Innovative R&D by NTT

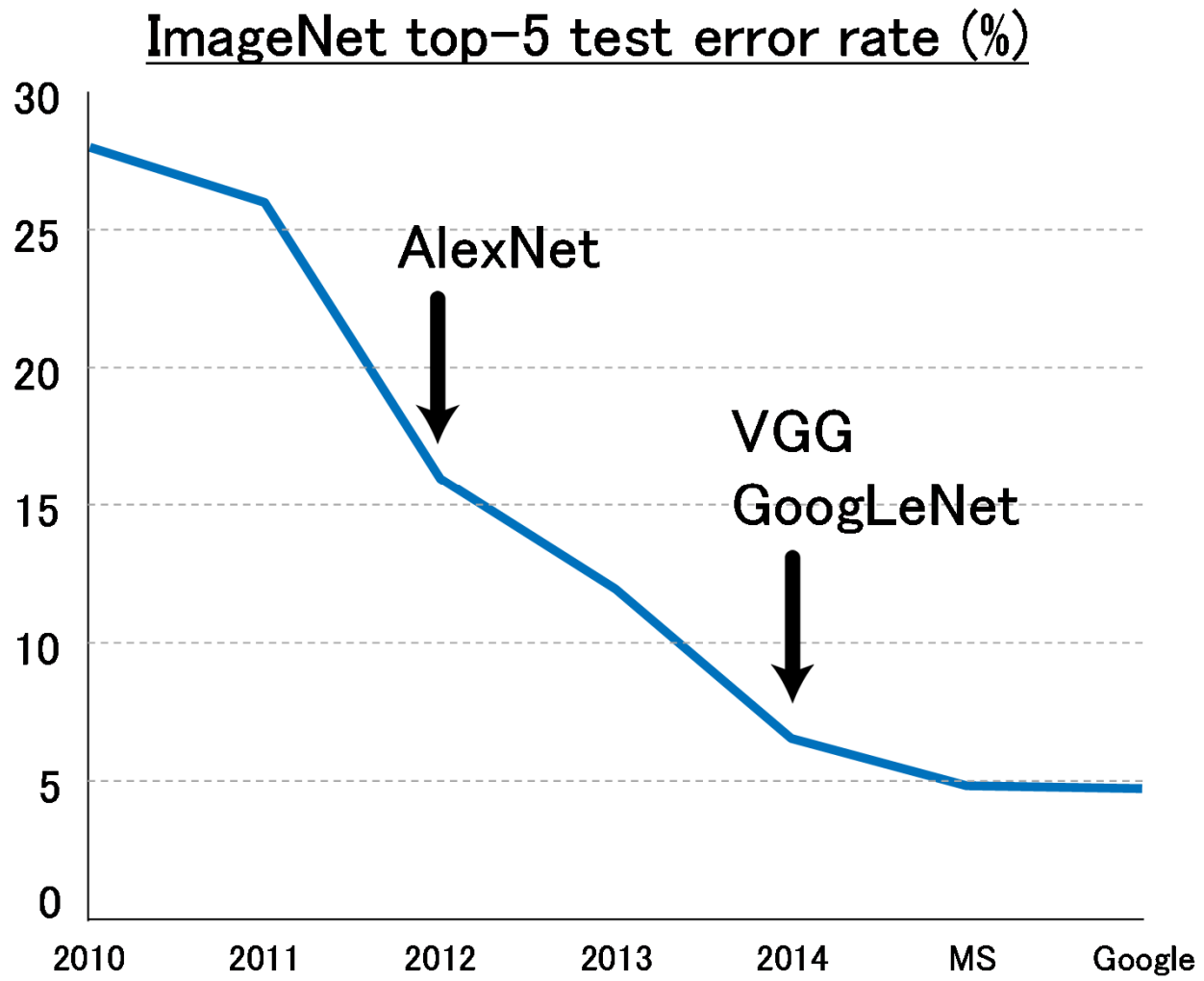
ディープ・ラーニングって何が凄いの？

日本電信電話株式会社
コミュニケーション科学基礎研究所
藤本　雅清

2015/09/16
日本音響学会2015年秋季研究発表会　ビギナーズセミナー

巷で大流行

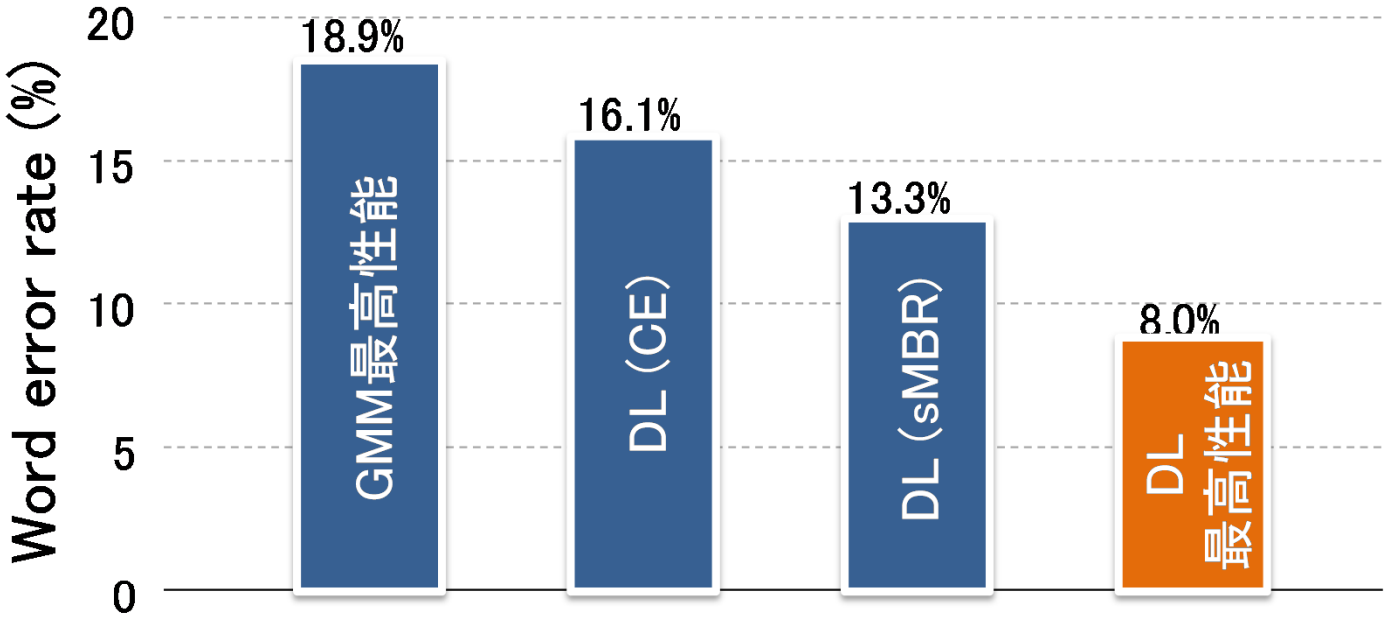
画像認識の性能がうなぎ上り



音声認識も負けていない

Switchboardタスクの結果

話者独立, one-pass, 300h
[Kingsbury, et al., Interspeech'12]



話者依存, multi-pass, 2000h
[Saon et al., 2015, arXiv:1505.05899]

GMM: 混合正規分布を用いた従来技術
CE:クロスエントロピー最小化学習
sMBR:系列識別学習の一つ

概要

1. ディープニューラルネットワークの基礎
2. 音声研究への応用
3. 様々なネットワーク
4. DNNのツールキット、開発環境
5. まとめ

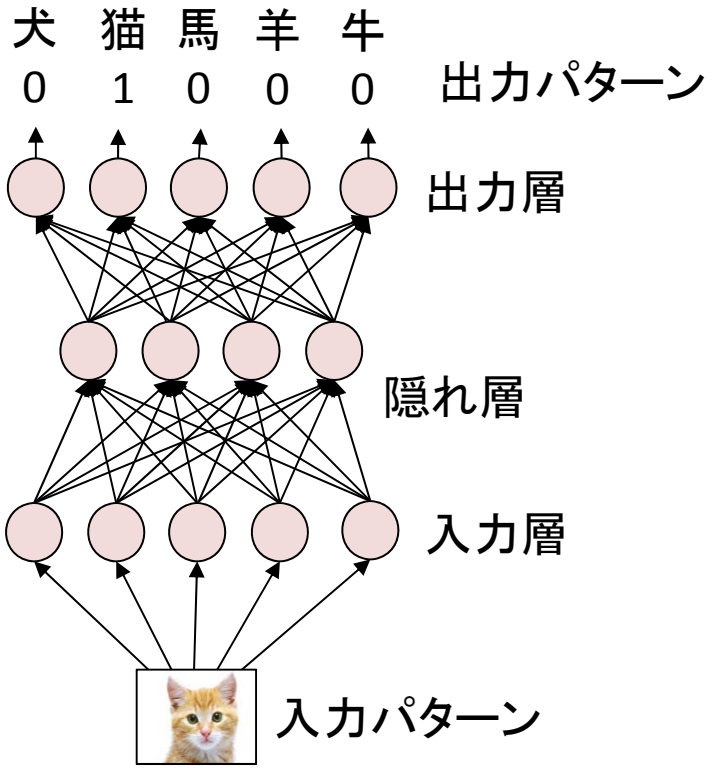
概要

1. ディープニューラルネットワークの基礎
2. 音声研究への応用
3. 様々なネットワーク
4. DNNのツールキット、開発環境
5. まとめ

深層学習(ディープラーニング)
≒(ディープな)ニューラルネットワーク

ニューラルネットワーク(多層パーセプトロン: MLP)

