論　文

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高信頼度部分の推定

宇津呂武仁[†]　西崎　博光[††]　小玉　康広[††]　中川　聖一[††]

Estimating Highly Confident Portions Based on Agreement among Outputs of Multiple LVCSR Models

Takehito UTSURO[†], Hiromitsu NISHIZAKI[††], Yasuhiro KODAMA[††], and Seiichi NAKAGAWA[††]

あらまし　本論文では，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，複数の音声認識システムによる認識結果のうち複数のシステムの間で共通となっている部分を用いる方法を提案し，その有効性を示す．具体的には，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，音響/言語スコアの重み，挿入ペナルティなど，様々な設定が少しずつ異なっている二つの大語彙連続音声認識モデルによる認識結果について，そ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，二つの大語彙連続音声認識モデルによる認識結果の共通部分の信頼度を評価する．新聞読上げ音声及びニュース音声を対象として，2 種類のデコーダを用いて行った評価実験の結果では，デコーダ及び音響モデルが異なる二つの大語彙連続音声認識モデルについて，認識結果の共通部分の信頼度を評価したところ，非常に高い性能が達成された．また，同一のデコーダを用いた場合にも，音響モデルの特徴の違いと信頼度との相関を網羅的に評価することにより，デコーダが異なる場合の性能をやや下回るものの，ほぼそれに匹敵する性能を達成した．特に，混合連続分布 HMM に基づく音響モデルの場合では，無音モデルの有無，音響モデルの種類（トライフォンや音節モデルなど）の違いといった特徴が高い信頼度に寄与していることがわかった．

キーワード　大語彙連続音声認識，信頼度尺度，複数モデル混合，音響モデル，認識誤り検出

1. まえがき

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) の研究が行われている．例えば，連続音声認識では，音響安定度 (acoustic stability) を用いるもの [7]，単語グラフ中のエッジ接続数 [14] や仮説密度 (hypothesis density) [7] を用いるもの，音響・言語ゆう度 [13]，あるいは，事後単語確率 [19] を用いるものなどをはじめとして，数多くの研究が行われている．ここで，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 — Recognizer Output Voting Error Reduction) [3] も提案され，一定の効果が報告されている（例えば，文献 [15] など）．

本論文では，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ことを目的として，複数の音声認識システムによる認識結果のうち複数のシステムの間で共通となっている部分を用いる方法を提案し，その有効性を示す．具体的には，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，音響/言語スコアの重み，挿入ペナルティなど，様々な設定が少しずつ異なっている

† 京都大学大学院情報学研究科知能情報学専攻，京都市
Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto-shi, 606–8501 Japan
†† 豊橋技術科学大学工学部情報工学系，豊橋市
Department of Information and Computer Sciences, Toyohashi University of Technology, Tempaku-cho, Toyohashi-shi, 441–8580 Japan

二つの大語彙連続音声認識モデルによる認識結果について，そ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，二つの大語彙連続音声認識モデルによる認識結果の共通部分の信頼度を評価する．

ここで，より高い適合率を実現するために，3 種類以上のモデルを用いてその共通部分を利用するという方式も考えられるが，予備調査の結果，2 種類のモデルの組合せにおける最も高い適合率が上限となり，これを大きく上回る適合率は達成できないことを確認している [9]．また，3 種類以上のモデルを用いた多数決法でも，2 種類のモデルの組合せにおける最も高い適合率を上回ることはないことも確認している [16]．したがって，本論文では，二つのモデルの組合せを網羅的に評価するという手順をとる．

以上の考えに基づき，二つのモデルの出力の共通部分の信頼度を実験的に評価したところ，デコーダ及び音響モデルが異なる二つのモデルについて，最も高い性能が達成された．その性能は，新聞読上げ音声の場合，正解単語の約 87%を 99%近くの精度で予測でき，また，ニュース音声の場合，正解単語の約 64%を 95%近くの精度で予測できるという，非常に高いものであった．また，同一のデコーダを用いた場合にも，音響モデルの特徴の違いと信頼度との相関を網羅的に評価することにより，デコーダが異なる場合の性能をやや下回るものの，ほぼそれに匹敵する性能を達成した．具体的には，新聞読上げ音声及びニュース音声に共通して，音響モデルが無音モデルをもつか否か，あるいは，音響モデルの種類（トライフォンか音節モデルかなど），といった要因が高い信頼度に寄与する度合が大きいことがわかった．更に，これらの重要な要因を単独で用いるよりも，その他の様々な要因と組み合わせて用いることで，より高い信頼度が達成できるという結果が得られた．この評価結果より，混合連続分布 HMM に基づく音響モデルが通常もつとされる各種特徴の組合せについて，今回用いたデコーダの範囲内では，網羅的な評価を行うことができた．本論文では，新聞読上げ音声及びニュース音声を対象として，2 種類のデコーダを用いて行った評価実験を踏まえて，これらの分析結果の詳細について報告する．

2. 大語彙日本語連続音声認識モデル

本章では，本論文で用いた大語彙日本語連続音声認識の各モデルについて述べる．本論文では，デコーダとしては，IPA「日本語ディクテーション基本ソフトウェアの開発」プロジェクト [6] から提供された Julius，及び，豊橋技術科学大学工学部情報工学系中川研究室で開発された SPOJUS [1] を用いている．各々のデコーダで用いた音響モデル・言語モデルの一覧を表 1〜表 3 にそれぞれ示す．

2.1 音響モデル

音響モデルとしては，混合連続分布 HMM に基づくモデルを用い，特に，音素を基本単位とする HMM モデル，及び，音節を基本単位とする HMM モデルの二種類のモデルを評価対象とした．デコーダの実装の都合上，Julius と SPOJUS とでは，異なった音響モデルを用いている．以下では，各々のデコーダで用いている音響モデルについて簡単に説明する．

2.1.1 Julius で用いた音響モデル

Julius では，表 1 に示すように，音素を基本単位とする HMM モデル，及び，音節を基本単位とする HMM モデルを用いた．

音素モデルは IPA「日本語ディクテーション基本ソフトウェアの開発」プロジェクト [6] から提供されたものを用いた．これらのモデルは，無音モデルをもち，無音を含む訓練用音声データを用いて学習されたものである．本論文では，これらの音素モデルをそのまま用いて，無音モデルをもつ音響モデルとして認識を行った．音節モデル [11] についても，訓練データ，学習方式，HMM が自己遷移ループをもつ，無音モデルをもつ，などの諸条件は音素モデルの場合と同じである．なお，音素モデル，音節モデルのいずれにおいても，認識時に無音を考慮せず認識を行うことにより，

表 1　音響モデルの特徴（デコーダ：Julius）
Table 1　Specifications of acoustic models. (Decoder: Julius)

<table>
<tr><td rowspan="3">音響モデル</td><td colspan="3">音素モデル
（音素数：43（無音あり)/42（無音なし)，5 状態（3 出力状態））</td><td>音節モデル
（音節数：124（無音あり)/123（無音なし)，</td></tr>
<tr><td>モノフォン
（16 混合，総状態数：
129（無音あり)/126（無音なし））</td><td>トライフォン
（16 混合，
総状態数：2000）</td><td>PTM（64 混合，
（総状態数：
3000/129）</td><td>母音，促音，撥音，無音は 5 状態（3 出力状態)，
他 7 状態 (5 出力状態),16 混合，
総状態数：600（無音あり)/597（無音なし））</td></tr>
<tr><td colspan="4">16 kHz サンプリング，25 ms ハミング窓，フレーム周期 10 ms，性別依存（男性），対角共分散行列，自己遷移ループ</td></tr>
<tr><td>特徴ベクトル</td><td colspan="4">MFCC（12 次元）＋ ΔMFCC ＋ ΔPOW（計 25 次元）</td></tr>
</table>

表 2　音響モデルの特徴（デコーダ：SPOJUS）
Table 2　Specifications of acoustic models. (Decoder: SPOJUS)

