品詞情報と単語内位置情報を用いた話し言葉音声認識のための状態クラスタリング*

◎五十川賢造　西本卓也　篠田浩一　嵯峨山茂樹　(東大情報理工)

1　はじめに

本研究では話し言葉音声認識の精度向上のため、品詞と単語内位置を用いた決定木による状態クラスタリングを提案する。

現在、講演音声や対話音声等の話し言葉 (spontaneous speech) に対する認識性能は未だ十分な域に達していない。その原因として、話し言葉音声は朗読音声より発声変形が大きいことが挙げられている [1]。

我々は、“局所的な話速と発声変形との間に関係がある” という仮説のもと、話速を反映した音響モデリングを行う方法を検討している。そして、本稿では特にその実現方法としてクラスタリングの単位に品詞を用いた状態クラスタリングを提案する。青野ら [2] は、会話音声に対して、朗読発話で学習した音響モデルが与える尤度と自然発話で学習した音響モデルが与える尤度とを品詞別に比較し、両モデルの尤度差が助詞や助動詞で大きくなることを示した。この研究から我々は、品詞を元に話速を推定し、音響モデルを切り替えることを検討した。

また妹尾ら [3] は、単語のモーラ数と位置別に母音モデルを作成することで、単語認識精度を向上させた。この研究は音素の単語位置と発声変形の間には関係があることを示していると考えることができる。そこで我々はまず、音素の単語内位置を話し言葉の連続音声認識に利用することを検討した。

2　品詞クラスタ別音響モデル

話速によって音響モデルを切り替えるためには、認識の段階で使用できる事前情報により話速を予測しなければならない。品詞別の話速 (図 1) を実験で求めると、品詞毎に話速が異なることが分かった。そこで品詞をもとに、異なる話速に対応する複数の音響モデルを作成する事を考える。

そのためにまず、品詞を話速の大きい “Fast” クラスとそれ以外の “Slow” クラスに分割する。話し言葉のデータベースから品詞毎の話速 (mora/sec) を求め、話速が大きい方から上位 n 個までの品詞を Fast クラス、それ以外を Slow クラスとする。n は、データベース内での両クラスに属する単語ののべ出現単語数に占める Fast クラスに属する単語ののべ出現数の割合をもとに、実験で最適化する。

次に、Slow、Fast 両クラスタに対し初期モデルを作成し、この初期モデルを元に以下の 2 種類の音響モデルを作成する。

(1) 品詞クラスタ別音響モデル: 各初期モデルに対して音韻決定木を用いた状態クラスタリングを行う音響モデル。Slow、Fast 両クラスタに対応する音響モデルの間でパラメタ共有は行われない。

(2) 品詞クラスタ依存音響モデル: 両初期モデルを合わせたモデルに対して決定木による状態クラスタリングを行う音響モデル。音韻に関する質問に加え「Fast クラス用のエントリであるか?」「先行コンテキストは Fast クラス用のエントリであるか?」「後続コンテキストは Fast クラス用のエントリであるか?」の 3 つの質問を用いて、決定木を作成する (図 2)。Slow、Fast 両クラスタに対応する音響モデルの間でパラメタの共有が行われる。

