音声の構造的表象を用いた大語彙音声認識の識別的リランキング*

◎鈴木 雅之 (東大)，倉田 岳人，西村 雅史 (IBM)，峯松 信明，広瀬 啓吉 (東大)

1 はじめに

音声認識をはじめ，構文解析，機械翻訳などのタスクにおいて，識別的リランキング（Discriminative reranking）が，さらなる精度向上を実現する手法として注目されている [1, 2, 3]．識別的リランキングとは，何らかの手法で複数の解候補を選んだ後，識別モデルを使って解候補をリランキングする手法である．

識別的リランキングの一つの利点は，任意の特徴量を利用できることである．例えば音声認識の音響モデルに HMM を使う場合，特徴量としては PLP やその Δ，ΔΔ など，時間窓ごとに得られる局所的特徴量しか利用できない．しかし識別的リランキングでは，局所的特徴量に限らず，長時間単位にまたがる特徴も利用できる．従来の研究では，単語 n-gram カウントが識別的リランキングの特徴量として多く用いられ，これは「識別的言語モデル」という名称でも知られている [4]．最近では，言語的な特徴の他にも，duration n-gram などの音響的特徴も利用されるようになった [5]．

本研究では，大語彙音声認識の識別的リランキングにおいて，音声の構造的表象を特徴量として利用することを提案する．構造的表象は，一発声に含まれる音素と音素の距離から構成されるもので，非言語的特徴のミスマッチに頑健であることが知られている [6]．しかし非局所的な特徴量であるために，大語彙音声認識のデコーディングアルゴリズムと相性が悪く，これまで大語彙音声認識に用いられた例は存在していなかった．一方，識別的リランキングの枠組みを用いれば，デコーディングは前段階で終了しているため，構造的表象のような非局所的特徴も利用することが可能となる．

2 関連研究

2.1 音声の構造的表象

音声の構造的表象の概念図を Fig. 1 に示す．ある二つの分布に任意の一対一対応変換を施しても，その分布間距離尺度の一種である f-divergence は不変であることが証明されている [7]．音声の構造的表象とは，ある話者の発声を音素などの単位で特徴量空間上で分布化し，それらすべての分布間の f-divergence を

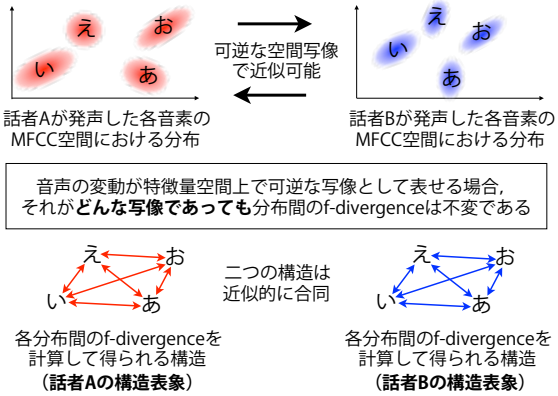

Fig. 1 Invariant structures

計算することで得られる構造のことである．

ここで，話者の違いはある一対一対応変換で近似することができる．例えば声質変換に関する研究では，話者の違いを特徴量の非線形変換と仮定して対応関係を学習している．また，音声認識における MLLR 適応でも，話者の違いを線形変換と仮定して近似している．構造的表象は，このような変換に不変であるため，話者の違いに高い頑健性を持つ．なお話者の違いの他にも，例えばマイクの違いや伝送特性の違いなども，特徴量の線形変換として近似できるため，構造的表象は高い頑健性を持つ．

2.2 大語彙音声認識の識別的リランキング

大語彙音声認識の識別的リランキングのモデル化には数多くのやり方があるが，本稿では，N 個の認識結果の仮説を入力とし，線形識別モデルを用いてリランキングを行うタイプのものを用いる．

