論　文

大規模な日本語話し言葉データベースを用いた講演音声認識

南條　浩輝[†]　加藤　一臣[†]　李　晃伸[††]　河原　達也[†]

Lecture Speech Recognition Using Large Corpus of Spontaneous Japanese

Hiroaki NANJO[†], Kazuomi KATO[†], Akinobu LEE[††], and Tatsuya KAWAHARA[†]

あらまし　開放的融合研究「話し言葉工学」プロジェクトにおいて構築されている日本語話し言葉コーパスを用いて講演音声の認識を行った．話し言葉は書き言葉の読上げ音声と大きく性質が異なるため，それに合致したモデル化と認識手法の検討が必要となる．音響モデルについては発話スタイルとデータ量の影響を調べた．言語モデルについては，話し言葉コーパスのデータ量不足を補うために他のコーパスと混合する方法，特に混合重みの最適化手法を提案する．また認識に際して，事前の発話のセグメンテーションが容易でないため，ショートポーズの自動認識に基づいて区分化と認識結果の確定を行う逐次デコーディング方式を提案・実装した．10名の話者による講演音声の認識実験で提案手法の有効性を示し，平均66.2%の認識率を得た．

キーワード　話し言葉，音声認識，音響モデル，言語モデル，逐次デコーダ

1. まえがき

ディクテーションシステムに代表される「書き言葉」の音声認識は，数万語レベルの大語彙連続発声でも90%程度の単語認識率が達成されるようになった[1]．しかしこれらは，発話が辞書の表記どおりめいりょうであり，文法的にも正しいという前提に基づくものである．一方，人間同士の会話のような「話し言葉」を対象とした場合，ふめいりょうな発話や口語的表現，間投詞や言いよどみなど，音声認識の上で多くの問題が存在する．大語彙の話し言葉音声認識の技術は，講演の書き起こしや会議の議事録作成，同時通訳などの基盤となる技術であり，その実現に対する要望も大きい．

日本語の話し言葉に対する音声認識の研究は，これまで主にドメインを限定したタスクで行われており，大語彙に関しては十分な研究がなされていなかった．その主な原因としてデータベースが整備されていなかったことが挙げられる．これに対して，平成11年度より始まった開放的融合研究「話し言葉工学」（代表：古井貞熙教授）のプロジェクトでは，独話による講演を主な対象として，700万形態素を目標に大規模なコーパスの構築を進めている．本論文では，この構築されつつある日本語話し言葉コーパス (CSJ: Corpus of Spontaneous Japanese) [2], [3] を用いて，話し言葉音声認識の基礎的な研究の立場から，認識の実時間性を考慮しないで，精度を優先した方法を探求する．また，汎用的な音声認識を目的とし，個々の話者や講演の話題へのモデル適応は行わない．その上で，音響モデル・言語モデル・デコーダの各々について，話し言葉音声認識における問題点を検討しながら，対処法について述べる．

まず，音響モデルについては学習データ量と発話スタイルの影響を調べた．このような検討は，大規模な日本語話し言葉コーパスにおいて，これまでなされていない．

言語モデルに関しては，学習データが書き言葉に比べると少ないので，他の言語資源の利用を検討した．特に本論文では講演録の利用を考え，それとのスタイルの違いの影響を調べた上で，効果的にテキストデータを混合する手法を提案する．

更に，話し言葉では発話が文単位でなされず，区切りが明確でないので，認識の際に大きな問題となる．本論文ではまず，ポーズのモデル化について検討を行った上で，認識とセグメンテーションを並行して行

[†] 京都大学情報学研究科，京都市
Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan
[††] 奈良先端科学技術大学院大学情報科学研究科，生駒市
Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Ikoma-shi, 630–0192 Japan

うデコーディング手法を実現する．このような逐次デコーディングは従来のディクテーションやニュース音声の認識においても行われているが，それらは主に文単位の発話を仮定した上で実時間処理を目的としていた．これに対して本論文で提案する手法は，話し言葉音声において探索を安定にし，認識精度の改善を目指すものである．

2. 講演音声データベース (CSJ)

まずCSJについて簡単に説明し，次に本論文で実験に使用した学習データとテストセットについて述べる．

CSJは主に学会講演と模擬的な講演からなるコーパスであり，音声データと人手による書き起こしテキストから構成される．収集・書き起こしの途上であり，現在は音声・言語関連の学会発表が中心である．表1に収集された学会の一例を示す．音声データはヘッドセットマイクで収録され，16 kHz，16 bitでサンプリングされている．書き起こし作業は詳細なマニュアル[3]に基づいて行われているため，作業者によらず高い精度で行われている．

表1　CSJの講演の一部
Table 1　Lectures in CSJ.