生物の神経回路網を模した数理モデル

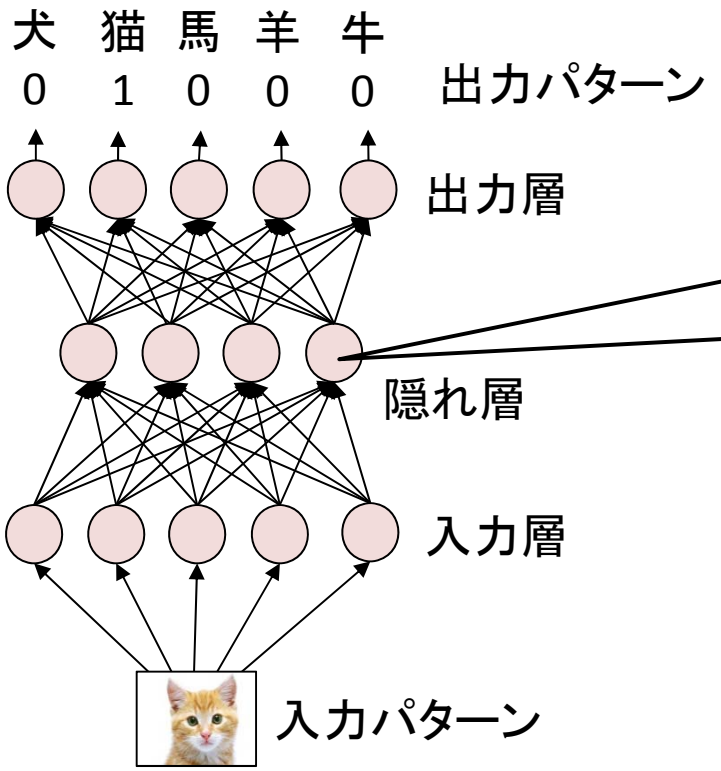


データから任意の識別関数
を学習可能

深層学習（ディープラーニング）
≒（ディープな）ニューラルネットワーク

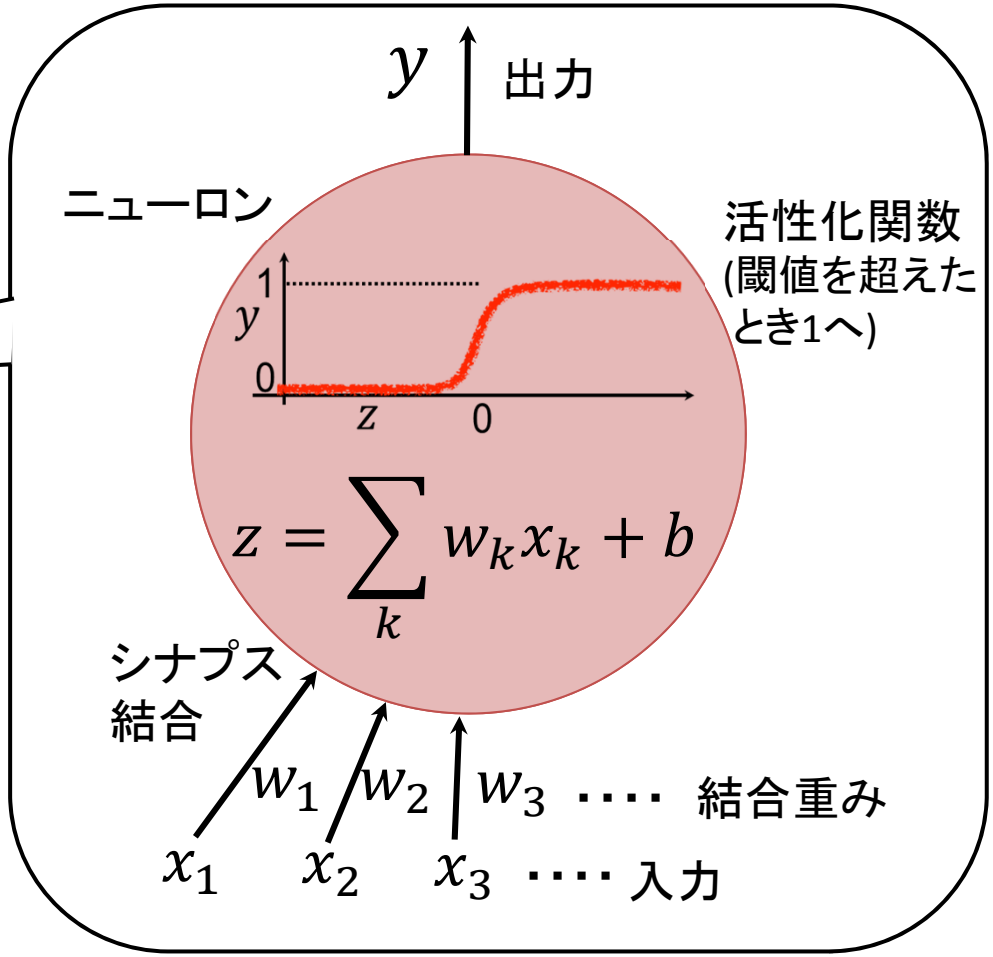
ニューラルネットワーク（多層パーセプトロン: MLP）

生物の神経回路網を模した数理モデル



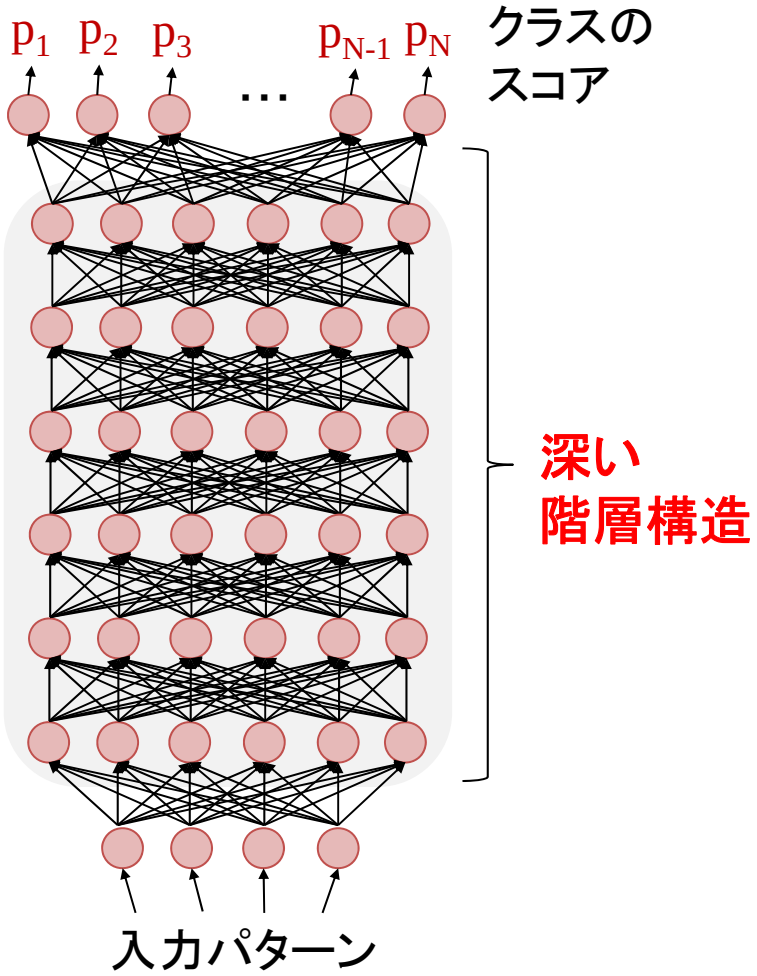
データから任意の識別関数
を学習可能

線形2クラス分類が可能



ディープニューラルネットワーク(DNN)

パターン識別用DNN



■ 深くなっただけ!?

隠れ層が5～8層　(従来は2～3層)

■ 幅も広い

ニューロン数 2000以上 (従来は数百程度)
(e.g. パラメータ数 5000万)

■ なぜ今なのか？

・学習アルゴリズムの進歩
　層毎の事前学習、確率的勾配降下法 …
・計算機の進歩
　GPGPUによる超並列計算、演算ライブラリの充実
・更なる発展
　Deep CNN, Deep RNN/LSTM,…

概要

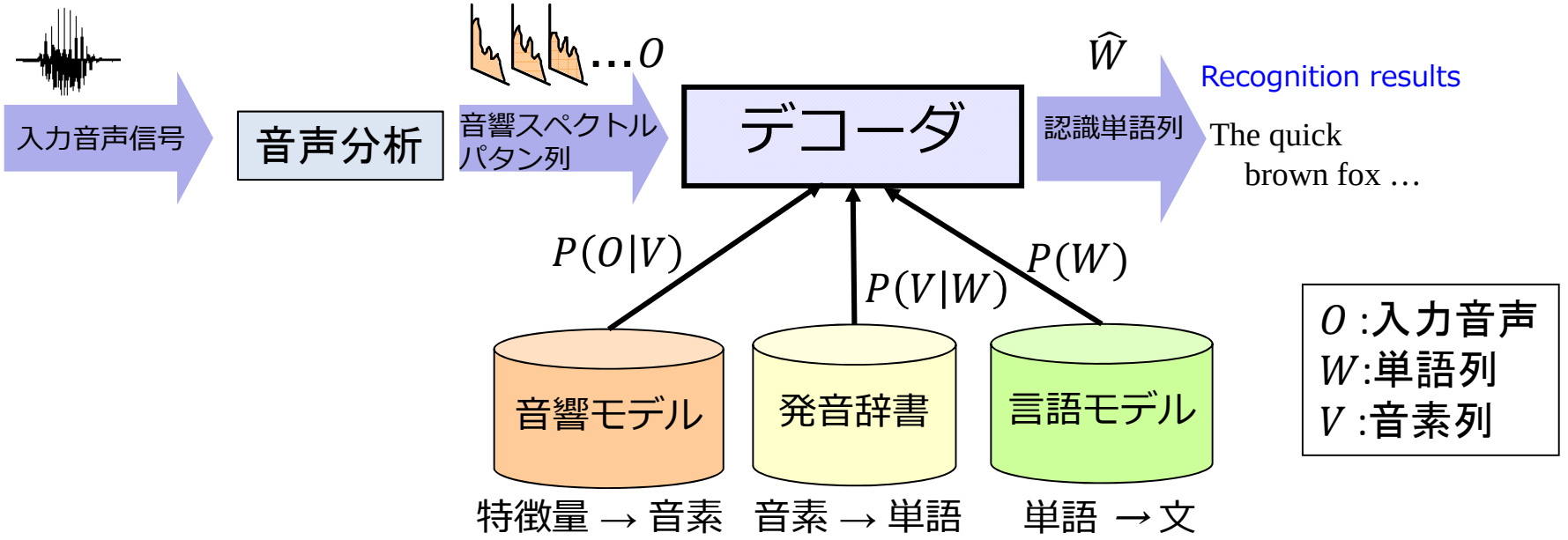
1. ディープニューラルネットワークの基礎
2. 音声研究への応用
3. 様々なネットワーク
4. DNNのツールキット、開発環境
5. まとめ