まず，線形識別モデルに用いる特徴量を定義する．具体的には，観測した音声 x，仮説 y から，任意の特徴量ベクトル $\mathbf{\Phi}(x, y)$ を定義する．例えば

$$\mathbf{\Phi}(x,y) = \begin{bmatrix} \text{“a” という単語が } y \text{ に含まれる数} \\ \text{“the” という単語が } y \text{ に含まれる数} \\ \vdots \end{bmatrix}. \quad (1)$$

のように，仮説に含まれる単語のカウントを特徴量として利用できる．また他にも，例えば PLP の時間平均など，x と y に依存する任意の特徴量が利用できる．

*Discriminative reranking for LVCSR using invariant structure by M. Suzuki (Tokyo Univ.), G. Kurata, M. Nishimura (IBM), N. Minematsu and K. Hirose (Tokyo Univ.)

Input: Training samples $(x_i, \overline{y_i}, \underline{y_i})$ for $i = 1 \ldots I$
Initialization: $\boldsymbol{\alpha}_0^I = \mathbf{0}$
1: **for** $t = 1 \ldots T$ **do**
2: $\quad \boldsymbol{\alpha}_t^0 = \boldsymbol{\alpha}_{t-1}^I$
3: $\quad$ **for** $i = 1 \ldots I$ **do**
4: $\quad\quad$ **if** $\boldsymbol{\alpha}_t^{i-1} \cdot \boldsymbol{\Phi}(x_i, \overline{y_i}) + \phi_0(x_i, \overline{y_i})$
$\quad\quad\quad > \boldsymbol{\alpha}_t^{i-1} \cdot \boldsymbol{\Phi}(x_i, \underline{y_i}) + \phi_0(x_i, \underline{y_i})$ **then**
5: $\quad\quad\quad \boldsymbol{\alpha}_t^i = \boldsymbol{\alpha}_t^{i-1} + \lambda \left(\boldsymbol{\Phi}(x_i, \underline{y_i}) - \boldsymbol{\Phi}(x_i, \overline{y_i}) \right)$
Output: $\boldsymbol{\alpha} = \Sigma_{i,t} \boldsymbol{\alpha}_i^t / IT$

Fig. 2 A variant of averaged perceptron algorithm

次に，上記で定義した任意の特徴量を用い，以下の線形識別モデルで最終的な認識結果 y^* を選択する．

$$y^* = \underset{y \in \mathrm{NBEST}(x)}{\operatorname{argmax}} \ \boldsymbol{\alpha} \cdot \boldsymbol{\Phi}(x, y) + \phi_0(x, y) \qquad (2)$$

ここで，NBEST(x) は，x から HMM ベースの音声認識などを用いて得られる N 個の仮説の集合である．$\phi_0(x, y)$ は，HMM ベースの音声認識の際に得られる対数尤度である．$\boldsymbol{\alpha}$ は線形識別モデルのパラメタであり，次元数は $\boldsymbol{\Phi}(x, y)$ と同じで，各特徴に対する重みとなっている．$\boldsymbol{\alpha} = \mathbf{0}$ であれば，N 個の仮説を出力したモデルの対数尤度 $\phi_0(x, y)$ がそのままスコアになるため，リランキングを行わないことに相当する．

$\boldsymbol{\alpha}$ は，本稿では Fig. 2 に示す averaged perceptron の一種を用いて学習する．$\overline{y_n}$ と $\underline{y_n}$ は，観測音声 x_n に対する N 個の仮説の中で，最もエラーレートが高かった仮説，低かった仮説をそれぞれ表す．I が学習データの個数，T が学習の繰り返し回数を表す．λ は学習率で，本稿では一定の値に予め設定した．このアルゴリズムの中心的なアイディアは，エラーレートが最も高い仮説にペナルティを与え，逆にエラーレートが最も低い仮説に報酬を与える，というものである [9]．アルゴリズムの最後に $\boldsymbol{\alpha} = \Sigma_{i,t} \boldsymbol{\alpha}_i^t / IT$ と平均をとってるが，これはマージンを大きくし汎化性能を向上させる効果がある [10]．

