Internal Use Only（非公開）

TR－SLT－0017

複数の大語彙連続音声認識モデルの出力の共通部分を用いた信頼度
—信頼度を利用した複数モデルの出力の混合—

A CONFIDENCE MEASURE BASED ON AGREEMENT AMONG MULTIPLE LVCSR MODELS
—Mixing Output of Multiple model based on confidence measure—

渡邊 友裕　　小窪 浩明
Tomohiro Watanabe　　Hiroaki Kokubo,
山本 博史　　菊井 玄一郎
Hirofumi Yamamoto　　Genichiro Kikui

平成14年8月19日

概要

大語彙日本語連続音声認識において、デコーダ、音響モデル、言語モデルなど、設定が少しずつ異なっている複数モデルによる認識の結果について、その認識結果の共通部分が正解となっている割合を測定することにより、複数モデルの出力の共通部分の信頼度を評価した。これらの多数の要因の各々が、複数モデルの出力の信頼度の性能に寄与する度合を評価し、高い信頼度に寄与する度合の大きい重要な要因について分析を行った。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2　TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

目 次

1 はじめに

近年，音声認識結果の正解部分と誤り部分を分離することを目的として信頼度 (Confidence Measure) の研究が行なわれている (例えば，連続音声認識では [4, 14, 8, 7] など)．これまで提案されてきた信頼度尺度の多くは，いずれも，単一の認識エンジン・認識モデルが出力する認識結果を用いて，その正解部分と誤り部分を分離するというものであった．一方，連続音声認識の認識率そのものの向上を目的とする研究においては，複数の認識システムの出力を統合する方式 (ROVER 法 [2, 9]) も提案され，一定の効果が報告されている．我々は，ROVER 法のような (重み付き) 多数決法が認識率の改善に効果的であることを考慮して，音声認識結果の正解部分と誤り部分を分離するための信頼度尺度として，複数の音声認識システムの出力の共通部分を用いる方法を提案し，その有効性を示した [12, 10]．評価実験の結果では，デコーダおよび音響モデルが異なる二つのモデルについて，出力の共通部分の信頼度を評価したところ，最も高い性能が達成された．これまでの研究 [12, 10] における評価音声データは，新聞読み上げ音声とニュース音声のみであったが，本論文では，新たに，ATR 旅行会話音声において評価実験を行った．今回、実験に用いた ATR 旅行会話音声の set01 では，単語正解率を 20%犠牲にすることにより，97%近い適合率を達成し，set02 でも，単語正解率を 20%犠牲にすることにより，97%近い適合率を達成した．また，同一のデコーダを用いた場合は，set01 で 92%程度，set02 でも 92%程度という適合率であった．

さらに [11] では，デコーダ，音響モデルの異なる 26 種類の大語彙日本語連続音声認識モデルについて，機械学習 (決定リスト学習 [15]) の枠組を用いて，単語の品詞ごとあるいは音節数ごとに，最も適合率の高いモデルの組合せを選択的に組み合わせる規則を学習し，この混合規則を適用することにより，単独モデルの適合率・再現率を改善できた．また，ROVER 法のような (重み付き) 多数決法との比較においても，機械学習を用いた混合法の性能が上回ることを示した．しかし，以前の研究では大語彙連続音声認識モデルのデコーダは SPOJUS [1] と Julius [3] の 2 種類のみで実験を行ったため，本論文では新たに SPREC というデコーダを追加して同様の実験を行い，どのような結果が得られるかの調査を行った．更に，本論文では機械学習として決定リスト学習だけでなく，SVM [13] も用いて評価を行った．

