NTT

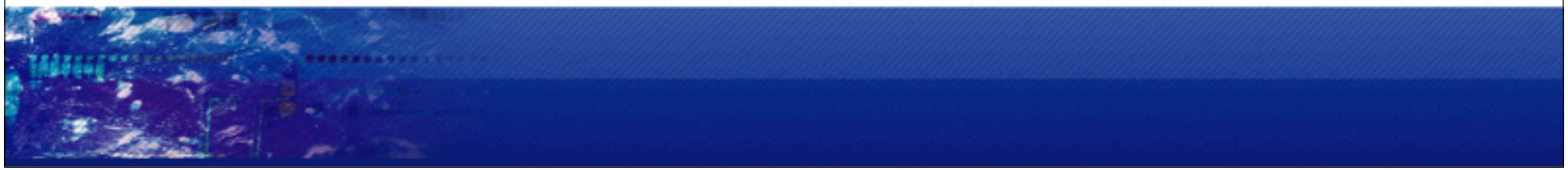
NTT Communication Science Laboratories

[IBIS 2013 企画セッション2：ディープラーニング]

音声認識分野における深層学習技術の研究動向

久保 陽太郎

NTT コミュニケーション科学基礎研究所

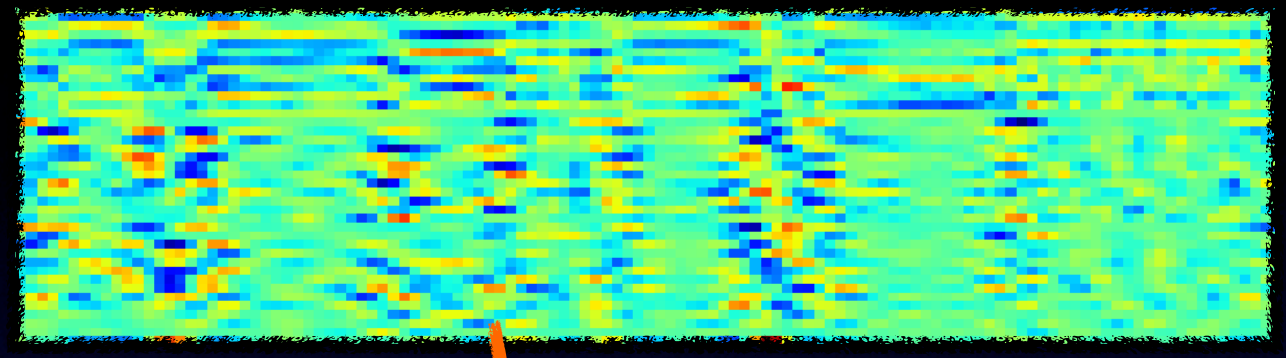


音声認識とは

NNと音声認識

音声認識分野でのDeep Learning

統計的音声認識



“あらゆる 現実 を 全て 自分 の ほう

$$\mathbf{X} \overset{\text{def}}{=} \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_t, \cdots | \boldsymbol{x}_t \in \mathbb{R}^D\}$$

入力

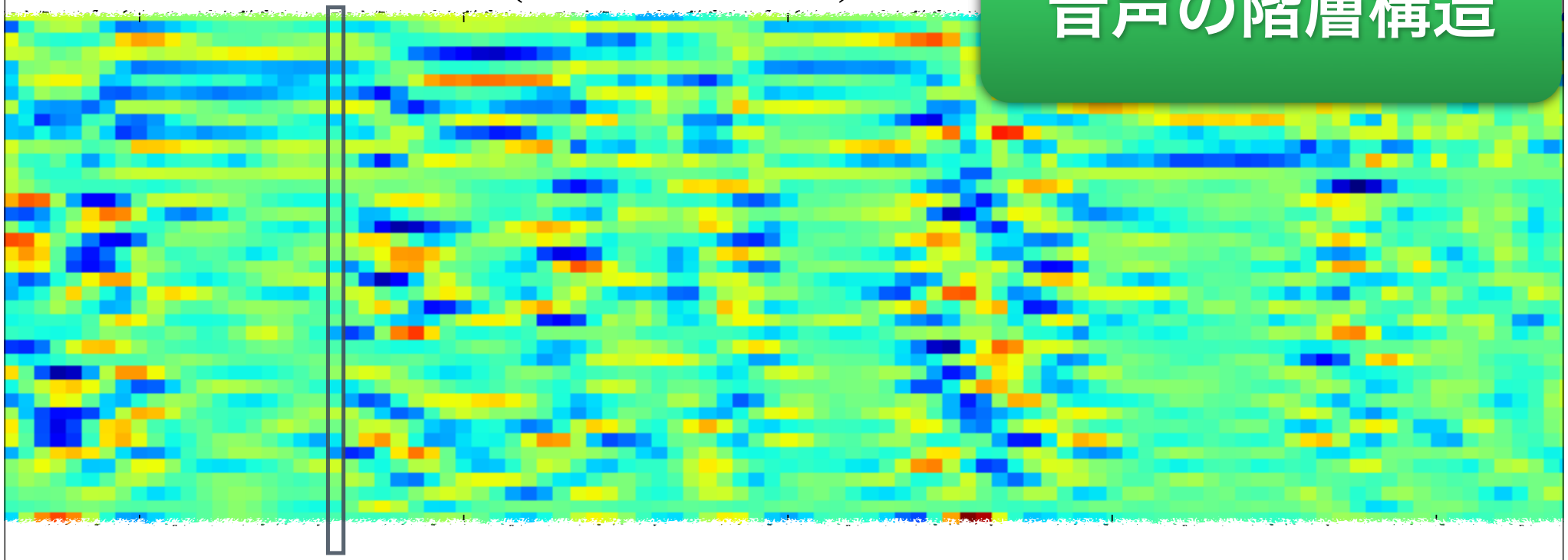
ℓ

ラベル

(単語)

$$\hat{\ell} \overset{\text{def}}{=} \text{argmax}_{\ell} P(\ell | \mathbf{X})$$

音声の階層構造



a+r-a1　a+r-a2　a+r-a3

sil　sil+a-r　a+r-a　r+a-y　y+u-r　r+u-g　g+e-

a+y-u　u+r-u

sil　a　r　a　y　u　r　u　g

無音　あらゆる　現実

音声の階層構造の利用

$$P(\ell|\mathbf{X}) = P(\mathbf{X},\ell)/P(\mathbf{X})$$

隠れ状態モデル　　言語モデル

$$P(\mathbf{X},\ell) = \sum_{q,m} P(\mathbf{X}|q)P(q|m)P(m|\ell)P(\ell)$$

音響モデル　　辞書モデル

$$\hat{\ell} \stackrel{\text{def}}{=} \text{argmax}_{\ell} P(\ell|\mathbf{X})$$

$$= \text{argmax}_{\ell} \sum_{q,m} P(\mathbf{X}|q)P(q|m)P(m|\ell)P(\ell)$$

$$\approx \text{argmax}_{\ell} \max_{q,m} P(\mathbf{X}|q)P(q|m)P(m|\ell)P(\ell)$$

音声の階層構造の利用

$$P(\mathbf{X}, \ell) = \sum_{q,m} P(\mathbf{X}|q) P(q|m) P(m|\ell) P(\ell)$$

隠れ状態モデル

隠れ状態系列が定まると
音素が定まるように設計される

隠れ状態の確率を理想的に推定できれば
同音異義語以外のエラーは起こらない

音声認識とは

NNと音声認識

音声認識分野でのDeep Learning

1989: Time delay neural networks

1995: Hybrid MLP-HMM

2000: MLP-HMM Tandem

2009: DNN-HMM

言語モデルを使った
大語彙音声認識のはじまり

話者適応技術
識別学習技術の発展

Deep Learningのはじまり

1989: Time delay neural networks

1995: Hybrid MLP-HMM

2000: MLP-HMM Tandem

2009: DNN-HMM

言語モデルを使った
大語彙音声認識のはじまり

話者適応技術
識別学習技術の発展

Deep Learningのはじまり

IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. 37, NO. 3, MARCH 1989

Phoneme Recognition Using Time-Delay Neural Networks

ALEXANDER WAIBEL, MEMBER, IEEE, TOSHIYUKI HANAZAWA, GEOFFREY HINTON,
KIYOHIRO SHIKANO, MEMBER, IEEE, AND KEVIN J. LANG

(現在ではHMMとNNの組み合わせで解消されている)

