社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

制約条件付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

下村　直也† 　朝川　智† 　峯松　信明† 　広瀬　啓吉††

† 東京大学大学院新領域創成科学研究科 〒277–8561 千葉県柏市柏の葉 5–1–5
†† 東京大学大学院情報理工学系研究科 〒113–0033 東京都文京区本郷 7–3–1
E-mail: †{shimo,asakawa,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　音声技術の高度化を図る場合，音素境界ラベルが付与された音声コーパスが要求されることがある．通常，HMM などの音響モデルを用いた強制切り出し（forced alignment）を行なうことが多いが，この場合，学習データと切り出しデータの発話スタイルが異なると精度は劣化する．一方，音響モデルを必要としない，音響分析のみに基づいた（例えばスペクトル遷移最大点に着眼した）音素境界抽出も試みられている．この場合，学習データが存在しないため，上記の問題は原理的に回避できる．本稿では，時間制約を持たせたボトムアップクラスタリングにより，音響モデルを用いずに，音声イベント境界を検出する手法について検討する．その際，入力音声に含まれる音素数を自動推定する方法，及び，提案手法がイベント群の階層構造を推定することに着眼し，音素やモーラなど異なる言語単位での境界検出の可能性について考察する．更に先行研究と比較することで，本手法の頑健性を示す．
キーワード　音素境界検出，クラスタリング，時間制約，頑健性，階層構造

Detection of acoustic event boundaries from continuous speech based on constrained clustering

Naoya SHIMOMURA†, Satoshi ASAKAWA†, Nobuaki MINEMATSU†, and Keikichi HIROSE††

† Grad. School of Frontier Sciences, Univ. of Tokyo, 5–1–5, Kashiwanoha Kashiwa, Chiba, 277–8561 Japan
†† Grad. School of Info. Sci. and Tech., Univ. of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo 113–0033 Japan
E-mail: †{shimo,asakawa,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　Speech databases with accurate phoneme labeling are often required to improve speech technologies. Although HMM-based forced alignment is widely used, the performance easily decreases when the input speech data has a different speaking style from that of the speech data used for training HMMs. An alternative is a speech segmentation method only based on acoustic analysis. For example, the speech segmentation based on measuring spectral transition was proposed previously. In this report, another analysis-based method is proposed which uses constrained clustering on a time series of frames. Some discussions will be done on estimation of the number of phonemes and detection of mora boundaries. Finally, by comparing the phoneme boundary detection performance between the previous method and the proposed method, the higher robustness of our proposal is shown.
Key words　phoneme boundary detection, clustering, temporal constraint, robustness, hierarchical structure

1. はじめに

音素，モーラ，音節などの言語単位の時間境界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不可欠なものである．しかしながら，大規模データに対して高精度なイベント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，従来，多くの自動セグメンテーション手法が提案され，利用されている．

特に HMM 音響モデルを用いた自動ラベリング手法は多く報告されている [1]〜[3]．HMM ベースの手法は高い精度を示すものの，事前学習が必要であり，学習に用いた音声データと同一発話スタイルのデータに対してのみ，高精度な処理が可能となる．例えば，非母国語者の音声は，その話者の母国語，及び学習対象言語の音響モデルのみでは精度に限界があり [4]，また，感情の入った音声や歌声等の特殊なデータに対しては読み上げ音声から構築された音響モデルでは精度が落ちる．更には，音声データベースの構築が困難な（つまり，音響モデル構築が困

難な）幼児音声のラベリングなど,「事前に大量の音声データが用意できない」状況に遭遇することは珍しいことではない．適応技術によりデータに対して頑健な処理を行なう研究が報告されているものの，根本的な解決には至っていない [5] [6]．音声認識技術は大量の音声データに対する数理統計的な手法によってその精度向上を実現してきたが，これは逆に，大量のデータを準備できない問題に対しては，その精度向上が困難とならざるを得ない方法論であると言える．

また，音響モデルが「音」のモデルである以上，それは年齢，性別などの非言語的情報に依存したモデルとなる．近年，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 [7]，連続母音認識という簡素なタスクではあるが，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装している [8]．この場合においても，音素やモーラなどの言語単位に準拠する必要性は無いが，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，前処理としてのセグメンテーションが必要である．

