時間制約を持つクラスタリングによる連続音声の自動セグメンテーション*

◎下村直也, 朝川智, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

ラベリング情報が付与された音声コーパスは，音声認識・合成システムを含む，多様な音声処理技術の実現に必須となる場合が多い．また，明示的にラベル情報を必要としない場合でも，音声ストリームをある評価関数の下で自動区分化して処理系を組む場合もある．例えば，年齢，性別，個人性などの非言語情報が除去された音声表象が提案され [1]，連続母音認識というタスクにおいて，非常に少数の学習話者のみで，高精度の不特定話者連続音声認識を実現している [2]．この場合，音声ストリームを一旦分布系列へと変換する (類似したフレーム列を 1 ブロックとして扱う) 必要があり，前処理としての区分化が必要である．

しかしながら，大規模データに対して高精度な音素境界情報を得るには，専門家による手動ラベリングが必要であるため，莫大な時間的，経済的コストがかかる．これらを解決するため，HMM 音響モデルを用いた自動ラベリング手法 [3, 4] や，スペクトル変化を捉え音素境界と定める手法 [5]，またはベクトル量子化歪みをコスト関数とするトップダウンクラスタリングによる境界検出手法 [6] 等が提案されてきた．

本稿では学習データを必要としない，音響分析に基づくセグメンテーション手法として，音声に不可避な時系列を制約条件として反映させたボトムアップクラスタリングによる音素境界検出手法を提案する．まずはボトムアップクラスタリングの各手法の内，最も音素境界検出精度の高い手法を実験的に明らかにする．更にトップダウンクラスタリングを用いた先行研究と異なり，連続音声中に含まれる音素数を自動推定する手法についても検討する．

2 ボトムクラスタリングによる境界検出

2.1 評価用音声コーパス

頑健性の評価のために，2 つの異なる言語の評価用音声コーパスを用意する．一方は日本語読み上げ音声コーパス ATR データベースの setA，連続発声の dsa パートの女性 3 名，男性 2 名，各 115 発話，計 575 発話を用いた．計 33,607 個の境界を含む音声記号層ラベルデータ（音素レベルに相当）を使用した．他方は，米語読み上げ音声コーパス TIMIT の training データベースである．462 人の各 10 発話，計 4,620 発話で構成され，172,460 個の音素境界を有している．サンプリング周波数は，ATR が 20kHz であり，TIMIT が 16kHz である．そのため，ATR の音声データは全て 16kHz にダウンサンプリングした上で処理した．

両データベースとも，ラベリングは音声学の専門家がスペクトログラムの視察に基づき行なっている．正解ラベルの記号（音素）種類数は，ATR が 51 であり，TIMIT が 61 である．ATR の音声記号層は発声を母音部と子音部に分割して記述しており，境界が決定できない場合は分割は行なわれていない．そのため TIMIT より若干粗いラベリングであると言える．

2.2 時間制約条件付きクラスタリング

連続発声された音声の各音素内には，スペクトルの安定区間が存在する．本稿ではその安定区間を捉えるために，音響的に類似した連続するフレーム同士をマージする形で纏め上げ，音声ストリームの大局的な階層構造を抽出することを考える．この構造抽出の結果，副産物として音声ストリームの区分分割，即ち，音素境界位置が得られる．ただし，全フレームに対する距離行列を求め，通常のボトムアップクラスタリングを行なうと，時間的に離れたフレーム（クラスタ）がマージされることが頻繁に起こる．このようなマージ操作は本タスクにおいては意味を成さないため，時間的に連続する 2 クラスタのみをマージ対象とした制約付きクラスタリングを考える．

ボトムアップクラスタリングの手法として，最短距離法，最長距離法，群間平均法，重心法そして Ward 法が一般に知られている．本稿ではそれらの手法に対して時間制約を課し，音素境界を最も精度よく抽出可能である手法を実験的に明らかにしていく．ここではクラスタリングの時系列制約という問題を，Ward 法に適用した例を取り上げて解説を行う．

