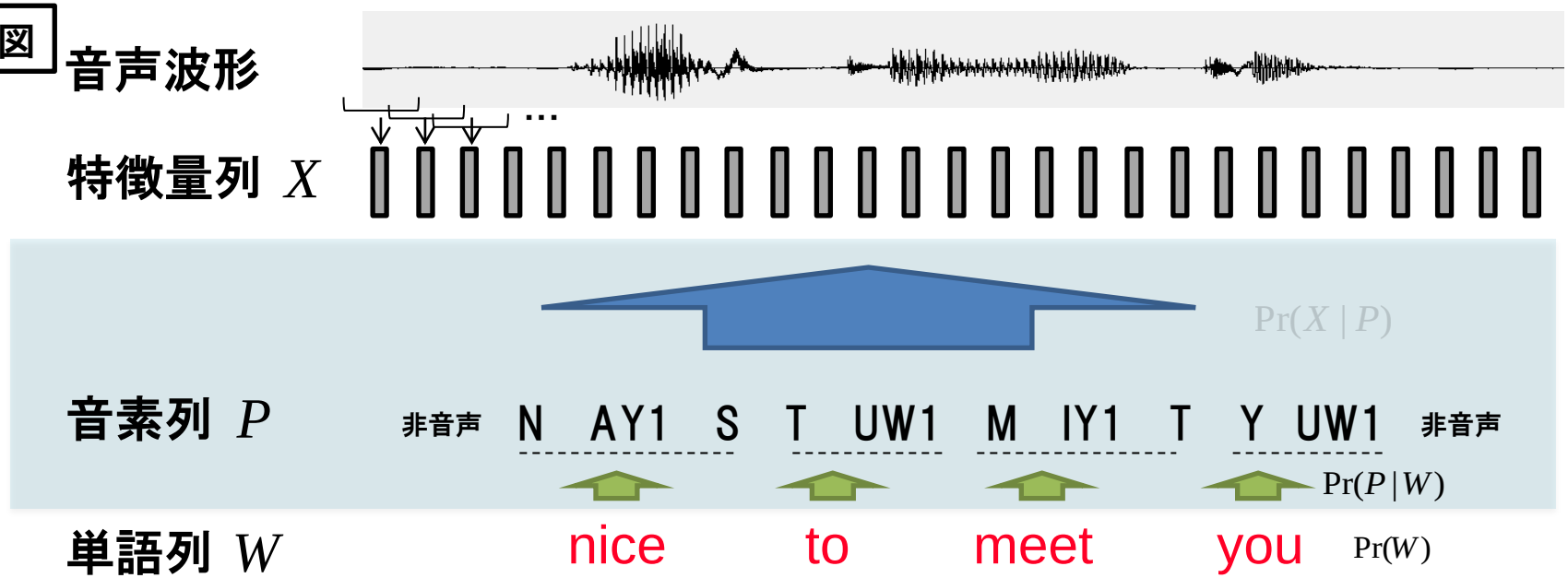
数式

$$\tilde{W} = \arg\max_{W} \Pr(X \mid W) \Pr(W)$$

$$\cong \arg\max_{W} \{\max_{P \in \Psi(W)} \Pr(X \mid P) \Pr(P \mid W)\} \Pr(W)$$

音響モデル　発音モデル　言語モデル

模式図



生成モデルに基づく音声認識

数式

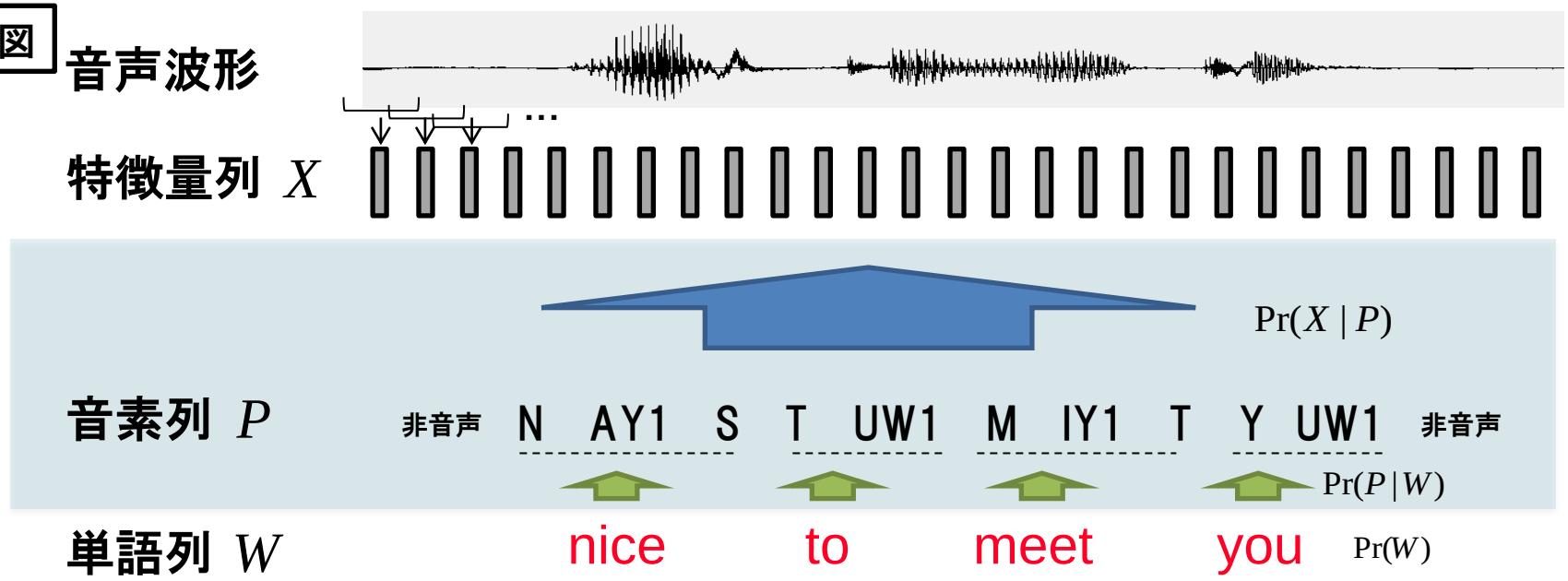
$$\tilde{W} = \arg\max_{W} \Pr(X \mid W)\Pr(W)$$

$$\cong \arg\max_{W} \{\max_{P \in \Psi(W)} \Pr(X \mid P)\Pr(P \mid W)\}\Pr(W)$$

$$\cong \arg\max_{W} [\max_{P \in \Psi(W)} \{\max_{S \in \Phi(P)} \Pr(x_t \mid s_t)\Pr(s_t \mid s_{t-1})\}\Pr(P \mid W)]\Pr(W)$$

出力確率　遷移確率　発音モデル　言語モデル

模式図



生成モデルに基づく音声認識

数式

$$\tilde{W} = \arg\max_{W} \Pr(X \mid W)\Pr(W)$$

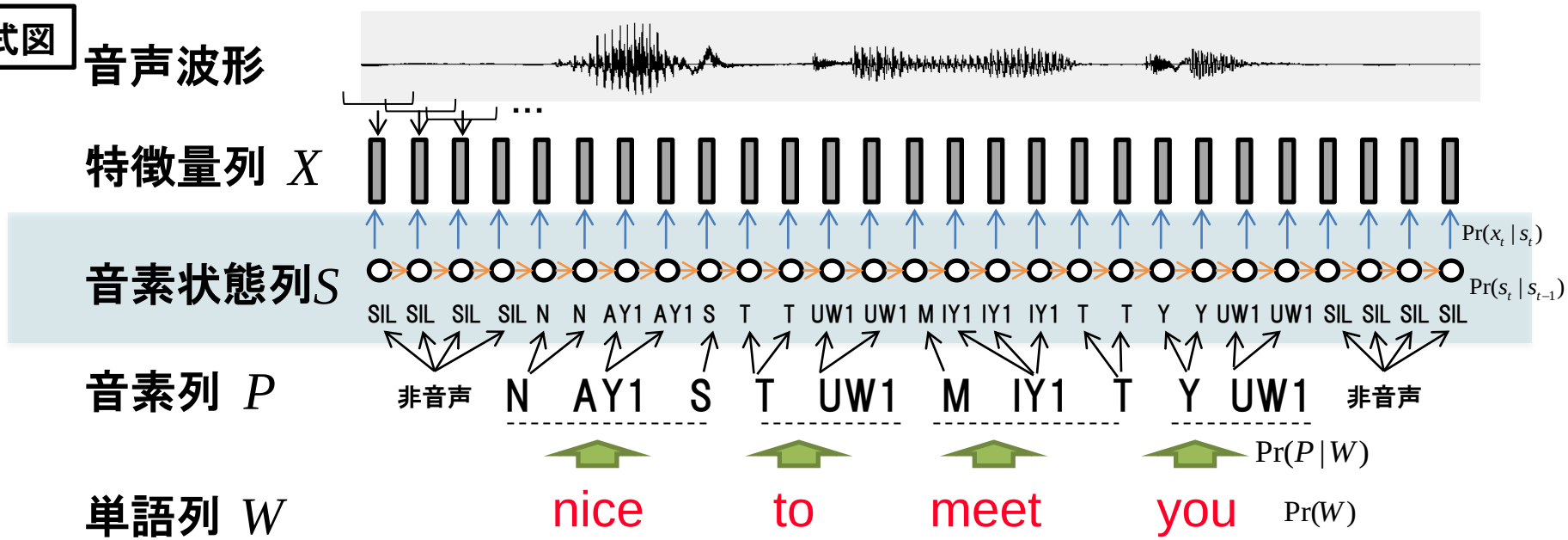
$$\cong \arg\max_{W} \{\max_{P \in \Psi(W)} \Pr(X \mid P)\Pr(P \mid W)\}\Pr(W)$$

$$\cong \arg\max_{W} [\max_{P \in \Psi(W)} \{\max_{S \in \Phi(P)} \Pr(x_t \mid s_t)\Pr(s_t \mid s_{t-1})\}\Pr(P \mid W)]\Pr(W)$$

出力確率　遷移確率　発音モデル　言語モデル

出力確率が音とシンボルを結びつける鍵

模式図

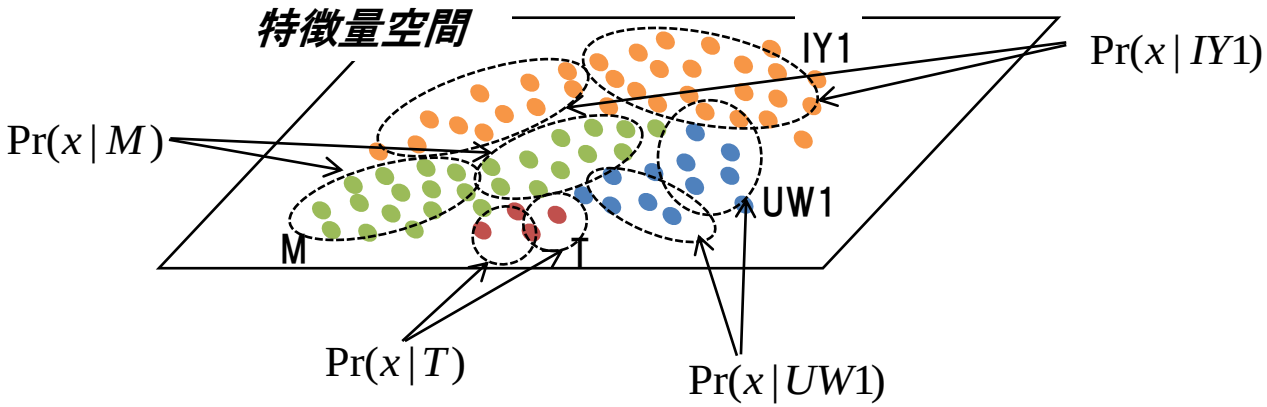


音響モデルの鍵：出力確率Pr(x|s)

HITACHI
Inspire the Next

- “深層学習以前”のモデル化: 混合ガウスモデル(GMM)

音素状態ごとに，特徴量の分布を混合正規分布で表わす



混合重み　状態sのk番目の平均　状態sのk番目の分散

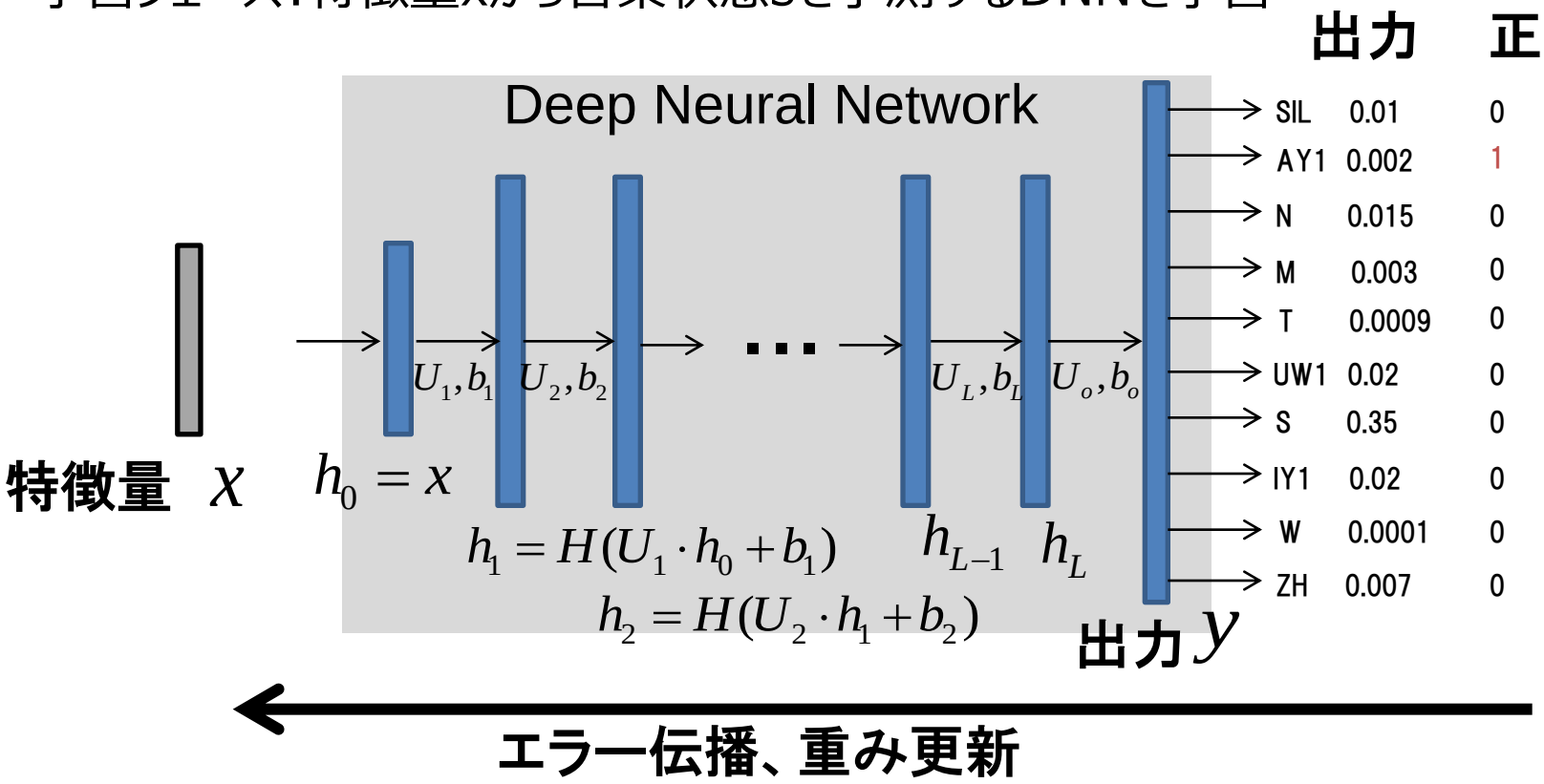
$$\Pr(x \mid s) = \sum_k w_{s,k} N(x; \mu_{s,k}, \Sigma_{s,k})$$