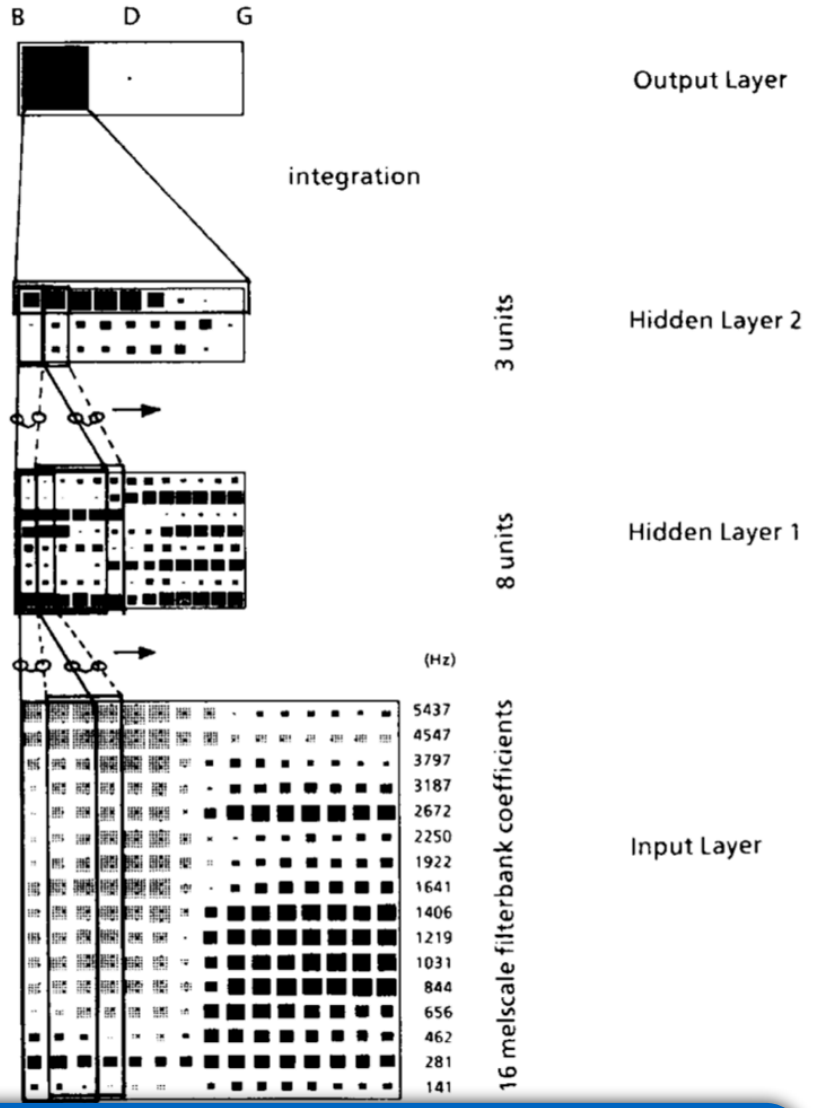
時間方向シフトをどう解消するか?

畳み込みNNの一種

- 入力: 音声分析結果のセグメント
- 出力: 音素識別結果

- CNNで2Dフィルタを使う代わりに，1D多変量フィルタを使う

- Pooling層は基本的になし
  最終層でSum-poolingをする



固定長セグメントの分類には効果的

系列ラベルへの拡張や言語モデルとの統合が難

1989: Time delay neural networks

1995: Hybrid MLP-HMM

2000: MLP-HMM Tandem

2009: DNN-HMM

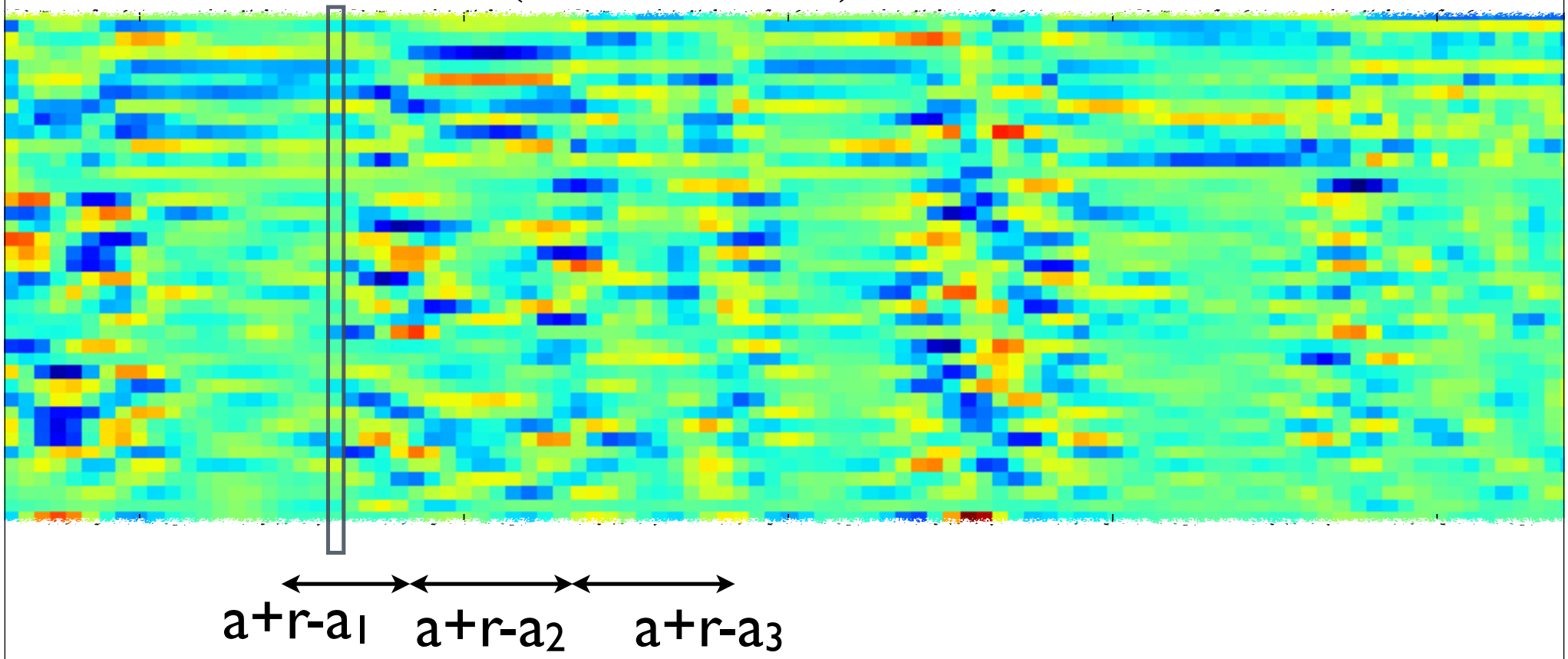
言語モデルを使った
大語彙音声認識のはじまり

話者適応技術
識別学習技術の発展

Deep Learningのはじまり

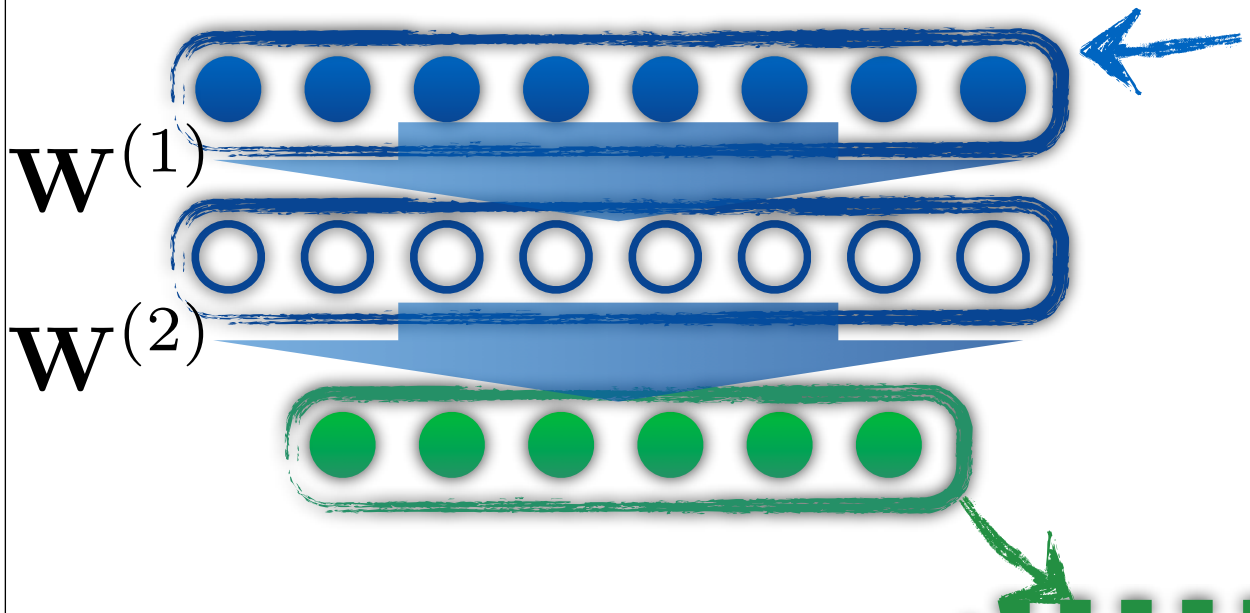
$$P(\mathbf{X}, \ell) = \sum_{\boldsymbol{q},\boldsymbol{m}} P(\mathbf{X}|\boldsymbol{q})P(\boldsymbol{q}|\boldsymbol{m})P(\boldsymbol{m}|\ell)P(\ell)$$

t=132 (1.32sec時点)



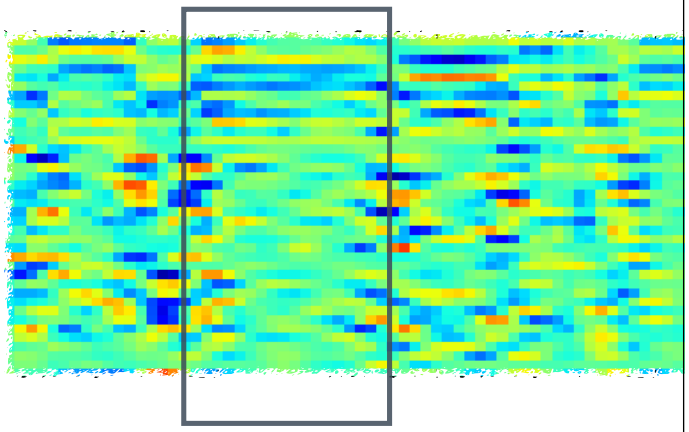
$$P(\boldsymbol{x}_t|q_t) = \sum_{k} \pi_{q_t,k}\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{q_t,k}, \mathbf{S}_{q_t,k})$$