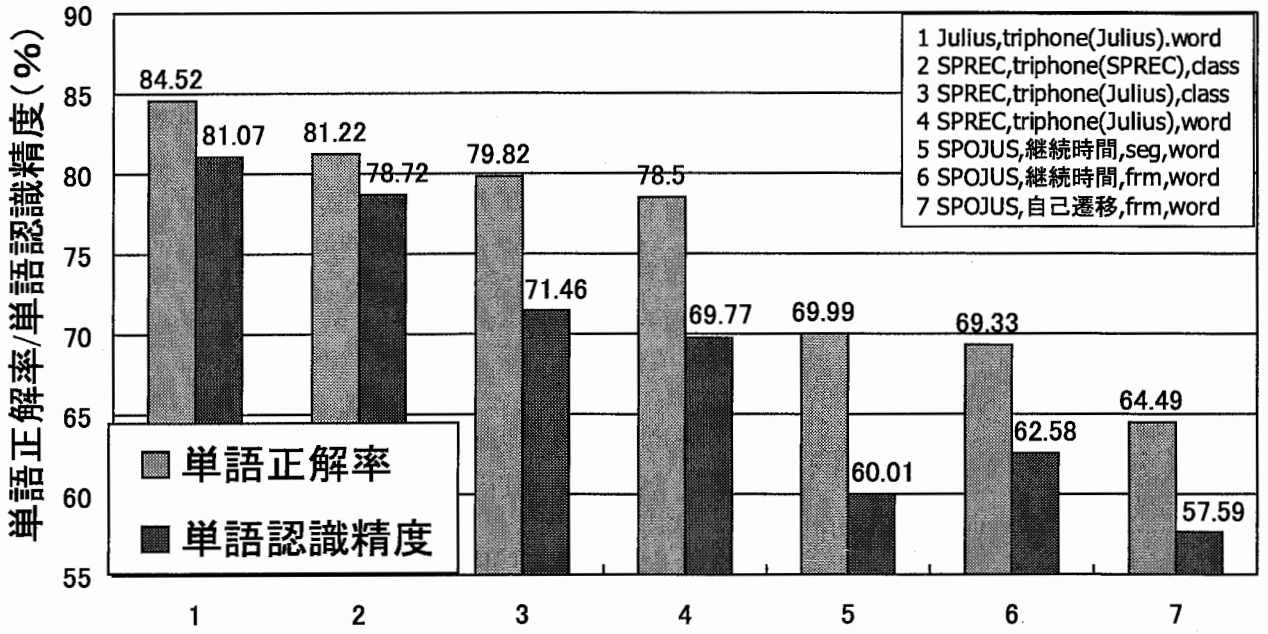
2 実験条件

2.1 日本語大語彙連続音声認識モデル

デコーダ (および音響モデル) としては，SPOJUS [1] (音響モデル [6] は，16kHz サンプリング，フレーム周期 10ms，特徴ベクトルはセグメント単位/フレーム単位の MFCC の二種類，音節モデル, 無音モデル有り，全共分散行列，継続時間制御/自己遷移ループ，等合計 3 種類) および Julius [3] (音響モデルは，16KHz サンプリング，フレーム周期 10ms，特徴ベクトルはフレーム単位の MFCC，トライフォンモデル，無音モデル有の 1 種類) を使用した．さらに，SPREC(音響モデルは，Julius で用いるトライフォンモデルと，SPREC で用いる，フレーム周期 10ms，トライフォンモデル，無音モデル有り等合計 3 種類) を使用した．言語モデルは，Julius と SPOJUS では，tri-gram モデルを用いた．SPREC では，class-tri-gram と tri-gram モデルの 2 種類を用いた．認識モデルのデコーダ，音響モデル，デコーダの詳細な組み合せは，以下の 7 通りの設定で評価した．

i) デコーダは Julius を用い，音響モデルは Julius のトライフォン，言語モデルは単語 tri-gram モデルを用いる．
ii) デコーダは SPREC を用い，音響モデルは SPREC のトライフォン，言語モデルは class-tri-gram を用いる．
iii) デコーダは SPREC を用い，音響モデルは Julius のトライフォン，言語モデルは class-tri-gram を用いる．
iv) デコーダは SPREC を用い，音響モデルは Julius のトライフォン，言語モデルは単語 tri-gram を用いる．
v) デコーダは SPOJUS を用い，音響モデルは継続時間制御 (特徴ベクトル MFCC-seg)，言語モデルは単語トライグラムを用いる．
vi) デコーダは SPOJUS を用い，音響モデルは継続時間制御 (特徴ベクトル MFCC-frm)，言語モデルは単語 tri-gram を用いる．
vii) デコーダは SPOJUS を用い，音響モデルは自己遷移ループ (特徴ベクトル MFCC-frm)，言語モデルは単語 tri-gram を用いる．

(a) 旅行会話音声 (set01)



(b) 旅行会話音声 (set02)

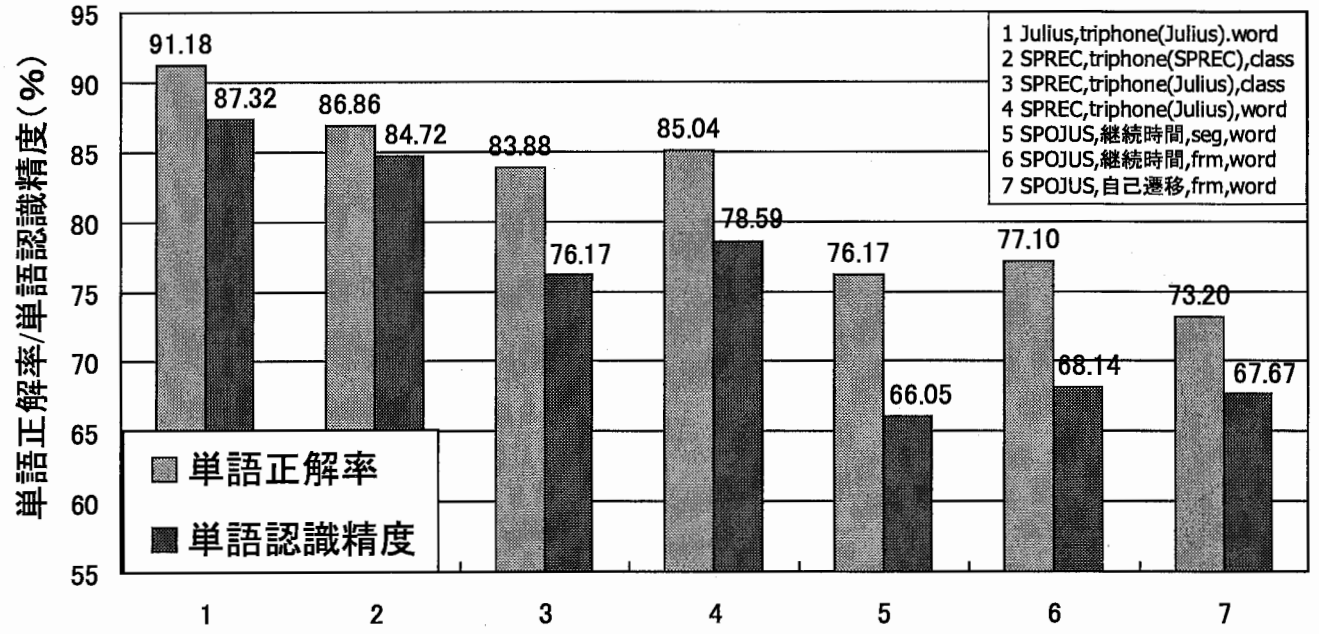


図 1: 単独モデルの単語認識率

2.2 評価データ

評価データとしては ATR 旅行会話音声 (phrasebook の V0 を用いたバージョン) の set01,204 文 (男性話者のみ,1362 単語)set02,306 文 (男性話者のみ,2145 単語) の音声の二種類を使用した.

set01 の単語認識率は，SPOJUS で単語正解率 69.99~64.49%，単語正解精度 62.58~57.59%，Julius で単語正解率 84.52%，単語正解精度 81.07%，SPREC で単語正解率 81.22~78.50%，単語正解精度 78.72~69.77%，set02 の単語認識