3 提案手法

音声認識における識別的リランキングの一つの利点は，HMM ベースの音声認識では扱えない，非局所的な特徴量を利用できる点にある．しかしこれまで，識別的リランキングのモデルに関する研究は盛んに行われているものの，特徴量に関する研究はあまり行われてこなかった．特に，音響特徴量の非局所的特徴量に関しては，ほとんど検討がなされていない．

本研究では，識別的リランキングの特徴量として，非局所的な音響特徴量である構造的表象を利用することを提案する．構造的表象は，PLP 等とは異なり，音声の相対関係を捉えたものである．そのため，HMM ベースの音声認識で使われる情報とは異なる情報を持っていると考えられ，識別的リランキングで利用することが認識精度の向上に有効だと考えられる．

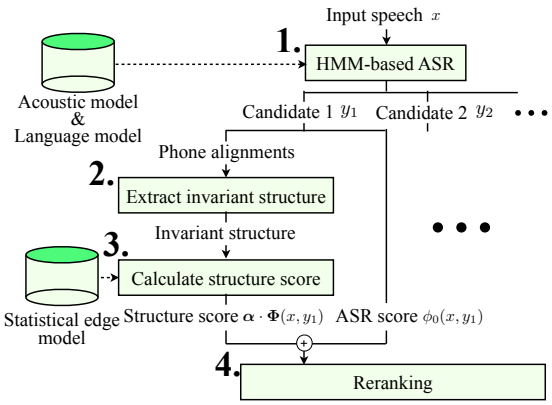

Fig. 3 A framework of discriminative reranking for LVCSR leveraging invariant structure

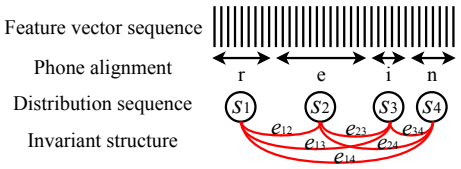

Fig. 4 A procedure of extracting an invariant structure from a phone alignment

提案手法の概略を Fig. 3 に示す．Fig. 3 の 1〜4 の番号は，以下のサブセクションに対応している．

3.1 HMM ベースの音声認識

まず，広く利用されている HMM ベースの音声認識システムを用い，対数尤度の高い上位 N 個の仮説を得る．ここでそれぞれの仮説ごとに，対数尤度 (ASR score $\{\phi_0(x, y_n)\}_{n=1 \cdots N}$) と，音素アライメントの結果を保存しておく．

3.2 音声の構造的表象の抽出

次に，N 個の仮説それぞれから構造的表象を抽出する．Fig. 4 に，仮説 [r e i n] のから構造的表象を抽出する方法を示す．まず，音素アライメント結果を利用して，M 個の仮説内音素ごとにガウス分布を推定する．次にそれらのガウス分布間の f-divergences ($\{e_{ij}\}_{1<i<j<M}$,「エッジ」と呼ぶ) を計算することで，構造的表象を抽出する．

3.3 構造スコアの計算

次に，構造的表象を構成する各エッジの尤もらしさのスコアを特徴量として，識別的リランキングを行う．

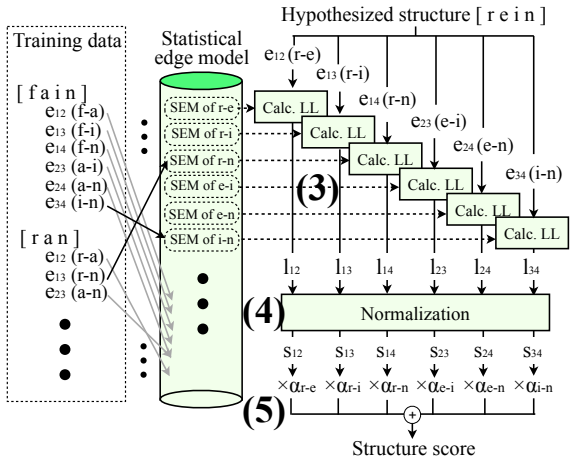

Fig. 5　Process for building statistical edge models (SEMs) and process for calculating a structure score. Log likelihood is abbreviated as LL. (3), (4), and (5) in this figure correspond to those of equations in this paper.

