音声言語情報処理　31－11
（2000．6．2）

種々のタスクにおける大語彙連続音声認識システムの性能評価と診断

南條 浩輝　加藤 一臣　三村 正人　李 晃伸　河原 達也

京都大学 情報学研究科 知能情報学専攻
〒606-8501 京都市 左京区 吉田本町
e-mail: nanjo@kuis.kyoto-u.ac.jp

あらまし　大規模な統計的モデルに基づく大語彙連続音声認識における認識誤りの原因を自動的に同定・診断する手法を，読上げ音声，講演音声，対話音声の三つの異なる認識システムに適用し，現状の音声認識システムの評価を行う．この手法は，正解文と認識結果のスコアを比較し，区間と要因ごとに分解するものであり，誤り単語を含む区間ごとに対して，原因モジュールを音響モデル・言語モデル・デコーダのいずれかに同定し，その傾向を分析する．読上げ音声認識システムにおいては，サーチエラーを自動同定・分析することにより，実際に効率的にデコーダの改善を行うことができた．講演や対話のような話し言葉の音声認識では，間投詞やショートポーズに関する言語モデルや，典型的なつなぎ語や文末表現における音素コンテクストのモデルに問題があることが明らかになった．また，同一の発声スタイルのデータで音響モデル・言語モデルを構築することの重要性・有効性が確認された．

キーワード　大語彙連続音声認識，自動診断，話し言葉，音響モデル，言語モデル，デコーダ

Diagnosis and evaluation of various LVCSR systems

Hiroaki Nanjo, Kazuomi Kato, Masato Mimura,

Akinobu Lee, Tatsuya Kawahara

Graduate School of Informatics Kyoto University, Kyoto 606-8501, Japan
e-mail: nanjo@kuis.kyoto-u.ac.jp

Abstract　We have proposed a method to automatically diagnose recognition errors in large vocabulary continuous speech recognition. For every erroneous segment, it identifies the error-causing module among acoustic model, language model and the decoder. More detailed diagnosis is also output. This method has been applied to three LVCSR systems: read speech, lecture speech and dialogue speech recognition system. In the read speech recognition system, the diagnosis is useful to improve our decoder Julius. In the lecture and dialogue recognition systems, we have made clear the significance of matching training database in its speaking style to the task.

keywords: large vocabulary continuous speech recognition, automatic diagnosis, conversational speech, acoustic model, language model, decoder

1 はじめに

大語彙連続音声認識は音声対話システムや種々の音声インターフェースの基盤となる技術である[1][2]. 大語彙の特に自然な発話についてはまだ十分な技術は確立されていないが, 第一に高い認識精度が要求される. そのような高精度のシステムの構築には段階的なシステム改善が必要であり, 認識誤りの原因を同定し分析することが有益である. 認識誤りの傾向からデバッグやアルゴリズム改善のヒント, さらには今後の研究やデータ収集の指針を得ることが期待できる. しかしながら一般の大語彙連続音声認識システムにおいては, 音響モデル・言語モデルともに大規模かつ統計的なモデルであり, 探索過程でも膨大な仮説を扱うために, 誤り原因を人手で同定することは容易でない.

我々は既に, HMMによる音響モデル, 単語N-gramモデルによる言語モデル, 探索アルゴリズムに基づくデコーダからなる確率的な大語彙連続音声認識の枠組みにおいて, 認識誤りがどのモジュールに起因するかを自動的に分類・同定し, その分析をおこなう手法を提案し実装した[3]. 本稿では, これを実際に種々の大語彙連続音声認識システムに適用した結果の報告を行い, 同時にシステムの評価も行う. まず, 読上げ音声を対象とした認識システムを構成するIPA「日本語ディクテーション基本ソフトウェア」[4][5]に適用した結果を示し, 次に, 話し言葉音声を対象とした認識システム, 具体的には講演音声認識システム[6]と対話音声認識システム[7]に適用した結果を示す.

