音声言語情報処理 27−6
(1999. 7. 23)

大語彙連続音声認識における認識誤り原因の自動同定

南條 浩輝　李 晃伸　河原 達也

京都大学大学院 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　音声認識誤りの原因が同定されればシステム改善のための指針を得ることができ、今後の研究の指針やデータ収集時のヒントが得られる。しかし、大語彙連続音声認識においては構成要素となる音響モデル、言語モデルが大規模、かつ統計的モデルであるため認識誤りの原因が何に起因するかを人手で同定するのは容易ではない。本稿では、認識誤りの原因を自動的に同定する手法を提案する。具体的には、正解文を与え、音響モデル、言語モデルから計算されるスコアを認識結果の音響スコア、言語スコアと比較し、認識誤りの原因を音響モデル、言語モデル、探索アルゴリズムのいずれかに同定する。また、一文全体でのスコア比較はモデル改善の指針としては不十分であるため、誤りを含む数個の区間に分割し、区間ごとに原因の同定を行う。探索誤りの場合はそれと提示するにとどめるが、音響モデルが原因であるときは、クラスタリングされて学習されたtriphone、スコアの低いtriphoneを原因と同定し、言語モデルが原因であるときは、低次のN-gramから推定された3-gram,2-gramを原因として同定する。

キーワード　音声認識, 認識誤り, 統計的音響モデル, 統計的言語モデル, 探索アルゴリズム

Automatic Diagnosis of Recognition Errors in Large Vocabulary Continuous Speech Recognition System

Hiroaki Nanjo　Akinobu Lee　Tatsuya Kawahara

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

Abstract　High-quality recognition is required for speech recognition system.If the causes of recognition errors are specified, it is usefull for improvement and next researches. In this paper,we propose the method diagnose errors in the framework of LVCSR.Apply the recognized and correct sentences to an acoustic and a linguistic models,calclate and compare scores,and specify the cause.To be more usefull,separate a sentence to some parts including wrong word,and specify triphones trained together with other triphones and 3-grams and 2-grams not sufficiently trained.

key words　speech recognition, recognition error, statistical acoustic model, statistical linguistic model, search algorithm

1 はじめに

大語彙連続音声認識すなわち任意語彙のディクテーションの研究開発が盛んに行われており [1][2] 、この研究の成果は話題同定や音声理解など、他の関連研究を推進する基盤ともなっている。

大語彙連続音声認識には一般的に、統計的な音響モデルと統計的言語モデルが用いられる。これらのモデルは大規模かつ複雑であり、統計モデルであるため人手による誤り原因診断は困難である。音声認識システムの設計段階において、認識誤り原因を人手で同定し、システムの改善を行うために費やしている労力は設計者にとって大変大きく負担となっている。音声認識誤りの原因を自動的に同定することができればシステム設計者の負担は軽減され費やす時間も短くなり、システム改善に大いに有効である。また、研究者にとっては、認識誤りを起こす要因を分類するによってどのような研究を今後の対象にしていけばよいのかという指針にもなり、データ収集時にどのようなことに注意を払うべきかといった指針にもなる。

こういった背景をもとに本研究では、認識を誤り起こす要因を分類し、現在最もよく使用される大語彙連続音声認識システム、すなわち統計的音響モデルと統計的言語モデルを使用した音声認識システムについて、認識誤り原因の自動同定を試みる。認識結果（以下、認識文）と実際に読み上げられた文（以下、正解文）それぞれを音声データに適用したときのスコアを音響モデルのスコア（音響スコア）と言語モデルのスコア（言語スコア）に分解し、それぞれを比較することにより、誤りが探索過程、音響モデル、言語モデルのいずれに起因するのか大別し、それぞれについて詳細に認識誤りの原因を自動同定する。