まず「各エッジの尤もらしさ」を計算するために，統計的エッジモデル（Statistical Edge Model; SEM）を予め学習する．Fig. 5 の左側に，SEM の学習プロセスを示す．SEM の学習では，二つの音素ペアをラベルとして，1 次元の f-divergence を Gaussian Mixture Model（GMM）により学習する．音素が P 種類あった場合，SEM は $P(P-1)/2$ 個になる．

Fig. 5 の右側に，ある仮説と音素アライメントから得た構造的表象が入力されたときに，それに対応する structure score を計算するプロセスを示す．まず，構造的表象のそれぞれのエッジ e_{ij} に対して，以下のように SEM の対数尤度を計算する．

$$l_{ij} = \log \sum_{k=1}^{K} w_{p_{ij}}^{k} \mathcal{N}(e_{ij}; \mu_{p_{ij}}^{k}, \sigma_{p_{ij}}^{k}) \tag{3}$$

ここで p_{ij} は，e_{ij} に対応する音素ペア ID（Phoneme-Pair ID; PPID）を表す．$w_{p_{ij}}^{k}$, $\mu_{p_{ij}}^{k}$, $\sigma_{p_{ij}}^{k}$ は，PPID が p_{ij} の SEM（K 混合の GMM）の k 番目のコンポーネントの重み，平均，分散共分散行列である．そして得られた $\{l_{ij}\}_{1<i<j<M}$ を，duration，音素数 M で以下のように正規化する．

$$s_{ij} = \frac{f_i + f_j}{M-1} l_{ij} \tag{4}$$

ここで f_i, f_j は i 番目，j 番目の音素にアライメントされたフレーム数で，音素アライメント結果から簡単に計算することが出来る．

最後に，structure score を以下のように計算する．

$$\text{structure-score} = \boldsymbol{\alpha} \cdot \boldsymbol{\Phi}(x, y) \tag{5}$$

ここで，$\boldsymbol{\alpha}$ と $\boldsymbol{\Phi}$ は，常に $\frac{P(P-1)}{2}$ 次元となるベクトルであり，それぞれの要素は PPID に対応する．$\boldsymbol{\Phi}$ は，それぞれの PPID に対応する正規化エッジスコア $\{s_{ij}\}_{1<i<j<M}$ の和で，以下のように定義される．

$$\boldsymbol{\Phi}(x, y) = \begin{bmatrix} \sum_{i,j} s_{ij} \text{ if } p_{ij} = 1 \text{ , otherwise } 0 \\ \sum_{i,j} s_{ij} \text{ if } p_{ij} = 2 \text{ , otherwise } 0 \\ \vdots \\ \sum_{i,j} s_{ij} \text{ if } p_{ij} = \frac{P(P-1)}{2} \text{ , otherwise } 0 \end{bmatrix} \tag{6}$$

もし特定の PPID が仮説 y の構造的表象に含まれなければ，その PPID に対する特徴は 0 になるため，$\boldsymbol{\Phi}(x, y)$ は一般的にスパースなベクトルとなる．また $\boldsymbol{\alpha}$ は，先に説明した Fig. 2 のアルゴリズムで予め学習しておく．そのため $\boldsymbol{\alpha}$ は，エラーレートを削減するためにどの PPID のスコアを重要視すればよいかの度合い，と解釈できる．