2 誤り原因モジュール同定と診断

2.1 確率的音声認識の枠組みとモジュール同定

確率的音声認識の概要を図1に示す. 音声認識は入力音声 X に対して確率 $P(W/X)$ を最大とするような単語列 W を求める問題とみなせる. ここで事後確率 $P(W/X)$ はベイズ則により,

$$P(W/X) = \frac{P(X/W) \cdot P(W)}{P(X)} \qquad (1)$$

となる. $P(W/X)$ を最大とする W を求めるには, 式(1)の右辺の分子 $P(X/W) \cdot P(W)$ を最大とする W を求めればよい. 実際の認識システムにおいて

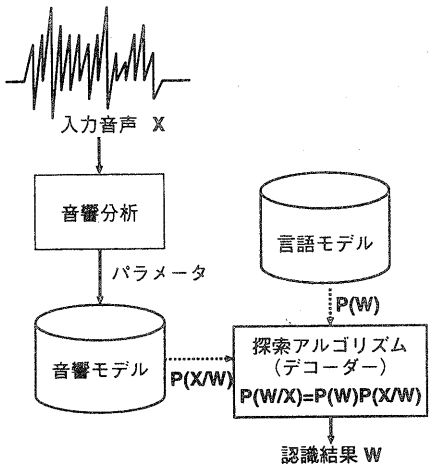

図 1: 確率的音声認識の概観

は, $P(W)$ は言語モデルのスコア, $P(X/W)$ は音響モデルのスコアに相当し, デコーダは両モデルを効率よく適用して $P(X/W) \cdot P(W)$ を最大とする仮説 W を求める機構である.

ここで, 認識システムが出力する認識結果の単語列を W_r（認識文）, 正解の単語列を W_c（正解文）と表記すると, W_rが W_cに一致しないとき認識誤りとなる. このとき, 図2に示す決定木に従って, 認識誤りの原因となったモジュールを分類できる. すなわち, 正解文の方がスコアが高いにもかかわらず, これを見つけることができなかった場合は, デコーダによるサーチエラーである. そうでない場合は, 認識結果のスコアを高くしたモデルが原因である. ただし, 認識文が正解文よりスコアの高い場合にも, よりスコアの高い解を見つけられなかった可能性があるが, 真の最適解を知ることはできないので, この場合はサーチエラーに分類していない. また実際には, 言語モデルスコアには重みがつけられ, 単語挿入ペナルティなどで補正されるが, 図2では簡単のためにこれらのパラメータを省略している. これらは実装の際には考慮している. 未知語に関する誤りは, 単語辞書に起因するものと分類する. 言い淀みや単語の一部分の言い直しなども未知語とみなされる.

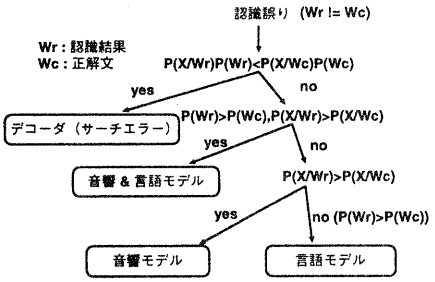

図 2: 誤り原因モジュール分類決定木

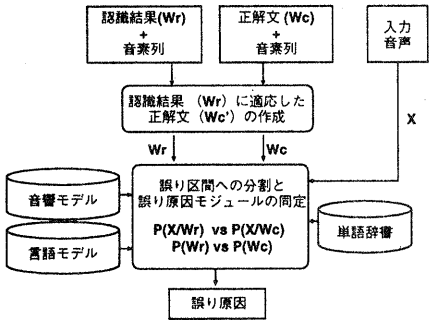

図 3: 原因同定の処理過程

2.2 実装上の問題点

誤り原因モジュール同定の処理過程を図 3 に示す. 図 2 の決定木を適用する際には, 厳密なスコアの比較が必要である. そのため, 認識結果に適合した正解文を作成する必要があり, 具体的には以下の処理を行う.

