論文 / 著書情報
Article / Book Information

論題(和文)	話し言葉音声認識における認識率の変動要因の分析と認識単位の設計
Title	
著者(和文)	篠崎 隆宏, 古井 貞熙
Author	Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	第2回 話し言葉の科学と工学ワークショップ講演予稿集, Vol. , No. , pp. 59-64
Journal/Book name	, Vol. , No. , pp. 59-64
発行日 / Issue date	2002, 3

話し言葉音声認識における認識率の変動要因の分析と認識単位の設計

篠崎　隆宏†　　古井　貞熙†

† 東京工業大学 大学院情報理工学研究科
〒 152–8552 東京都目黒区大岡山 2–12–1

E-mail: †{staka,furui}@furui.cs.titech.ac.jp

あらまし　日本語は単語の単位が明確ではないことから、言語モデルの作成において先ず単語単位の認定が必要となる。単語単位の設計により単語の長さや学習セット中での出現回数が影響を受けるが、これらは音声認識における単語の認識難易度と関係することが示されている。音声認識の立場からは、システムの認識率が最大となるように設計を行うことが望ましい。単語の長さが長く出現回数が多いほど認識に有利となると考えられるが、これらはトレードオフの関係にあり双方を独立に制御することは出来ない。そこで本稿ではまず、個々の単語の単語長や出現回数が認識難易度に及ぼす影響を詳しく分析し、モデル化を行う。次いで、得られた認識難易度モデルに基づく評価値を最大とするように単語を順次併合する、単語単位の最適化手法を提案する。本手法は単語長と出現回数を同時に考慮した最適化手法である。講演音声を対象とした認識実験において、本手法が有効に働くことを確認した。

キーワード　話し言葉音声認識，　日本語話し言葉コーパス，　形態素，　単語単位，　単語長，　出現回数

Error analysis of spontaneous speech recognition and its application to word unit optimization

Takahiro SHINOZAKI† and Sadaoki FURUI†

† Graduate School of Information Science and Engineering, Tokyo Institute of Technology
Ookayama 2–12–1, Meguro-ku, Tokyo, 152–8552 Japan

E-mail: †{staka,furui}@furui.cs.titech.ac.jp

Abstract　For making language models for Japanese ASR (automatic speech recognition), morphological analysis to split sentences into words (morphemes) is indispensable, since there is no spacing between words in the written text and there is even no clear definition of words. It is desirable that the words be designed so as to maximize the recognition accuracy. In this paper, we first analyze the relationship between word correctness of recognition results and two word attributes: each word length and word frequency in the training corpus. Next we propose a method that optimizes the word set in terms of recognition correctness considering both the word length and the word frequency before and after concatenating consecutive words. This method is based on a word correctness model trained by the above-mentioned analysis. We performed a recognition experiment and found that the proposed method improved recognition correctness and accuracy.

Key words　spontaneous speech recognition, Corpus of Spontaneous Japanese, morpheme, word unit, word length, word frequency

1. はじめに

日本語では文章は一つづきの文字列として表記され、単語の明確な定義が存在しない。これは英語などの、単語と単語がスペースで区切られる言語との大きな違いである。現在の標準的な音声認識手法では音声に含まれる言語情報の言語的制約をモデル化するために、言語モデルを用いている。言語モデルは単語のマルコフ連鎖として実現されている。このため言語モデルの作成に際し、日本語のように分かち書きされない言語の場合は先ず単語単位を決め、テキストを単語の列に直す作業が必要となる。(ここでは、「単語」を「形態素」の意味で用いている。)

単語単位の設計により、単語の長さや出現回数などが影響を受ける。すなわち、個々の単語の長さを長くすると、全体の語彙数は増え、各単語の出現回数は減少する。既に単語の長さや出現回数はその単語の認識難易度に影響することが示されており [1]、音声認識の立場からは単語単位をシステムの認識率を最大化するように決めるのが望ましい。

また、単語の単位が明確な言語においても、音声認識の性能の向上の観点から認識に用いる単位を最適化する試みが行われている。文献 [2] では Ngram においてパラメータの増加をさけながら実質的に N を大きくすることを目的として、パープレキシティを基にフレーズを同定し 1 つの単位としてモデル化する手法が示されている。