音声認識への応用

確率モデルによる音声認識 [Jelinek 75]

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|O) \approx \underset{W}{\operatorname{argmax}}\, P(O|V)P(V|W)P(W)$$

音響尤度　発音確率　言語確率

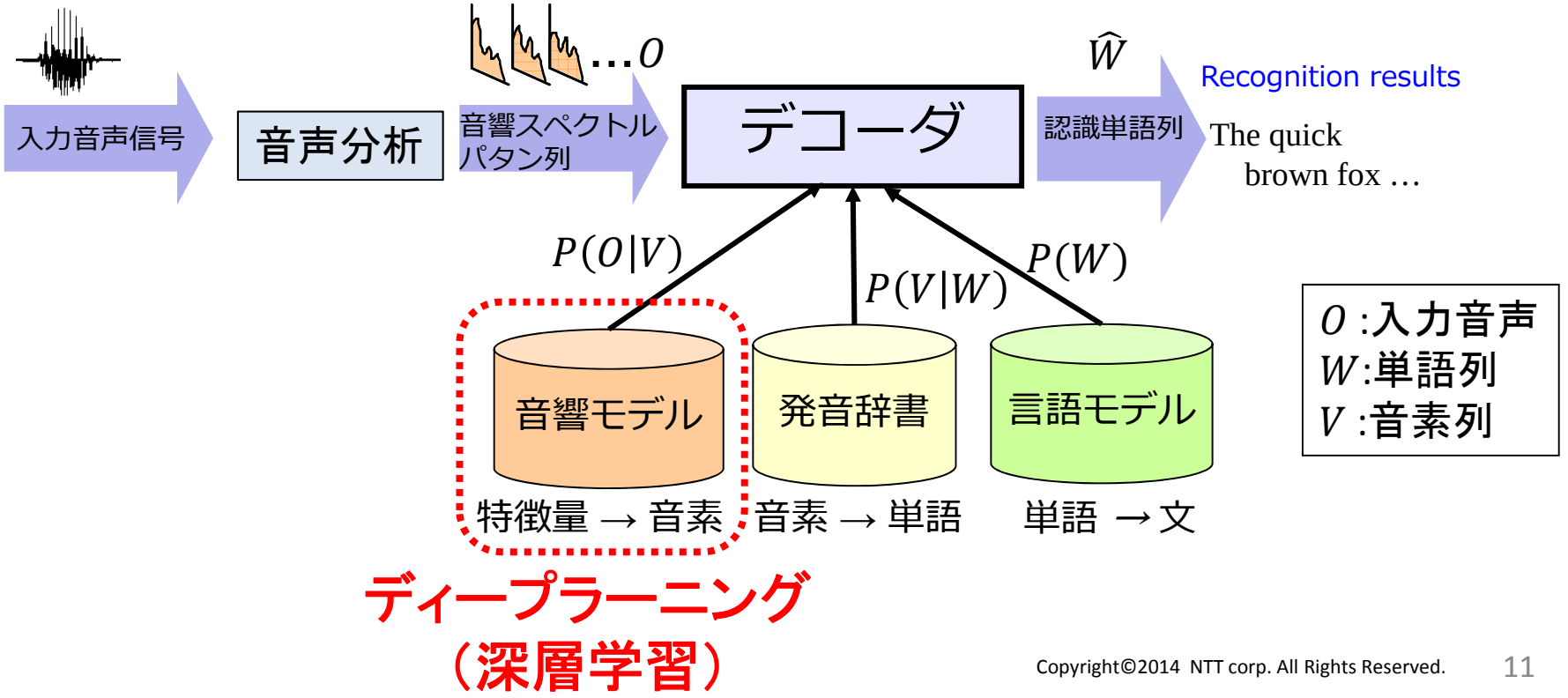


音声認識への応用

確率モデルによる音声認識 [Jelinek 75]

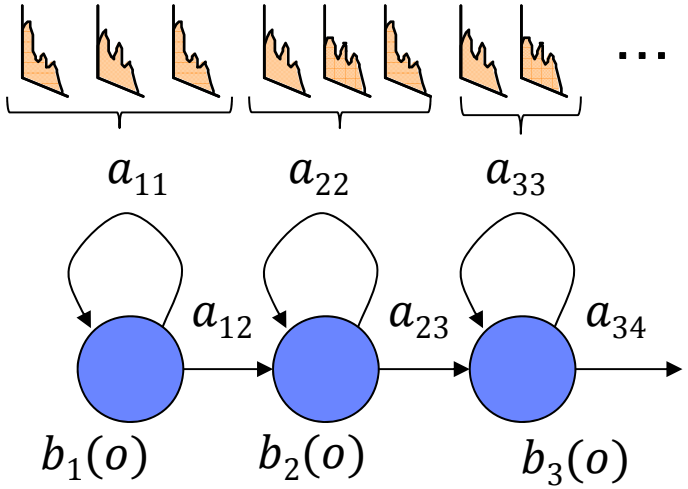
$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|O) \approx \underset{W}{\operatorname{argmax}}\, P(O|V)P(V|W)P(W)$$

音響尤度　発音確率　言語確率



音響モデル

隠れマルコフモデル(Hidden Markov model: HMM)



… 音響スペクトルパタン列

a_{ij}:状態遷移確率

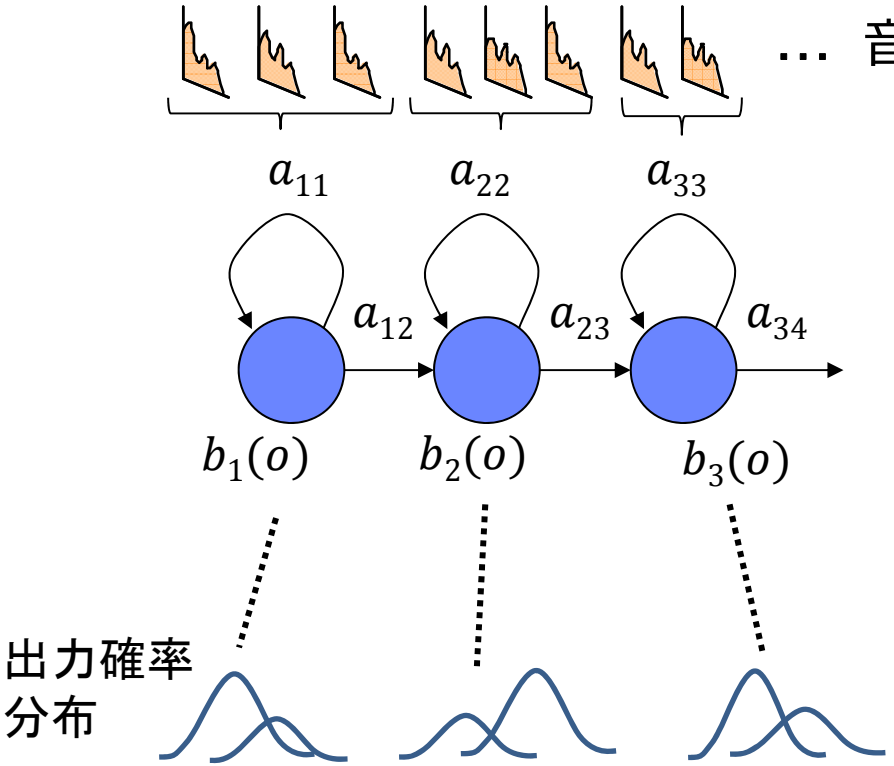
$b_j(o)$:出力確率分布

音響尤度の計算

$$P(O|V) = \Sigma_{S_{V,T}} \Pi_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(o_t)$$

音響モデル

隠れマルコフモデル(Hidden Markov model: HMM)



a_{ij}:状態遷移確率

$b_j(o)$:出力確率分布

音響尤度の計算

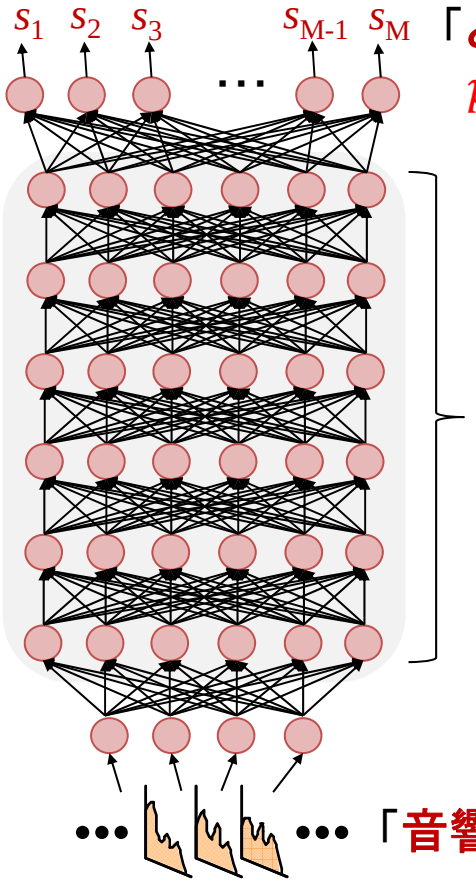
$$P(O|V) = \Sigma_{S_{V,T}} \Pi_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(o_t)$$

混合ガウス分布(Gaussian Mixture Model: GMM)

$$b_j(o) = \Sigma_k c_{jk} N\left(o \middle| \mu_{jk}, \Sigma_{jk}\right)$$

“GMM-HMM” と呼ばれる

DNNに基づく音響モデル [Hinton 2012]



「**どのHMM状態か**」
$p(s_j|o_t)$を出力

HMMの出力確率にDNNを利用*

$$b_j(o_t) = p(o_t|s_j) = \frac{p(s_j|o_t)p(o_t)}{p(s_j)}$$

$p(s_j)$: 状態のユニグラム確率として推定
$p(o_t)$: 識別に影響しないので無視

*ANN-HMM [Morgan 93] と同じ

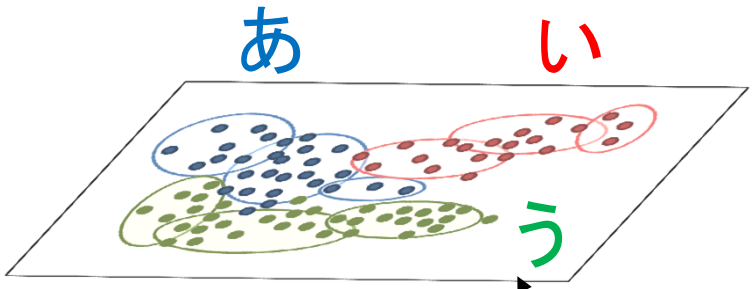
“DNN-HMM” と呼ばれる

GMM vs. DNN

特徴ベクトル
空間
(MFCC/FBank等)