1. 正解文の表記の揺れへの対処

一般的な音声認識システムでは単語を単位として認識を行っている [8]. しかし, 日本語は分かち書きされないため単語の区切りが明確でない. 例えば「千九百九十九」は「千_九百_九十_九」のように 1 単語であるとも 4 単語であるとも解釈できる. また,「従って」と「したがって」など同一単語に対する漢字表記とかな表記や,「百」と「100」などの様々な数字の表記法による曖昧性も存在する. 認識率の算定の際には, 認識結果を正解文の表記にマッチングするように補正することが多い [9]. しかし誤り原因の診断を行う際には, 認識時に言語モデルスコアを求めた単語列を保持して厳密な比較を行うために, 正解文の方を認識文にあわせて補正する.

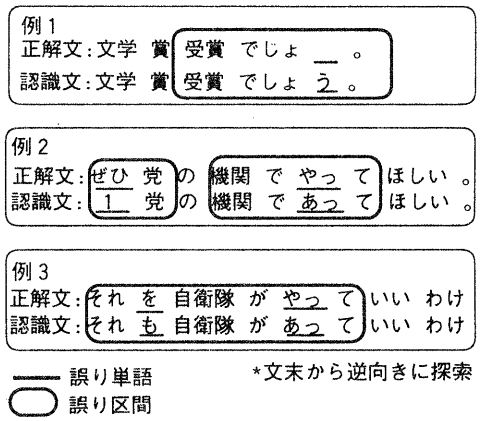

図 4: 誤り区間への分割

2. 正解文の発音の揺れへの対処

単語辞書中には, 例えば「自動同定」に対して /jidoudoutei/ と /jido:do:te:/ といったように複数の発音エントリを持つ単語が多数存在する. 通常の認識率の算定の際には, このような発音の違いは許容されている. しかし, 同じ単語でも発音が異なると音響スコアが異なるため, スコアの比較を厳密に行うためには, 認識文と正解文を発音レベルで対応をとる必要がある. ここでは, 複数の発音をもつ単語に対しては, 認識結果で用いられた方を採用する.

3. 正解文への句読点の挿入

通常, 認識率を求める場合には句読点を考慮しない. しかし, 句読点を言語モデルに含めている場合には, 言語モデルのスコアに影響する. また, 句読点は音響的にショートポーズ (sp) に対応づけられるので, その有無が音素環境依存モデルの選択や音響モデルのスコアにも影響する. したがって, 認識結果を基に句読点を正解文に挿入する.

2.3 誤り区間への分割

入力の複数の区間に誤りがある場合, 一文全体に対して図 2 の分類を適用するよりも, 誤り単語ごとに原因モジュールを同定できる方が望ましい. しかし, ある単語が誤るとマッチング長も異なってくるため, 前後の単語の音響スコアに影響を与える. 通常

表 1: 読上げ音声認識システムと評価データ

	デコーダ	音響モデル	言語モデル	評価データ
5K システム	Julius rev.1.1	性別依存 triphone 2000 状態 16 混合 (IPA モデル)	5000 語 2-gram 0.57M, 3-gram 2.0M (cutoff 1-2) 新聞記事コーパス 65M 単語 より学習	IPA-97-TestSet 男性話者 10 名, 合計 100 文
	Julius rev.2.1			
20K システム	Julius rev.2.1		20000 語 2-gram 1.7M, 3-gram 7.4M (cutoff 1-1) 新聞記事コーパス 118M 単語より学習	IPA-98-TestSet 男性話者 23 名, 合計 100 文 (IPA-97-TestSet と異なる 100 文)
	Julius rev.3.0			

はこの影響は 1 単語程度と考えられるので, 前後の単語を誤り区間に含める. また, 言語モデルに単語 3-gram を用いる場合は, 誤り単語の直後の 2 単語まで言語スコアに影響が及ぶので, これらも含める. そこで, この一連の単語列, 具体的には誤り単語の 1 単語前から 2 単語後までを誤り区間と定義する. 複数の誤り区間が重なる場合はマージする. 誤り区間生成の例を図 4 に示す. なお, 本研究では文末から逆向きに探索を行っている.

このように分割された各区間ごとに対して, 図 2 の決定木に従って誤り原因モジュールを同定する.

2.4 誤り原因の診断

各誤り区間に対して, その原因となったモジュールが同定されるが, さらに詳細な誤り診断を行う.