音響モデル	音節モデル（音節数：116（無音あり)/114（無音なし)， 5 状態（4 出力状態）-4 混合（全共分散行列)，6 状態（5 出力状態）-32 混合（対角共分散行列)， 総状態数 464（無音あり)/456（無音なし)，性別依存（男性))，自己遷移ループ/継続時間制御	
	12 kHz サンプリング，21.33 ms ハミング窓， フレーム周期 8 ms, 全共分散行列	16 kHz サンプリング，21.33/25 ms ハミング窓 フレーム周期 8/10 ms, 対角/全共分散行列
特徴ベクトル	MFCC（10 次元 × 4 フレームを KL 展開で 20 次元に圧縮） + ΔCEP + ΔΔCEP + ΔPOW + ΔΔPOW（計 42 次元） （以下，MFCC-seg と略記）	MFCC（12 次元）+ ΔMFCC + ΔΔMFCC + ΔPOW + ΔΔPOW（計 38 次元） （以下，MFCC-frm と略記）
	LPC·MEL·CEP（10 次元）+ ΔLPC·MEL·CEP + ΔΔLPC·MEL·CEP + ΔPOW + ΔΔPOW（計 32 次元） （以下，LPC-frm と略記）	MFCC（12 次元〜4 フレーム を KL 展開で 24 次元に圧縮） + ΔCEP + ΔΔCEP + ΔPOW + ΔΔPOW（計 50 次元） （以下，MFCC-seg と略記）
	LPC·MEL·CEP（10 次元 ×4 フレームを KL 展開で 20 次元に圧縮） + ΔCEP + ΔΔCEP + + ΔPOW + ΔΔPOW（計 42 次元） （以下，LPC-seg と略記）	

表 3　言語モデルの特徴
Table 3　Specifications of language models.

新聞読上げ音声	毎日新聞（45ヵ月分）から作成した tri-gram モデル（語彙数 2 万， パープレキシティ 33.4（句読点あり)/ 43.4（句読点なし)，カバレージ 99.54（句読点あり)/ 99.49（句読点なし）%)
ニュース音声	NHK 汎用ニュース原稿（5 年分）から作成した tri-gram モデル（語彙数 2 万， パープレキシティ 56.5（句読点あり)/ 72.2（句読点なし)，カバレージ 96.91（句読点あり)/ 96.62（句読点なし）%)

無音モデルをもたない音響モデルを擬似的に実現して認識を行った．具体的には，まず，入力音声に対してパワーのしきい値を設定し，あらかじめ，しきい値以下の音声区間を除去した音声に対して認識を行う．そして，認識の際には，無音を表す音素・音節を含む認識結果を出力しないように制約をかける．ただし，音響モデルにおける無音モデルの有無は，言語モデルにおける句読点の有無に対応しているので，パワーがしきい値以下の音声区間を除去して認識を行う場合は，語彙として句読点を含まない言語モデルを用いる(注1)．

2.1.2　SPOJUS で用いた音響モデル

SPOJUS で用いた音響モデルは，音節を基本単位とする HMM モデルで，デコーダと同様に，豊橋技術科学大学工学部情報工学系中川研究室で開発されたモデル [12] を用いた．音響モデルの特徴の概要を表 2 に示す．詳細については文献 [12] を参照されたい．

無音モデルの有無については，原則として，あらゆる認識モデルで，無音モデルをもつものともたないものの両方を用意して評価した．SPOJUS の場合は，無音をもたない音節モデルについても，パワーがしきい値以下の音声区間を除去した音声データを訓練データとして実際にモデル学習を行ってモデルを実現した．また，無音モデルをもたない音節モデルを用いた認識の際には，Julius の場合と同様に，入力音声から無音区間を除去した上で，語彙として句読点を含まない言語モデルを用いて認識を行う．

実際に実験を行った条件の組合せとしては，最も認識率の高い組合せ（サンプリング周波数 16 kHz，フレーム周期 10 ms，全共分散行列，特徴ベクトル MFCC-seg，継続時間制御）を中心として，個々の条件が一部だけ異なるモデルを用意しそれらの評価を行った．組合せの詳細は表 2 及び表 4 に示すとおりである．なお，無音モデルの有無については，予備実験の結果，複数のモデルの出力の共通部分を用いた信頼度の性能に大きな影響をもつと推定されたので，あらゆる音節モデルの設定において，無音モデルをもつものともたないものの両方の音節モデルを用意し評価した(注2)．

2.2　言語モデル

言語モデルの一覧を表 3 に示す．言語モデルとしては，語彙数 2 万の単語 bi-gram 及び単語 tri-gram（Julius では逆向き単語 tri-gram）を用いた．Julius，SPOJUS どちらのデコーダにおいても，2 パス探索により認識を行い，1 パス目では単語 bi-gram を，2 パ

(注1)：入力音声中の促音の無音区間のうち，無声摩擦音/s/の前以外の促音は除去される．Julius における無音モデルをもたない音響モデルは，促音のモデルをもっているが，SPOJUS における無音モデルをもたない音響モデルでは，無声摩擦音/s/の前以外の促音のモデルをもたない．一方，言語モデル用の単語辞書においては，Julius，SPOJUS とも，促音は除去されていない．

(注2)：特徴ベクトルとして LPC-frm あるいは LPC-seg を用いたものについては，MFCC-frm あるいは MFCC-seg を用いたものと比較して認識率が低いため，ここでは，あくまで参考データとして，無音モデルをもたないモデルによる結果だけを示すにとどめている．

表 4　単一の大語彙連続音声認識モデルの単語認識率 (%)
Table 4　Word recognition rates of individual LVCSR models (%).

デコーダ	無音モデル・句読点	音響モデル（継続時間制御/自己遷移ループ，サンプリング周波数，フレーム周期，対角/全共分散行列，特徴ベクトル）					新聞読上げ音声		ニュース音声	
							Cor	Acc	Cor	Acc
Julius	有	トライフォン（1 パス目）					85.1	82.2	59.0	50.2
		トライフォン（2 パス目）					**93.9**	**91.3**	66.5	57.5
									（言語モデル：新聞 45ヵ月分）	
									(62.3)	(56.1)
		PTM					92.7	91.3	62.3	58.2
		モノフォン					83.0	80.6	56.5	48.6
		音節モデル					91.9	90.4	**72.4**	**69.2**
	無	トライフォン（2 パス目）					84.3	79.4	62.6	51.0
		PTM					87.0	82.8	56.9	50.5
		モノフォン					73.8	70.3	50.4	40.8
		音節モデル					86.0	83.3	68.9	63.6
SPOJUS (音節モデル)	有	継続時間制御	16 kHz	10 ms	全	MFCC-seg	**91.1**	**84.3**	**71.5**	**63.9**
						MFCC-frm	90.6	86.2	70.6	62.7
					対角	MFCC-seg	83.9	55.3	61.4	57.0
				8 ms	全		87.5	70.2	71.8	56.1
			12 kHz				80.7	64.3	66.2	53.3
		自己遷移ループ	16 kHz	10 ms	全	MFCC-seg	87.7	86.0	65.8	60.7
						MFCC-frm	86.0	81.0	62.9	58.5
						MFCC-seg	87.0	81.9	68.5	62.8
				8 ms			79.5	73.2	60.6	55.8
	無	継続時間制御	16 kHz	10 ms			88.3	84.1	67.9	45.7
						MFCC-frm	89.0	84.8	66.2	39.9
					対角	MFCC-seg	82.3	77.9	58.2	43.0
				8 ms			88.0	82.9	68.3	38.9
			12 kHz				90.4	84.9	70.3	58.4
					全	LPC-seg	86.1	82.9	62.7	53.8
						LPC-frm	80.1	77.3	55.6	49.2
		自己遷移ループ	16 kHz	10 ms		MFCC-seg	85.6	82.7	63.4	56.9
						MFCC-frm	85.4	81.9	61.3	50.7
				8 ms		MFCC-seg	87.1	84.4	67.6	60.0
			12 kHz				88.1	84.9	68.2	60.2