NN/ HMM Hybrid approach



従来のMFCCより

リッチな情報を入力する



$$P(\boldsymbol{x}_t|q_t) = \frac{P(q_t|\boldsymbol{x}_t)}{P(q_t)} P(\boldsymbol{x}_t)$$

音響モデルのスコア

$P(\boldsymbol{x}_t)$ 定数

別の方法で求めておく

高次元かつ相互相関する特徴ベクトルを

扱える点でGMMより優位

最小Cross-entropy規準による学習

学習データ: $\mathcal{D} \stackrel{\text{def}}{=} \{(\mathbf{X}_1, \boldsymbol{\ell}_1), (\mathbf{X}_2, \boldsymbol{\ell}_2) \cdots\}$

音声特徴ベクトル列　単語列

$\mathbf{X}_n \stackrel{\text{def}}{=} \{\boldsymbol{x}_{n,1}, \boldsymbol{x}_{n,2}, \cdots\}$

推定を容易にするためGMM等を用いて最尤HMM状態を推定

$$\boldsymbol{q}_n \stackrel{\text{def}}{=} \arg\max_{\boldsymbol{q}} P(\boldsymbol{q}|\mathbf{X}_n, \boldsymbol{\ell}_n)$$

$$\text{maximize} \sum_n (\log P(q_{n,t}|\boldsymbol{x}_{n,t}, \Theta))$$

HMM状態と単語列は多対一対応なので正解単語列に

対応するHMM状態を復元できれば正しい単語列が推定できる

1989: Time delay neural networks

1995: Hybrid MLP-HMM

2000: MLP-HMM Tandem

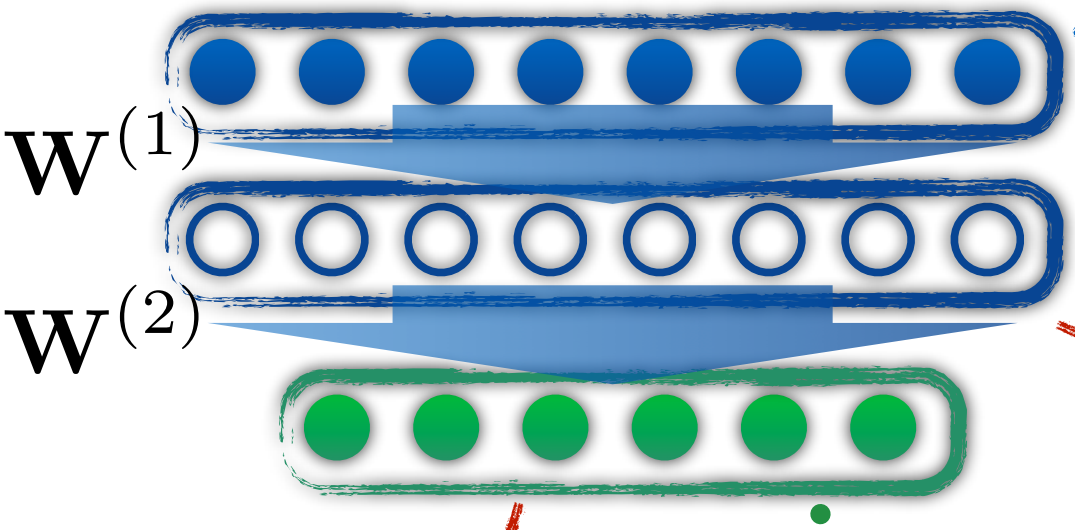
2009: DNN-HMM

言語モデルを使った
大語彙音声認識のはじまり

話者適応技術
識別学習技術の発展

Deep Learningのはじまり

Tandem approach



従来のMFCCより
リッチな情報を入力する
(たとえば前後5フレームの
MFCCも追加で入力)

GMM/HMMの
観測データとして使う

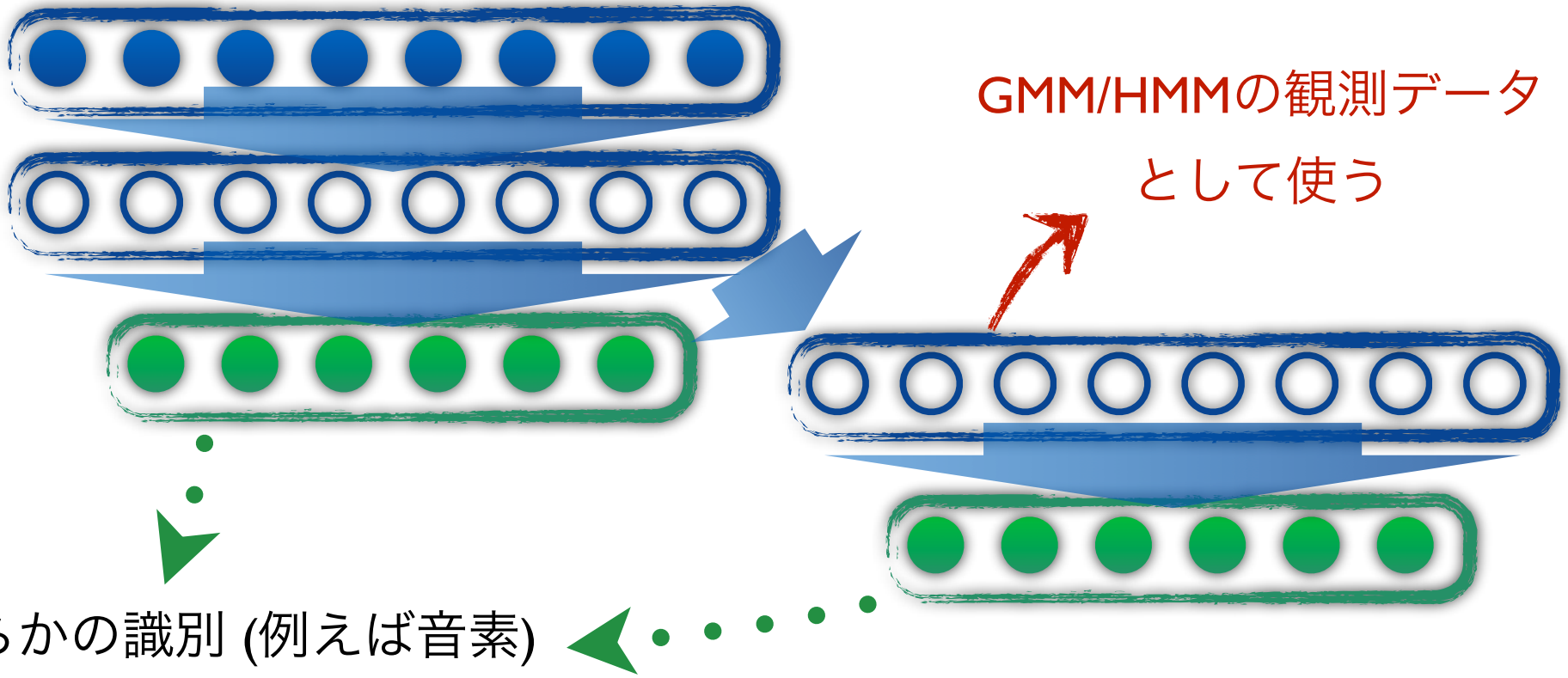
普通のGMM/HMMの
観測データとして使う

なんらかの識別
(例えば音素識別)

$$P(\boldsymbol{x}_t|q_t) = \sum_k \pi_{q_t,k} \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{q_t,k}, \mathbf{S}_{q_t,k})$$

GMM/ HMMで展開されてきた様々な技術
(識別学習／話者適応) がそのまま利用可能

Hierarchical tandem processing



困難であった深いNNの学習を
なんとか回避して用いていたのでは?