2 大語彙連続音声認識の枠組み

2.1 連続音声認識の概観

音声認識は、入力された音声を最もよく説明できる仮説を探しだす問題である。音声認識には文節や単語ごとに区切って発声してもらい、文節や単語ごとに認識を行う離散音声認識と、連続的に発声された音声に対して認識を行う連続音声認識とがある。文節や単語ごとに区切って発声するのは話者に大変な負担がかかるので、そのような制約を課さずに連続的に発声された音声に対して認識を行う連続音声認識が望ましい。しかし、一般に連続音声認識において許される発話の仮説の数は膨大になるので、仮説空間を全て探索することは実質的には不可能である。そこで、種々のモデル（知識）を効率よく適用して、最尤解を見つける機構が必要となる。連続音声認識システムは以下のような主な構成要素からなる。

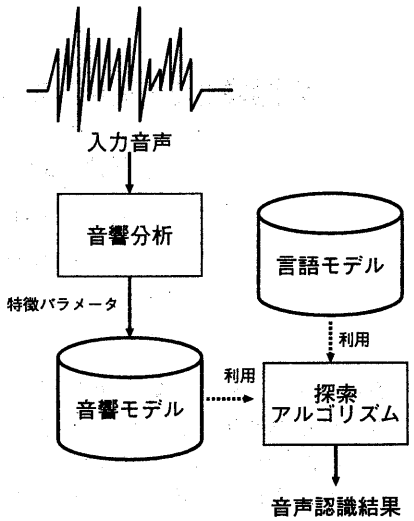

図 1: 音声認識概観

- 音響分析: 音声波形から音声の特徴パターンを抽出する。
- 音響モデル: 音素などの単位の音響的性質をモデル化する。
- 言語モデル: 言語的性質をモデル化する。
- 探索アルゴリズム: 探索空間の中から最適解を効率よく見つける。

これらの関係を図 1 に示す。

2.2 確率論的モデルによる連続音声認識

連続音声認識は確率論的定式化を行うことができる。すなわち、音声認識とは音声パターン X を観測したときに発声内容が W である確率 (事後確率)$P(W/X)$ を最大とするような発話の仮説 W を探

しだすことである。ベイズの定理を用いればこの事後確率 $P(W/X)$ は、

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

となる。

式 (1) の左辺 $P(W/X)$ を最大とする単語列 W を求めるので右辺の分母 $P(X)$ は計算する必要ない。したがって、式 (1) の右辺の分子を最大とする単語列 W を探し出せばよい。これらの確率の要素は、連続音声認識システムの構成要素と以下のように対応する。

- 音響分析:観測した音声波形からパターン X の抽出
- 音響モデル:$P(X/W)$ を求めるモデル
- 言語モデル:$P(W)$ を求めるモデル
- 探索アルゴリズム:積 $P(X/W)P(W)$ を最大とする単語列 W の探索

2.3 スコア比較に基づく誤り原因の分類

音声認識システムに音声データを与えて出力された単語列 W_r(以下、認識文)、事前に用意した、実際に読み上げられた単語列 W_c(以下、正解文) をとして、$P(X/W)$ (音響スコア) と $P(W)$ (言語スコア) をもとめる。$P(X/W_r)$ を認識文の音響スコア、$P(W_r)$ を認識文の言語スコアとよび、正解文に対する $P(X/W_c)$, $P(W_c)$ をそれぞれ正解文の音響スコア、正解文の言語スコアとよぶことにする。

これらを用いて、認識誤りの原因を、音響モデルに起因するもの、言語モデルに起因するもの、探索過程に起因するものと大きく三つに分類する。この分類は、図 2 の決定木に従って行う。すなわち認識スコアより正解スコアが高いにも関わらず正解文が出力されない場合は探索が失敗した、すなわちサーチエラーである。それ以外の場合は音響スコア、言語スコアをそれぞれ比較してエラーの原因を音響モデル、言語モデル、もしくは両モデルに分類する。

提案手法では、文全体でのスコアの比較、判断ではモデル改善の指針としては不十分であるため、誤り単語を同定しそれらを適切な部分区間に分類しその部分ごとに図 2 の決定木に従い、誤り原因を分類し詳細な誤り同定を行う。