ス目では単語 tri-gram を，それぞれ使用する．言語モデル訓練用のコーパスとしては，以下の 2 種類のものを用いた．

（1） 45 カ月分の毎日新聞記事

（2） 5 年分の NHK ニュース原稿（約 12 万文）．

言語モデルの作成においては，IPA「日本語ディクテーション基本ソフトウェアの開発」プロジェクト [6] から提供されている言語モデル作成ツールを利用した．なお，いずれのモデルにおいても，句読点を含んだものと含まないものの両方を学習し，認識の際には，音響モデルにおける無音モデルの有無と対応させて用いる．

2.3　評価用音声データ

評価用音声データとしては，音声認識が比較的容易な新聞読上げ音声，及び，相対的に音声認識が容易でないニュース音声の 2 種類を用いる．

（1） IPA「日本語ディクテーション基本ソフトウェアの開発」プロジェクト [6] において，新聞記事読上げ音声データベース (JNAS) [5] から選定した 100 文（男性話者 10 人，1,565 語）．

（2） NHK のニュース「ニュース 7」と「おはよう日本」（1996 年 6 月 1 日）の 175 文（男性話者 10 人 — アナウンサー 2 人，レポーター 8 人，6,813 単語）．

いずれの評価データも，音響モデル及び言語モデルの学習には用いていない．

2.4　単語認識率

新聞読上げ音声，及び，ニュース音声をそれぞれ評価用音声データとした場合の，単一の大語彙連続音声認識モデルの単語正解率 (word correct rate, “Cor”)，単語認識精度 (word accuracy, “Acc”) を表 4 に示す(注3)．ここで，言語モデルとしては，新聞読上げ音声の認識の場合は，新聞記事から作成したモデルを，ニュース音声の認識の場合は，ニュース原稿から作成し

(注3)：正解文の単語数を N，認識結果における正解単語数を C，置換誤り単語数を S，挿入誤り単語数を I，脱落誤り単語数を D とすると，単語正解率は $C/N = (N - S - D)/N$，単語認識精度は $(N - S - D - I)/N$ として定義される．

たモデルを，それぞれ用いた．言語モデルは，Julius, SPOJUS のデコーダ間では共通のものを用いている．認識時の音響/言語スコアの重み，挿入ペナルティの設定については，単語認識精度が最大となった結果を採用した．なお，デコーダとして Julius を用い，音響モデルとして無音モデルをもつトライフォンモデルを用いた場合には，1 パス目及び 2 パス目の両方の単語認識率を示すが，特に，ニュース音声の認識については，言語モデルとして，毎日新聞記事 45 カ月分から作成したモデルを用いた場合の 2 パス目の単語認識率も括弧内に示す．表中では，デコーダとして Julius を用いた場合，及び，SPOJUS を用いた場合について，単語正解率・単語認識精度の両方から判断して最も高いと考えられる単語認識率を太字で示す(注4)．

3. 信頼度の評価尺度

本章では，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．

まず，n 個の大語彙連続音声認識モデルの出力を，それぞれ $Hyp_1, \ldots, Hyp_n$ とする．ただし，各出力 Hyp_i は，認識結果の単語の列で表現される．次に，DP マッチングにより，n 個の認識結果の単語の列 $Hyp_1, \ldots, Hyp_n$ の対応付けを行い，n 個の認識結果すべてに含まれる単語を集め，これを「一致単語リスト」と呼ぶ．

例えば，$n = 2$ の場合，二つのモデルの出力 Hyp_1 及び Hyp_2 が以下のように表現されるとする．

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

このとき，一致単語リストは，同一の単語 w_{1i} と w_{2j} $(w_{1i} = w_{2j})$ のうち，DP マッチングによって対応づけられたものを集めることにより構成される．

(注4)：最大の単語正解率付近において，有意水準 5%での有意な認識率の差は，新聞読上げ音声では 1.8%程度，ニュース音声では 1.6%程度であった．

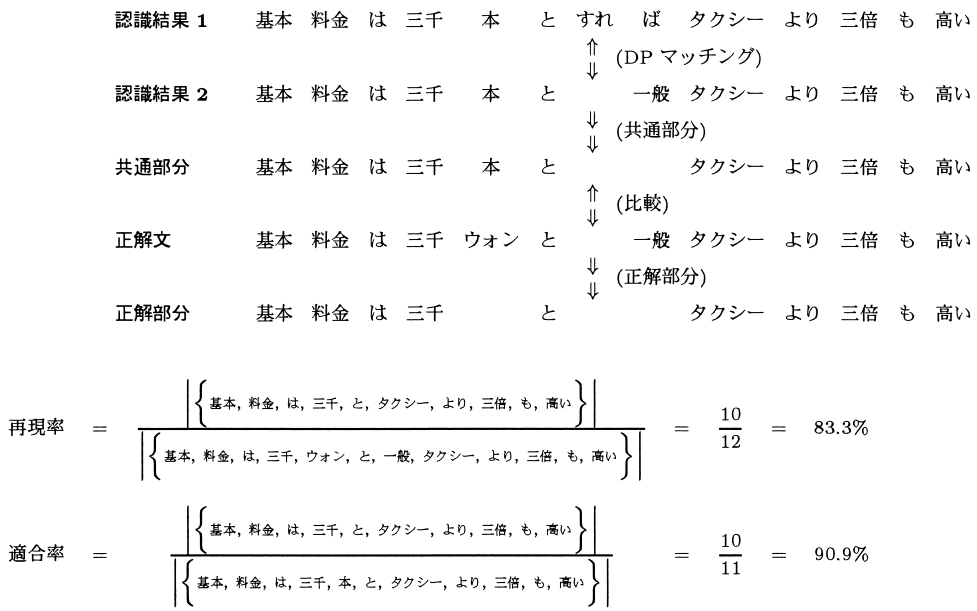


図 1 二つのモデルの認識結果の共通部分の再現率・適合率の例
Fig. 1 An example of recall/precision of agreement between hypotheses by two models.

そして，一致単語リストと正解文を比較して，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．

$$\text{再現率}=\frac{\text{一致単語リスト中の正解単語数}}{\text{正解文の単語数}}$$

$$\text{適合率}=\frac{\text{一致単語リスト中の正解単語数}}{\text{一致単語リストの単語数}}$$

図 1 に，二つのモデルの認識結果の共通部分の再現率及び適合率を計算する例を示す．図中では，認識結果同士の比較，あるいは，共通部分と正解文の比較において共通の単語であると判定された部分を四角で囲んでいる．

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．

4. 音響モデルの差異の評価

本章及び次章では，二つの大語彙連続音声認識モデルの出力の共通部分について，その信頼度を測定した結果を示し，高い信頼度に貢献する要因を分析する．まず，本章では，デコーダが異なる場合も含めて，**2.1** で説明した音響モデルのあらゆる可能な二つの組合せについて，その出力の間の共通部分の再現率・適合率を評価した結果について述べる．

4.1 デコーダの組合せの評価

まず，**2.** の表 4 に挙げたほぼすべてのモデルの認識結果の組合せについて，その出力の間の共通部分の再現率・適合率を評価した．その組合せの数は，表 4 中のトライフォン（1 パス目），トライフォン（ニュース音声の認識において新聞用言語モデルを用いた場合），LPC-seg，LPC-frm を除く 26 種類の認識結果のすべての組合せ（325 通り），及び，LPC-seg，LPC-frm を含む数種類の認識結果の組合せである．そして，共通部分の信頼度の性能を概観するために，再現率に下限（新聞読上げ音声：80%，ニュース音声：50%）(注5) を設け，再現率がこの下限値を上回る組合せを適合率順に並べた結果をプロットしたものを図 2（(a) 新聞読上げ音声，(b) ニュース音声）に示す．図 2 では，横軸に適合率の順位を，縦軸に適合率をとり，デコーダの組合せと信頼度の性能の相関を調べるために，デコーダの組合せ (Julius-SPOJUS, Julius-Julius, SPOJUS-SPOJUS) ごとのプロットを示す．また，表 5 には，