本稿ではまず、単語の長さと出現回数が単語の認識難易度に及ぼす影響を定量的に分析する。単語が正しく認識される確率を単語の長さと出現回数の関数としてモデル化する。次に、この認識難易度のモデルを用いて認識単位の最適化を行う手法を提案する。単語の長さは認識処理において音声信号とのマッチングという音響的な要因、単語の出現回数はその出現確率の推定精度という言語的な要因と特に関係していると考えられ、本手法は音響的および言語的要因の双方を考慮した認識単位の最適化手法といえる。単語の認識難易度を認識結果の分析から直接モデル化し、単語単位設計時の評価関数として用いる点が本手法の特徴である。

本稿では、先ず 2 章において実験条件を示す。次に 3 章において単語長および単語出現回数が単語の認識難易度に及ぼす影響の分析、モデル化を行う。4 章において単語単位の最適化手法の提案、および提案手法を用いた学習セットの最適化を行う。5 章においてベースラインシステムと単語単位を最適化したシステムの比較を行い、その有効性を確認する。6 章において実験結果の考察を行い、7 章において全体のまとめを行う。

表 1　テストセット a（44 名）の概要

Conference name	No. presentations
人工知能学会	32
日本音響学会	9
国語学会	3

表 2　テストセット b（7 名）の概要

Presentation ID	Short name	Conference name	Length [min]
A01M0035	M035	日本音響学会	28
A01M0007	M007	日本音響学会	30
A01M0074	M074	日本音響学会	12
A02M0117	M117	国語学会	57
A03M0100	M100	言語処理学会	15
A05M0031	M031	音声学会	27
A06M0134	M134	社会言語科学会	23

2. 実験条件

2.1 認識タスク

本稿では日本語話し言葉コーパス (CSJ) の学会講演よりなる、2 つのテストセットを使用する。

テストセット a は認識誤りの分析に用いる。これには 44 講演を用いた。表 1 に講演の概要を示す。分析に使用したのは各講演のはじめの約 10 分間である。

テストセット b は提案手法の効果を評価するために用いる。こちらは 7 人の話者による講演で、認識には講演の全体を用いた。表 2 に概要を示す。

どちらのテストセットとも学習セットに登場する話者による講演は含まれていない。また 2 つのテストセットの中の講演は、全て異なる話者による。

2.2 ベースライン認識システム

言語モデルは CSJ の書き起こしテキストを用いて作成した。形態素解析には、NTT で開発された形態素解析ツール JTAG を使用した。学習セットとして使用したのは 610 講演であり、内訳は多い順に模擬講演が 336、音響学会が 139、言語処理学会が 63 講演、その他 72 講演となっている。形態素数にすると約 1.5M のサイズがある。形態素とその発音のペアをモデル化の単位として用い、2gram と逆向き 3gram を作成した。語彙サイズは 30k である。語彙は学習データ中の出現頻度が上位のものから選択した。3gram のカットオフは 1 とした。フィラーは単に通常の単語としてモデル化した。言い直しは Ngram で有効にモデル化することが難しいことから、学習テキストから取り除きモデル化は行わなかった。言

語モデルの学習には、CMU SLM Tool Kit v2.05 を使用した。

音声は 16kHz で標本化、16 ビットで量子化を行った。音響パラメータは MFCC12 次元、Δケプストラム 12 次元、対数パワーの 1 次差分の計 25 次元で、平均ケプストラムによる正規化 (CMS) を行った。音響モデルは言語モデルの学習に使用した講演のうちで男性話者による 338 講演、約 59 時間から学習したモデルを使用した。音響モデルの学習には HTK2.2 を使用した。

音声認識デコーダは Julius3.1 [3] を使用した。認識実験の際、言語重み、挿入ペナルティはテストセット全体で正解精度が最大になるように調節した。話者間では共通の値を用いている。

音声認識には、エネルギーを用いて切り出した単位をもとに、単語の途中などで切れないように人手により修正した音声単位を用いた。認識単位の切れ目はおよそ 500ms 以上の無音に対応する。

3. 単語の長さおよび出現回数と単語正解率

個々の単語の認識難易度に影響する要因としては様々なものが考えられるが、単語に含まれる音素数および言語モデルの学習セット中での出現回数の影響が大きいことが観察されている [1]。音素数が少ない単語や出現回数の低い単語は認識誤りとなりやすい傾向がある。理由としては一般に、短い単語は認識時に音声とのマッチングが不安定となること、出現回数の低い単語は正確な言語尤度の推定が難しいことなどが言われている。