講演種	略称
日本音響学会春季&秋季研究発表会	AS
電子情報通信学会音声研究会	SP
言語処理学会年次大会	NL
国語学会	JL
音声学会全国大会	PS
社会言語科学会	SG
国立国語研究所内での種々の研究会	KK
融合研究の会合	YG
模擬講演	AC, IG, ST

表2　音響モデル用学習セットの概要
Table 2　Data sets for acoustic model training.

	使用講演（講演数）	データ量
set-1a	AS(102)	13.2 時間
set-2a	AC(134) + IG(78) + ST(27)	23.9 時間
set-3a	set-1a + SP(11) + NL(45) + JL(9) + PS(17) + KK(6) + YG(5)：計 195	35.3 時間
set-4a	set-1a + set-2a：計 341	37.1 時間
set-5a	set-2a + set-3a：計 534	59.2 時間
IPA [1]	JNAS コーパス	約 40 時間

表3　言語モデル用学習セットの概要
Table 3　Data sets for language model training.

	set-1b 2000/10	set-2b 2001/02	set-3b 2001/02+	Web
講演数	186	316	612	81
単語総数	466842	759512	1480834	1692802
異なり語数	17172	21381	29939	37462

+: 書き起こしのチェックが行われていないものを含む

本論文で音響・言語モデル学習用に用いたデータを表2及び表3に示す．音響モデルの学習には，基本的に2000年10月時点のものを用いた．データ量はJNASコーパス[4]と同程度である．女性話者の講演データが少量であったため，男性用の性別依存モデルのみを作成した．言語モデルの学習には，CSJの2000年10月と2001年2月時点のもの及びWebから収集した講演録（Web講演録）を用いる．

次にテストセットを表4に示す．分野・学会のバランスも考慮しながら，講演に熟練した話者により，原稿を用いずに話されている講演を中心に選定した．最後の3講演は学会ではない非公開の会合でのものである．最初の4講演は音声関連のものであるが，本論文の大半の実験でこれらを用いている．

なお本研究では，形態素解析システムとしてChaSen ver2.02を用いており，単語の定義はそれに基づく．また，単語総数の計数においてポーズを含めていない．

3. 音響モデルの比較

3.1 音響モデルの仕様

音響モデルは混合連続分布HMM（対角共分散）でありHTK [5]で作成した．音声データ（16 kHz，16 bit）をフレーム長25 msのハミング窓，フレーム周期10 msで音響分析を行った．各フレームごとにMFCC（12次元），ΔMFCC（12次元），ΔPower（1次元）を計算し，計25次元の特徴量ベクトルを求めた．

音素は43種類とし，各音素は3状態left-to-right HMM（飛び越し遷移なし）でモデル化した．音素環境依存モデル，具体的には通常の状態共有triphoneモデルとPTM (Phonetic Tied-Mixture) triphoneモデル[6]も作成した．その際，決定木に基づく状態共

表4　テストセットの概要
Table 4　Test set.

講演名（略称）	単語数	時間（分）
A01M0035 (AS22)	6294	28
A01M0007 (AS23)	4391	30
A01M0074 (AS97)	2508	12
A05M0031 (PS25)	5372	27
A02M0117 (JL01)	9833	57
A03M0100 (NL07)	2644	15
A06M0134 (SG05)	4460	23
KK99DEC005 (KK05)	6536	42
YG99JUN001 (YG01)	2759	14
YG99MAY005 (YG05)	3108	15

有を行い，状態数 1000，2000，3000 のモデルを作成した．これらの仕様はおおむね IPA モデル [1] と同じである．

音響モデルの学習に必要な音素ラベルの作成は人手による書き起こし [7] に基づいて行った．その際，書き起こしの仮名表記欄に仮名のみ，または仮名と以下のタグのみを含む音声区間を学習データとした．

言い直し (D) (D2)，言い淀み (F)，口語 (S)，発音の転化・なまけ (W)，個人名・差別語・誹謗中傷など (R)，メタ的引用 (M)，漢字表記不可 (K)

ただし，これらのタグ情報は利用していない．

3.2 学習セット

ここでの評価データは表 4 上段に示す，主に日本音響学会での講演（AS）である．そこで学習セットに

set-1a: 日本音響学会の講演（AS）のみ
set-2a: 模擬講演（AC+IG+ST）のみ
set-3a: AS+学会講演（JL+KK+NL+PS+SP+YG）
set-4a: AS+模擬講演（AC+IG+ST）
set-5a: AS+学会講演+模擬講演

