N-best 音声認識方式に基づく高度音声入力インタフェース

趙　國

Advanced Speech Input Interface Based on N-best Speech Recognition Methods

Kook Cho

論文要旨

音声は人間がコミュニケーションを行なうための最も有効な手段の一つであることから，音声認識技術の進歩とともに人間と機械との音声による対話を実現するインタフェースへの期待が高まっている．カーナビなど一部の音声インタフェースでは，今や実用に足る有用性を得たと言えるが，音声インタフェースが日常社会に普及した例はまだ少ない．近年，音声認識の研究を行なう上での環境が大きく進歩し，さまざまなアプリケーションで音声認識技術が研究されている．しかしながら，音声認識技術はまだ十分ではないと言える．その最大の要因は，音声認識技術の性能と利用形態に関わる制約が大きい点にある．本論文の目的は，音声認識システムの性能を向上させ，人間にとって使いやすいインタフェースを構築するための要素技術を開発することである．

本論文では，まず，音響尤度の算出方法を高度化することによって音声認識システムの性能を改善する手法について検討した．音声認識では，話者の違いによる音響スペクトルの多様性に対応した音響尤度の算出手法が必要となる．音素の音響的特徴は話者の違いにより広い範囲に変動するが，音素間の相対的な関係には話者によらず強い依存性があることが知られている．従来の音声認識における音響尤度の算出ではこのような音素間の相関が考慮されておらず，音素間の相関を利用することによって認識率の改善が期待できる．本研究では，不特定話者音声認識において音素間の相関を考慮することにより，適応データを用いることなく話者の特徴を利用する認識手法を開発した．多数話者による日本語５母音の物理的特徴の広がりを，主成分分析によって次元を圧縮することにより，ある話者の５母音が存在しうる特徴部分空間（話者空間）を構成しモデル化する．音声認識での各文候補において５母音からなる特徴を特徴ベクトル空間上に配置し話者空間までの距離を求め各文候補の妥当性とする．複数の文候補を得る N-best の認識結果に対して，各候補の音響スコアに話者空間までの距離を付け加え再評価し，第１位候補の認識誤り率を 25%削減した．

次に，ユーザにとって使いやすい認識結果の候補提示手法について検討した．近年，統計的手法の利用によって音声認識技術が大きく向上したものの，認識結果を一つだけ提示する場合，正しい結果が常に得られるとは限らないのが現状である．そのため，

認識結果を複数候補提示し，利用者が候補の中から正解を選択する音声インタフェースの研究が行われている．このような N-best 方式の音声認識に基づく音声インタフェースでは，提示する候補数の決定が重要な問題となる．認識候補を多く表示すれば正解が含まれる確率は高くなるが，ユーザが正解を探す手間も増える．そこで，N-best候補の認識スコアの分布を利用して候補提示数を動的に決定する手法を提案した．N-best 候補における認識スコアと候補間の認識スコアの差を利用して候補提示数を決定すれば，正解が含まれる割合（正解提示率）の減少を平均で 1%以内に抑えながら，提示する候補数を平均で 73%以上減らせることを示した．

キーワード

音声認識，N-best，音素間相関，主成分分析，話者空間，母音の正規化，認識スコア，ヒューリスティックス

Abstract

Since spoken language is one of the most effective means for communication for humans, speech interface is favorable to realizing natural and friendly human-machine interface. Improvement of large vocabulary continuous speech recognition can bring certain usefulness to the speech interface. However, there are only few speech interface systems running in our daily life. In recent years, a great progress has been made in the speech recognition technology. The speech recognition is applied to the human-machine interface of various fields. However, the current state of the speech recognition technology is still immature, because there are a large number of constraints on the use of the technology. The purpose of this thesis is to improve the performance of speech recognition, and to develop a fundamental method to construct the interface that humans can effectively use.

Firstly, a technique for improving the performance of the speech recognition by introducing a new method of calculating acoustic scores was investigated. Acoustic features of phonemes spread over a wide range due to different speakers. However, the relative relation among phonemes tends to be kept for different speakers because there are strong dependencies among the phonemes. In a standard speech recognition system, the generative probabilities of the observed feature vectors are estimated independently for each phoneme, and the dependencies between phonemes are ignored. This thesis proposes a novel method of speech recognition using inter-phoneme dependencies which are trained with speech data of a large number of speakers using no adaptation data from the target speaker. PCA restricts the space of acoustic features to a sub-space, called the speaker space, in which a set of feature vectors of a speaker exists. A feature vector set of Japanese five vowels is constructed from recognition result candidates of input sentence and is projected into the original space of acoustic features. A recognition result is represented as a point in the acoustic feature space. Introducing a distance of the point to the speaker space into scoring sentence hypotheses for an N-best scheme

improves the speech recognition performance. The proposed method reduced the word error rate by 25%.

Secondly, a technique for effectively showing recognition results to users was investigated. Recently statistical techniques have greatly improved the performance of speech recognition. The correct result is not always obtained when only the most probable recognition result is shown to the user. Therefore, some speech interface systems display several candidates of speech recognition to the user, and they ask him/her to choose the correct answer from the candidates. In such a speech interface system based on the N-best speech recognition, the determination of the number of candidates to be shown becomes an important problem. When many recognition candidates are displayed, the probability that the correct answer is included becomes high, but the user needs much time and much effort to find the correct answer. This thesis describes a technique of determining the number of candidates dynamically using the distribution of the recognition scores for the N-best speech recognition. The proposed method reduced the number of candidates to be shown by 73%, without degrading the speech recognition rate.

Keywords:

Speech recognition, N-best, Inter-phoneme dependency, PCA, Speaker space, Normalization by phonemic context, Recognition score, Heuristics

目次

目次

図一覧

表一覧

表一覧

第１章

序論

1.1　研究背景と目的

様々な分野でのコンピュータ利用が進むにつれて利用者層が拡大し，誰にでも容易に操作できるインタフェースの開発が求められている．これまで主として用いられてきたキーボード，マウス，CRT などの入出力装置を置き換える，あるいは補う手段として音声の利用が挙げられる．音声は人間にとって最も自然なコミュニケーション手段の一つであることから，音声をコンピュータで処理する音声認識/合成などの技術を確立することによって，自然で使いやすいインタフェースの実現が期待できる．特に，小型化された機器における入出力や，高齢者や障害者などの情報弱者のためのインタフェースでは音声の利用が有効と考えられる．

これら音声インタフェースに関する研究はすでに多く提案されており，1980 年代から 1990 年代初頭には実際に数多くのシステムが構築された[1]．しかし，当時は音声認識の実時間処理は困難であり，実用化に成功したシステムはほとんど無い．ところが，その後の数年間での計算機の爆発的発展により，実時間に近い処理スピードでの音声認識が可能となった．

近年，大規模データに基づく統計的手法の導入により音声認識の性能が大きく向上し，発話内容を文字テキストに変換するディクテーションと呼ばれる処理を行なうソ

フトウェアが市販されている．IBMのViaVoice[2] やDragon Speech[3]など商用音声認識プログラムの登場は，インタフェースを構築する際の音声認識システム導入の壁を低くした．音声インタフェースに対する世間の期待が再び高まったのは，このような背景にもよる．

近年の音声インタフェース実用化の例としては，MIT によるコールセンターを想定した電話音声対話システムが先駆的である[4][5]．日本では京都大学のグループによるバス運行情報案内システムがある[6]．カーナビゲーションでの音声インタフェースの利用も活発である[7]．以上のような研究から，解決しなければならない多くの課題が残されているものの，音声インタフェースは実用化に近づきつつあると言える．しかし，一般の人々が気軽に利用できる音声インタフェースが，我々の日常生活に導入された例はまだ少数にすぎない．その理由として，まず，現在の認識システムは，比較的きれいに発声された読み上げ音声なら高精度で認識するものの，人間同士が普段会話している時のようなくだけた調子の音声は十分な精度で認識できない．そのため，厳密な正確さが必要とされるシステムへの音声インタフェースの導入は現状では難く，さらなる認識率の向上が必要である．また，現状の音声認識技術では避けられないと思われる認識誤りへの対応もインタフェースシステムとして解決しなければならない重要な問題となっている．

本論文では，音声認識の性能を向上させ，人間にとって使いやすいインタフェースを構築するための要素技術の開発について述べている．

現在の音声認識システムは，単語の並びの特徴を表現する言語モデル，単語に含まれる音素の物理的特徴を表現する音響モデル，発話される可能性のある文（単語列）から言語モデルと音響モデルに基づいて最適な文を探索する認識エンジンの３つのモジュールで構成されている．音声認識を行い，その認識結果を一つだけ提示する場合，正しい結果が常に得られるとは限らない．そのため，認識結果を複数候補提示し，利用者に候補の中から正解を選択してもらう音声インタフェースの研究が行われている．認識候補を多く表示すれば正解が含まれる確率は高くなるが，ユーザが正解を探す手間も増える．本論文では，音響尤度の算出方法を高度化することによって音声認識の性能改善を行なう手法と N-best 候補の認識スコアの分布を利用して候補提示数を動的に決定する手法について述べる．

1.2 本論文の構成

本論文の構成を以下に述べる．本論文は５章から構成される．研究の主目的である N-best 音声認識システムにおける音声入力インタフェース高度化については，第３章の話者空間モデルに基づいた音素間相関を用いる手法と第４章の N-best 音声認識における候補提示手法を通じて述べる．

第２章では，本研究の基盤となる連続音声認識技術に関する基礎知識などについて述べる．

第３章では，話者空間モデルに基づいた音素間相関を用いる音声認識について述べる．ここでは，主成分分析を用いて話者空間を構成し，母音の音素コンテキストによる正規化を行なうことにより，適応データを用いることなく高精度で認識を行なう手法を提案する．また，提案手法と従来法との比較を行なって，有効性を示す．

第４章では，音声認識における候補提示数の決定について述べる．認識スコアの分析手法とその結果，単語信頼度の分析とその結果について検討する．また，それらを用いた認識候補提示数の決定手法について提案し，有効性を示す．

最後に第５章で本論文を総括し，結論とする．

参考文献

［1］ 速水悟，菅村昇，“音声対話システムの研究と実用化の動向，” 日本音響学会誌，Vol.50, No.7, pp.574–580, July 1994.

［2］ IBM ViaVoice, http://www-6.ibm.com/jp/voiceland/

［3］ http://www.scansoft.com/naturallyspeaking/, http://www.scansoft.co.jp/naturallyspeaking/

［4］ S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, V. Zue, “Galaxy-II: A Reference Architecture for Conversational System Development,” Proc. 5th International Conferences on Spoken Language Processing (ICSLP98), pp.931–934, 1998.

［5］ V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T.J. Hazen, L. Hetherington, “JUPITER: A Telephone-Based Conversational Interface for Weather Information,” IEEE Trans. Speech and Audio Processing, Vol.8, No.1, pp.100–112, 2000.

［6］ 安達史博，河原達也，奥乃博，岡本隆志，中嶋宏，“VoiceXML の動的生成に基づく自然言語音声対話システム，” 情報処理学会研究報告（音声言語情報処理研究会），2002-SLP-40-23, 2002.

［7］ 有田正剛，島津秀雄，“カーナビゲーションシステム用音声対話インタフェース，” 人工知能学会研究会資料，SIG-SLUD-9502-1, 1995.

参考文献

第２章

統計的手法に基づいた音声認識

2.1 はじめに

音声認識技術は，その認識対象とする単語辞書の大きさによって小語彙（数十単語程度），中語彙（数百単語程度）と大語彙（五千単語以上，数万単語程度）に分類することができる．一般に，特定のタスクドメインを定めることによって，音声認識での認識の対象とする語彙を限定することはある程度可能であり，そのタスク操作は音声インタフェースの利便性に大きな影響を与える．音声インタフェース内で音声認識を利用することを考えると，ユーザが発話した自然な自由文章を認識できなければならない．しかし，ある特定の事柄を指し示す発話においても，その文章表現は多種多様に変化することが自然であり，その中で使用する単語を小語彙に限定することは難しい．また，システムの単語辞書に無い未知語を含んだ発話が入力されたとき，現在の音声認識では認識不可能な未知語が認識誤りを起こすのみではなく，その前後にまで影響が派生することが知られている．結果として，辞書に含まれる単語で構成される誤った文章が出力され，認識後の対話処理に悪影響を与えかねない．例えば，2 万語の単語辞書を用いることで 75 か月分の新聞記事に出現する全単語の約 97%を被覆することができる[1]．固有名詞など残りの単語の対処が必要になる場合もあるが十分な被覆率であると言えよう．つまり，大語彙単語辞書は，実際上必要なほとんどの

語彙をカバーすることができ，これは音声インタフェースを構築する際には有利に働く．一方，初期の音声認識には，単語単位などの離散発話を強制するシステムも存在したが，これでは利用者の負担が大きい．よって，自然文章発話を認識できる連続音声認識は音声インタフェースでは必須であると考える．以上をふまえ，本論文においては連続音声認識を音声インタフェースシステムの基盤技術として扱うものである．

本章は，その連続音声認識に関する基礎知識の導入を目的とする．以下，2.2 節，2.3節及び2.4節では，音声認識の原理や音声認識の重要な構成要素である音響モデル及び言語モデルについて概説する．現在，最も広く用いられている HMM 音響モデルと N-gram 言語モデルなどを取り上げる．2.5 節では，音声認識アルゴリズムを，2.6 節では，話者適応を取り上げる．

2.2　音声認識の情報理論的定式化

入力音声をテキスト化する音声認識技術は，入力音声の特徴ベクトルの時系列パターン X（フレーム数: n）を

$$X = x_1, x_2, \cdots, x_n \tag{2.1}$$

としたとき，X を観測して最も尤度の高い単語列である

$$W = w_1, w_2, \cdots, w_m \tag{2.2}$$

を探索する問題として考えることができる（単語数: m）．

つまり，音声認識は事後確率 $P(W/X)$ を最大にする単語列 W を探す問題となる．ここで $P(W/X)$ に対してベイズの定理を用いると次のように変形ができる．

$$P(W \mid X) = \frac{P(W) \cdot P(X \mid W)}{P(X)} \tag{2.3}$$

（2.3）式の分母 $P(X)$ は，入力パターン自体の生起確率であり，単語列 W には無関係である．よって，音声認識は（2.3）式より，

$$P(W)\cdot P(X\mid W) \tag{2.4}$$

を最大にする W を求める問題と考えることができる．（2.4）式の $P(W)$ は，単語列の事前確率であり，入力 X とは無関係な確率である．この単語列の出現確率を与えるモデルが言語モデルである．$P(X/W)$ は，単語列 W を発生したときに，特徴ベクトル時系列 X が観測される確率で，この計算に用いるモデルは音響モデルと呼ばれる．

2.3 音響モデル

音響モデルは，音素の並びを統計的に学習したものであり，HMM（Hidden Markov Model;隠れマルコフモデル）によるモデル化が広く使われる[2]．HMM は特徴ベクトル時系列の確率モデルであり，自己遷移を持つ複数の状態間を遷移することで，音声のような長さの一定しない時系列信号を効率良くモデル化することが可能である．

HMM は c 個の隠れ状態の有限集合 $\Omega=\{w_1,\cdots w_c\}$，出力記号の有限集合 $S=\{v_1,\cdots v_m\}$，状態遷移確率 $A=\{a_{ij}=P(w_j(t+1)\mid w_i(t))\}$，記号出力確率分布 $B=\{b_{ij}=P(v_k(t)\mid w_j(t))\}$，初期状態確率分布 $\pi=\{\pi_i\}$．以上の 5 項組 $\lambda=\{\Omega,S,A,B,\pi\}$ で表される．HMM は確率的に状態遷移を繰り返しながら，記号を出力する．a_{ij} は状態 w_i から w_j への遷移確率を示し，b_{ij} は状態 w_i から出力記号 v_j を出力する確率である．

モデルが観測系列 $V^T=\{v(1),v(2),\cdots,v(T)\}$ を生成する確率は，HMM の定義により，

$$P(V^T)=\sum_{r=1}^{r_{\max}}P(V^T\mid w_r^T)P(w_r^T) \tag{2.5}$$

ただし，各 r は T 個の隠れ状態の特定の系列 $w_r^T=\{w(1),w(2),\cdots,w(T)\}$ を示す．c 個

の隠れ状態を持つ一般の場合には，$r_{\max} = c^T$ 個の可能な項が存在する．

HMM の学習には EM（Expectation Maximization）アルゴリズムと呼ばれる最尤パラメータ推定手法が用いられる．

基本となる音響モデルは，単純に音素ごとにモデル化を行なった monophone モデルである．しかし，一般に音素の音響的特徴は，前後の音素環境により大きく変化することが知られている．そのため，ひとつの音素に対して，その先行・後続音素（音素環境）に依存して複数のモデルを用意する．このようなモデルをコンテキスト依存モデルと呼ぶ．コンテキスト依存モデルとしてモデル化することで，音響的特徴をより精密にモデル化できると考えられ，多くの音声認識システムにおいて利用されている[3]．コンテキスト依存モデルは例えば，「この研究室の歴史を知りたい」の「歴史」の部分が次のような音素の系列になるとする．

r　e　k　i　sh　i

このとき，先行・後続音素を考慮したモデルの系列は以下のように表現することができる．

o-r+e　r-e+k　e-k+i　k-i+sh　i-sh+i　sh-i+o

真ん中の中心音素を基準に「－」の前の音素は先行音素を「＋」の後の音素は後続音素を意味する．例えば，**o-r+e** は，中心音素が **r** で，その一つ前の音素が **o**，一つ後の音素が e であることを示している．

このようなモデルは，音素の 3 つ組み（triphone）に依存するため，triphone モデルと呼ばれる．各 triphone モデルは，中心の 1 音素分の時間長だけをモデル化する．

前後の音素環境を考慮するため triphone モデルは高い認識精度を得ることができるが，通常，音素は数十種類あるため，組み合わせにより triphone モデルの総数は膨大なものになる．それにともない，各モデル当たりの学習データは極端に少なくなり，適切なモデルパラメータを推定することが難しくなる．更に，大量の学習用音声データを用意しても，すべての triphone がデータ中に出現することは期待できず，学習データに存在しない triphone に対応するモデルをつくることができないという問題が起こってくる．そこで，状態ごとに音響的特徴が近い triphone を共有し，モデル数を削減した状態共有 triphone が用いられる．一方，HMM では状態ごとに混合正規分布を用いて出力確率分布を構成するため，異なるモデルや状態間で混合正規分

布のための正規分布を共有することで効率的なモデルを構成することができる．これを Tied-Mixture モデルと呼ぶ．特に中心音素が同一である triphone の間だけで分布を共有する手法は，PTM（Phonetic Tied-Mixture；音素内タイドミクスチャ）モデル[4]と呼ばれる．PTM モデルは，monophone モデルから出力確率分布を，状態共有 triphone モデルから状態共有構造を抽出し，出力確率分布を状態共有構造に重みづけを行い作成される．

2.4 言語モデル

音響モデルが話者性や音声入力環境などの音声認識における音響的特徴を担うものであるのに対し，言語モデルと単語辞書は，言い回しなどの文章表現や認識対象単語などの言語的特徴を定めるものである．言い換えれば，言語モデルは音声認識システムにおいて認識対象となるタスクを決定する要素である．言語モデルと単語辞書の中で定義されていない文章表現や単語を音声認識では受理することは困難なので，音声インタフェース内で使用することもできない．多様な言語的特徴を柔軟に受理できる言語モデルを使用することが求められる．

音声認識の言語モデルには，文脈自由文法などの有限状態ネットワークで記述された記述文法やコーパスから統計的な手法によって確率推定を行なう統計的言語モデルが用いられる．通常，自動販売機用の音声認識タスクなどの狭い認識対象に限定しても良いタスクの場合には，文法記述型の音声認識を用いることが多い．しかし，記述文法では，システムはあらかじめ想定された文法内の発話のみしか受理できず，発話の文章表現や語尾などの発話様式までも限定されてしまう．認識対象語彙が比較的小さくなる事や複雑な文法を開発者が記述する必要があるなどの問題点も知られている．一方，統計的言語モデルを用いた大語彙連続音声認識では，認識結果を開発者があらかじめ決定的に定義することは難しく，一見するとアプリケーションに組み込むのには向かない．しかし，柔軟に様々な発話を受理することが可能であるため利用する価値は高い．

2.4.1 単語 N-gram

現在，音声認識において最も良く利用される統計的言語モデルは単語 N-gram モデ

ルである．単語 N-gram モデルは単語連鎖のマルコフモデルで構成され，単純なモデルでありながら効果が大きい[5].

言語モデル *P(W)* による *n* 単語からなる単語列 $w_1 w_2 \cdots w_n$ の生起確率は以下の(2.6)式で表すことができる.

$$P(w_1 w_2 \cdots w_n) = \prod_{i=1}^{n} P(w_i \mid w_1 w_2 \cdots w_{i-1}) \tag{2.6}$$

しかし，この確率を推定するのは現実的には不可能であるため，N-gram モデルでは，ある単語の生起が直前の N-1 単語の生起にのみ依存するという近似によって，単語列の生起確率を推定する．つまり，（2.6）式は以下のように近似することができる.

$$P(w_1 w_2 \cdots w_n) = \prod_{i=1}^{n} P(w_i \mid w_{i-N+1} \cdots w_{i-1}) \tag{2.7}$$

各々，N=1 のときは，1-gram（ユニグラム），N=2 のときは，2-gram（バイグラム），N=3 のときは，3-gram（トライグラム）と呼ばれ，現在の音声認識の言語モデルでは 2-gram または 3-gram モデルを使用することが多い.

2.4.2 有限状態文法

言語モデルは大きく分けて，決定的な文法と確率統計的なモデルに分類され，前者は通常人手で記述され，後者はテキストコーパスから自動的に学習される．文法では，最も単純な有限状態文法（有限オートマトン）がよく用いられ，事前にオートマトンに展開することにより，効率的な照合・予測・枝刈りが可能になる．有限状態文法は認識したい言葉をテキストとして用意する．有限状態文法は，記述能力はあまり高くなく，そのため比較的単純なモデルにのみ使われる．有限状態文法には決定性（deterministic）のものと非決定性（nondeterministic）のものがある．決定性有限状態文法とは解析過程にあいまいさがないようなオートマトンを言う.

例としてここでは「どこの研究室ですか」「山下研です」などの発話を受理する「研究室検索システム」用の有限状態文法を取り上げる．まず，単語間の構文制約（単語

間の接続に関する制約）と単語を記述する．記述した文法と単語で構成された文なら認識できるが，記述されてない文法と単語で構成された文は認識できない．登録単語として「どこ，ここ，の，研究室，研，山下，寺井，です，ですか」が，文法として「～研」「～研究室」の前は「どこ，ここ，の，山下，寺井」が，後は「です，ですか」が接続できると記述されたとする．この場合，例えば，「どこの研究室ですか」「山下研です」以外に「ここの研究室です」「寺井研ですか」などの文は認識できるが，「向こうの研究室ですか」「そこの研です」などの文は認識できない．有限状態文法の例を図 2.1 に示す.