これらの影響を詳しく調べるため、各単語に対する属性を次のように定義する。

Cor　単語正解率 (%)
NP　単語の音素数
WF　単語の学習セット中での出現回数
LF　出現回数の対数 log10(*WF*+1)

Cor は各単語に対して 100(正しく認識された) か 0(間違って認識された) かどちらかの値をとる。間違いとするのは他の単語と置き換わって認識された場合（置換誤り）および対応する単語が出力されなかった場合（削除誤り）で、認識出力のみが存在し対応する正解単語が存在しない場合（挿入誤り）は対象としない。

表 3 にテストセット a の 44 講演を用いて求めたこれらの属性値の平均と標準偏差を示す。単語正解率の計算において正解単語列と認識結果のアライメントをとる際、複合語処理は行っていない。単語正解率 *Cor* の標準偏差は、*Cor* の値域が 2 値であり、かつ平均が中央寄りにあるため大きな値となっている。単語出現回数 *WF* の平均が 14221 と大きいのは、テストセット中に出現する単語の大部分は出現回数の多い単語であるためである。また標準偏差が平均値より大きくなっている。対数単語出現回数 *LF* でみると、平均値より標準偏差の方が小さくなっている。

表 3　各属性の平均と標準偏差

	Cor	*NP*	*WF*	*LF*
Mean	64.52	3.71	14221	3.20
Standard deviation	47.85	2.18	19829	1.29

認識正解率に対する単語長と出現回数の影響を見るため、図 1 に音素数で層別した単語出現回数と正解率を、図 2 に単語出現回数で層別した音素数と正解率の関係を破線で示す。図 1 において正解率は対数単語出現回数 *LF* を区分し、各区間に属す単語集合において単語正解率 *Cor* の平均を求めプロットしたものである。図 2 は音素数 *NP* を区分し、各区間において同様に平均単語正解率をプロットしたものである。

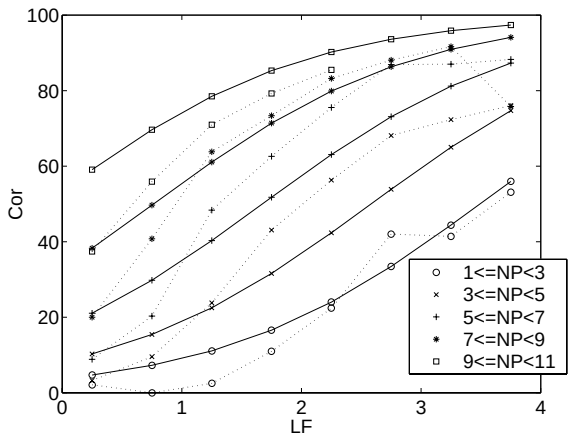

図 1　単語長毎に見た、対数単語出現回数と単語正解率

図 1, 図 2 より、平均単語正解率が音素数と出現回数に応じて大きく変化する様子が観察される。また変化は比較的連続であり、かつほぼ単調であることが分かる。

平均単語正解率は、単語が正しく認識される確率であると解釈することが出来る。この確率をモデル化することを考える。モデル化は以上の観察から、音素数と、出現回数の対数を説明変量とするロジットモデルにより行った。モデルパラメータは最尤推定により求めた。結果を式 (1) に示す。

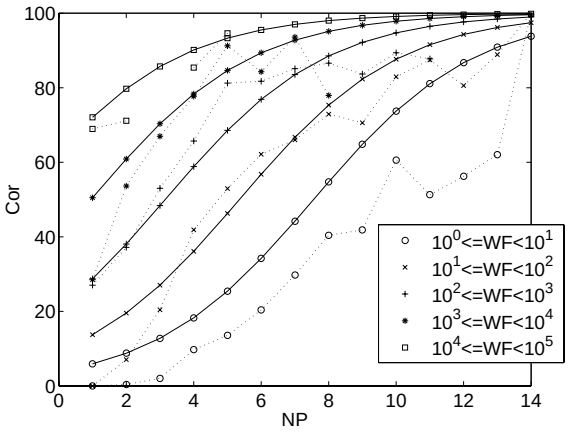

図 2　単語出現回数毎に見た、単語長と単語正解率

$$P(Cor = 100|NP, LF) = \Lambda(0.42NP + 0.94LF - 3.92) \quad (1)$$

ここで Λ はロジスティック関数であり、次式で表される。

$$\Lambda(x) = \frac{e^x}{1+e^x}.$$

また、図 1 および図 2 に単語正解確率モデル (1) により単語が正解する確率を求めた結果を実線で示す。ロジットモデルにより単語正解確率の大局的な特徴がとらえられていることが分かる。なお、単語正解確率モデルのあてはまりの程度を反映する対数尤度は、単語正解確率を定数でモデル化した場合で -59992.2、音素数と出現回数の対数を説明変量とする式 (1) を用いた場合で-52727.9 であった。

