音声の構造的表象を用いた連続数字音声認識*

☆鈴木 雅之 (東大)，倉田 岳人，西村 雅史 (IBM)，峯松 信明 (東大)

1 はじめに

音声認識（Automatic Speech Recognition; ASR）には，年齢，性別，マイク，背景ノイズなどの非言語的情報に起因する，学習データと評価データのミスマッチへの頑健性が求められる．ミスマッチに頑健なASR を実現するために，これまで正規化，適応など，多種多様な手法が提案されてきている．

近年，このような手法の一つとして，音声の構造的表象が提案された [1]．音声の構造的表象は，音と音との相対関係のみを捉えたものであり，特徴量空間の一対一対応の空間写像に理論的に不変性を持つ．そのため，そのような空間写像で近似できる非言語的特徴に高い頑健性を持つ．音声の構造的表象を利用してこれまで，孤立単語音声認識，外国語発音評定などが提案され，その頑健性が示されている [2]．

しかし，これまで音声の構造的表象は，連続音声認識には応用されていない．これは，音声の構造的表象を利用する適切なデコーディングアルゴリズムが存在しないからである．Hidden Structure Model (HSM) と短時間特徴量系列のボトムアップクラスタリングを用いることでこの問題を解決しようとする研究もあるが，現実的なタスクでの有効性は示すことができていない [3]．

そこで本研究では，N-best リランキングの枠組み [4] に音声の構造的表象を用いることで，連続音声認識を実現する手法を提案する．まず，広く用いられている HMM ベースの ASR [5] を用い，N-best リストと音素アライメントを出力する．次に，N-best リストに含まれる仮説ごとに，音素アライメントから音声の構造的表象を抽出する．そして，構造的表象ベースのスコアを算出し，それと HMM ベースの ASR で算出されたスコアと組み合わせて N-best リランキングを行う．この手法を用いれば，デコーディング自体は HMM ベースのアルゴリズムを利用できるため，連続音声認識に応用することが可能になる．

提案手法の評価のため，日本語の連続数字音声の認識実験を行った．実験の結果，HMM ベースの ASR から，数字誤り率を 17.4 % 削減することができた．

2 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

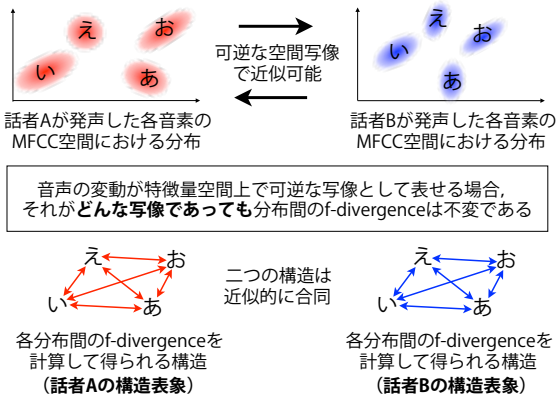

Fig. 1 Invariant structures

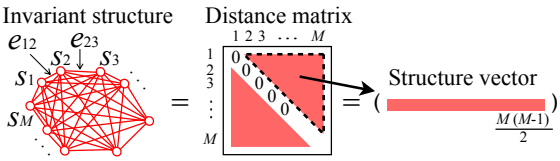

Fig. 2 An invariant structure can be represented as a distance matrix and a structure vector

分布間の f-divergence は不変であることが証明されている [6]．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，ASR 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，このような変換に理論的に不変であるため，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．

構造的表象に関する用語を Fig. 2 を用いて定義する．構造的表象は，M 個のノードからなる．それぞれのノードを，$\{s_i\}_{i=1}^{M}$ で表す．それぞれのノードは，音響イベントの分布であり，具体的には音素 HMM の各状態などが対応する．次に構造のエッジ長を，$\{e_{ij}\}$ $(1 \leq i \leq M, i < j \leq M)$ で表す．エッジ長は，二つのノード間の f-divergence である．音声の構造的表象は，数学的には距離行列として表現できる．もし，f-

*Continuous digits recognition leveraging speech structure. by M. Suzuki (The Univ. of Tokyo), G. Kurata, N. Masafumi (IBM) and N. Minematsu (The Univ. of Tokyo)

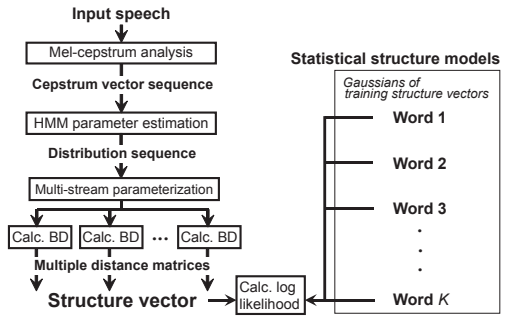

Fig. 3 Framework of the structure-based isolated word recognition [2].