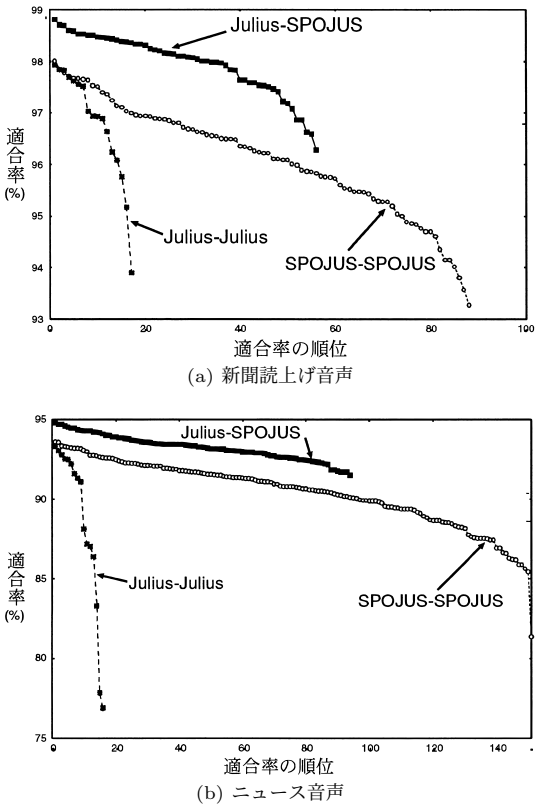


図 2 二つのシステムの出力の共通部分の適合率の分布（デコーダの組合せごと）
Fig. 2 Distribution of precision of agreement between two systems. (For each pair of decoders)

デコーダの組合せごとに，上段に適合率最大となるモデルの組合せを，下段に単語正解率が最大のモデルの組合せ（同一デコーダの組合せの場合は，単語正解率が 1 位と 2 位の組合せ）を，それぞれ示す．表中では，二つのモデル間で異なる特徴を下線で示す．

図 2 からわかるように，適合率最大のモデルの組合せの性能から判断すると，新聞読上げ音声及びニュース音声のいずれにおいても，Julius-SPOJUS という

(注5)：再現率の下限値は，各単独モデルの認識率の分布などを考慮して，一応の目安として決めた値である．再現率がこの下限値を上回る組合せの割合は，新聞読上げ音声については約半数，ニュース音声については約 8 割であった．ニュース音声については，各単独モデルの認識率の分散が大きいため，再現率の下限値もやや低めに設定した．なお，**4.** における評価結果は，すべて，この下限値を上回る結果だけを対象としている．また，最大の適合率付近において，有意水準 5%での有意な適合率の差は，新聞読上げ音声では 0.9%程度，ニュース音声では 0.8%程度であった．

表 5　適合率最大のモデルの組合せと単語正解率最大のモデルの組合せ（デコーダの組合せごと）

Table 5　Pairs of models with highest precision/highest word correct rates. (for each pair of decoders)

	デコーダの組合せ	モデルの組合せ（上段：適合率最大の組，下段：単語正解率最大の組）	単語正解率	再現率	適合率
新聞読み上げ音声	Julius SPOJUS	J: 無音モデルあり，トライフォン S: 無音モデルあり，MFCC-seg，16 kHz，8 ms，全，継続	**93.9** 87.5	**84.1**	**98.8**
		J: 無音モデルあり，トライフォン S: 無音モデルあり，MFCC-seg，16 kHz，10 ms，全，継続	**93.9** **91.1**	87.1	98.7
	Julius Julius	J: 無音モデルあり，トライフォン J: 無音モデルなし，PTM	**93.9** 87.0	84.9	97.9
		J: 無音モデルあり，トライフォン J: 無音モデルあり，PTM	**93.9** 92.7	90.2	96.9
	SPOJUS SPOJUS	S: 無音モデルあり，MFCC-frm，16 kHz，10 ms，全，継続 S: 無音モデルなし，MFCC-frm，16 kHz，10 ms，全，継続	90.6 89.0	84.6	98.0
		S: 無音モデルあり，MFCC-seg，16 kHz，10 ms，全，継続 S: 無音モデルあり，MFCC-frm，16 kHz，10 ms，全，継続	**91.1** 90.6	87.6	94.9
ニュース音声	Julius SPOJUS	J: 無音モデルあり，PTM S: 無音モデルあり，MFCC-seg，16 kHz，8 ms，全，継続	62.3 **71.8**	**55.1**	**94.8**
		J: 無音モデルあり，音節モデル S: 無音モデルあり，MFCC-seg，16 kHz，8 ms，全，継続	**72.4** **71.8**	63.8	94.5
	Julius Julius	J: 無音モデルあり，音節モデル J: 無音モデルなし，トライフォン	**72.4** 62.6	56.5	93.3
		J: 無音モデルあり，音節モデル J: 無音モデルなし，音節モデル	**72.4** 68.9	64.8	87.2
	SPOJUS SPOJUS	S: 無音モデルなし，MFCC-seg，16 kHz，10 ms，全，自己 S: 無音モデルあり，MFCC-seg，12 kHz，8 ms，全，継続	63.4 66.2	53.7	93.6
		S: 無音モデルあり，MFCC-seg，16 kHz，8 ms，全，継続 S: 無音モデルあり，MFCC-seg，16 kHz，10 ms，全，継続	**71.8** 71.5	64.8	89.5

太字: 単語正解率最大（デコーダごと），適合率最大

異なるデコーダの組合せの場合に，高い信頼度が達成できている(注6)．また，表 5 に示すように，最も高い適合率を達成した組合せでは，新聞読上げ音声で 99%程度，ニュース音声で 95%程度の高い適合率を達成しており，信頼度尺度としては非常に有用なものであるといえる．更に，表 5 で，Julius-SPOJUS という異なるデコーダの組合せの場合に，単語正解率が最大のモデルの組合せの再現率・適合率を評価した結果では，最高の適合率をわずかに下回るものの，新聞読上げ音声で約 87%，ニュース音声で約 64%という高い再現率を達成している．この結果を，単一のモデルの単語正解率（再現率と同等の尺度）と比較すると，新聞読上げ音声では，最大単語正解率（94%程度）から 7%程度下回る再現率となっているものの，99%近い適合率を達成していることがわかる．また，ニュース音声でも，最大単語正解率（72%程度）から 9%弱程度下回る再現率となっているものの，95%近い適合率を達成している．一方，表 5 中で，Julius-Julius，SPOJUS-SPOJUS といった同一のデコーダの組合せにおいて最も高い適合率を達成した組合せでは，新聞読上げ音声で 98%程度，ニュース音声で 93〜94%程度という適合率であった．異なるデコーダの組合せの最高の適合率からは約 1%程度劣るものの，かなり高い信頼度であるといえる．しかし，同一のデコーダの組合せにおいて，単語正解率が 1 位と 2 位のモデルの組合せでは，ほとんどの場合，適合率がかなり低くなっており，どのような特徴をもった二つのモデルを組み合わせるかによって，適合率が大きく左右されることがわかる．

4.2　デコーダが同一の場合の評価

次に，本節では，音響モデルの特徴における個々の差異が，二つのモデルの出力の共通部分の適合率に与える影響を調べるために，同一のデコーダの組合せの

(注6)：同様の傾向は，ATR 旅行会話音声を対象として，本論文で用いた Julius 及び SPOJUS のほかに，第 3 のデコーダとして ATR-SPREC を追加して行った評価 [18] においても確認している．ただし，実際に，デコーダの特性の様々な相違の網羅的な評価を行うためには，音響モデル・言語モデル等の他の条件を同じにした上で，デコーダの様々な特性を少しずつ調整して，信頼度へ与える影響との相関を評価することが不可欠である．一方，本論文の実験では，デコーダが異なると音響モデルにも何らかの差異が生じるため，現段階では，デコーダのみの差異の評価はできていない．そこで，本節では，今回用いた二つのデコーダに関する実験的事実を述べるにとどめる．