4. 認識単位の最適化

前章では単語の音素数と学習セット中での出現回数から単語がデコーダにより正しく認識される確率を有効に見積もることができることを示した。本章ではこの見積もりを基に認識単位を最適化する手法を提案する。最適化は単語を順次併合することで行う。

4.1 最適化手法

テストセットの母集団における単語の出現を言語モデルの学習セットでの出現で近似する。このとき、認識された単語が正解となる回数の期待値 c は、単語正解確率モデル (1) を用いることで式 (2) のように求めることが出来る。認識システムの単語正解率は単語が正しく認識される確率値の期待値 $E[Cor]$ として、式 (3) より見積もることが出来る。ここで $\alpha(x)$ は正解率実効係数である。α は異なった認識単位を持つシステムを統一的に評価するために導入した。たとえば、正解文が「音声認識実験」であるとき、単純な単語正解率を考えると「音声 認識 事件」「音声認識 事件」は同一の結果であるにもかかわらず、正解率はそれぞれ 2/3, 1/2 となり、前者の方が良い結果となってしまう。そこで、「音声」「認識」の重みをそれぞれ 1、「音声認識」の重みを 2 とすると、公平な比較が出来る。

次に、認識語彙中から単語対 $\langle w_1, w_2\rangle$ を 1 組選び、学習セット中でこれらの単語がこの順で連続して出現する箇所を全て 1 単語 $w_{1,2}$ にまとめることを考える。単語対 $\langle w_1, w_2\rangle$ を併合して新しい単語 $w_{1,2}$ に置き換えたときの、$w_{1,2}$ の各属性値は次のように求めることが出来る。

- $NP(w_{1,2}) = NP(w_1) + NP(w_2)$
- $WF(w_{1,2}) = w_1$ と w_2 が学習セット中でこの順に並んでいる回数

$\alpha(w_{1,2})$ は $\alpha(w_1) + \alpha(w_2)$ と定義する。また w_1, w_2 の属性値も、併合の影響を受ける。併合後の語彙中の w_i $(i = 1 \text{ or } 2)$ を w_i' とすると、新しい属性値は以下のように求めることが出来る。

- $NP\left(w_i'\right) = NP(w_i)$
- $WF\left(w_i'\right) = WF(w_i) - WF(w_{1,2})$

併合操作を行う前の $\alpha(w)$ の初期値としては一律に 1 を割り当てる方法、単語 w の文字数を割り当てる方法などが考えられる。前者は併合前の単語単位を基にした正解率、後者は文字を単位とした正解率を使用することに対応する。

単語の併合操作を行う前と後で言語モデルを作成し使用した場合の、システムの単語正解率の比較を考える。$\alpha(w)$ の定義より式 (3) の分母は不変である。また、分子の単語 w に関する和で、変化するのは $w_1, w_2, w_{1,2}$ に関係する項だけである。そこで 2 つのシステムの比較においては、併合前および併合後それぞれについて式 (4), (5) に示す部分和 $pre(w_1, w_2)$, $pos(w_1, w_2)$ のみを考えれば良い。以上のことから、単語対 $\langle w_1, w_2\rangle$ を併合することに対する評価値を式 (6) に示す $\Delta(w_1, w_2)$ で定義する。単語長、出現回数ともテキスト中での単語のコンテキストにはよらないため、評価値は効率的に計算することができる。

