002

TR-S-0010

旅行会話タスクにおける連続音声認識システムの診断

Diagnostic Assessment of ATR Spontaneous Speech Recognition Task

河原 達也
Tatsuya Kawahara

2000.9.7

大語彙の連続音声認識は、大規模な音響モデル・言語モデル、複雑なデコーダから構成されるため、改善のための指針を得るのが容易でない。本稿では、旅行会話タスクの自然発話の認識システムにおいて、認識誤りの原因を診断を行なう。その結果、デコーダによるサーチエラーはほとんどなく、音響モデル、特に話し言葉特有の機能語に関連する音素環境に起因する誤りが多いことが確認された。

1 研究の背景と目的

音声認識システムがその学習データも含めて大規模になるにつれて、システマティックな研究・開発が要求される。音声認識は、認識エンジン (デコーダ)・音響モデル・言語モデルの3つのモジュールから構成されるが、誤りがいずれのモジュールに起因しており、さらに個々のモジュールのどの部分が不十分であるかが同定できれば、システムのデバッグ、今後の研究テーマ、さらにはデータ収集のための指針が得られると考えられる。

そこで著者は、確率的音声認識の枠組において、認識誤りの原因となっているモジュールを同定するとともに、その詳細な診断を行なう方法について提案してきた。本稿では、ATR音声翻訳通信研究所で策定された旅行対話に関する自然発話音声認識タスクに対する自動診断手法の適用について報告する。

2 自動診断手法の概要 [1][4]

2.1 誤り原因モジュールの同定

確率的な音声認識の枠組は、図1のようになる。すなわち入力 X に対して、音響モデルによるスコア $P(X/W)$ と言語モデルによるスコア $P(W)$ の積 $P(X/W)P(W)$ を最大にするような W をデコーダにより見つけるものである。ただし実際には、言語モデルスコアの重みや insertion penalty などにより補正されるが、図では簡単のためこれらのパラメータを略している。

ここで、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) と表記し、それぞれに対する音響スコアと言語スコアを求めると、図2に示すような決定木によって、認識誤りを分類することができる。すなわち、正解文の方がスコアが高いのにこれを求められなかった場合はデコーダのサーチエラーであり、そうでない場合は誤った認識結果のスコアを高くしたモデルが原因である。なお未知語による誤りについては、単語辞書に起因するものと分類する。

ただしこの際には、下記の点を考慮して処理を行う必要がある。

1. 正解文の発音の揺れへの対処

   複数の読み (発音) を持つ単語に対しては、認識結果で用いられた方を採用する。

2. 正解文への句読点の挿入

   認識率算定の際には句読点を考慮しなくても、これらは言語スコアや音響スコアに影響を与えるので、認識結果を基に句読点を挿入する。

3. 誤り区間の分割

   入力 (文) 全体に対して図2の処理を適用するよりも、誤り単語毎に原因を同定できる方が望ましい。ただし、言語スコア・音響スコアともに隣接する単語の影響を受ける。単語 3-gram を用いている場合は、誤り単語の前2単語までを考慮して誤り区間とする。これにより複数の誤りの区間が重なる場合はマージする。

なお厳密には、図2の決定木はサーチエラーに関して正しくない。すなわち、正解文のスコアが認識結果より低い場合でも、認識誤りがより少なく、かつスコアが認識結果より高い仮説が存在しうるからである。この問題に対しては、誤り区間を分割することによって実質的に軽減されると期待できる。