音響モデルに起因する誤り区間に対しては, 顕著に含まれる音素コンテクストを列挙し, その傾向や偏りを調べる. 言語モデルに起因する誤り区間に対しても同様に, 顕著に含まれる単語コンテクストを列挙し, またそれらがバックオフされていないかも調べる. これらはモデル化や学習データ収集の参考になる.

サーチエラーに関しては, どの単語のどのモデルスコアの低下により生じたかという分析や, 誤り区間のフレーム単位のスコアの遷移を出力することにより改善の指針を得る.

3 読上げ音声認識システムへの適用

まず本手法を, IPA「日本語ディクテーション基本ソフトウェア」[4][5] で構成される読上げ音声認識システムに適用した. 語彙サイズ 5K と 20K のシステムのそれぞれに対して評価・診断を行った. いずれも認識精度を重視した高精度版を用いた. 使用した音響・言語モデル, デコーダと評価データは表 1 に示す通りである.

特に著者らは 2 パス探索を行うデコーダ Julius[10] の開発を行っているので, その改良の指針を得ることを主な目的とした.

まず, 5K システムに対して適用, 診断を行った. 診断結果を表 2 に示す. [1] Julius rev.1.1 (97 年度版) [4] を用いたシステムでは単語誤り率は 7.6%で, その約半数がサーチエラーであった. サーチエラーと同定された区間をトレースすることにより, 第二パスの探索における単語の接続フレームが, 第一パスの結果の候補の単語境界と必ずしも一致しないことがわかった [11]. そこで, 接続範囲を数フレーム分許容することによって大幅にエラーが減った. その他にも, 実際に誤りサンプルに基づいて効率的に修正が行え, Julius rev.2.1 (98 年度版) [5] では, 単語誤り率を 5.4%まで改善できた.

次に 20K システムに対して適用, 診断を行った. 表 3 に結果を示す. 表 2 と比較すると, 5K から 20K に語彙が大きくなったことにより探索空間が増大し, サーチエラーの割合が増加している. 表 4 に Julius rev.2.1 (98 年度版) によるサーチエラーの一覧を示す.

これから一音節の単語 (表 4 の下線部) に関する誤りが多いことがわかる. これは Julius rev.2.1 の第一パスでは単語内の triphone のみを扱い, 単語間の triphone を適用しないため, 単語の遷移において音響モデルの精度が落ち, 最適解が失われていることを示している. 特に, 一音節単語が続く場合はこの影響が大きい. そこで, 単語間 triphone を第一パス

[1] 認識率や未知語率が過去の報告と若干異なるのは人手による修正を行っているからである.

表 2: 5K システムの診断結果

デコーダ	音響	言語	両方	探索	合計
Julius-1.1	0.8%	1.4%	1.5%	4.0%	7.6%
Julius-2.1	0.5%	1.6%	1.7%	1.7%	5.4%

単語誤り率

表 3: 20K システムの診断結果

デコーダ	音響	言語	両方	探索	未知語	合計
Julius-2.1	0.8%	1.8%	0.9%	4.1%	0.06%	7.6%
Julius-3.0	0.9%	1.2%	1.5%	2.0%	0.06%	5.7%

単語誤り率

で近似的に扱うように改善した．具体的には，可能な単語間 triphone のスコアの最大値で代用した [12]．その結果，Julius rev.3.0（99 年度版）ではサーチエラーが大きく削減された．

なおデコーダのみを修正したにもかかわらず，音響・言語モデルに起因する誤りが増減しているのは，主として，よりスコアの高い解が得られるようになった結果による．

以上のように，サーチエラーの区間を自動同定，分析することにより，デコーダ Julius の改善を効率的に行うことができた．

4 講演音声認識システムへの適用

講演音声は，“書き言葉” の特徴を有する “話し言葉” であり [13]，文章がある程度丁寧で整っている半面，間投詞・言い直しなどが頻繁に出現し，つなぎ語や文末表現にも特徴がある．また，講演には特有の話題が存在し，話題に関する特殊な単語キーワードが当該講演においてのみ繰り返し用いられる．このため，講演音声のディクテーションには，講演独特の言い回しを反映したモデル化と講演話題への適応が必要である．これらを考慮した講演音声認識システムは一定の成果をおさめているが [6][14]，読上げ音声認識システムに比べると認識精度は下回っている．