一方，事前に用意された音響モデルを使用せず，音響分析のみによってセグメンテーションを行なう手法が提案されている．この場合，事前データを必要としないため，学習データ依存性は原理的に無い(注1)．例えば [9] では，音響分析により求まったスペクトル遷移のピークを音素境界候補とし，後処理による候補削減を通して，高精度の音素境界検出を報告している．このような音響分析のみに基づく方法論は，明示的に音素書き起こしを与えないため，挿入／削除誤りは，HMM を用いた音素アライメントに比較すれば格段に増加するが，検出された境界の時間的精度は高い値を示す．また，上記したような学習データの依存性に対して高いロバスト性が期待できる．

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，クラスタリングによるイベント境界検出手法を提案する．単にクラスタリングを行なうのではなく，音声に不可避な時系列としての制約を反映させたクラスタリングを提案する．先行研究 [9] と異なり，クラスタ数を増減することによって対象とするイベントの「粒度」を制御可能である．この特性を活かし，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し，更には，異なる言語単位に対する境界検出精度についても実験的に考察する．また，異言語音声に対する境界検出の頑健性についても報告する．

2. 制約付きクラスタリングによる境界検出

2.1 制約付きクラスタリング

異なる音素が連続的に発声された時，各音素間にはある音響事象から異なる音響事象への遷移が存在する．多くの場合，この遷移はスペクトル包絡の変化として観測される．音響モデルを用いない音素セグメンテーション手法として，スペクトル遷移のピークを抽出することで音素境界を検出する手法が提案されている [9]．スペクトル遷移の定量化を Δ ケプストラムの絶対値として行ない，この値が前後フレームより大きい時点を音素境界候補時点とし，その後，後処理による候補削減を行ない最終結果としている（後処理については後述）．以降，この手法を MST（Maximum Spectral Transition）法と記す．

さて，連続音声の各音素内に注目すると，逆に，スペクトルの安定区間が存在する．MST 法はスペクトルの局所的変化を積極的に検出する方法であるが，本稿では，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（ボトムアップクラスタリング），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，イベント境界位置が得られる．

この場合，注意すべきことがある．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．

本稿で用いるボトムアップクラスタリングは Ward 法 [10] であり，これに対して時間制約を課す．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

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

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

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

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．

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

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．

本手法は，初期フレーム系列長が N の場合，最終的に N 段の階層構造（樹形図）を生成する．この階層構造は N 通りの粒度におけるセグメンテーション結果を提供する．Ward 法では樹形図の縦軸は「偏差平方和の増分」の総和となり，これは，サンプル群を，与えられたコードブックサイズでベクトル量子

(注1)：但し，種々の閾値を必要とする場合，閾値決定は使用したデータに依存する可能性がある．

表 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
帯域	fullband

化した際の量子化歪みに相当する．結局，許容する量子化歪みを与えれば，それに対応する粒度で階層構造を提供し，結果的に，対応する粒度のイベント境界位置を提供する．複数の粒度を考えた場合，それが，音素やモーラなど複数の言語単位の境界に相当するか否かについては，実験的に検討する．

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．

2.2　使用した音声コーパス

定量的な解析・評価は，クラスタリングにより自動検出された音素境界と，手動ラベリングによる音素境界とを比較することで行なう．2 つの異なる言語の音声コーパスを用意した．一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(注2)．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．他方は，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．サンプリング周波数は，TIMIT が 16kHz であり，ATR は 20kHz 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，TIMIT が 61 であり，ATR が 51 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．

2.3　分 析 条 件

分析条件を表 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 [11] 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．

2.4　音素境界の自動検出とその精度

TIMIT データベースに対して，自動で音素境界を検出した結果の一例を図 1 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．本手法の特徴として/k/や/t/等の破裂音に伴う休止区間/kcl/，/tcl/，またはポーズは検出されやすい．逆に母音が二つ続いた際はその 2 母音間の境界を検出しにくい(注3)．図 2 は TIMIT と ATR の全データに対して（音素境界数を与えて）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）との絶対時間誤差の程度に応じた検出率により評価を行なっている．共に 20msec 以内の誤差で 70%，30msec 以内の誤差で 75%程度の精度となっている．また，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．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．