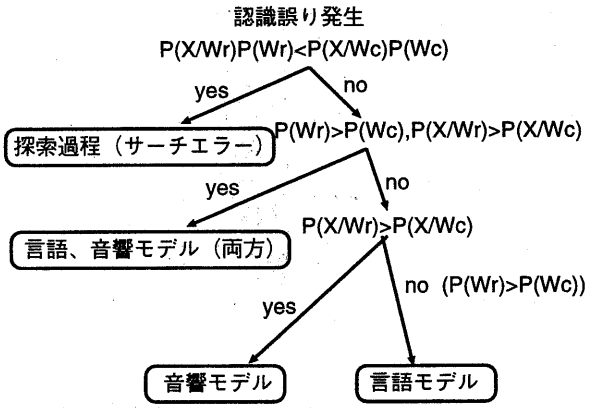

図 2: 誤り原因分類決定木

3 本手法で扱う誤り原因の分類

3.1 音響モデルによる誤り

大語彙連続音声認識で最もよく用いられる統計的な音響モデルは HMM であるが、その学習には、本質的に精度と学習データ量の兼ね合いが問題となる。統計的音響モデルに起因する認識誤り原因には以下のような問題がある。

音響モデルのモデリングの単位として音素を単位とするのが一般的であるが、精度が不十分であり高精度の認識には向いていない。精度が低い理由は、音素が前後の音素の影響をうけパターンが変化することが最大の原因であると考えられている。そのため前後の音素を考慮したモデル (triphone) が広く使用される。しかし、日本語の音素数は 43 程度であり、triphone を構成しようとするとその総数は 43 の 3 乗 (約 8 万, 実際の日本語で許される組は 2 万程度) と膨大になってしまう。全ての triphone を学習を行うためのデータ収集や学習にかかる時間などを考えると現実的でない。したがって、通常は類似した音素環境はまとめて学習をおこなうが、それが誤り原因である可能性が高いといえる。これらを同定、提示することによって今後どのようなデータを収集すべきかといった指針を得ることができる。

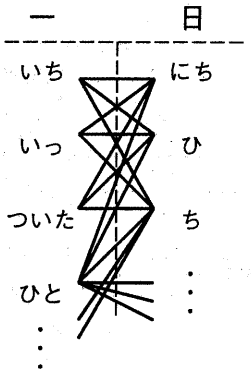

図 3: 様々な読みが許される例

3.2 言語モデルによる誤り

言語モデルに起因する誤りとして以下が挙げられる。統計的言語モデルを使用するため、やはり本質的に精度と学習データ量の関係が問題となる。

- 未知語
  発話に未知語が含まれているかを判定することは困難であり、現在の認識システムでは未知語に対しては必ず認識誤りとなる。
  本手法では、正解文を与えるため、未知語、すなわち辞書にエントリのない単語は直ちに同定できる。

- 学習量の不足
  統計的なモデルである単語N-gramモデルを使用すると推定するN個の単語連鎖(N-tuple)の数は語彙数のN乗になる。これは5000語レベルの語彙でも3-gramだと 125×10^9 個のtupleが、20000語レベルだと 8×10^{12} 個のtupleが必要となる。
  したがって、N-gramモデルは本質的に学習量が不足する。そのため、3-gramが存在しないときは2-gram,1-gramから推定(バックオフ)されるがこれが誤り原因である可能性が高い。

- 単語の読み付与を考慮しないための誤り
  N-gramモデルでは単語の読みを考慮していないと異なる読みに対しても同じ出力確率を割り当ててしまう。図3に形態素列「一 日」の例をあげる。

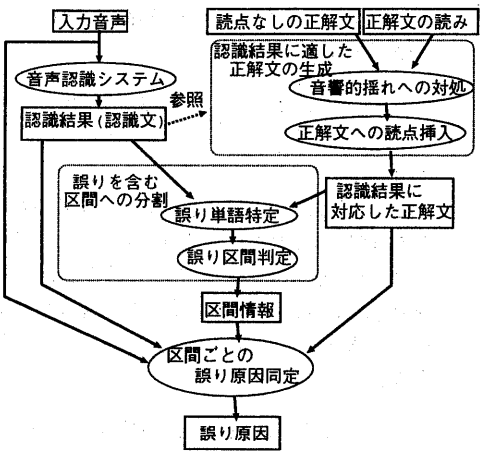

図 4: 原因同定の処理過程

図3中の全ての読みの組合わせを許してしまうため次のような問題が生じる。「一 日」の出現確率が「一 致」の出現確率よりもはるかに高い場合「いっち」と発声された入力音声は、「一 日」と認識される。
これは、読みを考慮した言語モデルを作成することや、読みが異なれば、単語を別のエントリにするといった方法で改善が望めるので、誤りがおきた単語について読みを出力し、開発者に教示する。