率は，SPOJUSで単語正解率77.10~73.20%，単語正解精度68.14~66.05%，Juliusで単語正解率91.18%，単語正解精度87.32%，SPRECで単語正解率86.86~83.88%，単語正解精度84.72%~76.17%，であった (これらの単語認識率は，単語・品詞・読みの三項組によって評価したものであり，単語のみで評価した認識率よりも1%程度低い値となっている)．詳しくは，図1に示す．

3 信頼度の評価尺度

本節では，本論文で用いる信頼度の評価尺度を定義する．一般には，大語彙連続音声認識モデルが出力する認識結果の各単語の信頼度を推定するタスクは，どの単語が正しく認識されていて，どの単語が誤認識であるかを推定することである．しかし，本論文では，正解単語がどの程度の精度で検出できるかに焦点を当て，複数の大語彙連続音声認識モデルの出力の共通部分が正解単語であると仮定した場合の，正解単語の再現率・適合率によって，複数モデルの出力の共通部分の信頼度を評価する．具体的には，二つのモデルの出力を Hyp_1 および Hyp_2 とする．

$$
\begin{aligned}
Hyp_1 &= w_{11}, \ldots, w_{1i}, \ldots, w_{1k} \\
Hyp_2 &= w_{21}, \ldots, w_{2j}, \ldots, w_{2l}
\end{aligned}
$$

このとき，DPマッチングによって対応付けられ，しかも，その表層形が同一の単語 w_{1i} と w_{2j} $(w_{1i}=w_{2j})$ を集めることにより「一致単語リスト」を構成する．そして，一致単語リストと正解文を比較し，各単語の表層形と出現位置を考慮して，一致単語リスト中の正解単語を判別し，次式によって再現率・適合率を算出する[1]．

$$
\begin{aligned}
\text{再現率} &= \frac{\text{一致単語リスト中の正解単語数}}{\text{正解文の単語数}} \\
\text{適合率} &= \frac{\text{一致単語リスト中の正解単語数}}{\text{一致単語リストの単語数}}
\end{aligned}
$$

本論文では，信頼度に対する要件として，正解単語をいかに高い精度で推定するかという点を重視し，再現率よりも適合率に重点をおいて，一定以上の再現率のもとでどれだけ高い適合率を達成できるかを重視して評価を行う．

図2は単一の認識モデルのなかで認識率最大のものと，あらゆる可能な二つのモデル組の出力の間の共通部分のなかで，適合率が最大のものとを比較したものである．異なるデコーダ間というのは2つのモデル組のデコーダが異なっていることを示し，同一デコーダ間というのはデコーダが同じであることを示している．set01では，単語正解率を20%犠牲にすることにより，97%近い適合率を達成し，set02でも，単語正解率を20%犠牲にすることにより，97%近い適合率を達成した．また，同一のデコーダを用いた場合は，set01で92%程度，set02でも92%程度という適合率であった．

4 信頼度を用いた複数の大語彙連続音声認識モデルの出力の混合

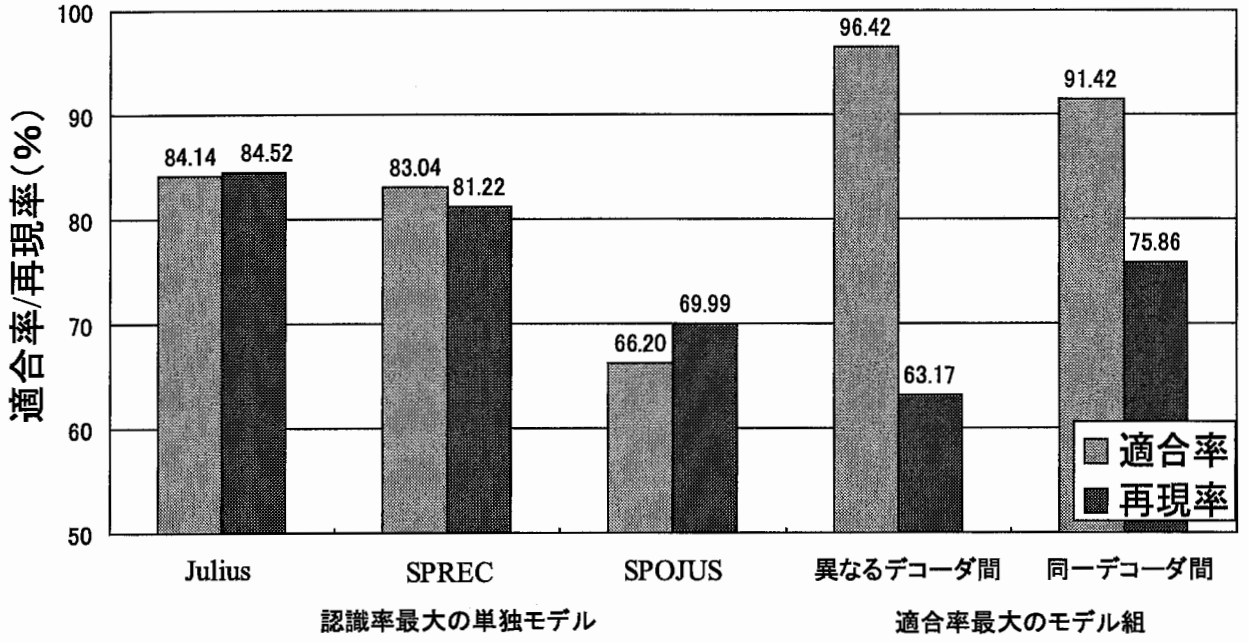
単語の品詞，音節数，および各単語を出力したモデルの情報を素性として，機械学習を行った．そして，機械学習の手法として，決定リスト学習 [15] と SVM [13] を評価した．

