音声の構造的表象による頑健な教師無し語彙獲得に関する実験的検討*

☆尾崎洋輔，齋藤大輔， 峯松信明，広瀬啓吉 (東大)

1 はじめに

教師無しパターン発見はラベルの付与されていない音声信号から頻出パターンを獲得するというタスクであり、認知ロボティクスの分野では，幼児の単語獲得と関連付けて議論されている．この研究領域は未知言語音声に対するタグ付けやキーワード抽出，言語獲得のシミュレーションなどを内包しており，どのような問題設定を行うかによって使用できるリソース／事前知識が変わってくる．本研究はその中でも，特に乳幼児の発達段階に焦点を当てた言語獲得のシミュレーションをテーマとし，零リソースからの語彙獲得を最終的な目標としている．

教師無しパターン発見の先行研究で代表的なものとして Segmental DTW (以下 S-DTW) を用いたアルゴリズムが挙げられる [1]．これは DTW ベースでパターン間類似性を計算して収集し，それを基にパターン発見を行うものである．しかし距離基準にケプストラムのユークリッド距離を用いてる為，発話交替のある複数話者による音声信号を入力とすると，同じ発話内容にも関わらず複数のクラスタが形成されるという問題が発生する（教師無しパターン発見における複数話者問題）．この問題に対して，ベクトル時系列中の任意の二時刻間のベクトル間距離により構成される自己類似度行列 Self-Similarity-Matirx (以下 SSM) に変換して比較を行うといった手法が提案されており，その有効性が示されている [2]．音響特徴の絶対量ではなく相対量を話者の変化に頑健な特徴として採用する SSM の考え方は，峯松らの音声の構造的表象に非常に近い [3]．SSM は系列をフレーム単位で計算しているのに対して，構造的表象では特徴量系列を一旦分布系列に変換して，それらの分布間の距離行列を特徴としている．この分布間距離に f-divergence を用いると任意の連続かつ可逆な変換に対して不変であることが数学的に保証されている．そこで本研究では従来のパターン発見アルゴリズムへの構造的表象の適用によりパターン発見の高精度化，特に複数話者音声の問題への対応を目指す．

2 関連研究

2.1 S-DTW を用いた教師無しパターン発見 [1]

このアルゴリズムでは大きく分けて以下の 3 つのステップにより音声信号に頻出するパターン/キーワードの発見を試みている．

1. S-DTW を用いた類似パターン候補の発見
2. 発見されたパターン候補の始終端の決定（得られた音声区間をノードと呼ぶ）
3. ノードクラスタリングに基づく語（クラスタ）の同定

ここで距離尺度としてケプストラムのユークリッド距離を採用している為，複数話者データに対して頑健に動作する事は保証されておらず，如何にしてこの問題を解決するのかが課題となる．

これを解決するために，Glass らによって入力特徴として Gaussian Posteriogram (GP) や Universal Phone Posterior (UPP) を用いた拡張が行われている [4, 5]．これらの研究では事後確率空間で比較を行う事により上記の問題の解決を図っている．しかし，多量のデータからバックグラウンドモデルを作成しておく必要があり，本研究のように予め事前知識を仮定できないタスクにおいては使用するのは難しい．

2.2 SSM に基づくノードクラスタリング [2]

Glass らは特徴量の段階で話者情報を消すというアプローチを行っていたが，Muscariello らはノード間距離尺度として話者情報の影響の少ない手法を採用する事でこの問題に対処している．ここで使用されている SSM という特徴は，ケプストラム時系列の任意の二時刻間のフレーム距離群であり，こうして得られる行列は音声の相対量に基づくものなので話者の違いによる影響が比較的少ない．この SSM 同士を比較した値をパターン間の距離尺度として用いる事で，より頑健性が高いクラスタリングを実現している．

2.3 音声の構造的表象による話者の変化に頑健なパターン表現 [3]

音響特徴の相対量を話者の変化に頑健な特徴として採用する SSM の考え方は，峯松らの音声の構造的表象に非常に近い．音声の構造的表象は話者性を音声から分離し，言語的側面だけを表象することを目的とした枠組みである．