Ward 法はユークリッド距離に基づくクラスタリングであり，2 つのクラスタをマージした際の「群内偏差平方和の増加量」を非類似度と定義している．この非類似度が最も低い（即ち最も類似している）クラスタ同士をマージさせ，最終的には 1 つのクラスタへと纏め上げる．今，クラスタ p とクラスタ q をマージして新しいクラスタ $r(=p\cup q)$ を作ることを考える．特徴

*Automatic segmentation of continuous speech based on time-constrained bottom-up clustering.
by Naoya Shimomura, Satoshi Asakawa, Nobuaki Minematsu, Keikichi Hirose (Univ. of Tokyo)

Table 1　音響分析条件

サンプリング	16bit / 16kHz
フレーム幅	32msec
窓	ハミング窓 32msec
フレームシフト長	10msec
音響パラメータ	MCEP 1〜12 次元

量として m 次元のベクトル $(x_1, x_2, \ldots, x_m)$ を考え，p の全サンプルに対する重心を $(\overline{x}_1^{(p)}, \overline{x}_2^{(p)}, \ldots, \overline{x}_m^{(p)})$ とすると，p の偏差平方和 $E(p)$ は

$$E(p) = \sum_{i=1}^{n_p} \sum_{j=1}^{m} (x_{ij}^p - \overline{x}_j^p)^2 \quad (1)$$

となる．n_p は p のサンプル数である．p，q をマージし，r を生成した際の偏差平方和の増分 $\Delta E(p,q)$ は

$$\Delta E(p,q) = E(r) - E(p) - E(q) \quad (2)$$

となる．各段階でクラスタのマージによる偏差平方和の増分 $\Delta E(p,q)$ が最小となる p と q をマージする．ただし，上記した様に時間制約を設ける．ある段階でのクラスタを時間軸に沿って $\{p_0, p_1, \ldots, p_t, \ldots, p_T\}$ とおくと，(2) 式は下記となる．

$$\Delta E(p_t, p_{t+1}) = E(p_t \cup p_{t+1}) - E(p_t) - E(p_{t+1}) \quad (3)$$

即ち，連続する 2 クラスタに対して，偏差平方和の増分が最小となるクラスタを対象としてマージする．この時，より類似した連続 2 フレームをマージするのではなく，より類似した連続 2 クラスタをマージする点，即ち，フレームの部分系列をより大きな纏まりとして捉える手法である点に注意すべきである．

本手法の計算コストは初期フレーム系列長 N に対して $O(N^2)$ である．しかし，連続音声を無音区間ごとに区切れば N を小さくすることは十分可能である．

2.3　分析条件

分析条件を Table. 1 にまとめる．スペクトル変化を捉える音響特徴量として，聴覚特性を考慮したメルケプストラムを用いる．本稿では，スペクトルの安定区間を纏めることで音素境界を得ることを検討しているため，パワー（MCEP の 0 次項）は用いなかった．しかし，強勢，弱勢の存在する英語ではパワーを用いることでの精度向上が予備実験により確認されている．