4.1 決定リスト学習

決定リストは，ある素性のもとでクラスを決定するという規則を優先度の高い順にリスト形式で並べたもので，適用時には優先度の高い規則から順に適用を試みていく．本論文では，各規則の優先度として，素性 f の条件のもとでの，クラス c の条件付き確率 $P(c \mid f)$ を用い，この条件付き確率順に決定リストを構成する．本論文では，複数モデルの出力の混合の問題に対して決定リスト学習を適用するにあたっては，まず，評価音声データ (ATR旅行会話音声 set01，204文，set02，306文) を，訓練データセットと評価データセットに話者が重複しないように分割する (set01，102文ずつ，set02，153文ずつ)．そして，訓練データセットを用いて学習した決定リストを評価データセットに適用するこ

[1]再現率は単語正解率と同じ評価式となっている．

(a) 旅行会話音声 (set01)



(b) 旅行会話音声 (set02)

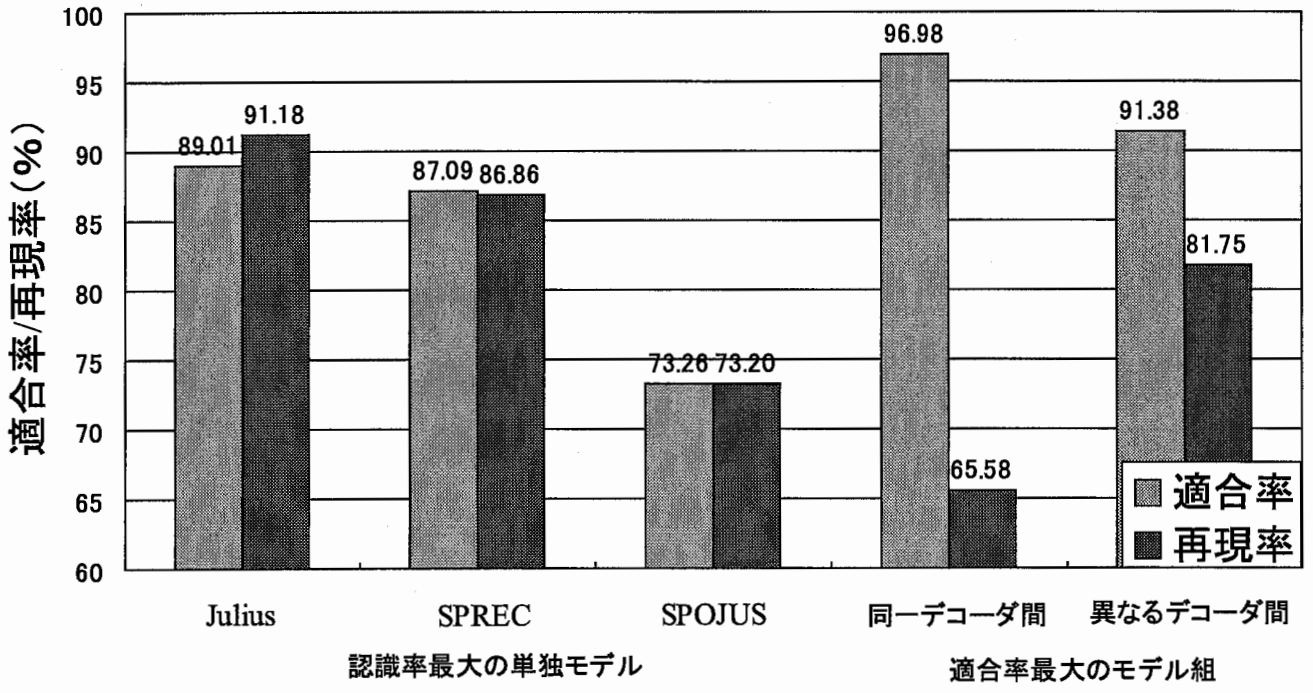


図 2: モデルの出力の共通部分の信頼度

とにより，複数モデルの出力の混合の評価を行う．決定リストの学習・適用に際しては，まず，混合の対象となる複数のモデルの出力の単語列に対して DP マッチングを行い，単語ラティスを構成する．そして，単語ラティス中の各単語の正誤を決定リストのクラス c とし，このクラスを決定するための決定リストを学習する．決定リスト学習の素性としては，以下の 4 種類の素性を全て用いた．

i) 各単語を出力した二つのモデルの組を用いる[2]

[2]本論文の素性の設定においては，複数モデルの出力の混合において，ある単語が出力として選択されるためには，少なくとも二つ以上のモデル

ii) 各単語を出力した二つのモデルの組・その単語の品詞 (「茶筌」 [5] の最も粗い 9 品詞) を結合したものを用いる
iii) 各単語を出力した二つのモデルの組・その単語の音節数を結合したものを用いる
iv) 各単語を出力した二つのモデルの組・その単語の品詞・その単語の音節数を結合したものを用いる

また，決定リストの適用の際には，決定リストの各規則に対して，素性 f の条件のもとでのクラス c の頻度 $freq(f, c)$ の下限，および，条件付き確率 $P(c \mid f)$ の下限を設け，単語ラティス上の競合する単語中で，これらの下限を満たす単語が存在すれば，その中で最も優先度 (条件付き確率 $P(c \mid f)$) の高い単語を選択する．単語ラティス上の競合する単語中で，頻度および条件付き確率の下限を満たす単語が存在しない場合には，その区間には高信頼度の単語は存在しないとして，単語を出力しない．

4.2 SVM

SVM(サポートベクトルマシーン) [13] は，一つ一つの属性を次元として考えると，単語の持つ属性の組み合せをベクトルとして考えることができるので，これを利用して 2 クラスの分類 (正解か誤りかなど) を行うための座標変換式を学習し，これによって，クラス分類を行う方法である．クラス分類時に計算される 2 クラスの境界からの距離 (SVM では，そのベクトルが座標変換後ある境界面からどちら側に変換されたかでクラス分類を行うのだが，その際の境界面からの距離) を信頼度とすることで，音声認識性能 (再現率・適合率) の向上を図る．今回，ツールとしては Tiny-SVM(http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/) における SVM を利用した．今回，データ形式によりモデルの性能にどのような違いが現れるのかを調査するために，2 つのデータ形式を用いた．