GMM vs. DNN

特徴ベクトル空間上のサンプルに
ガウス分布を当てはめて識別



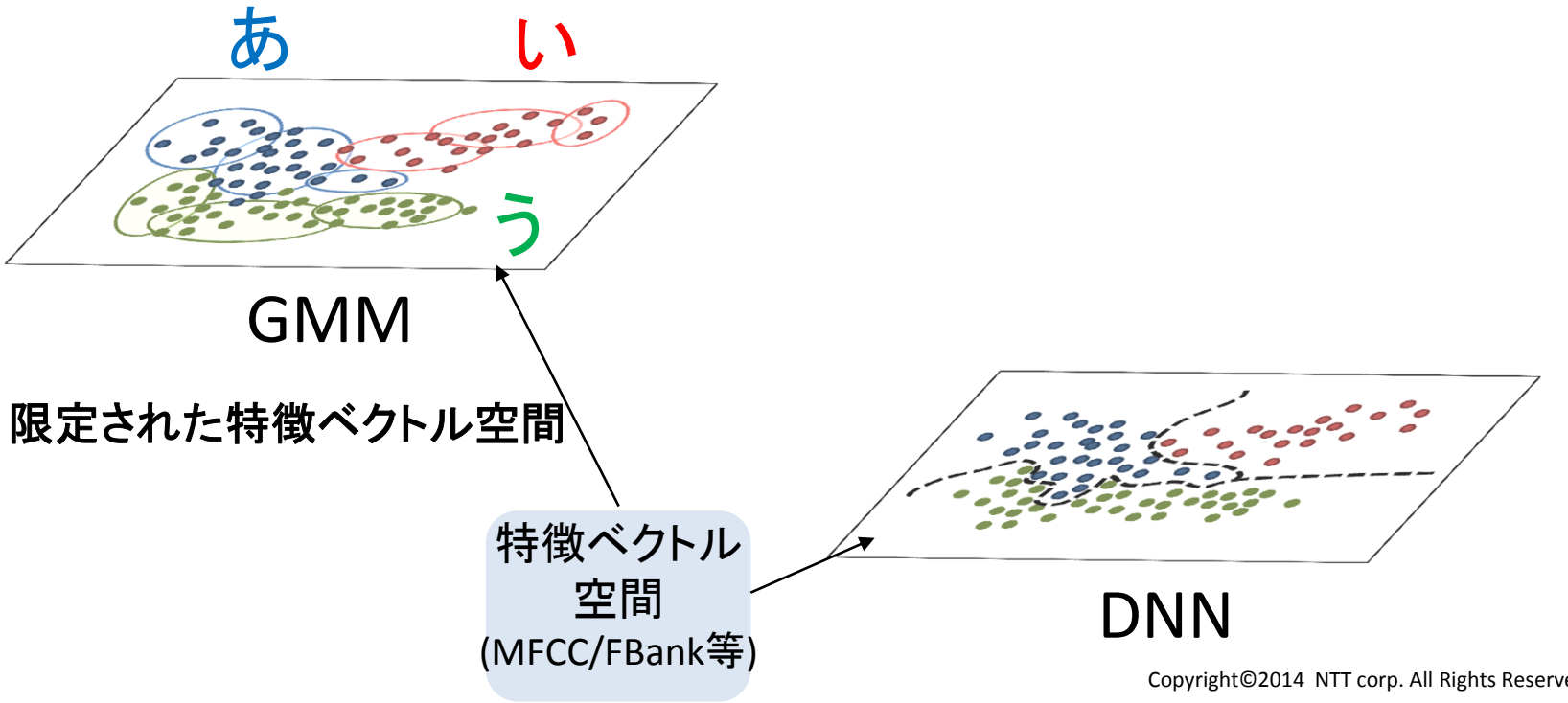
GMM

限定された特徴ベクトル空間

特徴ベクトル
空間
(MFCC/FBank等)