3.4　リランキング

最後に，HMM ベースの音声認識から出力されたスコア $\phi_0(x, y_n)$ と，構造的表象から得た structure score を組み合わせ，式 (2) を用いてリランキングを行う．

4　実験

4.1　実験条件

提案手法の有効性を検証するために，日本語の連続数字音声認識，大語彙音声認識の二つの実験を行った．実験条件を Table 1, Table 2 に示す．HMM ベースの音声認識システムとして [11] で示されているシステムを用い，10-ベストの仮説と，音素アライメントを出力した．HMM は音素単位で学習され，決定木を用いて状態クラスタリングを行った．構造的表象を得るための分布の特徴量には，13 次元の PLP 特徴量を使い，3 状態 left-to-right HMM の 2 状態目に対応する部分のみを用いてガウス分布の平均を ML 推定した．ガウス分布の分散に関しては，少ないデータから分布を学習するため，音素ごとに分散を予め設定しておき，それを常に利用した．f-divergence としては，バタチャリヤ距離の平方根を用いた．SEM には，16 混合の GMM を用いた．

4.2　結果

Fig. 6 と Fig. 7 に，連続数字音声認識の単語誤り率（Word Error Rate; WER）と，大語彙連続音声認識の文字誤り率（Character Error Rate; CER）を示す．CER を用いる理由は，日本語は単語分割に曖昧性があるためである．そのため $\boldsymbol{\alpha}$ の学習時の $\overline{y_n}$, $\underline{y_n}$ には，それぞれ，10 ベストの中で WER・CER が最も高い・低いものを利用した．図の横軸は $\boldsymbol{\alpha}$ 学習の繰り返し回数 T を表す．$\boldsymbol{\alpha}$ の初期値を $\mathbf{0}$ としてい

Table 1　Experimental condition for Japanese continuous digits recognition.

発声内容	1 から 11 回の日本語連続数字読み上げ
HMM 学習データ	27.5 時間 / 667 話者 / 17316 発声
SEM 学習データ	27.5 時間 / 667 話者 / 17316 発声
$\boldsymbol{\alpha}$ 学習データ	5.0 時間 / 520 話者 / 3977 発声
テストデータ	1.5 時間 / 100 話者 / 7382 発声
HMM 状態数	500
HMM ガウシアン数	15000
monophone 数 (P)	18
monophone-pair 数	136
言語モデル	0 から 9 の 10 の数字，終端記号を当確率で出力するユニグラム
ベースライン WER	1.09% (S=67, I=140, D=14 / 20303)*
10 ベストオラクル	0.75% (S=59, I=85, D=9 / 20303)*

* S: 置換誤り数, I: 挿入誤り数, D: 削除誤り数

Table 2　Experimental condition for Japanese LVCSR

発声内容	日本語読み上げ音声
HMM 学習データ	352 時間 / 1325 話者 / 196475 発声
SEM 学習データ	24 時間 / 100 話者 / 13112 発声
$\boldsymbol{\alpha}$ 学習データ	30 時間 / 164 話者 / 16733 発声
テストデータ	1.5 時間 / 20 s 話者 / 600 発声
HMM 状態数	5000
HMM ガウシアン数	150000
monophone 数 (P)	57
monophone-pair 数	1596
言語モデル	Modified Kneser-Ney smoothing [12] を用いた単語 2-gram
単語数	104262
ベースライン CER	3.59% (S=422, I=56, D=64 / 15096)*
10 ベストオラクル	1.32% (S=161, I=15, D=24 / 15096)*