音声の音響特徴量に混入する非言語特徴量による歪みは，ケプストラム領域に対するアフィン変換で近似される，これらの歪みの混入は不可避であるので，非言語情報によらない認識を行うにはアフィン変換に対して不変な特徴量が必要となる．線形・非線形を問わずあらゆる可逆な変換・写像に対して不変な特徴量として，分布間距離である f-divergence が挙げられる．特に f-divergence の実装としてバタチャリヤ距離の平方根を用いると，二つの構造（距離行列）間の幾何学的差異が行列をベクトルと見なした場合に計算されるユークリッド距離に近似的に比例することが知られている．そこでケプストラム時系列を K 個のガウス分布の時系列で表現されるとすると，分布間距離によって張られる $K \times K$ の距離行列が得られる．こうして得られる距離行列は先の定義より非言

*Toward Robust Unsupervised Pattern Discovery Using Speech Structure by Y. Ozaki, D. Saito, N. Minematsu and K. Hirose (The University of Tokyo)

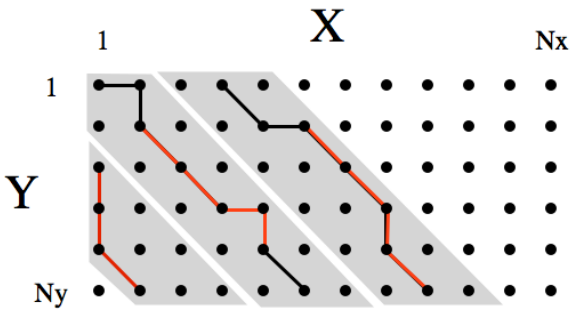

Fig. 1　Segmenatl DTW と局所アライメント（赤線部分）

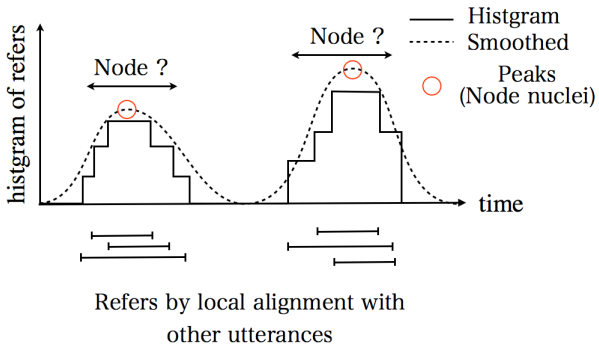

Fig. 2　局所アライメント情報による発話からのノードの抽出．発話毎に参照のヒストグラムを作成し，ノードの位置と始点・終点を確定する．

語特徴量に対して凡そ不変な音響的特徴量となり，これを音声の構造的表象と呼ぶ．これを特徴として用いる事で話者や環境の変化に左右されない頑健な音声パターンの表現が可能となる．

3　提案手法

3.1　概要

従来手法 [1] によって得られたノード情報を基に，新たなノード間距離尺度として構造的表象を導入し，複数話者データに対する精度の向上を試みる．そのため，提案手法の流れは以下のようになる．

1. S-DTW を用いた類似パターン候補の発見
2. 発見されたパターン候補の始終端の決定
3. 構造的表象を距離基準としたノードクラスタリングに基づく語（クラスタ）の同定

以下ではこれらの実装について詳しく説明する．

3.2　S-DTW

DTW とは時間方向に伸縮する二つの系列間の距離を計算する手法であり，同時に二系列の時間的対応も得られるので，現在では時系列間のアライメントやゲノム解析のシンボル間のパターン検出などにも用いられている．S-DTW もその一種であり，長時間の音声信号の中から類似するパターン対を見つけ出す事が目的となる．

