対話音声における類似音声パターンの発話テンポ*

○芦村 和幸 (JST/CREST)　ニック・キャンベル　(ATR, JST/CREST)

1　はじめに

人は，日常会話において，文字には表われない情報を韻律や声質などによって示すことにより，同じ言葉を意図や態度に応じたさまざまな意味に使い分けている [1].

発話様式と，意図や態度との関係を解明するためには，大規模自然音声対話データベースの収集と分析が必要不可欠であるが，大規模音声データのテキスト書き起こしを全て人手により作成することは，量的・質的に困難である.

本稿では，テキスト情報に依存せず音声情報のみから対話中に繰り返し出現する類似音声パターンを自動抽出する方法，および同手法を利用して算出する新しい発話テンポについて提案する.

2　音声試料

音声試料として，JST/CREST 発話様式プロジェクトで作成中の自由対話データベース [1, 2, 3, 4] を用いた.

このデータベースは，以下のような特徴を持つ.

- 電話を用いた自由対話
- 特定の女性話者 1 名
- 2 年間にわたって継続的に収録 (収録時間: 250 時間程度)
- 1 収録あたり 6 分 ～ 30 分程度の対話
- 多様な対話者 (父母，配偶者，子供，親戚，友人，他人)

今回の分析にあたっては，全対話音声データ 1,412 件のうち，発話の開始・終了時間が付与されているもの 792 件を対象に，145,152 発話に分割した上で分析を行なった．なお，提案手法の処理は音声波形情報と発話の開始・終了時間のみにもとづいて行なわれるため，書き起こしテキスト情報は不要である.

3　類似音声パターン自動抽出手法

近年，対話音声データ分析への音声認識技術の応用が試みられているが，対話データ中には，語彙辞書や言語モデルに登録されていない音声パターンが多く含まれ，認識精度低下の一因となっている．対策として，書き起こしテキストにもとづいて語彙を辞書や言語モデルへ追加することが考えられるが，対話の内容は話者や状況により異なるため，収録データ量を単に増やすだけでは，網羅的な語彙セットを得られる保証はない．また，対話音声においては，長音や促音の混入，音素の欠落などの発話変形が多くみられ，認識に適した発音情報を書き起こしテキストにより常に表現できるとは限らない．例えば，従来，発話テンポは一発話ごとに継続時間長をモーラ数で割ったものとして規定されてきたが，長音や促音の混入があった場合，モーラ数を規定することが困難となる．一方，対話音声中から類似する音声パターンを抽出することができれば，同一パターン同士の継続時間長を直接比較することにより，一発話内における発話テンポの推移を分析することが可能と考えられる.

そこで，本稿では，自然対話データにおいて意図や態度に応じて多様に使い分けられる発話様式を分析するために必要な技術として，対話音声中で繰り返し利用される音声パターンに着目した上で，テキスト情報に依存せず音響情報のみにもとづいて繰り返しパターンを自動抽出する手法を提案する (図 1).

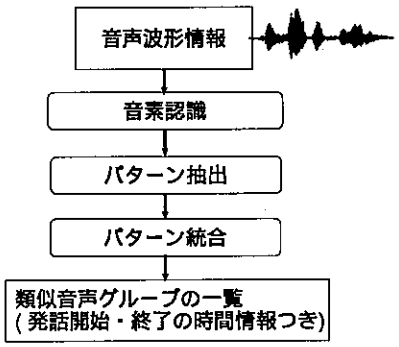

図 1　類似音声パターン抽出方法

まず，大語彙連続音声認識エンジン julius[5] を利用して音素認識 [6, 7] を行ない，音響特徴量[1]にもとづいて，音声波形情報を音素文字列に対応づける．なお，本手法においては，音素文字列を発話内容のテキスト情報としてではなく，音響特徴量の時系列パターンをインデクシングするための単なるシンボルとしてとらえているため，認識結果文字列と書き起こしテキストとの対応関係にもとづく音声認識精度については言及しない.

次に，Multigram Package [8] [2]を用いて，全対話分の認識結果音素文字列中に繰り返し現れる可変長パターンのうち 5 音素 ～ 10 音素に相当するものを，類似音声パターンとして抽出する．その上で，同じパターンに対応する音声区間同士を対話データ全体に渡って統合することにより，類似音声パターンのグループを作り，各グループの一覧を出力する．各類似音声パターンには，それぞれの発話開始・終了時間が付与されている.

*Speech rate of similar speech pattern in dialogue speech, by Kazuyuki ASHIMURA (JST/CREST) and Nick CAMPBELL (ATR, JST/CREST)

[1] 12 次元 MFCC + 12 次元 Δ MFCC + Δ log パワー

[2] Multigram Package は，入力文字列テキストに含まれる文字列パターンの中から，指定した長さより短く，指定した頻度よりも多く出現するものを可変長に抽出した上で，それらの出現確率を EM アルゴリズムにもとづいて計算するツールキットである.