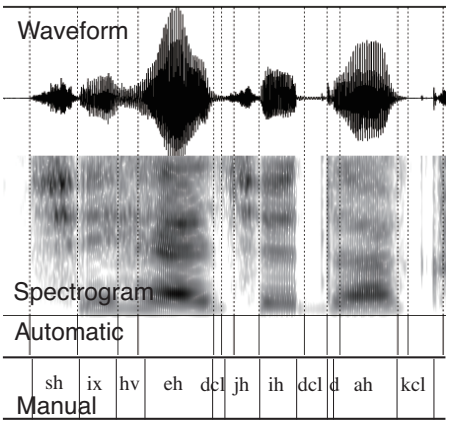


図 1　自動セグメンテーション結果の一例

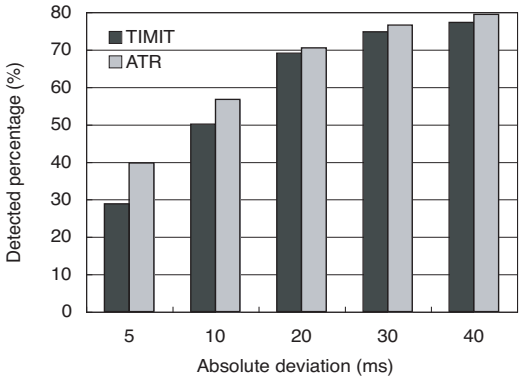


図 2　自動セグメンテーション結果

なお，音素数のみならず，個々の音素の種類，及び各音素の音響的特徴を，大規模音声データベースから計算される統計量としてシステムに与えた場合（即ち，HMM による強制アライメント），種々の後処理の導入により，20msec の誤差で，約 90%の精度が報告されている [1]〜[3]．音響分析に基づく音素境界抽出は，先行研究を含め，通常の読み上げ音声を対象とすれば，HMM による強制切り出しに基づく方法論とは比較できないほど精度は悪い．本稿ではまだ十分に検討していないが，

(注2)：先行研究 [9] が，TIMIT の training データを評価データとしているため，本研究でも評価実験時にこれを用いた．

(注3)：英語の二重母音（diphthong）を考えれば，この方がより正しいラベリングを行なっていると解釈することもできる．

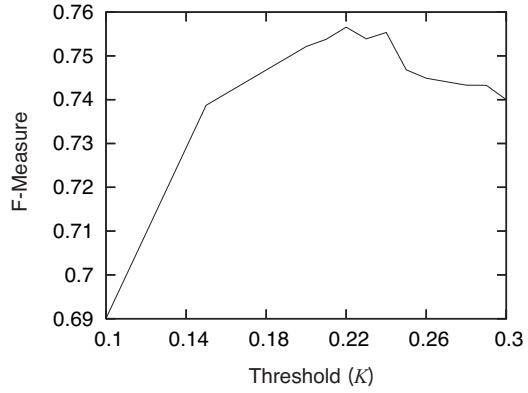


図 3　閾値 K に対する F 値

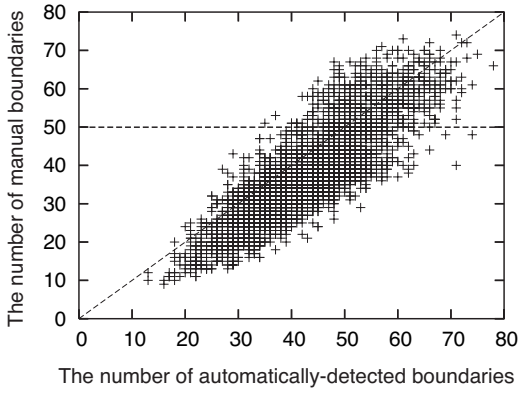


図 4　自動推定音素数と正解音素数との関係

音響モデルの学習データとその音響的特性が大きく外れた音声（例えば乳児が泣き叫ぶ音声，笑い転げる音声）などを本来は対象とすべき手法であろう．今後早急に検討する予定である．なお，次節以降では「読み上げ音声」を対象として，提案手法の各種特性を先行研究と比較しつつ，実験的に検討する．

3.　音素数自動推定手法

第 2 節では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，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，音素数を自動的に推定する手法について検討する．そして，自動推定結果の妥当性を検討した上で，先行研究と比較し，本手法の頑健性を示す．

3.1　閾値によるクラスタリングの自動停止

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考える．ここでは，以下の考察から，更新コストに対する閾値操作を検討する．

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(注4)．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 [9] では困難な方法論である．

図 3 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．今後，ATR データも含め，閾値 $K = 0.22$ を用いてクラスタリングを自動停止させる．