3.3 サーチエラー

モデルが正しくても探索過程で失敗することがある。それは、枝刈りされたなどの理由で最尤解が見つからないことがあげられる。しかし、よく間違えるような例を抽出できれば、それをを提示するだけでも探索プログラム(認識エンジン)開発者にとって改善時の指針となり有益である。

4 原因同定手法

認識誤り原因同定プロセスの概観を図4に示す。図4の正解文作成の部分では、正解文の音響的揺れに対処を行うとともに正解文の適当な位置に読点を挿入する。読点とそれに対応するポーズの有無が言語スコア、音響スコアともに影響を及ぼすからであ

る。誤り位置の特定部ではどの単語が挿入、置換、削除誤りとなったかを特定し、誤り区間判定部では認識文、正解文ともに特定された誤り位置を含む適当ないくつかの区間に分割する。最後に各区間に対して誤り原因を同定し出力する。以下に詳しく述べる。

- 正解文の表記の揺れへの対処
  正解文と認識文の間に表記のずれがあると、スコアの比較が困難になる。例えば、「だいたい そんな 様子 です」という文字列が「大体 そんな ようす です」と認識された場合、正解文も認識文と同じにするべきである。このような表記の揺れに対して自動的に対処するツールが作成されているが、人間が対処した場合と比べて 0.5%ほど単語誤り率が増加するので [3] 機械的に対処できない部分には人手で対処する。

- 正解文の音響的揺れへの自動的対処
  日本語では同じ読みに対しても様々な音響的揺れが存在する。例えば、「同定」は「d o u t e i」や「d o: t e i」「d o: t e:」などのように様々に発音される。正確な比較のためには、正解文を認識文と同じにする必要がある。

- 正解文への句読点挿入
  現在の大語彙連続音声認識においては、句読点の有無は誤りとして考えられていない。したがって、句読点は任意の位置にあってよいが、句読点の有無は 3-gram による言語スコアに影響する。また音響モデルでは sp(ショートポーズ) として扱われ、やはり音響スコアにも影響を及ぼす。正確な比較のために、認識文の句読点の位置に対応する正解文の位置に句読点を挿入することによって音響スコア、言語スコアともに認識文と正解文の対応をとる必要がある。

- 誤り区間の判定、分割
  誤り区間は、誤り単語がスコア計算に影響を及ぼす範囲と定める。誤り区間生成の例を図 5 に示す。本研究で使用する認識エンジン Julius[4][5][6] は逆向きの 3-gram を使用しているので図 5 も逆向き 3-gram で示してある。音響モデルが誤り単語の前後の単語のスコアに影響を及ぼし、言語モデルは逆向き 3-gram (Julius で使用) の場合、直前の単語 2 つまでのスコアに影響を及ぼすので、ある誤り単語に対する誤り区間は誤り単語の前 2 単語から後

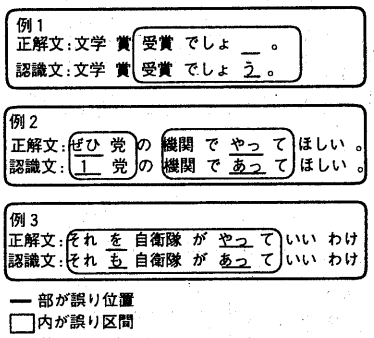

図 5: 誤り区間生成の例

1 単語とする。誤り区間が重なった場合は、それらの誤り単語は同じ区間に含まれるとして区間をマージする。

- 区間ごとの誤り原因同定
  分割された区間ごとに図 2 の決定木に従い誤り原因を分類し原因の同定を行う。区間ごとの詳細な原因の同定は以下に記す通り行う。
  サーチエラーであるものはその旨を出力するにとどめるがそれ以外の場合は誤り単語の読みを出力する。さらに音響モデルが原因である場合、まとめて学習され、別の triphone で代用されている triphone を原因として同定する。言語モデルが原因である場合、存在しない 3-gram,2-gram(より低次の N-gram から推定されている) を原因と同定する。