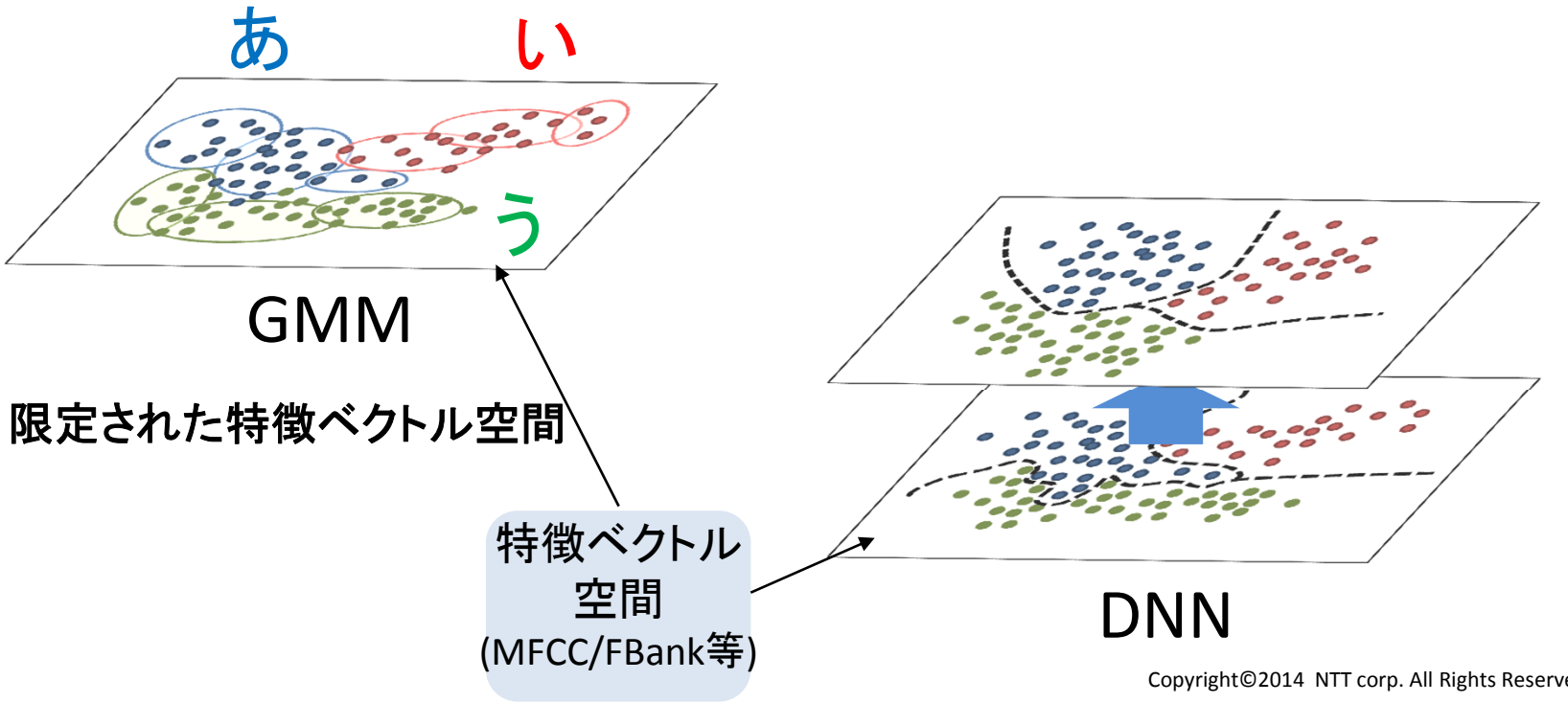
GMM vs. DNN

特徴ベクトル空間上のサンプルに
ガウス分布を当てはめて識別



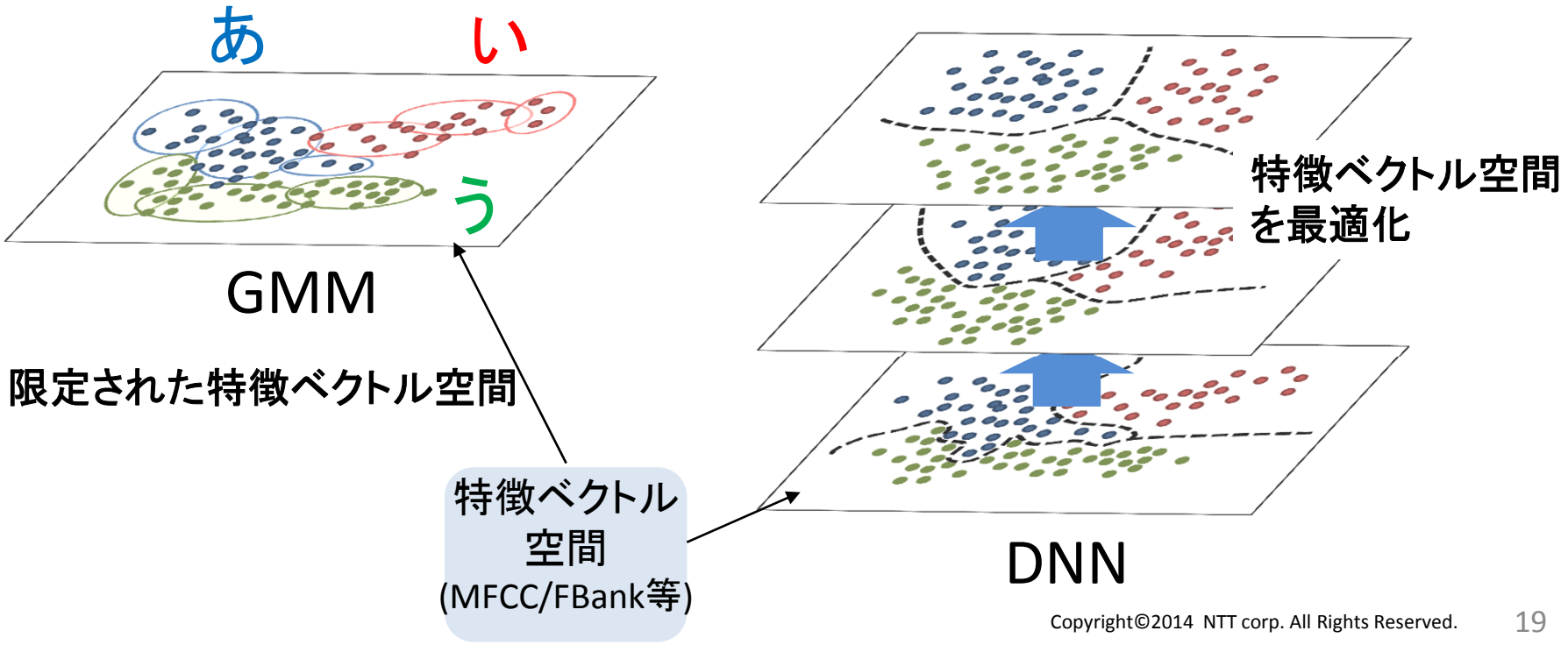
GMM vs. DNN

特徴ベクトル空間上のサンプルに
ガウス分布を当てはめて識別



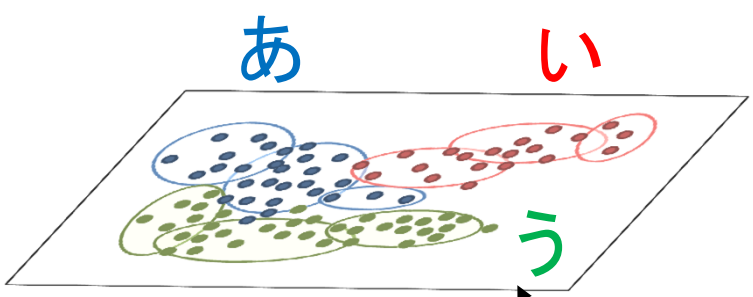
GMM vs. DNN

特徴ベクトル空間上のサンプルに
ガウス分布を当てはめて識別



GMM vs. DNN

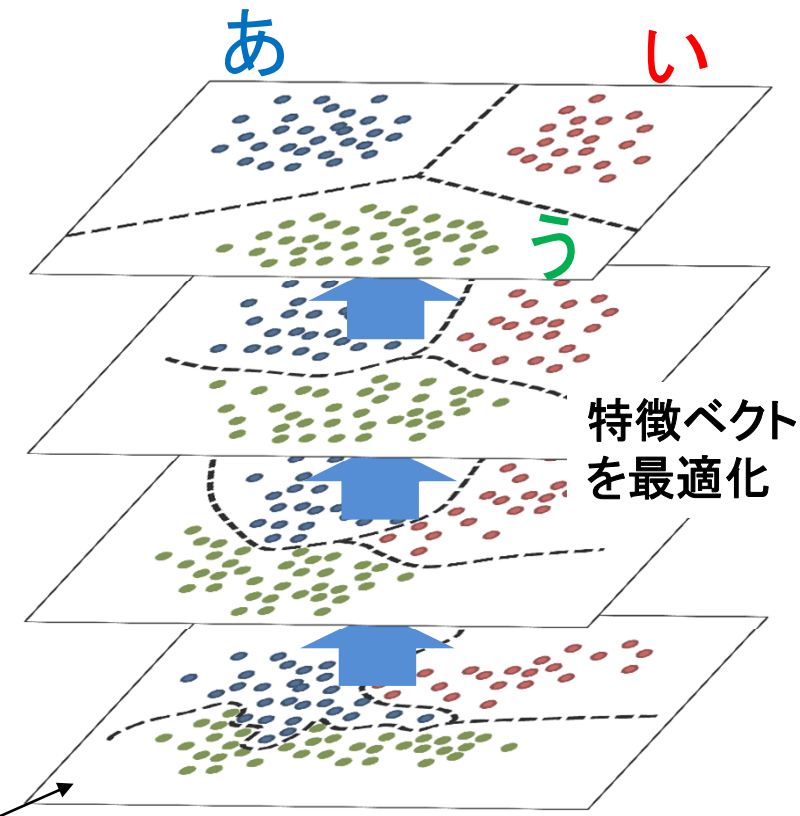
特徴ベクトル空間上のサンプルに
ガウス分布を当てはめて識別



GMM

限定された特徴ベクトル空間

多段階の非線形変換により
音素の識別に有効な高次元空間を
構成して識別



特徴ベクトル空間
を最適化

DNN

特徴ベクトル
空間
(MFCC/FBank等)

音声認識だけ？
音声合成への応用 [Zen 2013]

http://static.googleusercontent.com/media/research.google.com/ja//pubs/archive/41539.pdf