場合について，音響モデルの特徴における個々の差異と適合率の相関について分析する．

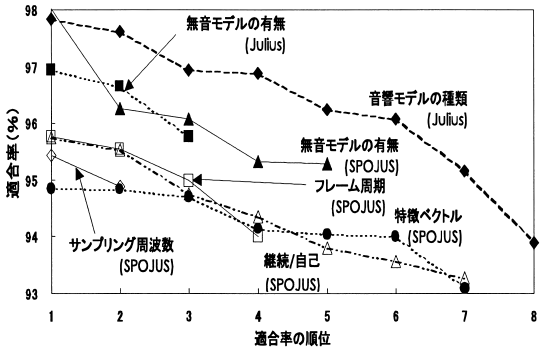
4.2.1 単一の特徴だけが異なる場合

まず，音響モデルの個々の特徴の差異がそれぞれ単独で適合率に与える影響を分析するために，音響モデルの差異がただ一つの特徴だけであるモデル同士で，出力の共通部分の再現率・適合率を評価した．音響モデルの特徴ごとに，値の大きい順に適合率をプロットした結果を図 3（(a) 新聞読上げ音声，(b) ニュース音声）に示す．分析の対象とした音響モデルの特徴は以下のとおりである．

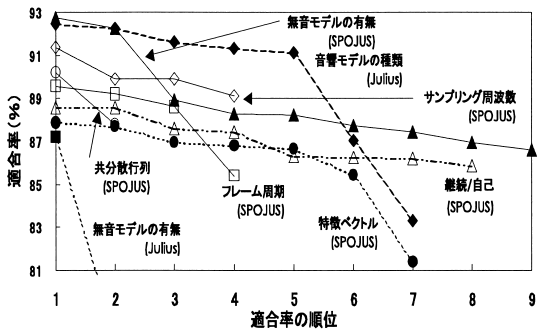
（1） 無音モデルの有無（デコーダの組合せ：Julius-Julius，SPOJUS-SPOJUS）

（2） 音響モデルの種類（デコーダの組合せ：Julius-Julius，モノフォン/トライフォン/PTM/音節モデル）

（3） 特徴ベクトル（デコーダの組合せ：SPOJUS-SPOJUS，MFCC-seg / MFCC-frm / LPC-seg / LPC-frm）



(a) 新聞読上げ音声



(b) ニュース音声

図 3 二つのシステムの出力の共通部分の適合率の分布（モデル間の差異が単一の特徴の場合，単語認識）

Fig. 3 Distribution of precision of agreement between two systems. (Difference in a single feature, word recognition)

（4） サンプリング周波数（デコーダの組合せ：SPOJUS-SPOJUS，16 kHz/ kHz）

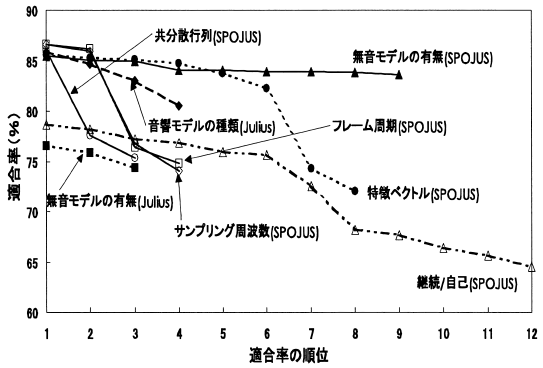
（5） フレーム周期（デコーダの組合せ：SPOJUS-SPOJUS，10 ms/8 ms）

（6） 共分散行列（デコーダの組合せ：SPOJUS-SPOJUS，全共分散行列/対角共分散行列．図 3 (a) では，サンプル数が一つであったため，スペースの都合上，プロットを省略している．）

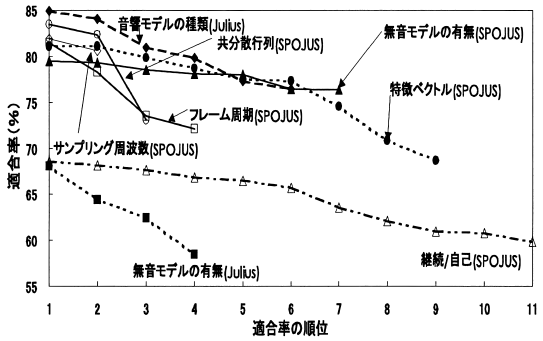
（7） 自己遷移ループをもつか継続時間制御か（デコーダの組合せ：SPOJUS-SPOJUS）

各特徴ごとの最大の適合率を比較すると，新聞読上げ音声，ニュース音声とも，SPOJUS における無音モデルの有無の組合せの性能が最も高く，Julius における異種の音響モデルの組合せもほぼ同等の性能を示している．新聞読上げ音声においては，Julius における無音モデルの有無の組合せがこれらに続いているが，ニュース音声における性能はあまりよくない．

ここで，無音モデルの有無における違いを，他の音響モデルの特徴の差異と比較する際に留意すべき点として，無音モデルの有無に差異がある場合には，言語モデルにおいても句読点の有無に違いがある点が挙げられる．他の音響モデルの特徴の差異は厳密に音響モデルのみの違いであることを考えると，この点は非常に重要である．そこで，言語モデルにおける差異の効果を排除して，音響モデルにおける差異の効果のみを評価するために，本論文の実験において対象とした音響モデルの組合せについて，言語モデルを用いずに音節認識実験を行った結果について，モデル組の出力の共通部分の信頼度（音節単位での再現率・適合率）の実験的評価を行った．特に，図 3 の場合と同様に，音響モデルの個々の特徴の差異がそれぞれ単独で適合率に与える影響を分析するために，音響モデルの差異がただ一つの特徴だけであるモデル同士で，音節出力の共通部分の再現率・適合率を評価した．音響モデルの特徴ごとに，値の大きい順に適合率をプロットした結果を図 4（(a) 新聞読上げ音声，(b) ニュース音声）に示す．この結果における「無音モデルの有無 (Julius)」及び「無音モデルの有無 (SPOJUS)」を，それぞれ，図 3 (単語認識）における「無音モデルの有無 (Julius)」及び「無音モデルの有無 (SPOJUS)」と比較するとわかるように，(a) 新聞読み上げ音声，及び，(b) ニュース音声のいずれにおいても，音節認識においては，無音モデルの有無の違いの効果は，単語認識における効



(a) 新聞読上げ音声



(b) ニュース音声

図 4　二つのシステムの出力の共通部分の適合率の分布
（モデル間の差異が単一の特徴の場合，音節認識）
Fig. 4　Distribution of precision of agreement between two systems. (Difference in a single feature, syllable recognition)

果ほど目立ったものではない．このことから，単語認識における無音モデルの有無においては，言語モデルにおける句読点の有無が重要な役割を担っているといえる．

4.2.2　無音モデルの有無の評価

前節の（単語認識に関する）分析から，デコーダが同一の場合には，無音モデルの有無（デコーダ：SPOJUS）及び音響モデルの種類の違い（デコーダ：Julius）が高い適合率の重要な要因になっていることがわかった．そこで，本項では，これらの特徴の差異の有無と，二つのモデルの出力の共通部分の適合率の間の相関について分析を行う．具体的には，デコーダの組合せがJulius-Julius の場合，及び，SPOJUS-SPOJUS の場合のそれぞれについて，i) 無音モデルをもつ/もたないモデルの組合せ，ii) 無音モデルをもつモデル同士の組合せ，iii) 無音モデルをもたないモデル同士の組合せ，の 3 通りについて，適合率の分布をプロットして比較した．この結果を図 5 に示す．なお，このうち，デコーダの組合せが Julius-Julius の場合については，ii) 及び iii) におけるモデル間の差異は，音響モデルの種類のみの違いとなっている．また，i) におけるモデル間の差異は，無音モデルの有無のみの違い，あるいは，無音モデルの有無と音響モデルの種類の違いの両方となっている．図 5 の結果から，最高の適合率については，Julius-Julius，SPOJUS-SPOJUS のいずれのデコーダの組合せにおいても，モデル間の差異として無音モデルの有無が異なる方がよいことがわかる．