5 実験及び結果

5.1 診断ツールの仕様

本診断ツールは Julius のモジュールを使用しているため Julius で使うことのできるモデルであればすべて使用できる仕様となっている。入力は認識対象の音声データと認識文、及び正解文とそれらの読みである。

本実験で診断対象とした言語モデル、音響モデル、認識エンジンは IPA のプロジェクトの「日本語ディクテーション基本ソフトウェア 97 年度版及び 98 年度版」[7][8][9] であり言語モデルには語彙数 5000 のものと 20000 のものを用意した。以下にそれらの詳

表 1: Julius(改善前) による認識誤りの分類
Julius(rev.1.1), 語彙数 5000

原因	音響	言語	両方	探索	合計
誤り単語数	8	13	14	38	73
単語誤り率	0.8%	1.4%	1.5%	4.0%	7.6%

表 2: Julius(改善後) による認識誤りの分類
Julius(rev.2.0), 語彙数 5000

原因	音響	言語	両方	探索	合計
誤り単語数	5	15	16	16	52
単語誤り率	0.5%	1.6%	1.7%	1.7%	5.4%

細を示す。

- 音響モデル [10]
  音素モデルには、男性話者 triphone モデル (状態数 2000、混合数 16) を使用した。
- 言語モデル
  語彙は毎日新聞の 91 年 1 月～94 年 9 月までの 45 か月分の記事データにおいて高頻度の形態素 (単語) から構成される。言語モデルの学習には語彙数 5000 には前述の 45 か月分の記事データのものを、語彙数 20000 には、91 年 1 月～94 年 9 月+95 年 1 月～97 年 6 月までの 75 か月分の記事データのものを使用した。
- デコーダ
  5000 語には Julius(rev.1.1) と Julius(rev.2.0) を 20000 語には Julius(rev.2.0) を使用した。
- テストセット
  語彙数 5000 には IPA-97-TestSet を語彙数 20000 には IPA-98-TestSet を使用した。これらは異なるものであるが、いずれも日本音響学会の新聞記事読み上げ音声コーパス (ASJ-JNAS) のうち、音響モデルの学習に用いていないセット合計 100 文の発声からなる。

5.2 実験結果

語彙数 5000 での Julius(rev.1.1) による誤り原因の分類を表 1 に示す。ここで誤り単語数は、挿入、置換、削除された単語の総数である。改善後の Julius(rev.2.0) による誤り原因の分類は表 2 のようになった。また、音響モデル、言語モデルの原因だと同定された誤りに対しても表 4,5 のような代用されていた triphone、3-gram,2-gram が見つかった。

また、語彙数 20000 語のときの Julius(rev.2.0) による誤り原因の分類を表 3 に示す。

表 3: Julius による認識誤りの分類
Julius(rev.2.0), 語彙数 20000

原因	音響	言語	両方	探索	未知語	合計
誤り単語数	7	27	8	70	1	113
単語誤り率	0.4%	1.7%	0.5%	4.4%	0.06%	7.2%

5.3 考察

まず、サーチエラーと同定された単語が多いことについて考察を加える。サーチエラーと同定された区間内でも局所的にマッチした誤り単語により探索が失敗した可能性が考えられる。すなわち音響、言語モデルの誤りが探索誤りを引き起こしている例があり、それらがすべてサーチエラーにカウントされているためにサーチエラーの数が多くなったといえる。