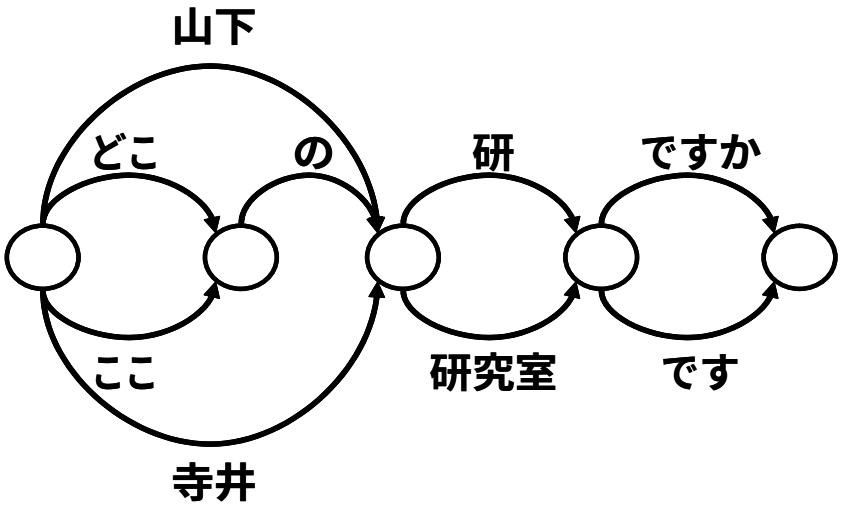


図 2.1: 有限状態文法の例

Figure 2.1: An example of finite-state automaton.

2.4.3 パープレキシティ

音声認識システムの評価結果はタスクの複雑さに依存しており，異なったタスクにおける評価結果を直接比較することはできない．そこである尺度に基づいてタスクの複雑さを規定する必要がある．現在，一般にタスクの複雑さを表す尺度としてパープ

レキシティ（Perplexity）が広く用いられている．パープレキシティは以下の式で求められる．まず，言語 L における単語列 $w_1 \cdots w_n$ の生成確率を $P(w_1 \cdots w_n)$ とすると，言語 L の単語あたりのエントロピーは，

$$H(L) = -\frac{1}{n} \sum_{w_1 \cdots w_n} P(w_1 \cdots w_n) \log_2 P(w_1 \cdots w_n) \tag{2.8}$$

と表すことができる．（2.8）式は，言語から生成される単語を特定するために必要な情報量であり，ある時点での単語の後に等確率で接続する $2^{H(L)}$ 個の単語の候補があることを示している．よって，

$$PP = 2^{H(L)} \tag{2.9}$$

は，情報理論的な意味での単語の平均分岐数を表しており，パープレキシティと呼ばれる．

音声認識の言語モデルの評価にはテストセットの書き起こしテキストに対するパープレキシティが用いられる．これをテストセットパープレキシティと呼ぶ．パープレキシティが低いということは，テストセットに含まれる単語列の出現する確率が高く，テストセットに対して高い性能を持つ言語モデルであると言える．

ただし，パープレキシティによる評価は必ずしも音声認識の性能には結び付かない．これはパープレキシティがある時点に生起した単語に等確率に接続する単語候補数を表したものであり，そのある時点での単語自体の間違いやすさという基準が含まれないためである．一般に単語数が小さいほど一つの単語に割当てられる確率は大きくなるからパープレキシティは低下する．一方，未知語率が小さい言語モデルではパープレキシティが増大することが多い．厳密には単語数や未知語率が異なる言語モデルをそのまま比較することは難しい．この問題を対処した未知語率を考慮する補正パープレキシティも提案されている[3].

2.5 音声認識アルゴリズム

音声認識システムは音声分析部（Feature Extraction）とデコーダ（Decoder）から構成される．音声分析部では入力された音声波形から短時間周波数分析によって特徴ベクトルを抽出する．近年の音声認識では，特徴量として MFCC（Mel-Frequency Cepstrum Coefficient）を用いることが多い．

デコーダは抽出された特徴ベクトルを入力として，言語モデル（Language Model），音響モデル（Acoustic Model），単語辞書（Word Dictionary）を用いた尤度計算により入力音声のテキスト化を行なうプログラムである．デコーダは膨大な認識候補の中から最も尤度の高い解を探索するアルゴリズムによって構成されている．連続音声認識では，単語音声認識などの連続ではない離散発声音声認識と比較すると，仮説の時間方向への曖昧性から出現する仮説数が莫大になる．その探索は複雑かつ困難なものになり，実装には高度なアルゴリズムが要求される．音声認識システムおいて，入力文に対して全ての可能性を計算し，N 個までの複数の文候補を求め，認識結果として最も尤度の高い順に文候補を出力する N-best 方式[6][7][8]が多く利用されている[9]．利用者はその複数の文候補の中から正解を選択する．

2.5.1 Julius と Julian

本研究では，デコーダとしてフリーソフトウェアとして広く利用されている大語彙連続音声認識エンジン Julius と Julian を用いた[9]．Julius と Julian では，認識を 2 段階マルチパスに分けて処理することで実時間に近い実行時間での高い認識性能を得ている．Julius と Julian の第 1 パスでは，フレーム同期ビーム探索を行なう．第 2 パスでは，第 1 パスの認識スコアを先読み情報（ヒューリスティック）として利用し，A*探索によりさらに高精度な認識を行なっている．また，文単位の単語履歴を考慮した仮説の評価を行なえるが，第 1 パスと第 2 パスで言語モデル・音響モデルともに異なり，第 2 パスのモデルの方が高いスコアを与える可能性があるために，A*適格性は満たしていない．そのため最初に得られる解が最尤解である保証はない．したがって，N 個の候補を出力してから，それらをスコアでソートして最尤解を求める．Julius では第 1 パスで 2-gram モデル，第 2 パスで逆向き 3-gram モデルを用いる．

Julius は，数万語の語彙を対象とし，統計言語モデルである単語 N-gram を用いて

認識を行なう．認識率は 20,000 語彙の読み上げ音声で 90%以上である．Julian は，記述文法に基づく認識を行なう．文のパターンを人手で記述した認識用文法（有限状態文法）を用いることで，小語彙の音声対話システムや音声コマンド入力など比較的小規模な音声認識システムを容易に構築できる．

Julius と Julian はオープンソースプログラムとしてそのソースコードとともに仕様は公開されており，その利用実績は多く信頼性も高い．他プラットフォームへの移植や改造も容易で，音声インタフェースの研究のベースとなるプログラムとして適していると言える．

2.5.2 単語信頼度

単語信頼度（確信度，Confidence Measure）とは，認識システムの出す認識結果をどれだけ信頼してよいかを表す尺度である．信頼度の数値が高いほど信頼度が高い，すなわちその単語に競合するような他の候補が見あたらなかったことを示し，低いほど信頼度が低い，すなわち認識過程においてその単語のほかにも似たスコアをもつ単語の候補が多く競合していたことを示す．ただし，この信頼度はあくまで，認識エンジン自身が与えられた言語モデル・音響モデルの元で算出するものにすぎない．信頼度と認識精度の間にある程度高い相関はあるが，信頼度が高いことが必ず正解ということではない．

Julius と Julian では単語の事後確率を用いて信頼度を計算している[10]．認識処理の結果得られた単語グラフ，あるいは N-best 候補のリストにおいて，その中のある単語仮説 w が入力のフレーム τ から t に存在するとき，その単語仮説 $[w;\tau,t]$ の入力音声系列 X に対する事後確率 $p([w;\tau,t]|X)$ は，その仮説をパス上に含むすべての文仮説の出現確率の和より求められる．すなわち，

$$p([w;\tau,t]|X) = \sum_{W \in W_{[w;\tau,t]}} \frac{p(X|W)p(W)}{p(X)}$$

$$= \sum_{W \in W_{[w;\tau,t]}} \frac{e^{g(W)}}{p(X)} \qquad (2.10)$$

ただし，$W_{[w;\tau,t]}$は単語仮説$[w;\tau,t]$をパス上に含む全文仮説の集合であり，$g(W)$はデコーダより得られる文仮説Wの言語モデル上および音響モデル上の出現確率の対数尤度である．$p(X)$はその N-best 候補リストもしくは単語グラフにおける全ての文仮説の出現確率の和として計算できる．この事後確率から，単語仮説$[w;\tau,t]$の信頼度$C[w;\tau,t]$は以下のように定義される．

$$C([w;\tau,t]|X)=\sum_{W\in W_{[w;\tau,t]}}\frac{e^{\alpha\cdot g(W)}}{p(X)} \tag{2.11}$$

ただし，αはスムージング係数$(0<\alpha\le 1)$である．一般に音声認識においでは，対数尤度の値が非常に大きいダイナミックレンジを持つため，少量の上位単語仮説の値によって事後確率の値が支配される傾向がある．係数αはこの尤度のダイナミックレンジを補正するために用いられる．

2.6 話者適応

不特定話者音声認識システムは，多数話者の発声データを用いて学習されるため，話者の発声の多様性に対して頑健であり，特定話者音声認識システムに比べて，使用者が事前に発声する必要がないという利点がある．しかしながら，このような不特定話者音声認識システムは，ある話者の音声だけを事前に学習した特定話者音声認識システムに比べて一般に認識精度が低い．また，一部の話者に対して極端に認識精度が低くなる現象が見られる．これらの問題に対処するために，話者適応に関する研究が行なわれている[11][12][13][14][15][16]．話者適応は，不特定話者音声認識の性能を向上させるために特定話者が事前に発声した適応データを用いて不特定話者音声認識の音響モデルを特定話者へ近づける手法である．この話者適応の代表的な手法として，以下の二つが挙げられる．一つは，多数話者のデータで学習された不特定話者モデルを事前知識として利用し，適応データ量に応じた効率的なパラメータ推定法として，最大事後確率推定法（Maximum A Posteriori probability estimation；MAP 推定法）[11][12][13][14]が用いられている．もう一つの方法は，少量の適応データしか得られ

ない場合に，適応データから学習できない多数のパラメータが存在するので，これらの未学習パラメータを補間する方法として，MLLR（Maximum Likelihood Linear Regression）[15]，移動ベクトル場平滑化法（Vector Field Smoothing；VFS 法）[16]が用いられている．話者適応では音声認識システムの利用者の負担を軽減するために，必要な適応データ量の削減が重要な研究課題となっており，かなり少量のデータでの適応も可能になっているが，これをなくすことはできない．本研究では，不特定話者音声認識において音素間の相関を考慮することにより，適応データを用いることなく話者の特徴を利用する．

2.7　まとめ

本章では，連続音声認識の原理を述べ，音声認識システムの構成や本研究に関わる手法などの説明を通じて本論文で必要になる基礎的知識の導入を行なった．

参考文献

［1］ 広瀬良文，伊藤克亘，鹿野清宏，中村哲，“日本語ディクテーションシステムにおける被覆率の高い言語モデル，” 電子情報通信学会論文誌，Vol.J83-D-II, No.11, pp.2300–2308, Nov. 2000.

［2］ 中川聖一，確率モデルによる音声認識， コロナ社，1988.

［3］ 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.

［4］ 李晃伸，河原達也，武田一哉，鹿野清宏，“Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識，” 電子情報通信学会論文誌，Vol.J83-DII，No.12, pp.2517-2525, Dec. 2000.

［5］ F. Jelinek, “Self-Organized Language Modeling for Speech Recognition,” Language Processing for Speech Recognition, pp.450–506, Mercel Dekker, Inc., 1990.

［6］ Richard Schwartz and Yen-Lu Chow, "The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses," Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'90), Vol.1, pp. 81-84, Albuquerque, NM, Apr. 1990.

［7］ Richard Schwartz and Steve Austin, "A Comparison of Several Approzimate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses," Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'91), Vol.1, pp. 701-704, Toronto, Canada, May 1991.

［8］ Frank K. Soong and Eng-Fong Huang, "A Tree-Trellis Based Fast Search for Finding the N-Best Sentence Hypotheses in Continuous Speech Recognition," Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'91), Vol.1,

pp. 705-708, Toronto, Canada, May 1991.

［9］ http://julius.sourceforge.jp/

［10］ 李晃伸，河原達也，鹿野清宏，“２パス探索アルゴリズムにおける高速な単語事後確率に基づく信頼度算出法,” 電子情報通信学会技術研究報告(音声研究会), SP2003-160, NLC2003-97 (SLP-49-48), 2003.

［11］ C. -H. Lee, C. -H. Lin, and B. -H. Juang, “A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models,” IEEE Trans. Acoustic. Speech and Signal Process., Vol.ASSP-39, No.4, pp.806-814, April 1991.

［12］ J. -L. Gauvain and C. -H. Lee, “Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” IEEE Trans. Speech Audio Process., Vol.2, No.2, pp.291-298, April 1994.

［13］ 中川聖一，越川忠，“最大事後確率推定法を用いた連続出力分布型 HMM の適応化,” 日本音響学会誌，Vol.49, No.10, pp.721-728, Oct. 1993.

［14］ 中川聖一，鶴見豊，“最大事後確率推定法と認識結果を用いた連続出力分布型 HMM の教師なし話者適応化,” 電子情報通信学会論文誌，Vol.J78-DII，No.2, pp.188-196, Feb. 1995.

［15］ C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, Vol.9, pp.171-185, 1995.

［16］ 大倉計美，杉山雅英，嵯峨山茂樹，“混合連続分布 HMM 移動ベクトル場平滑化話者適応方式,”電子情報通信学会論文誌，Vol.J76- DII，No.12，pp.2469-2476, Dec. 1993.

第３章

話者空間モデルに基づいた音素間相関を用いる音声認識

3.1　はじめに

音声認識の実用化において，不特定話者の音声を高精度に認識することが必要となってくる．不特定話者音声認識では，話者による音声のスペクトル特徴の変動をいかに吸収して高い認識精度を達成するかが重要な問題である．話者による変動に対応して不特定話者音声認識で高い認識精度を得るための手法として隠れマルコフモデル（Hidden Markov Model；HMM）を用いた不特定話者音声認識システムに関する研究が近年盛んに行なわれている[1][2]．不特定話者音声認識システムは，多数話者の発声データを用いて学習されるため，話者の発声の多様性に対して頑健であり，特定話者音声認識システムに比べて，使用者が事前に発声する必要がないという利点がある．しかしながら，このような不特定話者音声認識システムは，ある話者の音声だけを事前に学習した特定話者音声認識システムに比べて一般に認識精度が低い．この問題に対処するために，話者適応に関する研究が行なわれている[3][4][5][6]．話者適応は，2.6 節でも述べたように，不特定話者音声認識の性能を向上させるために特定話者が事前に発声した適応データを用いて不特定話者音声認識の音響モデルを特定話者へ近

づける手法である．話者適応では音声認識システムの利用者の負担を軽減するために，必要な適応データ量の削減が重要な研究課題となっており，かなり少量のデータでの適応も可能になっているが，これをなくすことはできない．

現在良く用いられている音声認識システムにおける音響モデルは，前後の音素環境別に推定された HMM と，その HMM の各状態に割り当てられた混合正規分布によって表現される[7][8][9][10]．HMM を用いた不特定話者音声認識システムにおける音響モデルは，話者の違いによる変動に対応するため，広がりの大きいモデルとなっている．このため，ある特定の話者に対しては一般に冗長であり，認識精度劣化をもたらす場合がある．これに対して，話者ごとに適切な特徴パラメータ空間を表現することができれば，認識率の向上が期待できる．ある音素の特徴は発話ごとに変動し，もし話者が異なればその変動はさらに大きくなる．このような事実にもかかわらず，人間が良好に音声を認識できる理由の一つとして，音素間の相対的な関係が比較的安定していて，これを有効に使っていることが考えられる．

通常の音声認識では，観測された特徴ベクトルの生成確率を音素ごとに独立して計算しているため，音素の相対的な関係は無視される．しかし，図 3.1 に示されているように F1－F2 平面での母音のホルマントの相対的な関係が話者によらず比較的保存されているように，音素間に相関があることはよく知られている．このような音素間の相関を音声認識に取り入れることにより，認識率の改善が期待できる[11][12][13]．

本章では，話者の適応データの代わりに多数話者のデータに基づいた音素間の相関を不特定話者音声認識に組み込むことにより，一人の話者の音声として妥当な認識を行なう手法について述べる．ある話者の特徴を 5 母音の特徴を連結して構成した特徴ベクトル（以後，スーパーベクトルと呼ぶ）で表現し，多数話者のスーパーベクトルを主成分分析（PCA）することによって部分空間（以後，話者空間と呼ぶ）を構成する．ある話者のスーパーベクトルはこの話者空間内に存在するはずである．認識結果から得られるスーパーベクトルと話者空間への距離を用いて，認識結果の妥当性を評価する[14][15][16]．

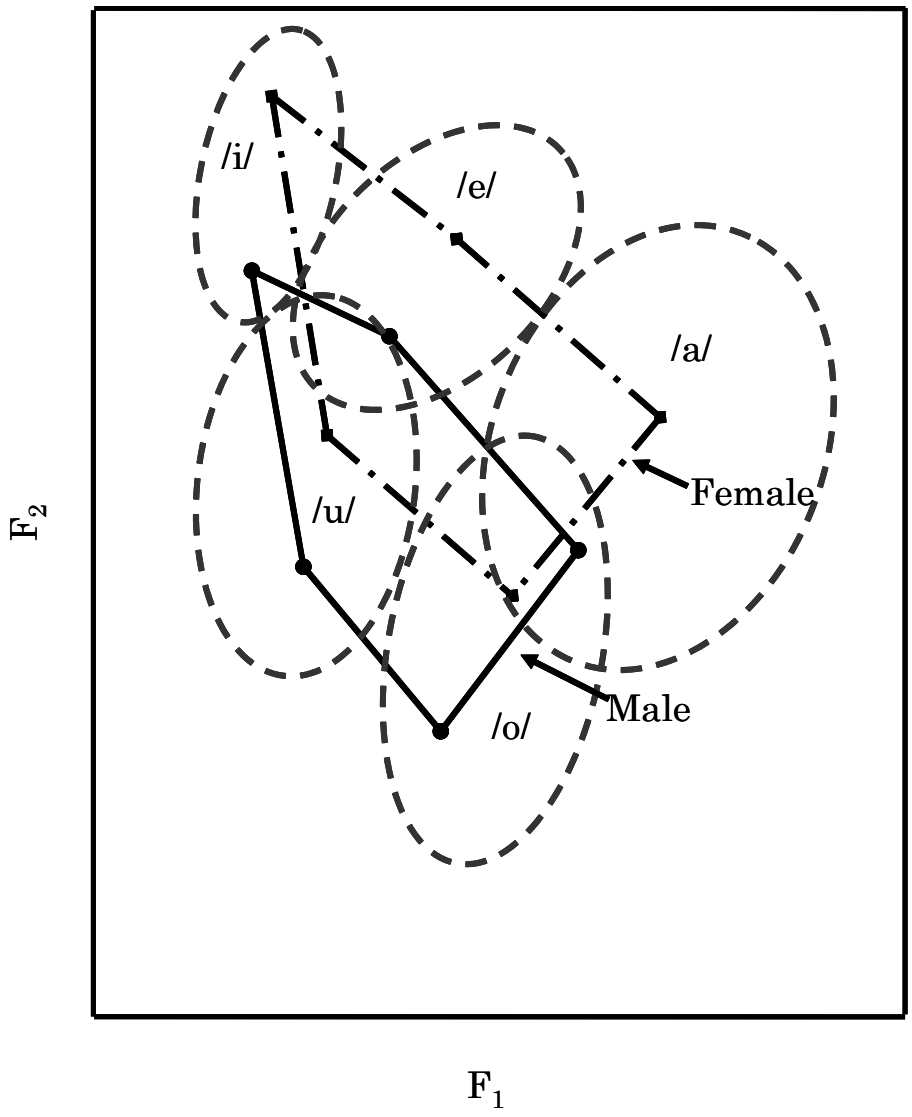


図 3.1: 日本語母音のホルマント

Figure 3.1: Formants of 5 Japanese vowels.

3.2 手法

3.2.1 話者空間

Eigenvoices[17][18] は，顔認識のために用いられる固有値の手法（eigenface）を

基にしている．多数話者のデータによる物理的特徴の広がりを主成分分析を用いて次元を圧縮することにより，話者を表現する特徴部分空間を構成する手法であり，少量の学習データで話者適応をすることが可能となる．一人の話者はこの部分空間上の一点として表現される．図 3.2 (a) は，一次元の特徴パラメータで表現された音素 /a/, /i/ だけから話者空間が構成される場合を模式的に示したものである．図中の一つの点が一人の話者を表す．この場合，話者の特徴ベクトル空間は二次元で，主成分分析により話者空間は一次元の直線として得られている．この結果，各話者は話者空間（この場合は直線）上に射影した点となり，全話者は直線上に分布すると考えることになる．

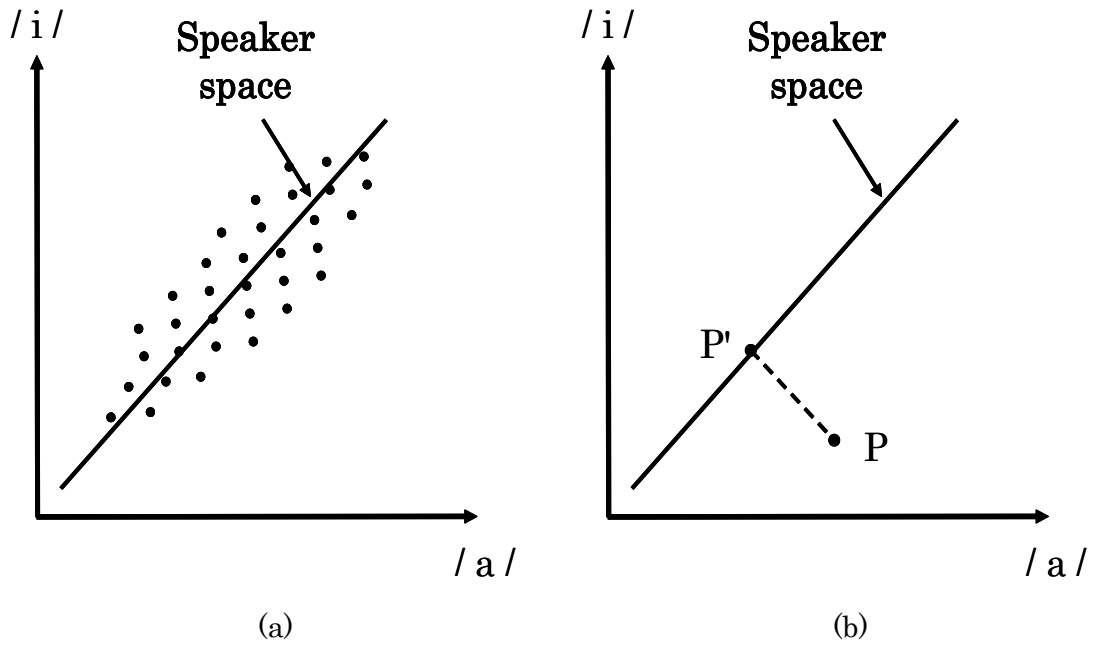


図 3.2: 話者空間

Figure 3.2: Speaker space.