多変量正規分布

深層学習による出力確率Pr(x|s)の計算

HITACHI
Inspire the Next

- 学習フェーズ：特徴量xから音素状態sを予測するDNNを学習

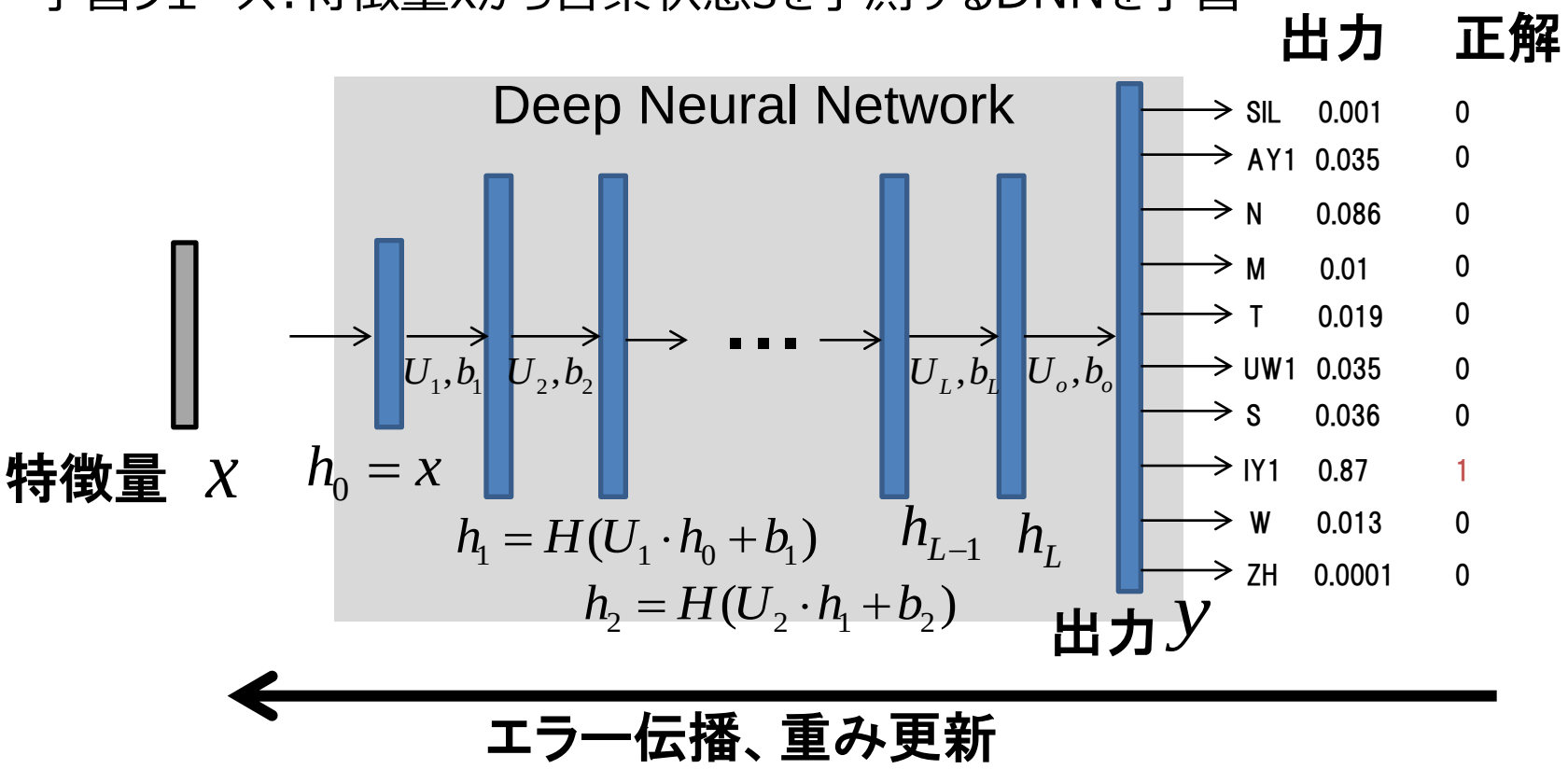


これを

深層学習による出力確率Pr(x|s)の計算

HITACHI
Inspire the Next

- 学習フェーズ:特徴量xから音素状態sを予測するDNNを学習

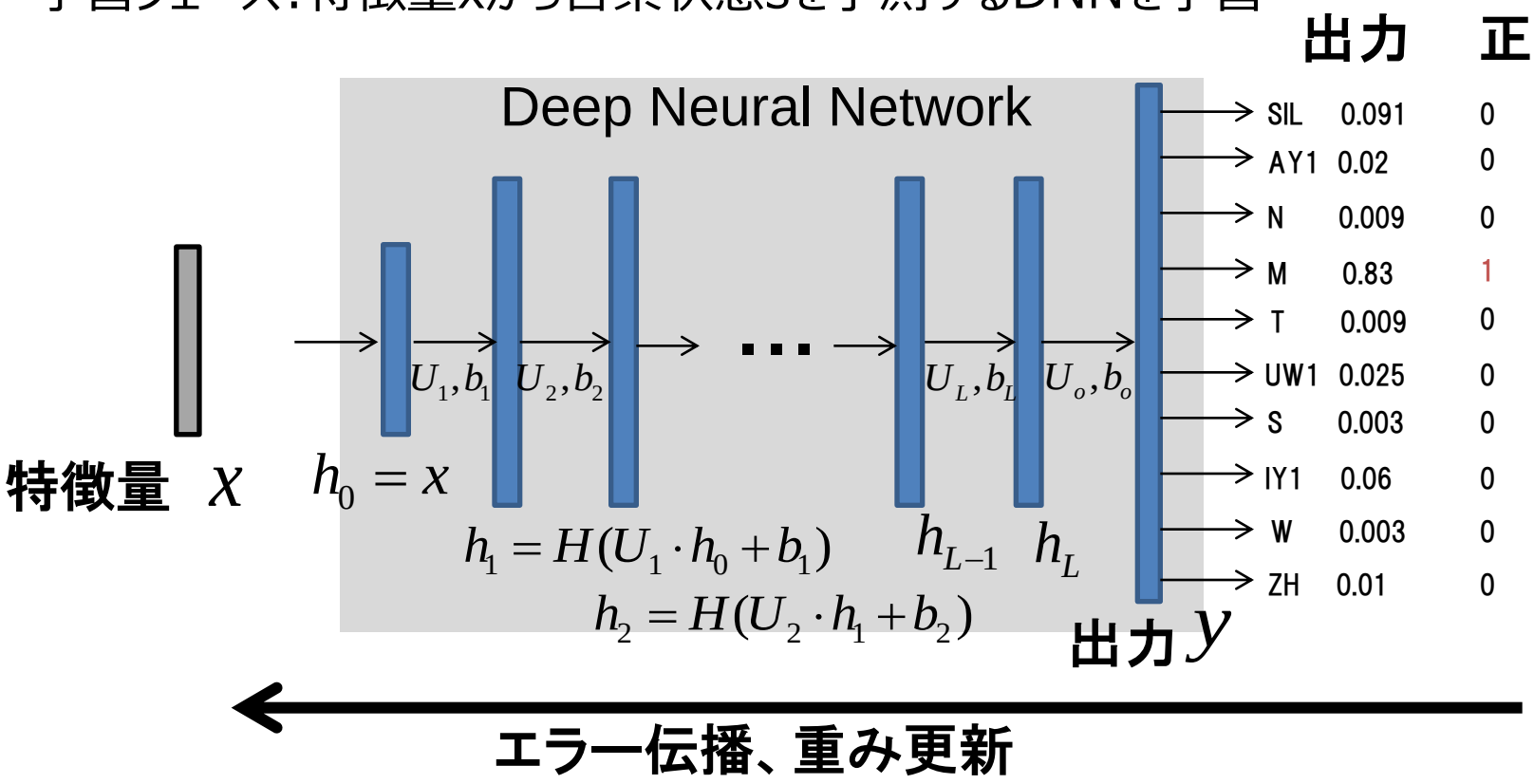


これを　収束するまで

深層学習による出力確率Pr(x|s)の計算

HITACHI
Inspire the Next

- 学習フェーズ：特徴量xから音素状態sを予測するDNNを学習

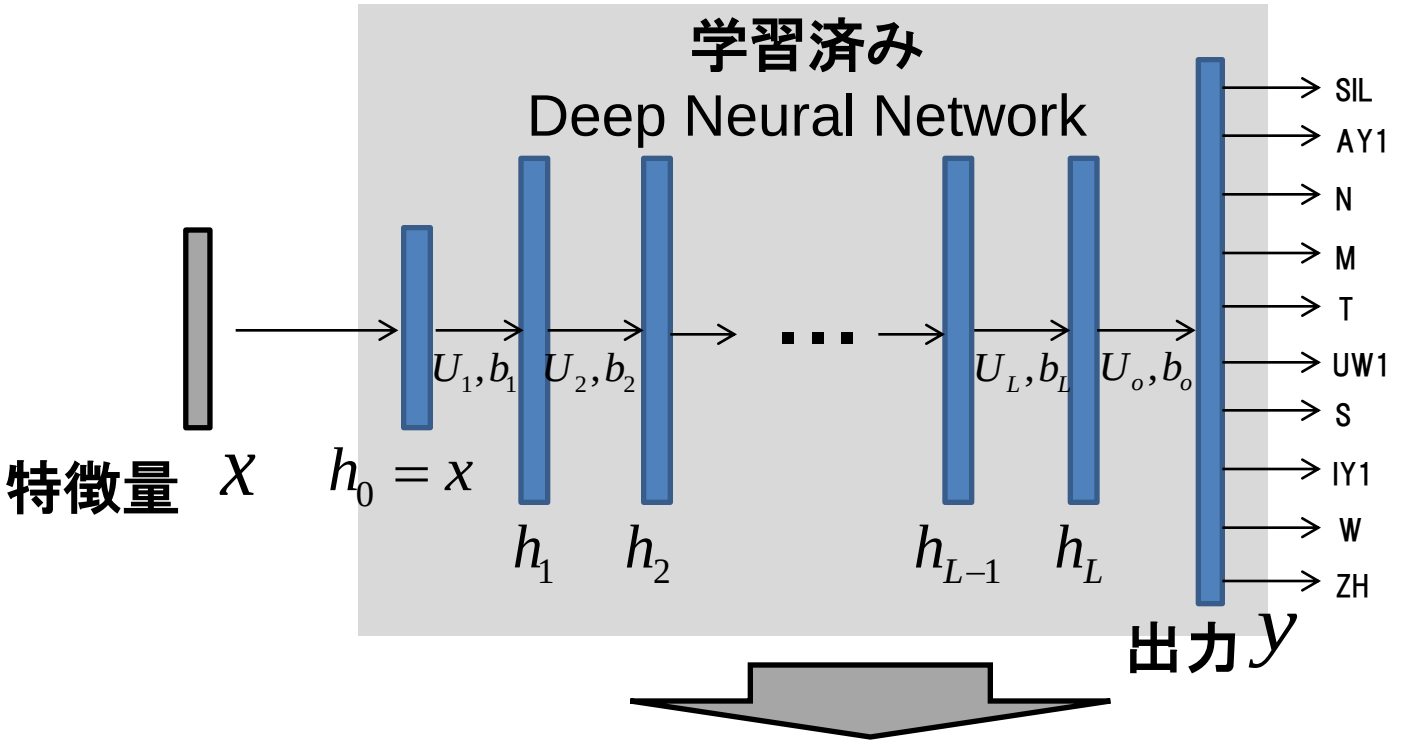


これを　収束するまで　反復

深層学習による出力確率Pr(x|s)の計算

HITACHI
Inspire the Next

- 認識フェーズ：学習したDNNを使い出力確率を計算



値の解釈

$\Pr(SIL \mid x)$
$\Pr(AY1 \mid x)$
$\Pr(N \mid x)$
$\Pr(M \mid x)$
$\Pr(T \mid x)$
$\Pr(UW1 \mid x)$
$\Pr(S \mid x)$
$\Pr(IY1 \mid x)$
$\Pr(W \mid x)$
$\Pr(ZH \mid x)$

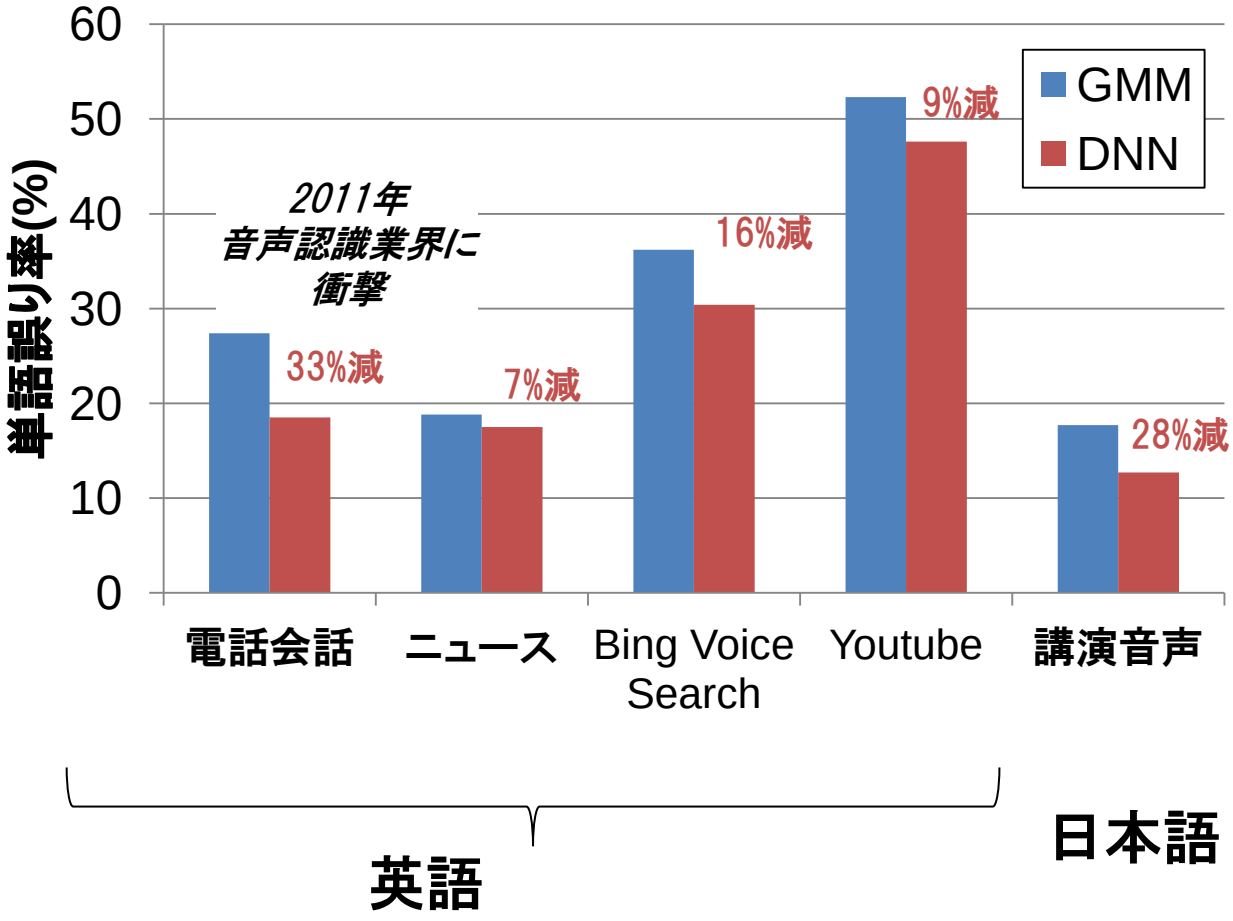
出力確率

$$\Pr(x \mid s) = \frac{\Pr(s \mid x)\Pr(x)}{\Pr(s)} \propto \frac{\Pr(s \mid x)}{\Pr(s)}$$

ニューラルネットの出力値

状態の事前確率（別途学習で求める）

音声認識性能の改善（DNN登場時）



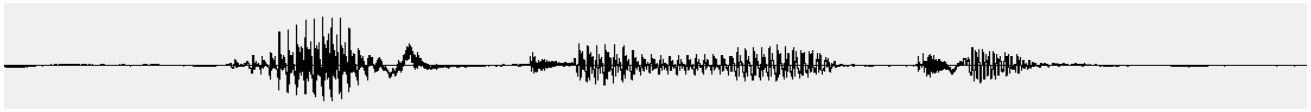
■F. Seide et al., "Conversational speech transcrip-tion using context-dependent deep neural networks,"Proc. Interspeech, pp. 437-440 (2011).
■G. Hinton, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." Signal Processing Magazine, IEEE 29.6 (2012): 82-97.
■N. Kanda, et al. "Elastic spectral distortion for low resource speech recognition with deep neural networks." *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013.