(注4)：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．

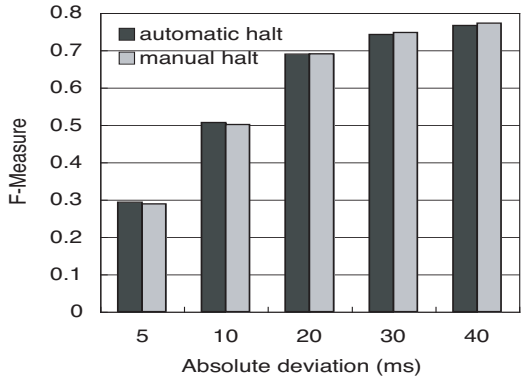
3.2　提案する音素数自動決定手法の妥当性の検討

open データである TIMIT の training データの各発声に対して音素数推定実験を行なった．自動推定音素数と正解音素数の関係を図 4 に示す．自動推定音素数と正解音素数の相関係数は 0.84 となり，強い相関が確認できた．しかし，音素数推定という観点から見ると，正解音素境界数 50 に対して，検出境界数は 40 から 65 ほどの幅があり，検出漏れや過検出も多い．前説で述べた，「乳児が泣き叫ぶ音声，笑い転げる音声」などを対象とした場合，明確な音素書き起こしが単にラベラーの音韻化想像力を示すのみとなり，意味を持たないことも多い．これらの音声を対象とした音声区分化ツールを構築する場合，クラスタリングの更新コストに対する閾値をスライダー指定する GUI を用意し，ラベラーが検出境界の是非を判断しながらラベリング初期値を決定するような場面では，応用可能性がある．

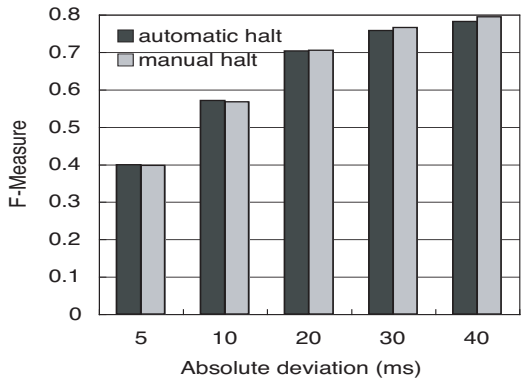
図 5 は閾値処理によりクラスタリングを自動停止させた場合の，音素境界検出精度と，正解音素数を与えた場合の境界検出精度を F 値で比較した結果である．正解音素数を与えたとき，F 値は Precision(適合率) 及び Recall(再現率) と一致する．クラスタ数を自動推定しても音素数を与えた場合とほぼ同等の結果を出していることが分かる．これらの結果から，閾値処理を用いることで，F 値ベースの評価においては事前知識を用いずとも音素数を与えた場合に近い出力が可能であり，提案手法は妥当性の高い手法であると言える．よって，この閾値をマニュアル操作しながら適切な音声区分化の初期値を求めることは可能であると考えられる．

3.3　スペクトル遷移に着眼した先行研究との比較実験

前節までで，制約条件付きクラスタリングによる音素境界検出手法を音素境界数を与えることなく実行する手法へと拡張した．本節では，音響分析に基づく（事前学習を閾値設定以外では必要としない）先行研究である，MST 法との比較を行なうことで提案手法の有効性を確認する．MST 法とはスペクトルの遷移度（Δ ケプストラムの絶対値）がピークとなる時点を音素境界候補とし，後処理で候補を適宜削減する手法である [9]．ここで言う後処理とは，具体的には次のものを指す．

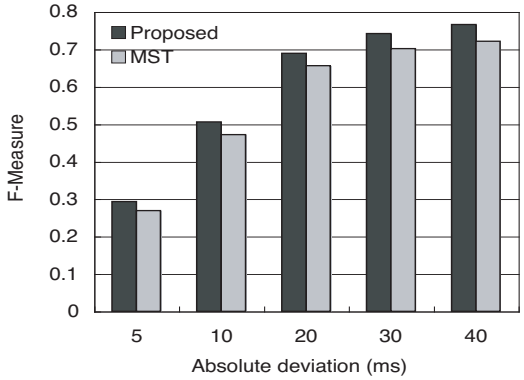


(a) TIMIT

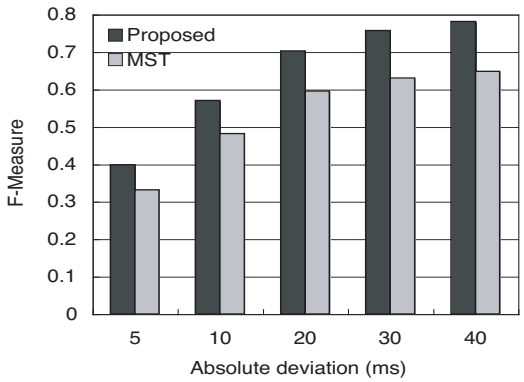


(b) ATR

図 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



(a) TIMIT



(b) ATR

図 6　スペクトル遷移に着眼した先行研究との比較

- あるピークと，時間的に隣り合っている遷移度との差分が，ある閾値以下の場合そのピークを除外する．この閾値は発話の最大ピーク値の 1%とする．
- あるピークに対して，その前後に存在する “谷”に着目する．ピークと谷との差分がピークの 10%以下なら，そのピークは排除する．