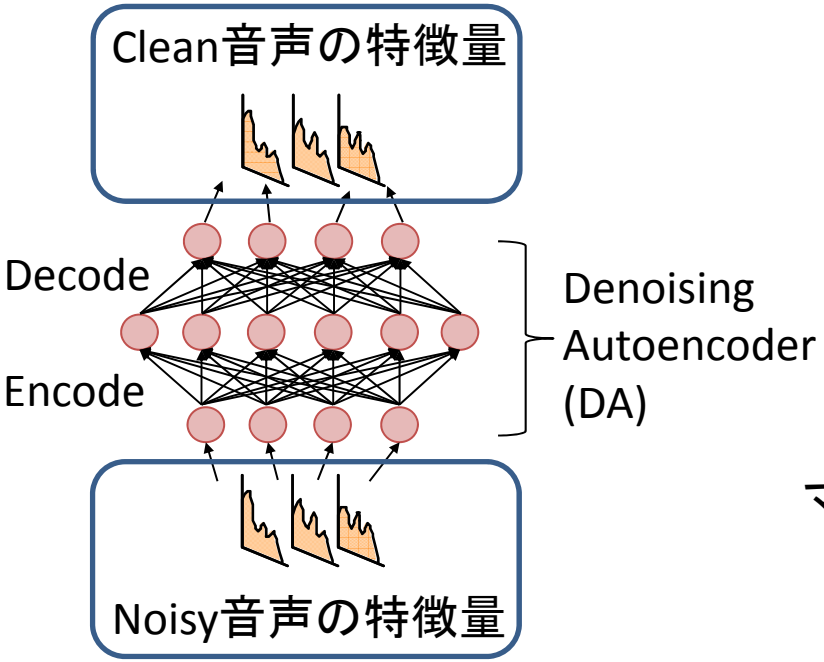
音声認識だけ？
音声強調、雑音抑圧への応用

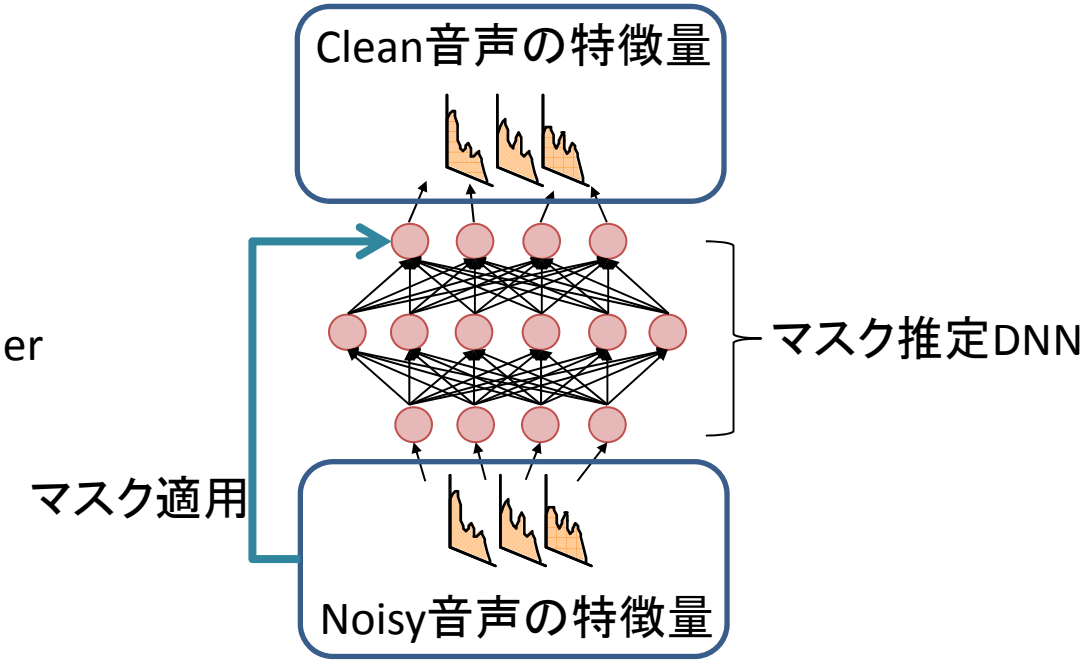
特徴量強調

- 入力: Noisy特徴量
- 出力: Clean特徴量

バイナリマスク推定

- 入力: Noisy特徴量
- 出力: バイナリマスク





概要

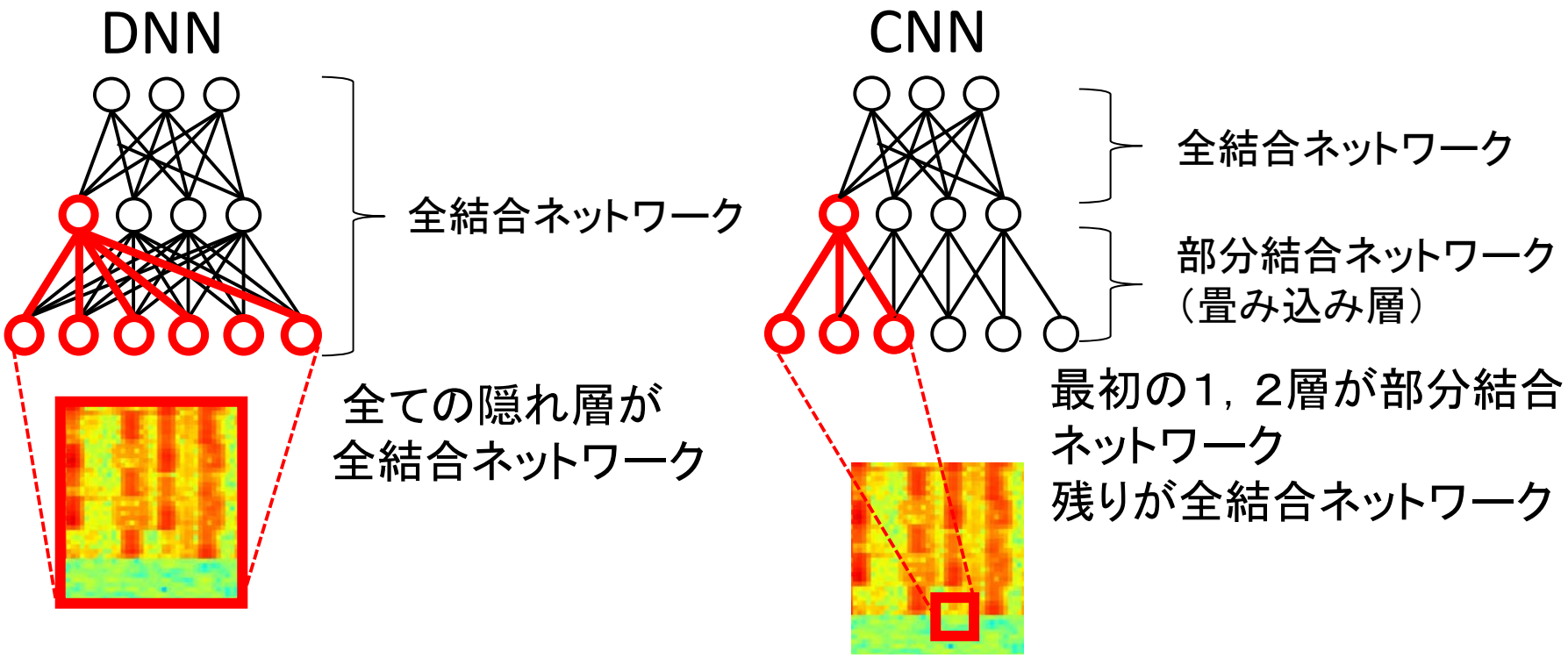
1. ディープニューラルネットワークの基礎
2. 音声研究への応用
3. 様々なネットワーク
4. DNNのツールキット、開発環境
5. まとめ

様々なネットワーク

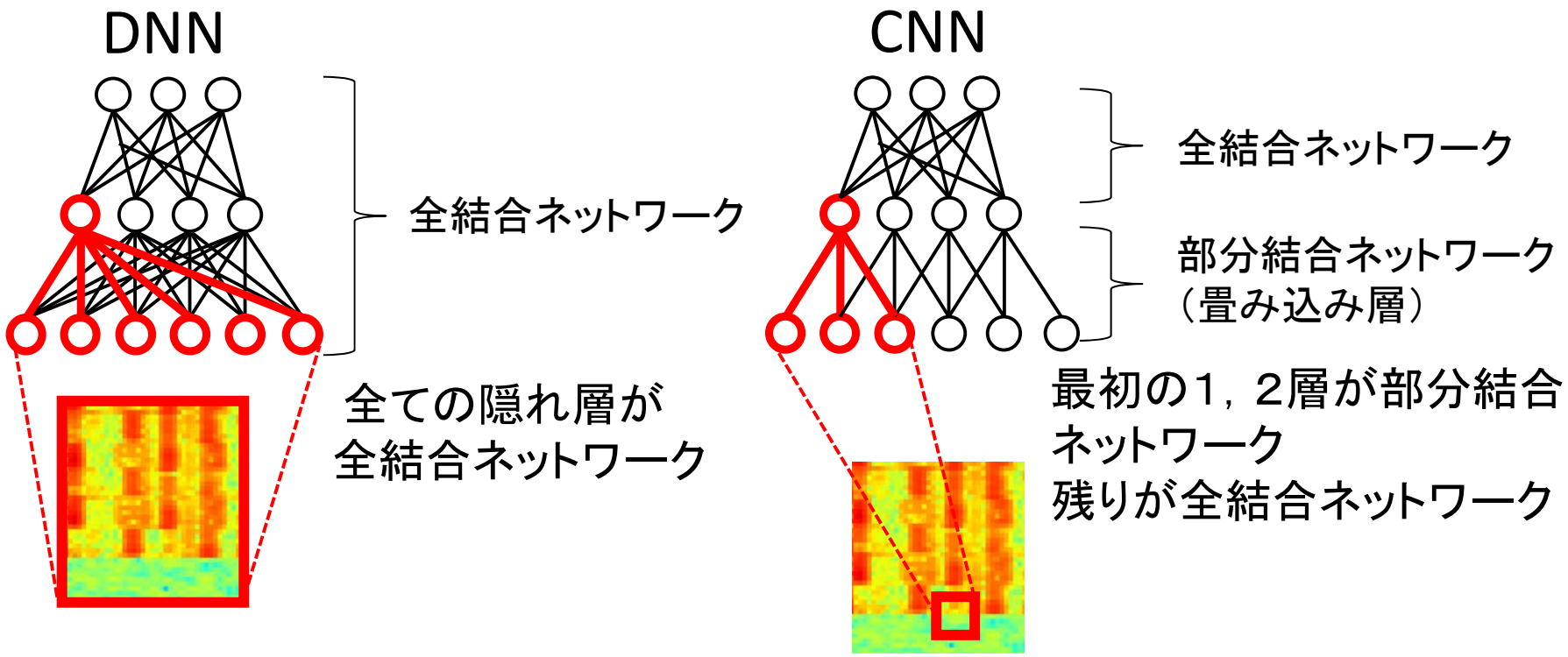
雑音の影響を軽減したり、時系列解析に適したネットワークの有効性が示されつつある

- 畳み込みニューラルネットワーク（Convolutional NN: CNN）
  - 画像認識分野での応用
- リカレントニューラルネットワーク（Recurrent NN: RNN）
  - （双方向）時系列解析

畳み込みニューラルネットワーク(CNN) [Abdel-Hamid 2012]



畳み込みニューラルネットワーク(CNN) [Abdel-Hamid 2012]