divergence として対称な距離尺度を選べば，距離行列の上三角成分だけで情報をすべて表現できることになる．このような f-divergence として，Bhtaccharyya Distance (BD) の平方根がよく用いられる．距離行列の上三角成分をベクトルに並べ直したものを，構造ベクトルと呼ぶ．構造ベクトルは，$\frac{M(M-1)}{2}$ 次元のベクトルである．

2.1 構造的表象を用いた孤立単語音声認識

次に，音声の構造的表象を用いた孤立単語音声認識を，Fig. 3 を用いて説明する．Fig. 3 の左側は，入力の孤立単語発声から構造ベクトルを抽出する方法を示している．まず，音響分析により，入力された孤立単語発声から MFCC などの短時間特徴量系列を抽出する．次に，得られた一つの短時間特徴量系列のみから，left-to-right HMM を学習する．この際，HMM の状態数は M 個，各状態の出力する分布はガウシアンとする．そして，学習された HMM 各状態の持つガウシアンを，構造的表象のノードとみなして利用する（状態遷移確率は捨てる）．ここで，少ないデータから分布を推定したことによる不安定性を取り除くため，MAP 推定を導入する．次にガウシアン系列を，特徴量の次元方向に分割し，マルチストリーム化する．この処理は，構造的表象の強すぎる不変性に制約条件を加え，音声認識に必要な情報を増やす効果がある．これ関しては，[2] を参照されたい．その後，各ストリームごとに，f-divergence を計算することで構造的表象を抽出し，それを構造ベクトルにする．ここで，マルチストリーム化をしているため，構造ベクトルの次元数は $\frac{M(M-1)}{2}$ のストリーム数倍となることに注意する．これで入力の孤立単語発声から構造ベクトルを抽出することができたので，これを単語ごとのモデルと比較し，最も対数尤度が大きいモデルに対応する単語を音声認識結果として出力する．

Fig. 3 の右側は，構造的表象の音響モデルである．我々はこのモデルを構造統計モデル（Statistical Structure Model; SSM）と呼ぶ．今回は，単語の数を K として，K 個の単語ごとの SSM を用意する．SSM として，構造ベクトルのガウシアンを用いる．構造ベクトルの抽出は，先に示したものと同じ方法を用いる．ここで，SSM の学習に用いる構造的ベクトルも，入力に用いる構造的ベクトルも，まったく同じ次元数，すなわち同じノード数 M かつ同じストリーム数を持たなければならない制限がある．

2.2 従来手法の問題点

従来の方法では，M の数を予め固定しなければならない．そのため，入力の長さが大きく変化する連続音声認識では，従来の枠組みを利用することは不可能である．この問題に対処するためには，なんらかのデコーディングアルゴリズムが必要である．しかし，デコーディングは，短時間特徴量系列に適切なアライメントを行うものだが，そもそも音声の構造的表象を抽出するためには，短時間特徴量系列に対するアライメントが必要になる．この「鶏と卵」の問題により，構造的表象をデコーディングに利用することは難しい．

一つの可能性としては，HSM を使う方法がある．HSM を用いた ASR では，まず，ケプストラム系列をボトムアップクラスタリングなどの手法により分布系列に変換する．HSM には，この分布系列に対してデコーディングを行うための各種アルゴリズムが定式化されている [3]．しかしながら，HSM は計算量が高すぎるため，人工データにしか適応されておらず，精度も十分ではない．

従来の方法には，さらにもう一つ問題がある．従来の枠組みでは，孤立単語ごとに音響モデルを用意していたため，単語サイズ K が大きくなると，非常に膨大なデータが必要になる．そのため，単語単位などではなく，音素単位など，もっと小さい単位でモデルを用意することが望まれる．

3 提案手法

従来手法の問題点を解決するため，本研究では HMM ベースの ASR から出力した N-best リストを構造的表象を用いてリランキングする手法を提案する．提案手法の概略を Fig. 4 に示す．Fig. 4 の中で付けられている 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-based ASR

まず，広く利用されている HMM ベースの ASR システムを用い，尤度の高い上位 N 個の仮説を N-best として出力する．ここでそれぞれの仮説ごとに，ASR のスコアと，音素アライメントの結果を保存しておく．