1989: Time delay neural networks

1995: Hybrid MLP-HMM

2000: MLP-HMM Tandem

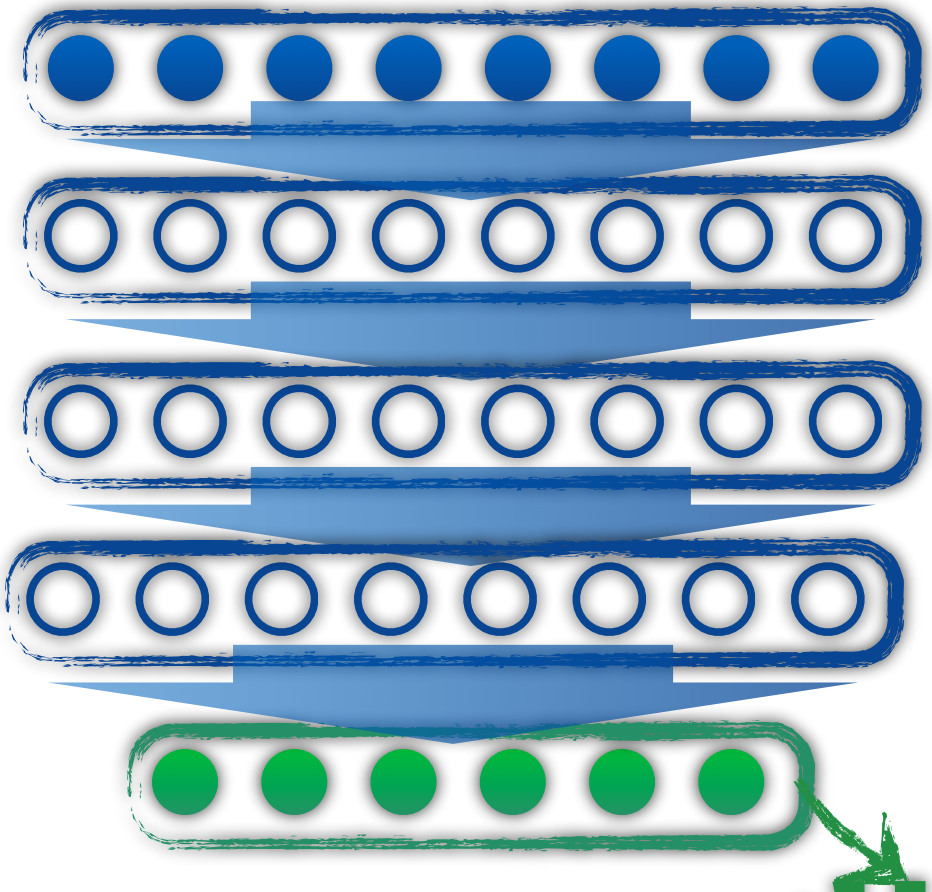
2009: DNN-HMM

言語モデルを使った
大語彙音声認識のはじまり

話者適応技術
識別学習技術の発展

Deep Learningのはじまり

DNN-HMM



Hybridアプローチと
全く同じ!!
(単に深いNNを使う
ようになっただけ)

$$P(\boldsymbol{x}_t|q_t) = \frac{P(q_t|\boldsymbol{x}_t)}{P(q_t)} P(\boldsymbol{x}_t)$$

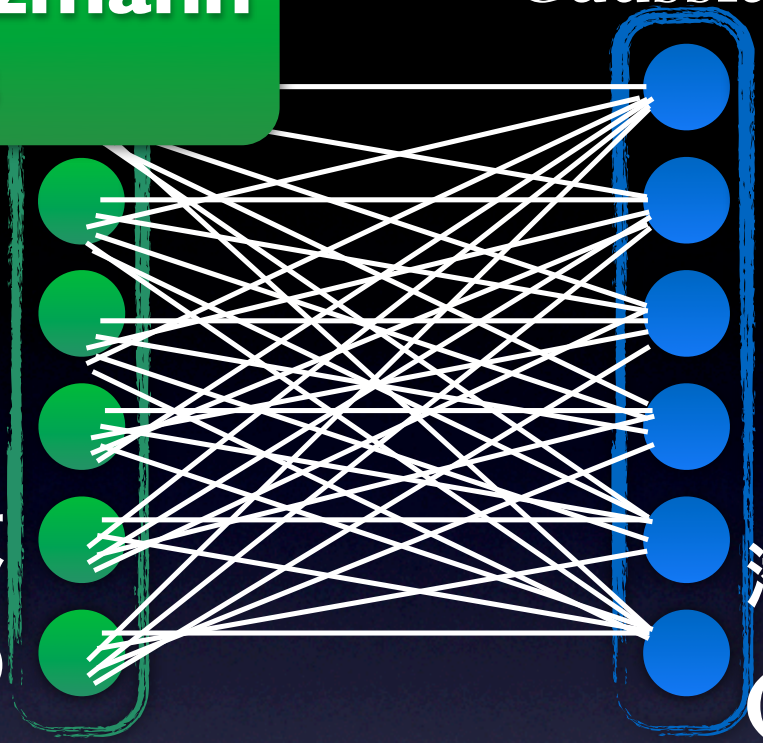
音響モデルのスコア

定数

別の方法で求めておく

Restricted Boltzmann Machines

Gaussian-Bernoulli RBM



観測変数
(実数ベクトル)

潜在変数
(バイナリベクトル)

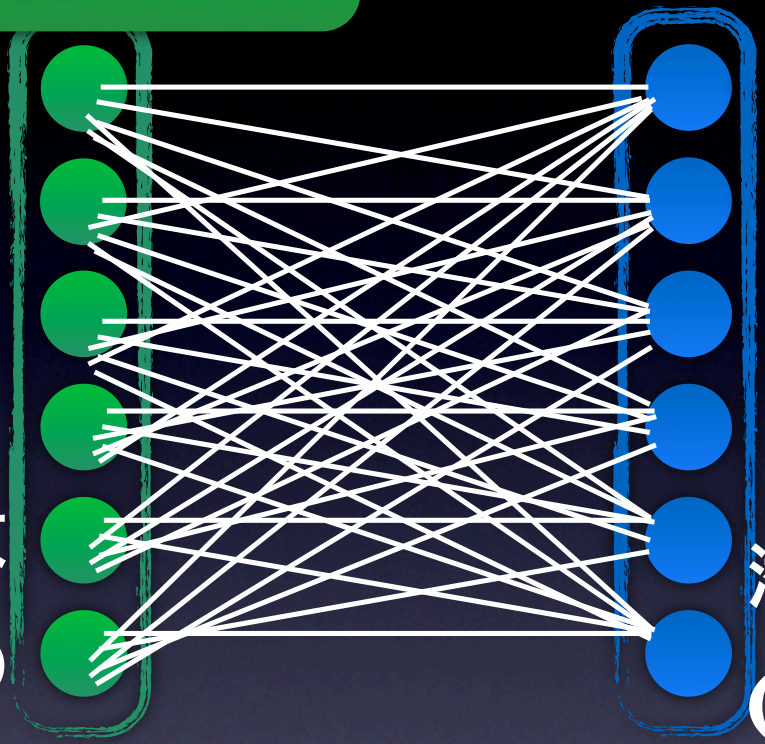
$$P(\boldsymbol{x}, \boldsymbol{h}|\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}) \propto \exp\left\{-\boldsymbol{h}^T\mathbf{W}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{h} - \frac{1}{2\sigma^2}||\boldsymbol{x}-\boldsymbol{c}||_2^2\right\}$$

$$P(x_j|\boldsymbol{h}) = \mathcal{N}\left(x_i; \sum_i w_{i,j}h_i + c_j, \sigma^2\right) \qquad P(h_i = 1|\boldsymbol{x}) = f\left(\sum_j w_{i,j}x_j + b_j\right)$$

潜在変数で条件付けた観測変数の分布が正規分布

Restricted Boltzmann Machines

Bernoulli-Bernoulli RBM

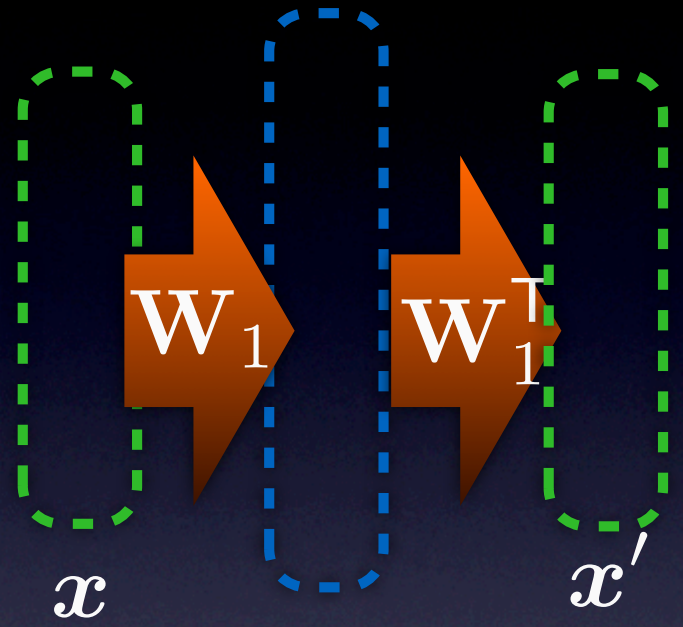


観測変数
(バイナリベクトル)

潜在変数
(バイナリベクトル)

$$P(\boldsymbol{x}, \boldsymbol{h}|\mathbf{W}, \boldsymbol{b}, \boldsymbol{c}) \propto \exp\left\{-\boldsymbol{h}^T \mathbf{W} \boldsymbol{x} - \boldsymbol{b}^T \boldsymbol{h} - \boldsymbol{c}^T \boldsymbol{x}\right\}$$

$$P(x_j = 1|\boldsymbol{h}) = f\left(\sum_i w_{i,j} h_i + c_i\right) \qquad P(h_i = 1|\boldsymbol{x}) = f\left(\sum_j w_{i,j} x_j + b_j\right)$$



Gaussian-Bernoulli RBMをContrastive Divergence学習

Restricted Boltzmann Machines

最尤推定

SGDによる最適化を考える

$$\nabla_W f_t = \frac{1}{\sum_{\boldsymbol{h}_t} P(\boldsymbol{x}_t, \boldsymbol{h}_t | \mathbf{W}, \boldsymbol{b}, \boldsymbol{c})} \sum_{\boldsymbol{h}_t} \left(\nabla_W \left[\frac{1}{Z(\mathbf{W}, \boldsymbol{b}, \boldsymbol{c})} \exp\{-E(\boldsymbol{x}_t, \boldsymbol{h}_t; \mathbf{W}, \boldsymbol{b}, \boldsymbol{c})\} \right] \right)$$

$$= \langle -\nabla_{\mathbf{W}} E(\boldsymbol{x}_t, \boldsymbol{h}') \rangle_{P(\boldsymbol{h}' | \boldsymbol{x}_t, \mathbf{W}, \boldsymbol{b}, \boldsymbol{c})} - \langle -\nabla_{\mathbf{W}} E(\boldsymbol{x}', \boldsymbol{h}') \rangle_{P(\boldsymbol{x}', \boldsymbol{h}' | \mathbf{W}, \boldsymbol{b}, \boldsymbol{c})}$$