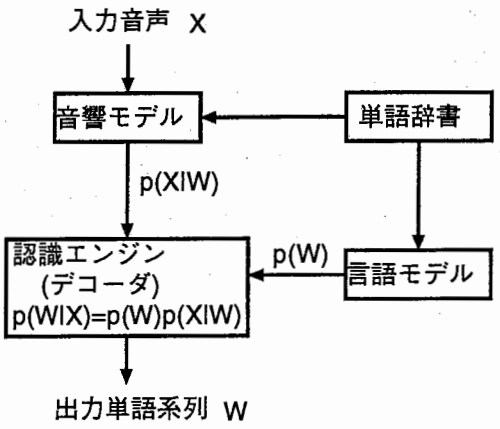

図 1: 確率的音声認識の枠組

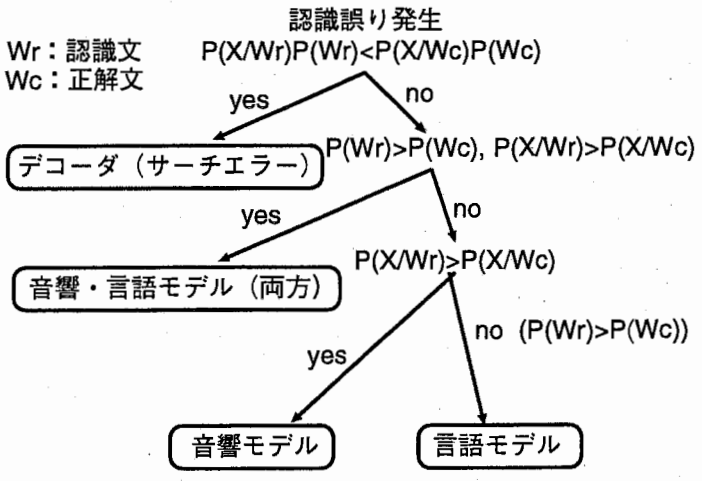

図 2: 誤り原因分類決定木

2.2 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、さらに詳細な診断を行う。音響モデルに関しては、誤りの多い (triphone) 音素コンテクストを求める。またそれらが学習データ不足等により、クラスタリング (代用) されていないかも調べる。言語モデルに関しては、誤り区間に顕著な単語コンテクストがないか調べ、またそれらがバックオフされていないかも調べる。

ただし現状では、サーチエラーを含めて、誤りサンプルを提示し、人手で確認したりトレースしないと詳細な原因はわからない場合が多い。

3 ATR 旅行会話タスクへの適用

3.1 モジュールの仕様

京都大学で作成された診断ツールは、できるだけ汎用性を指向しているが、下記に関して制約があり、対処を行なった。

1. 入力ファイル

   入力の特徴量 (MFCC) ファイルが HTK フォーマットしか扱えない。これは変換プログラムを作成した[1]。CMN についてもできるだけ同じ条件になるように留意した[2]。

2. 音響モデル

   HTK フォーマットのものしか扱えない。音響スコアを計算するモジュールが HTK フォーマットにしか対応していないためである。これについては、ATR の音響モデル (HMnet) を HTK フォーマットの triphone 集合に変換することにより対処した[3]。

3. 言語モデル

   ARPA フォーマットのものしか扱えない。ARPA フォーマットは、CMU-Cambridge ツールキットで用いられている。言語スコアを計算するモジュールが ARPA フォーマットにしか対応していないためである。

   クラス N-gram を扱えないため、ATR の言語モデル (複合 N-gram) をこのフォーマットに変換することができない。そこで学習テキストから単純な単語 3-gram モデルを学習した。カットオフ係数は 2-gram, 3-gram ともに 1 で、バックオフスムージングを用いている[4]。しかしこれにより、言語モデルに関しては ATR のシステムの評価を行なうことができなかった。

4. デコーダ

   認識結果の出力形式の問題であり、軽微な修正で SPREC のエンジンにも対応できるはずであるが、上述の扱える言語モデルの問題により、一貫性のある診断が行なえない。

[1] a2mfc.c
[2] cmsmfc.c
[3] 中村篤氏作成の HMnet2HTK
[4] Mike Schuster 氏による情報提供

表 1: 実験条件

音響分析	26 次元 (MFCC_E_D), CMN なし MFCC (12) + ΔMFCC (12) + power (1) + Δpower (1)
単語辞書	13456 単語, 32400 ベースフォーム (主に語末の sil の有無による)
言語モデル	2-gram: 60717, 3-gram: 108694
音響モデル	3269 triphone models, 1278 states, 9155 Gaussians

そこで京都大学で開発されたデコーダ Julius を用いることとした。参考のために、同一の音響モデル・言語モデルを用いてデコーダ nozomi と認識率を比較した結果、1% 未満の違いであったので、デコーダによる差はほぼ無視できると考えられる。