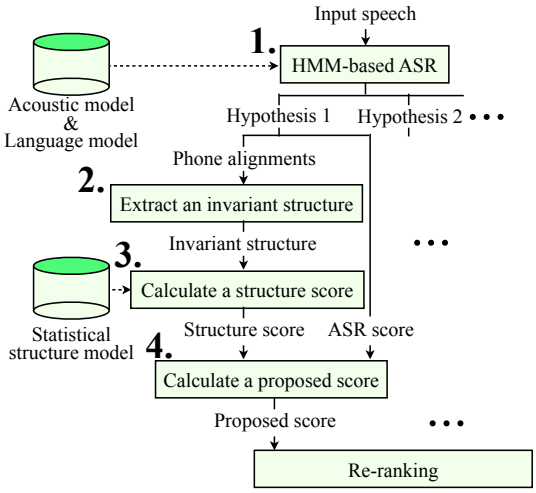

Fig. 4 N-best re-ranking leveraging invariant structure

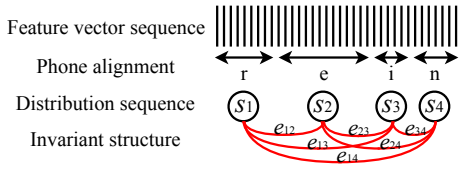

Fig. 5 A procedure of extracting an invariant structure from a phone alignment. A hypothesized word is [r e i n].

3.2 Extract an invariant structure

次に，N 個の仮説それぞれから構造的表象を抽出する．Fig. 5 に，音素アライメント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から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 を計算することで，構造的表象を抽出する．

3.3 Calculate a structure score

次に，構造的表象のモデルの対数尤度を用いて，構造的表象のスコアを計算する．従来の構造的表象を用いた孤立単語音声認識では，単語ごとに構造ベクトルのガウシアンを学習して構造統計モデル（SSM）として用いていた．しかし，音声認識の対称が大語彙になった場合，単語単位で音響モデルを用意するためには，莫大な学習データが必要になってしまう問題がある．

この問題を解決するため，単語単位ではなく，エッジ単位で音響モデルを利用することを考える．これは，[7] でも利用されている手法である．我々は，このエッジ単位の音響モデルを，Statistical Edge Model (SEM) と呼ぶことにする．

Fig. 6 の左側に，SEM を学習データから学習プロセスを示す．右側に，ある構造的表象の仮説が入力さ

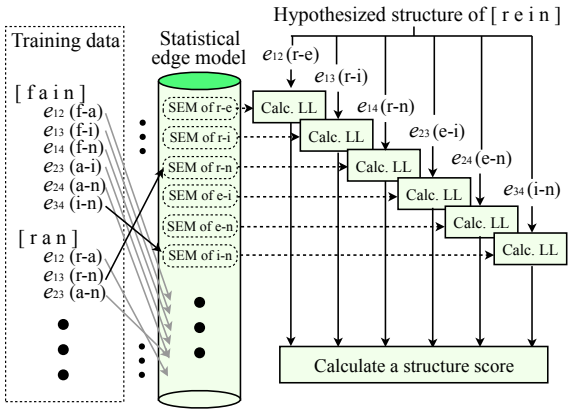

Fig. 6 Process for building statistical edge models (SEMs) and process for calculating log likelihoods using SEMs. Log likelihood is abbreviated as LL.

れたときに，それに対応する SEM の対数尤度を計算するプロセスを示す．SEM の学習では，二つの音素ペアをラベルとして，SEM をガウシアンもしくは Gaussian Mixture Model（GMM）により学習する．音素が P 個あった場合，SEM は $P(P-1)/2$ 個必要になる．SEM が学習できれば，入力の構造のエッジ e_{ij} ごとに，対数尤度を計算することができる．

n 番目の仮説 h^n の構造スコアは，すべてのエッジの対数尤度を，以下の式により統合して計算する．

$$\mathrm{score}_{\mathrm{structure}}(h^n) = \frac{\sum_{i=1}^{O}\sum_{j=i+1}^{O} L_{ij}(e_{ij}^n)}{O} \quad (1)$$

ここで $L_{ij}(e_{ij}^n)$ は n 番目の仮説から得られるエッジの対数尤度であり，O は n 番目の仮説に出現した音素の数である．

3.4 Calculate a proposed score

最後に，HMM ベースの ASR から出力されたスコアと，SEM から計算した構造スコアを組み合わせることで新たなスコアを算出し，それを用いて N-best リストのリランキングを行う．最終的なスコアは，以下の式で計算する．

$$\mathrm{score}_{\mathrm{proposed}}(h^n) = \mathrm{score}_{\mathrm{ASR}}(h^n) + w\ \mathrm{score}_{\mathrm{structure}}(h^n) \quad (2)$$

ここで w は構造スコアの重みであり，あらかじめ決定しておく．

今回提案した手法は，HMM ベース ASR で利用するデコーディングアルゴリズムを利用しているため，連続音声認識に問題なく適用することができる．また，HMM ベース ASR のスコアも使っているため，仮に w を 0 にすれば，HMM ベースの ASR の音声認識率自体は保証される．提案手法は，HMM ベースのスコアと，音素間の相対的関係の確からしさを示す構造スコアを組み合わせている．