全ての可能なバイナリベクタの例に関する

期待値→計算不可能！！！

Restricted Boltzmann Machines

最尤推定

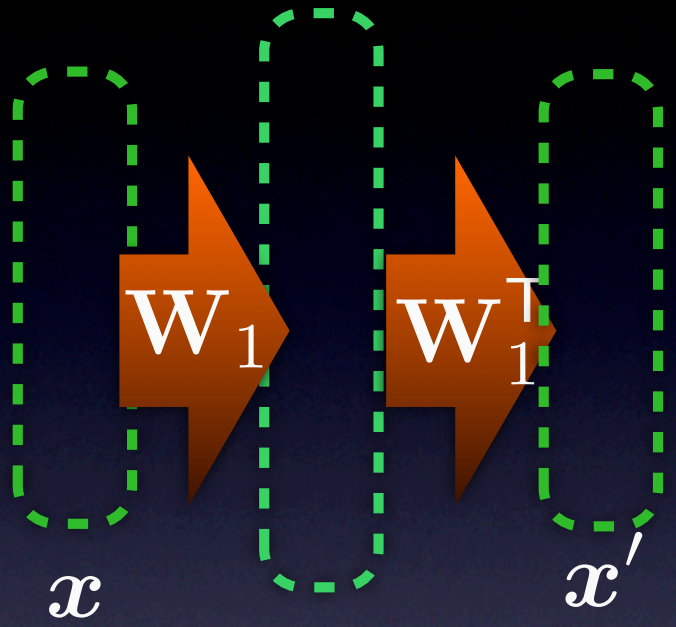
SGDによる最適化を考える

$$\nabla_W f_t = \frac{1}{\sum_{\boldsymbol{h}_t} P(\boldsymbol{x}_t, \boldsymbol{h}_t | \mathbf{W}, \boldsymbol{b}, \boldsymbol{c})} \sum_{\boldsymbol{h}_t} \left(\nabla_W \left[\frac{1}{Z(\mathbf{W}, \boldsymbol{b}, \boldsymbol{c})} \exp\{-E(\boldsymbol{x}_t, \boldsymbol{h}_t; \mathbf{W}, \boldsymbol{b}, \boldsymbol{c})\} \right] \right)$$

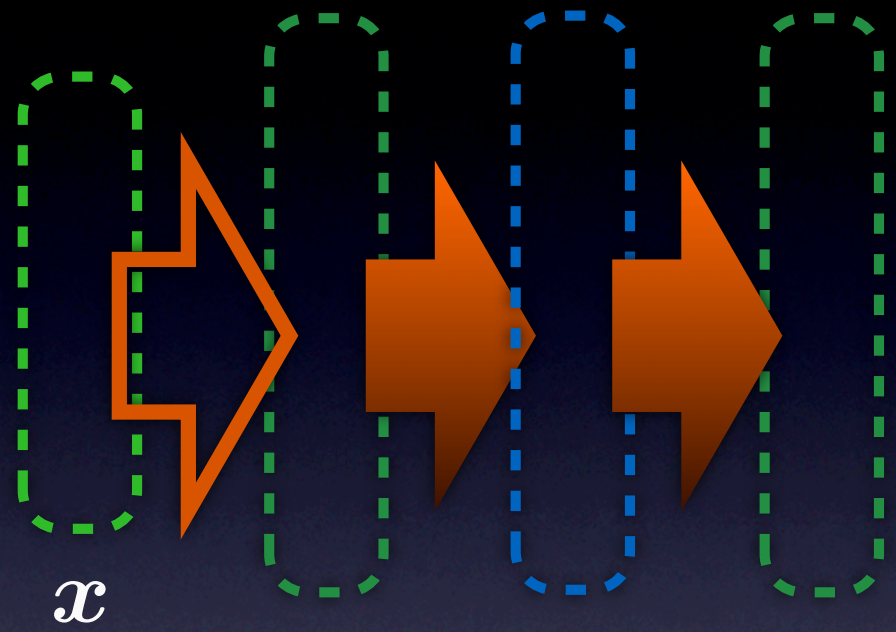
$$= \langle -\nabla_{\mathbf{W}} E(\boldsymbol{x}_t, \boldsymbol{h}') \rangle_{P(\boldsymbol{h}' | \boldsymbol{x}_t, \mathbf{W}, \boldsymbol{b}, \boldsymbol{c})} - \langle -\nabla_{\mathbf{W}} E(\boldsymbol{x}', \boldsymbol{h}') \rangle_{P(\boldsymbol{x}', \boldsymbol{h}' | \mathbf{W}, \boldsymbol{b}, \boldsymbol{c})}$$

サンプリングで期待値を近似することを考える

→ Contrastive divergence

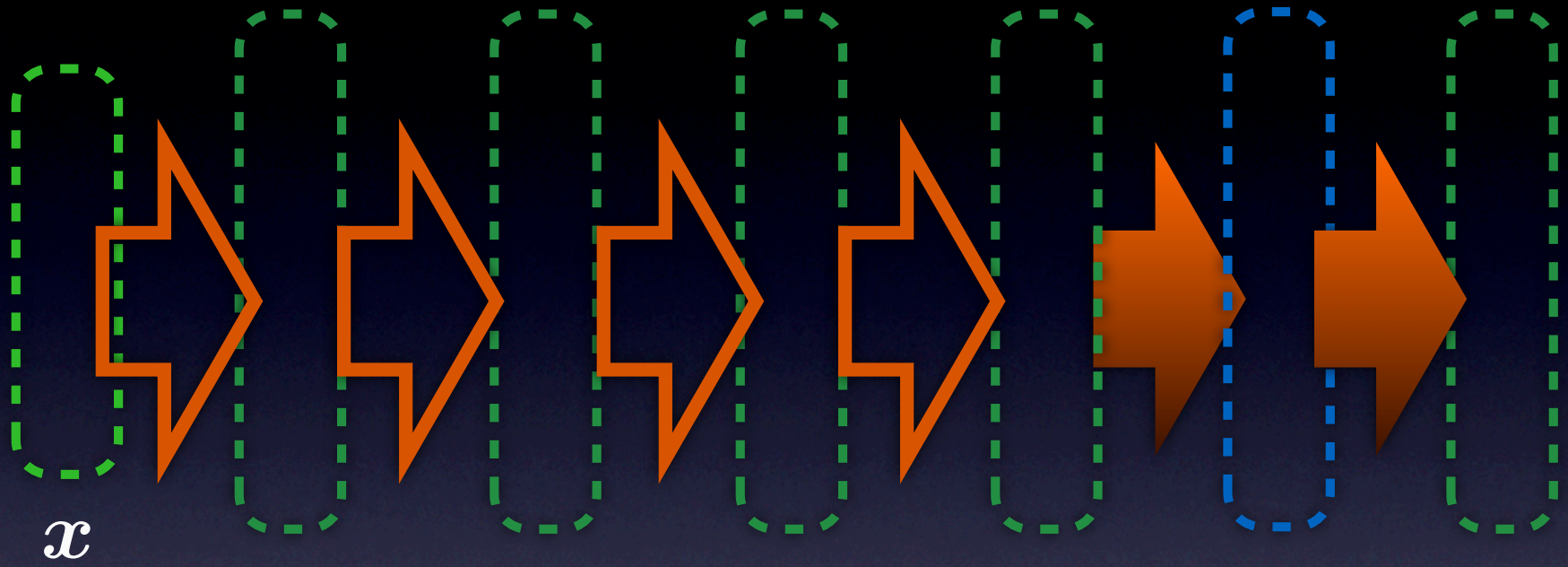


学習されたGaussian-Bernoulli RBMを用いて，学習データ内の各サンプルに対応する潜在変数をサンプリング



サンプルされた潜在変数を観測変数とする

Bernoulli-Bernoulli RBMを学習



所望の層数まで繰り返す

得られたRBMのパラメタをNNの初期値として用いる

最後の潜在変数とラベルの間の重みは乱数で初期化

音声認識のためのレシピ

- 入力：

  (MFCC or LogMelFB) 11～17-frames (+ Δ + ΔΔ)

- 隠れユニット数：2048

- 層数：L=5 for MFCC, L=8 for LogMelFB

  (開発データセットで決める)

- 近年ではRBMの代わりに

  Denoising Auto-encoderが使われることも

使用データ	研究機関	データ量 (時間)	従来法 エラー率	DNN エラー率	エラー 削減率
CSJ 日本語講義	京大	250h	20.0	17.5	12.5
MIT-OCW 英語講義	NTT	104h	28.2	22.5	20.2
SWB1 英語電話	Microsoft	309h	27.4	18.5	32.5
YouTube 動画音声	Google	1400h	52.3	47.6	9.0
Broadcast 動画音声	IBM	50h	18.8	17.5	6.9