大局的特徴抽出

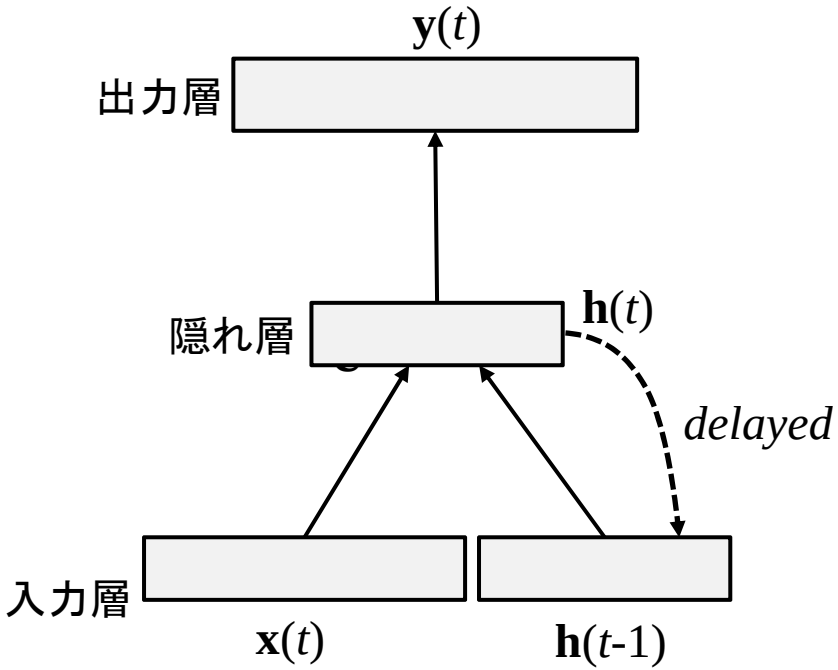
全周波数帯域の情報が
全てのノードに伝搬

局所的特徴抽出

近傍の情報のみが伝搬
局所的な特徴量表現

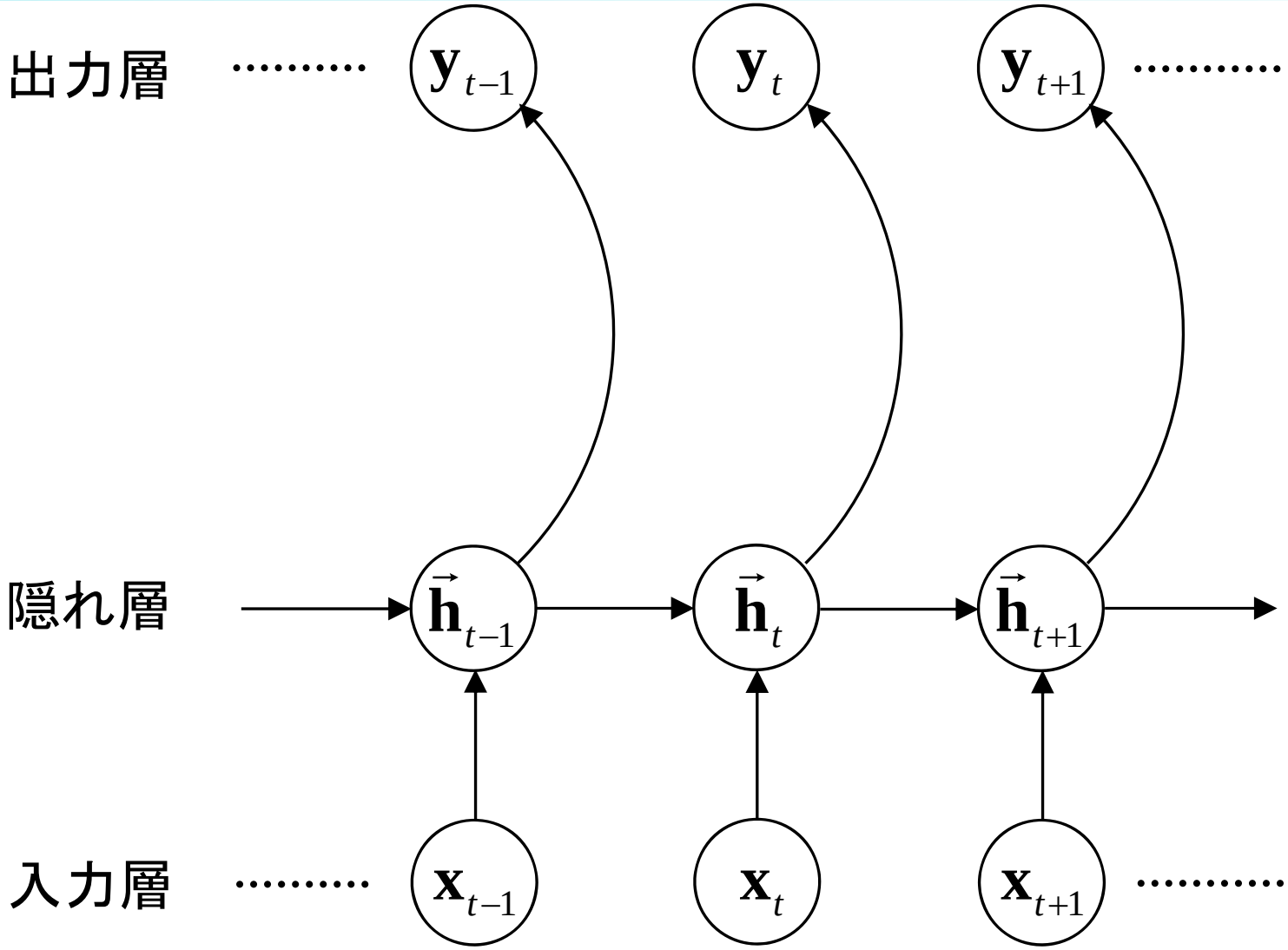
リカレントニューラルネットワーク

リカレントニューラルネットワーク(RNN)
・時系列解析
・言語モデル



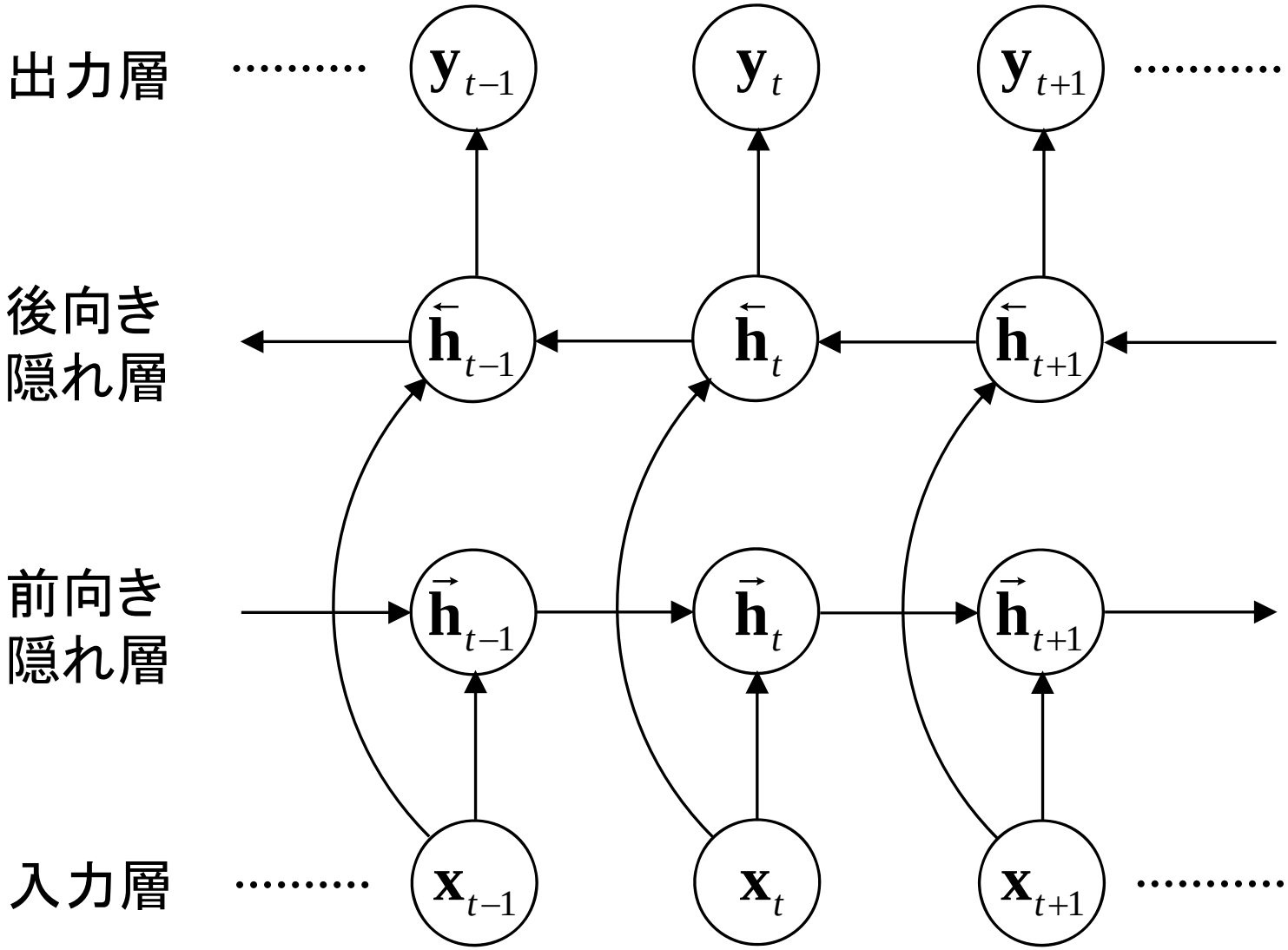
中間層の活性ベクトルを
入力層へフィードバックすることで
中間層は過去の全履歴を表現

Forward（前向き）RNN



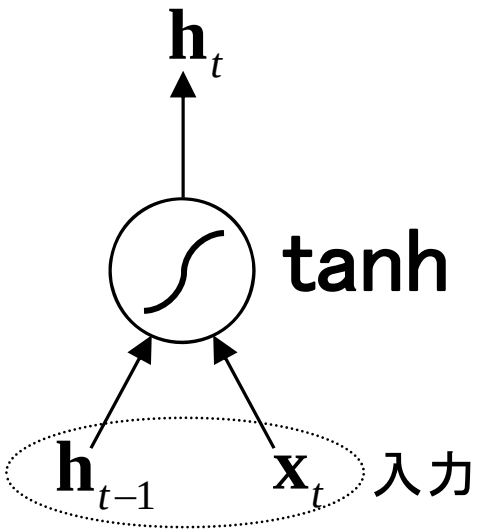
言語モデルに利用、長い文脈に有効 [Mikorov 2012]

Bidirectional RNN (B-RNN) [Shuster 1997]

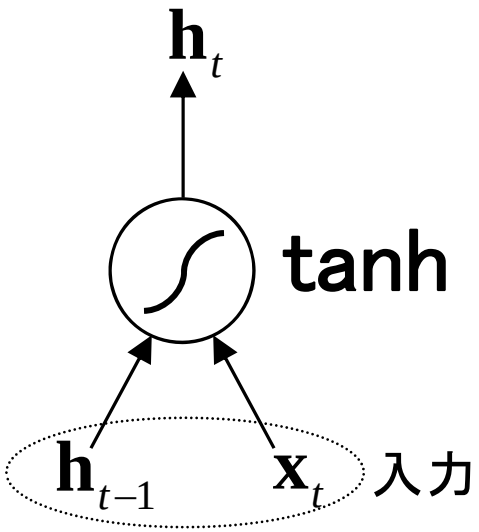


Denoising Autoencoder 、マスク推定に効果的

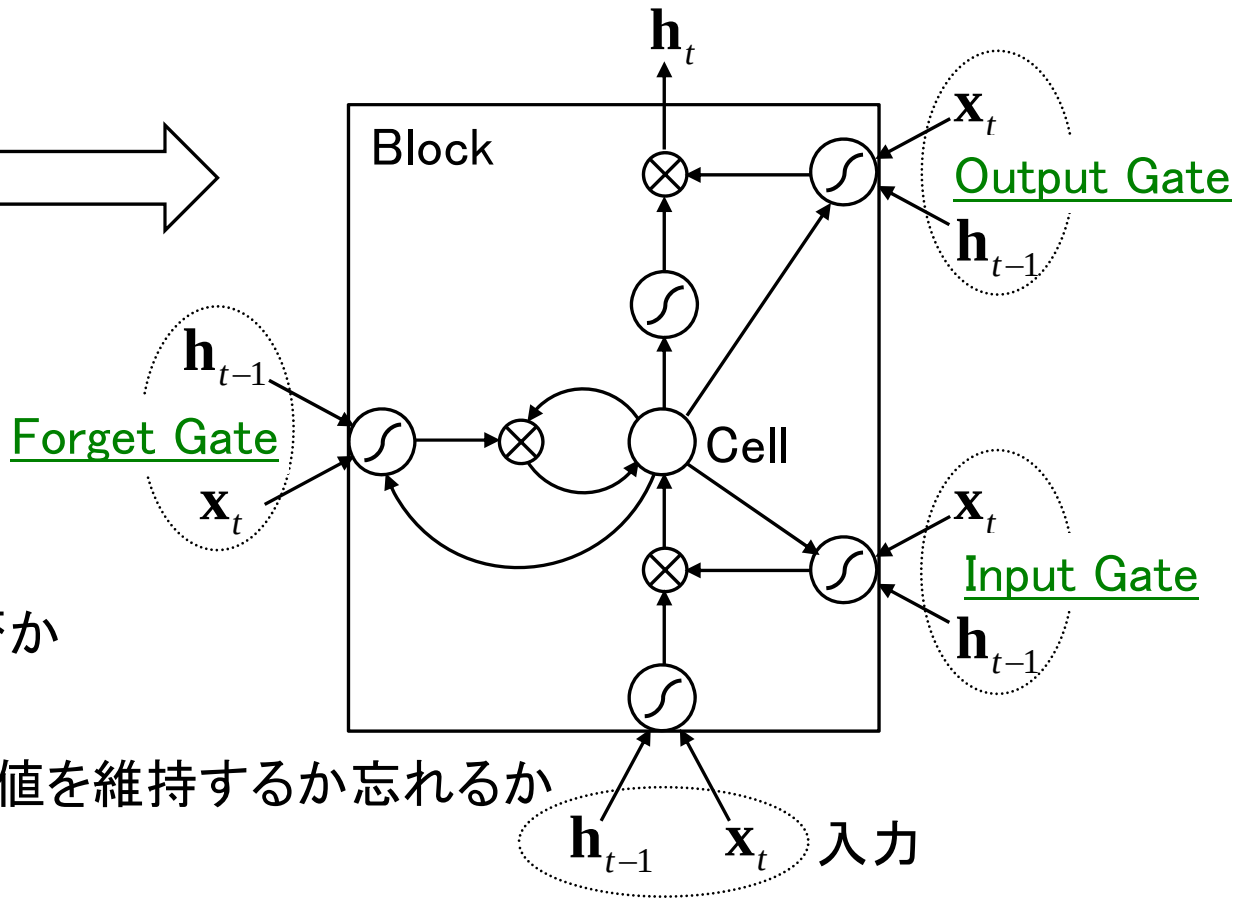
Long Short Term Memory [Hochreiter 1997]