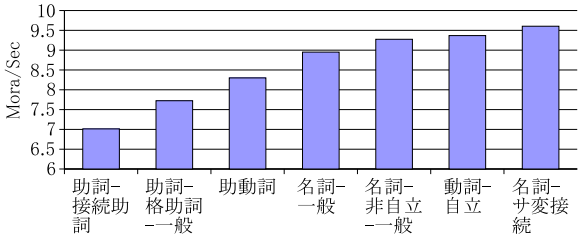

図 1: 品詞別話速。データベース内での出現数が多いもの上位 7 個を話速でソートしている

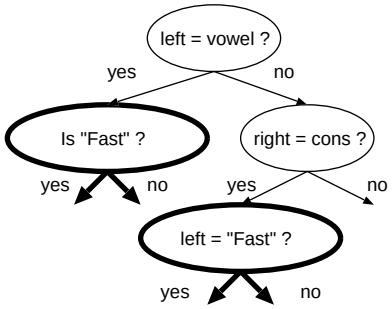

図 2: 品詞クラスタ依存音響モデル用決定木例

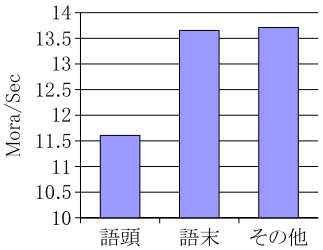

図 3: 単語内位置別継続長。1 モーラ単語は語頭として集計している

3　単語内位置別音響モデル

単語内位置による発声変形をモデル化するため、単語内位置で状態を分けることを考える。ここでは単語内位置として特に語頭・語末に着目する。実験の際に求めた単語内位置別の話速を語頭、語末、その他について図 3 に示す。この結果は話速の観点から、語頭の音声特徴量が他の部分と異なる可能性が裏づけられたと考えることが出来る。

まず、語頭と語末とそれ以外の音素とを区別して初期モデルを作成する。次に、この初期モデルを用いて以下の 2 種類の音響モデルを作成する。

(3) 単語内位置別音響モデル: 各初期モデルに対して、決定木による状態クラスタリングを行う音響モデル。対応する単語内位置が異なる音響モデルの間でパラメタ共有を行わない。

---

* “HMM’s state clustering using word contexts for spontaneous speech recognition” by Kenzo Isogawa, Takuya Nishimoto, Koichi Shinoda and Shigeki Sagayama, Department of Information Physics and Computing, Graduate School of Information Science and Technology, The University of Tokyo.

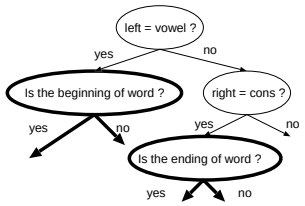

図 4: 単語内位置依存音響モデル用決定木例

表 1: 実験条件

継続長推定用データ	CSJ コーパスの全 391 講演
分析条件	16kHz サンプリング, 25msec 窓幅, 10msec フレームシフト
特徴量	MFCC12 次元 + ΔMFCC12 次元 + Δ 対数パワー
学習データ	CSJ コーパスより 200 講演 (テスト話者を除く男性の講演)
HMM のパラメタ数	1500 状態、16 混合、対角共分散
テスト話者 (話者適応用データ)	CSJ コーパスより 7 講演分 a01m0007, a01m0035, a01m0074, a02m0117, a03m0100, a05m0031, a06m0134

(4) 単語内位置依存音響モデル: 全初期モデルを 1 つにまとめたモデルに対して決定木による状態クラスタリングを行う音響モデル。音韻に関する質問に加え「先行コンテキストは語頭であるか?」「中心コンテキストは語頭であるか?」「中心コンテキストは語末であるか?」「後続コンテキストは語末であるか?」の 4 つの質問を用いて決定木を作成する (図 4)。対応する単語内位置が異なる音響モデルの間でパラメタ共有が行われる。

4 音声認識実験

講演原稿の書き起こしを目的として、対角共分散混合ガウス分布トライフォン HMM による講演音声の認識実験を行った。表 1 に実験条件を示す。モデルの学習・テストのためのデータには CSJ(Corpus of Spontaneous Japanese) コーパスモニタ版 (2001)[4] を用いた。認識エンジンには Julius3.1p2[5] を用い、形態素解析には Chasen ver. 2.02[6] を用いた。言語モデルは CSJ コーパスに付属している講演音声用言語モデル (2002-06; 京都大学) を用い、辞書についても CSJ コーパス付属の辞書を用いた。

品詞情報を利用するためには、単語境界の情報と品詞情報とを持つ書き起こしテキストが必要なので、以下の手順で作成した。まず、CSJ コーパス付属の書き起こしテキストの転記基本単位をもとに、コーパスの音声データを分割した。ただし、1 つの形態素が複数の転記基本単位に分割されているものは[1]、それらを接続し 1 つの転記基本単位として扱った。次に、各転記基本単位のテキストデータを形態素解析した。Chasen が出力する片仮名を利用した DP を用いて、転記基本単位内の形態素に CSJ コーパスの読み仮名を割り付けた。読み仮名の割り付けができない単語が存在した転記基本単位と、未知語を含む転記基本単位は、全て取り除いた。Chasen によって、用いたデータセットの 98.0%の単語に正しい読み仮名を付与できた。