次に、探索アルゴリズム改善後に音響モデル、言語モデルそれぞれによる誤り単語数が変化したことについて考察を加える。言語モデルや言語モデル+音響モデルに原因があると同定された単語が増加した理由は、探索アルゴリズムが改善されたため最尤スコアをもつ結果 (正解文よりもスコアがよい) が得られたという可能性が考えられる。また、音響モデルが原因となっている誤り単語数が減少しているのは、もともと正解文と認識文のスコアが近くアルゴリズム、つまり計算手順を変えたためにその計算誤差により正解文の方がスコアがよくなったためであると考えられる。また先ほどと同様に、別の正解ではない最尤の認識結果が求められ、言語モデルや、言語モデル+音響モデルの原因に分類されたとも考えられる。

6 まとめ

大語彙連続音声認識における認識誤りの原因の自動同定の手法を提案した。

自動同定の方法としては、誤り単語を含む適切な大きさの誤り区間に分割し、それぞれの区間に対し

表 4: 存在しない triphone の例

あるべき triphone	代用 triphone
ky-o+u	y-o+u
sh-u:+k	y-u:+k
j-i+j	y-i+j

語彙数 5000

表 5: 存在しない 3-gram,2-gram の例

一人/で い (tri)
で/い (bi)
ライバル/は 多い (tri)
多く/が 言った (tri)

語彙数 5000

て、正解文と認識文それぞれの出力値（確率）の計算を行うことによって、音響モデル、言語モデル、探索過程のいずれに原因があるかを分類した。さらに、それぞれについて認識システムの性能向上に有用であろう情報を提示した。

本研究室で開発された大語彙連続音声認識エンジン Julius を使用して評価実験を行ったところ、探索過程における誤りを減らすための労力が軽減された。実際に、音響モデル、言語モデルその他の条件を同じにした場合に、Julius の単語認識精度が 92.4%から 94.6%になり、デコーダの性能向上に貢献できた。

今後の課題としては、本ツールを Julius に組み込むことや、サーチエラーとされた区間の中でもそれが、単にサーチエラーなのか、もしくは音響モデル、言語モデルが悪いのかをより詳細に分け、さらに後者の場合、何が悪かったかを同定することが挙げられる。

同様に、言語モデルや音響モデルの精度向上における過程においても開発者の負担を軽減させ、結果、全体として大語彙連続音声認識システムとしての性能向上に貢献していく予定である。

謝辞：本研究では、音響モデル、言語モデル、認識エンジンに情報処理振興事業協会 (IPA) の「日本語ディクテーションの基本ソフトウェア 97 年度版、98 年度版」のものを使用した。

参考文献

[1] 吉田 航太郎, 松岡 達雄, 大附 克年, 古井 貞熙 単語 trigram を用いた大語彙連続音声認識. 情報処理学会研究報告 96-SLP-14-14

[2] 西村雅史, 伊東伸泰, 山崎一孝, 荻野紫穂 単語を認識単位とした日本語の大語彙連続音声認識情報処理学会研究報告 98-SLP-20-3

[3] 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. 音講論,3-Q-19, 春季 1999

[4] 河原 達也, 李 晃伸, 伊藤 克亘, 伊藤 彰則, 宇津呂 武仁, 小林 哲則, 清水 徹, 田本真詞, 荒井 和博, 峯松 信明, 山本 幹雄, 竹沢 寿幸, 武田 一哉, 松岡 達雄, 鹿野 清宏 大語彙日本語連続音声認識研究基盤の整備. -評価用連続音声認識プログラムの開発.- 情報処理学会研究報告 97-SLP-18-1

[5] 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大語彙連続音声認識エンジン JULIUS. 信学技報 SP98-3

[6] 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,pp1-9,1999.

[7] 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, 情報処理学会研究報告 98-SLP-21-10

[8] 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ェア (97 年度版) 音響誌 Vol. 55,No. 3,pp175-180,1999.

[9] 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, 情報処理学会研究報告 98-SLP-26-6

[10] 武田 一哉, 峯松 信明, 伊藤 彰則, 伊藤 克亘, 宇津呂 武仁, 河原 達也, 小林 哲則, 清水 徹, 田本真詞, 荒井 和博, 山本 幹雄, 竹沢 寿幸, 松岡 達

雄, 鹿野 清宏 大語彙日本語連続音声認識研究基盤の整備. -汎用音素モデルの作成.- 情報処理学会研究報告 97-SLP-18-3