特徴量抽出方法の変遷１

波形から直接学習!![Hoshen,2015][Sainath,2015] ➡ 多ch波形から雑音抑圧も学習!!![Sainath,2016]

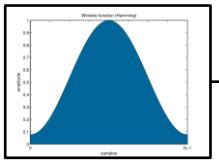
音声波形

16,000sample/秒



25msec = 400sample

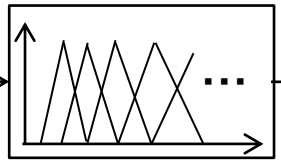
フィルタバンクも学習![Sainath,2013]



窓関数

高速フーリエ変換

Power(・)



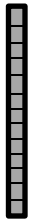
メルフィルタバンク

log(・)

離散コサイン変換

40次元 MFCC

高次成分カット



13次元 MFCC

離散コサイン変換と高次成分カットはしないほうがよい！[Mohamed,2012]

GMMで有効だった特徴量抽出が次々と不要に・・

特徴量抽出方法の変遷 1

HITACHI
Inspire the Next

- ただし“音声波形“からの認識の効果はまだ限定的
  - ネットワーク構造の高度なチューニングが必要
  - （現在のところ）従来特徴量と同等の認識性能

Hrs	WER-raw	WER-log-mel
666	18.8	18.4
1,333	17.1	17.3
2,000	16.2	16.2
40,000	15.5	15.4

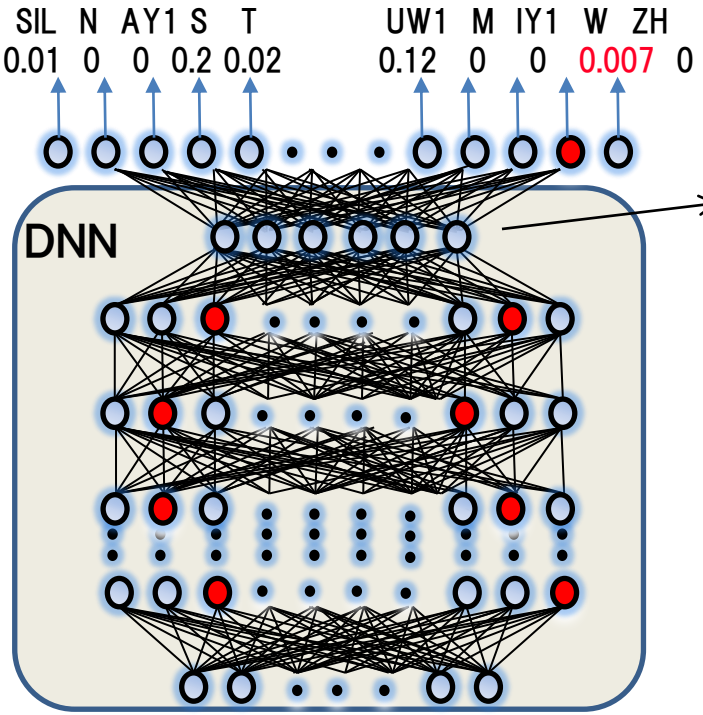
Table 5: WER for Different Amounts of Data

T. Sainath et al., ”Learning the speech frontend with raw waveform CLDNNs,” Proc. Interspeech (2015).より引用

- 従来の特徴量も相変わらずよく利用されている。
  - 高次カットをしないMFCC特徴量なども。

特徴量抽出方法の変遷２ （もうひとつの深層学習適用法）

HITACHI
Inspire the Next



音声特徴量 X_t

ノード数を絞った層
（ボトルネック層）を挿入し，
この値を特徴量として利用

出力確率は
従来どおりGMMで表現

ボトルネック特徴量 → GMM音響モデル

Tandem(縦列)型
と呼ばれる
[Hermansky,2000]

GMM用に培われた
並列学習、話者適応などの技術がそのまま利用できる

DNNで出力確率を計算する方法とは異なる挙動をするので
2つの認識結果を組み合わせると相補的に高精度化ができる

HITACHI
Inspire the Next

識別モデルによる音声認識

２つの基礎数式

HITACHI
Inspire the Next

- 基礎数式１（識別モデル）

単語列

音声特徴量列

音声認識結果
単語列

$$\tilde{W} = \arg\max_{W} \Pr(W \mid X)$$

ベイズ則

Deep Learningによる
モデルが登場！

- 基礎数式２（生成モデル）

$$\tilde{W} = \arg\max_{W} \frac{\Pr(X \mid W)\Pr(W)}{\Pr(X)}$$

Wの最適化に無関係

$$= \arg\max_{W} \Pr(X \mid W)\Pr(W)$$

識別モデルによる音声認識

- 系列長の違う音響特徴量列Xと単語列Wの関係を，ニューラルネットワークでEnd-to-Endに学習

$$\tilde{W} = \arg\max_{W} \Pr(W \mid X)$$

単語列　　音響特徴量列

- End-to-Endモデルへの期待
  - 学習がシンプル
  - 良い認識性能（一部で報告されている）
  - 高速な実行
    - 大きな時間シフトで入力を与えても、勝手に最適なモデルを学習してくれる。
    - 従来10msec単位で動作していたものが30msec単位で動作 ⇒ 3倍高速化!

識別モデルのためのEnd-to-Endニューラルネットワーク

HITACHI
Inspire the Next

- 実際には多くの場合、単語列ではなくサブワード列（文字、カナ、音素等）を利用

$$\tilde{L} = \arg\max_{L} \Pr(L \mid X)$$

サブワード列　　音響特徴量列

- これは、以下の理由による
  - 単語でモデルを作ってしまうと，後から新規語彙を追加するのが難しい
  - 数十万単語レベルのモデルを作ろうとすると、巨大で非効率的
  - 高精度な言語モデルを作るには、通常の音声コーパスの書き起こしは小さすぎる

- 以下の例ではサブワードをベースに紹介します。

End-to-Endモデル：主に２つのモデルが検討されている

HITACHI
Inspire the Next

- Connectionist Temporal Classification (CTC)
- Attention Encoder Decoder

Connectionist Temporal Classification (CTC)[Graves,2006]

HITACHI
Inspire the Next

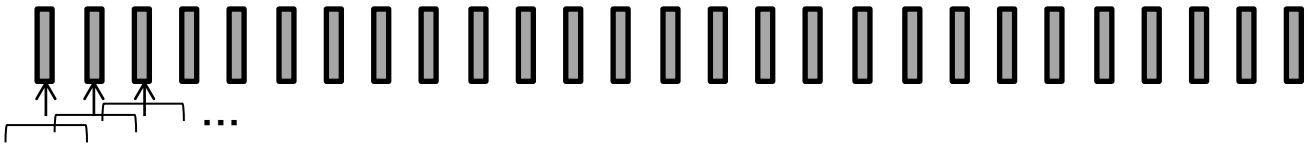
数式

$$\Pr(L \mid X) = \ ?$$

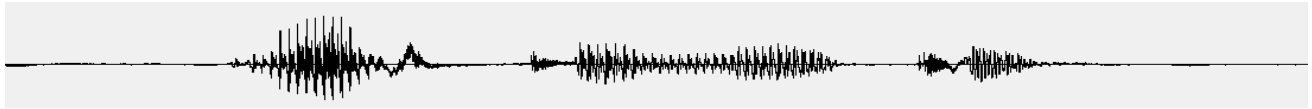
模式図

文字列 L n i c e _ t o _ m e e t _ y o u

特徴量列 X



音声波形



Connectionist Temporal Classification (CTC)[Graves,2006]

HITACHI
Inspire the Next

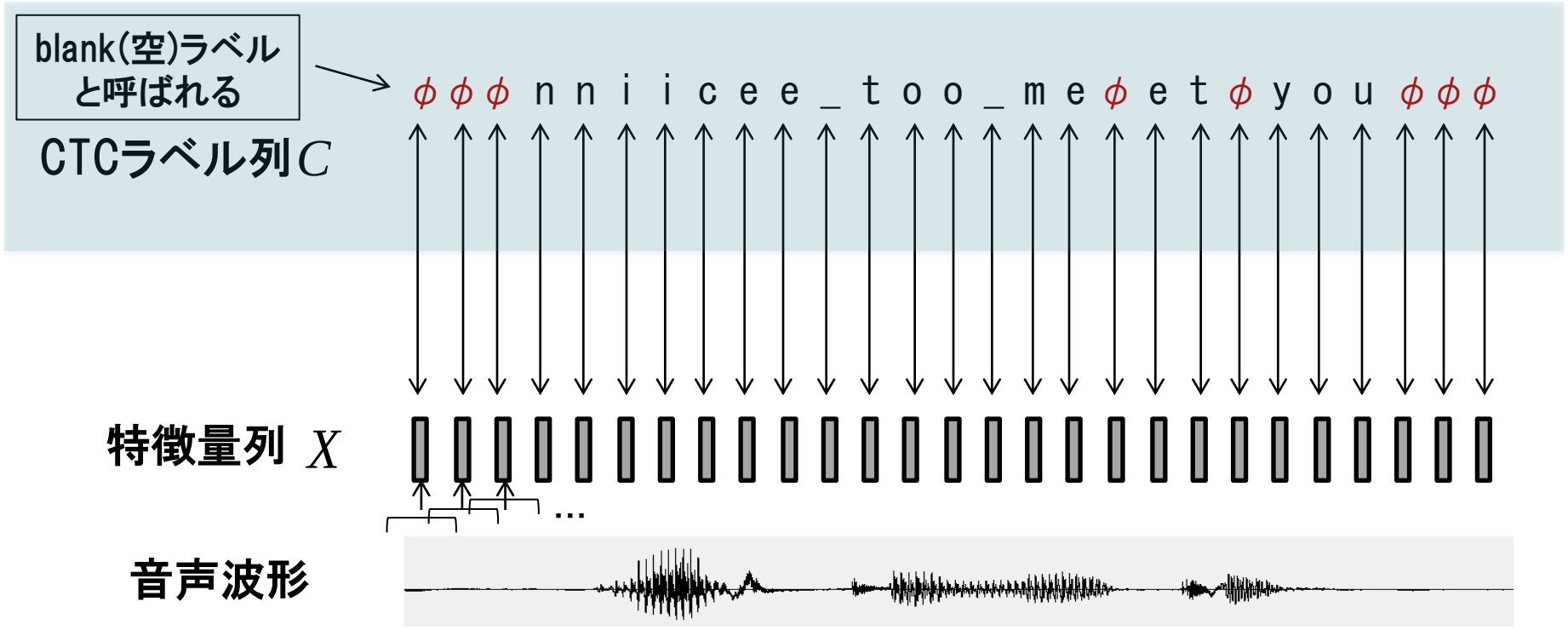
数式

$$\Pr(L \mid X) = \ ?$$

模式図

文字列 L n i c e _ t o _ m e e t _ y o u

blank(空)ラベル と呼ばれる



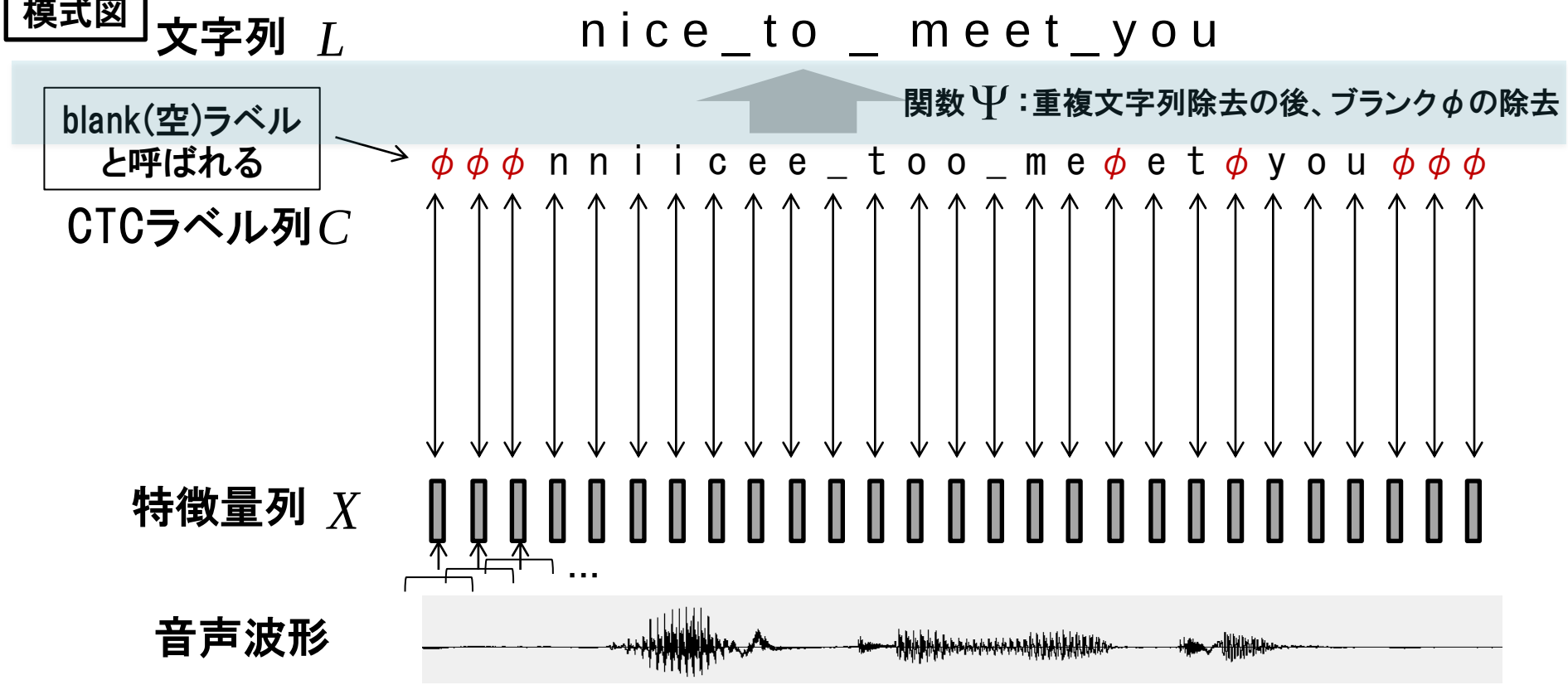
Connectionist Temporal Classification (CTC)[Graves,2006]