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実験を行った．HMM ベースの ASR として [5] で示されているシステムを用い，ASR スコア，N-best リスト，音素アライメントを出力した．HMM の学習には，27.5 時間，667 名分のデータを用いた．それぞれの発声は，数字を 1 〜 11 回連続して読み上げたもので，音素数は 18 ある．HMM は，context-dependent，3 状態，left-to-right HMM である．HMM の状態は決定木でクラスタリングされており，状態数は 500，ガウシアン数は 15,000 である．また言語モデルには，0 から 9 の 10 単語もしくは文終了記号を同確率で出力する文法を用いた．

SEM を学習するには，HMM の学習データのサブセットとなる 2.5 時間，67 名分のデータを用いた．ノードの単位には，18 個の monophone を用いた．そのため，SEM は ${}_{18}C_2 = 136$ 個学習される．Monophone のガウス分布を抽出する際には，13 次元の PLP 特徴量を使い，3 状態 left-to-right HMM の真ん中の状態に対応する部分のみを用いて ML 推定した．ただし，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，同一音素内では分散パラメータを共有して推定した．このガウス分布を 12 ストリーム化し [2]，f-divergence として BD の平方根を用いて構造を抽出した．SEM として利用する分布には，混合数 1，2，4，8，16 の GMM を用いた．

評価時には，HMM ベースの ASR を用いて N-best リストを抽出し，それをリランキングした．ここで，HMM ベースの ASR において，少なくとも 2 つ以上の N-best リストが得られたデータのみを評価に利用した．最終的に評価データは，1.0 時間，95 名分のデータとなった．リランキングを行う際の構造重み w は，1-person-leave-out の 95-fold クロスバリデーションにより決定した．

4.2 結果

Fig. 7 に単語 (数字) 誤り率を示す．図には，HMM ベースの ASR の結果（$w = 0$ の場合）と，N-best オラクルの結果も表示した．提案手法は，SEM の混合数にどれを用いても，ベースラインの結果から精度が向上している．最も精度が良いのは，混合数 4 の場合で，単語誤り率は 1.17 % となった．これは，ベースラインから 17.4% の誤り削減となっている．また，オラクルの場合でも単語誤り率は 0.87% であることを考えると，41.2% の誤りを削減できたことになる．

ここで理論的に述べれば，構造のエッジは非言語的特徴に頑健であるので，SEM の混合数は非常に小さくなるはずである．しかし今回，ノードとして monophone を利用しているため，コンテキストに依存してエッジが多様性を持つ可能性がある．そのため，最適な混合数は，1 ではなく，4 となったと考えられる．

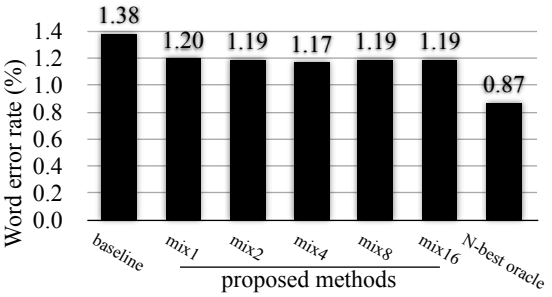

Fig. 7 Word error rate

5 結論

本論文では，音声の構造的表象を連続数字音声認識に応用するために，N-best リランキングを用いる手法を提案した．提案手法は，構造的表象を初めて実用的な連続音声認識に適応させた手法となっている．実験の結果，単語誤り率で HMM ベースの ASR から 17.4% の誤り削減を実現することができた．

今後の展望としては，HMM ベースの ASR を，識別学習や適応などを含むより高度なものに入れ替えることがある．これにより，より正確な音素アライメントが得られるため，構造スコアもより適切なスコアが得られるようになると考えられる．加えて，今回は連続数字音声認識実験だけで評価を行っているため，大語彙音声認識での評価を行いたい．今回の提案手法では，単語単位でなく，エッジ単位の音響モデルである SEM を用いているため，大語彙音声認識にも適用可能であると考えている．

参考文献

[1] N. Minematsu, *Proc. ICASSP*, pp.585–588, 2004.

[2] N. Minematsu, *et.al.*, *Jornal of New Generation Computing,* Vol. 28, No. 3, pp.299–319, 2010.

[3] Y. Qiao, *et.al.*, *Proc. ASRU,* pp.118–123, 2009.

[4] B. Roark, *et.al.*, *Proc. ICASSP,* pp.749–752, 2004.

[5] S. Chen, *et.al.*, *IEEE Transactions on Speech and Audio Processing,* 2006.

[6] Y. Qiao *et.al.*, *IEEE Trans. on Signal Processing*, Vol 58, No.7, pp.3884–3890, 2010.

[7] 齋藤 他, 信学技報, No.77, pp.7–12, 2010.