また，本節では各発話に含まれる音素数をクラスタ数として与えている．音素数を自動推定し，制約付きクラスタリングを自動的に中止する方法については次節で述べる．

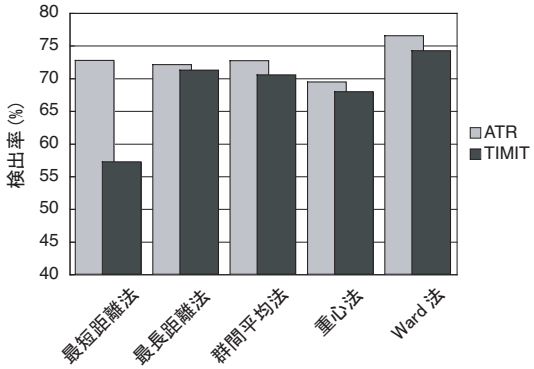


Fig. 1　各手法による音素検出精度の違い

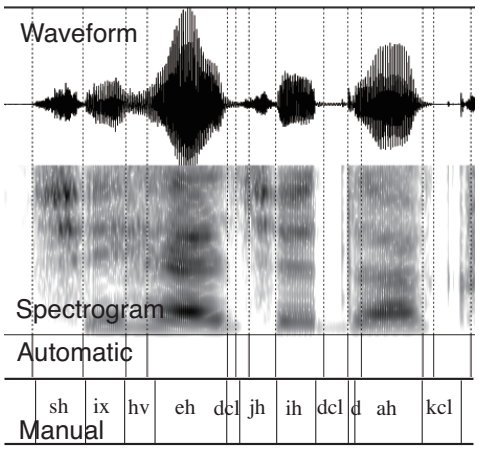


Fig. 2　Ward 法によるセグメンテーション結果の一例

2.4　各クラスタリング手法の境界検出精度比較

Fig. 1 は ATR 及び TIMIT データに対して，各クラスタリング手法によって（音素境界数を与えながら）音素検出処理を行なった結果である．自動で検出された境界（以後，検出境界と呼ぶ）とラベラーにより定められた境界（以後，正解境界と呼ぶ）との絶対時間誤差が 30msec 以内となる境界検出率により評価を行なっている．2 つの言語共に，音素境界検出というタスクにおいて Ward 法を用いることが適切であることが実験的に求まった．また Ward 法を用いる方法論は，Svendsen らによって提案されたトップダウンクラスタリングによる音素境界検出 [6] と，『コスト関数にベクトル量子化歪みを用いている』という点で一致している．今後記載している実験は全て Ward 法のみを用いて行なっている．

また，Fig. 1 で TIMIT より ATR の方が検出精度は高い．ATR のラベリングはスペクトログラム特徴に基づいて行なわれるが，境界を決定できない区間は無理に分割を行なっていない．そのことから ATR の方が纏まりを捉える本手法と，より近い処理をしていると考えられる．

Ward 法による TIMIT データベースに対して自動

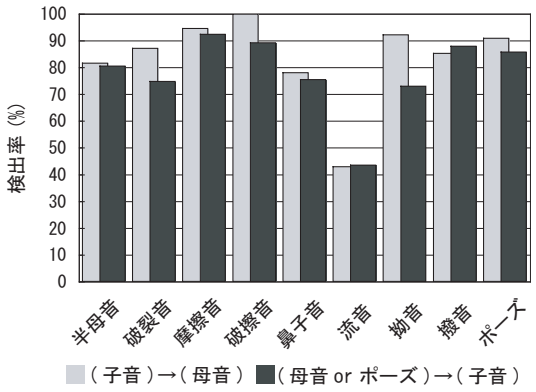


Fig. 3　日本語の各音素接続間の検出率

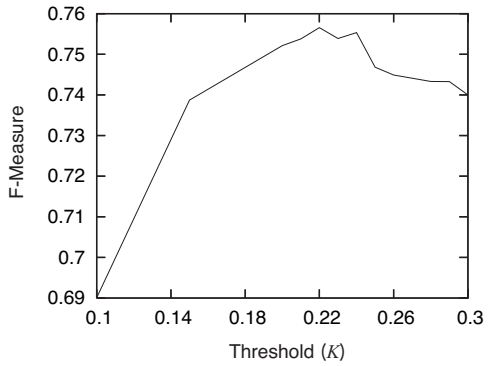


Fig. 4　閾値 K に対する F 値

で音素境界を検出した結果の一例を Fig. 2 に示す．原波形，スペクトログラム，自動検出結果及び手動でのラベリング結果を併記する．/jh/及び/kcl/の音素中に誤挿入が存在することが確認できる．次に，各音素毎に異なると予想される検出率を定量的に評価し，本手法の特徴を洗い出していく．