Kuhn らはこの話者空間を用いて，少ない適応データで安定して話者適応する手法を提案している[17][18]．図 3.2（b）は，少ない適応データから得られたある話者の

特徴が点 P となったことを示している．このとき，その点に存在する話者はないと考え，大量学習データで安定して得られた話者空間上の最も近い点 P'に話者の特徴をとる．

提案手法でも同様に，話者空間を用いることにより話者の特徴ベクトルが存在することができる特徴部分空間を構成する．音声認識結果が正しい場合，話者の特徴ベクトルはこの話者空間内に位置するはずである．図 3.2（b）の例のように，音声認識結果から得られた /a/，/i/ の特徴から話者が P の点に配置されたとする．話者空間までの距離 ‖P-P'‖ が大きければ，認識結果における /a/ あるいは /i/ が間違っている可能性が高いと判断できる．

3.2.2 話者空間モデルを用いた音声認識手法

3.2.1 節で述べた話者空間を用いることにより，音声認識における音響尤度の算出に音素間の相関を取り込むことを考える．音声認識結果が正しければ，その結果から得られる音素の特徴に基づいて話者を特徴ベクトル空間に配置すれば，話者はその部分空間である話者空間内に存在するはずである．そこで，特徴ベクトル空間内に配置された話者の位置から話者空間までの距離を算出することにより，認識結果の妥当性を評価する[14][15][16].

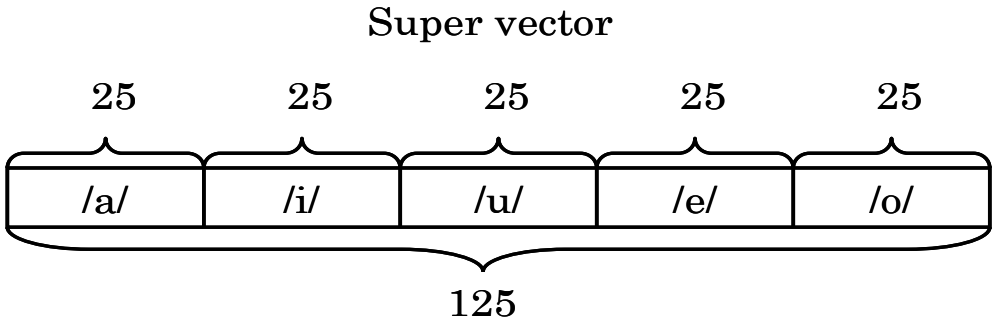


図 3.3: スーパーベクトル

Figure 3.3: Super vector.

認識に先立ち多数話者のデータを用いて話者空間を構成しておく．本手法では，図 3.3 のように話者の特徴をその話者の日本語 5 母音 /a/，/i/，/u/，/e/，/o/ の平均的な特徴を連結したスーパーベクトルで表現する．話者の平均的な母音はその話者の多数の母音区間の中間フレームにおける MFCC（12 次），ΔMFCC，Δ パワーの 25 次元の特徴を母音ごとに平均して用いる．多数話者のデータは強制アライメントにより自動ラベリングされており，そのラベルを基に母音区間の中間フレームを決定する．多数話者のスーパーベクトルを平均を引いて正規化した後，主成分分析を行ない話者空間を得る．

話者空間モデルを用いた音声認識の処理方式を図 3.4 に示す．通常の音声認識により，N-best 候補を生成する．このとき，入力文の各 N-best 候補は音響スコアと言語スコア，音素ごとのアライメント結果を出力する．それぞれの文候補において，音素ごとのアライメントにより母音区間の中間フレームを求め，母音ごとに特徴量を平均し，それらを連結することによってスーパーベクトルを得る．従って，本手法では，入力発話中に 5 母音すべてが出現していることが前提となる．スーパーベクトルと話者空間との距離 D を文候補それぞれの音響スコア S に重み付けして加えた値によって再評価された音響スコア S' を次式のように得る[14][15][16].

$$S' = S - w \times D \tag{3.1}$$

ここで，w は重み係数である．さらに，通常の音声認識と同様に言語スコアを加えて文候補の再評価値を得る．

話者のスーパーベクトルを X，多数話者によるスーパーベクトルの平均ベクトルを $\overline{X}$，主成分を表現する固有ベクトルを並べた話者空間への変換ベクトル C を

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kp} \end{bmatrix}$$

とすると，Xから話者空間への距離 Dは

$$D = \left\| (I - C^T C)(X - \overline{X}) \right\| \tag{3.2}$$

で与えられる．ここで，p はスーパーベクトルの次元数で p=125，k は主成分数，I は $p \times p$ の単位行列である．

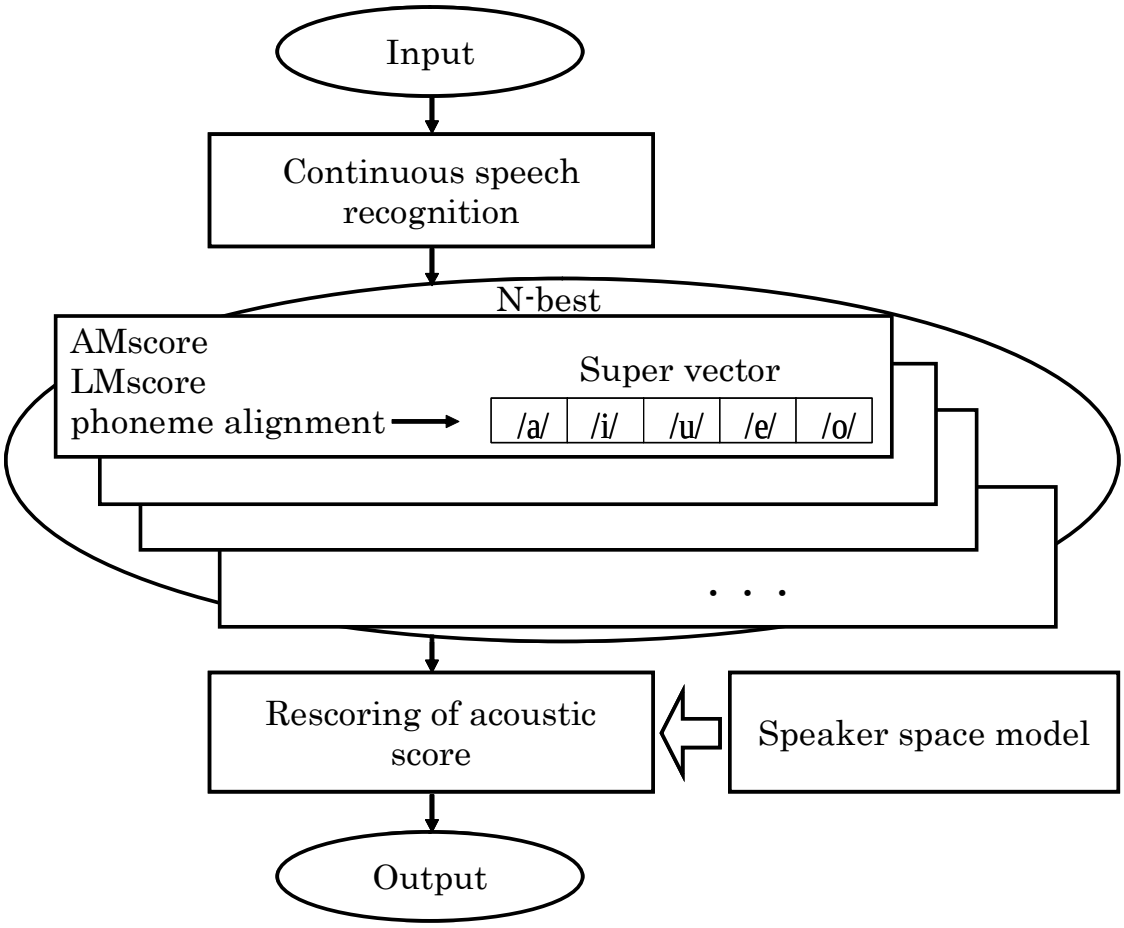


図 3.4: 認識処理手順

Figure 3.4: A flow of speech recognition using the speaker space.

3.2.3 母音の音素コンテキストによる正規化

母音の特徴ベクトルのばらつきは話者の違いだけではなく音素コンテキストの違いによっても生じる．提案手法では話者を平均的な母音で表現するため認識結果において音素コンテキストの影響を取り除いた特徴ベクトルを用いてスーパーベクトルを構成し，話者空間との距離を求めることが好ましい．話者空間を構成するための多数話者データでは，一人の話者における各母音のデータ数が多いため，母音の平均ベクトルにおける音素コンテキストの影響は少ないが，入力音声では，一般に一発話中の各母音の数が少ないため音素コンテキストの影響を大きく受ける．そこで，入力音声に対するスーパーベクトルの作成においてのみ各母音の特徴から音素コンテキストの影響を取り除く．このため，あらかじめ作成されている HMM による音素の monophone モデルと triphone モデルの差に注目して母音特徴から音素コンテキストの影響を取り除いて正規化することを考える．

認識結果（仮説）における母音 v の中間フレームの特徴ベクトルを V とし，先行音素，後続音素を p, f とする．母音 v の monophone HMM および triphone HMM の中間状態の出力確率分布の平均ベクトルをそれぞれ Vv, $Vp\text{-}v\text{+}f$ としたとき，音素コンテキストの影響を取り除いた正規化後の特徴ベクトル V' を

$$V' = V + \left(V_v - V_{p-v+f} \right) \tag{3.3}$$

とする．入力音声の母音区間の音素コンテキストは音声認識によって出力された各 N-best 候補の音素列から得られる．

3.3　実験

提案手法の有効性を確認するために評価実験を行なった．N-best 候補を生成する連続音声認識システムとしては，CSRC2000 年度版のシステムを用いた[19]．デコーダは Julius rev.3.2[20]のスタンダード版，言語モデルは語彙数 20K の 3-gram，音響モデルは「新聞記事読み上げ音声コーパス」で学習した 64 混合分布，3000 状態の PTM モデル[9]を用いた．

3.3.1 使用データ

話者空間を構成するための多数話者データとして，JNAS コーパス[21]中の 306 人分（男・女各 153 人分）の新聞記事読み上げ音声を用いた（1 人当り 102 文～162 文）．本手法では，入力音声中に 5 母音すべてが出現していることを前提としているため，音声認識の評価には，JNAS コーパス中のテスト文，男・女各 100 文の中で 5 母音すべてが出現する文の中からさらに，N-best の候補文にも 5 母音すべてが出現している 139 文（男：71 文，女：68 文）だけを評価実験に用いた．

母音の特徴量の正規化のために用いる monophone，triphone モデルとしては連続音声認識コンソーシアムによって提供されている monophone model（gender-independent, 16 mixtures）および，triphone model（gender-independent, 16 mixtures, 2000 states）を用いた．

3.3.2 母音の音素コンテキスト正規化の効果

母音の音素コンテキストによる正規化が正しく行なわれているならば，実際観測された話者の各音素の特徴ベクトルは正規化前より正規化後の方が話者の各音素の平均特徴ベクトルとの距離が小さくなるはずである．

図 3.5 は 4 人の話者の文中の母音の特徴ベクトルとその話者の同じ母音の平均特徴ベクトルの差の平均を母音の音素コンテキストによる正規化の有無で比較した結果である．話者番号 M1，M2 が男性で F1，F2 が女性である．

これより，正規化後の母音の特徴ベクトルが正規化前の母音の特徴ベクトルよりその話者の平均的な母音に近くなっていることがわかる．

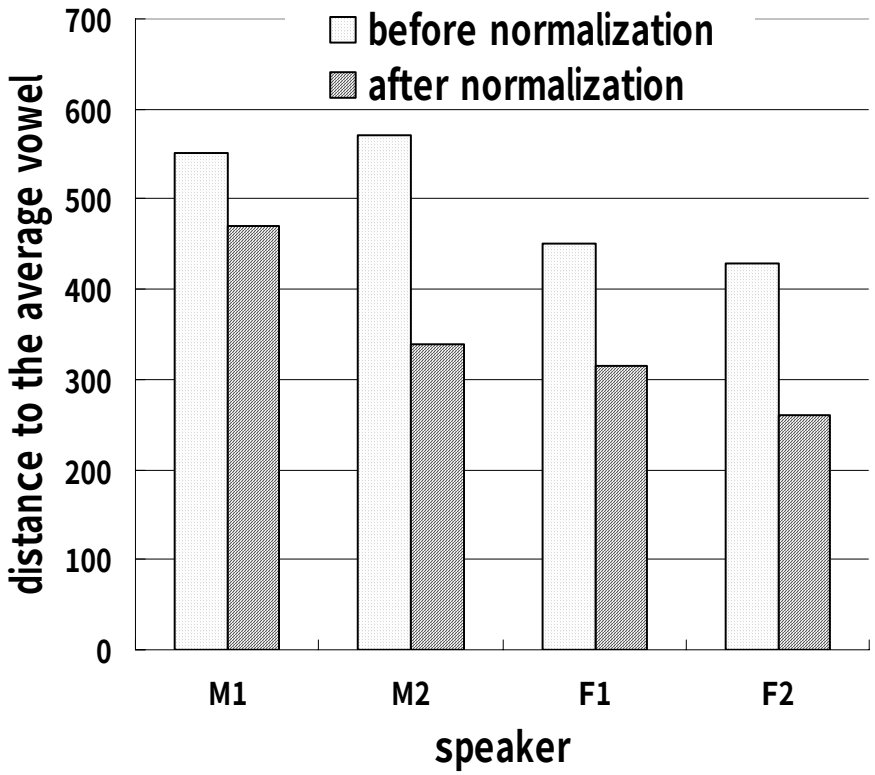


図 3.5: 発話文の母音と平均母音との距離

Figure 3.5: Distance from vowels in utterances to the average vowels.

3.3.3 主成分分析

主成分分析により得られる話者空間の次元数の目安を得るために主成分分析における次元数と累積寄与率との関係を調べた．結果を図 3.6 に示す.

3.2.2 節で述べたように話者の特徴は 125 次元のスーパーベクトルで表現されているが，初めの 25 の主成分により 90%以上の累積寄与率が得られており，パラメータ間の相関が大きいことがわかる.

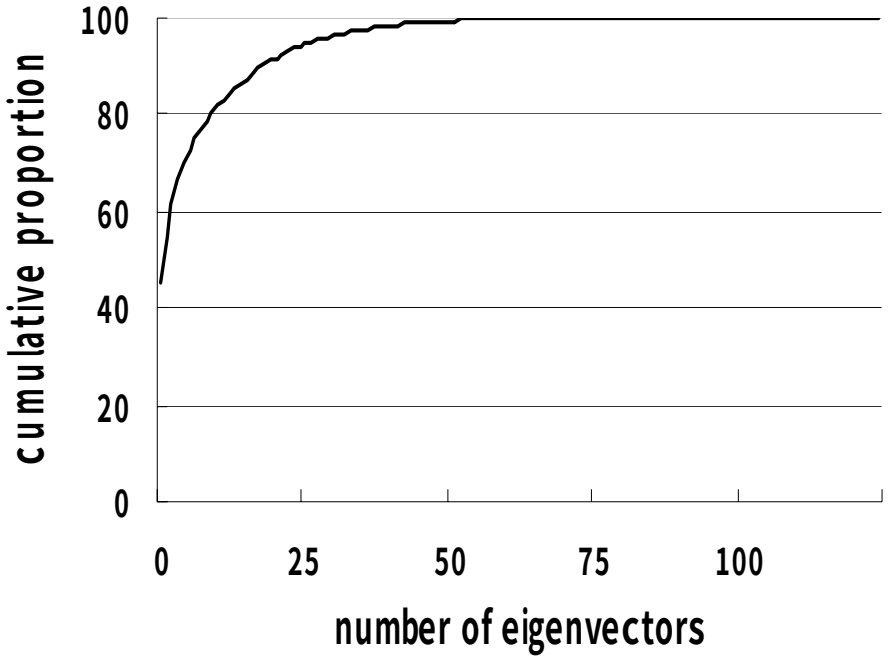


図 3.6: 主成分分析における累積寄与率

Figure 3.6: Cumulative proportions for the number of eigenvectors.

3.3.4 提案手法の妥当性の予備検証

話者空間は各話者のスーパーベクトルが存在できる制限された特徴ベクトルの部分空間である．話者空間がうまく構成されれば，実際に観測された話者のスーパーベクトルは，話者空間との距離が小さくなるはずである．用いる主成分数を変えたときに，各話者のスーパーベクトルと話者空間との距離がどのように変化するかを調べた．結果を図 3.7 に示す．（a）は各話者のスーパーベクトルから話者空間への平均距離である．（b），（c）は，それぞれ /o/，/i/ の音素を全て /a/ の特徴パラメータに置き換えた時の話者空間への平均距離で，入力音声中の /o/，/i/ が全て /a/ に誤認識された状況をシミュレーションしている．

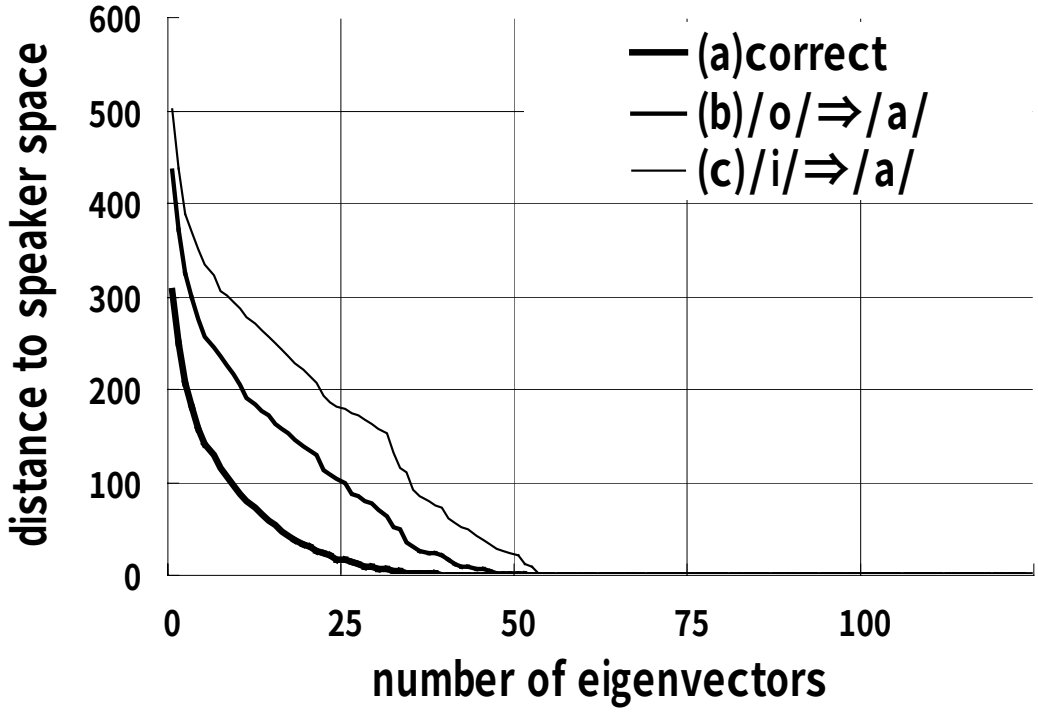


図 3.7: 各話者と話者空間の距離

Figure 3.7: The distance to the speaker space.

主成分数が 20～40 では（a）は距離が小さいが，それに比べて（b），(c）では距離が大きくなっている．この結果は，提案手法により誤認識の可能性を表現できることを示しており，通常の音響スコアに話者空間への距離を導入することによる音声認識システムの改善が期待できる．

3.3.5 連続音声認識実験

Julius の N-best 候補の設定を 100 個とし，各 N-best 候補のそれぞれの特徴ベクトル空間での位置を求め，特徴ベクトル空間でのその各候補の位置から話者空間への距離を求めた．音響スコアの再評価は式（3.1）によって，求められた話者空間への距離を 100-best の結果の各文に付け加える．

適切な主成分数を用いて話者空間を構成し，また適切な重みを設定することにより，有効な音響スコアの再評価が行われると考えられる．適切な主成分数と重みを設定するために主成分数と重みを変えた場合の音声認識率の変化を調べた．主成分数を固定し，重みを変化させた場合と，重みを固定し，主成分数を変化させた場合の連続音声認識実験を行なった．

ここで母音の正規化を行なわない場合を方法 1，母音の特徴量の正規化を行なう場合を方法 2 とする．図 3.8（a）と（b）は，方法 1 および方法 2 の場合に主成分数を固定して重みを変化させたときの音声認識率の変化をそれぞれ示している．

連続音声認識実験での性能評価の評価尺度には，(3.4) 式で表す単語正解率（Word Correctness）と（3.5）式で表す単語正解精度（Word Accuracy）を用いる．ここで W は単語数，S は置換誤り，D は脱落誤り，I は挿入誤りの単語数を表す．