の 5 種類を選んだ（表 2）．

set-1a はテストセットと講演の話題まで含めてほぼ一致しているが，データ量は 13.2 時間とほかに比べて少ない．set-2a は発話スタイルの異なる模擬講演のみで構成する．データ量は 23.9 時間であり，set-1a に比べると多い．set-3a は set-1a に他の学会講演を加えたものであり，発話スタイルは学会講演という点で一致している．データ量も 35.3 時間と比較的多い．set-4a は set-1a に発話スタイルの異なる模擬講演を加えたものであり，データ量は 37.1 時間と set-3a とほぼ同程度になっている．set-5a は利用できる学会講演と模擬講演のすべてであり，データ量は 59.2 時間と最も多い．

これらにより，学習データ量の効果及びテストデータとの一致度の効果を調べる．

3.3 音響モデルの評価

各々の学習セットで種々の音響モデルを作成し，テストセット上段 4 講演に対して認識実験を行った．比較のため，読上げ音声の JNAS コーパスで学習された IPA モデルでも評価を行った．言語モデルは，CSJ（set-3b）で学習したもの（後述）を用いている．デコーダは Julius 3.1 [8] であり，事前にテストセットの音声データをパワーと零交差数に基づいてファイルに分割してから，認識を行っている．表 5 にテストセット 4 講演分に対する単語正解精度を示す．

set-1a から学習したモデルは認識率が低く，学習量不足が原因と考えられる．PTM は通常の triphone に比べてパラメータ数が少ないため，13 時間程度の学習量でも triphone より高い性能を示している．一方，set-2a（模擬講演のみ）から学習したモデルは全体的に性能が低く，学会講演に対しては高い認識率を得ることができなかった．読上げ音声 40 時間で学習した IPA モデルでは，更に低い認識率しか得られなかった．

次に，set-1a に他の学会講演や模擬講演を加えて学習データを増加させた場合について考察する．set-1a に他の学会講演を加えた set-3a（約 35 時間）から学習したモデルは，set-1a のみで学習したモデルに比べて高い認識率を得た．学会講演という発話スタイルが一致したデータを加えて学習データ量を増加させた効果と考えられる．一方，set-1a に模擬講演を加えた set-4a（約 37 時間）から学習したモデルを用いた場合は，set-3a と学習データ量が同程度であるものの，認識率は低く，set-1a と比べても認識率にほとんど差は

表 5 テストセット 4 講演に対する単語正解精度（%）
Table 5 Word accuracy for four test-set lectures (%).

モデル	学習セット					
	set-1a	set-2a	set-3a	set-4a	set-5a	IPA
monophone 129×32	55.3	50.6	55.4	53.6	54.9	-
monophone 129×64	56.6	52.4	57.3	55.5	56.7	-
triphone 1000×16	61.1	59.2	64.9	62.4	64.4	53.5
triphone 2000×16	59.3	56.4	63.9	61.2	64.1	53.5
triphone 3000×16	57.5	54.1	62.8	58.1	62.6	52.6
PTM 129×64(1000)	62.6	58.8	64.3	62.5	63.3	-
PTM 129×64(2000)	62.9	58.9	64.6	63.0	63.8	-
PTM 129×64(3000)	62.5	58.8	64.7	62.9	64.0	53.1

AS22，AS23，AS97，PS25 に対する平均

各モデルの後ろの数字は，コードブック数 × 混合数（状態数）を示す．
PTM 以外ではコードブック数は状態数と等しいため省略．

なかった．set-1a に学会講演と模擬講演の両方を加えた set-5a（59.2 時間）から学習したモデルを用いた場合でも，学会講演のみ (set-3a) から学習したモデルより高い認識率は得られなかった．この結果は，話し言葉においても，学会講演と模擬講演では発話スタイルが異なり，学会講演の音声のモデル化に対して，模擬講演のデータを加える効果はあまりないことを示している．

模擬講演は，趣味や旅行などのあまり堅苦しくない内容についてゆっくり語りかけるようなものが多い．また，学会講演の発声は多数の聴衆を前にした緊張した状況で行われているのに対して，模擬講演の発声は聴衆はいるもののリラックスした状況で行われている．このような発声状況の違いが心理的に発話スタイルに影響しているものと考えられる [9]．

次章以降では，最も安定して高い認識率を得ることができた set-3a で学習した PTM モデルを用いる．

4. 種々のコーパスを利用した言語モデル

次に言語モデルについて検討する．まず学習データ量の問題を考察した上で，これを補完するために他の言語資源，具体的には講演録コーパスの利用を考える．講演録と実際の講演の書き起こしの違いを調べながら，両者を効果的に混合する手法を提案する．

4.1 学習テキスト量の影響

最初に，言語モデルの学習テキスト量の影響について調べる．ここでは 3 種類の学習セット set-1b，set-2b，set-3b の比較を行った．

set-1b，set-2b は表記も含めてチェックが済んだ忠実な書き起こしのみで構成する．set-1b と set-2b の間では，書き起こし基準は統一されており，set-1b は set-2b のサブセットとなっている．両者の差は，CSJ が構築途上であり，書き起こしが追加された結果生じたものである．更に，set-2b に未チェックの書き起こしを加えた学習セット set-3b も比較する．set-3b は set-2b に比べてデータ量が 2 倍程度になっている．