2.5　各音素の境界検出精度

Fig. 3 は，ATR データベースに対する各音素間接続の検出実験結果である．モーラ構造をもつ日本語において，摩擦音や破擦音等の母音と大きく異なるスペクトル構造 (白色雑音性) をもつ音素は，その安定区間が母音と明確に異なるクラスタとして纏め上げられる傾向にあるため，境界が検出されやすいことがわかる．また，本稿ではパワーを特徴量として用いていないものの，破裂音に伴う休止区間 (/kcl/，/tcl/ 等) や，ポーズ等の無音区間も同様の理由から比較的検出されやすい．逆に，鼻子音や流音は顕著に検出精度が低くなっていることが実験的に検証された．興味深い事に，流音と撥音以外の音素では，(母音 or ポーズ) → (子音) という音素接続により精度が劣化している．このことから，音素単位以上に纏め上げを継続しても，モーラ単位での境界検出は困難となることが考察される [7].

3　音素数自動推定手法

3.1　閾値によるクラスタリングの自動停止

第 2 節ではトップダウンクラスタリングにより境界を検出する先行研究 [6] と同様に，発声に含まれる音素数を既知として与えていた．本節では，音素数を自動的に推定する手法について述べ，その自動推定結果の妥当性を実験的に示す．提案手法が基本的に階層的ボトムアップクラスタリングであることを考慮し，Ward 法の更新コスト（群内偏差平方和の増加量）に着眼することで実現する．以下の考察から，閾値処理により音素数を自動で決定すること，即ち，クラスタリングを自動停止することを検討する．

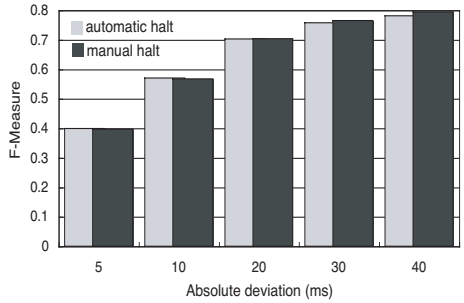
Ward 法では，クラスタ p，q に対して (2) 式の距離尺度を定義している．各段階で 2 クラスタのマージによる群内偏差平方和の増分 $\Delta E(p,q)$ が，最小となるクラスタをマージしようとする．ここで，ある段階において，各クラスタが凡そ各音素に対応している状態を考える．この場合次の操作で，異なる音素が強引にマージされることになる[1]．ある話者が生成する各音素の音響特徴を考える．この場合，任意の 2 音素間距離（重心間距離）の最小値は，凡そ話者非依存であると仮定する．その結果，どの話者の発声した音声であっても，音素に対応した形でクラスタリングされた状態に対する次のマージ操作は，比較的大きな更新コスト（群内偏差平方和の増加量）を呈するはずである（$E(p\cup q) >> E(p), E(q)$）．以上の考察から $\Delta E(p,q)$ に対応する閾値を実験的に定め，これを用いてクラスタリングの自動停止を検討する．このような手法はスペクトル遷移を局所的に走査する先行研究 [5] では困難な方法論である．

Fig. 4 は TIMIT の test データベースから無作為に 30 発話を選択し，$\Delta E(p,q)$ に対する閾値 K に対する絶対誤差 30msec 以内の F 値を求めた結果である．図より $K = 0.22$ の時，最大値（0.757）となった．ATR データも含め，閾値 $K = 0.22$ を用いてボトムクラスタリングを自動停止させる．

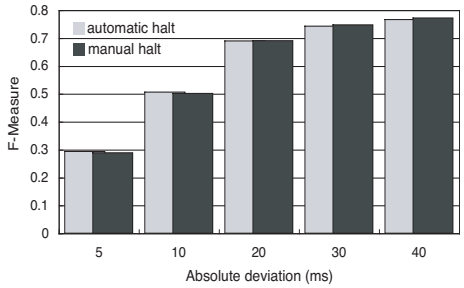
3.2　提案する音素数自動推定手法の妥当性の検討

open データである TIMIT の training データの各発声に対する自動推定音素数と正解音素数の相関を求めると，相関係数 0.84 が得られ，強い相関が確認できた．