以上をまとめると、実験条件は以下の通りである。

3.2 正解基準

自然発話においては、言い淀みなどが頻繁に生じるために、認識率などの算出の際の正解基準を定義するのが容易でない。特に ATR では自動翻訳を目的としているため、翻訳するのに関係のない箇所は認識率の算出から除外するという方針をとっているようである。それに対して、厳密な音声認識の枠組では、(それにどういう意味があるかは別として) すべての単語を正しく認識することを目標としている。

また日本語特有の形態素解析やかな漢字表記の問題を含めて、いくつかの技術的問題があり、正解基準に揺れが生じる。

1. かな漢字表記のゆれ

   「私」と「わたし」などの違いは、全く考慮・補正していない。ATR のテキストデータがかなり一貫性があるためか、あまり問題にならないようである。

2. 複合語の扱い

   「八時」と「八＋時」などの形態素区分による違いは、しばしば現れる。特に ATR では複合語の登録を多数行なっているものの、認識率算出の際は分解して集計を行なっているため、問題になった。ただし、診断ツールでは複合語をそのまま扱わないと N-gram による言語スコアが正しく求められない。そこで、認識率算出時に補正を行なった[5]。

   ただし、実際にこれによる認識率の差は 0.2% であり、無視できるほどであることがわかった。実際に多くの研究で、複合語の登録をしてもあまり認識率の改善に結び付いていない。

3. sil(無音) の有無の違い

   認識結果と正解との対応づける際に、単語間や文頭・文末での sil の有無だけが違う場合があるが、通常はこの差は認識誤りとはみなさない。しかし診断ツールでは、この差によって音響スコアが変わるために、厳密に考慮している。

   そこで文末については、認識結果に基づいて正解の方を自動修正した。この補正による認識率の差は 4.6% もあった。

[5]make_wtable.pl を修正

表 2: 誤り原因同定結果

	探索エラー	音響モデル	言語モデル	両モデル	合計
誤り数	159	355	212	439	1165
割合 (%)	3.1	6.8	4.0	8.5	22.5

表 3: 音響モデルが原因とされた誤りに頻出した triphone (出現数)

```
i-m+a (74), e-s+u (71), d-e+s (68), i-t+a (58), s-u+k (55),
a-s+u (54), sh-i+t (51), o-d+e (46), ng-d+e (42), o-k+a (36)
```

また単語辞書には、[a s u] と [a s u sil] のように、語末の sil の有無により 2 通りのベースフォームが登録されている。結果として ATR では、sil は単独で扱われていないようである。この違いは音響スコアは影響を受けるので、診断ツールでは区別して扱っているが、認識率算出のときに、ベースフォームが異なっていても単語 ID が一致していれば正解に数えるオプション (-ec) を用意した。この補正による認識率の差は 2.9% であった。

4. 無意味語リストの扱い (merge.list)

   ATR では翻訳に影響のしない「あのー」などの間投語は認識率の算出から除外している。しかし、ここでは音声認識自体の性能評価が目的であり、またこれらの誤認識が周辺の単語にも影響を与えうるので、特別扱いしないことにした。

   なお、これによる認識率の差 (低下) は約 2% であった。

4 診断結果と考察

対面対話 (SDB) の評価セット 551 文 (話者 42 名) に対する誤り率とその原因モジュールの同定結果を表 2に示す。なおテストセットの総単語数は 5185 で、未知語はない。

これから、ディクテーションなどのタスクに比べて、パープレキシティが小さいためか、探索エラーの割合が小さいことがわかる。これは Julius による結果であるが、Julius は単語間 triphone を第一パスで正しく扱っていないため、nozomi との比較から SPREC ではこれよりさらに 1% 程度少ないと推定される。それに比べて、音響モデルに起因する誤りが多いことがわかった。

そこで、音響モデルが原因とされた誤り区間に頻出した音素コンテキスト (triphone) の上位 10 個を表 3に示す。これらが、「～します」「～です」「～しました」「～ですか」「～ので」などの文末表現に関連するコンテクストであることがわかる。こういった区間はそもそも自然発話では明瞭に発声されないので認識が困難であるが、より精密なモデル化が必要であることを示唆している [3]。