Long Short Term Memory [Hochreiter 1997]



Long Short Term Memory (LSTM) Block

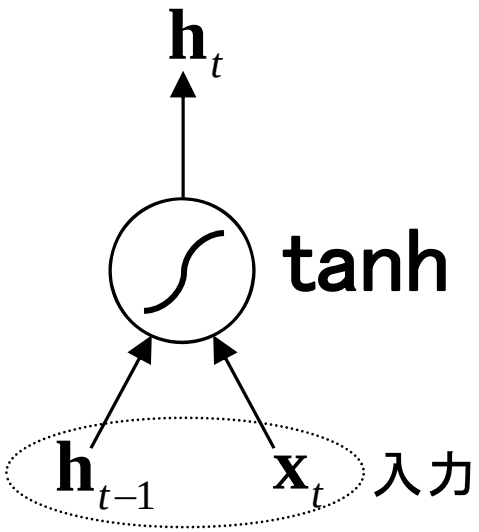


Input Gate:
入力をBlockに入れるか否か

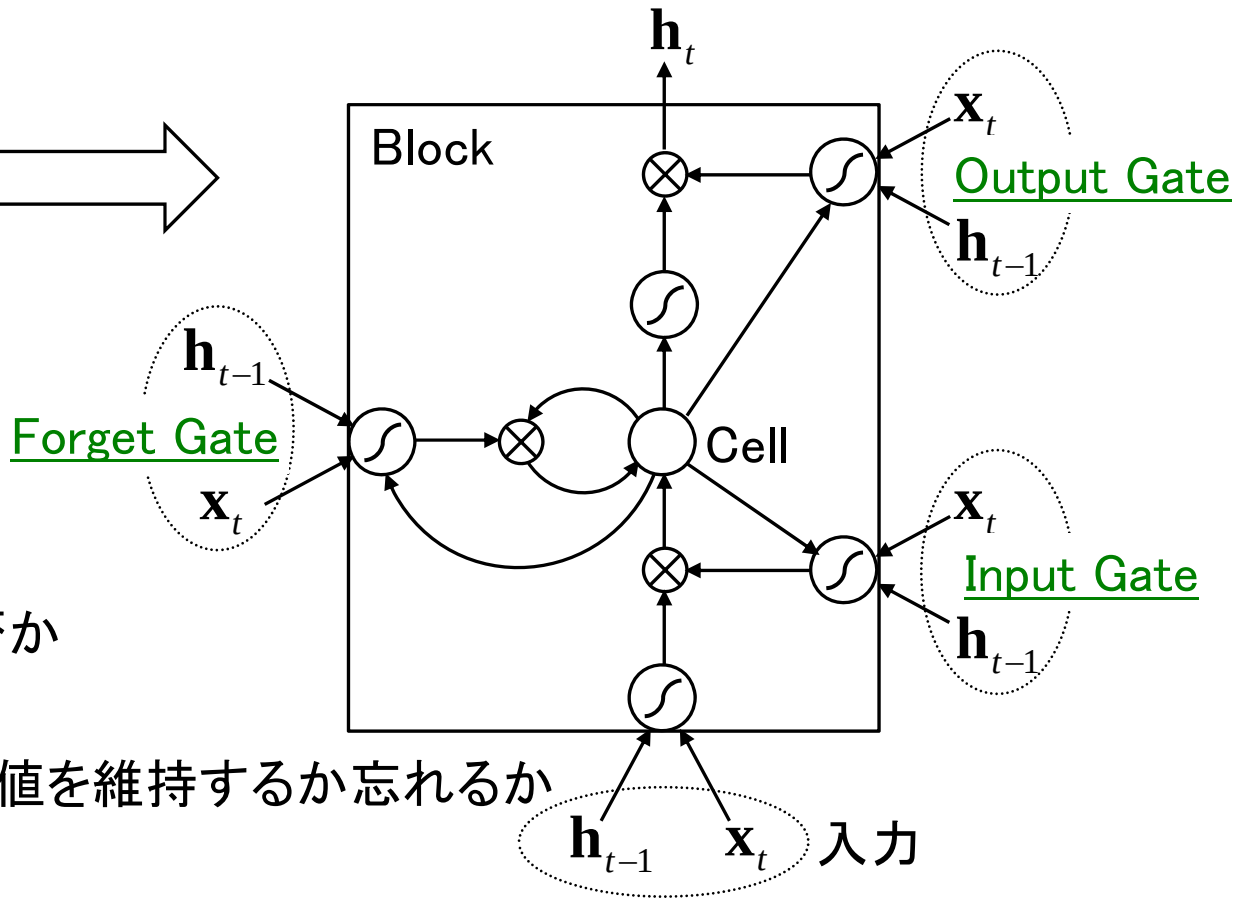
Forget Gate:
Blockで記憶している活性値を維持するか忘れるか

Output Gate:
Blockで記憶している活性値を出力するか否か

Long Short Term Memory [Hochreiter 1997]



Long Short Term Memory (LSTM) Block



Input Gate:
入力をBlockに入れるか否か

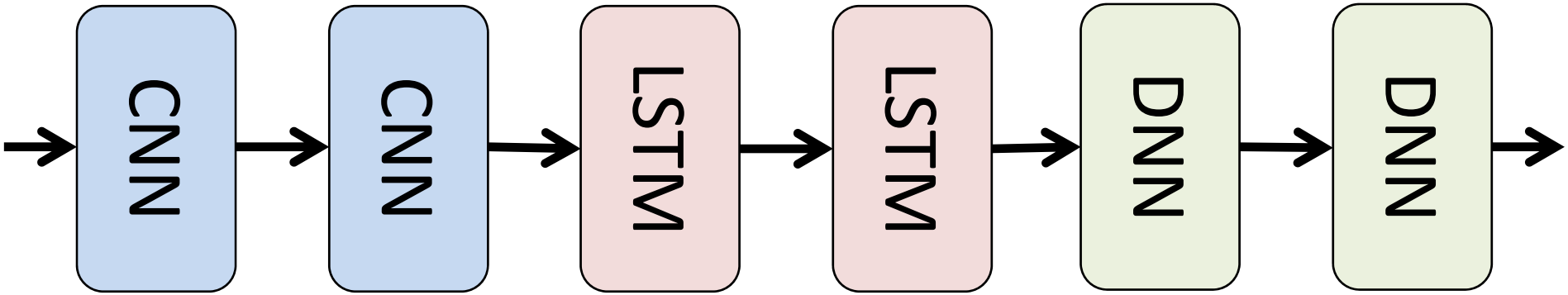
Forget Gate:
Blockで記憶している活性値を維持するか忘れるか

Output Gate:
Blockで記憶している活性値を出力するか否か

Bidirectional LSTM (B-LSTM) [Graves 2013]

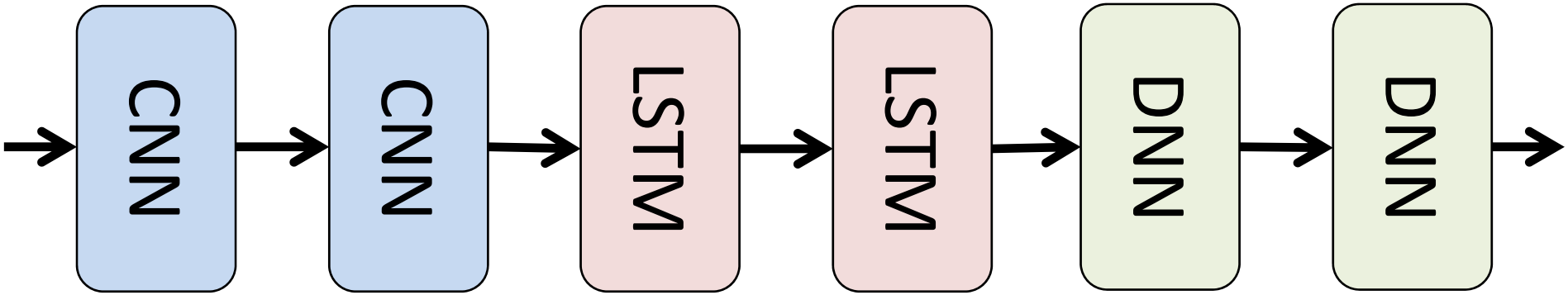
ネットワークの組み合わせ

CLDNN (CNN-LSTM-DNN) [Sainath 2015]



ネットワークの組み合わせ

CLDNN (CNN-LSTM-DNN) [Sainath 2015]



GoogleNet (22層CNN) [Szegedy 2014]

超大規模画像認識

C. Szegedy, et al., “Going deeper with convolutions,” arXiv:1409.4842v1, 2014より引用

概要

1. ディープニューラルネットワークの基礎
2. 音声研究への応用
3. 様々なネットワーク
4. DNNのツールキット、開発環境
5. まとめ

DNNのツールキット、開発環境

- Kaldi
  - 音声認識のデファクトスタンダード
  - 様々なレシピが公開
- CNTK
  - 複雑なネットワークをマクロで比較的に簡単に表現、学習可能
- RNNLib
  - RNN, B-RNN, LSTM, B-LSTMをサポート
  - CPUのみ
- Current
  - RNNLibのGPU版？
  - 基本的にLSTMをサポート

DNNのツールキット、開発環境

- Theano
  - Pythonベース、自動微分のサポート
  - 比較的簡単にDNNの学習機を作成可能
- Pylearn2
  - Python + Theano
- Chainer
  - Pythonベース、ReLUを標準サポート
  - Theanoよりとっつきやすい?
- cuDNN
  - NVIDIAが提供するCUDAベースのDNN開発キット
- Cafe, Torch, PyBrain...

概要

1. ディープニューラルネットワークの基礎
2. 音声研究への応用
3. 様々なネットワーク
4. DNNのツールキット、開発環境
5. まとめ

まとめ

- まだまだバブルは続く。。。と思う
- 何をやっているか何ができるかが少しずつ分かってきたような、そうでないような
- 既存ツールをうまく使ってディープにラーニング
  - 研究の個性をどこにだすかが難しくなったかも
- 頭が柔らかい若手のみなさんに期待