語彙は，各々の学習セット中で 2 回以上出現した単語で構成した．次に各語彙で単語 3-gram モデルを学習した．学習は，CMU-ToolKit ver.2 [10] を用いて行った．カットオフのしきい値はいずれも 0 で，バックオフスムージングを行っている．

これらによるテストセット 4 講演に対するカバレージ，テストセットパープレキシティを表 6 に示す．カバレージは，テストセットの単語のうち言語モデルの語彙に含まれるものの割合である．パープレキシティの計算は，未知語はコンテクストには含めるが評価の際には無視して行っている．パープレキシティの平均は，各テストセットの書き起こしを結合して一つのテキストにし，それに対して計算し求めている．更に，ベースラインの音響モデルとデコーダ (Julius 3.1) を用いて認識実験を行った結果も表 6 に示す．

表 6　各言語モデルによるカバレージ，パープレキシティ，単語正解精度
Table 6　Coverage, test-set perplexity and word accuracy by language models.

	set-1b	set-2b	set-3b
語彙サイズ	10346	13314	19158
カバレージ（%）	95.3	96.0	96.8
パープレキシティ	150.9	142.3	132.5
単語正解精度（%）	61.5	63.8	65.1

AS22，AS23，AS97，PS25 に対する平均

学習テキスト量の増加に伴って，語彙サイズは約 1 万から 2 万になり，カバレージが改善された．語彙が一致していないためパープレキシティの公正な比較は困難であるが，学習量の増加に伴って語彙サイズが増加したにもかかわらず，パープレキシティの改善も得られた．認識率も set-1b から set-3b まで着実に改善されており，未チェックの書き起こしを利用してでも学習量を増加させる効果が見られた．これらの結果は，N-gram 言語モデルを学習するために，話し言葉のコーパスがまだ十分でない可能性を示唆している．

4.2 種々のコーパスの重み付き混合

ディクテーションシステムにおける言語モデルの学習には新聞記事等が用いられてきた．そのテキストサイズは 1 億語を超える膨大なものであり，精度の高い言語モデルの学習を可能にしている [1]．一方，話し言葉のコーパスとしては大規模な CSJ でも，その目標テキストサイズ（700 万形態素）は新聞記事数か月分程度であり，set-3b では 1 か月分程度でしかない．したがって，他の言語資源との効果的な混合により，学習を強化・補完することを考える [11]．

4.2.1 Web 講演録

World Wide Web 上には膨大なテキストデータが存在し，言語モデルの構築にも利用できる [12]．ここでは，Web 上で公開されている講演録を収集した（表 3 の Web）．CSJ に比べて，講演数は少ないが，出現単語総数は 169 万語で set-3b を若干上回っている．ただし，これらの Web 講演録は主に政治家や著名人の講演から構成され，CSJ とは話題・分野が大きく異なる．

そこで，個々の講演に固有の話題依存の単語ではなく，話し言葉に普遍的な単語を抽出するために，話題との相互情報量に基づいて話題独立な語彙の選択 [13] を適用している．頻度による語彙選択と比較すると，上位8000語のうち21.5%が入れ換わっており，本論文と別のタスクにおいてカバレージや認識率が改善されることを確認している [14]．

4.2.2 言語モデルの混合手法

複数のテキストコーパスを組み合わせて言語モデルを構築するのに，各モデルを重み付けして混合する方法が効果的である [15]．ここでは，それぞれの学習コーパスで得られたN単語連鎖の出現頻度を重み付けして加えた後にN-gram言語モデルを作成する手法を採用する．

各テキストコーパス i での単語列 s の出現頻度を $\{C_i(s) \mid (i = 1, \ldots, n)\}$，出現頻度に対する重みを $\{\mu_i \mid (i = 1, \ldots, n)\}$ とすると，重み付け混合を行った後の単語 w の合計出現頻度 $C(w)$ は以下の式で与えられる．

$$C(w) = \sum_{i=1}^{n} \mu_i C_i(w)$$

単語履歴を h とすると $P(w|h) = \frac{C(hw)}{C(w)}$ に基づいて，統合された語彙やカットオフ条件のもとで，確率の推定及びバックオフスムージングを行える．

4.2.3 削除補間法による混合重みの自動推定

重み付け混合は効果的な手法であるが，最適な混合重みの推定が大きな問題である．従来，学習セットの一部を重み推定用に除外したり（ヘルドアウト法），あるいはテストセットを用いて事後的に調整したりすることが一般的であった．例えば文献 [16] では，ヘルドアウト法を用いて新聞記事から学習したモデルを放送大学の講義の書き起こしに適応を行っている．