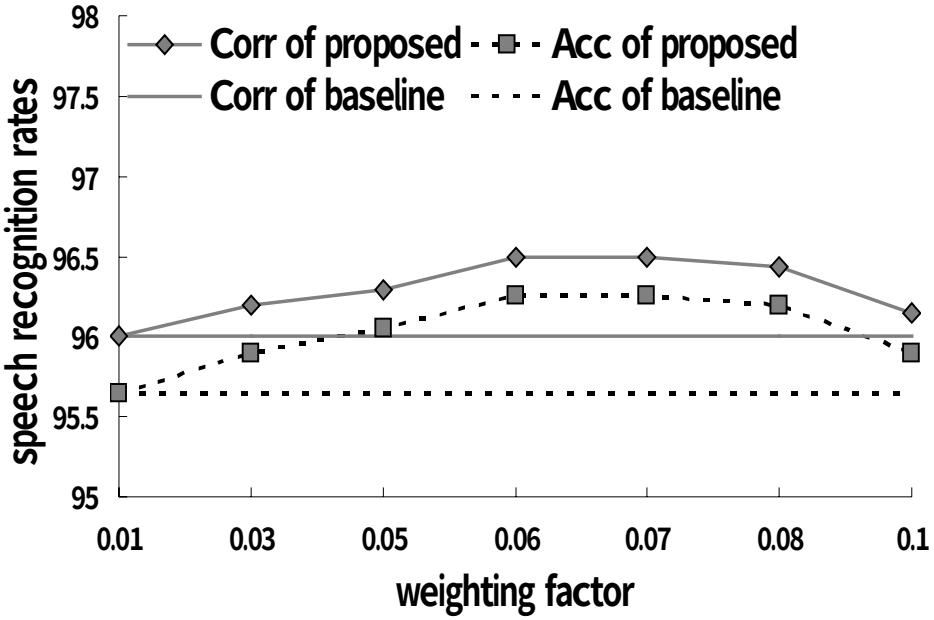
$$\text{単語正解率}（Corr.）= \frac{W - S - D}{W} \tag{3.4}$$

$$\text{単語正解精度}（Acc.）= \frac{W - S - D - I}{W} \tag{3.5}$$

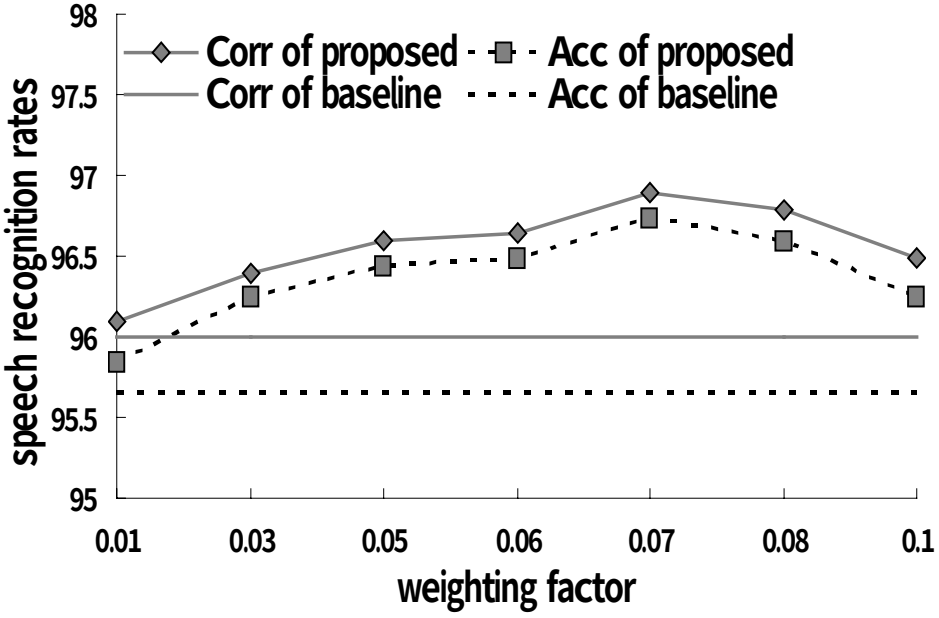
単語正解率と単語正解精度では，無音区間などにわき出した単語による挿入誤りを算出に含めるか否かが異なる．純粋に音声認識の評価をする場合，単語正解精度を用いることが多いが，音声インタフェースで良く使われるキーワードマッチの枠組みでは，挿入誤りが存在してもキーワードさえ正しく認識できれば良いと考えることもできる．よって，単語正解率と単語正解精度の両方を求める．

図 3.8（a）は主成分数を 32 に固定，図 3.8（b）は主成分数を 30 に固定し，重みを 0.01，0.03，0.05，0.06，0.07，0.08，0.1 と変化させた．そして直線と点線の水平な線はそれぞれ Julius の結果を示している．提案手法は重みの変化に対して広い範囲で音声認識率の改善がみられる．重みは方法 1 および方法 2 両方とも 0.07 の場合が最良の認識率を示している．

図 3.9（a）と（b）は，方法 1 および方法 2 の場合に重みを固定して主成分数を変化させたときの音声認識率の変化をそれぞれ示している．図 3.9（a）と（b）は重みを両方とも 0.07 に固定し，主成分数を 15，20，25，30，32，33，35 と変化させた．主成分数は方法 1 が 32，方法 2 が 30 のとき最もよい認識率を示している．Julius の結果と比較した場合，主成分数，重みを適切に設定すれば認識率が向上することがわかった．重みは 0.07 の場合最もよい認識率を示している．主成分数は母音の正規化前が 32，母音の正規化後が 30 の場合最もよい認識率を示した．以上の評価実験結果を表 3.1 にまとめる．方法 1 では主成分数を 32，重み（w）を 0.07，方法 2 では主成分数を 30，重みを 0.07 に設定した．認識精度評価における認識誤り率でみると，手法 1，2 ではそれぞれ 14%，25%の認識誤りが削減されている．



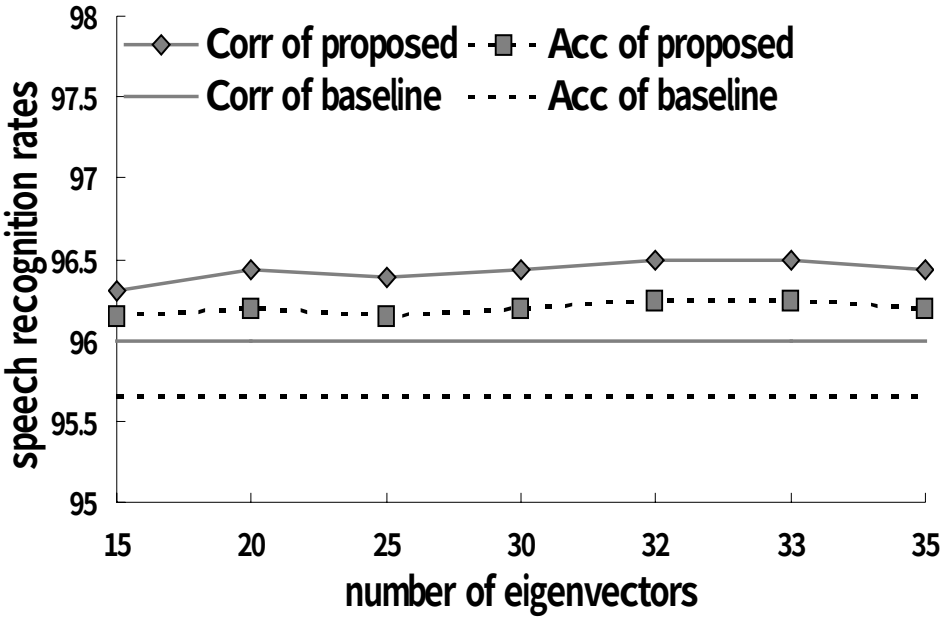
(a) method 1



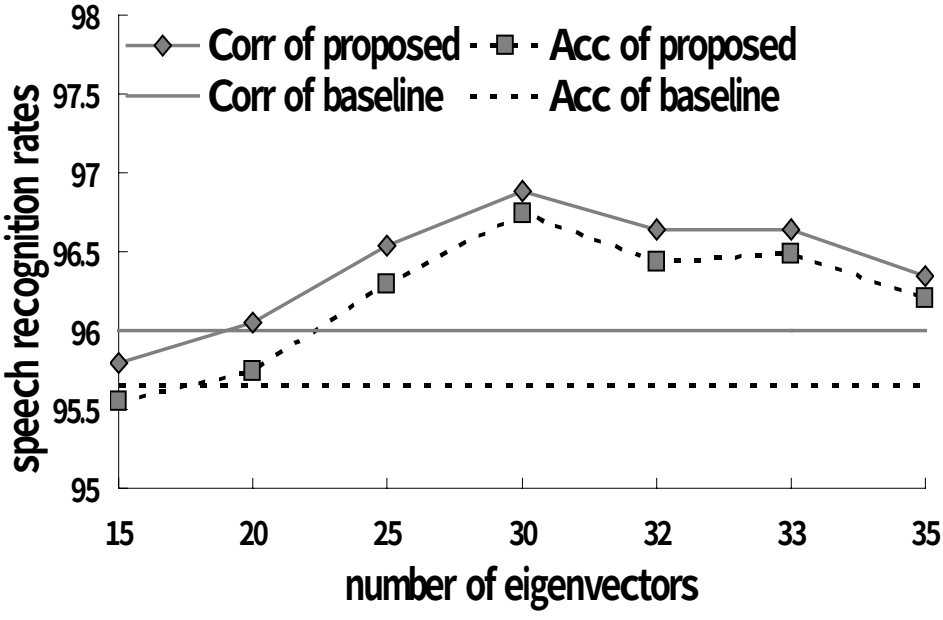
(b) method 2

図 3.8: 重みを変化させた場合の認識率の変化

Figure 3.8: The speech recognition rate for changing the weighting factor.



(a) method 1



(b) method 2

図 3.9: 主成分数を変化させた場合の認識率の変化

Figure 3.9: The speech recognition rate for changing PCA components.

表 3.1: 連続音声認識実験結果（単語認識率）

Table 3.1: Continuous speech recognition rates.

method	Correct (%)	Accuracy (%)
baseline	96.00	95.65
method 1	96.49	96.25
method 2	96.89	96.74

3.3.6 認識率改善の検定

Julius（ベースラインに相当）と提案手法の間に認識率の有意な差があるかどうか検定を行なった．認識率が 100%に近くかつ認識率の差が少ない場合は，有意差検定がほとんど不可能になってしまう．このような場合，符号検定ができればテストデータ数を大きく削減することができる．本研究では，両者のテストサンプルが同一であるため，一つ一つのテストデータに対する認識結果を比較することにより符号検定法を行なう[22].

ここで，Julius と提案手法の認識方法の結果が異なった数を調べると，提案手法の方法 1 では Julius が正解，方法 1 が誤認識の場合は 3 個，Julius が誤認識，方法 1 が正解の場合は 16 個で，符号検定表を参考にすると危険率 1%で提案手法の方法 1 の方が有意に性能が高いと言える．方法 2 では Julius が正解，方法 2 が誤認識の場合は 1 個，Julius が誤認識，方法 2 が正解の場合は 23 個で，同じく危険率 1%で提案手法の方法 2 の方が有意に性能が高いと言える．

3.4 まとめ

不特定話者音声認識において，話者の適応データを用いることなく音素間の依存性を利用する手法について述べた．多数話者のデータに基づいて構成する話者空間を利用して認識文候補における母音の相対的関係の妥当性を評価している．さらに，認識

文候補における母音特徴を前後の音素コンテキストを考慮した正規化を行なうことによって，精度を改善する手法を示した．Julius の N-best 結果に対して，提案手法による再評価を行なうことにより約 25%の認識誤りが削減された．本論文で提案した手法では，入力中に 5 母音全てが出現することが前提となっている．入力音声にすべての母音が含まれない場合の対応として，5 母音のうち一部が出現しなかった場合は多数話者による母音の平均値を利用することが考えられる．また，4 母音からなる話者空間を 5 つ構成するなど，少ない母音数による話者空間を用いる手法も考えられる．今後の課題としては，入力音声に 5 母音全てが出現しない場合への対応の他，母音特徴の音素コンテキストに対する正規化手法の改善などが挙げられる．

参考文献

［1］ 中川聖一，確率モデルによる音声認識， コロナ社，1988.

［2］ 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.

［3］ C. -H. Lee, C. -H. Lin, and B. -H. Juang, “A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models,” IEEE Trans. Acoustic. Speech and Signal Process., Vol.ASSP-39, No.4, pp.806-814, April 1991.

［4］ J. -L. Gauvain and C. -H. Lee, “Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” IEEE Trans. Speech Audio Process., Vol.2, No.2, pp.291-298, April 1994.

［5］ 中川聖一，越川忠，“最大事後確率推定法を用いた連続出力分布型 HMM の適応化，” 日本音響学会誌，Vol.49, No.10, pp.721-728, Oct. 1993.

［6］ 中川聖一，鶴見豊，“最大事後確率推定法と認識結果を用いた連続出力分布型 HMM の教師なし話者適応化，” 電子情報通信学会論文誌，Vol.J78-DII，No.2, pp.188-196, Feb. 1995.

［7］ B. -H. Juang and L. -R. Rabiner, “Mixture Autoregressive Hidden Markov Models for Speech Signal,” IEEE Trans. Acoust. Speech and Signal Process., Vol.ASSP-33, No.6, pp.1404-1413, Dec. 1985.

［8］ 永井明人，鷹見淳一，嵯峨山茂樹，ハラルドシンガー，“隠れマルコフ網と一般化 LR 構文解析を統合した連続音声認識，”電子情報通信学会論文誌，Vol.J77-DII, No.1, pp.9-19, Jan. 1994.

［9］ 李晃伸，河原達也，武田一哉，鹿野清宏，“Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識，” 電子情報通信学会論文誌，Vol.J83-DII，No.12,

pp.2517-2525, Dec.2000.

［10］ 中川聖一，“確率モデルによる音声認識，” 電子情報通信学会，July 1988.

［11］ Mari Ostendorf, Ashvin Kannan, Orith Ronen, “Tree-based Dependence Models for Speech Recognition,” in Computational Models of Speech Pattern Processing, pp.40-53, Springer,1997.

［12］ 李宝潔，広瀬啓吉，峯松信明，“音素ペアモデルによる音素間情報の表現に関する検討，” 情報処理学会研究報告（音声言語情報処理研究会），2000-SLP-32-2, pp.7-12, July 2000.

［13］ 李宝潔，広瀬啓吉，“音素間相互情報を利用した音素ペアモデルによる話者適応，” 電子情報通信学会技術研究報告（音声研究会），SP99-104, pp.67-71, Dec. 1999.

［14］ 趙國，一丸太一郎，山下洋一，“話者空間モデルに基づいた音素間相関を用いた音声認識，” 電子情報通信学会論文誌，Vol.J87-D-II, No.7, pp.1402-1408, July 2004.

［15］ K. Cho and Y. Yamashita, “Speech Recognition Using Inter-Phoneme Dependency,” Proc. of the Eighth Western Pacific Acoustic Conference (WESPAC8), MB32, April 2003.

［16］ K. Cho and Y. Yamashita, “Speech Recognition Using Inter-phoneme Dependency Based on a Speaker Space Model,” Proc. of the 18th International Congress on Acoustics (ICA2004), 5, pp.3507-3510, April 2004.

［17］ R. Kuhn, P. Nguyen, J. -C. Janqua, L. Goldwasser, N. Niedzielski, S. Finke, K. Field, and M. Contolini, “Eigenvoices for speaker adaptation,” Proc. ICSLP-98, pp.1771-1774, 1998.

［18］ R. Kuhn, J. -C. Janqua, P. Nguyen and N. Niedzielski, “Rapid Speaker Adaptation in Eigenvoice Space Robust Speech Recognition,” IEEE Trans. Speech Audio Process., Vol.8, No.6, pp.695-707, 2000.

［19］ http://www.lang.astem.or.jp/CSRC/

［20］ http://julius.sourceforge.jp/

［21］ http://www.milab.is.tsukuba.ac.jp/jnas

［22］ 中川聖一，高木英行，“パターン認識における有意差検定と音声認識システム評価法，” 日本音響学会誌，Vol.50, No.10, pp.849-854, Oct. 1994.

第４章

N-best 音声認識における候補提示数の決定

4.1　はじめに

序論でも述べたように，近年の音声認識技術の向上により高い認識率が得られるようになったため，音声認識を用いた対話インタフェースへの期待が高まっている[1][2][3][4][5]．音声認識の認識結果を一つだけ提示する場合，現状の音声認識技術では，正しい結果が常に得られるとは限らない．その場合には，正しい結果が得られるまで何度も同じ発話を繰り返すことになり，利用者にとって大きな負担となる．このような負担を軽減するために，認識結果を複数候補提示し，利用者に候補の中から正解を選択してもらう音声インタフェースへの期待が高まっている[6][7][8][9][10].

N-bestの音声認識結果に基づいて複数の正解候補を利用者に提示する場合には，通常，提示する候補数をあらかじめ決めておく．提示する候補数を多くした方が，少ない場合に比べて候補中に正解が含まれている確率が高くなる．しかし，多くの候補の中から正解を探す必要があるため，手間がかかってしまう．提示される候補数が少なく，かつ，候補中に正解が含まれていることが理想的である．このように，候補中で正解が含まれる割合を減少させずに提示する候補数を減らすには，あらかじめ候補数を決めておくのではなく，認識結果に基づいて候補数を動的に決定する必要がある．

そこで，本論文では，N-best 候補の認識スコアを分析し，認識スコアを利用した候

補提示数の決定手法について検討する．また，認識スコアと併せて単語信頼度を用いて認識候補提示数を決定する手法も検討する．さらに，提案手法の有効性を検証する．

対象とする音声対話のタスクとしては，例えば，予約や情報検索のように，ユーザの発声内容からキーワードを取り出してスロットフィリングを行なうようなタスクを考える．個々のタスクに依存した知識を利用することによって，ユーザへの結果提示を効率的に行なえることが示されているが[11]，本論文では，タスクに依存しない認識スコアだけを利用する手法について検討する．

4.2　認識スコアの分析

実際に音声認識を行なう場合，N-best 候補が少なく，しかも正解文が正しく含まれている候補が提示されることが理想的である．認識スコアは対数尤度を用いて音響モデルの尤度で表される．そこで，N-best の認識結果においてスコアの分布と正解順位の関係にどのような特徴があるか分析し，スコアに応じて候補提示数を決定する規則を見つける．

N-best 候補の認識スコアの例を表 4.1 に示す．「この研究室の歴史を知りたい」と発話し，音声認識システムを用いて音声認識を行なった結果である．

認識スコアは，文の長さで大きく値が異なるため，正規化する必要がある．本研究では，文の長さを文全体のフレーム数で表し，文の認識スコアを文全体のフレーム数で割り，正規化した値を用いて分析を行なっていく．表 4.1 には，認識スコアを文の長さで正規化した値が示されている．また，本研究では，各候補を上位から順に第 1 候補，第 2 候補，・・・と呼ぶことにし，各候補の認識スコアを第 1 候補のスコア，第 2 候補のスコア，・・・と呼ぶことにする．

表 4.1: 認識結果と認識スコア

Table 4.1: Recognition results and recognition scores.

n-best	recognition result	recognition score
1	この研究室 の 歴史 が 知りたい	-26.160279
2	この研究室 の 歴史 を 知りたい	-26.161865
3	この研究室 の 研究 を 知りたい	-26.332994
4	この研究室 の 研究 が 知りたい	-26.398474
5	この研究室 の 歴史 が 聞きたい	-26.442401
6	この研究室 の 歴史 を 聞きたい	-26.443183
7	この研究室 の 歴史 が 知りたい です	-26.460549
8	この研究室 の 歴史 を 知りたい です	-26.462135
9	この研究室 の 先生 を 知りたい	-26.471869

The correct answer: この研究室 の 歴史 を 知りたい

4.2.1 タスク

タスクは研究室のホームページ検索とした．タスク文法と単語辞書の語彙数を変えて 3 つのタスクで実験を行なう．まず，タスク 1 では単語辞書の語彙数は 94 単語で，タスク文法は検索時に発話すると思われる 28 パターンの文が用意されている．タスク 2 では単語辞書の語彙数が 184 単語で，文法として 41 の文パターン，タスク 3 では単語辞書の語彙数が 1360 単語で，文法として 45 の文パターンが用意されている．

2.4.3 節でも述べたように，テスト文が少ない場合においては，テスト文によって認識の難易度に差が生じるのでテスト文集合に対するパープレキシティで評価する必要がある．これをテストセットパープレキシティと呼ぶ．タスク 1，2，3 のテストセットパープレキシティはそれぞれ 26，140，1290 である．また，発話文は様々な長さの文を用意した．

4.2.2 分析データ

認識スコアの分析にあたって，20 人(男 11 人，女 9 人)に 20 文ずつ発話してもらい，この合計 400 文の音声を，N-best 候補数を 30 として音声認識を行なった．N-best 候補を生成する連続音声認識システムとしては，CSRC2001 年度版のシステムを用いた[12]．デコーダは Julian rev.3.3[13]のスタンダード版，音響モデル(HMM)は「新聞記事読み上げ音声コーパス」で学習した 64 混合分布，3000 状態の PTM モデルを用いた．Julian のオプション設定として，第 1 パスのビーム幅を 800，第 2 パスの最大スタック数を 700，第 2 パスの仮説オーバフローのしきい値を 1000，第 2 パスで見つける文の数と見つかった文のうち出力する数を共に 30 とし，これ以外はデフォルト値を用いた．

提示された N-best 候補の中に正解が含まれている確率を正解提示率と呼ぶことにする．400 文に対する正解提示率はタスク 1 が 100%，タスク 2 が 95.75%，タスク 3 が 42%であった．また，N-best 候補数は 30 に設定したが，実際には常に 30 の文候補が生成されたわけではなく，出力された平均提示候補数はタスク 1 が 9.79 個，タスク 2 が 27.8 個，タスク 3 が 24 個であった．

第 1 候補から第 30 候補まで，その候補が正解であった文がいくつあったかを表 4.2 に示す．

認識スコアの分析に用いる 400 文の発話内容はタスク 1，2，3 すべて同じで，それぞれのタスク文法での，400 文の発話に対する受理率は 100%である．タスクが異なっても同じ評価文を用いているが，語彙数の増加などによりタスク 1，2，3 と順に文法が複雑になるにつれて N-best 内に正解が含まれない“No correct answer”の数が表 4.2 に示すように増加している．平均的な単語数はタスク 1 が 4.9 個，タスク 2 が 6. 2 個，タスク 3 が 7.3 個である．同じ発話内容で平均的な単語数が異なるのは単語の定義がタスクによって異なるためである．単語境界を/で表して発話文の例を挙げると，タスク 1 では「いままで/の/修士論文テーマ/を/教えて/ほしい」，タスク 2 では「いままで/の/修士/論文/テーマ/を/教えて/ほしい」などとなる．発話内容を受理できる文の長さ（単語数）の範囲はタスク 1 で 8 個，タスク 2 で 22 個，タスク 3 で 22 個である．

表 4.2: 各候補における正解の文数

Table 4.2: The number of sentences of the correct answer for each candidate.

Rank of the correct answer	The number of sentences		
	Task 1	Task 2	Task 3
1	378	302	67
2	17	33	18
3	5	22	12
4		8	10
5		4	8
6		2	12
7		1	6
8		2	4
9			5
10		2	4
11			3
12		1	1
13		3	1
14			1
15			
16		1	4
17			1
18			2
19			1
20			1
21			
22		1	
23			2
24			
25			
26		1	1
27			
28			1
29			
30			3
No correct answer	0	17	232
Total	400	400	400

4.2.3 ヒューリスティックス 1 の分析

4.2.3.1 隣接候補間のスコア差

隣り合う候補（第 n 候補と第(n+1)候補）のスコアの差が大きい場合，第(n+1)候補以降に正解が含まれる可能性は低いことが予想でき，この場合第(n+1)候補以後を提示しないことを考える．これをヒューリスティックス 1 と呼ぶ[14][15].