単語の併合を評価値の合計が最大となるように繰り返すことで音声認識に適した認識単位が得られる

$$c(NP, WF, \alpha) = P(Cor = 100 | NP, log(WF + 1)) \cdot WF \cdot \alpha, \quad (2)$$

$$E[Cor] = \frac{\sum_w c(NP(w), WF(w), \alpha(w))}{\sum_w WF(w)\,\alpha(w)}, \quad (3)$$

$$pre(w_1, w_2) = c(NP(w_1), WF(w_1), \alpha(w_1)) + c(NP(w_2), WF(w_2), \alpha(w_2)), \quad (4)$$

$$\begin{aligned} pos(w_1, w_2) = {} & c(NP(w_1), WF(w_1) - WF(w_{1,2}), \alpha(w_1)) + \\ & c(NP(w_2), WF(w_2) - WF(w_{1,2}), \alpha(w_2)) + \\ & c(NP(w_1) + NP(w_2), WF(w_{1,2}), \alpha(w_1) + \alpha(w_2)), \end{aligned} \quad (5)$$

$$\Delta(w_1, w_2) = pos(w_1, w_2) - pre(w_1, w_2) \quad (6)$$

と期待できる。単語の併合を繰り返す場合、併合の順番は全体としての評価値に影響を及ぼすが、計算量の点から貪欲法を採用した。最適化手順を以下に optwordunit() としてまとめる。

```
procedure optwordunit() {
 for i = 1:maxiter {
   select word pair <w1,w2>
     which maximizes delta(w1,w2);
   merge(w1,w2);
 }
}
```

4.2 単語の併合

アルゴリズム optwordunit を用い、ベースラインシステムに使用した言語モデルの学習セットに対し、単語単位の最適化を行った。初期状態での異なり単語数は 34895 である。文頭記号、文末記号、読点は併合の対象としなかった。$\alpha(w)$ の初期値としては単語の文字数を用いた。併合は 1000 回行った。併合過程において併合されたはじめの 10 単語対を表 4 に示す。評価式の定義から、これらはじめに選択される単語対は出現回数が多いものとなっている。

表 4　併合された単語対 (はじめの 10 個)

単語 1	単語 2	評価値
と	いう	1573.1
って	いう	1474.4
んです	けど	1056.9
という	こと	958.5
んです	けれども	936.2
あり	ます	838.5
思い	ます	870.3
です	ね	793.9
んです	けども	771.5
おり	ます	766.7

この操作による学習セットの単語数の減少量は 162190 であった。評価値の累積値を α で重み付けした単語総数で割ることにより求めた、予測値としての単語正解率の増加の差分は 1.72%であった。1000 回の併合を行った際の 1 単語対を併合するときの評価値、およびその累積値を図 3 に示す。図より 1 単語対を併合する際の評価値は、併合の繰り返し回数と共に指数関数的に減少することが分かる。

表 5　単語単位最適化前 (*pre*) 後 (*pos*) の学習セットにおける単語属性の平均と標準偏差

	NP	*WF*	*LF*
Mean(*pre*)	3.79	13541	3.23
Standard deviation(*pre*)	2.14	19510	1.20
Mean(*pos*)	4.32	10034	3.00
Standard deviation(*pos*)	2.75	16217	1.19

単語単位の最適化前後での、学習セットにおける単語属性の平均と標準偏差を表 5 に示す。単語の音素数の平均は大きくなり、出現回数の平均は小さくなっている。また、音素数の標準偏差は増えているが、出現回数の標準偏差は減少していることが分かる。

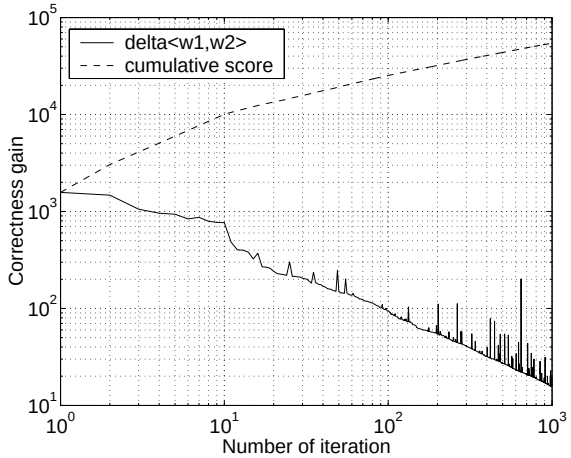

図 3　1000 回の繰り返しにおける評価値とその累積値

5. 単語単位の最適化の効果

単語単位の最適化を行った後の学習セットを用い、言語モデルを作成した。単語単位最適化言語モデルはベースラインシステムに使用したモデルと同様に

語彙数 30k であり、bigram と逆向き trigram を作成した。ベースラインシステムの言語モデルを新たに得られたモデルに置き換えたシステムを用い、7 人の話者よりなるテストセット b を対象として認識実験を行った。ベースラインシステムとの認識結果の比較は文字正解率および文字正解精度を用いて行う。認識実験を行う際の言語重みと挿入ペナルティは、ベースラインシステムと単語単位最適化システムそれぞれにおいてテストセット全体の文字正解精度が最大となるように選んだ。結果を図 4 に示す。