素性として i) と ii) と iii) を用いたものを形式 1 とし，i) と ii) と iv) を用いたものを形式 2 とする．

i) その単語の品詞
ii) その単語の音節数
iii) モデル情報 (その単語はどのモデルが出力したか)
iv) モデル組情報 (その単語はどのような 2 つのモデル組の出力の共通部分か)

2.2 節の評価音声データ (ATR 旅行会話音声 set01,204 文，set02,306 文) を，訓練データセットと評価データセットに話者が重複しないように分割する (set01,102 文ずつ，set02,153 文ずつ)．そしてこれらについて，形式 1，2 の 2 つの形式のデータを作成した．SVM モデルの学習を行わせた後，各々のモデルの評価を行った．今回 -t オプション (座標変換のため式の形を決定するもの) は (多項式カーネル)1 とし,-d オプション (座標変換のための式の形を決定するもの) は 1 と 2 を用い，-c オプション (大きいほど学習データに対しての識別誤りを許す) は下記に示す値を用いて評価を行った．今回設定したオプションとその意味は以下のようになる．

- -t オプション

  1 ...多項式カーネル $(\mathrm{sw}*\mathbf{x} + \mathrm{r})^n$ の形式の座標変換式の s, **w**, r を学習する．n の値は -d オプションにて決定する．
  ※ここで **x** は観測されたベクトルに相当し **w** は **x** と同じ次元を持つ． s, r はスカラー値．

- -d オプション

  1 ...$(\mathrm{sw}*\mathbf{x} + \mathrm{r})$ の座標変換式を使用する．(1 次)

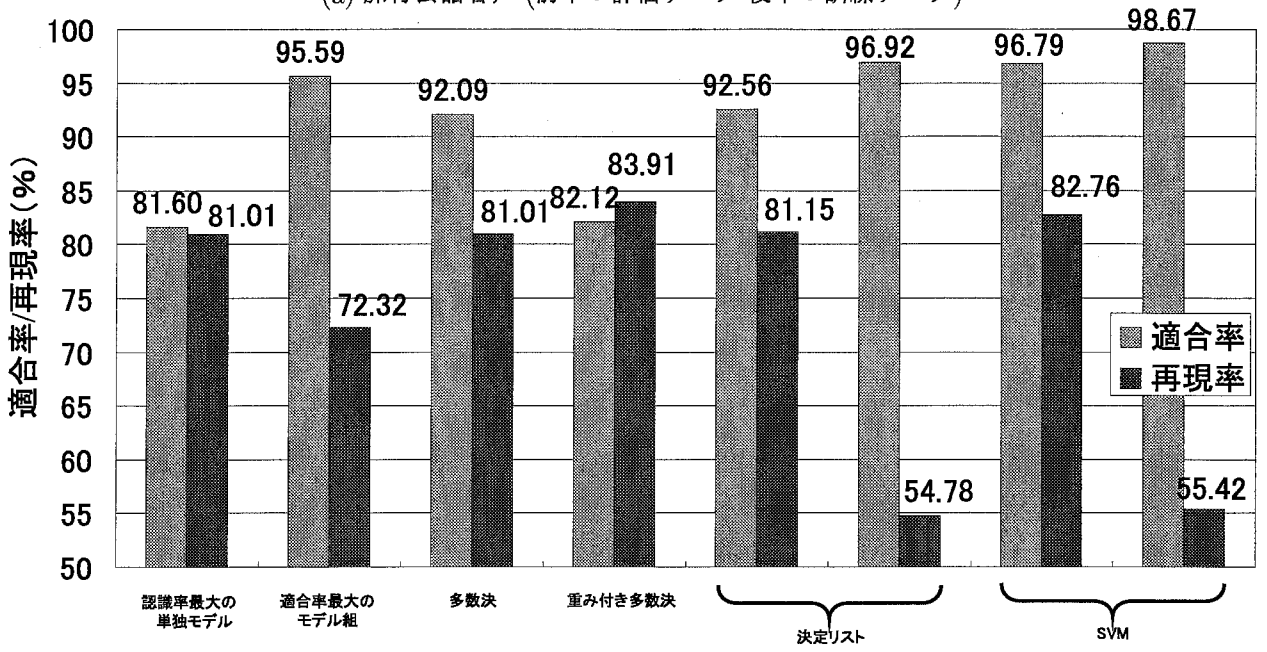
  2 ...$(\mathrm{sw}*\mathbf{x} + \mathrm{r})^2$ の座標変換式を使用する．(2 次)

- -c オプション

$$\frac{1}{\text{学習データの各々のベクトルにおいて「値が 0 でない次元数」の平均}}$$

によって出力される必要がある，という制約を課している．

(a) 旅行会話音声 (前半：評価データ 後半：訓練データ)



(b) 旅行会話音声 (前半：訓練データ 後半：評価データ)