4.2.3.2 分析結果

1 文ごとに第 n 候補と第(n+1)候補のスコアの差を求め，正解の順位とスコアの差にはどのような関係があるのか調べる．第 1 候補と第 2 候補のスコアの差を調べたタスク 2 の結果を図 4.1 に，タスク 3 の結果を図 4.2 に示す．横軸は，正解の順位，縦軸は第 1 候補と第 2 候補のスコアの差を示している．つまり，グラフ中の一つの点は，一つの文を認識した結果，何候補目が正解で，その時に第 1 候補と第 2 候補のスコアの差はどのくらい開いていたかを示しており，点の数は全部で 400 ある．尚，横軸の「0」は，30 候補中に正解がなかったことを意味しており，正解の順位それぞれに対する点の数は表 4.2 に対応している．

このグラフから分かるように，タスク 2 では，第 1 候補と第 2 候補のスコアの差が 0.05 以上開くような文は，第 1 候補に正解が多いことが分かった．また，タスク 3 では，第 1 候補と第 2 候補のスコアの差が 0.05 以上，タスク 1 では，第 1 候補と第 2 候補のスコアの差が 0.03 以上開くような文は，第 1 候補に正解が多いことが分かった．

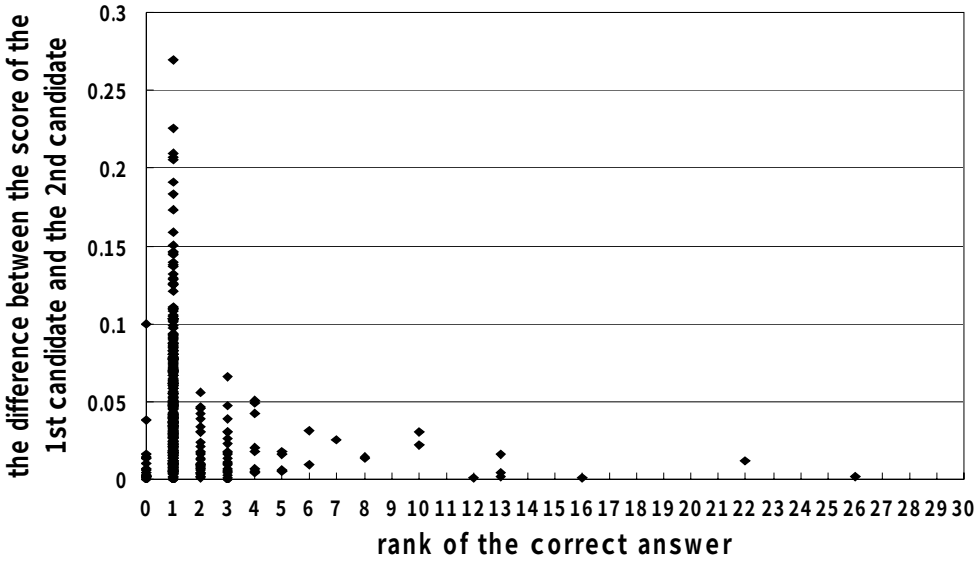


図 4.1: 第 1 候補と第 2 候補のスコアの差（タスク 2）

Figure 4.1: The difference of the score of the 1st candidate and the 2nd candidate.(Task2)

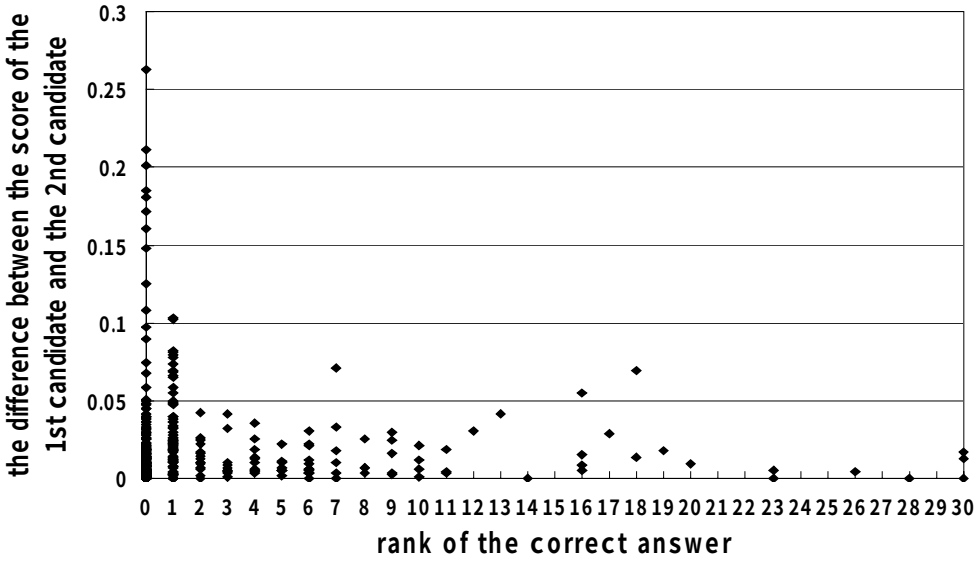


図 4.2: 第 1 候補と第 2 候補のスコアの差（タスク 3）

Figure 4.2: The difference of the score of the 1st candidate and the 2nd candidate.(Task3)

次に，第 2 候補と第 3 候補のスコアの差を調べたタスク 2 の結果を図 4.3 に，タスク 3 の結果を図 4.4 に示す．これにより，タスク 2 では，第 2 候補と第 3 候補のスコアの差が 0.03 以上開くような文は，第 2 候補以前に正解が多いことが分かった．また，タスク 3 では，第 2 候補と第 3 候補のスコアの差が 0.03 以上，タスク 1 では，第 2 候補と第 3 候補のスコアの差が 0.01 以上開くような文は，第 2 候補以前に正解が多いことが分かった．

次に，第 3 候補と第 4 候補のスコアの差を調べたタスク 2 の結果を図 4.5 に，タスク 3 の結果を図 4.6 に示す．同じく，タスク 2 では，第 3 候補と第 4 候補のスコアの差が 0.02 以上，タスク 3 では，第 3 候補と第 4 候補のスコアの差が 0.03 以上開くような文は，第 3 候補以前に正解が多いことが分かった．また，タスク 1 では，このような関係は見られなかった．

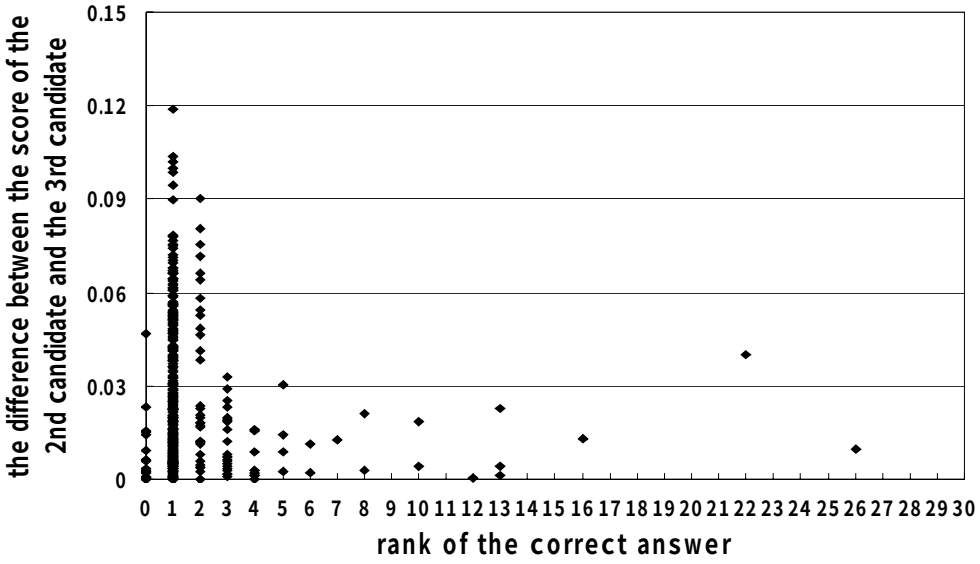


図 4.3: 第 2 候補と第 3 候補のスコアの差（タスク 2）

Figure 4.3: The difference of the score of the 2nd candidate and the 3rd candidate.(Task2)

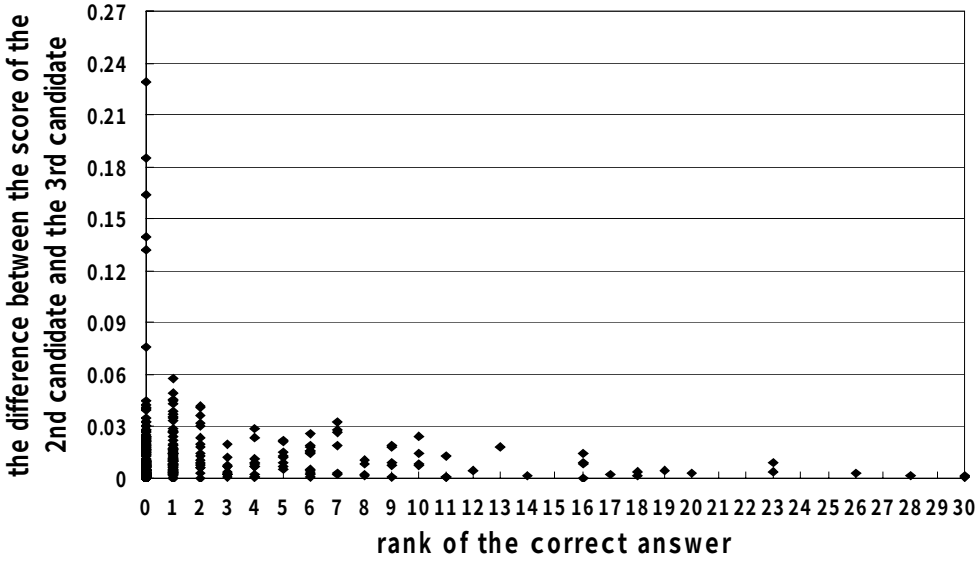


図 4.4: 第 2 候補と第 3 候補のスコアの差（タスク 3）

Figure 4.4: The difference of the score of the 2nd candidate and the 3rd candidate.(Task3)

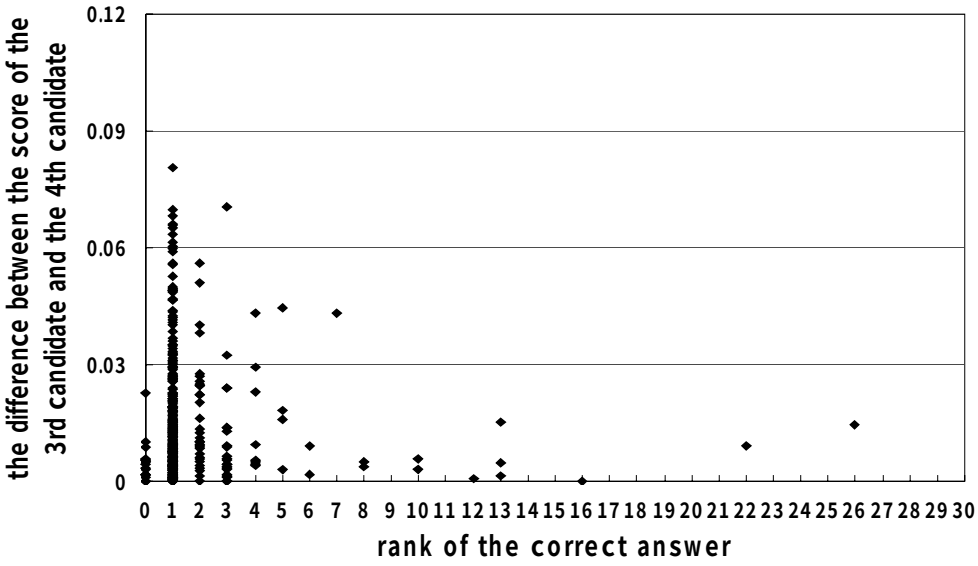


図 4.5: 第 3 候補と第 4 候補のスコアの差（タスク 2）

Figure 4.5: The difference of the score of the 3rd candidate and the 4th candidate.(Task2)

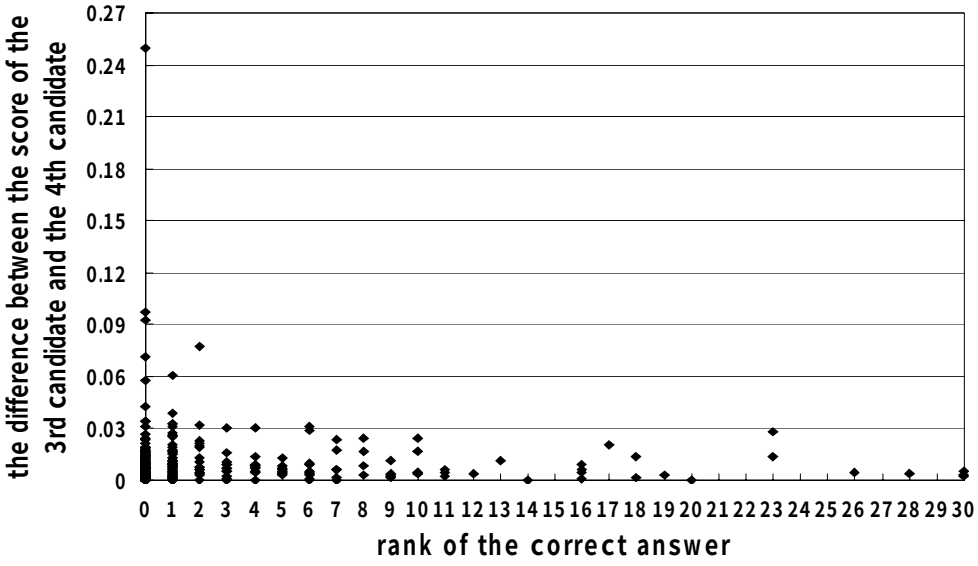


図 4.6: 第 3 候補と第 4 候補のスコアの差（タスク 3）

Figure 4.6: The difference of the score of the 3rd candidate and the 4th candidate.(Task3)

次に，第 4 候補と第 5 候補のスコアの差を調べたが，タスク 1 とタスク 3 では，このような関係はなく，タスク 2 だけ，第 4 候補と第 5 候補のスコアの差が 0.02 以上開くような文は，第 4 候補以前に正解が多いことが分かった．タスク 2 の結果を図 4.7 に示す.

尚，第 5 候補と第 6 候補，第 6 候補と第 7 候補，・・・，第 29 候補と第 30 候補のスコアの差も調べてみたが，このような関係はどのタスクでも見られなかった.

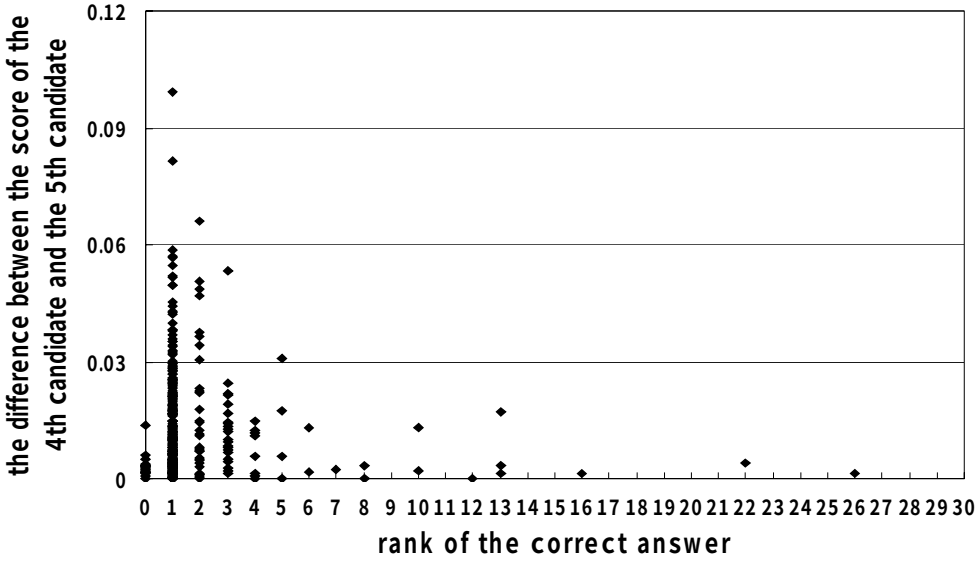


図 4.7: 第 4 候補と第 5 候補のスコアの差（タスク 2）

Figure 4.7: The difference of the score of the 4th candidate and the 5th candidate.(Task2)

4.2.4 ヒューリスティックス 2 の分析

4.2.4.1 第 1 候補とのスコア差

第 1 候補と第n候補のスコアの差が大きい場合，第n候補以降に正解が含まれている可能性は低いことが予想でき，この場合第n候補以降は提示しないことを考える．これをヒューリスティックス 2 と呼び，第 1 候補と第n候補とのスコア差の閾値を θ_1 と

する[14][15]．大語彙連続音声認識では数万語の単語の任意の組み合わせを認識対象とするため登場しうる文仮説数は膨大であり，最尤単語系列を素早く発見する高速な解探索アルゴリズムが重要となる．そのために，フレーム同期型ビームサーチ法[16]，枝刈り法[17][18]などといった認識処理の効率化手法においても同様のヒューリスティックスが用いられている．これは連続音声認識における検索空間の削減だけでなく，N-bestの提示候補数の決定にも有効であることが期待できる．θ_1を大きくした方が，提示される候補数が増えるため，候補中に正解が含まれている確率（正解提示率）も高くなっていく．しかし，θ_1を大きくしていっても，ある値を境に正解提示率は変化せずに一定で，平均提示候補数だけが増えていくのではないかと予想できる．

4.2.4.2 分析結果

θ_1を変化させた時の平均提示候補数と正解提示率を調べた．タスク 1 の分析結果を図 4.8 に，タスク 2 の分析結果を図 4.9 に示す．これにより，θ_1を大きくすると正解提示率が高くなっていくことがわかる．しかし，タスク 1 では，θ_1を 0.10 以上，タスク 2 では，θ_1を 0.12 以上にすると平均提示候補数は増加していくのに対し，正解提示率はあまり増加せず，ほぼ一定になっている．また，タスク 3 でも同じ傾向が見られ，θ_1を 0.14 以上にすると平均提示候補数は増加していくのに対し，正解提示率はあまり増加しなかった．

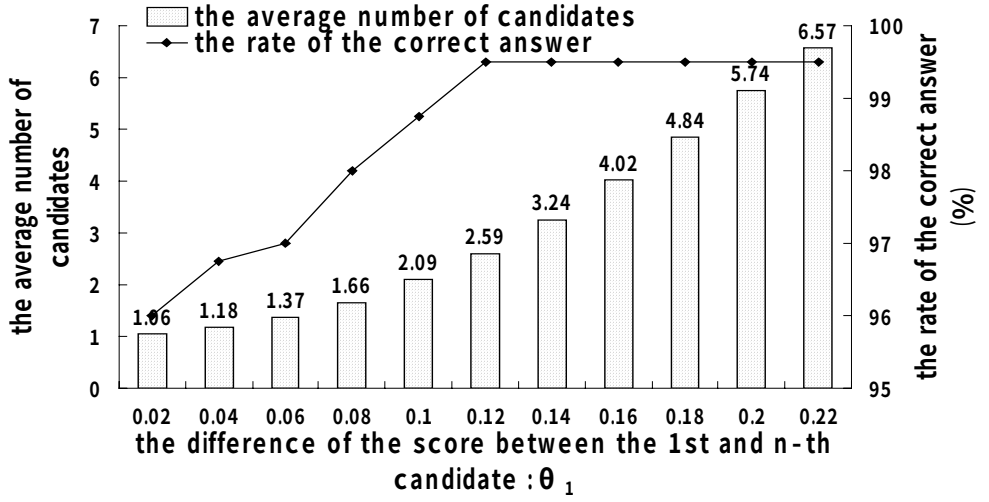


図 4.8: ヒューリスティックス 2 における正解提示率と平均提示候補数の変化（タスク 1）

Figure 4.8: Change of the rate of the correct answer and the average number of candidates for heuristics 2.(Task1)

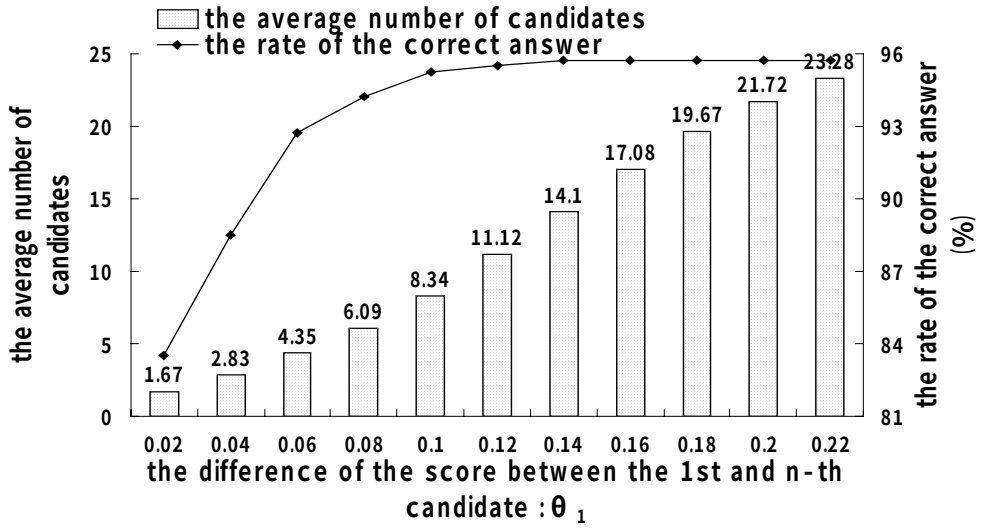


図 4.9: ヒューリスティックス 2 における正解提示率と平均提示候補数の変化（タスク 2）

Figure 4.9: Change of the rate of the correct answer and the average number of candidates for heuristics 2.(Task2)

4.2.5 ヒューリスティックス 3 の分析

4.2.5.1 第 n 候補のスコア

第n候補のスコアが小さい場合，第n候補以降に正解が含まれている可能性は低いことが予想され，この場合第n候補以降は提示しないことを考える．これをヒューリスティックス 3 と呼び，第n候補のスコアに対する閾値を θ_2とする[14][15].

4.2.5.2 分析結果

θ_2を変えた時の平均提示候補数と正解提示率の変化を調べた．タスク 2 の分析結果を図 4.10 に，タスク 3 の分析結果を図 4.11 に示す．これにより，それぞれ θ_2が-27 までは，θ_2を小さくすれば正解提示率が高くなっていく．しかし，タスク 2 と 3 では，θ_2を−27 以下以下にしても正解提示率はあまり増加せず，ほぼ一定になっている．また，タスク 1 でも θ_2を-26.5 以下にしても正解提示率はあまり増加しなかった.