このため，講演音声認識システムに提案手法を適用し，診断を行った．用いた評価データは，男性話者 3 名の卒業研究に関する OHP を用いた発表であり，システムの仕様とともに表 5 に示す．その診断結果を表 6 に示す．読上げ音声の認識と比べると，パープレキシティが大きいためサーチエラーも増加しているが，言語モデルと音響モデルに起因する誤りが大きく増えている．

表 4: Julius rev.2.1 によるサーチエラー区間 (20K)

認識文	正解文
今年 より 彼女	お 年寄り から
の 荷物 は 英 、	の 注文 に も 備え、
個人 に は 、	個人技 が 、
も 、 ひざ 、	も 、 いざ 、
安全 両 彼 の おかげ で 方向	安全 に お金 を かける 方向
変わっ て い た	変わっ て き た
九百 九十 四 年間 の	九百 九十 四 年 版 の
事業 計画 後 、	事業 計画 を 、
男子 より 書 が 使わ れ	男子 より 不利 に 扱わ れ
は 、 飛行 は 少ない	は 、 非行 が 少ない
部族 は 表 を 作る こと	部族 は 火 を つくる こと
神戸 市 は 演習 場 を 建設 .	人工 芝 練習 場 を 併設 .
二 人 家族 な 安心	二 人 で 肩 を 組ん だ 安心
ミリグラム に	ミリグラム に
迫る こと 足り	一 割 ほど 足り
の 役割 よう に なっ て	の 役割 を 担っ て
ほか 、 史上初 と	ほか 、 地上波 と
除草 剤 の	共通 素材 の
に し た が 、	に し ながら 、
一緒 に し なかっ	一緒 に 行け なかっ
時間 が 残る と	時間 が かかる と
五 キロ を 一 割	五 キロ 行け ば いい
側 も 十 抜き し	側 も 銃撃 し
区 、 肉 、	区 、 二 区 、
者 の を 、 、 氏名	者 の 、、 氏名
それ も 自衛隊	それ を 自衛隊
ペン で 知る 資料 を	ペン で 印 を

下線: 単語間の遷移での音響スコアが問題となる（一音節）誤り

言語モデルに起因するとされた誤り区間で見つかった 2-gram, 3-gram の存在しないエントリには，間投詞やショートポーズを含むものが顕著にみられた．それらの割合を表 7 に示す．2-gram については，その先行単語と後続単語について頻度順に表 8 に示す．間投詞と句読点で約半数を占めていることがわかる．残りを分析すると，言い直し・言い淀みなどが多数確認された（「表の表」「大き，大きい」など）．この言語モデルの学習には講演録コーパスを使用しているが，人手による書き起こし・編集の際に間投詞や言い淀みが削除・修正されているため，これらが正しくモデル化できていないと考えられる．

また，これらは言語的制約を弱めるので探索にも影響を与える．例えば「辞書-を-構成」という 3-gram エントリは存在するにもかかわらず，「辞書 , を , エー 構成」と発声された場合は，「エー-構成」「 ,-を」などの 2-gram エントリがないために，局所的に言語モデルスコアが低くなり認識を困難にする．

表 5: 講演音声認識システムと評価データ

デコーダ	音響モデル	言語モデル	評価データ				
			話者	発表時間	総単語数	間投詞数	PP
Julius rev.2.1	性別依存 triphone 2000 状態 16 混合 (IPA モデル)	8000 語 + 適応語 2-gram 130K, 3-gram 71K (cutoff 1-1) 講演録コーパス 840K 単語より学習	A	10 分	1077	48	60.5
			B	8 分	1363	96	95.7
			C	11 分	1829	221	85.2

PP: パープレキシティ

表 6: 講演音声認識システムの診断結果

話者	音響	言語	両方	探索	未知語	合計
A	2.5%	3.7%	5.1%	4.2%	2.2%	17.7%
B	5.2%	3.4%	4.5%	6.4%	4.5%	24.1%
C	5.0%	5.1%	9.4%	6.0%	2.8%	28.3%