これに対して，学習セットのすべてを利用する（テストセットは用いない）混合重みの最適化を考える．この例としては，ニュース原稿の重み付け混合手法 [17] があるが，これはニュース原稿テキスト間に時期的な依存性があることを利用し，前日の記事を擬似的なテストセットとするものである．本論文では，このような特殊な条件を仮定しない汎用的な手法を提案する．すなわち，タスクに合致したマッチドテキストをターゲットとする削除補間法を考える．ここでのターゲットは，実際の講演音声認識に最も合致するCSJとする．削除補間法は，学習データを分割して交互に重みを推定するもので，学習データを最大限に活用できる．

具体的な手順は以下のとおりである．

（1） マッチドコーパスの語彙と大規模コーパスの語彙との併合語彙を作成する．

（2） マッチドコーパス U を n 個の部分集合に分割し，各部分集合 U_j に対して（5）までを行う．

（3） U_j をマッチドコーパスから除外し，評価用テキストとする．

（4） マッチドコーパス $\{U_k \ (k \neq j, \ k = 1, \ldots, n)\}$ と大規模コーパスから併合語彙を用いて，それぞれ言語モデル（単語N-gram）を学習する．

（5） 各言語モデルを重み付けして混合を行う．この際に，評価用テキスト U_j に対するパープレキシティが最小になるように混合重み μ_{ij} を求める．

（6） μ_{ij} をすべての部分集合 U_j について平均し，最終的な混合重み μ_i とする．

手順（5）において，パープレキシティを最小とする μ_{ij} を見つけるには，まずパープレキシティ $PP(\mu_{ij})$ が $PP(a) > PP(b)$，$PP(c) > PP(b)$ となるような $a < b < c$ で囲い込みを行い，囲い込まれた区間に対して黄金分割法による推定を行う．

なお，言語モデル確率の混合の場合は，$\Sigma\mu_i = 1$ といった制約を入れることができ，実質的には μ_i の比を求めればよいが，出現頻度の混合の場合はバックオフスムージングや実数を整数化（切上げ）する際の丸め誤差の影響から，絶対値によって生成されるモデルが若干異なる．そのため μ_i 各々について探索する．

4.3 Web講演録との混合によるモデルの評価

CSJの学習セットset-1bとset-3bの各々についてWeb講演録との混合を行った．その際，削除補間法を適用するために，CSJを七つの集合に分割した．混合重みの推定の結果，set-1bではCSJ: 0.93，Web: 0.16，set-3bではCSJ: 0.95，Web: 0.066となった．CSJのデータ量が大きくなるにつれて，Web講演録の重みは小さくなっている．

これらの言語モデルによる，テストセットに対するカバレージ・パープレキシティ・単語正解精度を表7に示す．Web講演録単独で作成したモデルは，すべての尺度において他のモデルより性能が低い．これは，間投詞や口語的表現などが人手で修正されており，忠実な書き起こしでないことに起因する．しかし，このWeb講演録をCSJに混合することによってカバレージが若干向上した．パープレキシティの一様な上昇は，語彙サイズが大きくなったためである．両者の混合に

表 7　言語モデルにおける Web テキスト混合の効果
Table 7　Result of incorporating Web text in language modeling.

	Web のみ	set-1b	set-1b+Web		set-3b	set-3b+Web	
			1:1 混合	最適化		1:1 混合	最適化
語彙サイズ	8024	10346	13197		19158	20531	
カバレージ（%）	84.5	95.3	95.8		96.8	96.9	
パープレキシティ	229.3	150.9	182.3	168.0	132.5	144.6	138.2
単語正解精度（%）	50.4	61.5	61.1	63.0	65.1	64.8	65.6

AS22，AS23，AS97，PS25 に対する平均

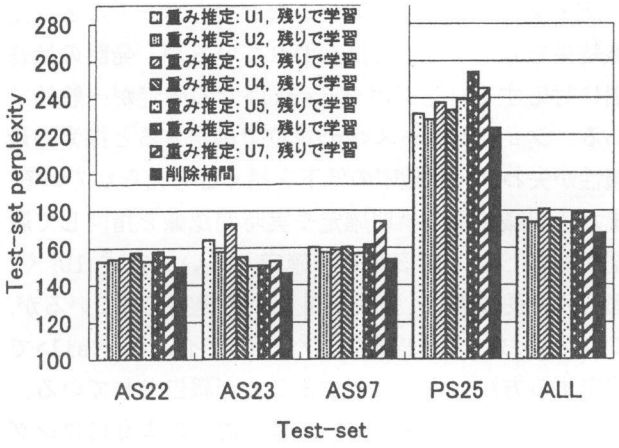


図 1　削除補間法とヘルドアウト法によるテストセットパープレキシティの比較
Fig. 1　Test-set perplexity using deleted interpolation and held-out method.