音声認識とは

NNと音声認識

音声認識分野でのDeep Learning

時系列の識別を意識したモデル／学習

音声の多様性を考慮したモデル構築

時系列の識別を意識したモデル／学習

音声の多様性を考慮したモデル構築

Sequential Discriminative Criterion

$$\text{maximize} \sum_n (\log P(q_{n,t}|\boldsymbol{x}_{n,t}, \Theta))$$

Frame-levelの識別だと音声認識のような時系列問題には不適

String-level MMI

単語列　特徴ベクトル列

$$\text{maximize} \sum_n \log P(\boldsymbol{\ell}_n|\mathbf{X}_n)$$

MPE

$$\text{minimize} \sum_m \sum_{\boldsymbol{\ell}'} P(\boldsymbol{\ell}'|\mathbf{X}_n)\mathcal{L}(\boldsymbol{\ell}_n;\boldsymbol{\ell}')$$

String-level MMI

$$\text{maximize} \sum_n \log P(\boldsymbol{\ell}_n | \mathbf{X}_n)$$

フレーム毎のHMM状態の推定エラーじゃなく,
単語列全体の一致を目指す学習が可能

MPE

$$\text{minimize} \sum_m \sum_{\boldsymbol{\ell}'} P(\boldsymbol{\ell}' | \mathbf{X}_n) \mathcal{L}(\boldsymbol{\ell}_n; \boldsymbol{\ell}')$$

単語列全体の一致を目指すだけではなく,
単語列間のエラー尺度 (例えば単語エラー数) を考え,
それを最小化するように学習可能

String-level MMI

単語列　特徴ベクトル列

$$\text{maximize} \sum_{n} \log P(\boldsymbol{\ell}_n | \mathbf{X}_n)$$

MPE

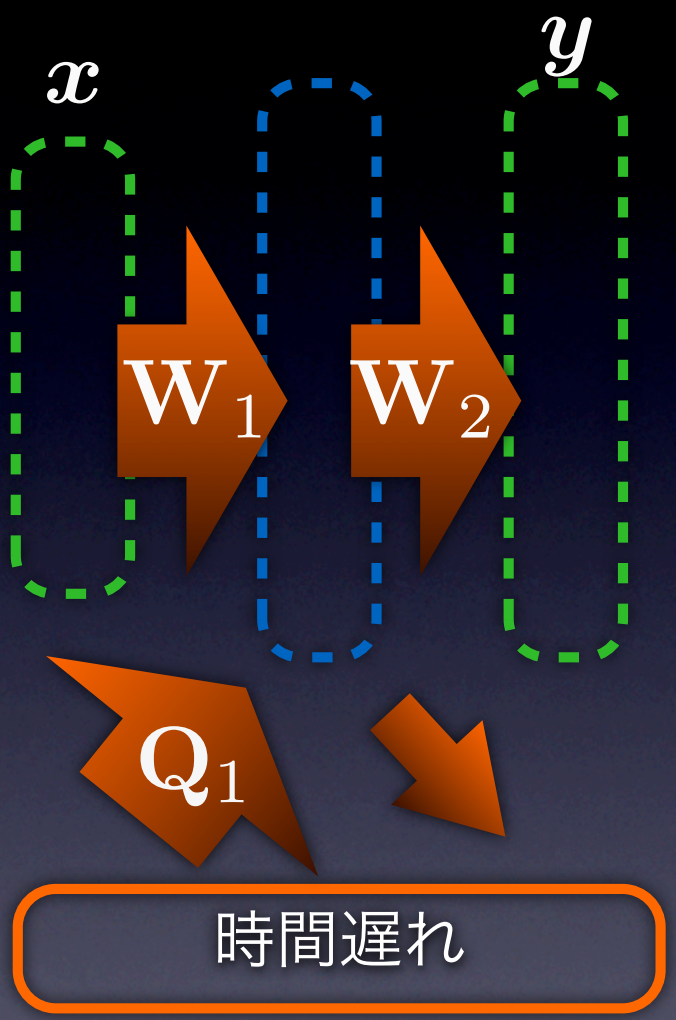
$$\text{minimize} \sum_{m} \sum_{\boldsymbol{\ell}'} P(\boldsymbol{\ell}' | \mathbf{X}_n) \mathcal{L}(\boldsymbol{\ell}_n; \boldsymbol{\ell}')$$

多くの場合，最適化の実現方法に難あり

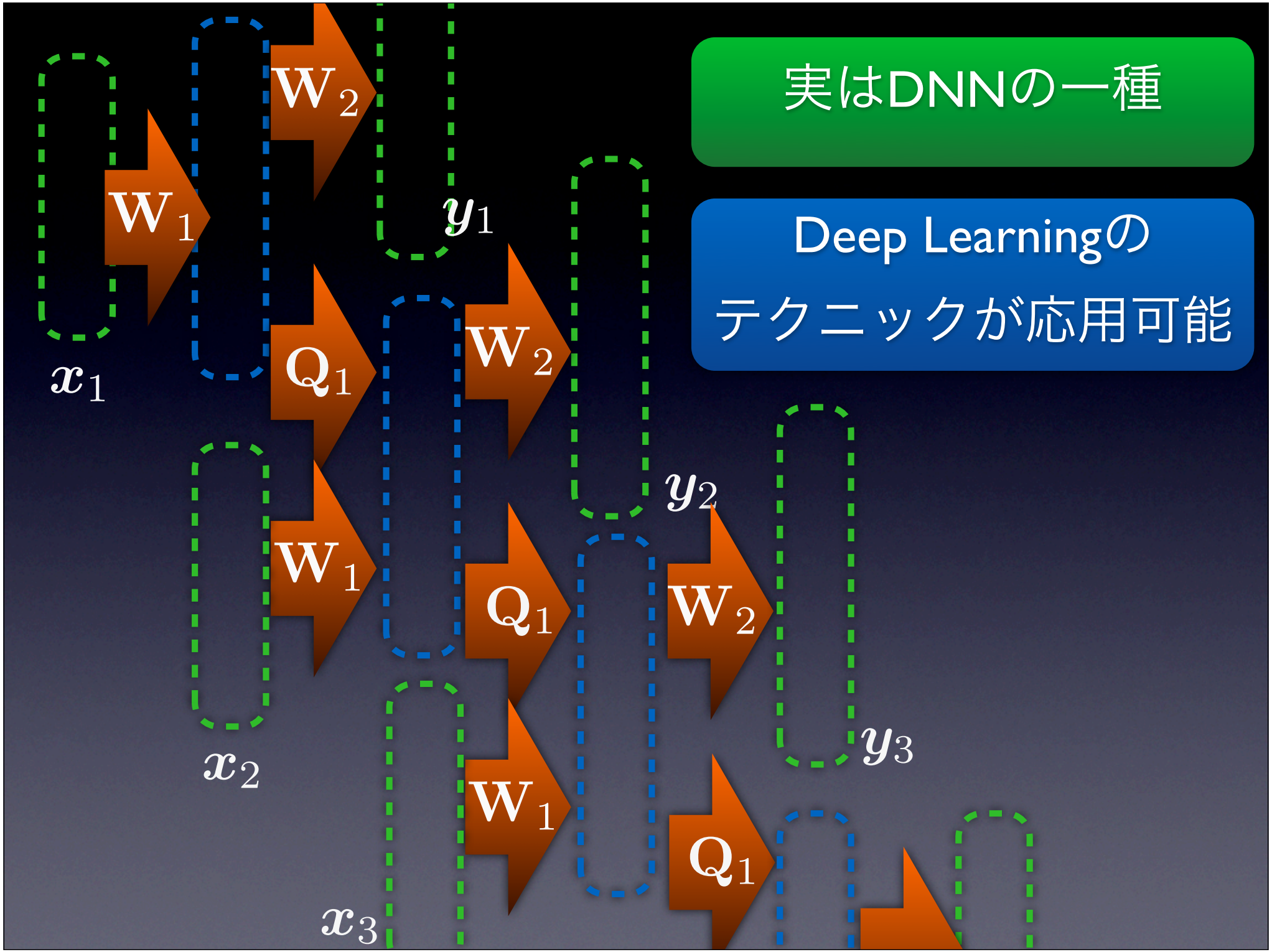
例えば総和($\sum_{\boldsymbol{\ell}'}$)の計算等...