* S: 置換誤り数, I: 挿入誤り数, D: 削除誤り数

るため，$T = 0$ は，ベースラインシステムのエラーレートを表す．λ は学習率であり，$0.1, 0.2, 0.5$ の三種類で実験を行った．

提案手法は，どのような λ，T においても，ベースラインを上回る精度となった．連続数字読み上げ音声の実験に関しては，$\lambda = 0.2$，$T = 4$ または 5 のとき，WER は 0.94% と最小になり，ベースラインシステムから 14.1% の WER 削減を達成した．このとき，置換誤りは 67 から 64 に，挿入誤りは 140 から 113 に，削除誤り 14 から 13 に，それぞれ減少した．大語彙音声認識に関しては，$\lambda = 0.2$，$T = 4$ のとき，CER は 3.35% と最小になりベースラインシステムから 6.69% の CER 削減を達成した．このとき，置換誤りは 422 から 401 に，挿入誤りは 56 から 44 に，削除誤り 64 から 61 に，それぞれ減少した．

5　結論

本論文では，大語彙音声認識の識別的リランキングにおいて，音声の構造的表象を利用する手法を提案した．日本語の大語彙音声認識実験の結果，HMM ベースのシステムから 6.69% の CER 削減を実現することができた．

今後の課題として，識別的リランキングの素性エンジニアリング（feature engineering）がある．構造的表象に限らず，例えば単語 n-gram カウントのような言語的特徴量も，識別的リランキングには有効であることが知られている．これらは異なる種類の情報を捉えていると考えられるため，同時に利用すれば，音声認識率の向上に有効ではないかと考えられる．

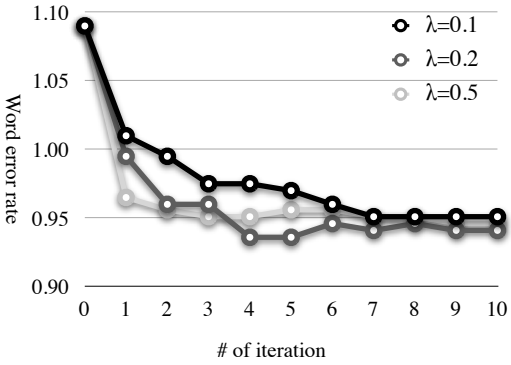

Fig. 6　Word error rate for Japanese connected digits recognition

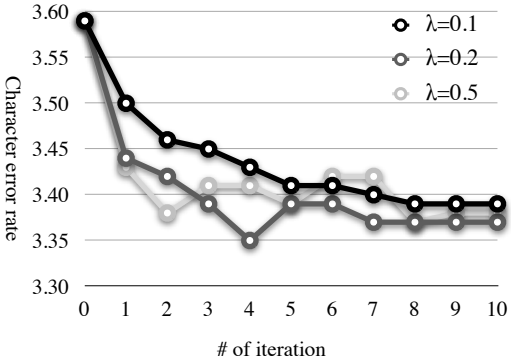

Fig. 7　Character error rate for Japanese LVCSR

参考文献

[1] B. Roark, *et.al. CSL*, vol.21, no.2, pp.373–392, 2007.
[2] M. Collins, *ICML*, pp.175–182, 2000.
[3] L. Shen *et.al. HLT-NAACL*, pp.177–184, 2004.
[4] T. Oba *et.al. ICASSP*, pp.5588–5591, 2011.
[5] M. Lehr and I. Shafran, *ICASSP*, pp.5542–5545, 2010.
[6] N. Minematsu, *et.al.*, *Jornal of New Generation Computing,* vol. 28, no. 3, pp.299–319, 2010.
[7] Y. Qiao *et.al.*, *IEEE TSP*, vol 58, No.7, pp.3884–3890, 2010.
[8] B. Roark, *et.al.*, *ICASSP,* pp.749–752, 2004.
[9] 大庭 他，SP2007-185-214, pp. 99–104, 2008.
[10] K.Crammer *et.al*, NIPS Workshop on Learning With Structured Outputs, 2005.
[11] S. Chen, *et.al.*, *IEEE TSAP,* vol. 14, no. 5, pp.1596-1608, 2006.
[12] S. Chen *et.al*, *CSL,* vol. 13, no.4, pp. 359–393, 1999.