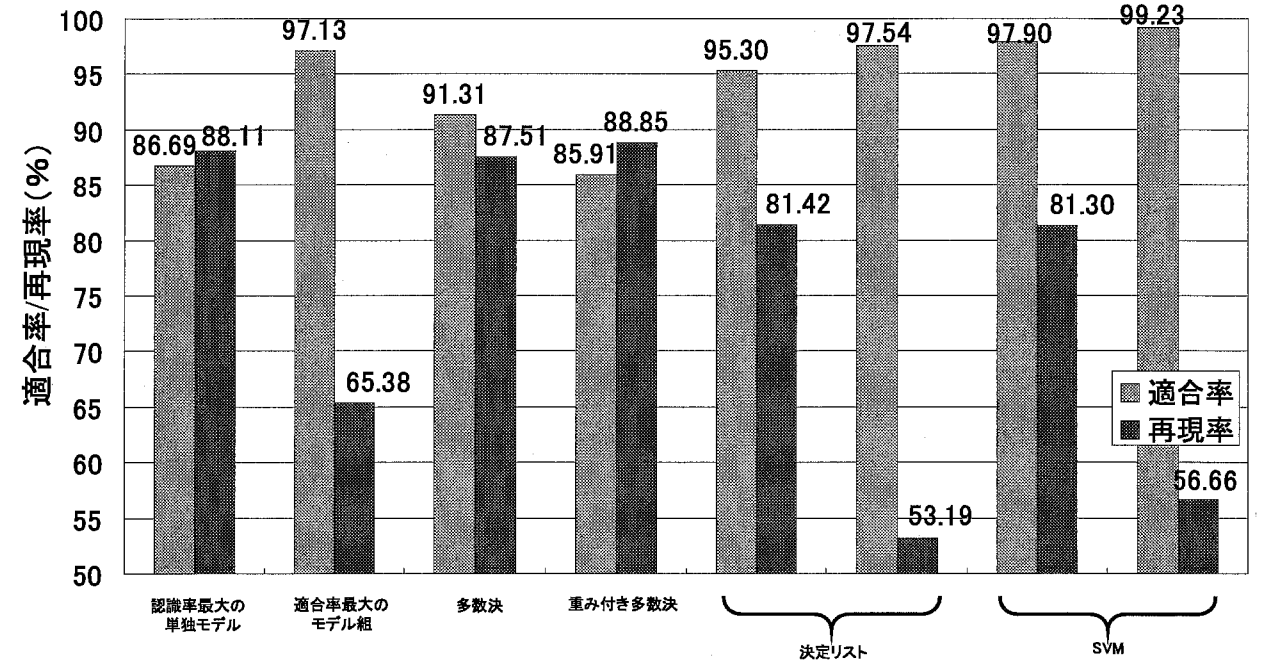
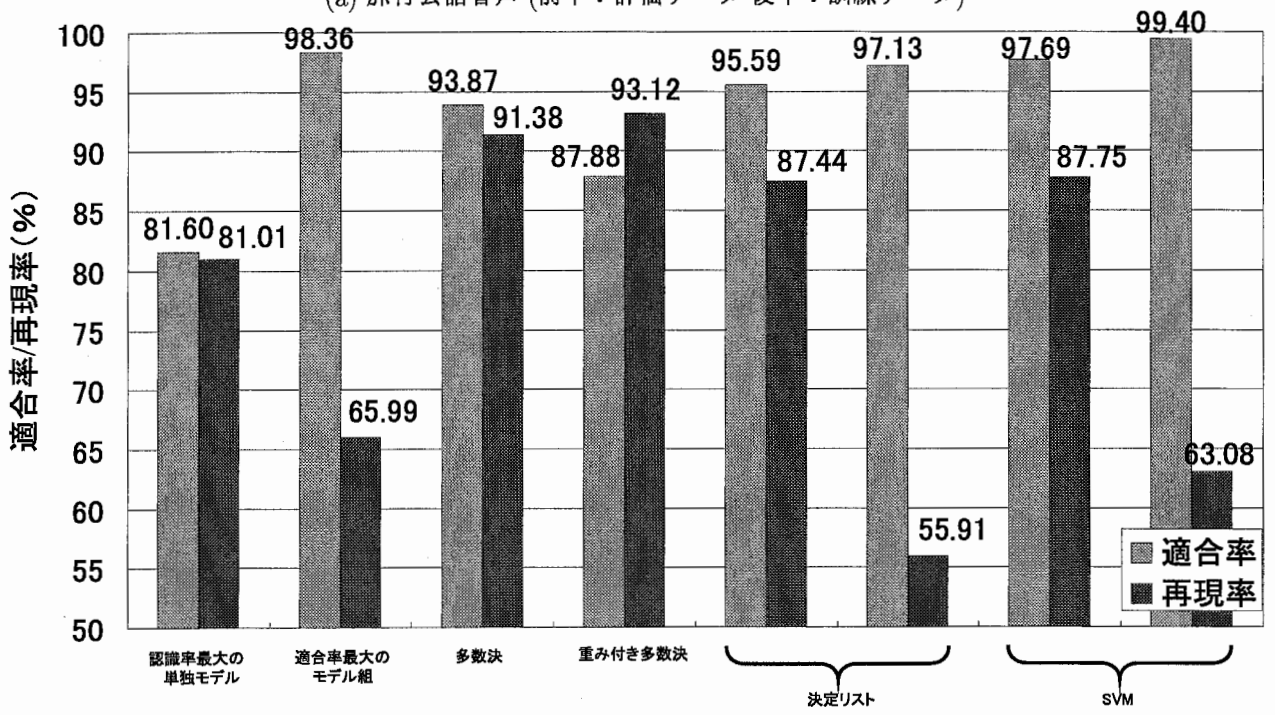


図 3: 複数モデルの出力の混合結果 (set01)

4.3 評価結果

本節では，前節で述べた決定リストおよびSVMにより複数モデルの認識結果を混合する実験を行った評価結果について述べ，その性能を，認識率最大の単独モデル，適合率最大のモデル組，(重み付き) 多数決，などによる性能と比較する．まず，混合の対象となる複数モデルとしては，2.1 節で述べた全 7 種類のモデルのうち，認識率の高い順に n 個 $(3 \leq n \leq 7)$ 選択して混合を行い，最も性能が良かった何通りかの結果を示す．(重み付き) 多数決においても，同様の

(a) 旅行会話音声 (前半：評価データ 後半：訓練データ)



(b) 旅行会話音声 (前半：訓練データ 後半：評価データ)