より作成した言語モデルでは，同じ語彙をもつので，混合重みにかかわらずパープレキシティは公正に比較できる．提案手法を用いて重みを自動推定することで，単純に 1：1 に混合したときに比べてパープレキシティが小さくなっている．

参考のために，削除補間法とヘルドアウト法の比較を行った．ヘルドアウト法では学習セットの一部を重みを推定するためだけに用いるが，この割合を小さくすると重みの推定の信頼性が低下し，逆に大きくすると N-gram 自体の学習データが少なくなる．ここでは set-1b について，削除補間に用いた 7 分割のうち一つをヘルドアウトした場合を 7 通り試行した．その結果を図 1 に示す．削除補間法に比べて，全般にパープレキシティが 5〜10% 大きくなっている．これは N-gram 自体の学習量と重みの推定精度が低下したためであると考えられる．

認識率に関しては表 7 に示すとおり，単純に 1：1 で混合した場合はベースラインのモデルよりも低下したが，提案手法による混合重みの推定により改善が得られた．特に，set-1b については平均で単語正解精度が 1.5% 向上した．単語誤り率の差の検定を行ったところ，有意水準 1% でこの改善は有意であり，提案手法を用いて Web 講演録を混合することの有効性が示された．

しかし，set-3b に Web 講演録を混合した場合はその混合重みが小さくなり，認識率の改善も有意でなかった．また，set-1b に新聞記事コーパスの混合も試みたが，新聞記事は性質が大きく異なるため，混合重みが非常に小さく推定されてしまい，効果が見られなかった [18]．具体的には，最もよい場合でも新聞記事コーパスの重みが CSJ の重みの 5% 程度で，パープレキシティは 4 講演平均で 178.8 であった．

これらの結果は，マッチドコーパスのテキスト量が大きくなるほど，混合の意義が薄れていくことを示している．どの程度の量のマッチドコーパスが得られれば混合の効果が得られなくなるかは，混合するコーパスの性質の一致度や，その量に依存すると考えられる．

一方，マッチドコーパスの量が少ない場合は，提案手法を用いて複数のコーパスを利用することは効果的である．タスクやドメインを限定して統計的言語モデルを構築する場合，現実には CSJ のような大量のデータを集めることは困難であり，本手法の応用範囲は広いと考えられる．

5.　認識時におけるポーズの扱い

話し言葉においては，発話の区切りは必ずしも文の単位と一致せず，文の認定も容易でない．実際に CSJ の書き起こしにおいては，原則として 200 ms 以上のポーズで発話を区切っており，ポーズ情報は継続時間とともに含まれているが，句読点は付与されていない．本研究においては，テストセット音声をパワーと零交差数に基づいて，約 500 ms のポーズ長で自動的に分割し，これを入力発話と扱っている．したがって，文単位で発声されることを想定し，句点を発話終端に，読点を発話内のショートポーズに対応づけることもできた読上げ音声と大きく異なる．このような日本語話し言葉の音声認識におけるポーズの扱いについては，

これまで十分に検討されていない．

5.1 ポーズのモデル化

そこでまず，ポーズのモデル化について検討を行った．発話の始終端のポーズに対応する<sil>と発話内のショートポーズ<sp>に関する言語モデル上の扱いに関して，下記の4通りについて比較を行った．

（1） ポーズ情報ラベルを用いず，始終端<sil>は言語モデル中の未知語カテゴリー<UNK>で代用し，ショートポーズ<sp>はモデル化しない

（2） 無音区間をすべてショートポーズ<sp>とする

（3） しきい値1000 ms以上のポーズを始終端<sil>に対応させ，しきい値未満をショートポーズ<sp>に対応させる

（4） しきい値500 ms以上のポーズを始終端<sil>に対応させ，しきい値未満をショートポーズ<sp>に対応させる

set-1bを用いて各々に対応する言語モデルを学習し，4名のテストセットに対する認識実験を行った結果を表8に示す．

（1）ではショートポーズが適切にモデル化されないため最も認識率が低く，（2）では始終端が適切にモデル化されないため，やはり若干認識率が低下している．これらはポーズのモデル化がある程度重要であることを示している．（3）と（4）はほぼ同等の性能を得ているが，最も高い認識率が得られた1000 msをしきい値として学習したモデルをベースラインとした．