4.2.3　ま と め

これまでの分析を総合して，デコーダが同一の場合について，音響モデルの各特徴の差異が，二つのモデルの出力の共通部分の高い適合率に寄与する度合をまとめる．

まず，**4.2.1** の分析から，無音モデルの有無（デコーダ：SPOJUS）及び音響モデルの種類の違い（デコーダ：Julius）が高い適合率の重要な要因になっていることがわかった．また，モデル間の差異が単一の特徴の場合と複数の特徴の場合を比較すると，複数の特徴で差異があるモデルの組合せの方が高い適合率が達成できる傾向にあることがわかっている [17]．そこで，ここでは，音響モデルの各特徴の差異の組合せを以下のように分類する．

まず，デコーダが Julius の場合については，以下の 3 通りに分類する．

（1）　音響モデルの種類，無音モデルの有無の両方が異なるモデルの組合せ

（2）　音響モデルの種類のみが異なるモデルの組合せ

（3）　無音モデルの有無のみが異なるモデルの組合せ

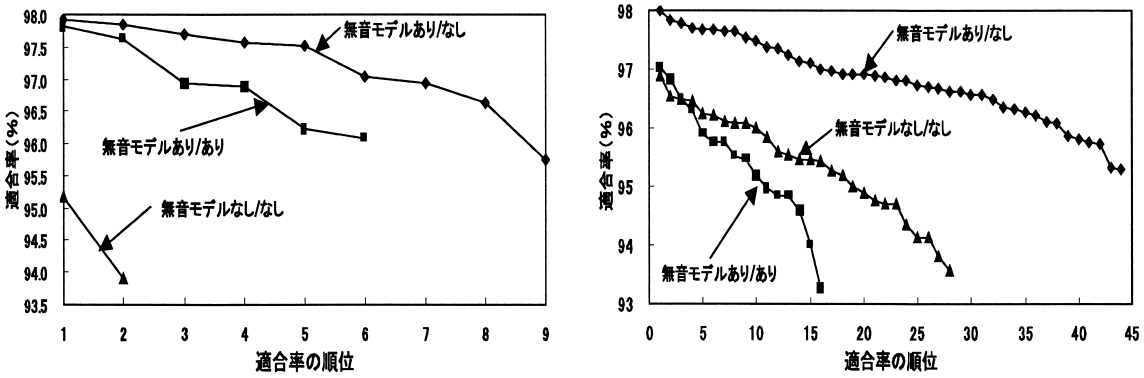
これらの各分類について，適合率が最大となるモデルの組合せの適合率を図 6 左側に示す．この結果について，特にニュース音声における性能の差を重視すると，高い適合率への寄与の度合は以下の不等式で表現できる．

$$(1) > (2) \gg (3)$$

（ただし，(1) と (2) との差は，有意水準 5%では有意ではない）また，デコーダが SPOJUS の場合については，以下の 5 通りに分類する．

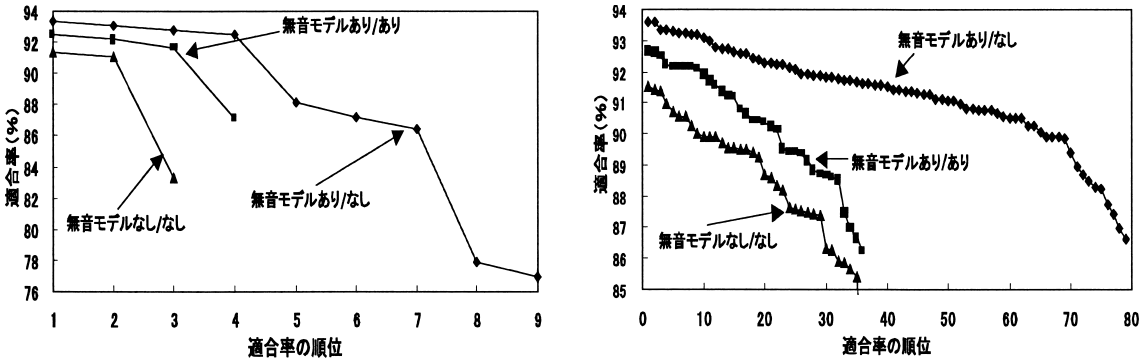
（4）　無音モデルの有無を含む複数の特徴に差異が

(1) 新聞読上げ音声



(a) デコーダの組合せ：Julius-Julius　　(b) デコーダの組合せ：SPOJUS-SPOJUS

(2) ニュース音声



(a) デコーダの組合せ：Julius-Julius　　(b) デコーダの組合せ：SPOJUS-SPOJUS

図 5　二つのシステムの出力の共通部分の適合率の分布（無音モデル { あり/なし，あり/あり，なし/なし } による分類）

Fig. 5　Distribution of precision of agreement between two systems. (Classified by {with/without, with/with, without/without} Short Pause Models)

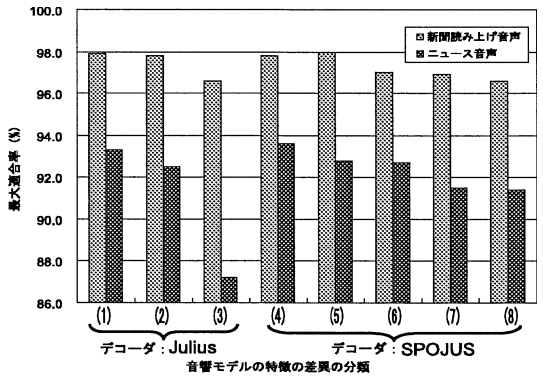


図 6　音響モデルの差異ごとの最大適合率の比較（単一デコーダの場合）

Fig. 6　Evaluation results of agreement between two acoustic models: summary.

あるモデルの組合せ

（5） 無音モデルの有無のみが異なるモデルの組合せ

（6） 無音モデルをもつモデル同士で，複数の特徴に差異があるモデルの組合せ

（7） 無音モデルをもたないモデル同士で，複数の特徴に差異があるモデルの組合せ

（8） 単一の特徴だけが異なるモデルの組合せのうち上記以外のもの

これらの各分類について，適合率が最大となるモデルの組合せの適合率を図 6 右側に示す．この結果についても，特にニュース音声における性能の差を重視すると，高い適合率への寄与の度合は以下の不等式で表現できる．

$$(4) > (5), (6) > (7) > (8) \quad (1)$$

(ただし，分類 (7) 及び (8) の優劣については，分類 (8) において，新聞読上げ音声とニュース音声の両方において分類 (7) と同等の適合率を達成したモデル組が存在しないため “(7) > (8)” と判定した．また，ニュース音声においては，(4) と (5) との差，及び，(6) と (7) との差は，それぞれ，有意水準 5%で有意である.)

5. 音響モデル以外の差異の評価

本章では，デコーダが同一の場合に，言語モデル，デコーダの設定など，音響モデル以外で複数のモデル間に差異がある場合に，それらの複数のモデルの出力の共通部分の再現率・適合率を評価した結果について述べる.

5.1 言語モデル

ニュース音声の認識の場合について，毎日新聞記事を用いて訓練された言語モデル，及び，ニュース原稿を用いて訓練された言語モデルの二つを用いて，2 通りの認識結果を作成し，その共通部分について，再現率・適合率を評価した結果を表 6 の「二つの言語モデル」の欄に示す．この場合の適合率は，音響モデルの特徴に差異がある場合の最高の適合率（表 5，デコーダが同一の場合の，適合率最大のモデルの組合せ）には及ばなかった.