S-DTW ではまず，Fig. 1 のように発話ペアから得られる距離行列をいくつかの領域に分割し，それぞれの領域で DTW によるアライメントを行う[1]．そうして得られた各領域のアライメントから，一定以上の長さを持ち且つ平均距離が最小となる局所アライメントを抽出する (Fig. 1)．この局所アライメントは「音声信号の中のこの部分とこの部分が似ている」という情報となる．特に平均歪みが比較的小さい局所アライメントについてはパターン対が同じ単語であるか，あるいは発音的に非常に似ているという事であり，そのような局所アライメントによって何度も繰り返し参照[2]されている区間は，語である可能性が高い．

3.3　ノードの決定

S-DTW で得られる局所アライメントにはパターンの始点・終点の時刻情報が含まれるが，ペアの選び方によって区間情報が異なる事が多く，場合によってはアライメントが発話内の全く異なる単語を参照している事も考えられる．そこで従来手法 [1] に従って，この参照情報を用いて発話からノードを抽出する．参照情報からノード決定までの流れは以下のようになる．

1. 平均歪みが閾値 (θ) 以上の参照を削除
2. 各時刻毎に参照回数をヒストグラム化[3]（Fig. 2）
3. 得られたヒストグラムの平滑化を行い，参照数が一定以上のピークをノードの核としてピックアップ
4. ピークを含む参照の平均開始時間と平均終了時間をノードの区間とする

こうして得られたノードは入力音声中に頻出するパターンであると考えられる．

3.4　構造特徴を用いたグラフクラスタリング

前項で得られたノードに対して，それぞれの距離情報を基にクラスタリングを行う．ノード間距離尺度として構造的表象を用いる事で，より頑健なクラスタリングが期待できる．

構造的表象の実装はいくつかあるが，分布系列となったサンプル間の時間的対応が必要という制約が存在する．そこで SSM の実装と同じように系列間のアライメントを予め行い，サイズ合わせを行った上で比較する．特徴時系列から構造的表象を抽出し比較するまでの流れをまとめると Fig. 3 のようになる．こ

[1]基本的に処理は発話単位であり，入力が長時間の連続音声の場合は `Voice Activity Detection` などの前処理により発話単位に区切られている事が前提となっている．

[2]参照とは本論文では，ある発声のある区間と別発声のある区間とが類似性により対応付けられていることを言う．

[3]実際は単なるヒストグラムではなく，参照の平均歪みで重み付けを行っている．

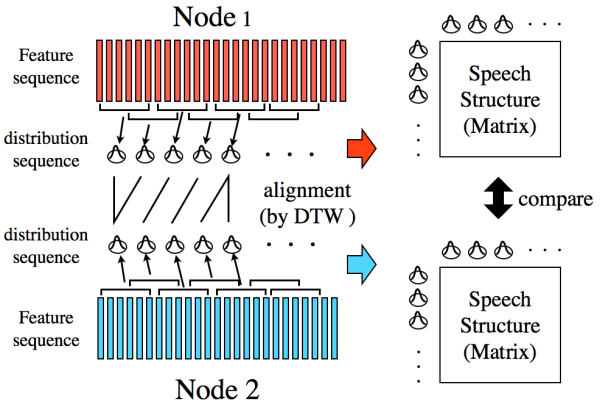

Fig. 3　構造的表象の実装

Table 1　音響分析条件

サンプリング周波数	8 bit / 16 kHz
窓幅&シフト長	25 ms length / 10 ms shift
特徴量	MFCC (12 dim.)
正規化	平均：発話単位
	分散：音声信号全体

Table 2　実験パラメータ

S-DTW	探索幅	700 [ms]
S-DTW	局所アライメントの最小長	800 [ms]
ノード抽出	局所アライメントの閾値	上位 10%に調整
ノード抽出	参照数の閾値	ノード数が 30 に調整
構造的表象	分布推定の分析幅	5
構造的表象	分布推定のシフト長	3
構造的表象	マルチストリーム [6]	ブロックサイズ 1

こでアライメントには分布の平均の値を用いており，構造間の比較にはユークリッド距離を距離行列のサイズで正規化したものを用いる．

4　実験

教師無しパターン発見アルゴリズムを複数話者タスクにおいて適用し，その結果を樹形図にして考察する．実験ではノードの抽出まではそれぞれの話者で行った上で，全話者に対して検出された全ノード間の距離を計測してクラスタリングを行う．