5.2 逐次デコーディング

次に，デコーダにおけるポーズの扱いについて検討する．話し言葉においては，発話の区切りが文の単位と一致しないので，単純に一定のポーズ長で音声を区分化すると，分割された音声区間が極端に長くなったり，逆に細切れになったりする．実際，テストセット音声に対して一定のポーズ長で区分化を行うと，50単語以上からなる極端に長い発話と，『まー』『えー』『なります』のようにフィラーや文末表現だけのような短い発話が多数生成された．特に，文末などでは発声が弱くなる傾向にあるため，パワーや零交差数に基づいて音声区間を検出することは容易でない．発話（入力音声）が長いと仮説数が膨大になるため，探索の失敗を引き起こす要因となる．逆に短すぎると，言語モデルの制約があまり作用しなくなる．これに対して，事前にセグメンテーションを行うことなく，逐次的に認識結果を確定していくデコーディングを検討する．

ディクテーションシステムなどでもポーズごとに認識結果を確定することは行われているが，発話の始終端に対応するロングポーズ<sil>での確定が一般的である．ショートポーズ<sp>単位で確定すると探索の最適性が失われ，認識率の低下を招くと考えられている．また，認識結果の早期確定や実時間認識を指向して履歴が変化しなくなる場合に確定していく方式[19]や，固定入力長に基づく方式[20]なども提案されているが，これらにおいても認識率はベースライン（=<sil>で確定する方法）より低下することが報告されている．

しかし，話し言葉においては，前述のようにロングポーズの挿入が不規則で発話長が不安定になる現象が顕著である．また，これまで逐次的にデコーディングを行う手法も十分に検討されていない．そこで，本論文では，探索を安定させるために，入力中のショートポーズ<sp>を認識・検出することによって，そこまでの区間を順次確定していく方法を提案する．

Juliusは前向き・後向きの2パス探索を行うが，ショートポーズの検出には，第1パスの認識における<sp>モデルのゆう度を用いる．第1パス処理中に各フレームの最ゆう単語仮説を調べて，<sp>モデルが最ゆうとなるフレームがしきい値以上連続する区間を無音区間と判定して，そこで第1パスを中断し，第2パスを実行する．区間の検出と第1パスの再開の様子を図2に示す．第1パスはポーズ区間を開始点までさ

表8 ポーズのモデル化の違いによる単語正解精度（%）
Table 8 Word accuracy with various modeling of pauses (%).

ポーズラベル情報	未使用	すべて<sp>	1000 ms	500 ms
AS22	50.4	55.3	55.5	55.5
AS23	61.0	66.8	68.1	67.7
AS97	64.2	67.2	67.8	67.6
PS25	55.8	59.0	60.3	59.8
平均	56.3	60.7	61.5	61.3

言語モデル学習セット：set-1b

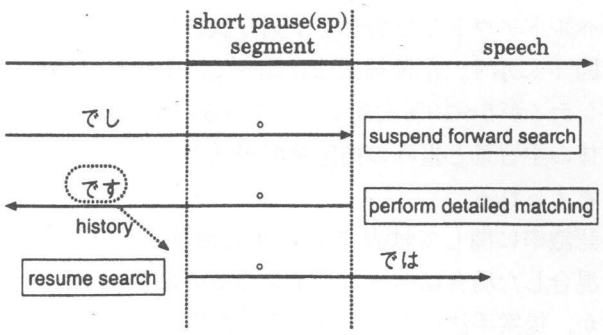


図2 逐次デコーディング
Fig. 2 Sequential decoding.

かのぼって認識を再開するが，その際に初期仮説として<sp>を，初期の単語履歴として直前の第2パスで確定したポーズ以外の単語を割り当てる．

これにより，第2パスの探索範囲を小さくでき，全体的な認識精度の向上が期待できる．ポーズの検出精度に依存する手法であるが，パワーや零交差数に基づいてセグメンテーションを行うよりは，音響的・言語的な情報が総合的に反映されており，音響モデルの適応も導入できる可能性があるので，信頼性は高いと考えられる．

この逐次デコーディングアルゴリズムを Julius-3.2 において実装し，テストセットの講演すべてに対して認識実験を行った．まず，音声区間を検出しセグメンテーションする際のしきい値（<sp>モデルの継続時間長）と認識率との関係を調べた．結果を図 3 に示す．話者によらず，しきい値を小さくした場合も大きくした場合も認識誤りが増えている傾向がわかる．しきい値を大きくすると探索が困難になり，小さくすると言語制約が弱くなるためである．話者ごとに最適なしきい値は異なるが，10 話者（講演）平均で最も高い認識率が得られたのは，しきい値を 120 ms に設定した場合であった．