5.2 デコーダの設定

デコーダの各種設定として，1 パス目と 2 パス目，音響/言語スコアの重みの複数の設定，挿入ペナルティの複数の設定，について，それぞれ，それらの出力の共通部分の信頼度を評価した．まず，音響モデル・デコーダとして，トライフォンモデル・Julius を用いた場合について，デコーダの 1 パス目及び 2 パス目の出力の間の共通部分の再現率・適合率を評価した結果を，表 6 の「1 パス&2 パス」の欄に示す.

また，信頼度尺度に関する先行研究において一定の性能が確認されている音響安定度を用いた信頼度 [7] との比較のために，音響/言語スコアの重みについて，最も高い単語認識率を示す値を中心とする 10 通りの設定について，その出力の間のあらゆる可能な部分集合の共通部分の再現率・適合率を評価した(注7)．音響モデル・デコーダとしては，トライフォンモデル・Julius を用いた場合と，音節モデル（12 kHz サンプリング，MFCC-seg）・SPOJUS を用いた場合の 2 通りを評価した．この結果のうち，再現率が最も高い場合及び適合率が最も高い場合の結果を，表 6 の「音響/言語スコアの重み」の欄に示す.

更に，挿入ペナルティについても，最も高い単語認識率を示す値を中心とする 10 通りの設定について，その出力の間のあらゆる可能な部分集合の共通部分の再現率・適合率を評価した(注8)．音響モデル・デコーダとしては，トライフォンモデル・Julius を用いた．この結果のうち，再現率が最も高い場合及び適合率が最も高い場合の結果を，表 6 の「挿入ペナルティ」の欄に示す.

これらのすべての結果における適合率は，音響モデルの特徴に差異がある場合の最高の適合率（表 5，デコーダが同一の場合の，適合率最大のモデルの組合せ）には及ばなかった.

(注7)：これらの重みを用いた場合の最も低い単語認識率は，新聞読上げ音声の場合，トライフォン・Julius で，86.9% (Cor) / 79.3% (Acc)，音節モデル・SPOJUS で，80.1% (Cor) / 77.3% (Acc)，ニュース音声の場合，トライフォン・Julius で，61.9% (Cor) / 45.9% (Acc)，音節モデル・SPOJUS で，52.3% (Cor) / 46.3% (Acc) である.
(注8)：これらの挿入ペナルティを用いた場合の最も低い単語正解率 (Cor) 及び単語認識精度 (Acc) は，新聞読上げ音声の場合，89.3% (Cor) / 84.7% (Acc)，ニュース音声の場合，60.5% (Cor) / 45.3% (Acc) である.

表 6 音響モデル以外が異なる複数のモデルの出力の共通部分の再現率/適合率 (%)
Table 6 Recall/precision (%) of agreement among multiple models. (Differences in features other than acoustic models)

モデルの組合せ		新聞読上げ音声	ニュース音声
二つの言語モデル（トライフォン，毎日新聞記事/ニュース原稿）		—	53.7 / 86.5
1 パス & 2 パス（トライフォン）		81.0 / 92.7	57.9 / 66.1
音響/言語スコアの重み	トライフォン	86.9 / 93.2 ～ 82.4 / 93.9	50.7 / 87.6 ～ 44.4 / 91.6
	音節モデル (MFCC-seg)	82.7 / 92.2 ～ 76.2 / 94.5	58.5 / 81.1 ～ 44.7 / 90.1
挿入ペナルティ	トライフォン	89.7 / 91.4 ～ 87.7 / 92.7	56.9 / 75.7 ～ 50.8 / 82.9
N-best 候補	トライフォン	82.4 / 93.1 ～ 66.0 / 94.5	64.0 / 82.0
	音節モデル (LPC-seg)	80.4 / 90.6 ～ 39.9 / 98.3	42.4 / 74.4

トライフォン：Julius，無音モデル・句読点あり，16 kHz，10 ms，対角共分散行列，自己遷移ループ
音節モデル： SPOJUS，無音モデル・句読点なし，12 kHz，8 ms，全共分散行列，継続時間制御

5.3 N-best 候補の間の共通部分

信頼度尺度に関する先行研究において一定の性能が確認されているものとして，単語グラフ中のエッジ接続数を用いた信頼度 [14] や仮説密度を用いた信頼度 [7] などが挙げられる．これらの従来の信頼度尺度の考え方を参考にして，単一モデルの出力の N-best (200best) 候補の間の共通部分を求め，その再現率・適合率を評価した．音響モデル・デコーダとしては，トライフォンモデル・Julius を用いた場合と，音節モデル（12 kHz サンプリング，LPC-seg）・SPOJUS を用いた場合の 2 通りを評価した．単一モデルの出力の N-best (200best) 候補について，そのうちの任意の i 個 ($2 \leq i \leq 200$) の共通部分の再現率・適合率を評価した結果を，表 6 の「N-best 候補」の欄に示す(注9)．この場合の適合率（新聞読上げ音声の場合は，再現率が 80%程度での適合率）は，音響モデルの特徴に差異がある場合の最高の適合率（表 5，デコーダが同一の場合の，適合率最大のモデルの組合せ）には及ばなかった．

6. むすび

本論文では，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，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．実験の結果，デコーダ及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，新聞読上げ音声では正解単語の約 87%を 99%近くの精度で予測でき，また，ニュース音声では正解単語の約 64%を 95%近くの精度で予測できるという，非常に高い性能が達成された．また，同一のデコーダを用いた場合にも，音響モデルの特徴の違いと信頼度との相関を網羅的に評価することにより，デコーダが異なる場合の性能をやや下回るものの，ほぼそれに匹敵する性能を達成した．また，特に，混合連続分布 HMM に基づく音響モデルの多種多様な特徴が，高い信頼度に寄与する度合を評価した結果，無音モデルの有無，音響モデルの種類（トライフォンや音節モデルなど）の違いといった特徴が重要であることがわかった．更に，無音モデルの有無においては，言語モデルにおける句読点の有無が重要な役割を担っていることがわかった．

今後は，大語彙連続音声認識モデルの振舞いを左右する要因のうち，デコーダ及び言語モデルについても，様々な特徴の組合せを網羅的に評価し，複数モデルの出力の共通部分を用いた信頼度における有効性について分析を行う予定である．例えば，言語モデルについては，統語構造を利用するモデル（例えば，文献 [2], [8] など）や話題の情報を利用するモデル（例えば，文献 [4], [8] など）のように，tri-gram モデルとは異なる仮説を優先すると思われるモデルも提案されているので，これらについて分析を行うことが有望であると思われる．

なお，今回の評価実験の結果をより一般化して議論し，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の有効性を理論的観点から検証するためには，認識結果における正解単語・誤り単語の分布を複数システム間で比較し，その傾向と信頼度との相関を詳細に分析する必要がある．更に，認識結果における正解単語・誤り単語の分布の差異と信頼度との相関のモデル化を行い，そのモデル化に基づいて，任意のシステムの組に対して，その出力の共通部分の信頼度を的確に予測できる必要がある．例えば，今回の評価実験では，最大で 28 通りもの認識モデルを構築し評価実験を行ったが，評価実験を網羅的に行っただけであるので，新聞読上げ音声・ニュース音声以外の音声データに対して，今回の評価実験の結果がそのまま当てはまるという保証はない．新たな種類の音声データや，新たな種類のデコーダ・音響モデル・言語モデルに対して，複数の音声認識システムの出力の共通部分を用いる本論文の手法を有効に活用するためには，どのような認識モデルをどの程度の種類用意すれば十分であるかについての指針を与える必要がある．

例えば，最も単純なモデル化として，二つのシステムの正解部分・誤り部分が完全に独立に分布すると仮定する．このとき，単独システムの単語正解率が 90%であるとすると，2 システムがともに正解する確率は $0.9 \times 0.9 = 0.81$，ともに誤る確率は $0.1 \times 0.1 = 0.01$ となり，再現率は 81%，適合率は $0.81/(0.81 + 0.01) = 98.8\%$ と予測される．一見すると，この予測は実験結果と合っているように見えるが，単独システムの単語正解率が 70%の場合には，2 システムがともに正解する確率が 0.49，ともに誤る確率が 0.09 より，再現率の予測値は 49%，適合率の予測値は $0.49/(0.49 + 0.09) = 84.5\%$ となり，実験結果からは大きく外れる．したがって，二つのシステムの正