4.1　データベース

実験データベースには日本語連続数字読み上げコーパス AURORA2J を用いた．語彙数は 11 あり，全て一桁の数字である[4]．話者は男性話者 MBD と女性話者 FNG の 2 人を使用する．

4.2　実験条件

実験に用いた音響特徴量の分析条件とパラメータを Table 1 と Table 2 に示す．また，ノードのクラスタリング手法には Newman 法 [7] を用いており，予めエッジは k-nearest neighbor (k=5) 基準で枝刈りを行ってる．今回は以下の 3 種類のノード間の距離尺度について検討を行った．

- 従来手法 1：DTW + ケプストラム歪み [1]
- 従来手法 2：SSM + ユークリッド距離 [2]
- 提案手法：構造的表象 + ユークリッド距離

4.3　結果と考察

クラスタリング結果をそれぞれ樹形図で示す (Fig. 4)．ここで樹形図のラベルにある M, F はそれぞれ男性話者, 女性話者である事を表しており，数字はノードで発声されている単語の種類となっている．DTW ベースの距離の場合は，発話内容よりも話者の違いに敏感にクラスタリングされている事が分かる．その一方でクラスタリングの距離基準に音響相対量を用いた SSM と構造的表象の場合には，男性の発話と女性の発話が比較的正しくクラスタリングされる部分も見られた．どちらの場合も/ni/や/nana/のクラスタや音韻的に近い/ichi/と/hachi/が話者の違いの影響を受けながらも樹形図上で集約されているのが分かる．しかし従来手法である SSM と提案手法を比較した場合には，提案手法の優位性は見られなかった．また，複数話者の問題は軽減する事ができたが，全体的なクラスタリングの安定性はベースラインの DTW を用いた場合が高かったが，これはこれらの特徴を用いて比較する際に絶対量によるアライメントが必要となる事がボトルネックになっている可能性が考えられる．特に現在のタスクでは前段から得られるノードの単語の一部分のみを捉える，前後の単語の一部分と連結しているという事が多く，不適切なアライメント結果が得られることが多くなる．

5　まとめ

ベースラインとして Park らの S-DTW アルゴリズムを構築し，それに話者の変化に頑健な構造的表象を組み込む事により，複数話者音声データに対する教師無しパターン発見の精度向上を試みた．一部のクラスタでは効果が見られたが，従来手法である SSM に対して構造的表象の優位であるとは言えない結果となった．

今後の予定としてはまず，大語彙音声への移行が考えられる．今回の実験では連続数字読み上げ音声を用いたが，出現する全ての語が頻出語となっていた為，ノードの抽出が安定しないという問題があった．大語彙であれば，この問題が緩和される事が期待できる．また今回の実装では構造的表象や SSM の比較ではユークリッド距離を用いたが，Muscariello らはヒストグラムを用いた SSM の比較も提案しており [2]，こちらの場合の結果との比較や構造的表象への適用を検証する必要がある．

参考文献

[1] A. Park *et al.*, IEEE Trans. audio, speech, and language processing, Vo. 16, No. 1, pp. 186–197, 2008.
[2] A. Muscariello *et al.*, ICASSP, pp. 5640–5643, 2011.
[3] 峯松他, 電子情報通信学会論文誌, J94–D, 1, pp. 12–26, 2011.
[4] Y. Zhang *et al.*, ASRU, pp. 398–403, 2009.
[5] Y. Wang *et al.*, ICASSP, pp. 8232–8236, 2013.
[6] S. Asakawa *et al.*, ICASSP, pp. 4097–4100, 2008.
[7] M. Newman, Physical review, pp. 69–74, 2004.

[4] /ichi/, /ni/, /saN/, /yoN/, /go/, /roku/, /nana/, /hachi/, /kyuH/, /zero/, /maru/の 11 種類.