なお参考のため、言語モデルが原因とされた誤りにおいて、2-gram エントリが存在しなかった例を表 4に示す。「あっ」「あの (ー)」「え (ー)」「え (ー)っと」などの間投語を含む組合せがきわめて多かった。これらは不規則に出現するため、十分にモデル化できないためである。また人名・地名などの固有名詞に関連するエントリも多数あった。これらの問題は複合 N-gram

表 4: 言語モデルが原因とされた誤りにおいて存在しなかった 2-gram の例

[] / [あっ] , [] / [え] , [] / [えーっと] ,
[A] / [ライン], [です] / [か] , [です] / [ね], [室] / [の]

では改善されているものと考えられる。なお表 4で「です」に関連するエントリは、品詞が異なるためであると思われる。

この実験の後、用いた音響モデルが、性別依存でなく、CMN を導入していない古いバージョンのものであることが判明したので、最新版での評価を試みた。しかし、ATR のモデルが同一の音素でもコンテキストによって状態数が異なるやや特殊なものであるため、Julius で単語間近似処理が行なえず、適切な評価が行なえなかった。上記の実験では、単語間の音素環境依存性の処理のためにに biphone と monophone が用意されていた。

SPREC による最新の結果 [2] と比べると、9% 程度低い数字になっているが、これは不要語リスト (merge.list) の扱いの差が 2%、デコーダの違いが 1% で、残りが言語モデルと音響モデルの違いであると推定される。

5　むすび

自然発話の連続音声認識システムの評価・診断を行なった。その結果、話し言葉に特有の言い回しに起因する音素コンテキストのモデル化に問題があることが示された。

この結果は予期されたことであり、トラジェクトリモデルなどの従来の HMM より精密な音響モデルや、セグメントモデルなどの音素を越えたモデルが世界的にも研究されているが、より高い精度の発音・音響モデルの研究を今後さらに進めていく必要がある。

参考文献

[1] 河原達也, 南條浩輝, 李晃伸. 大語彙連続音声認識における認識誤り原因の自動同定. 日本音響学会研究発表会講演論文集, 2-1-17, 秋季 1999.

[2] 内藤正樹, Harald Singer, 山本博史, 中嶋秀治, 松井知子, 塚田元, 中村篤, 匂坂芳典. 旅行会話タスクにおける ATRSPREC の性能評価. 日本音響学会研究発表会講演論文集, 3-1-9, 秋季 1999.

[3] 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. 電子情報通信学会技術研究報告, SP99-140, 2000.

[4] 南條浩輝, 加藤一臣, 三村正人, 李晃伸, 河原達也. 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. 情報処理学会研究報告, SLP-31-11, 2000.

(補遺) 京都大学モデルでの診断

京都大学においても、ATR 旅行対話データベースを利用して音響モデルを構築した [3]。ATR のモデルとの主要な違いは以下の通りである。

- HMM の学習に SDB のすべてを利用した。すなわち、学習データ量が大きい。
- 2000 状態 16 混合の triphone モデルを作成した。すなわち、モデルの複雑度が大きい。
- SSS ではなく、HTK の decision tree clustering を用いた。
- 単語 3-gram モデルの学習に SDB と SLDB のすべてを利用している。

その結果、男性のみであるが同一の評価セットに対して、88.5% の認識率を達成した。この結果に対する診断結果を表 5に示す。誤り率が若干大きいのは、発音の違いまで考慮しているためである [4]。

表 5: 京都大学モデルによる誤り原因同定結果

	探索エラー	音響モデル	言語モデル	両モデル	合計
割合 (%)	6.3	1.5	1.3	2.6	11.8

このように HMM の学習量と複雑度を大きくすることにより、音響モデルに起因する誤りは一見ほとんどなくなった。

探索エラーが増加したのは、音響モデルがよくなったことにより、正解のスコアは向上したが、探索が難しくなったケースがあるためである。特に挿入誤りが顕著に増加した。これは、「あの」という発声に対して「あの お」というように短い間投詞が挿入される場合が多い。これらは音響的には微妙な違いであり、局所的な近似誤差によって誤ったものと考えられる。