HITACHI
Inspire the Next

数式

$$\Pr(L \mid X) = \ ?$$

模式図



Connectionist Temporal Classification (CTC)[Graves,2006]

HITACHI
Inspire the Next

数式

$$\Pr(L \mid X) = \quad ?$$

模式図

文字列 L : n i c e _ t o _ m e e t _ y o u

関数 Ψ **：重複文字列除去の後、ブランク** ϕ **の除去**

blank(空)ラベル と呼ばれる

CTCラベル列 C :

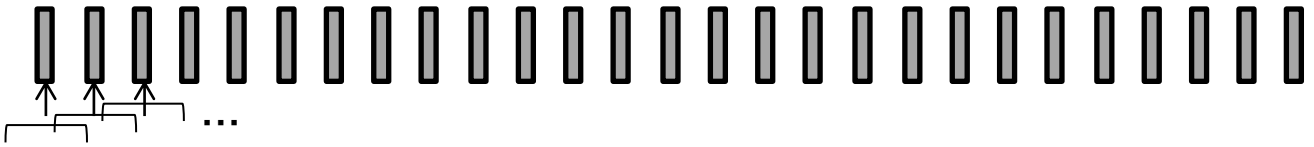
$\phi\ \phi\ \phi$ n n i i c e e _ t o o _ m e ϕ e t ϕ y o u $\phi\ \phi\ \phi$

$\phi\ \phi\ \phi\ \phi$ n ϕ i c ϕ e _ t o _ m e ϕ e t ϕ y o u $\phi\ \phi\ \phi\ \phi$

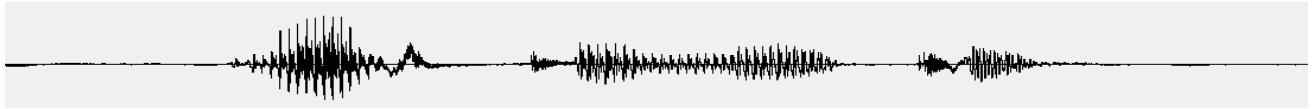
⋮

$\phi\ \phi\ \phi\ \phi$ n ϕ i c ϕ e _ t o _ m e ϕ e t ϕ y o u $\phi\ \phi\ \phi\ \phi$

特徴量列 X



音声波形



Connectionist Temporal Classification (CTC)[Graves,2006]

HITACHI
Inspire the Next

数式

$$\Pr(L \mid X) = \sum_{C \in \Psi^{-1}(L)} \Pr(C \mid X)$$

模式図

文字列 L : n i c e _ t o _ m e e t _ y o u

関数Ψ：重複文字列除去の後、ブランクφの除去

blank(空)ラベルと呼ばれる

CTCラベル列 C :

φ φ φ n n i i c e e _ t o o _ m e φ e t φ y o u φ φ φ

φ φ φ φ n φ i c φ e _ t o _ m e φ e t φ y o u φ φ φ φ

⋮

φ φ φ φ n φ i c φ e _ t o _ m e φ e t φ y o u φ φ φ φ

なんらかのモデルPr(C|X)

特徴量列 X

音声波形

Connectionist Temporal Classification (CTC)[Graves,2006]

HITACHI
Inspire the Next

数式

$$\Pr(L \mid X) = \sum_{C \in \Psi^{-1}(L)} \Pr(C \mid X) := \sum_{C \in \Psi^{-1}(L)} \prod_{t} y_{t,c_t}$$

NNの出力の積

模式図

文字列 L : n i c e _ t o _ m e e t _ y o u

関数Ψ：重複文字列除去の後、ブランクϕの除去

blank(空)ラベルと呼ばれる

CTCラベル列 C :

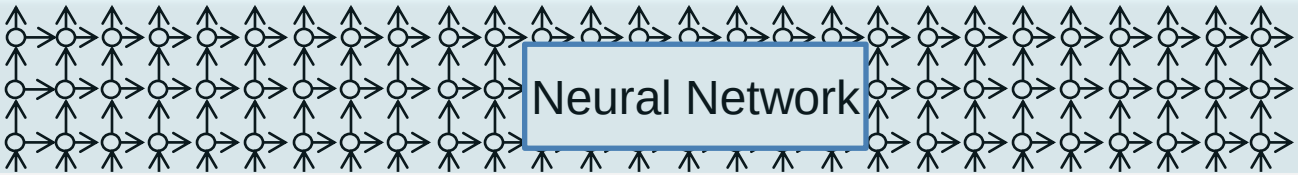
ϕ ϕ ϕ n n i i c e e _ t o o _ m e ϕ e t ϕ y o u ϕ ϕ ϕ

ϕ ϕ ϕ ϕ n ϕ i c ϕ e _ t o _ m e ϕ e t ϕ y o u ϕ ϕ ϕ ϕ

⋮

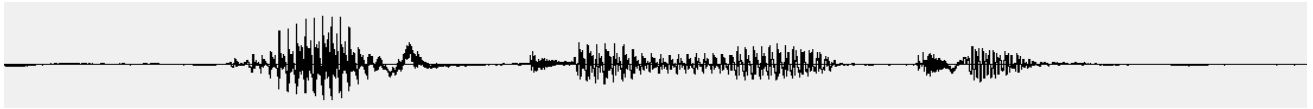
ϕ ϕ ϕ ϕ n ϕ i c ϕ e _ t o _ m e ϕ e t ϕ y o u ϕ ϕ ϕ ϕ

NNの出力 Y



特徴量列 X

音声波形



CTCの学習

- CTCの学習基準

$$F^{CTC} = \sum_u \log \Pr(L_u \mid X_u)$$

$$= \sum_u \log \sum_{C_u} \Pr(C_u \mid X_u) = \sum_u \log \sum_{C_u} \prod_t y_{t,c_{u,t}}$$

学習データの番号

最大化

ニューラルネットワークから得られる値が
$\Pr(L \mid X)$ を表すようになると期待される

CTCの学習

- CTCの学習基準

$$F^{CTC} = \sum_u \log \Pr(L_u \mid X_u)$$

$$= \sum_u \log \sum_{C_u} \Pr(C_u \mid X_u) = \sum_u \log \sum_{C_u} \prod_t y_{t,c_{u,t}}$$

学習データの番号

最大化

ニューラルネットワークから得られる値が $\Pr(L \mid X)$ を表すようになると期待される

- ニューラルネットワークの学習
  - エラーの計算（yで微分可能）
    - 出力層がSoftmaxの場合
  - 確率的勾配降下法などで最適化

$$e_u^{CTC}(c,t) = \frac{\partial \mathcal{F}^{CTC}}{\partial a_t^u(c)} = \sum_{c'} \frac{\partial \mathcal{F}^{CTC}}{\partial y_t^u(c')} \frac{\partial y_t^u(c')}{\partial a_t^u(c)}$$

$$= \frac{\sum_{\mathbf{c} \in \Phi^{-1}(\mathbf{s}_u)} \delta_{c_t,c} P(\mathbf{c}|\mathbf{X}_u)}{P(\mathbf{s}_u|\mathbf{X}_u)} - y_t^u(c)$$

yを計算するためのSoftmax関数に入力される，ラベルcのactivation

時刻tにラベルcを通る事後確率

動的計画法で計算可能

時刻tのラベルcに対するネットワークの出力

ＣＴＣの注意点

HITACHI
Inspire the Next

- いくつかの仮定
  - 入力長>出力長
  - 入力と出力の対応関係に、時間順序の逆転がないことが暗に仮定

- 系列の確率を、ニューラルネットワーク出力の積で表現しており、精度に限界がある可能性がある（independence assumption）

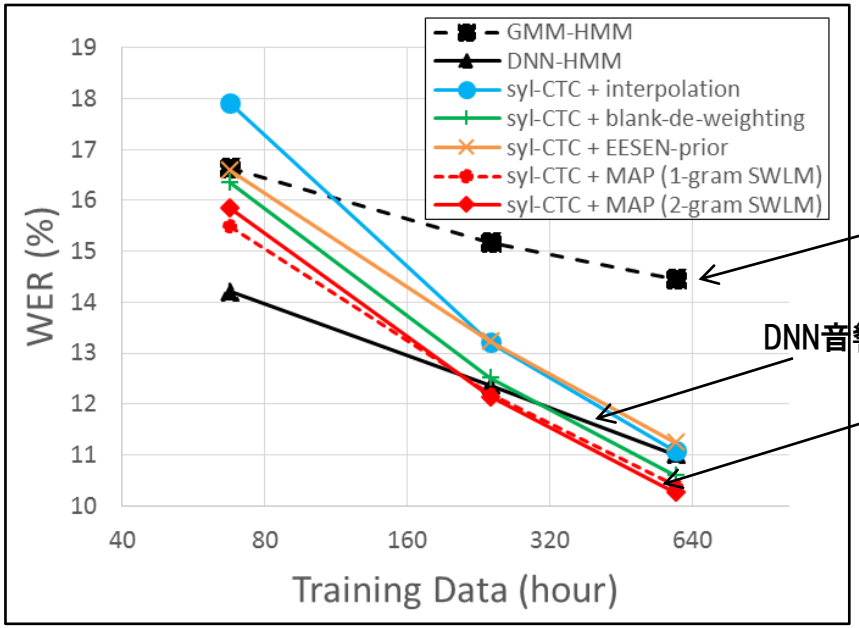
$$\Pr(L \mid X) = \sum_{C \in \Psi^{-1}(L)} \Pr(C \mid X) := \sum_{C \in \Psi^{-1}(L)} \prod_{t} y_{t,c_t}$$

NNの出力の積

ＣＴＣの注意点

- 現状、ほとんどの場合、別途言語モデルを学習して組み合わせないと良い精度が出ない
  - また，学習データと異なる言語ドメインで利用したいことも多い
- （少なくとも音声認識では）小規模なデータでは良い性能が出ない
  - ２点以上のデータサイズでの性能比較が重要（CTCの場合は特に顕著）

日本語話し言葉コーパス



GMM音響モデル

DNN音響モデル

カナCTC+単語言語モデル

[Kanda,2017]より引用

Naoyuki Kanda, Xugang Lu, Hisashi Kawai, Maximum A Posteriori based Decoding for End-to-End Acoustic Models, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017

End-to-Endモデル：主に２つのモデルが検討されている

HITACHI
Inspire the Next

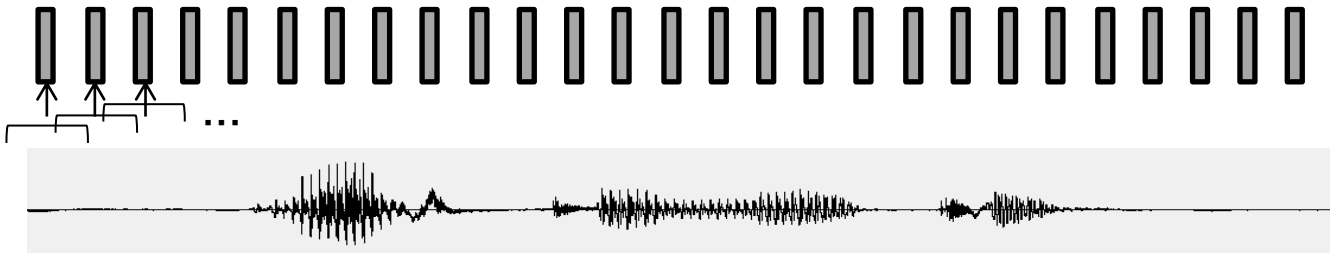
- Connectionist Temporal Classification (CTC)
- Attention Encoder Decoder

Attention Encoder Decoder [Chorowski, 2014][Chorowski, 2015]他

HITACHI
Inspire the Next

- 機械翻訳分野で生まれたSequence-to-Sequenceモデルを音声認識に応用

L n i c e _ t o _ m e e t _ y o u



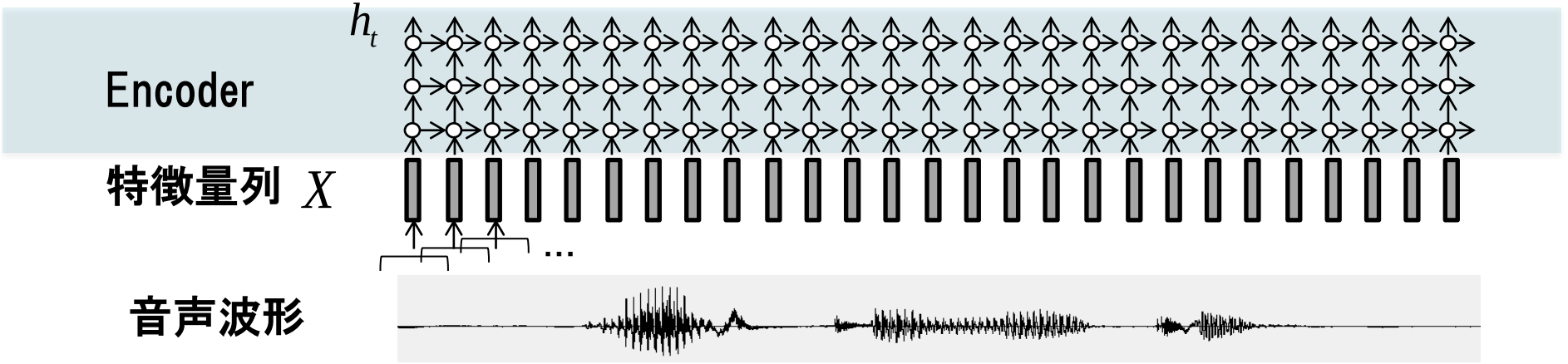
J. Chorowski, et al. "End-to-end continuous speech recognition using attention-based recurrent NN: First results." *arXiv preprint arXiv:1412.1602* (2014).