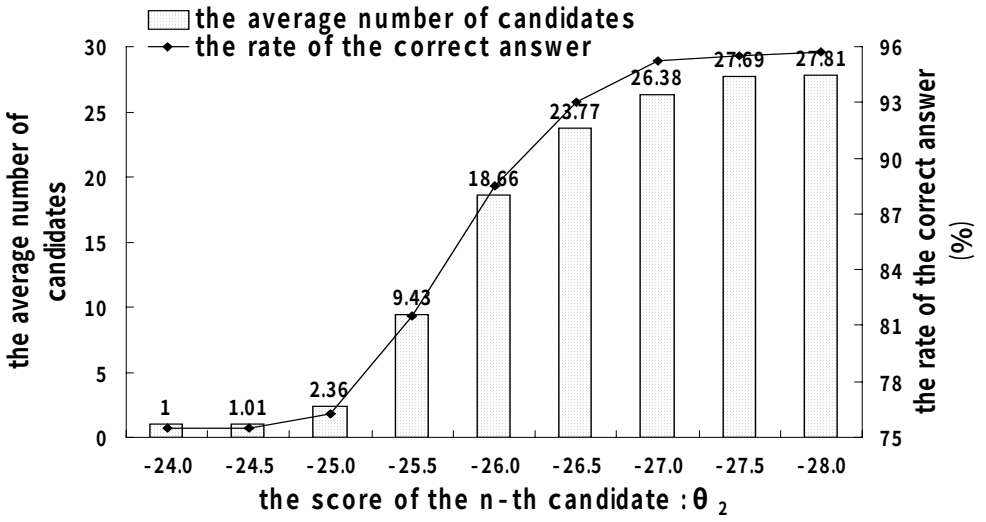


図 4.10: ヒューリスティックス 3 における正解提示率と平均提示候補数の変化（タスク 2)

Figure 4.10: Change of the rate of the correct answer and the average number of candidates for heuristics 3.(Task2)

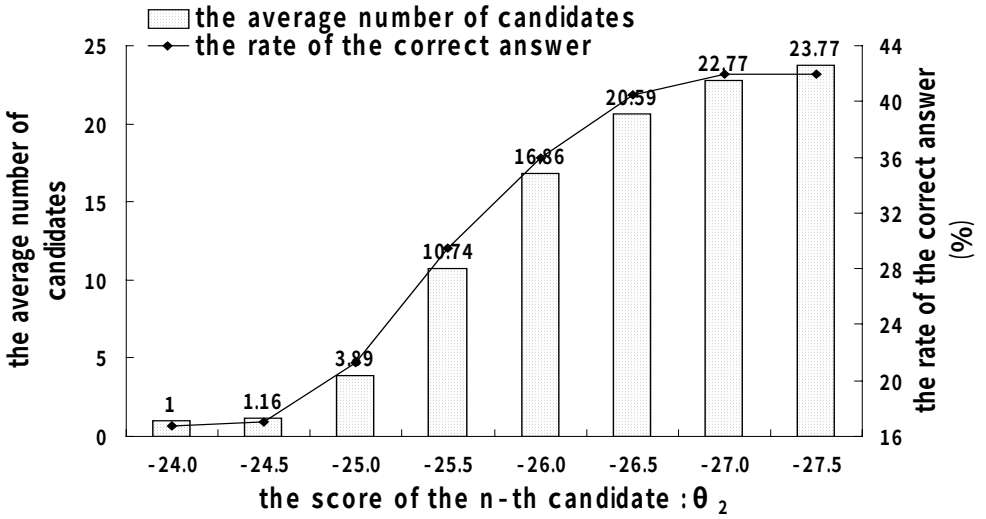


図 4.11: ヒューリスティックス 3 における正解提示率と平均提示候補数の変化（タスク 3)

Figure 4.11: Change of the rate of the correct answer and the average number of candidates for heuristics 3.(Task3)

4.3　認識候補提示数の決定

4.3.1 規則セット

4.2.3 節と 4.2.4 節，4.2.5 節の 3 つの分析結果より，候補提示数の決定に以下の規則を用いることにする．

タスク 1 は,

[規則 1]　第 1 候補と第 2 候補のスコアの差が 0.03 以上開く時，第 1 候補のみ提示する．

[規則 2]　第 2 候補と第 3 候補のスコアの差が 0.01 以上開く時，第 2 候補まで提示する．

[規則 3]　第 1 候補と第 n 候補のスコアの差が 0.10 以上開く時，第(n-1)候補まで提示する．

[規則 4]　第 n 候補のスコアが-26.5 以下の時，第(n-1)候補まで提示する（n=2，3，4，…，30）．

タスク 2 は,

[規則 1]　第 1 候補と第 2 候補のスコアの差が 0.05 以上開く時，第 1 候補のみ提示する．

[規則 2]　第 2 候補と第 3 候補のスコアの差が 0.03 以上開く時，第 2 候補まで提示する．

[規則 3]　第 3 候補と第 4 候補のスコアの差が 0.02 以上開く時，第 3 候補まで提示する．

[規則 4]　第 4 候補と第 5 候補のスコアの差が 0.02 以上開く時，第 4 候補まで提示する．

[規則 5]　第 1 候補と第 n 候補のスコアの差が 0.12 以上開く時，第(n-1)候補まで提示する．

[規則 6] 第 n 候補のスコアが-27 以下の時，第(n-1)候補まで提示する（n=2，3，4，…，30）．

タスク 3 は，

[規則 1] 第 1 候補と第 2 候補のスコアの差が 0.05 以上開く時，第 1 候補のみ提示する．

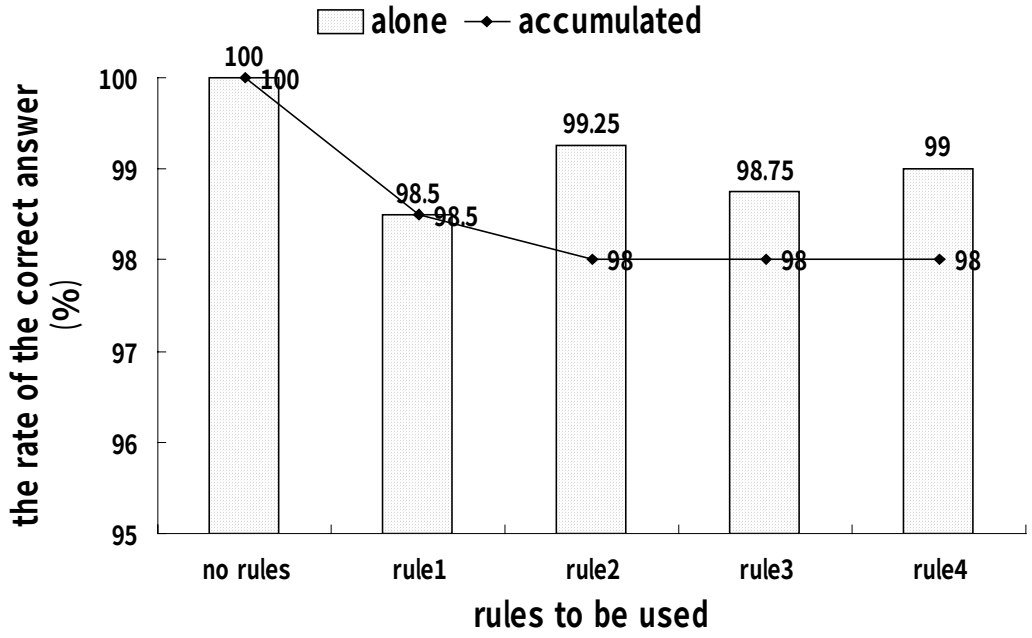
[規則 2] 第 2 候補と第 3 候補のスコアの差が 0.03 以上開く時，第 2 候補まで提示する．

[規則 3] 第 3 候補と第 4 候補のスコアの差が 0.03 以上開く時，第 3 候補まで提示する．

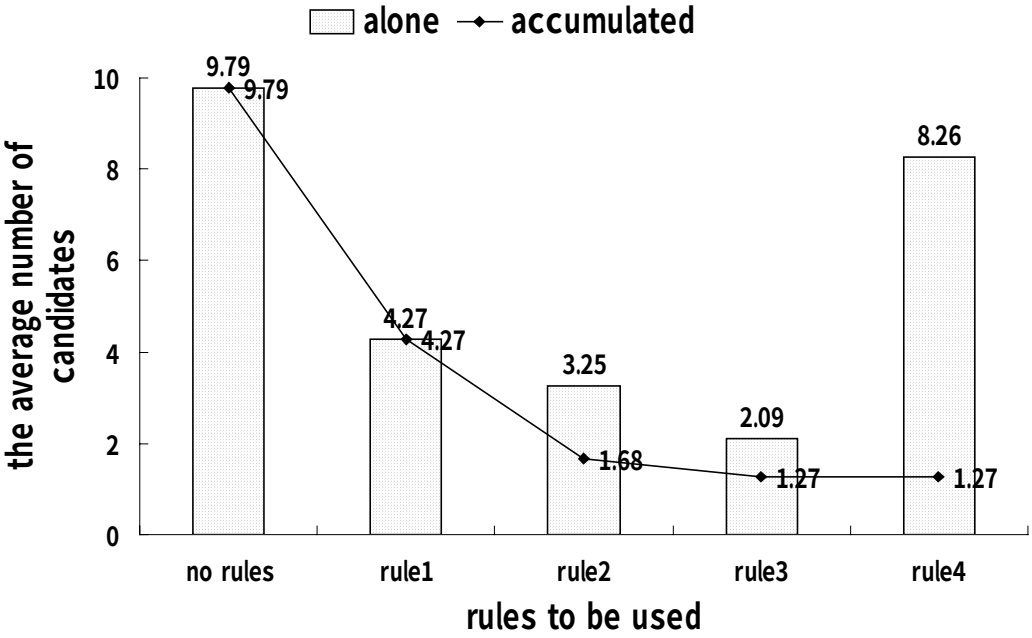
[規則 4] 第 1 候補と第 n 候補のスコアの差が 0.14 以上開く時，第(n-1)候補まで提示する．

[規則 5] 第 n 候補のスコアが-27 以下の時，第(n-1)候補まで提示する（n=2，3，4，…，30）．

4.2.2 節で示した 400 文をタスク 1，2，3 の文法で認識した時の 30-best の認識結果に，これらの規則セットを適応した結果をそれぞれ図 4.12，4.13，4.14 に示す．ここで「alone」は，各規則セット中の一つの規則を単独で用いる場合の効果を参考までに示している．「accumulated」は，規則セットの適用結果で，規則を規則 1 から順に照合していき，最初に条件が一致した規則を適用した結果である．規則セットを適用すると，30 候補提示した場合は正解があったにもかかわらず，提示数を減らしたために正解が提示されない場合がある．それぞれの規則を単独，あるいはセットで使用する場合，正解提示率はどのように変化するか調べた結果が図 4.12，4.13，4.14 の(a)である．規則をセットで使用した結果，正解提示率はタスク 1 が 98%，タスク 2 が 93.5%，タスク 3 が 40.5%となった．これは 30 候補提示した場合の正解提示率と比べてやや下がってはいるがあまり大差はないと思われる．



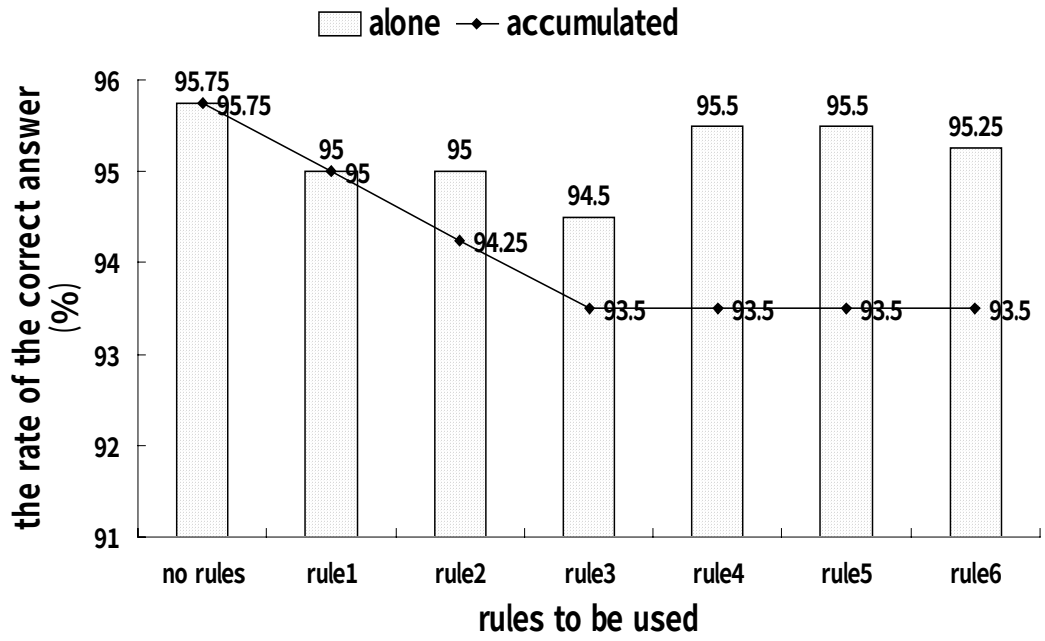
(a) Change of the rate of the correct answer.



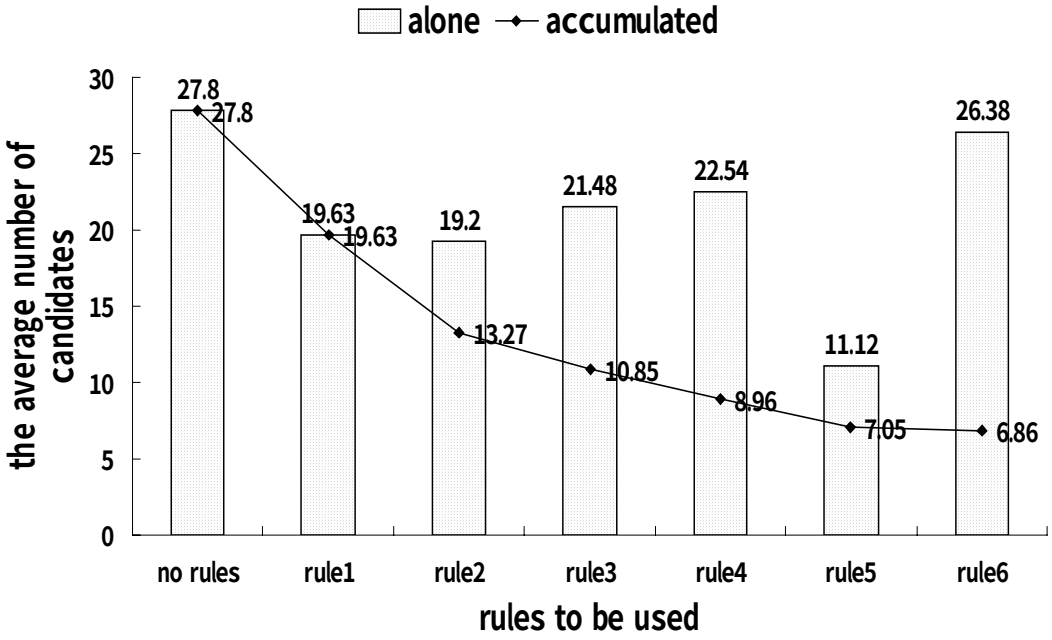
(b) Change of the average number of candidates.

図 4.12: 正解提示率と平均候補数（タスク 1）

Figure 4.12: The rate of the correct answer and the average number of candidates.(Task1)



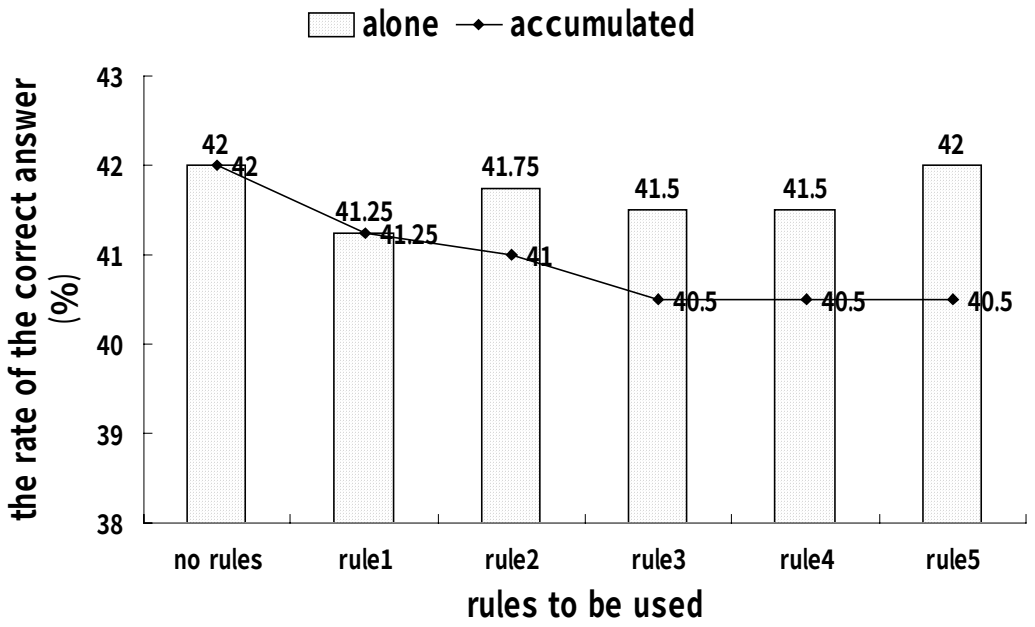
(a) Change of the rate of the correct answer.



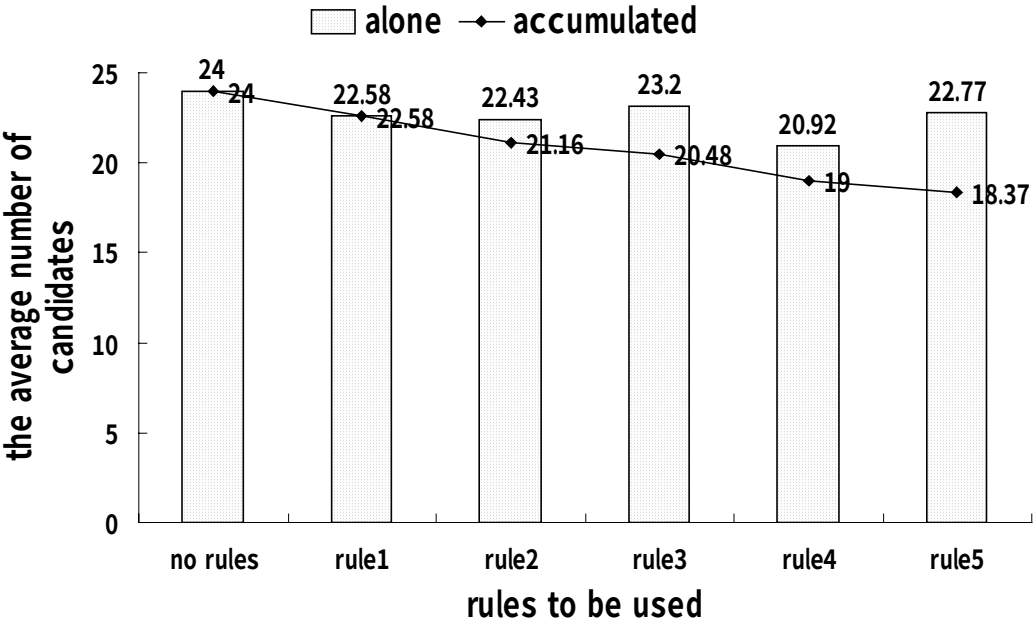
(b) Change of the average number of candidates.

図 4.13: 正解提示率と平均候補数（タスク 2）

Figure 4.13: The rate of the correct answer and the average number of candidates.(Task2)



(a) Change of the rate of the correct answer.



(b) Change of the average number of candidates.

図 4.14: 正解提示率と平均候補数（タスク 3）

Figure 4.14: The rate of the correct answer and the average number of candidates.(Task3)

一文あたりの平均提示候補数がそれぞれの規則を単独，あるいはセットで使用する場合，どのように変化するか調べた結果が図 4.12，4.13，4.14 の(b)である．それぞれの規則をセットで使った場合，平均提示候補数は規則を適用する度に減少し，最終的にはタスク 1 が 1.27，タスク 2 が 6.86，タスク 3 が 18.37 となった．これは，常に 30 候補提示する場合より大幅に減少している．

個々の規則を単独に用いた場合では，図 4.12，4.13，4.14 の alone の結果から，ヒューリスティクス 2 に関わる規則（タスク 1 での規則 3，タスク 2 での規則 5，タスク 3 での規則 4）の効果が大きいことがわかる．これらの規則を単独で利用するよりも，全ての規則を組み合わせて利用する（accumulated）ことによって，タスク 1 では 39%， タスク 2 では 37%の提示候補数の削減率が得られているものの，タスク 3 では 9%の削減率に留まっており，規則組合せの有効性については今後さらに検証する必要がある．また，ヒューリスティクス 2 に関わる規則だけでは，正解提示率の減少が小さいことなども考慮しながら，規則の組合せ方についても検討する余地があると思われる．

4.3.2 規則の一般化

認識スコアの分析結果により認識スコアを用いると正解提示率をあまり下げることなく，候補提示数を大幅に減らすことができた．しかし，これらの規則セットは３つのタスク，それぞれに適した規則であり，どの場合でも有効とは言えない．そこで，３つのタスクのそれぞれの規則の閾値を平均し，その平均値を基準に閾値を最適化し，タスクに依存しない規則セットとして以下を考えた．

[規則 1] 第 1 候補と第 2 候補のスコアの差が 0.06 以上開く時，第 1 候補のみ提示する．

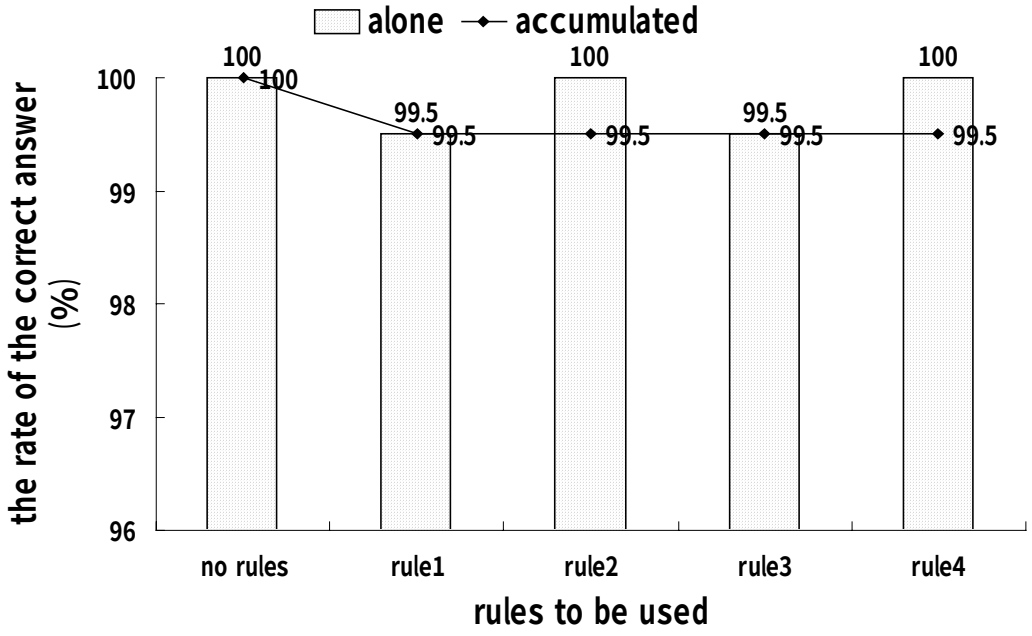
[規則 2] 第 2 候補と第 3 候補のスコアの差が 0.03 以上開く時，第 2 候補まで提示する．