Newton-CG法の援用や，GPUを活用したさらなる高速化などによって利用されはじめてきている

Recurrent Neural Nets

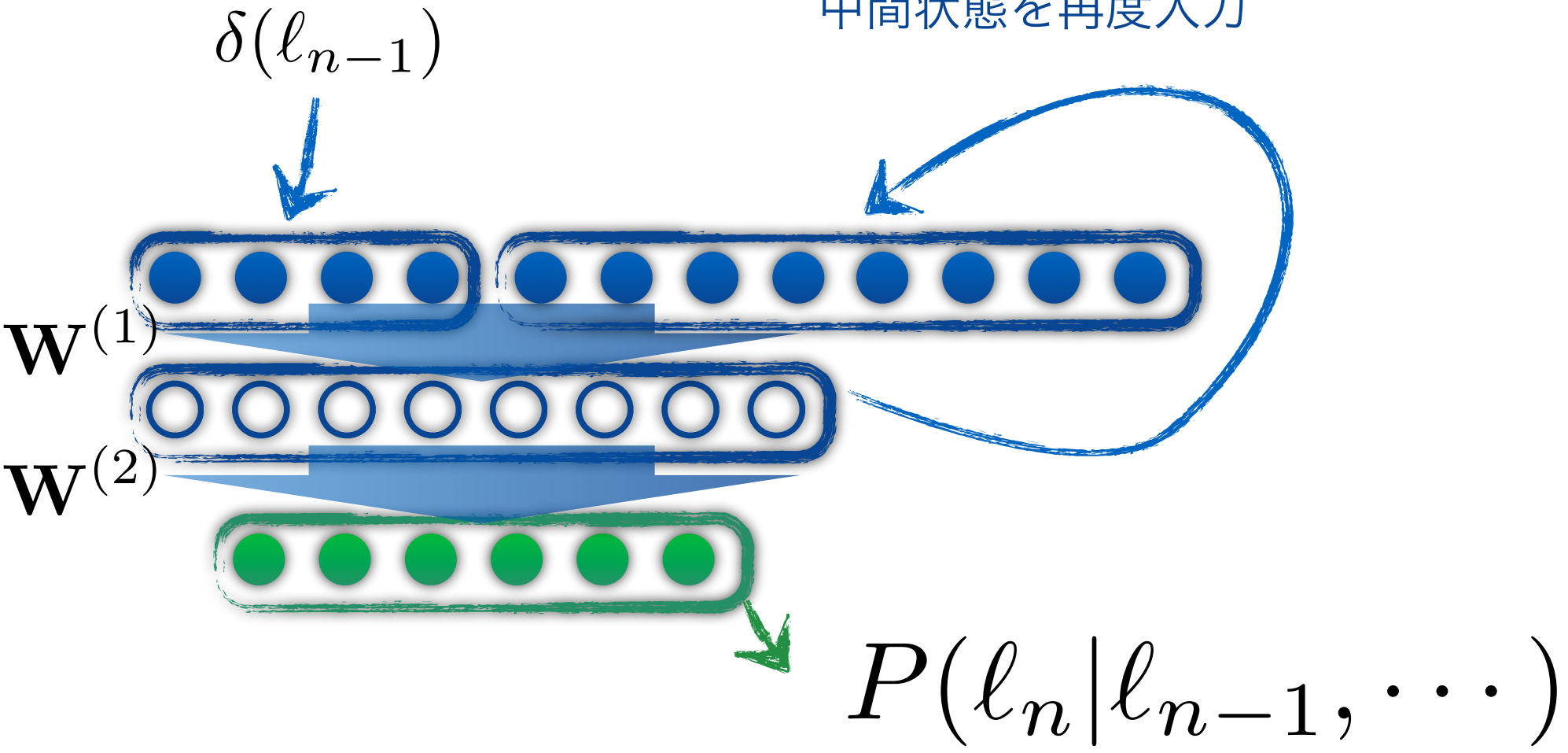


時系列データを扱うNN


実はDNNの一種
Deep Learningの
テクニックが応用可能
$\mathbf{W}_1$
$\mathbf{W}_2$
$\mathbf{Q}_1$
x_1
x_2
x_3
y_1
y_2
y_3

Recurrent NN言語モデル

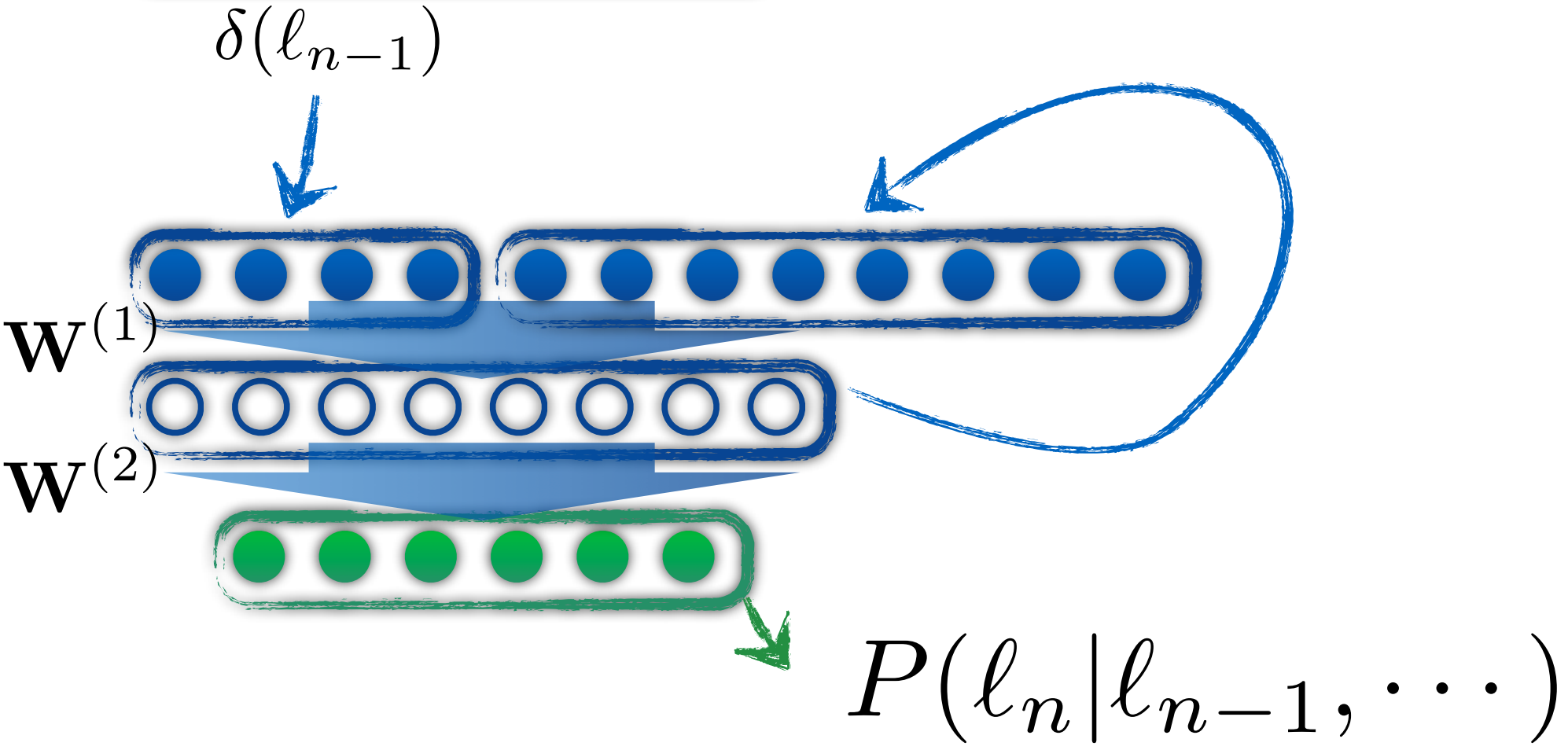
文脈を入力する代わりに、
直前の単語を処理した時の
中間状態を再度入力



Recurrent NN言語モデル

明示的に文脈長を
決める必要がない

確率値の計算が
非常に高コストな上,
学習も高コスト



Model	Dev WER[%]	Eval WER[%]
Baseline - KN5	12.2	17.2
Discriminative LM [14]	11.5	16.9
Joint LM [7]	-	16.7
Static RNN	10.5	14.9
Static RNN + KN	10.2	14.6
Adapted RNN	9.8	14.5
Adapted RNN + KN	9.8	14.5
All RNN	**9.7**	**14.4**

[Tomas 2011]

Model	Dev WER[%]	Eval WER[%]
Baseline - KN5	12.2	17.2
Discriminative LM [14]	11.5	16.9
Joint LM [7]	-	16.7
Static RNN	10.5	14.9
Static RNN + KN	10.2	14.6
Adapted RNN	9.8	14.5
Adapted RNN + KN	9.8	14.5
All RNN	**9.7**	**14.4**

[Tomas 2011]

Model	Dev WER[%]	Eval WER[%]
Baseline - KN5	12.2	17.2
Discriminative LM [14]	11.5	16.9
Joint LM [7]	-	16.7
Static RNN	10.5	14.9
Static RNN + KN	10.2	14.6

文と文の間にまたがる共起関係を
キャッシュモデルより高精度に学習できる

N-gramとの線形補完によって，より高度な言語モデル
を構築できる

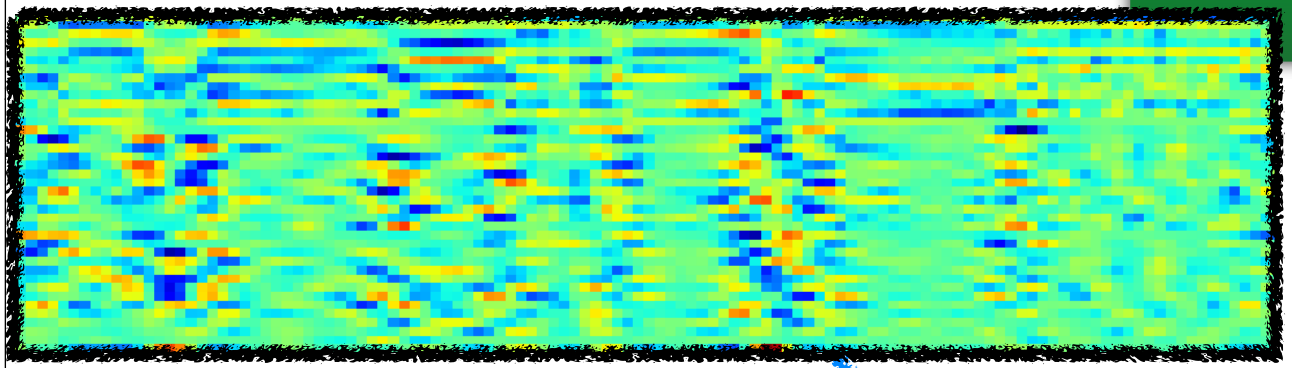
トピックモデル + *N*-gramとの差分は不明

Connectionist Temporal Classification

HMMによる時系列表現ではなく積極的にRNNを活用して音声認識を実現できないか?