以上の後処理の結果残ったスペクトル遷移ピークを，最終的な音素境界候補として出力している．正解音素境界数など一切与えていないため，40msec を許容誤差とした場合，挿入率 28.2%，脱落率 18.0% の精度を報告している．本稿では [9] の手法を，論文に沿って可能な限りそのアルゴリズムを忠実に実装した．なお, [9] で使用した特徴量は MFCC1〜10 次元であり，この点を除いて分析条件は同一である．

TIMIT と ATR に対して比較実験を行なった結果を図 6 に示す．2 つのデータベース共に提案手法により精度が改善されている．特に ATR に関しては，MST 法で精度が大きくて低下している．これには MST 法の後処理（特に閾値）が TIMIT に依存していることを強く示唆する結果である．それとは対照的に，本手法で提案した閾値は TIMIT データを参照して求められたにも拘らず，ATR に対して高い精度を呈している．本稿で提案する手法の言語依存性の低さを示していると考えている．

4.　様々な言語単位に対する境界検出

第 2 節で述べたように提案手法はボトムアップクラスタリングに基づいており，段階的な階層クラスタリングが結果として得られる．即ち，異なる粒度の音声区分化が得られる．本節では，ATR データベースに付属する複数の粒度の正解ラベルと，提案手法で得られる自動境界検出結果との比較を行なう．

4.1　様々な粒度の音声イベントラベル

ATR データベースに付属する複数粒度の正解ラベルを参照して，イベント層ラベル，TIMIT-level ラベル，音声記号層ラベル，モーララベルを定義した．各粒度に対して含まれる境界数を表 2 に表記する．音声記号層ラベル，イベント層ラベルは ATR に元から付属されているラベルである．前節までの実験では音声記号層ラベルのみを用いていた．イベント層ラベルとは，スペクトログラム特徴の変化に応じて音声記号層のセグメント内を更に複数の状態に分割したラベルである．休止区間，母音の開始部，終了部の過渡現象，及び何らかの原因でスペクトルパターンに乱れが生じている区間に境界を新たに設けている．モーララベルとは，音声記号層をモーラ単位に纏めて作成したラベルである．但し，音声記号層で区分化が不可能と判断されたラベルはそのままにしている(注5)．TIMIT-level ラベルとは，イベント層を TIMIT のラベリングの粒度と同様になるよう，纏めたラベルである．{>}, {<}, {∗ >} 等の母音過渡区間を各母音に吸収させて生成した．