品詞別の話速は、各単語の継続長を音声認識システム [5] 付属の性別非依存モノフォンモデル (混合数 16) を用いた viterbi アライメントから求めた。また、不特定話者モデルを初期モデルとして、7 名のテスト話者それぞれについて、同一話者のデータ全てを使用した MLLR(HTK. Ver. 3.0, HEAdapt[7]) による教師なし話者適応を行った。

実験結果は以下の様になった。Fast クラスに属する単語ののべ出現数が全体の 10%の場合に、品詞クラスタ別音響モデルの単語正解精度が最大となった。品詞クラスタ依存音響モデルでは、30%の場合に最大となった。前者より後者の方が高い単語正解精度が得られたが、両者ともベースラインの認識精度を越える単語正解精度は得られなかった。また、単語内位置別音響モデルでは 1.9 ポイント、単語内位置依存音響モデルでは 1.1 ポイント、それぞれベースラインより高い認識精度が得られた (図 5)。

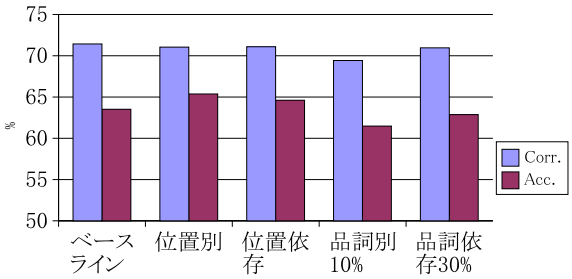

図 5: モデル別単語認識性能

単語内位置別音響モデルの精度が向上した理由は 2 つ考えられる。1 つは、語頭と語末を同時に含み複数の単語に跨るトライフォンにより単語境界のモデルの精度が向上したことにより、単語の挿入誤りが減少したこと、もう 1 つは、語頭語末を考慮したトライフォンにより、3 音素以下の単語に対して単語 HMM を用意したことと類似した効果が得られたことである。

5 おわりに

話し言葉音声における発声変形に対処するため、音響モデルの状態クラスタリングの際に、品詞の情報と音素の単語内位置の情報を用いる方法を提案した。品詞情報を用いる方法では単語正解精度の向上は見られなかったが、単語内位置情報を用いる方法では最大で 1.9 ポイント向上した。今後は、語頭・語末以外の単語内位置情報の検討、品詞を用いた状態クラスタリングと単語内位置情報を用いた状態クラスタリングの統合、本研究の手法の朗読音声に対する効果の検証等を行いたい。

参考文献

[1] 河原 達也, “話し言葉音声認識の概観,” 電子情報通信学会技術研究報告, Vol. 100, No. 523, SP95–116, pages 1–6, 2000.

[2] 青野 邦生, 安田 圭志, 竹野 寿幸, 山本 誠一, 柳田 益造, “言語情報を考慮した発話スタイル依存音響モデル自動選択の予備検討,” 日本音響学会 2002 年秋期研究発表会講演論文集, Vol. 1, pages 89–90, 2002.

[3] 妹尾 貴宏, 村上 仁一, 前田 智広, 池原 悟, “モーラ情報を用いた単語音声認識の研究,” 電子情報通信学会技術研究報告, Vol. 101, No. 234, SP2001–45, pages 1–5, 2001.

[4] 前川 喜久雄, 籠宮 隆之, 小磯 花絵, 小椋 秀樹, 菊地 英明, “日本語話し言葉コーパスの設計,” 音声研究, pages 51–61, 2001.

[5] 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, 山本 幹雄, “音声認識システム,” オーム社, 2001.

[6] 松本 裕治, 北内 啓, 山下 達雄, 平野 善隆, 松田 寛, 浅原 正幸, “日本語形態素解析システム「茶筌」version 2.0 使用説明書 第二版,” 奈良先端科学技術大学院大学, 1999, Information Science Technology Report NAIST–IS–TR99012.

[7] Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, Phil Woodland, “The HTK Book (for HTK Version 3.0),” Microsoft Corporation, 2000.

[1] CSJ コーパスの書き起こしテキスト内では< C >で表される