Attention Encoder Decoder [Chorowski, 2014][Chorowski, 2015]他

HITACHI
Inspire the Next

- 機械翻訳分野で生まれたSequence-to-Sequenceモデルを音声認識に応用

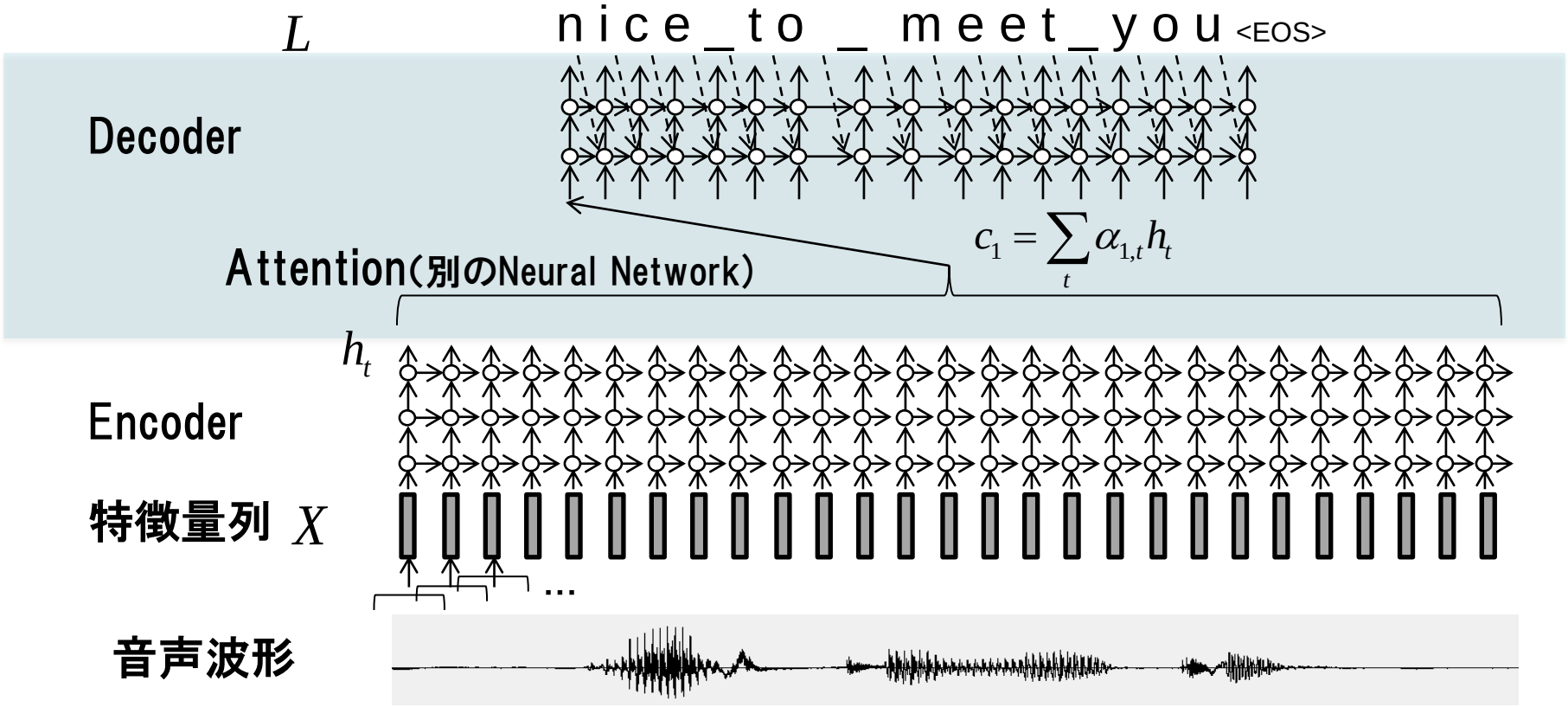
L n i c e _ t o _ m e e t _ y o u



J. Chorowski, et al. "End-to-end continuous speech recognition using attention-based recurrent NN: First results." *arXiv preprint arXiv:1412.1602* (2014).

HITACHI
Inspire the Next

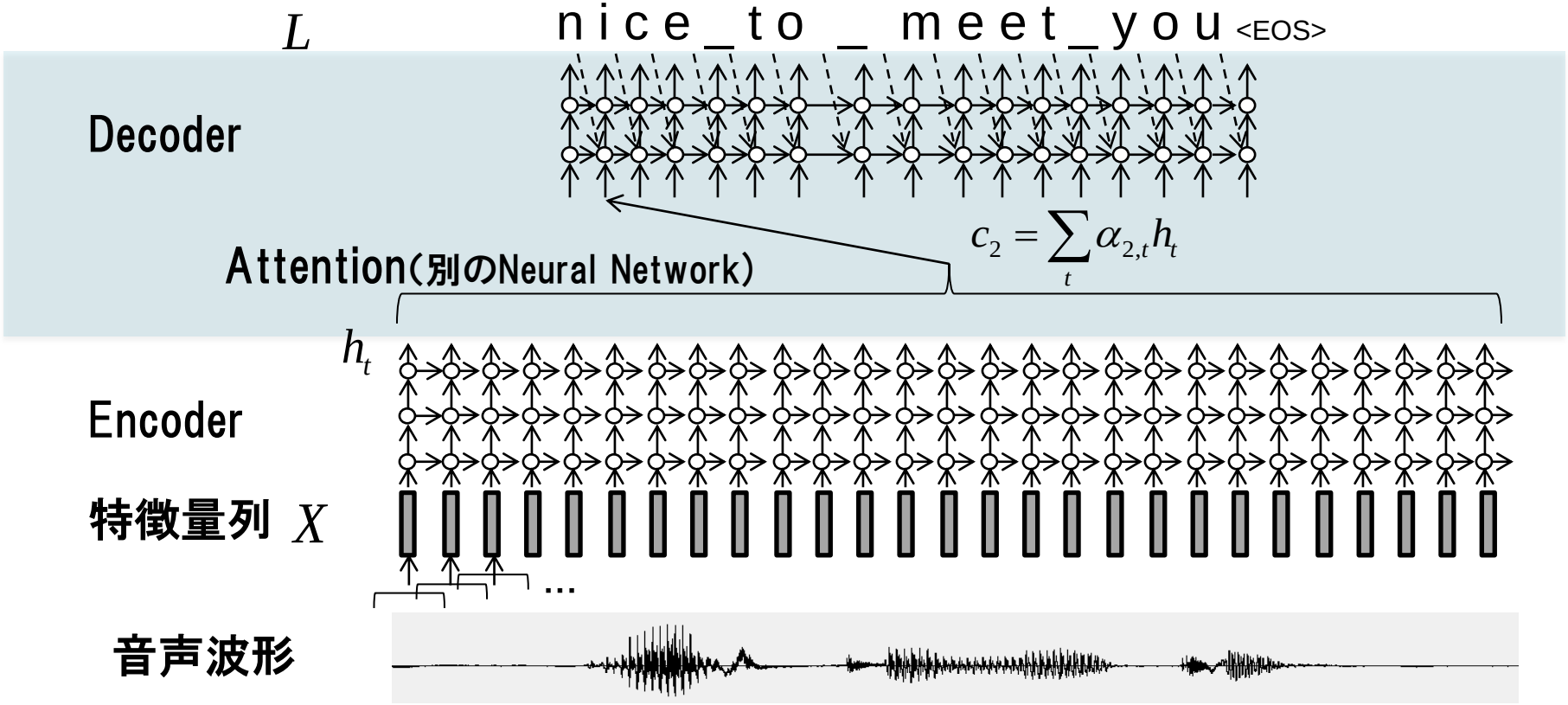
- 機械翻訳分野で生まれたSequence-to-Sequenceモデルを音声認識に応用



J. Chorowski, et al. "End-to-end continuous speech recognition using attention-based recurrent NN: First results." *arXiv preprint arXiv:1412.1602* (2014).

HITACHI
Inspire the Next

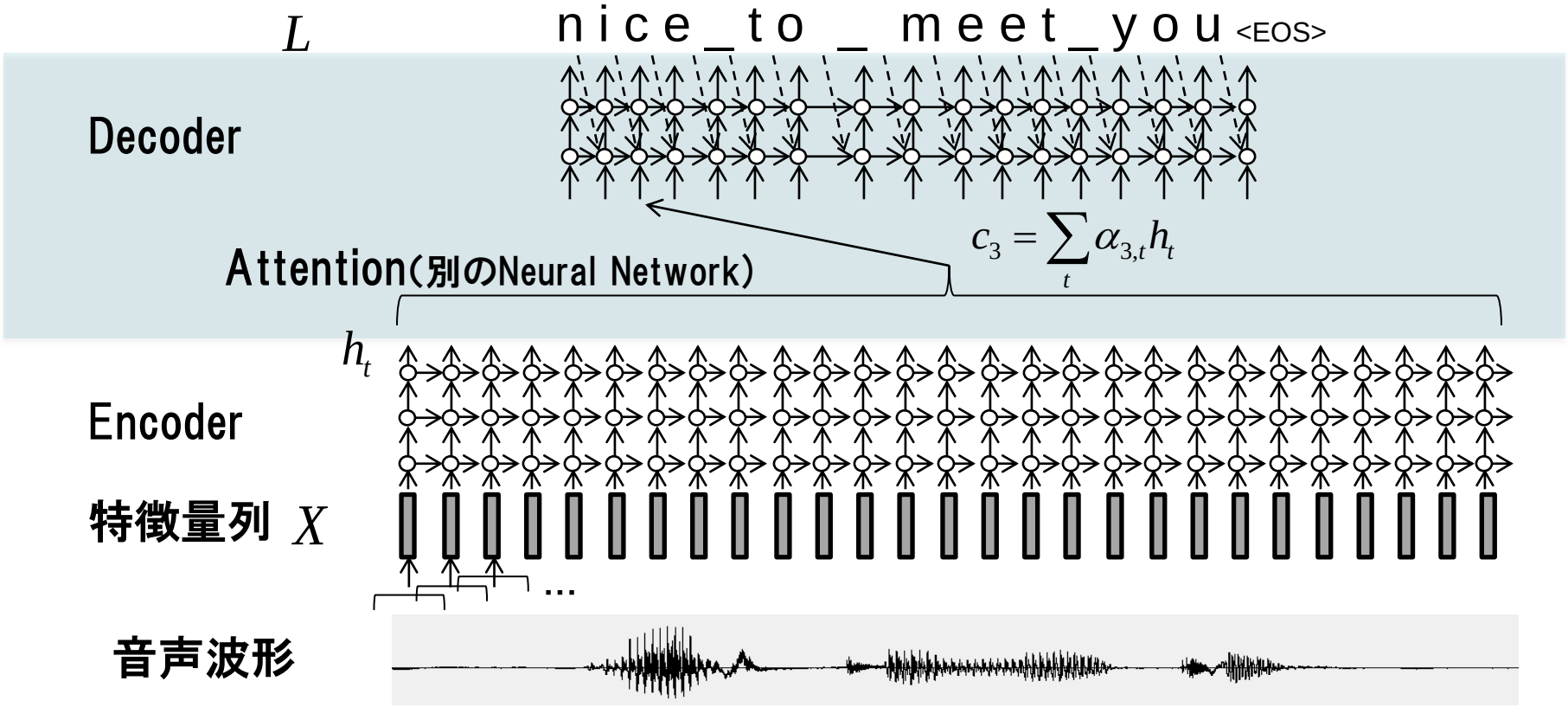
- 機械翻訳分野で生まれたSequence-to-Sequenceモデルを音声認識に応用



J. Chorowski, et al. "End-to-end continuous speech recognition using attention-based recurrent NN: First results." *arXiv preprint arXiv:1412.1602* (2014).

Attention Encoder Decoder [Chorowski, 2014][Chorowski, 2015]他

- 機械翻訳分野で生まれたSequence-to-Sequenceモデルを音声認識に応用



J. Chorowski, et al. "End-to-end continuous speech recognition using attention-based recurrent NN: First results." *arXiv preprint arXiv:1412.1602* (2014).

Attention Enc-Decの良い点と注意点

HITACHI
Inspire the Next

- 良い点
  - CTCで存在した様々な仮定(independence assumptionや入力長>出力長等）がない。

- 注意点
  - Attentionの自由度が高すぎるため，学習が難しい
    - Windowing: Attentionをかける範囲にhand-tunedな制約[Chorowski, 2015][Bahdanau, 2015]

  - 生成モデルに基づく最新手法と比較して良かった，という報告はまだない
    - とはいえ，ここ１～２年で大幅に性能改善しており，一気に従来法を抜く可能性も否定はできない。

ここまでのまとめ

HITACHI
Inspire the Next

- 音声認識の基本数式として生成モデルと識別モデルの2系統が存在
- 生成モデル
  - 出力確率をDNNで計算
  - Deep Learningの火付け役であり，現在でも主流の方式
- 識別モデル
  - End-to-Endモデル
  - CTC, Attention Encoder-Decoderなど
  - 近年，盛んに研究が進んでいる

HITACHI
Inspire the Next

Part2: 音声認識におけるDeep Learningの動向

注目トピックを順不同でご紹介します

HITACHI
Inspire the Next

さまざまなネットワーク

ネットワーク構造の変種

HITACHI
Inspire the Next

- 大量に提案・評価されており、他分野からの導入も多い
  - Simple Recurrent Neural Network
  - Time Delay Neural Network
  - Long short term memory (LSTM)
  - Bidirectional LSTM
  - Convolutional Neural Network
  - VGGNet
  - Network in Network
  - Highway Network
  - Residual Network
  - Attention Encoder Decoder
  - ……