単語誤り率

表 7: 存在しなかった N-gram エントリに間投詞やショートポーズが含まれる割合 (総数)

話者	2-gram	3-gram
A	48%(73)	71%(112)
B	34%(98)	56%(147)
C	61%(189)	78%(343)
合計	51%(360)	71%(602)

間投詞については, 話し言葉の忠実な書き起こしテキストから N-gram 確率を推定するか, 探索過程で透過単語として扱うなどして, 言語スコアの計算を工夫する必要がある [8][15]. ショートポーズについても間投詞と同様である. 言い直しや言い淀みに関しては, 現在の N-gram の枠組みではモデル化は難しい. また, 入力の始端と終端で N-gram 確率を与えられないために誤りが起きている例も, 少数ではあるが存在する. 講演のディクテーションにおいては, ポーズで区切る単位が文とは限らないので, 直前の履歴を参照して N-gram 確率を与えることも考えられる.

音響モデルに起因するとされた誤り区間に頻出した音素コンテクストを表 9 に示す. 「このように」「しまして」「あります」「ので」といった講演に頻出するつなぎ語や文末表現に起因するパターンが多くを占めている. また「~は」「~と」「~の」「~して」などの, 後ろにショートポーズが置かれる音素コンテクストも多く見られる. 講演音声は, 読上げ音声と比べて一般に発声速度が速く, ポーズを置く位置が

表 8: 存在しなかった 2-gram エントリ

先行単語 (頻度) 単語+*	後続単語 (頻度) *-単語
エ (56)	エ (22)
の (15)	, (17)
エー (14)	の (15)
、 (13)	エー (11)
ト (11)	. (9)
。 (9)	に (9)
で (8)	その (8)
が (6)	を (8)
に (6)	辞書 (7)
を (6)	に (6)
	この (6)

表 9: 音響モデルに起因する誤り区間に頻出する音素コンテクスト (頻度)

講演音声			
o-n+o (50)	sh-i+t (43)	a-sh+i (37)	i-t+a (30)
i-t+e (28)	i-m+a (28)	o-t+o (27)	w-a+sp (20)
k-o+t (19)	m-a+sh (19)	e-w+a (19)	t-o+sp (19)
a-k+u (19)	o-k+u (17)	o-d+e (17)	k-o+n (17)
a-r+i (17)	a-s+u (17)	o-k+o (16)	o-n+i (15)
i-sh+o (15)	m-a+s (15)		

多い上に, それらが不均一であるために, 読上げ音声から学習された音響モデルではこのようなコンテクストを十分に学習できていないと考えられる.

このように, 講演音声のような話し言葉の音声認識には, 当該発声スタイルの音声・言語データベースを収集して, それを用いて音響モデル・言語モデルを構築する必要性が明らかになった.

5 対話音声認識システムへの適用

最後に, 対話音声認識システムに適用して, 診断を行った. 対話音声は完全に "話し言葉" であるため,

表 10: 対話音声認識システムと評価データ

デコーダ	音響モデル	言語モデル	評価データ
Julius rev.3.0	性別依存 triphone (2000 状態 16 混合) ATR/SDB (旅行対話) 197 名 17 時間 性別依存 triphone (2000 状態 16 混合) ASJ/JNAS (読上げ) 130 名 48 時間	8600 語 2-gram 68K, 3-gram 180K (cutoff 1-1) ATR 旅行対話コーパス 780K 単語より学習	ATR/SDB 男性話者 17 名 合計 210 発話

表 11: 対話音声認識システムの診断結果

音響モデル 学習データベース	音響	言語	両方	探索	合計
対話 (ATR)	1.5%	1.3%	2.6%	6.3%	11.8%
読上げ (ASJ)	8.5%	2.7%	4.2%	7.3%	22.9%

単語誤り率

主語や助詞の省略や倒置などが頻繁に生じる. また, 間投詞や言い直し, 言い淀みも多い. そのため, 言語モデルは対話データから学習する必要がある. また, 発声が不明瞭な区間が多い上に, 話し言葉に多く出現する音素コンテクストは読上げ音声のそれと傾向が異なるため, 音響モデルも対話音声データから構築する必要がある [16].