[規則 3] 第 1 候補と第 n 候補のスコアの差が 0.12 以上開く時，第(n-1)候補まで提示する．

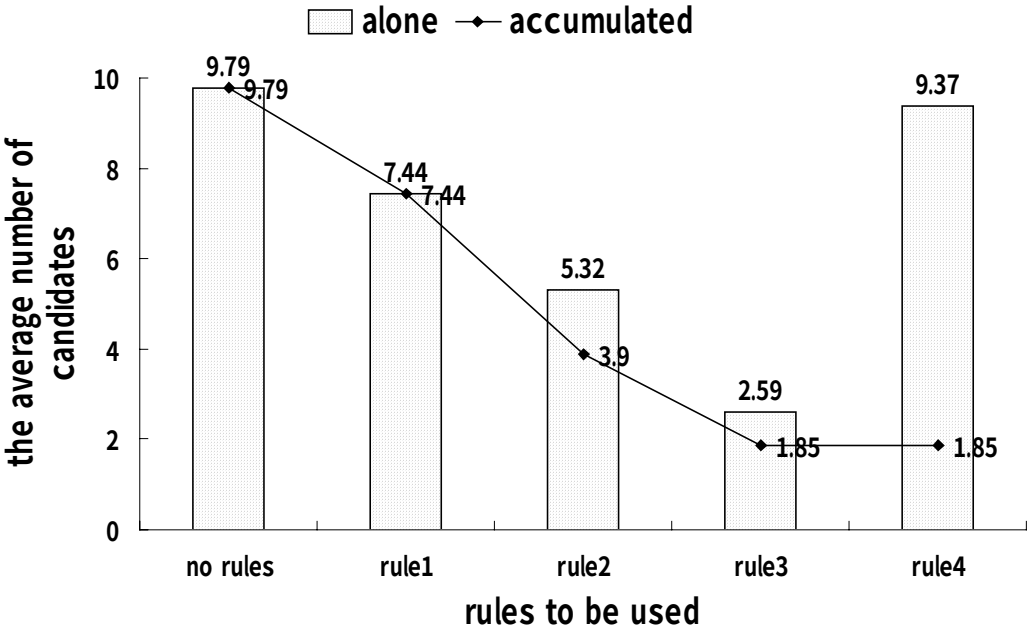
[規則 4] 第 n 候補のスコアが-27 以下の時，第(n-1)候補まで提示する（n=2，3，4，

…，30)．

それぞれの規則を単独，またはセットで使用した場合，タスク 1，2，3 の規則セットを適応した結果をそれぞれ図 4.15，4.16，4.17 に示す．この規則セットのそれぞれの規則をセットで用いると，正解提示率はそれぞれ 99.5%，94.75%，40%，平均提示候補数はそれぞれ 1.85，8.78，18 となった．この規則セットを用いても正解提示率をあまり下げることなく，提示する候補数を減らせることが分かった．



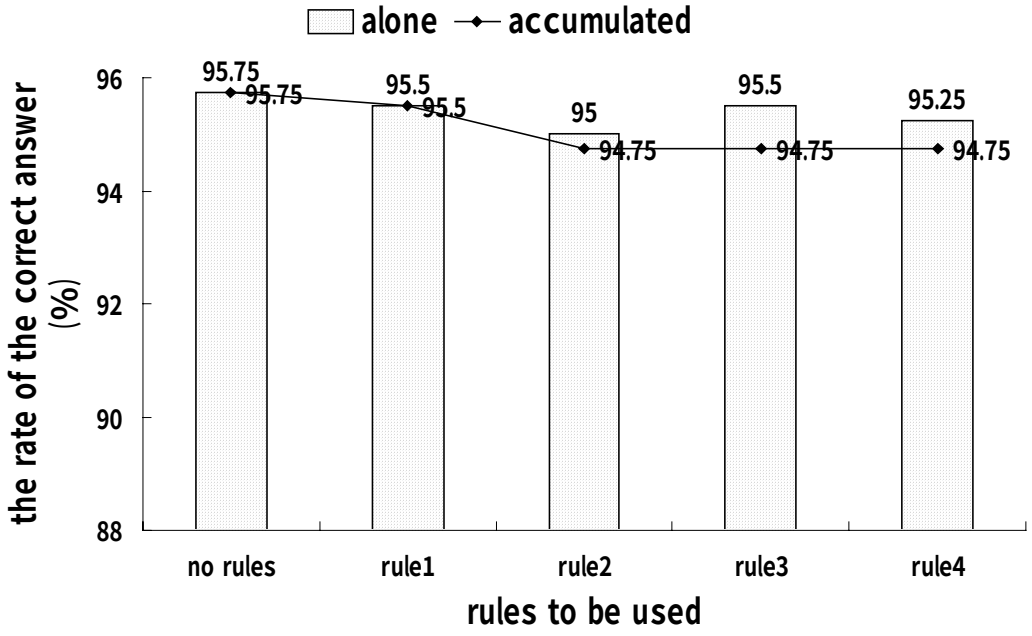
(a) Change of the rate of the correct answer.



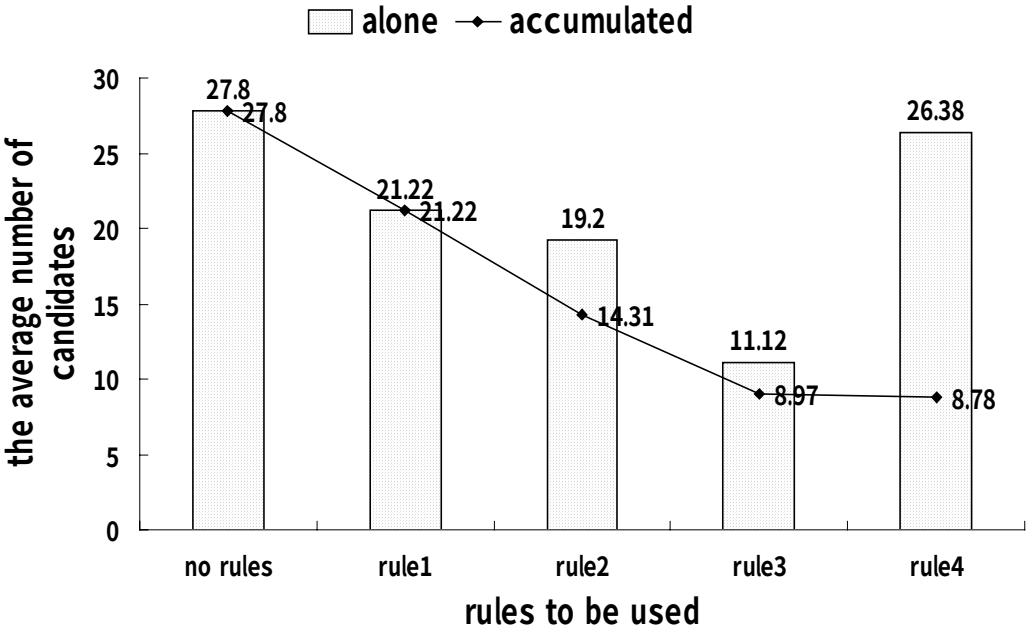
(b) Change of the average number of candidates.

図 4.15: 共通の規則セットによる正解提示率と平均候補数（タスク 1）

Figure 4.15: The rate of the correct answer and the average number of candidates with the common rule set.(Task1)



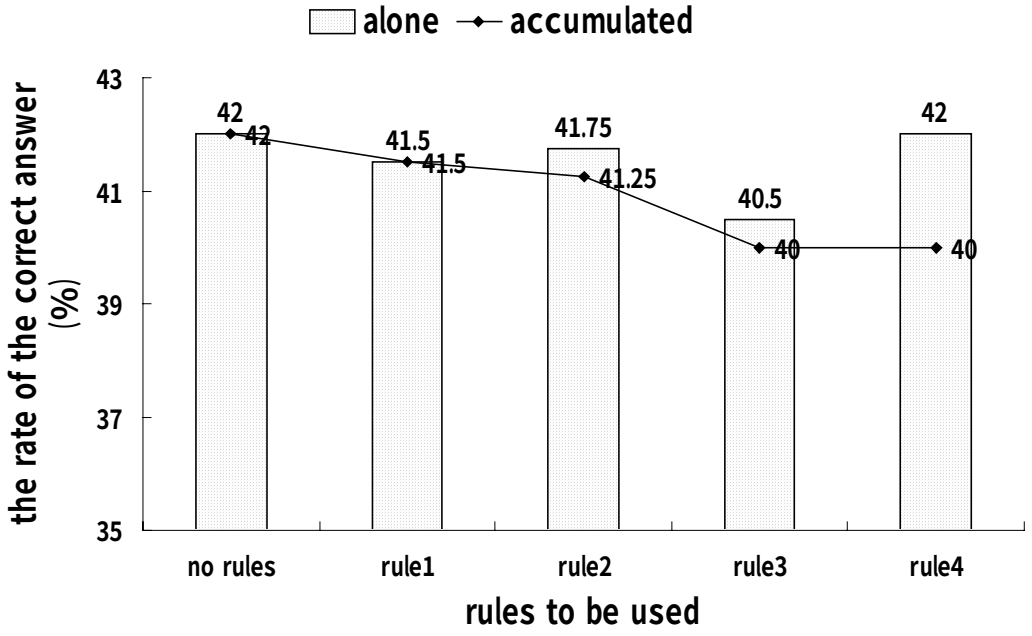
(a) Change of the rate of the correct answer.



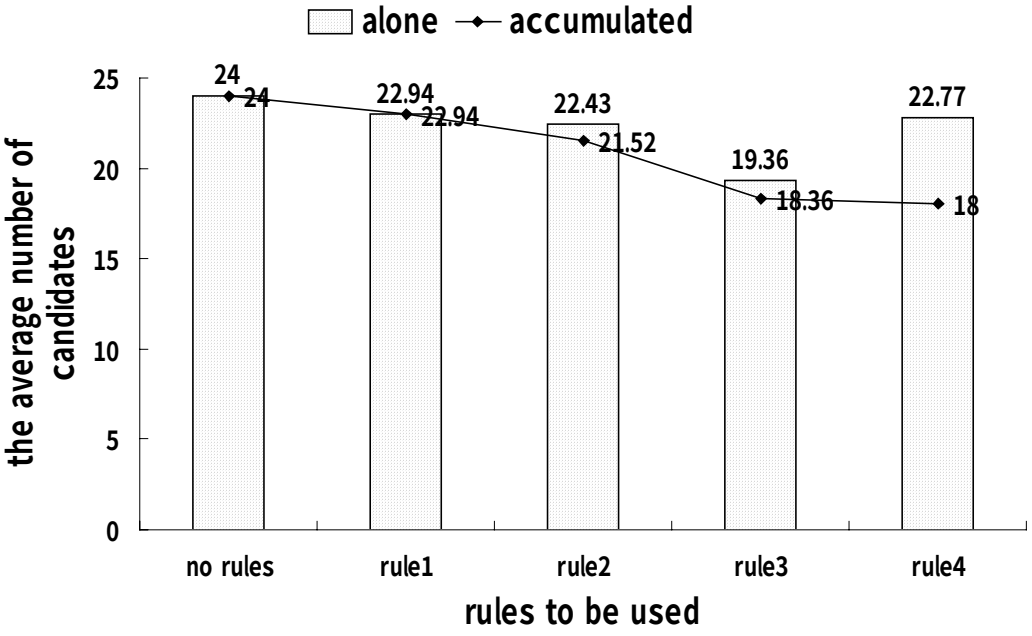
(b) Change of the average number of candidates.

図 4.16: 共通の規則セットによる正解提示率と平均候補数（タスク 2）

Figure 4.16: The rate of the correct answer and the average number of candidates with the common rule set.(Task2)



(a) Change of the rate of the correct answer.



(b) Change of the average number of candidates.

図 4.17: 共通の規則セットによる正解提示率と平均候補数（タスク 3）

Figure 4.17: The rate of the correct answer and the average number of candidates with the common rule set.(Task3)

4.3.3 一般化された規則の評価

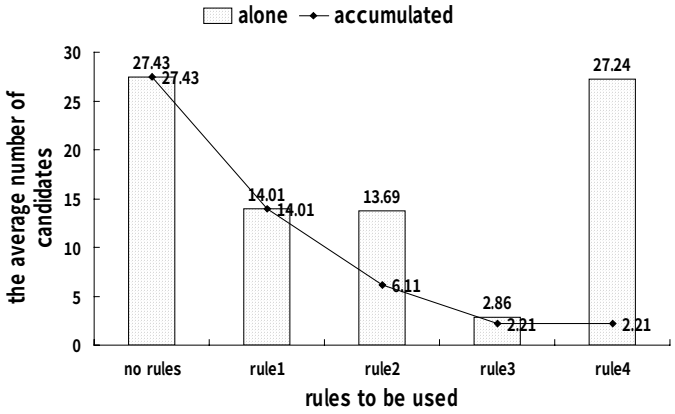
前節でタスクに依存しない規則セットとして規則を一般化した規則セットを考えたが，この規則セットがどのタスクの場合にも適した規則であり，有効であるか全く違うタスクを用いて確かめる必要がある．それを確かめるために連続単語の認識結果に，これらの規則セットを適用した．新たなタスクは人名検索とした．こちらも単語辞書に日本の上位姓（苗字）と人気上位名前（男性）を用いて語彙数を変えて３つのタスクで実験を行なう．まず，タスク 4 では単語辞書の語彙数は 710 単語，タスク 5 では 840 単語，タスク 6 では 1000 単語が用意されている．タスク 4，5，6 のテストセットパープレキシティはそれぞれ 257，313，400 である．そして，5 人(男 3 人，女 2 人)に 20 文ずつ発話してもらい，この合計 100 文をタスク 4，5，6 の文法で認識した時の 30-best の認識結果に，一般化された規則セットのそれぞれの規則をセットで使用した場合，タスク 4，5，6 の正解提示率と平均候補提示数の変化を表 4.3 に示す．また，それぞれの規則を単独，またはセットで使用した場合，タスク 4，5，6 の平均候補提示数の変化を図 4.18 に示す.

規則セットを用いる前の平均提示候補数はタスク 4 が 27.43 個，タスク 5 が 25.87 個，タスク 6 が 25.65 個であったが，規則セットを用いるとタスク 4 が 2.21 個，タスク 5 が 2.39 個，タスク 6 が 2.49 個であった．また，規則セットを用いる前の正解提示率はタスク 4 が 95%，タスク 5 が 91%，タスク 6 が 84%であったが，規則セットを用いるとタスク 4 が 93%，タスク 5 が 89%，タスク 6 が 82%であった．この規則セットを用いるとどのタスクでも正解提示率をあまり下げることなく，提示する候補数を減らせることが分かった.

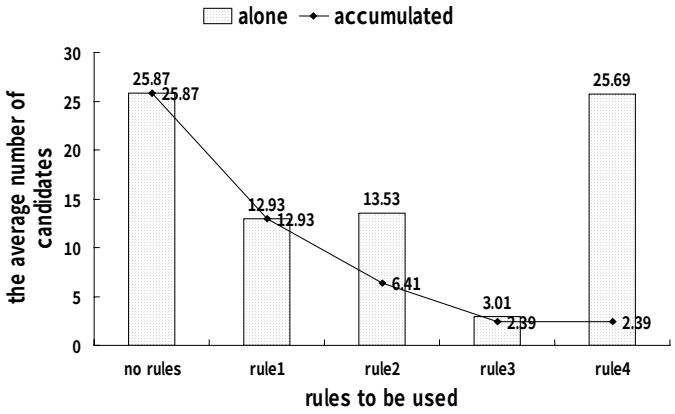
表 4.3：正解提示率と平均提示候補数

Table 4.3: The rate of the correct answer and the average number of candidates.

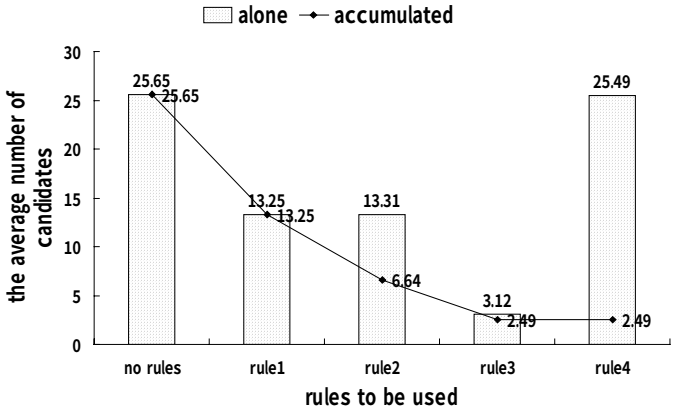
	The average number of candidates		The rate of the correct answer(%)	
	Before	After	Before	After
Task 4	27.43	2.21	95	93
Task 5	25.87	2.39	91	89
Task 6	25.65	2.49	84	82



(a) Task 4.



(b) Task 5.



(c) Task 6.

図 4.18: 共通の規則セットによる平均候補数

Figure 4.18: The average number of candidates with the common rule set.

4.4　単語信頼度を考慮した認識候補提示数の決定

4.4.1 実験条件

タスクは 4.2.1 節で述べた研究室のホームページ検索用タスク 1，2，3 を使用し，分析データは 4.2.2 節で述べた 20 人（男 11 人，女 9 人）に 20 文ずつ発話してもらった合計 400 文を使用する．

前節までの認識スコアを利用した候補提示数の決定では，N-best の評価として認識スコアをそのまま用いたが，本節では事後確率による正規化を行ないそのスコアを N-best の認識スコアとして用いる．また，認識スコアと併せて単語信頼度も用いて認識候補提示数を決定する．

連続音声認識システムとしては，CSRC2001 年度版のシステムを，デコーダは Julian rev.3.4 のスタンダード版，音響モデル(HMM)と Julian のオプション設定は 4.2.2 節の実験条件をそのまま用いた．また，N-best 候補数を 30 として音声認識を行なった．400 文に対する 30-best の正解提示率はタスク 1 が 100%，タスク 2 が 96.75%，タスク 3 が 70.50%であった．出力された平均提示候補数はタスク 1 が 6.87 個，タスク 2 が 23.53 個，タスク 3 が 26.25 個であった．

4.4.2 規則セット

4.4.2.1 認識スコアの規則セット

認識スコアの分析結果より，候補提示数の決定に以下の規則セットを用いることにする．

タスク 1 は，

[規則 1]　第 1 候補と第 2 候補のスコアの差が 0.06 以上開く時，第 1 候補のみ提示する．

[規則 2]　第 2 候補と第 3 候補のスコアの差が 0.01 以上開く時，第 2 候補まで提示する．

[規則 3]　第 1 候補と第 n 候補のスコアの差が 0.10 以上開く時，第(n-1)候補まで提示する．

[規則 4] 第 n 候補のスコアが 32 以下の時，第(n-1)候補まで提示する（n=2，3，4，…，30）．

タスク 2 は，

[規則 1] 第 1 候補と第 2 候補のスコアの差が 0.09 以上開く時，第 1 候補のみ提示する．

[規則 2] 第 2 候補と第 3 候補のスコアの差が 0.05 以上開く時，第 2 候補まで提示する．

[規則 3] 第 3 候補と第 4 候補のスコアの差が 0.04 以上開く時，第 3 候補まで提示する．

[規則 4] 第 4 候補と第 5 候補のスコアの差が 0.01 以上開く時，第 4 候補まで提示する．

[規則 5] 第 5 候補と第 6 候補のスコアの差が 0.01 以上開く時，第 4 候補まで提示する．

[規則 6] 第 6 候補と第 7 候補のスコアの差が 0.01 以上開く時，第 4 候補まで提示する．

[規則 7] 第 1 候補と第 n 候補のスコアの差が 0.11 以上開く時，第(n-1)候補まで提示する．

[規則 8] 第 n 候補のスコアが 32 以下の時，第(n-1)候補まで提示する（n=2，3，4，…，30）．

タスク 3 は，

[規則 1] 第 1 候補と第 2 候補のスコアの差が 0.05 以上開く時，第 1 候補のみ提示する．

[規則 2] 第 2 候補と第 3 候補のスコアの差が 0.05 以上開く時，第 2 候補まで提示する．

[規則 3]　第 3 候補と第 4 候補のスコアの差が 0.03 以上開く時，第 3 候補まで提示する．

[規則 4]　第 4 候補と第 5 候補のスコアの差が 0.02 以上開く時，第 4 候補まで提示する．

[規則 5]　第 5 候補と第 6 候補のスコアの差が 0.02 以上開く時，第 4 候補まで提示する．

[規則 6]　第 6 候補と第 7 候補のスコアの差が 0.03 以上開く時，第 4 候補まで提示する．

[規則 7]　第 1 候補と第 n 候補のスコアの差が 0.10 以上開く時，第(n-1)候補まで提示する．

[規則 8]　第 n 候補のスコアが 32 以下の時，第(n-1)候補まで提示する（n=2，3，4，…，30）．

4.4.2.2 単語信頼度の規則セット

単語信頼度の分析結果，候補提示数の決定に以下の規則セットを用いることにする．

タスク 1 は,

[規則 1]　信頼度が 0.05 以下の単語が含まれている候補は提示しない．

[規則 2]　各単語の信頼度の平均が 0.65 以下の場合，その候補は提示しない．

タスク 2 は,

[規則 1]　信頼度が 0.05 以下の単語が含まれている候補は提示しない．

[規則 2]　各単語の信頼度の平均が 0.55 以下の場合，その候補は提示しない．

タスク 3 は,

[規則 1]　信頼度が 0.05 以下の単語が含まれている候補は提示しない．

[規則 2]　各単語の信頼度の平均が 0.55 以下の場合，その候補は提示しない.

4.4.2.3 規則の一般化

それぞれの規則セットは３つのタスク，それぞれに適した規則であり，どの場合でも有効とは言えない．そこで，タスクに依存しない規則セットを考えた.
認識スコアは,

[規則 1]　第 1 候補と第 2 候補のスコアの差が 0.09 以上開く時，第 1 候補のみ提示する.

[規則 2]　第 2 候補と第 3 候補のスコアの差が 0.05 以上開く時，第 2 候補まで提示する.

[規則 3]　第 3 候補と第 4 候補のスコアの差が 0.04 以上開く時，第 3 候補まで提示する.

[規則 4]　第 4 候補と第 5 候補のスコアの差が 0.02 以上開く時，第 4 候補まで提示する.

[規則 5]　第 1 候補と第 n 候補のスコアの差が 0.10 以上開く時，第(n-1)候補まで提示する.

[規則 6]　第 n 候補のスコアが 32 以下の時，第(n-1)候補まで提示する（n=2，3，4，…，30）.

である.
単語信頼度は,

[規則 1]　信頼度が 0.05 以下の単語が含まれている候補は提示しない.

[規則 2]　各単語の信頼度の平均が 0.55 以下の場合，その候補は提示しない.

である.

4.4.2.4 認識スコアと単語信頼度を併せた規則セット

認識スコアと併せて単語信頼度を用いて３つのタスクそれぞれに適した規則セットと規則を一般化し，タスクに依存しない規則セットを考えた.

一般化した規則セットは,

単語信頼度が,

[規則 1]　信頼度が 0.05 以下の単語が含まれている候補は提示しない.