[1]時間的に連続する 2 クラスタのみをマージ対象としていることに注意．時間的に離れた 2 クラスタを対象とすれば，非常に類似性の高い 2 クラスタ（2 音素）がマージ対象となる．



(a) ATR



(b) TIMIT

Fig. 5　音素境界を自動停止した場合と，音素境界数を与えた場合の比較

また，Fig. 5 は閾値処理によりクラスタリングを自動停止させた場合の，音素境界検出精度と，正解音素数を与えた場合の境界検出精度を F 値で比較した結果である．正解音素数を与えたとき，F 値は Precision(適合率) 及び Recall(再現率) と一致する．クラスタ数を自動推定しても音素数を与えた場合とほぼ同等の結果を出していることが分かる．これらの結果から，閾値処理を用いることで，F 値ベースの評価においては事前知識を用いずとも音素数を与えた場合に近い出力が可能であり，提案手法は妥当性の高い手法であると言える．また，同一の閾値を異なる言語において用いたが，相応の結果であるため，定めた閾値は高い言語非依存性をもつことがいえる．

4　まとめ

自動音素境界検出というタスクを成すために，音声時系列上で隣接するクラスタのみをマージ対象とする制約条件付きボトムアップクラスタリングを提案した．クラスタリング手法を 5 つ紹介し，実験的に Ward 法が 30msec 以内の誤差での音素検出率が約 75% で，最も高いことを明らかにした．さらに，自動で音素数を推定する手法についても検討し，音素数を既知とした場合とほぼ同程度の F 値約 0.75 を得た．

また，音響モデルを用いない先行研究として，スペクトル特徴量に基づく検出手法 [5] においては F 値約 0.7，トップダウンクラスタリングによる検出手法では音素検出率約 75%という結果が報告されている [6]．前者においては精度で，後者に対しては事前知識として音素数を与える必要がない点で，本手法は優位性をもつ．

しかしながら，読み上げ音声を対象とした場合，HMM による強制切り出しに基づく手法では，約 95% の音素を 30msec 以内の誤差に検出することが可能であり，精度では遠く及ばない．その意味で本手法は，音響分析を基本とするセグメンテーションが効果的に利用できる場面を具体的に検討することは必須である．例えば，感情の入った音声や歌声等の特殊な発声，非母国語話者の音声，幼児音声，動物や虫の鳴き声など，そもそもデータの入手が困難なデータのラベリングに利用することが考えられる．また，近年，音声に不可避的に混入する非言語的特徴を削ぎ落とした形で音声を表象し，種々の応用に用いる研究が行なわれている [1, 2]．この話者／マイク不変の表象は，音声ストリームを分布系列へと変換する処理が必須であり，今後，制約付きクラスタリングの利用可能性について検討することを考えている．

参考文献

[1] N. Minematsu et al., “Theorem of the invariant structure and its derivation of speech Gestalt,” Proc. Int. Workshop on Speech Recognition and Intrinsic Variations, pp.47-52, 2006.

[2] S. Asakawa et al., “Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics”, Proc. EUROSPEECH, 2007. (accepted)

[3] H. Lo, H. Wang, “Phonetic boundary refinement using support vector machine”, Proc. ICASSP, pp.933-936, 2007.

[4] J. Keshet et al., “Phoneme alignment based on discriminative learning”, Proc. ICSLP, pp.2961-2964, 2006.

[5] S. Dusan, L., Rabiner, “On the Relation between Maximum Spectral Transition Positions and Phone Boundaries”, Proc. ICSLP, pp.645-648, 2006.

[6] T. Svendsen., F. K. Soong., “On the automatic segmentation of speech singnals”, Proc. ICASSP, pp.77-80, 1987.

[7] 下村直也他., “制約条件つきクラスタリングによる連続音声からのイベント境界検出”, 信学技報, SP2007-12, pp.25-30, 2007.