(注9)：新聞読上げ音声の場合については，適合率が最も高い場合，及び，再現率が 80%程度の場合の結果，ニュース音声の場合については，適合率が最も高い場合の結果を示す．

解部分・誤り部分は完全に独立に分布しているわけではなく，何らかの相関のもとで分布していると推測され，その分布の適切なモデル化が必要であるといえる．

また，本論文の高信頼度部分推定手法と，従来の，単一の認識エンジン・認識モデルのみを用いた信頼度尺度（例えば，[7], [13], [14], [19]）との比較においても，本論文の高信頼度部分推定手法に対して上述したようなモデル化を行い，そのモデル化と従来手法のモデル化を比較して，原理的にどのような性能の違いがあり得るのかの分析ができることが望ましい(注10)．そのような分析の結果を踏まえれば，本論文で用いた二つのモデルの出力の共通部分の情報と，従来の，単一の認識エンジン・認識モデルのみを用いた信頼度尺度において用いられている情報を併用することにより，より高性能な信頼度尺度を実現することも可能であると考えられる(注11)．

また，上記のことに関連して，複数システムの間で高信頼度の認識結果を相補的に組み合わせることにより，文全体の認識率を改善するという問題 [10], [16] においても，認識結果における正解単語・誤り単語の分布の差異と信頼度との相関のモデル化に基づいて，どのような認識モデルをどの程度の種類用意すれば十分であるかについての指針を与えられることが望ましい．

複数の音声認識システムの出力の共通部分を用いる本論文の手法の汎用性を高め，様々な状況においてその手法を有効に活用するためには，これらの理論的考察が不可欠である．しかし，紙数の都合上，本論文では，新聞読上げ音声及びニュース音声を対象として，2 種類のデコーダを用いた場合について，二つのモデルの出力の共通部分という非常に単純な指標だけでどの程度の性能が達成できるのかという点に焦点を当てて評価を行い，その結果についての分析と考察を行った．上記の理論的考察については別の機会に述べる．

謝辞　本研究に協力して頂いた豊橋技術科学大学工学部情報工学系中川研究室の関係者に深く感謝する．また，ニュース音声データベース，ニューステキストデータベースを提供して頂いた NHK 放送技術研究所の関係諸氏に感謝する．

(注10)：あくまで一事例にすぎないが，共通の評価データのもとでの文献 [13] の信頼度尺度との性能比較においては，本論文の手法の方が十分に高い性能を示している．

(注11)：手法の一例として，それらの様々な情報を用いて認識結果の正誤を判別するための規則を機械学習などの枠組みで学習するという方法が有効であると考えられる．実際，複数のモデルの出力を混合することにより単語認識率を改善するという問題において，そのような機械学習の枠組み (SVM) を適用した結果 [10] においては，単に単語認識率を改善するだけでなく，副産物として，信頼度尺度としても，本論文で示した性能を上回ることを確認している．

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，” 情処学会第 57 回全大，pp.35–36, 1998.

[2] C. Chelba and F. Jelinek, “Structured language modeling,” Comput. Speech Lang., vol.14, no.4, pp.283–332, 2000.

[3] J.G. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.347–354, 1997.

[4] R. Florian and D. Yarowsky, “Dynamic non-local language modeling via hierarchical topic-based adaptation,” Proc. 37th Annual Meeting of ACL, pp.167–174, 1999.

[5] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, “JNAS Japanese speech corpus for large vocabulary continuous speech recognition research,” J. Acoust. Soc. Jpn. (E), vol.20, no.3, pp.190–206, 1999.

[6] 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），” 音響誌（技術報告），vol.57, no.3, pp.210–214, 2001.

[7] T. Kemp and T. Schaaf, “Estimating confidence using word lattices,” Proc. 5th Eurospeech, pp.827–830, 1997.

[8] S. Khudanpur and J. Wu, “Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling,” Comput. Speech Lang., vol.14, no.4, pp.355–372, 2000.

[9] 小玉康広，宇津呂武仁，西崎博光，中川聖一，“複数の音声認識システムの出力の共通部分を用いた認識誤り検出，” 言語処理学会第 7 回年次大会論文集，pp.389–392, 言語処理学会，2001.

[10] 小玉康広，渡邊友裕，宇津呂武仁，西崎博光，中川聖一，“機械学習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，” 情処学研報，2003–SLP–45, pp.95–100, 2003.

[11] 諸戸正憲，松本　弘，“大語彙連続音声認識によるメル LPC 分析の評価，” 信学技報，SP2000-62, 2000.

[12] 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，” 信学論（D-II），vol.J83-D-II, no.6, pp.1412–1421, June 2000.

[13] 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，” 情処学研報，2001–SLP–36, pp.87–92, 2001.

[14] 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，” 情処学研報，2000–SLP–32, pp.1–6, 2000.

[15] H. Schwenk and J.-L. Gauvain, “Combining multiple speech recognizers using voting and language model information,” Proc. 6th ICSLP, vol.II, pp.915–918, 2000.
[16] 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,” 信学技報，SP2002-22, 2002.
[17] 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,” 信学技報，SP2001-128, 2002.
[18] 渡邉友裕，山本博史，小窪浩明，菊井玄一郎，西崎博光，小玉康広，宇津呂武仁，中川聖一，“機械学習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 — 旅行会話音声における評価,” 日本音響学会 2003 年春季研究発表会講演論文集，vol.I, 2003.
[19] F. Wessel, K. Macherey, and H. Ney, “A comparison of word graph and N-best list based confidence measures,” Proc. 6th Eurospeech, pp.315–318, 1999.

（平成 14 年 8 月 19 日受付，15 年 1 月 31 日再受付）

宇津呂武仁

1989 京大・工・電気工学第二卒．1994 同大大学院工学研究科博士課程電気工学第二専攻了．京都大学博士（工学）．同年，奈良先端科学技術大学院大学情報科学研究科助手．1999〜2000 米国ジョンズ・ホプキンス大学計算機科学科客員研究員．2000 豊橋技術科学大学工学部情報工学系講師，2003 京都大学大学院情報学研究科知能情報学専攻講師，現在に至る．自然言語処理，音声言語情報処理の研究に従事．情報処理学会，人工知能学会，日本ソフトウェア科学会，言語処理学会，日本音響学会，ACL 各会員．

西崎　博光　（正員）

1998 豊橋技科大・工・情報工学卒．2000 同大大学院工学研究科修士課程情報工学専攻了．2003 同大学院工学研究科博士後期課程電子・情報工学専攻了．博士（工学）．2003 山梨大学大学院医学工学総合研究部助手，現在に至る．音声言語情報処理に関する研究に従事．情報処理学会，日本音響学会各会員．

小玉　康広

2001 豊橋技科大・工・情報工学卒．2003 同大大学院工学研究科修士課程情報工学専攻了．現在，ソニー株式会社インテリジェント・ダイナミクス研究所インテリジェンスグループに勤務．在学中は，音声言語情報処理に関する研究に従事．

中川　聖一　（正員）

1976 京大大学院工学研究科博士課程了．同年京都大学工学部情報工学科助手．1980 豊橋技術科学大学工学部情報工学系講師．1990 同教授．1985〜1986 カーネギーメロン大学客員研究員．音声言語情報処理，自然言語処理，人工知能の研究に従事．工博．1977 電子通信学会論文賞，1998 年度 IETE 最優秀論文賞，2001 本会論文賞受賞．著書「確率モデルによる音声認識」（電子情報通信学会編），「音声・聴覚と神経回路網モデル」（共著，オーム社），「情報理論の基礎と応用」（近代科学社），「パターン情報処理」（丸善）など．