[規則 2]　各単語の信頼度の平均が 0.35 以下の場合，その候補は提示しない.

認識スコアが,

[規則 1]　第 1 候補と第 2 候補のスコアの差が 0.09 以上開く時，第 1 候補のみ提示する.

[規則 2]　第 2 候補と第 3 候補のスコアの差が 0.05 以上開く時，第 2 候補まで提示する.

[規則 3]　第 3 候補と第 4 候補のスコアの差が 0.04 以上開く時，第 3 候補まで提示する.

[規則 4]　第 4 候補と第 5 候補のスコアの差が 0.02 以上開く時，第 4 候補まで提示する.

[規則 5]　第 1 候補と第 n 候補のスコアの差が 0.10 以上開く時，第(n-1)候補まで提示する.

[規則 6]　第 n 候補のスコアが 32 以下の時，第(n-1)候補まで提示する（n=2，3，4，…，30）.

である.

4.4.3 実験結果

これらの規則セットの中，各タスクそれぞれの規則セットを適応した結果を表 4.4（平均候補提示数の変化）と表 4.5（正解提示率の変化）に，規則を一般化した規則セ

ットを適応した結果を表 4.6（平均候補提示数の変化）と表 4.7（正解提示率の変化）に示す．常に 30 候補提示する場合が baseline，認識スコアのみの場合が method1，単語信頼度のみの場合が method2，認識スコアと信頼度を併せて用いた場合が method3 である．

表 4.4：平均提示候補数

Table 4.4: The average number of candidates.

	The average number of candidates			
	baseline	method1	method2	method3
Task 1	6.87	1.24	1.72	1.12
Task 2	23.53	3.84	3.99	2.68
Task 3	26.25	11.83	11.90	9.36

表 4.5：正解提示率

Table 4.5: The rate of the correct answer.

	The rate of the correct answer (%)			
	baseline	method1	method2	method3
Task 1	100	99.75	99.75	99.75
Task 2	96.75	96.50	96.50	96.25
Task 3	70.50	69.50	68.75	69.50

表 4.6: 平均提示候補数

Table 4.6: The average number of candidates.

	The average number of candidates			
	baseline	method1	method2	method3
Task 1	6.87	1.71	1.75	1.50
Task 2	23.53	3.90	3.99	2.72
Task 3	26.25	12.34	11.90	9.48

表 4.7: 正解提示率

Table 4.7: The rate of the correct answer.

	The rate of the correct answer (%)			
	baseline	method1	method2	method3
Task 1	100	99.75	99.75	99.75
Task 2	96.75	96.50	96.50	96
Task 3	70.50	70	68.75	69.50

4.5　提示手法の被験者による評価

4.5.1 実験方法

認識スコアを用いて候補提示数を決定する手法の有効性を検証するため，以下の 3 つの候補提示方法をそれぞれ実際に利用してもらい，その結果，どの方法が最も使いやすかったかを評価してもらった．被験者は 7 人で，タスクはタスク 2 を用いた．

（提示方法 1） 常に 1 候補のみ提示する．

（提示方法 2） 常に 30 候補提示する．

（提示方法 3） 認識スコアを用いて候補提示数を決定する．

尚，今回は，候補提示数の変化による使いやすさの違いを見るため，自由な発話ではなく，決められた文を発話してもらった．被験者の発話をマイクで入力して音声認識を行い，認識結果の N-best 候補を提示方法に従って図 4.19 のように画面に表示する．図中の /silB/，/silE/ は文の先頭と末尾の無音部を表している．表示された N-best 候補の中から被験者が正しい認識結果をマウスで指示する．もし表示された N-best 候補の中に正しい認識結果が含まれていない場合は何度でも同じ文を発話してもらう．被験者の発話開始から，正しい認識結果を指示するまでの所要時間と発話回数は被験者とは別の実験者が手動で計測した．

評価実験において，実際に提示された候補数の平均は「常に 30 候補提示する」の場合が 28 個，「認識スコアを用いて候補提示数を決定する」の場合が 4.8 個であった．

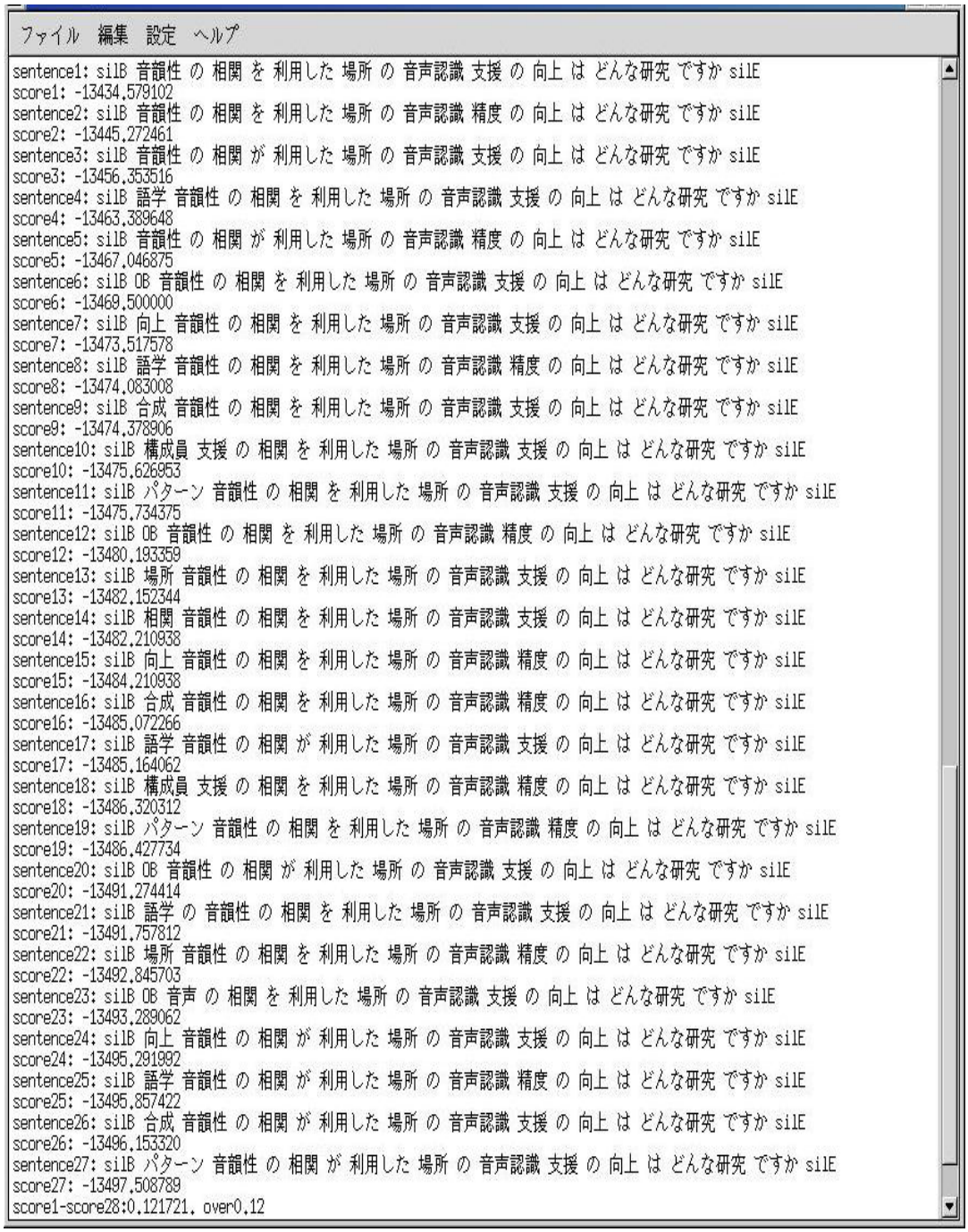


図 4.19: 認識結果の提示例

Figure 4.19: An example of displayed recognition results.

4.5.2 実験結果

3つの候補提示方法を被験者がそれぞれ実際に利用してもらった時の正解選択までの平均所要時間と平均発話回数を図4.20に示す．平均発話回数は提示方法2がもっとも少なかったが，正解選択までの平均所要時間は提示方法 3 がもっとも短くかかった．

また，各提示方法を実際に利用してみた評価として，

(1) 常に1候補のみ提示する場合

- 正解を探す必要がなく，すぐに結果がわかるので正解かどうかの判別が容易である．
- 正解に似た結果が出た時に，もう少し候補が提示されれば正解が得られたのではないかと思う．
- 正解が得られるまで何度も同じ発話を繰り返すのは負担である．

という意見があった．

(2) 常に30候補提示する場合

- 第 1 候補が正解でなくてもそれ以降に正解があることが多く，もう1度発話しなくても良い．
- 候補中に正解が含まれている確率は高いが，多くの候補の中から正解を探す手間がかかる．
- 候補数が多すぎて，正解を探すのが面倒である．

という意見があった．

(3) 認識スコアを用いて候補提示数を決定する場合

- 提示される候補数が少なく，候補中に正解が含まれている確率が高かったため使いやすい．
- 候補数が比較的少なくって正解を探しやすい．
- 第 1 候補が正解でなくてもそれ以降に正解があることが多く，もう1度発話し

なくてもよいため，常に 1 候補のみ提示する場合に比べて楽である．

- 上位候補が正解であるにもかかわらず，候補が多く提示されることがあり，認識スコアを用いた効果が見られない場合もある．

という意見があった．

３つの提示方法のうち，認識スコアを用いて候補提示数を決定する場合が最も使いやすかったと全員が答えた．これは，提示される候補数が少ないために正解を探しやすかったことと，提示された候補中に正解が含まれていることが多かったことにより，使いやすさを感じたためと思われる．残りの提示方法については常に 1 候補提示する場合の方が使いやすかったと答えた人がやや多かった．これは，同じ発話を繰り返すことより，多くの候補中から正解を探すことにわずらわしさを感じる人が多かったためと思われる．

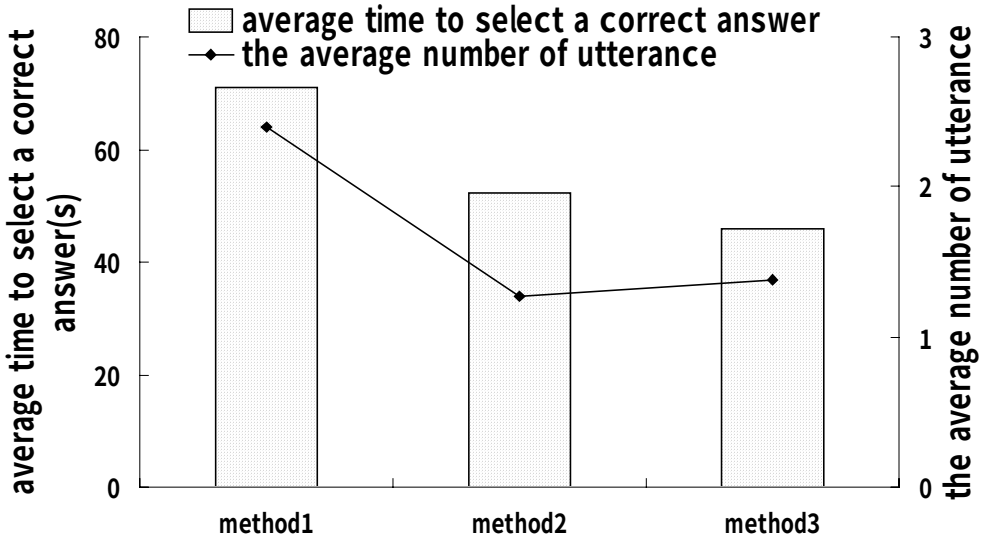


図 4.20: 正解選択までの平均所要時間と平均発話回数

Figure 4.20: The average number of utterance and average time to select a correct answer.

4.6 まとめ

本研究では，N-best 候補の認識スコアを利用して候補提示数を決定するため，認識スコアの分析を行なった結果，

- 第 n 候補と第(n+1)候補のスコアの差（n=1，2，3，4）
- 第 1 候補と第 n 候補のスコアの差
- 第 n 候補のスコア（n=2，3，4，…，30）

を利用して候補提示数を決定すれば，正解提示率をあまり下げることなく，提示する候補数を大幅に減らせることが分かった．また，認識スコアと併せて信頼度を用いて認識候補提示数を決定すれば，正解が含まれる割合（正解提示率）の減少を平均で 1%以内に抑えながら，提示する候補数を平均で 73%以上減らせることを示した．

そこで，評価実験として「常に 1 候補のみ提示する」場合，「常に 30 候補提示する」場合，「認識スコアを用いて候補提示数を決定する」場合の 3 つの候補提示方法をそれぞれ実際に利用してもらい，どの場合が最も使いやすいか比較を行なった．その結果，全員が，認識スコアを用いて候補提示数を決定する場合が最も使いやすいと評価した．しかし，第 1 候補が正解にもかかわらず，多く候補を表示してしまう場合や，提示した候補中に正解がない場合もあったため，まだまだ認識スコアを分析していく必要があると思われる．

参考文献

［1］ 桐山伸也，広瀬啓吉，“文献検索をタスクとした音声対話システムの応答生成,” 情報処理学会研究報告（音声言語情報処理研究会），SLP-27-16, pp.105-110, July 1999.

［2］ K. Kondo, C. Hemphill, “Surfin’ the World Wide Web with Japanese,” Proc. ICASSP’97, vol.2, pp.1151-1154, April 1997.

［3］ 近藤玲史，稲垣敬子，磯健一，三留幸夫，“音声インタフェースを用いた Web 新聞へのアクセス,” 情報処理学会研究報告（音声言語情報処理研究会），SLP-16-8, pp.43-48, May 1997.

［4］ 杉浦誠，中村哲，鹿野清宏，“音声キーワードによるネットサーフィンの実現,” 情報処理学会研究報告（音声言語情報処理研究会），SLP-20-12, pp.69-74, Feb. 1998.

［5］ 井上武士，山下洋一，“コマンド入力システムを対象とした音声入力インタフェース,” 日本音響学会 2001 年春季研究発表会講演論文集，3-P-27, pp.217-218, Mar. 2001.

［6］ 中野崇広，甲斐充彦，中川聖一，“WWW 上のフォーム型情報検索サービスのための音声インタフェースの検討,” 情報処理学会研究報告（音声言語情報処理研究会），SLP-25-1, pp.1-6, Feb. 1999.

［7］ 西本卓也，小林豊，新美康永，“WWW 上のデータベース検索のための汎用音声インタフェース,” 日本音響学会 1997 年春季研究発表会講演論文集，2-Q-20, pp.179-180, Mar. 1997.

［8］ 山本幹雄，高木三功，中川聖一，“メニューに基づく音声対話システムとその評価,” 電子情報通信学会技術研究報告（音声研究会），SP93-130, pp.17-24, Jan. 1994.

［9］ R. Lau, G. Flammia, C. Pao, and V. Zue, “WEBGALAXY - Integrating Spoken Language and Hypertext Navigation,” Proc. of EUROSPEECH’97, vol.2, pp.883-886, Sept. 1997.

［10］ Sunil Issar, “A Speech Interface for Forms on WWW,” Proc. of EUROSPEECH’97, vol.3, pp.1343-1346, Sept. 1997.

［11］ 翠輝久，駒谷和範，河原達也，奥乃博，木戸冬子，“音声対話によるソフトウェアサポートのための確認戦略，” 情報処理学会研究報告（音声言語情報処理研究会），SLP-47-11, pp.53-58, July 2003.

［12］ http://www.lang.astem.or.jp/CSRC/

［13］ http://julius.sourceforge.jp/

［14］ 趙 國，宮山 章子，山下 洋一，“N-best 音声認識における認識スコアを利用した候補提示数の決定，” 電子情報通信学会論文誌(D-II)，掲載決定.

［15］ Cho and Y. Yamashita, “Determination of the Number of Candidates Using Recognition Scores for N-best Based Speech Interface,” Proc. of the 6th IASTED International Conference on Signal and Image Processing (SIP2004), 444-196, Aug. 2004.

［16］ 迫江博昭，藤井浩美，吉田和永，亘理誠夫，“フレーム同期化，ビームサーチ，ベクトル量子化の統合による DP マッチングの高速化，” 電子情報通信学会論文誌，Vol.J71-D, No.9, pp.1650-1659, Sept. 1988.

［17］ 沢井秀文，米山正秀，中川聖一，“ベクトル量子化された語中 VCV 音節パターンと後処理を用いた大語彙音声認識法，” 日本音響学会誌，Vol.43，No.7, pp.752-763, July 1987.

［18］ 中川聖一，甲斐充彦，“文脈自由文法制御による One Pass 型 HMM 連続音声認識法，” 電子情報通信学会論文誌，Vol.J76-D-II，No.7，pp.1337-1345，July 1993.

第５章

結論

本論文では，音声認識システムの性能を向上させ，人間にとって使いやすいインタフェースを構築するための要素技術として，音響尤度の算出方法を高度化することによって音声認識システムの性能を改善する手法とユーザにとって使いやすい認識結果の候補提示手法について提案し，その有効性を検証した．

第 3 章では，不特定話者音声認識において，話者の適応データを用いることなく音素間の依存性を利用する手法について述べた．音素の音響的特徴は話者の違いにより広い範囲に変動するが，音素間の相対的な関係には話者によらず強い依存性がある．従来の音声認識における音響尤度の算出ではこのような音素間の相関が考慮されていない．多数話者のデータに基づいて構成する話者空間を利用して認識文候補における母音の相対的関係の妥当性を評価した．さらに，認識文候補における母音特徴を前後の音素コンテキストを考慮した正規化を行なうことによって，精度を改善する手法を示した．Julius の N-best 結果に対して，提案手法による再評価を行なうことにより約 25%の認識誤りが削減された．本論文で提案した手法では，入力中に５母音全てが出現することが前提となっている．入力音声にすべての母音が含まれない場合の対応として，５母音のうち一部が出現しなかった場合は多数話者による母音の平均値を利用することが考えられる．また，４母音からなる話者空間を５つ構成するなど，少ない母音数による話者空間を用いる手法も考えられる．今後の課題としては，入力音声に５母音全てが出現しない場合への対応の他，母音特徴の音素コンテキストに対する正

規化手法の改善などが挙げられる．

第 4 章では，N-best 候補の認識スコアを利用して候補提示数を決定する手法について述べた．現在，音声認識技術が大きく向上したが，認識結果を一つだけ提示する場合，正しい結果が常に得られるとは限らないのが現状である．そのため，認識結果を複数候補提示し，利用者が候補の中から正解を選択する音声インタフェースの研究が行われている．このような N-best 方式の音声認識に基づく音声インタフェースでは，提示する候補数の決定が重要な問題となる．認識候補を多く表示すれば正解が含まれる確率は高くなるが，ユーザが正解を探す手間も増える．そこで，N-best 候補の認識スコアの分布を利用して候補提示数を動的に決定する手法を検討した．認識スコアの分析を行なった結果，認識スコアを利用して候補提示数を決定すれば，正解提示率をあまり下げることなく，提示する候補数を大幅に減らせることが分かった．また，認識スコアと併せて信頼度を用いて認識候補提示数を決定すれば，正解が含まれる割合（正解提示率）の減少を平均で 1%以内に抑えながら，提示する候補数を平均で 73%以上減らせることを示した．しかし，第 1 候補が正解にもかかわらず，多く候補を表示してしまう場合や，提示した候補中に正解がない場合もあったため，まだまだ認識スコアを分析していく必要があると思われる．

音声インタフェースが我々の日常社会に今後もっと普及するためには，まだまだ，様々な技術的な問題や制約を解決してしなければならない．しかし，人間に優しく，使いやすい音声インタフェースがもっと身近に来ることを信じて願っている．

謝辞

本論文を執筆するにあたり，懇切丁寧な御指導と御鞭撻を賜わった立命館大学情報理工学部メディア情報学科，山下洋一教授に心から感謝致します．さらに，本論文に関して，御意見及び御指導を下さった立命館大学情報理工学部メディア情報学科，樋口宜男教授，八村広三郎教授に深く感謝致します．大学院での研究活動を成し遂げることができたのは，山下洋一教授の研究に対する深い御理解と的確かつ暖かい御助言に励まされたおかげです．また，理工学部と理工学研究科博士課程前期課程において，御指導を頂いた立命館大学理工学部電気電子工学科，寺井秀一教授に心から感謝致します．

本研究を進めるにあたり，立命館大学理工学部情報学科在籍時に実験に協力していただいた宮山章子氏，一丸太一郎氏，小橋修一氏に深く感謝致します．また，本研究を通し，様々な点で援助して下さった音声言語研究室の諸氏に感謝致します．

また，学会，研究会等でも，多くの方々に活発な議論をして頂きました．改めて皆様に感謝致します．

最後に，本論文作成にあたっていろいろな協力してくれた家族に感謝の意を表します．

謝辞

研究業績

本論文に関わる発表文献

論文

1. 趙國，一丸太一郎，山下洋一，“話者空間モデルに基づいた音素間相関を用いた音声認識，” 電子情報通信学会論文誌(D-II)，Vol.J87-D-II，No.7，pp.1402-1408，July 2004.

2. 趙 國，宮山 章子，山下 洋一，“N-best 音声認識における認識スコアを利用した候補提示数の決定，” 電子情報通信学会論文誌(D-II)，掲載決定.

国際会議（査読付き）

1. K. Cho and Y. Yamashita, “Speech Recognition Using Inter-Phoneme Dependency,” Proc. of the Eighth Western Pacific Acoustic Conference (WESPAC8), MB32, April 2003.

2. K. Cho and Y. Yamashita, “Speech Recognition Using Inter-phoneme Dependency Based on a Speaker Space Model,” Proc. of the 18th International Congress on Acoustics (ICA2004), 5, pp.3507-3510, April 2004.

3. K. Cho and Y. Yamashita, “Determination of the Number of Candidates

Using Recognition Scores for N-best Based Speech Interface," Proc. of the 6th IASTED International Conference on Signal and Image Processing (SIP2004), 444-196, Aug. 2004.

研究会

1. 趙國，山下洋一，“N-best 音声認識における認識スコアを利用した候補提示数の決定，” 電子情報通信学会技術研究報告，SP2003-182，pp.43-48，Jan. 2004.

国内学会

1. 趙國，山下洋一，小川均，“音素間の相関を用いた音声認識，”日本音響学会 2001 年春季研究発表会講演論文集，3-P-15，pp.193-194，Mar. 2001.

2. 趙國，山下洋一，“話者空間のモデルに基づいた音声認識，”日本音響学会 2001 年秋季研究発表会講演論文集，3-1-5，pp.101-102，Oct. 2001.

3. 趙國，山下洋一，“音素コンテキスト情報を考慮した話者空間モデルに基づく音声認識，”日本音響学会 2002 年秋季研究発表会講演論文集，3-5-3，pp.111-112，Oct. 2002.

4. 趙國，山下洋一，“N-best 認識スコアを利用した音声認識における候補数提示数の決定，”日本音響学会 2004 年春季研究発表会講演論文集，3-8-11，pp.143-144，Mar. 2004.