4.2　様々な粒度に対するイベント境界の検出実験

実験は ATR データベースの各イベントラベルに対して，正解境界数を既知として行なった．図 7 はラベル粒度を変化させ

(注5)：{k,a,u} や {s,u} 等の表記部分であり，専門家がスペクトログラム上で分割できなかった区間である．

表 2　各粒度のイベントラベルの境界数

粒度	境界数（個）
モーラ	20,621
音声記号層	33,607
TIMIT level	39,356
イベント層	46,212

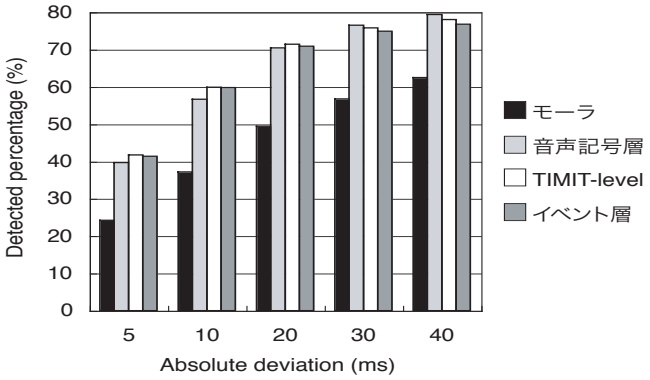


図 7　各粒度のイベント境界に対する境界検出結果

た時の検出結果である．この結果より，音素単位もしくはそれよりも細かい粒度に対する境界推定は可能であるが，より大きいモーラ単位でのセグメンテーションでは精度が落ちる結果となった．その理由として，音素単位以上でのクラスタリングでは母音同士，または，母音+半母音+母音など，スペクトル形状の近い音素同士のマージが優先され，/k/や/t/等の破裂音に伴う休止区間等が最後まで残されてしまうことが理由の一つであると考察される．しかし，例えば英語音声学の中には/aʊɚ/（tower の t 以降の発声）を三重母音として認める考え方があるなど，自動ラベリング結果が必ずしも誤りであると言い切れない場合もある．そもそも日本人であれば英語の二重母音は，二つの異なる母音連鎖と感覚するのに対し，英語母語話者は一つの母音として感覚する．このような事実を考えれば，絶対的な正解ラベルを用意することの是非すら問われかねない．既存のデータベースに付与されている「ある知覚特性を持った人間が恣意的に付与したラベル」を正解とした評価よりも，本手法が呈する境界情報の利用価値を認める応用場面を模索した方が，意味のある評価となるのかもしれない．

しかし，本稿で検討した分析条件を変更することで，モーラ境界検出の精度を上げることも可能であろう．例えば，最終的に得られる境界に対して，境界間時間長の最小値を設ける，各クラスタの音声事象が有する全エネルギーをある範囲内に収める，など種々の実験的検討が可能である．或いは，音素レベルまでクラスタリングを行なった後，パワーの大小により母音を表すクラスタをまず特定する．日本語においてはモーラに含まれる子音は 1 つであるから，母音と推定された各クラスタの前のクラスタがもし母音と推定されればばマージせず，母音でないと推定された場合は対応づけるといった後処理によりモーラ境界を検出することも可能であろう．

5. ま と め

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クラスタリングを提案した．本手法はスペクトル成分を特徴量として使用することで，同一音素内フレームの纏め上げが可能であり，音素（音声イベント）セグメンテーションとしての機能を取り上げ検討した．日本語，英語という音声学的にはかなり異なる二言語の音声データベースに対して，先行研究以上の精度（30msec 以内の誤差で，F 値は約 0.75）の値を得た．またクラスタリングの自動停止も，更新コストに対する閾値処理として実装し，その頑健性も実験的に示した．推定音素数と正解音素数の相関係数は約 0.84 であった．

しかしながら，読み上げ音声を対象とした場合は，HMM による強制切り出しに基づく手法には遠く及ばない．その意味で，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．更には，音素よりも大きな言語単位での切り出し精度の向上も，種々の制約，分析条件の最適化などを用いて検討する予定である．

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 [7], [8]．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，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．

文　献

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

[2] J. Keshet, S. Shalev-Shwartz, Y. Singer, and D. Chazan, “Phoneme alignment based on discriminative learning”, Proc. ICSLP, pp.2961-2964, 2006.

[3] J. -W. Kuo and H. -M. Wang, “Minimum boundary error training for automatic phonetic segmentation”, Proc. ICSLP, pp.1217-1220, 2006.

[4] Y. Tsubota, T. Kawahara, and M. Dantsuji, “Recognition and verification of English by Japanese students for computer-assisted language learning system”, Proc. ICSLP, pp.1205-1208, 2002

[5] 菊池英明, 前川喜久雄, “自発音声に対する音素自動ラベリング精度の検証”, 春音講論, 2-5-12, pp.97-98, 2002-3.

[6] 米澤朋子, 水野秀之, 阿部匡伸, “HMM 音素モデルによる自動ラベリングのロバスト性の検討”, 信学技報, SP2002-74, pp.17-22, Aug 2002.

[7] N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, “Theorem of the invariant structure and its derivation of speech Gestalt,” Proc. Int. Workshop on Speech Recognition and Intrinsic Variations, pp.47-52, 2006.

[8] S. Asakawa, N. Minematsu, and K. Hirose, “Automatic recognition of connected vowels only using speaker-invariant representation of speech dynamics”, Proc. EUROSPEECH, 2007. (accepted)

[9] Sorin, D., Lawrence, R., “On the Relation between Maximum Spectral Transition Positions and Phone Boundaries”, Proc. ICSLP, pp.645-648, 2006.

[10] 宮本定明, “クラスター分析入門 ファジィクラスタリングの理論と応用”, 森北出版, 1999.

[11] 音声信号処理ツールキット SPTK. http://www.sp.nitech.ac.jp/˜tokuda/SPTK/index-j.html