空文字 ϕ を挿入して系列長の違いを吸収

音声特徴量系列



音素列

{a, ϕ, ϕ, ϕ, r, ϕ, ϕ, y, ϕ, ϕ, ϕ ... }

Connectionist Temporal Classification

学習時：

空文字 ϕ が何処に挿入されるかについて
全てを考慮して勾配を計算する必要がある

HMMの前向き後ろ向きアルゴリズムと
同種のアルゴリズムで実現可能

認識時：

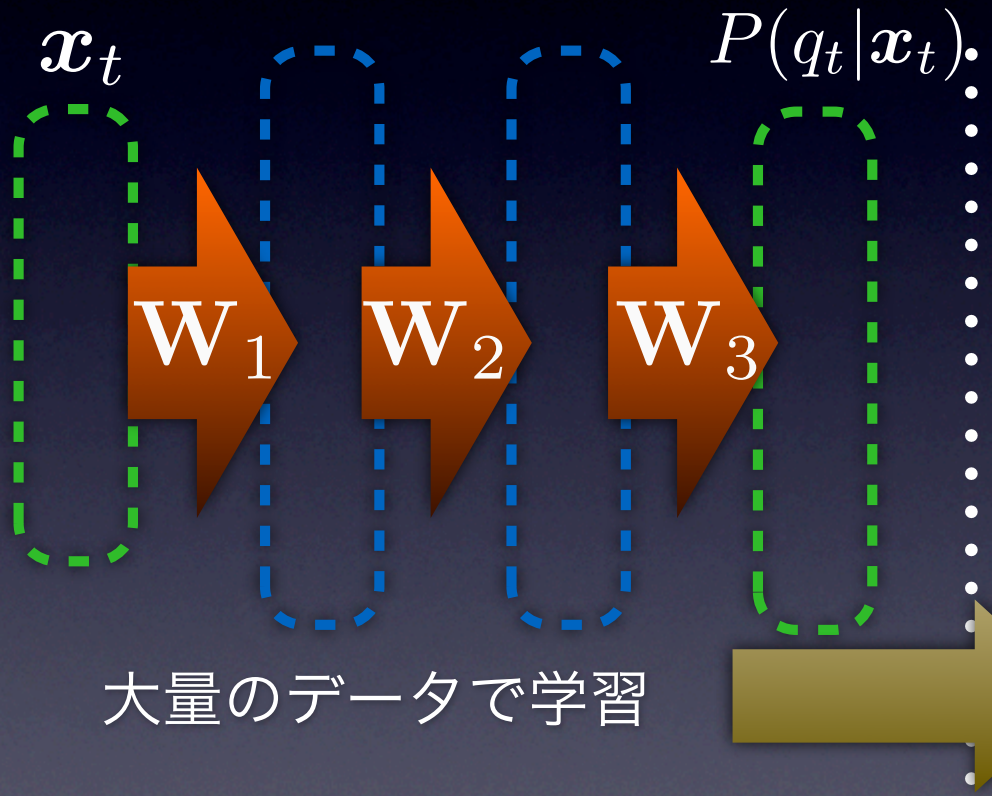
全ての 空文字 ϕ パターンについて周辺化した予測を
得る必要がある (Max近似はあまり上手く動かない)

時系列の識別を意識したモデル／学習

音声の多様性を考慮したモデル構築

Adaptation: param. transformation

GMM/ HMMの時代よりパラメタを特定話者データに基づく
推定量で変換し適応するという試みが見られた



$$\hat{\mathbf{W}}_\ell = \underset{\mathbf{W}'}{\mathrm{argmax}} \left(F(\Theta) \,|_{\mathbf{W}_\ell = \mathbf{W}'} \right)$$

特定の層のパラメタを
特定話者データ (少量) で
再学習 (もしくは変換)

何処を再学習すれば良いのか，明確な基準はない

Adaptation: KL-regularization

$$\hat{\Theta} = \mathrm{argmax}_{\Theta'} \left((1-\rho) F(\Theta') - \rho \sum_t \mathrm{KL}\left[P(q_t|\boldsymbol{x}_t, \Theta) || P(q_t|\boldsymbol{x}_t, \Theta') \right] \right)$$

少量・特定話者の適応データに
対する識別率を上げていく

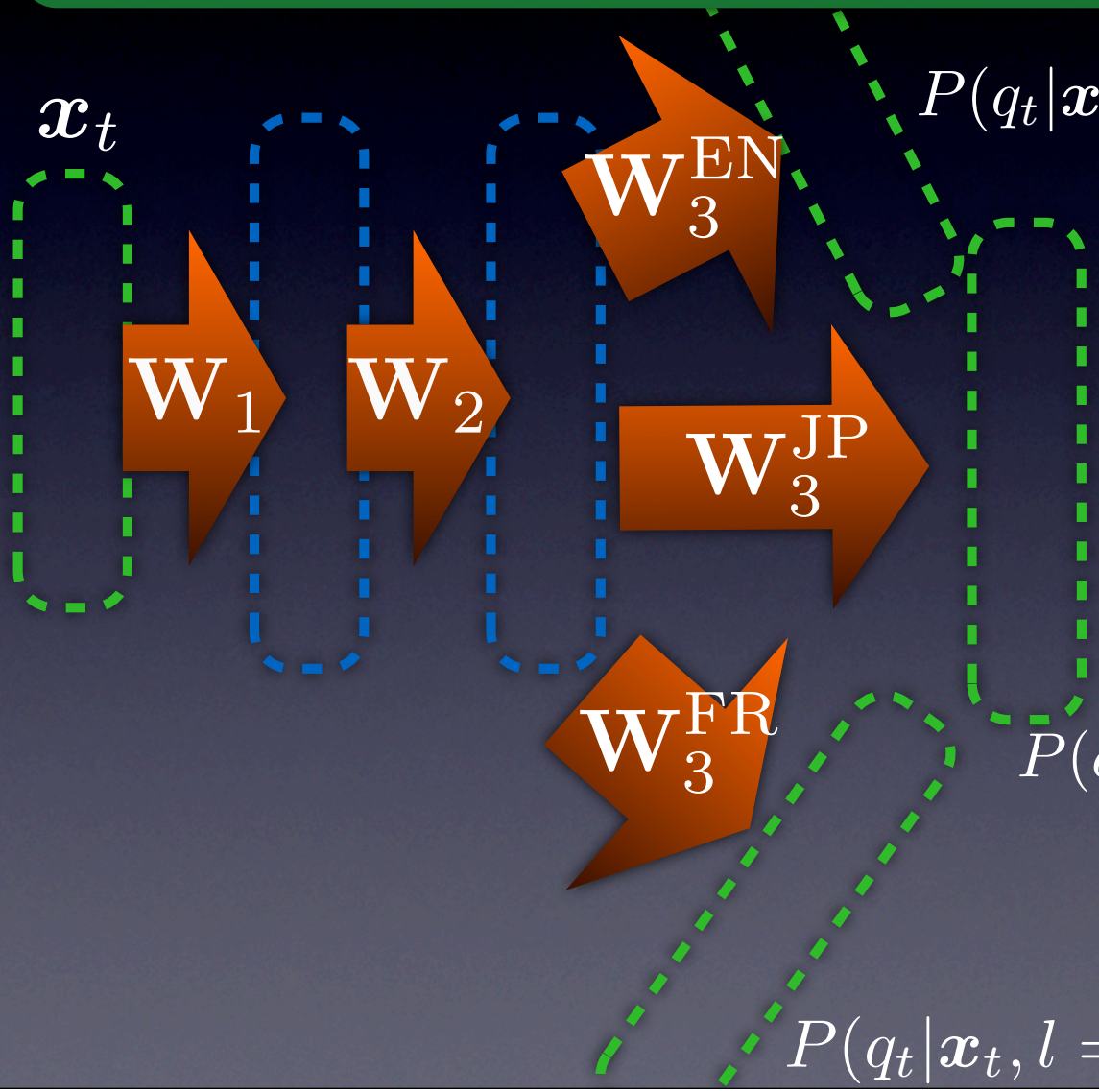
適応データについての
予測分布が元の (不特定話者の)
モデルの予測分布から逸脱しない

少量データで全ての層のパラメタの学習に成功

従来の枠組みから大きく逸脱

Multilingual Speech Recognition

例えば，日本語の音声認識器を作成する際，
英語やフランス語のデータセットは活用できるか？



複数の研究機関が前段の特徴抽出を共有することで多言語認識の精度向上が可能なことを確認

[Ghoshal 2013, Heigold 2013, Tüske 2013]

まとめ

まとめ

音声認識におけるニューラルネットワークの歴史を紹介

- TDNN
- Hybrid NN/ HMM
- GMM/ MLP-tandem

音声認識分野で研究されている深層学習技術を紹介

- DNN-HMM
- Sequential discriminative training
- Recurrent neural network language model
- Connectionist temporal classification
- Speaker adaptation
- Multilingual acoustic models