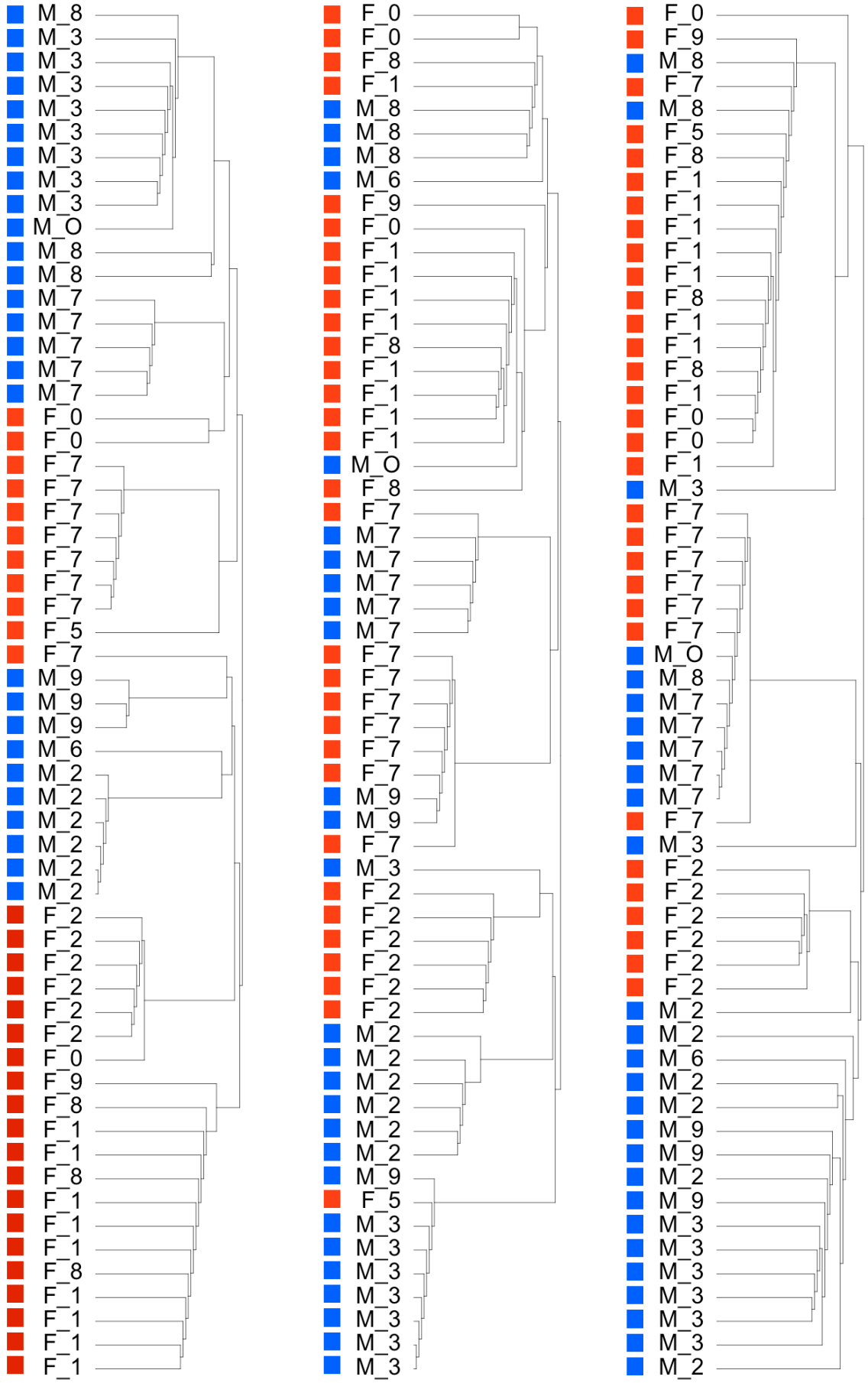

Fig. 4　左から従来手法 1（DTW ベースの距離尺度），従来手法 2（SSM ベースの距離尺度），提案手法（構造的表象ベースの距離尺度）のクラスタリング結果..