対話音声については, ATR 音声翻訳通信研究所で作成された大規模な自然発話データベースが利用できるので [17], これを用いて音響・言語モデルを学習している [7]. 表 10 に認識システムの仕様と評価データを示す. タスクドメインが限定されているため, 単語パープレキシティは 20 程度である. なお比較のために, 音響モデルに新聞記事読上げ音声データ (ASJ/JNAS) から学習した IPA モデルを使用したシステムも用いる.

診断結果を表 11 に示す. 対話データで構築した音響モデルが, 読上げ音声から構築された音響モデルと比べて誤りを約半分にしており, 発声スタイル・タスクを一致させることの効果がわかる. 表 12 に読上げ音声から学習された音響モデルに起因する誤り区間に頻出した音素コンテクストを示す. これらは主に, 「~します」「~です」「~んで」といった対話音声に特徴的な言い回しや, 「あの」「それ」などの指示表現などに含まれるパターンである. 前者は新聞記事にはほとんど出現しない音素コンテクストであり, 後者は新聞記事でも用いられるが, 読上げ音声と対話音声では発声が異なる音素パターンである.

表 12: 音響モデルに起因する誤り区間に頻出する音素コンテクスト (頻度)

対話音声

i-m+a (18)	d-e+s (16)	e-s+u (15)	o-sh+i (11)
m-a+s (11)	N-d+e (10)	sh-i+t (10)	i-t+e (9)
i-t+a (9)	q-t+o (9)	s-u+k (8)	o-k+u (7)
a-k+u (7)	i-sh+i (7)	e-d+o (7)	g-a+i (7)
a-i+sh (7)	a-r+a (7)	u-n+o (7)	o-r+e (7)
o-n+o (7)	s-o+r (7)	t-e+r (7)	a-s+u (7)

音響モデル学習データベース: 読上げ (ASJ/JNAS)

実際の認識結果においても, 読上げ音声モデルでは, 「~を」が「~を お を」, 「ない」が「ない い」, 「あの」が「あの お」と誤認識されたりする例が顕著に見られた. これらは語尾の母音が怠けたり間延びしたりする現象をモデル化できていないためと考えられる. 例えば, 「あの」という単語の音素表記が /ano/であっても, 対話音声中の語尾の母音/o/は読上げ音声の長母音/o:/に近いかさらに長い. そのため, 読上げ音声モデルでは/ano o/にマッチングし, 「あの お」と認識される.

これに対して, 対話音声から学習した音響モデルでは, 口語表現に頻出する音素パターンや語尾で間延びする母音が正しく認識できており, 音響モデルに起因する誤りが大幅に減っている. しかし, 「~します」「です」「んで」などは変形が大きく, モデル化が十分でない.

次にパープレキシティが低いにもかかわらず, サーチエラーが多いことについて考察する. サーチエラーの区間には, 上記の例のように一音節の間投詞が連続して出現する誤りパターンが多く存在し, 挿入誤りのカウントが大きく増えている. この言語モデルは自然発話データベースで学習されているため, 間投詞も組み込まれており, 間投詞の出現確率は比較的高い. そのため, このような間延びする語尾の音声区間に対して音響的なマッチングがよくないと, 間投詞が挿入されやすくなる. 人名や固有名詞など言

語スコアが低くなる区間でも，同様の誤り傾向が見られた．

このように，対話音声のような話し言葉では同一の発声スタイルのデータベースでモデルを学習することの有効性を確認できた．同時に，変形の著しい区間の音響モデル化の難しさや，間投詞を単純にモデルに組み込むことの副作用も明らかになった．

6 まとめ

大語彙連続音声認識における認識誤り原因の自動同定の手法を読上げ音声認識システム，及び講演音声と対話音声の認識システムに適用した．これにより，現在のシステムの問題点が明らかになり，今後の改善の指針を得ることができた．特に読上げ音声認識システムにおいては，実際にデコーダの開発，改善にとって有用な誤りパターンを効率よく見つけ出すことに成功し，改善につながった．