時系列の依存関係を表現

画像分野からの導入

言語処理分野からの導入

得られる効果はほぼ分野非依存

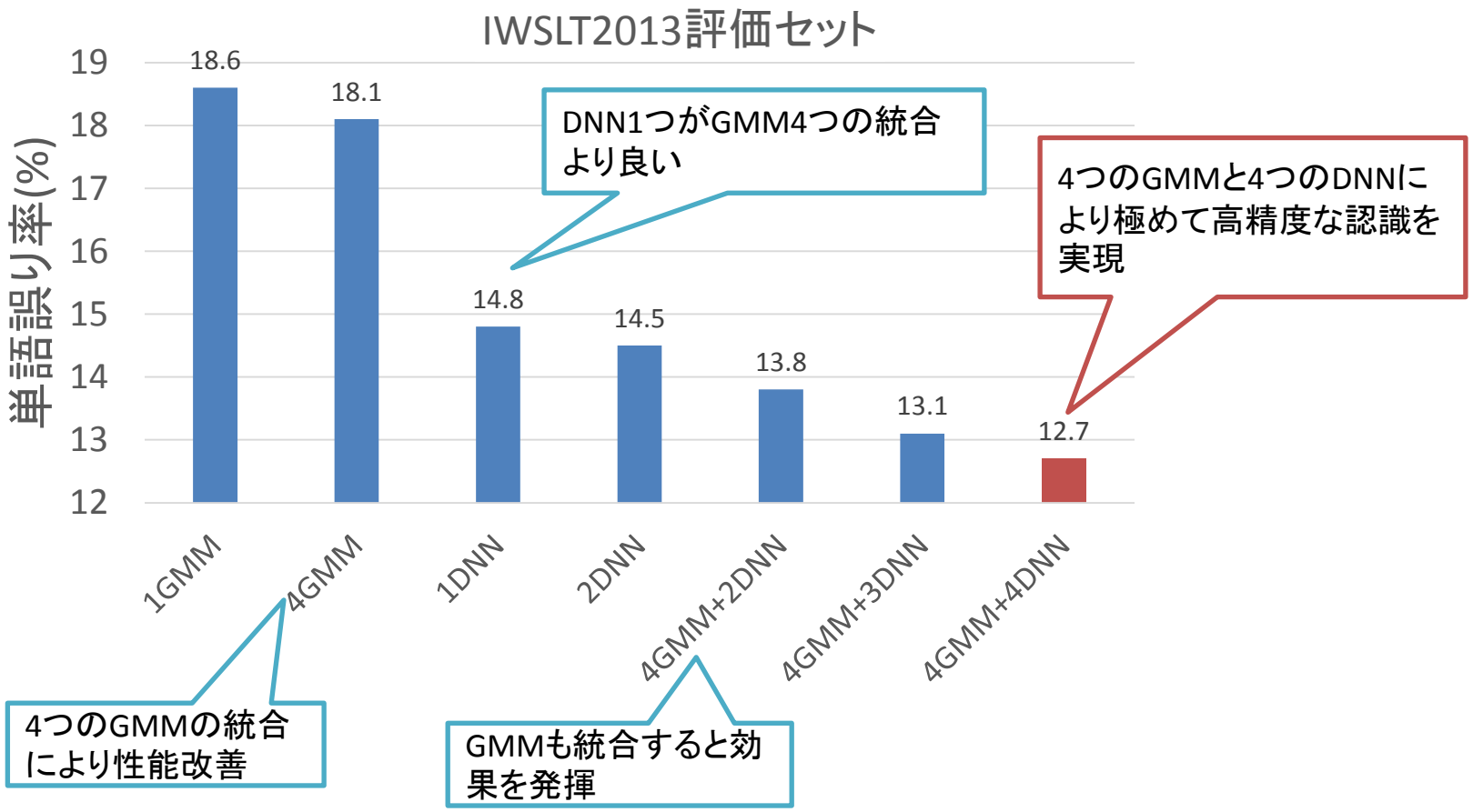
適切に組み合わせるのがBest Practice

多数の音響モデル群の統合

HITACHI
Inspire the Next

1種類の音響モデルではどうしても得意不得意が生じる

⇒ IWSLT2014コンペ優勝システム：4種類のGMMモデルと4種類のDNNモデルの計8モデルを組み合わせることにより極めて高精度な音声認識を実現



P. Shen, X. Lu, N. Kanda, M. Saiko, C. Hori: The NICT ASR system for IWSLT 2014, Proc. IWSLT 2014.

HITACHI
Inspire the Next

ネットワークの自動適応

音響モデルの自動話者適応

HITACHI
Inspire the Next

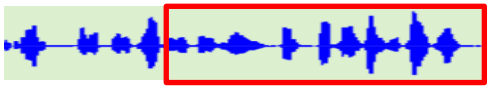
「正解していそうな箇所」に適合するようネットワークを更新

音声認識結果

「と今日オリンピック開催」

①正解していそうな箇所の算出

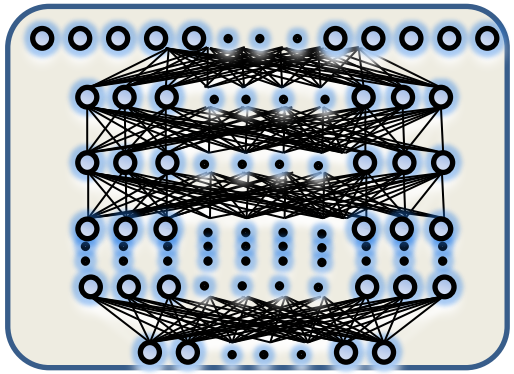
正解していそうな箇所



モデル統合後に得られる
信頼度を利用

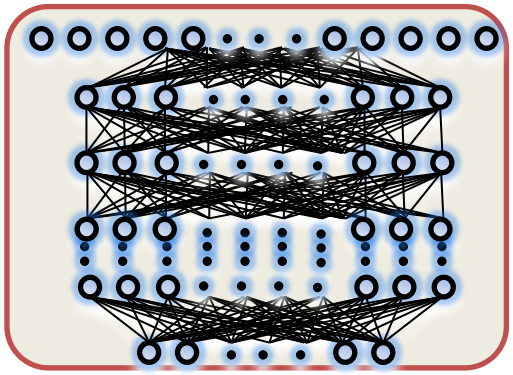
「と今日オリンピック開催」

②ネットワークの更新



音響モデル

更新



適応音響モデル

更新式

$$\Delta \mathbf{w}_t = -\epsilon \nabla_w E(\mathbf{w}_t) + \alpha \Delta \mathbf{w}_{t-1} - \beta(\mathbf{w}_{t-1} - \mathbf{w}_0)$$

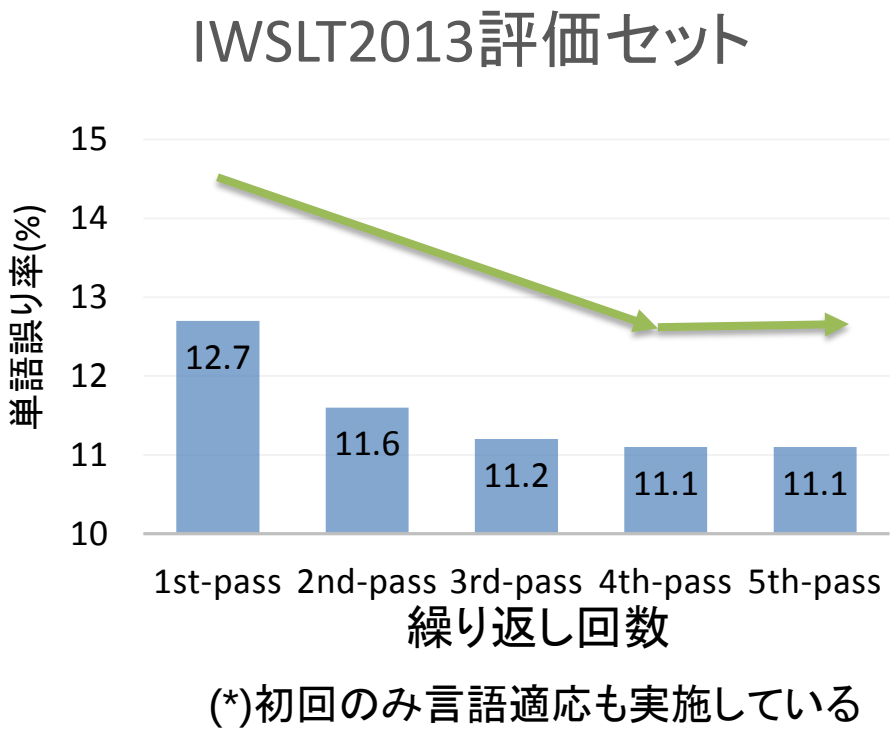
音声に適合する
ための項

更新を促進する
ための項

元の音響モデルから
離れすぎないための項

音響モデルの自動話者適応

「認識⇒自動適応」を繰り返すことで精度向上[Shen, 2014]



P. Shen, X. Lu, N. Kanda, M. Saiko, C. Hori: The NICT ASR system for IWSLT 2014, Proc. IWSLT 2014.

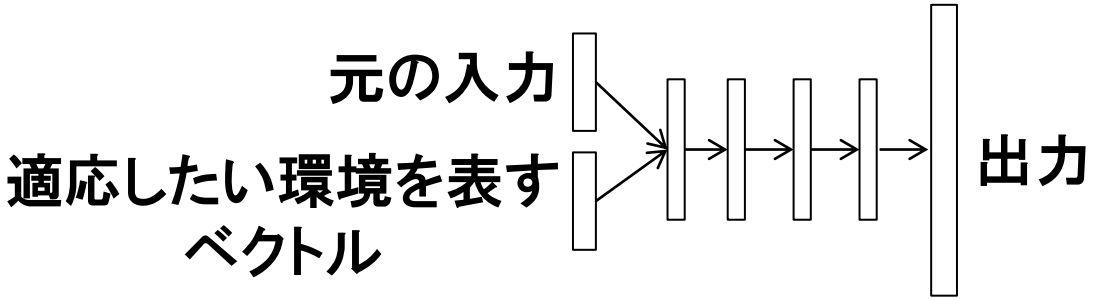
音響モデルの自動適応のトピック

HITACHI
Inspire the Next

- 過学習の抑制
  - KL正則化：出力の分布を、元のモデルの分布に近づける[Yu, 2013]
  - L2正則化：重みを、元の重みに近づける[Liao, 2013]

- 少ない発話で最大の性能アップを目指す
  - ノードの出力にゲートを設け、そのゲートだけ調整[Swietojanski, 2014]
  - 予め複数モデルを用意しておき，その混合重みだけを調整[Delcroix, 2015]

- 簡易に適応を行う
  - 特徴量に“ノイズを特徴量化したものの平均”を追加する[Seltzer, 2013]
  - 特徴量に”話者特徴量を抽出したもの“を追加する[Saon, 2013]



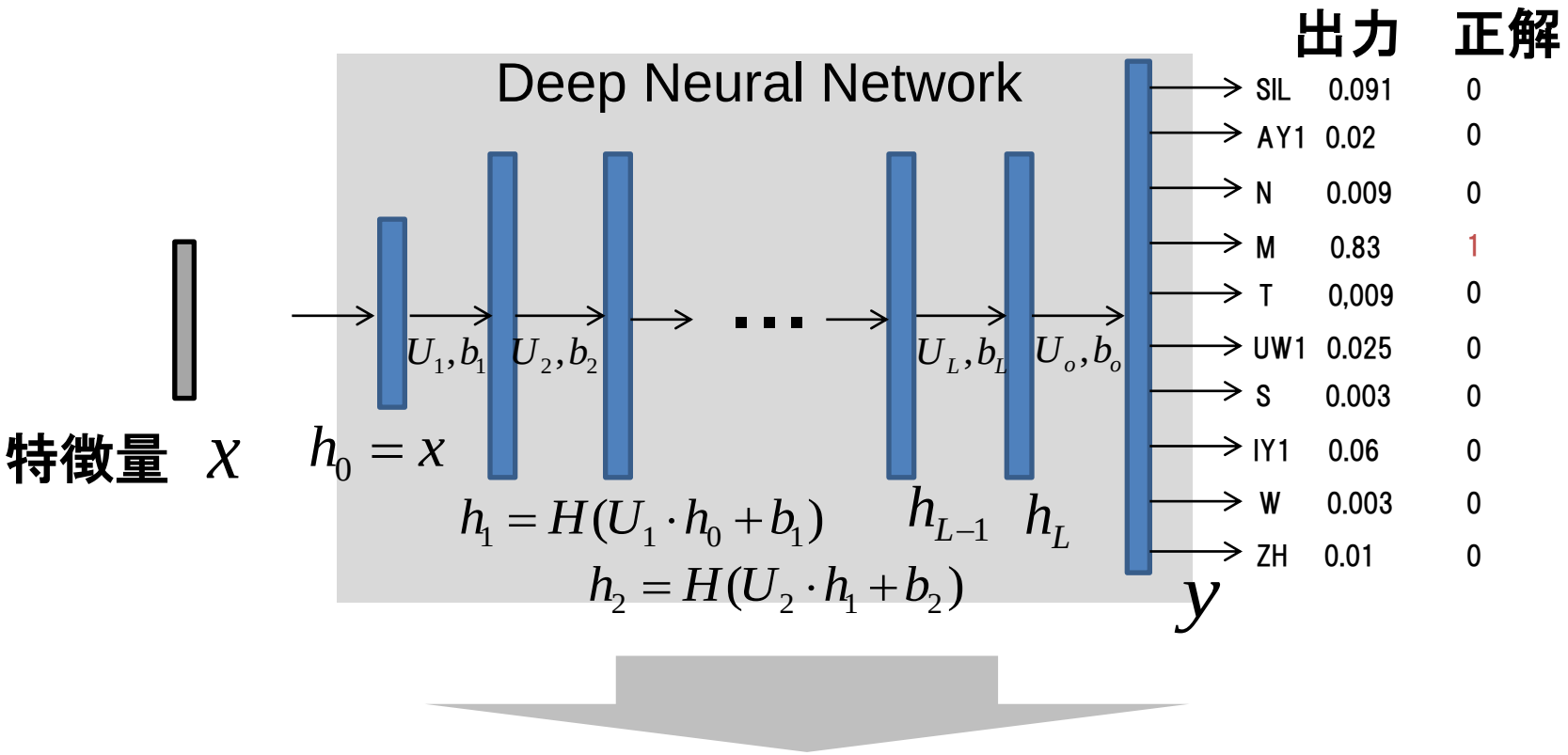
HITACHI
Inspire the Next

さまざまな学習方法

系列識別学習[Su, 2013][Vesely, 2013]

HITACHI
Inspire the Next

- 生成モデルでは、状態の予測精度が最大化するようにDNNを学習していた



- 期待音声認識率を最大化するようにDNNを学習できないか？→系列識別学習

H. Su, et al. Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013. p. 6664-6668.

K. Vesely, et al. Sequence-discriminative training of deep neural networks. In *Interspeech* (pp. 2345-2349), 2013.

系列識別学習[Su, 2013][Vesely, 2013]

HITACHI
Inspire the Next

- 期待音声認識率の最大化（state-level Minimum Bayes Risk学習）

$$F^{sMBR} = \sum_u \sum_W P(W \mid X_u) Acc(W, R_u)$$

現在のパラメータ下で
学習用音声Xから仮説Wが生成される確率

正解ラベルRと比較した
仮説Wの精度

- 確率的勾配降下法で学習可能
- 音声認識における必須技術のひとつ。
  - その他に、相互情報量最大化(MMI)基準やboosted MMI基準などが知られている。

Table 3: Results (% WER) of the DNNs trained on the full 300 hour training set using different criteria.

System	Hub5 '00 SWB	Hub5 '00 CHE	Hub5 '00 Total	Hub5 '01 SWB	Hub5 '01 SWB2P3	Hub5 '01 SWB-Cell	Hub5 '01 Total
GMM BMMI	18.6	33.0	25.8	18.9	24.5	30.1	24.6
DNN CE	14.2	25.7	20.0	14.5	19.0	25.3	19.8
DNN MMI	12.9	24.6	18.8	13.3	17.8	23.7	18.4
DNN sMBR	12.6	24.1	18.4	13.0	17.7	22.9	18.0
DNN MPE	12.9	24.1	18.5	13.2	17.7	23.4	18.2
DNN BMMI	12.9	24.5	18.7	13.2	17.8	23.5	18.3

GMM
DNN
sMBR DNN

[Vesely, 2013]より引用

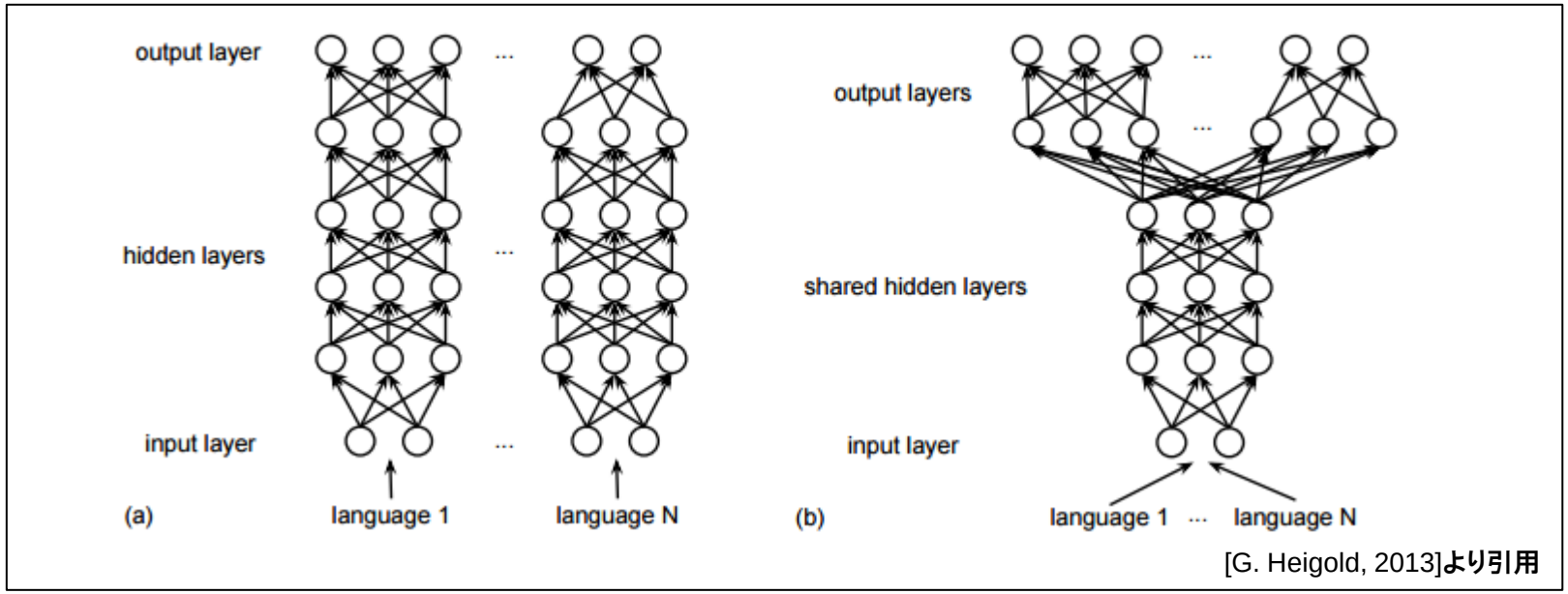
H. Su, et al. Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013. p. 6664-6668.