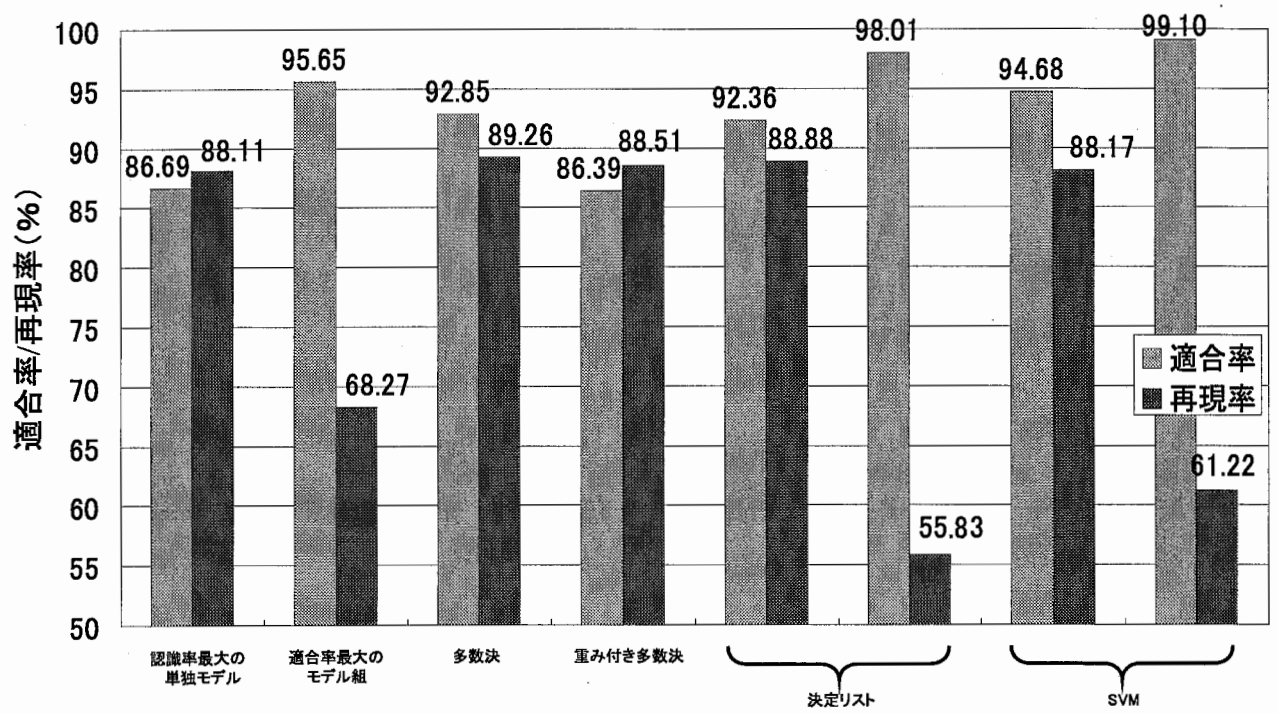


図 4: 複数モデルの出力の混合結果 (set02)

方法で n 個 $(3 \leq n \leq 7)$ のモデルの出力の混合を行い，最も性能が良かった結果を示す．ただし，重み付き多数決においては，各モデルの認識率を重みとして用い，重みなし多数決においては，モデル数が同数の場合にその区間では単語を出力しない，という方法をとっている．