4 結果と考察

4.1 類似音声パターンの抽出

パターン抽出の元データとして以下の 2 種類の HMM による音素認識結果を用いて，提案手法により類似音声パターンを自動抽出した，

HMM1: 音素バランス文および新聞文読み上げ音声から構築されたモノフォンモデル [5]

HMM2: HMM1 に対して，音声試料と同じ話者の音素バランス文読み上げ音声により話者適応を行なったモノフォンモデル

自動抽出された類似音声パターン，および各類似音声パターンに対応する音声セグメントの件数を表 1 に示す．なお，表 1 には，参考情報として，提案手法と同様の手法で書き起こしテキストから抽出した類似音声パターンおよび対応音声セグメントの件数も示す．

表 1　類似音声パターン抽出結果

件数	元データ		書き起こし (参考情報)
	HMM1	HMM2	
類似音声パターン	153	518	1,000
対応音声セグメント	37,382	147,228	521,258

表 1 に示す通り，HMM2 による音素認識結果から得た音声パターンの方が，HMM1 による音素認識結果から得た音声パターンよりも，音声パターン数，対応する音声セグメント件数ともに多く，書き起こしテキストから得た件数に近くなっている．

先行研究において，書き起こしテキストにもとづく julius 音素セグメンテーションを行なう際，話者適応した音響モデルを用いることにより，話者非依存の音響モデルを用いるよりも若干精度を向上できることが報告されている [9] が，音素認識においても，同様に認識精度が若干向上したため，よりきめ細かな類似音声パターンを抽出できていると考えられる．そこで，以下では，HMM2 を用いて抽出した類似音声パターンにもとづき，発話テンポについて考察する．

4.2 類似音声パターンの発話テンポ

対話音声は多様な発話様式を含むため，同一のテキスト情報を伝達する音声セグメントであっても，長音化など発話変形の影響により，従来のモーラテンポ (=発話継続時間長/モーラ数) では発話テンポを規定することが困難であると考えられる．そこで，同一の類似音声グループに割り振られた音声セグメント同士に着目し，発話テンポの指標として，音声パターン継続時間長の偏差値を求め，書き起こしから求めたモーラテンポと比較した (図 2)．なお，継続時間長は，類似音声パターンの抽出処理が完了し開始・終了時間を決定した時点で測定可能であるため，本稿では，対話データに現れる多様な発話様式を分析する第一歩として，まず発話テンポに着目する．

図 2 に示す通り，一発話内の発話テンポ変化に着目した場合，従来のモーラテンポは一定値となるのに対して，提案手法により求めた類似音声パターン継続時間長の偏差値は，ときおり偏差値 50 以上の外れ値がみられるものの，大部分のサンプルにおいて一発話内での発話テンポの滑らかな推移を示している．なお，外れ値は，句末音の引き延ばしなどに対応していると考えられるが，詳細分析については今後の課題とする．

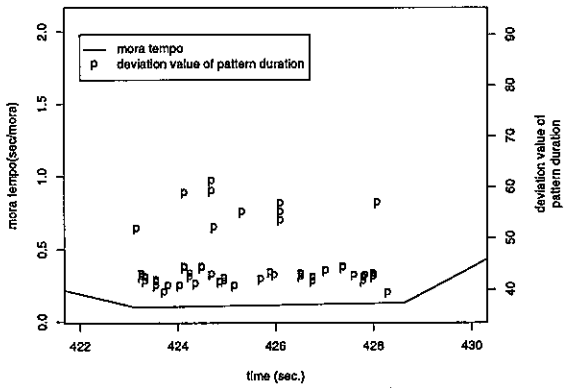

図 2　発話テンポ分析結果

分析結果の一例として，20 分程度の対話データのうち，対話開始から 423.139 秒 ～ 428.362 秒の間に行なわれた発話のテンポを示す．発話内容は，「なんかつかまり立ちみたいな、こう、さしたったら、して喜ぶねんけどな、でもあたしはほんまはハイハイさせたいから」．
なお，「モーラテンポ」および「パターン継続時間長の偏差値」が大きくなるほど，発話テンポは遅くなる．

5 おわりに

テキスト情報に依存せず音声情報のみから，対話中に繰り返し出現する類似音声パターンを自動検出する方法，および同手法を利用して算出する新しい発話テンポについて提案した．対話音声データにおいては，従来，長音化などの影響により発話テンポを求めることが困難であったが，提案手法により，一発話内における発話テンポの推移を分析することが可能になると期待される．

今後は，提案手法により求めた類似音声パターン継続時間長の偏差値に見られる外れ値について詳細を確認した上で，相手や談話機能に応じた発話テンポの変化に関する分析を行ないたい．

謝辞

いつも有益なアドバイスをくださる，ATR-SLT の柏岡秀紀主任研究員，および本プロジェクトのパーハム・モクタリ研究員，石井カルロス寿憲研究員，木村美名子氏に感謝いたします．

参考文献

[1] Campbell, Mokhtari: "Voice Quality, the 4 th prosodic dimension", Proc ICPhS 2003, pp.2414–2420 (2003).

[2] Campbell: "Recording techniques for capturing natural every-day speech", Proc LREC 2002, pp.2029–2032 (2002).

[3] Campbell, Mokhtari: "DAT vs. MD", 音講論 春季 1-P-27, pp.405–406 (2002).

[4] Campbell: "Labelling natural conversational speech data", 音講論 秋季 1-10-22, pp.273–274 (2002).

[5] 河原, 李, 小林, 武田, 峯松, 嵯峨山, 伊藤, 山本, 山田, 宇津呂, 鹿野: 日本語ディクテーション基本ソフトウェア (99 年度版) 日本音響学会誌, Vol.57, No.3, pp.210–214 (2001).

[6] 深田, 匂坂: 発音ネットワークに基づく発音辞書の自動生成, 情処研報 1996-SLP-14-3, pp.15–22 (1996).

[7] 大脇, シンガー, 鷹見, 樽松: 音素配列構造の制約を用いた音素タイプライタ, 信学技報 SP93-113, pp.71–78 (1993).

[8] Deligne, Bimbot: "LANGUAGE MODELING BY VARIABLE LENGTH SEQUENCES : THEORETICAL FORMULATION AND EVALUATION OF MULTIGRAMS", Proc ICASSP 95, pp.169–172 (1995).

[9] 米良, 李, 猿渡, 鹿野: Julius を用いた自由発話の自動ラベリングにおけるラベル尤度の統計的分析, 音講論 秋季 2-1-5, pp.57–58 (2001).