K. Vesely, et al. Sequence-discriminative training of deep neural networks. In *Interspeech* (pp. 2345-2349), 2013.

マルチリンガルモデル[G. Heigold, 2013][J. Huang, 2013][S. Matsuda, 2013]

HITACHI
Inspire the Next

- 多言語データをshareすることで強力な特徴量抽出能力を獲得



[G. Heigold, 2013]より引用

G. Heigold, et al. Multilingual acoustic models using distributed deep neural networks. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013. p. 8619-8623.

J. Huang, et al. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013. p. 7304-7308.

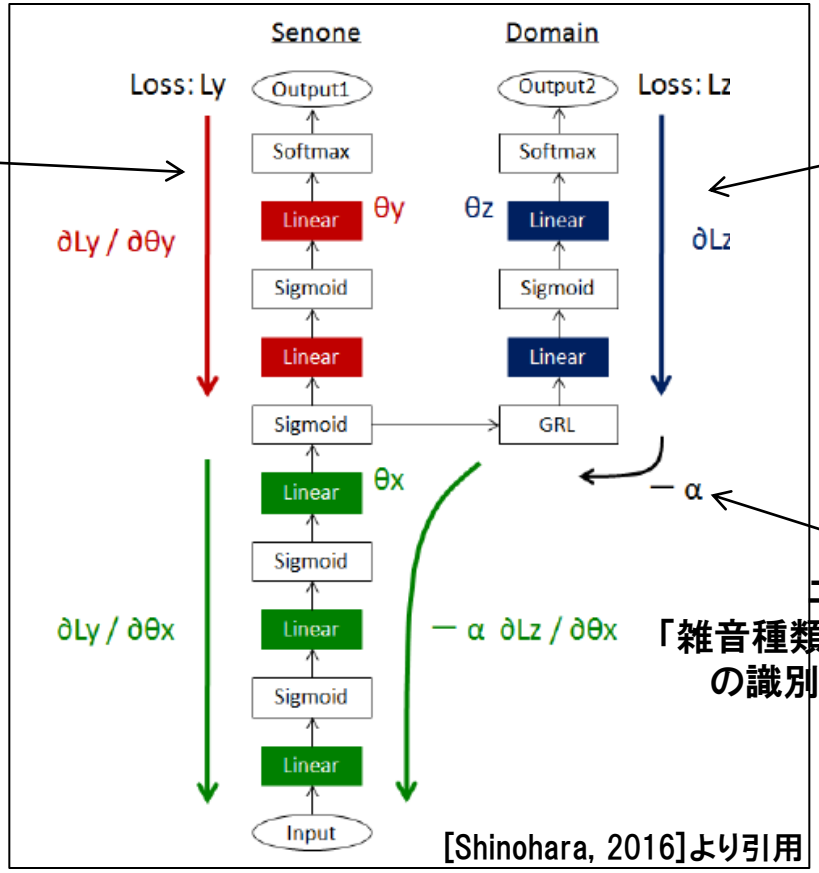
S. Matsuda, et al. Automatic localization of a language-independent sub-network on deep neural networks trained by multi-lingual speech. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013. p. 7359-7362.

音響モデルの敵対的学習 [Shinohara, 2016]

HITACHI
Inspire the Next

- 「雑音種類識別率を最大化するネットワーク」の識別率を最小化するように音響モデルを学習→雑音の変動に頑健に

音素識別率を最大化するネットワーク



雑音種類識別率を最大化するネットワーク

エラーを反転させることで
「雑音種類識別率を最大化するネットワーク」
の識別率を最小化しようと学習が進む

Shinohara, Y., 2016. Adversarial Multi-task Learning of Deep Neural Networks for Robust Speech Recognition. *Interspeech 2016*, pp.2369-2372.

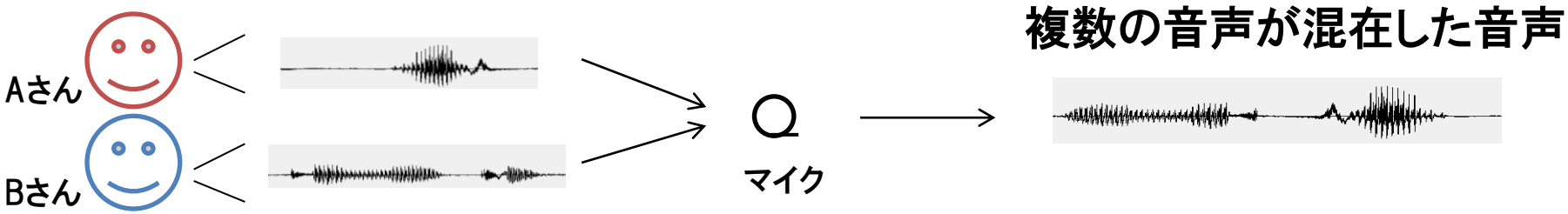
HITACHI
Inspire the Next

混合音声の分離

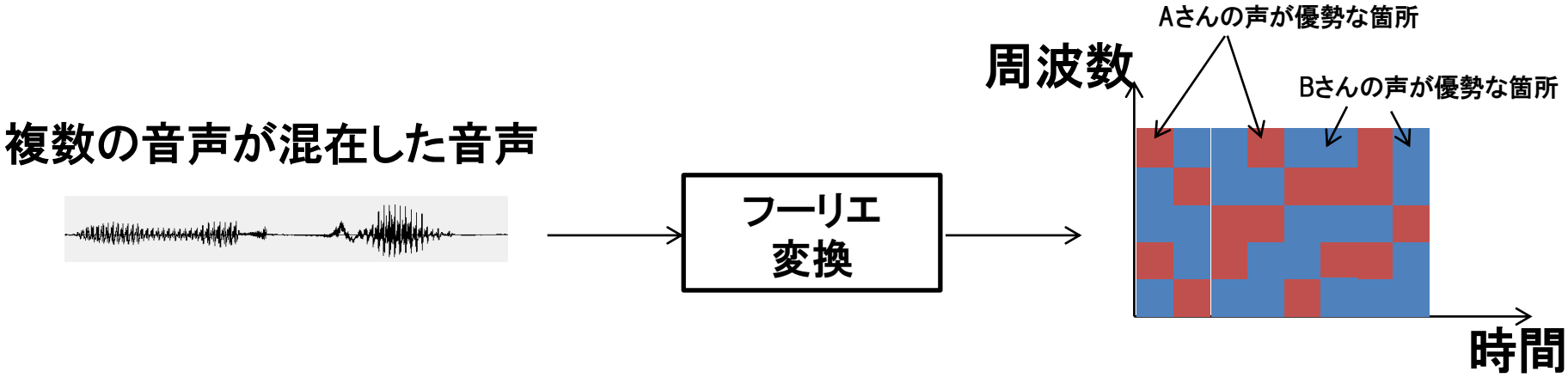
Deep Clustering [Hershey, 2016]

HITACHI
Inspire the Next

- 複数の音声が混在した音声から、もとの音声を取り出したい



- 時間周波数上でのクラスタリング問題と考える

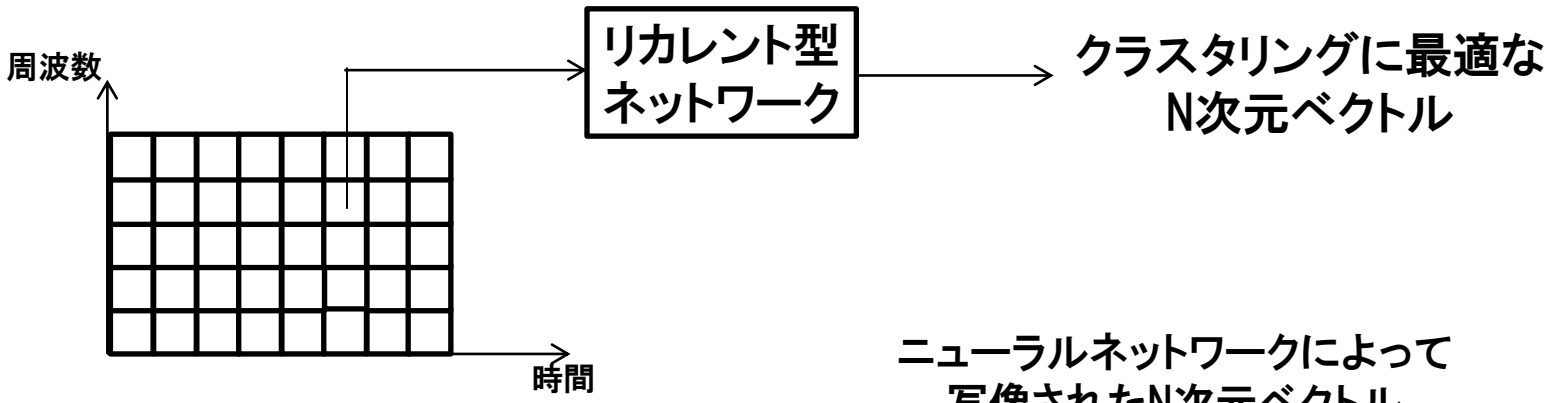


Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016, March). Deep clustering: Discriminative embeddings for segmentation and separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on* (pp. 31-35). IEEE.

Deep Clustering [Hershey, 2016]

HITACHI
Inspire the Next

- クラスタリングに最適な空間へ写像するニューラルネットワークを学習



- 学習基準

ニューラルネットワークによって
写像されたN次元ベクトル

正解(iとjが同じクラスなら1、違うなら0)

$$\mathcal{C}_Y(V) = \|VV^T - YY^T\|_{\mathrm{F}}^2 = \sum_{i,j} \left(\langle v_i, v_j \rangle - \langle y_i, y_j \rangle\right)^2$$

最小化

**同じクラスなら内積が1、違うクラスなら内積が0になるような
N次元ベクトル空間への写像が学習される**

Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016, March). Deep clustering: Discriminative embeddings for segmentation and separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on* (pp. 31-35). IEEE.

Deep Clustering [Hershey, 2016]

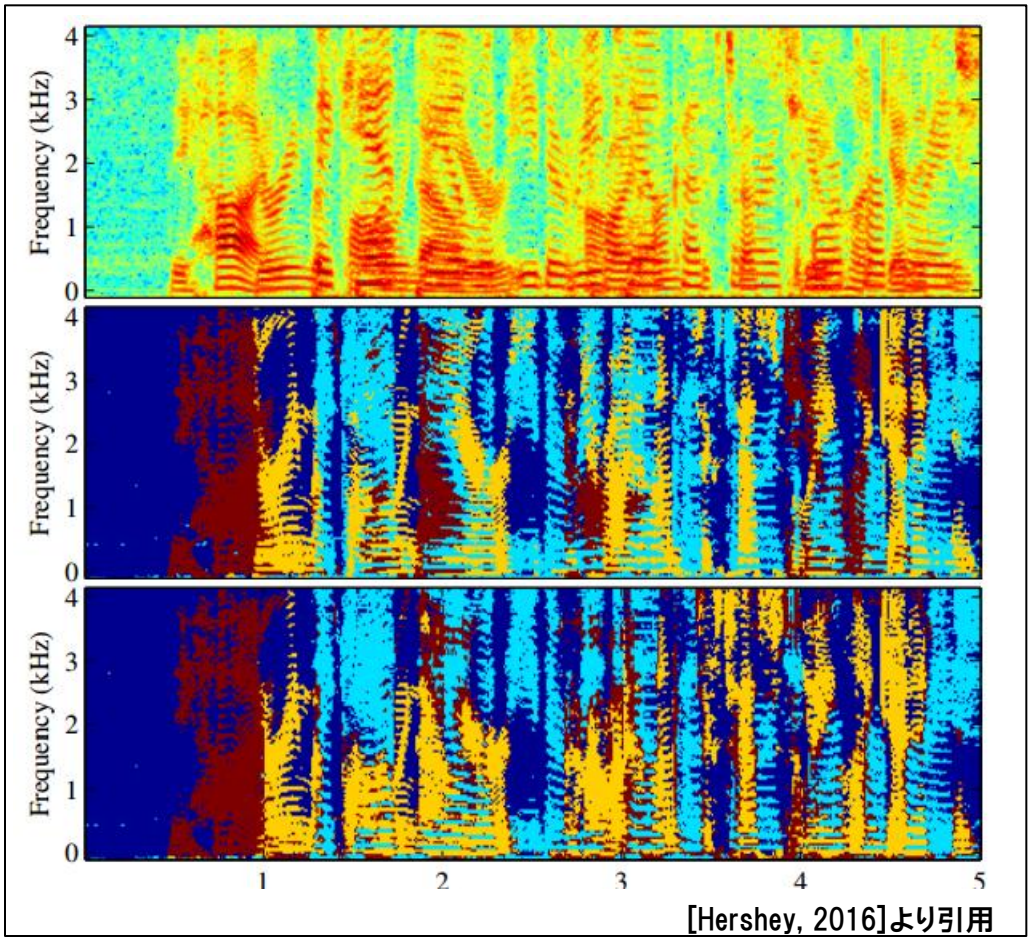
HITACHI
Inspire the Next

- 3話者混合音声も分離可能

3話者混合音声

理想的な
クラスタリング結果
（正解）

Deep Clustering



[Hershey, 2016]より引用

Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016, March). Deep clustering: Discriminative embeddings for segmentation and separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on* (pp. 31-35). IEEE.