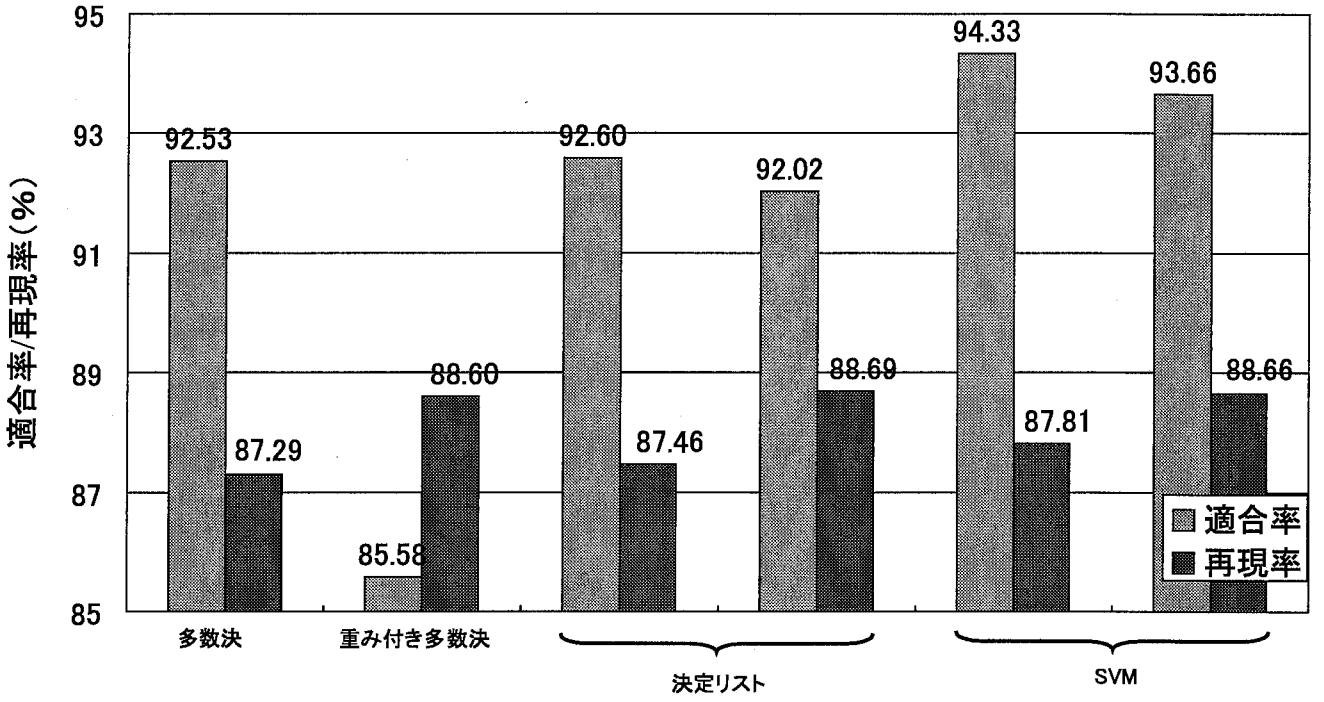


図 5: 信頼度の平均値

まず，決定リストによる混合，SVMによる混合，認識率最大の単独モデル，適合率最大のモデル組，(重み付き)多数決の性能を比較した結果を図 3，4に示す．

また，ここで用いているSVMによる混合の評価結果は4つの方法(形式1の1次，形式1の2次，形式2の1次，形式2の2次)のなかで最も高い性能が得られたものを用いた．

認識率最大の単独モデルと機械学習の比較では再現率は劣るものの，適合率は大幅に改善できていることが分かる．また，適合率最大のモデル組と比較すると，最大の適合率からはやや劣るものの，再現率を改善できることが分かる．

機械学習(決定リスト，SVM)は閾値を変化させることにより，いくつかの値が得られる．図 3，4に示されている値は適合率を重視し，再現率はあまり考慮していない値であるため，(重み付き)多数決と機械学習との比較は難しい．そのため，閾値を変化させ，(重み付き)多数決と比較して再現率が同等の値をとる点について，図 3(a)(b)，図 4(a)(b)の4種類の実験結果の適合率，再現率の平均をとって比較を行った．その結果を図 5に示す．この結果より，(重み付き)多数決より決定リストの方が有効であり，決定リストよりSVMのほうが有効であることが分かる．

(重み付き)多数決 < 決定リスト < SVM

5 おわりに

本論文では，複数の音声認識システムの出力の共通部分を用いる信頼度尺度 [12, 10]，および信頼度を用いた複数モデルの出力の混合 [11] について，新たにSPRECというデコーダを追加し，機械学習の枠組を用いて，どのような結果が得られるかの調査を行った．そして，これまでの研究 [12, 11, 10] における評価音声データは，新聞読み上げ音声とニュース音声のみであったが，新たに，ATR旅行会話音声において評価実験を行った．また，(重み付き)多数決との比較においては，これを上回る性能が達成できた．さらに，(重み付き)多数決，決定リスト，SVMについてさらに詳細な信頼度についての評価を行った結果，(重み付き)多数決より決定リストの方が有効であり，決定リストよりSVMのほうが有効であることが分かった ((重み付き)多数決 < 決定リスト < SVM)．なお，今回の評価実験では，異なる三種類のデコーダを用いた，7種類のモデル全てを混合の対象とした．

今後，混合に最も適したモデルの組合せを自動的に選択する過程の最適化手法について，詳細な検討を行う予定である．

参考文献

[1] 赤松裕隆, 花井建豪, 甲斐充彦, 峯松信明, 中川聖一. 新聞・ニュース文をタスクとした大語彙連続音声認識システムの評価. 情報処理学会第 57 回全国大会講演論文集, pp. 35–36, 1998.

[2] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. ASRU*, pp. 347–354, 1997.

[3] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 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 (99 年度版). 日本音響学会誌 (技術報告), Vol. 57, No. 3, pp. 210–214, 2001.

[4] T. Kemp and T. Schaaf. Estimating confidence using word lattices. In *Proc. 5th Eurospeech*, pp. 827–830, 1997.

[5] 松本裕治, ほか. 日本語形態素解析システム『茶筌』version 2.2.7 使用説明書, 2001.

[6] 中川聖一, 花井建豪, 山本一公, 峯松信明. HMM に基づく音声認識のための音節モデルと triphone モデルの比較. 電子情報通信学会論文誌, Vol. J83–D–II, No. 6, pp. 1412–1421, 2000.

[7] 中川聖一, 堀部千寿. 音響尤度と言語尤度を用いた音声認識結果の信頼度の算出. 情報処理学会研究報告, Vol. 2001, No. (2001–SLP–36), pp. 87–92, 2001.

[8] 緒方淳, 有木康雄. 信頼度を組み込んだデコーディングによる 音声認識の検討. 情報処理学会研究報告, Vol. 2000, No. (2000–SLP–32), pp. 1–6, 2000.

[9] H. Schwenk and J.-L. Gauvain. Combining multiple speech recognizers using voting and language model information. In *Proc. 6th ICSLP*, Vol. II, pp. 915–918, 2000.

[10] T. Utsuro, T. Harada, H. Nishizaki, and S. Nakagawa. A confidence measure based on agreement among multiple LVCSR models — correlation between pair of acoustic models and confidence —. In *Proc. 7th ICSLP*, Vol. I, pp. 701–704, 2002.

[11] 宇津呂武仁, 原田哲志, 渡邊友裕, 西崎博光, 中川聖一. 複数の大語彙連続音声認識モデルの出力の共通部分を用いた信頼度 — 信頼度を利用した複数モデルの出力の混合 —. 電子情報通信学会技術研究報告, SP2002-18～23, pp. 25–30, 2002.

[12] 宇津呂武仁, 西崎博光, 原田哲志, 小玉康広, 中川聖一. 複数の大語彙連続音声認識モデルの出力の共通部分を用いた信頼度の性能分析. 電子情報通信学会技術研究報告, SP2001-125～135, pp. 25–32, 2002.

[13] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[14] F. Wessel, K. Macherey, and H. Ney. A comparison of word graph and N-best list based confidence measures. In *Proc. 6th Eurospeech*, pp. 315–318, 1999.

[15] D. Yarowsky. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. In *Proc. 32nd ACL*, pp. 88–95, 1994.