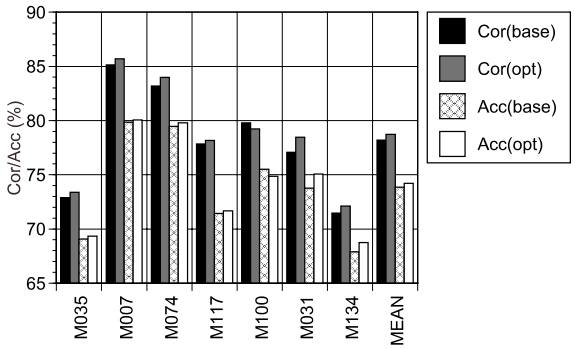

図 4　単語単位最適化前後の文字正解率と文字正解精度

7 講演中 6 講演で改善が見られ、平均して文字正解率で 0.52%、文字正解精度で 0.36%の改善が得られた。

6. 考　　察

認識実験では文字正解率にして 0.52%の改善が得られたが、見積もられた文字正解率 1.72%と比べると、小さな改善となった。

原因の一つとして、単語単位最適化アルゴリズムで使用している単語正解確率の予測誤差が考えられる。ロジットモデル (1) のパラメータ数は定数を含めて 3 と少なく、ある程度の認識結果を用いれば比較的正確なモデルパラメータの推定が行われると期待できるが、平均からはずれた属性値をもつ単語に対しては単語正解確率の予測誤差が問題になると考えられる。この影響は単語の併合過程において単語の長さが長くなるに従い大きくなると考えられる。さらに、ロジットモデル (1) では単語の音素数と出現回数のみの関数として単語の正解確率を求めており、これ以外の要因による影響は考慮していない。また音素数や出現回数の影響は、単語の併合過程において変化しないと仮定している。

これらの予測誤差に対する対策としては、単語の認識難易度に影響の大きい他の要因を単語正解確率のモデルに組み込むことが考えられる。また本稿では単語正解確率のモデルは、ただ 1 つの形態素解析結果を用いて作成した認識システムの認識結果から学習を行ったが、縮約規則を変更した種々の単語単位 (形態素) を用いた複数の認識結果を基に単語正解確率のモデルを学習することが考えられる。これにより、様々な単語単位をもつ複数のシステムにおける単語属性と単語正解確率の関係をモデル化出来ると考えられる。評価式 (6) では単語正解確率の予測値に、その単語の出現回数が掛けられている。このため出現回数の多い単語では誤差が大きく拡大される点を考慮した評価式の改良も必要と考えれる。

予測誤差以外の原因としてはモデル (1) では挿入誤りを考慮していないことが挙げられる。これは正解単語の属性からその単語の正誤を予測するモデルを用いているためである。この対策としは、認識結果の単語の属性からその単語の正誤を予測するモデルを同様に学習し、単語対の撰択に組み込むことが考えられる。

7. ま と め

日本語話し言葉コーパス（CSJ）の講演音声を用いて、単語の音素数および学習コーパス中での出現回数が単語の認識難易度に与える影響を詳しく調査した。単語の正解確率は音素数や出現回数に応じて比較的連続的かつ単調に変化することを示した。さらに、ロジットモデルにより単語が正しく認識される確率をこれらの属性の関数としてモデル化した。この単語正解確率のモデルを用いることで認識単位を最適化する手法を提案し、認識性能の向上に有効であることを示した。今後の課題としては評価関数を改良することにより、単語正解率、単語正解精度の向上をはかることが挙げられる。

文　　献

[1] T. Shinozaki and S. Furui, “Error analysis using decision trees in spontaneous presentation speech recognition,” Proc. ASRU, Madonna di Campiglio, a01ts039, 2001.

[2] E.P.Giachin., “Phrase bigrams for continuous speech recognition,” Proc. ICASSP, Detroit, pp.225-228, 1995.

[3] A. Lee, et al., “An efficient two-pass search algorithm using word trellis index,” Proc. ICSLP, Sydney, pp.1831-1834, 1998.