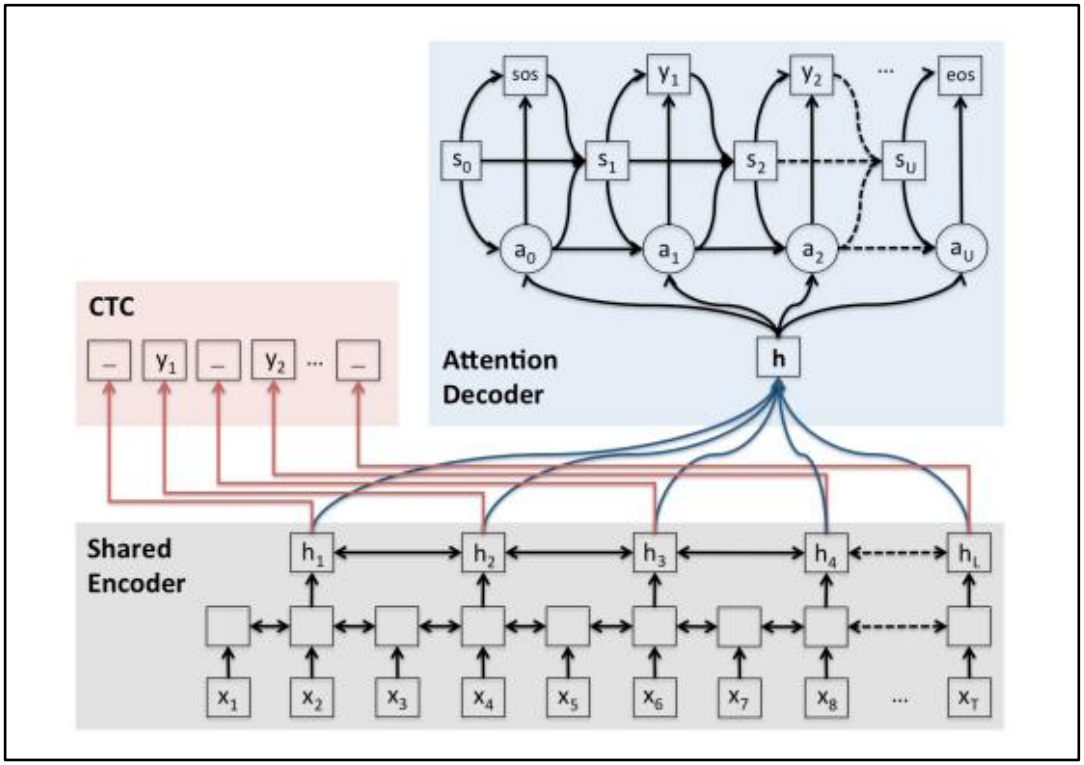
HITACHI
Inspire the Next

End-to-Endの進展

Attention Enc-Dec + CTC [Kim, 2017][Watanabe,2017]

HITACHI
Inspire the Next

- CTCとAttention Decoderのマルチタスクラーニング
  - Attention Decoderのほうが主体
  - CTCの学習効率の良さを利用しつつ、Enc-Decのモデル化の良さを利用



[Kim, 2017]より引用

Kim, S., Hori, T., & Watanabe, S. , Joint ctc-attention based end-to-end speech recognition using multi-task learning. In Proc Interspeech 2017.

Attention Enc-Dec + CTC [Kim, 2017][Watanabe,2017]

HITACHI
Inspire the Next

文字ベースの日本語音声認識
日本語話し言葉コーパス　文字誤り率(%)

Model (hours)	task1	task2	task3
Attention (147h)	20.1	14.0*	32.7
MTL (147h)	16.9	12.7*	28.9
Attention (236h)	17.2	12.4*	25.4
MTL (236h)	13.9	10.2*	22.2
Attention (581h)	11.5	7.9*	9.0
MTL (581h)	10.9	7.8*	8.3
MTL2 (581h)	**9.5**	**7.0**	**7.8**
GMM-discr. [3] (236h for AM, 581h for LM)	11.2	9.2	12.1
DNN-hybrid [3] (236h for AM, 581h for LM)	9.0	7.2	9.6
CTC-syllable [4] (581h)	9.4	7.3	7.5

← 漢字カナ Enc-Dec + CTC

← sMBR-DNN+単語言語モデル
（データサイズが少し小さいことに注意）

← カナCTC+単語言語モデル

[S. Watanabe, 2017]より引用

形態素解析、発音辞書、言語モデルなしで、極めて高い精度

S. Watanabe, et al., End-to-end Japanese ASR without using morphological analyzer, pronunciation dictionary and language model, 日本音響学会2017年春季講演論文集

MAPデコーディング [Kanda, 2016][Kanda, 2017]

HITACHI
Inspire the Next

- とはいえ、言語モデルを組み合わせたい状況は結構ある
  - 音声コーパスとは異なる言語ドメインで利用したい
  - 新規語彙を追加したい
  - 精度向上したい
    - 学習テキストのほうが大量にあるため、より高精度な言語モデルが学習できる

- End-to-Endモデルと言語モデルの組み合わせを考える必要
  - 従来は単純なlog-linear

単語列　　サブワード列（文字、カナ、音素等）
音声特徴量列

$$\tilde{W} = \arg\max_{W,L} \log \Pr(W) + \alpha \cdot \log \Pr(L \mid X)$$

言語モデル　　End-to-Endモデル

理論的根拠がない

MAPデコーディング [Kanda, 2016][Kanda, 2017]

HITACHI
Inspire the Next

- End-to-Endモデルに向けた，第3の数式

$$\tilde{W} = \arg\max_{W} \Pr(W \mid X)$$ ←―― **第1の基礎数式**

MAPデコーディング [Kanda, 2016][Kanda, 2017]

HITACHI
Inspire the Next

- End-to-Endモデルに向けた，第3の数式

$$\tilde{W} = \arg\max_{W} \Pr(W \mid X)$$

← 第1の基礎数式

$$= \arg\max_{W} \sum_{L} \Pr(W \mid L) \Pr(L \mid X)$$

$$\cong \arg\max_{W} \{ \max_{L} \Pr(W \mid L) \Pr(L \mid X)^{\alpha} \}$$

← 第3の数式：
MAPデコーディング方式

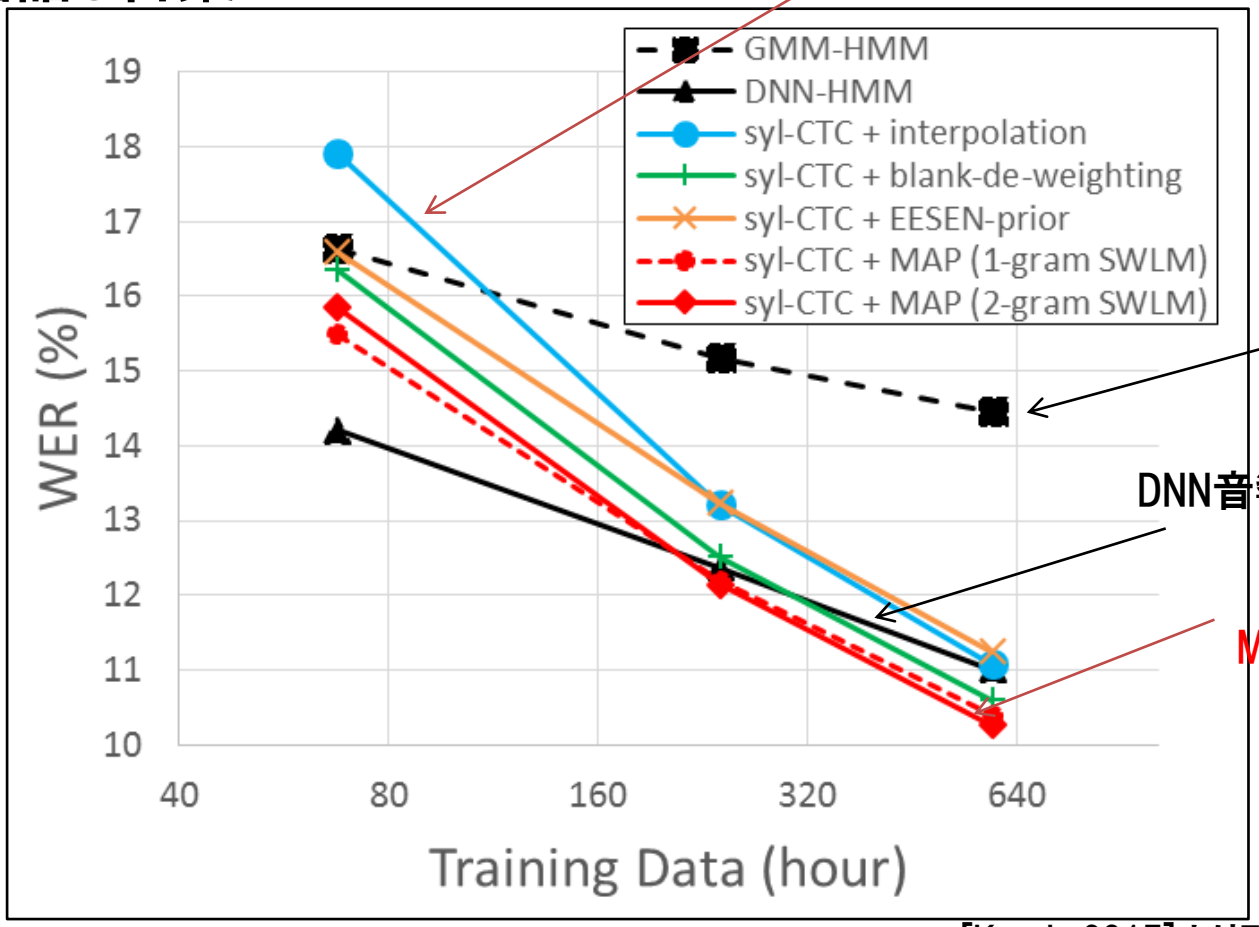
サブワード列Lが与えられたときの単語列Wの確率

サブワードEnd-to-End音響モデルのスコア

MAPデコーディング [Kanda, 2016][Kanda, 2017]

HITACHI
Inspire the Next

日本語話し言葉コーパス



[Kanda,2017]より引用

Naoyuki Kanda, Xugang Lu, Hisashi Kawai, Maximum A Posteriori based Decoding for End-to-End Acoustic Models, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017

まとめ

HITACHI
Inspire the Next

- Deep Learningに基づく音声認識について紹介
  - 生成モデルアプローチ→DNNによる出力確率の計算
  - 識別モデルアプローチ→End-to-End（CTC, Attention Enc-Dec)
- 音声認識におけるDeep Learningの動向紹介
  - 多種のネットワーク、話者適応、系列識別学習、敵対的学習、多言語データの活用、Deep Clustering、CTC+Attention Enc-Dec、CTCのデコーディング方式・・
- 今後の展望
  - End-to-Endモデルの発展
  - 話者適応の高速化
  - マルチモーダル
  - より強い残響、雑音(CHiMEチャレンジ)
  - 言語ドメインの違いへの対応(MGBチャレンジ)

参考文献

HITACHI
Inspire the Next

- DNN-HMM
  - A. Mohamed et al., ”Deep belief networks for phone recognition,” In NIPS workshop on deep learning for speech recognition and related applications, volume 1, page 39 (2009).
  - D. Yu et al., ”Roles of pre-training and fine tuning in context-dependent DBN-HMMs for real world speech recognition,” Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2010).
  - F. Seide et al., ”Conversational speech transcription using context-dependent deep neural networks,” Proc. Interspeech, pp. 437-440 (2011).
  - G. Dahl et al., ”Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” IEEE Trans. SAP, 20(1):30-42 (2012).

  (結果抜粋)
  - G. Hinton, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." Signal Processing Magazine, IEEE 29.6 (2012): 82-97.
  - N. Kanda et al., ”Elastic spectral distortion for low resource speech recognition with deep neural networks,” Proc. ASRU, pp. 309-314 (2013).

  (特徴量)
  - A. Mohamed et al., ”Understanding how deep belief networks perform acoustic modelling,” Proc. ICASSP, pp. 4273-4276 (2012).
  - T. Sainath et al., ”Learning filter banks within a deep neural network framework,” Proc. ASRU, pp. 97-302 (2013).
  - Y. Hoshen et al., ”Speech acoustic modeling from raw multichannel waveforms,” Proc. ICASSP, pp. 4624-4628 (2015).
  - T. Sainath et al., ”Learning the speech frontend with raw waveform CLDNNs,” Proc. Interspeech (2015).
  - T. Sainath, et al., "Factored spatial and spectral multichannel raw waveform CLDNNs," Proc. ICASSP, pp. 5075-5079 (2016).
  - H. Hermansky et al., ”Tandem connectionist feature extraction for conventional hmm systems,” Proc. ICASSP, volume 3, pp. 1635-1638 (2000).

参考文献

HITACHI
Inspire the Next

- CTC
  - A. Graves et al., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," Proc. ICML, pp. 369-376. ACM (2006).
  - A. Maas et al., ”First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs,” arXiv preprint arXiv:1408.2873 (2014).
  - A. Hannun et al., ”Deepspeech: Scaling up end-to-end speech recognition,” arXiv preprint arXiv:1412.5567 (2014).
  - H. Sak et al., ”Learning acoustic frame labeling for speech recognition with recurrent neural networks,” Proc. ICASSP, pp. 4280-4284 (2015).
  - Y. Miao et al., ”EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding,” Proc. ASRU, pp. 167-174 (2015).
  - N. Kanda et al., ”Maximum a posteriori based decoding for CTC acoustic models,” Proc. Interspeech, pp. 1868-1872 (2016).
  - N. Kanda, et al., "Maximum a posteriori based decoding for end-to-end acoustic models," IEEE/ACM Trans. on ASLP, 2017 (to appear)
  - N. Kanda, et al., “Minimum Bayes risk training of CTC acoustic models in maximum a posteriori based decoding framework”, Proc ICASSP, pp. 4855-4859 (2017).

参考文献

HITACHI
Inspire the Next

- Attention Enc Dec
  - J. K. Chorowski, et al. "End-to-end continuous speech recognition using attention-based recurrent NN: First results.“, arXiv preprint arXiv (2014).
  - J. K. Chorowski, et al., “Attention-based models for speech recognition,” Proc. NIPS , pp. 577-585 (2015).
  - D. Bahdanau, et al. End-to-end attention-based large vocabulary speech recognition. “ Proc. ICASSP, p. 4945-4949 (2016).
  - W. Chan, et al. "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” Proc. ICASSP p. 4960-4964 (2016).
  - Kim, S., Hori, T., & Watanabe, S. , “Joint CTC-attention based end-to-end speech recognition using multi-task learning”, Proc Interspeech (2017).
  - S. Watanabe, et al., End-to-end Japanese ASR without using morphological analyzer, pronunciation dictionary and language model, 日本音響学会2017年春季講演論文集 (2017).

参考文献

HITACHI
Inspire the Next

- Topics

(コンビネーション)

- P. Shen, et al. "The NICT ASR system for IWSLT 2014," Proc. IWSLT (2014).

(識別学習)

- K. Vesely et al., ”Sequence-discriminative training of deep neural networks,” Proc. Interspeech, pp. 2345-2349 (2013).
- H. Su et al., ”Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription,” Proc. ICASSP, pp. 6664-6668 (2013).

（マルチリンガル）

- S. Matsuda et al., ”Automatic localization of a language-independent sub-network on deep neural networks trained by multi-lingual speech,” Proc. ICASSP, pp. 7359-7362 (2013).
- J. Huang et al., ”Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” Proc. ICASSP, pp. 7304-7308 (2013).
- G. Heigold et al., ”Multilingual acoustic models using distributed deep neural networks,” Proc. ICASSP, pp. 8619-8623 (2013).

（敵対的学習）

- Shinohara, Y., 2016. Adversarial Multi-task Learning of Deep Neural Networks for Robust Speech Recognition. Proc. Interspeech, pp.2369-2372 (2016).

（Deep Clustering)

- J. R Hershey, et al. Deep clustering: Discriminative embeddings for segmentation and separation. Proc. ICASSP, pp. 31-35 (2016)

参考文献

HITACHI
Inspire the Next

- Topics

  （適応）

  - H. Liao, ”Speaker adaptation of context dependent deep neural networks,” Proc. ICASSP, pp. 7947-7951 (2013).
  - D. Yu et al., ”KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition,” Proc. ICASSP, pp. 7893-7897 (2013).
  - T. Ochiai et al., ”Speaker adaptive training using deep neural networks,” Proc. ICASSP, pp. 6349-6353 (2014).
  - M. Delcroix et al., ”Context adaptive deep neural networks for fast acoustic model adaptation,” Proc. ICASSP, pp. 4535-4539 (2015).
  - P. Swietojanski, et al., "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," Proc. SLT, p. 171-176 (2014).
  - G. Saon, et al. Speaker adaptation of neural network acoustic models using i-vectors. Proc. *ASRU*, p. 55-59 (2013).
  - M. Seltzer, et al., An investigation of deep neural networks for noise robust speech recognition. Proc. ICASSP, p. 7398-7402 (2013).


HITACHI
Inspire the Next