診断を行った三つのタスクには，ともに他のタスクで出現しない言い回しや特徴的な語彙の存在が，言語モデルや音素コンテクストのモデリングにおいて問題となる．また同一表記の単語でも，音響的に異なることが多い．そのため，言語モデル・音響モデルともに，当該タスク・発声スタイルに一致したデータベースを用いて学習を行うことの重要性が明らかになった．音響モデルに関しては，対話音声認識において一致させることの効果を示した．言語モデルに関しては，話し言葉では間投詞やショートポーズが問題となること，及び間投詞を単純にモデルに組み込むと副作用があることがわかった．

このように，現状の大語彙連続音声認識システムの課題が明らかになったので，今後はこの診断結果を，自然な話し言葉の音声認識の高度化に役立てていきたい．

謝辞 本研究の一部は情報処理振興事業協会 (IPA)「日本語ディクテーションの基本ソフトウェアの開発」の援助を受けて行われた．関係各位に感謝します．

参考文献

[1] 鹿野清宏．日本語ディクテーション基本ソフトウエア．春季音講論，2-8-1，2000.

[2] 西村雅史．日本語ディクテーションの現状と今後の課題．情処学研報，99-SLP-29-2，1999.

[3] 南條浩輝，李晃伸，河原達也．大語彙連続音声認識における認識誤り原因の自動同定．情処学研報，99-SLP-27-6，1999.

[4] 河原達也，李晃伸，小林哲則，武田一哉，峯松信明，伊藤克亘，伊藤彰則，山本 幹雄，山田篤，宇津呂武仁，鹿野清宏．日本語ディクテーション基本ソフトウエア (97 年度版) の性能評価．情処学研報，98-SLP-21-10，1998.

[5] 河原達也，李晃伸，小林哲則，武田一哉，峯松信明，伊藤克亘，山本幹雄，山田篤，宇津呂武仁，鹿野清宏．日本語ディクテーション基本ソフトウエア (98 年度版) の性能評価．情処学研報，99-SLP-26-6，1999.

[6] 加藤一臣，河原達也．講演ディクテーションのための話題独立言語モデルと話題適応．春季音講論，2-8-14，2000.

[7] 三村正人，河原達也．ディクテーションと対話音声認識における音響モデルの差異．春季音講論，2-8-4，2000.

[8] 西村雅史，伊藤伸泰，山崎一孝．単語を認識単位とした日本語大語彙連続音声認識．情処学論，Vol. 40，No. 4，pp. 1395–1403，1999.

[9] 伊藤克亘，山本俊一郎，鹿野清宏，中村哲．ディクテーションにおける日本語の特質を考慮した単語正解率判定ツール．春季音講論，3-Q-19，1999.

[10] 李晃伸，河原達也，堂下修司．単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識．信学論，Vol. J82-DII，No. 1，pp. 1–9，1999.

[11] 小川厚徳，野田喜昭，松永昭一．マルチパス探索における第 2 パス探索法．情処学研報，2000-SLP-30-11，2000.

[12] 李晃伸，河原達也．大語彙連続音声認識エンジン Julius における A*探索法の改善．情処学研報，99-SLP-27-5，1999.

[13] 峯松信明，片岡嘉孝，中川聖一．講演調の話し言葉に対する言語的解析．情処学研報，95-SLP-8-7，1995.

[14] 加藤一臣，李晃伸，河原達也．講演ディクテーションのための話題独立言語モデルと話題適応．情処学研報，99-SLP-26-2，1999.

[15] A. Stolcke and E. Shriberg. Statistical language modeling for speech desfluencies. In *Proc. of ICASSP*, pp. 405–408, 1996.

[16] 三村正人，河原達也．対話音声認識を指向した音響モデルの構築．信学技報，SP00-10，2000.

[17] T.Takezawa, T.Morimoto, and Y.Sagisaka. Speech and language database for speech traslation research in atr. In *In Proc. Oriental COCOSDA workshop*, pp. 148–155, 1998.