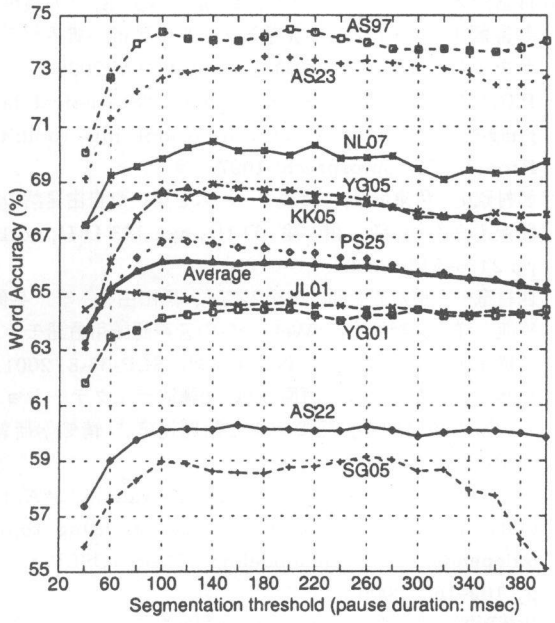


図 3　セグメンテーションしきい値と単語正解精度
Fig. 3　Accuracy vs. segmentation threshold.

次に，発話ごとにセグメンテーションを行ってから認識を行う従来の方法との比較を 表 9 に示す．

これまでの実験で用いてきた音声の自動切出しに加えて，人手による書き起こしの時間ラベル（ポーズ情報）に基づく切出しを用いた場合も比較した．なおこの際に，音響モデルも 2001 年 2 月時点のデータで作成し直した（224 講演，37.9 時間）．

従来の認識法において，音声を自動切出しした方が，時間ラベルに基づいて分割した場合に比べて発話数は少ない．自動切出しでは，促音や文末の発声の弱い区間で誤って分割されてしまい，当該区間で認識誤りとなる現象も確認された．一方，手動ラベルを用いた場合は，多くの単位に区切られるが，言語モデルの制約が作用しなくなる．

これに対して逐次デコーディングでは，人手による時間ラベルに基づいて分割した場合とほぼ同数の発話に切出すことができ，しかも，前の発話の履歴を引き継ぐため，従来手法（自動切出しの場合）と比べて認

表 9　逐次デコーディングによる単語正解精度（%）
Table 9　Word accuracy using sequential decoding (%).

	従来の認識法		逐次デコーディング（提案手法）	パープレキシティ
音声セグメンテーション	事前，自動	事前，手動 *	認識時，自動	
AS22	59.9 (349)	59.4 (691)	60.2 (943)	133.5
AS23	73.3 (247)	72.2 (802)	73.0 (998)	107.5
AS97	73.5 (120)	71.8 (253)	74.2 (372)	117.2
PS25	66.1 (273)	66.0 (490)	66.9 (506)	164.4
JL01	64.3 (507)	64.1 (1336)	64.8 (1452)	186.8
NL07	69.1 (153)	69.6 (352)	70.3 (448)	94.5
SG05	57.2 (197)	58.8 (431)	58.9 (470)	111.8
KK05	66.5 (426)	69.1 (740)	68.8 (925)	127.7
YG01	58.3 (145)	64.0 (239)	64.3 (303)	125.5
YG05	68.3 (176)	68.1 (275)	68.5 (337)	117.8
10 講演平均	65.0 (2593)	65.6 (5609)	66.2 (6754)	135.1

言語モデル学習セット：set-3b，新しい音響モデル，() 内は分割された発話の数
手動 *：人手による書き起こしの時間ラベル使用

識率の向上（1.2%）を得ることができた．単語誤り率の差の検定を行ったところ，有意水準1%でこの改善は有意であった．これは話し言葉においては，逐次デコーディングが探索の最適性を損なうよりも，探索空間を絞り込み，探索を安定化させる効果が高いことを示すものである．また，事前の音声セグメンテーションの際には，無音検出のためのパワーや零交差数などのパラメータを，話者ごとに適応的かつ事後的に変化させるなど，調整にかなりの労力を要していたので，セグメンテーションを同時並行して行う逐次デコーダの意義は大きい．

6. むすび

講演音声のような話し言葉の音声の認識において，音響・言語モデル学習に関する種々の検討とデコーダの改善法について述べた．

音響モデルに関しては，話し言葉音声と読上げ音声では発話スタイルが異なることを確認し，また，話し言葉においても学会講演と模擬講演では若干発話スタイルが異なること，及び発話スタイルの一致した音声でモデルを構築することの重要性を示した．

言語モデルに関しては，話し言葉のデータ量が非常に重要であることを示し，更に他のコーパスと効果的に混合する手法を提案し，有効性を示した．

話し言葉では文の区切が明確になされないことから，ショートポーズを認識・検出することによりセグメンテーションを同時並行する逐次デコーディングの方法を提案・実装した．事前のセグメンテーションの手間を省くことができ，また認識精度の改善につながった．

謝辞 本研究は，開放的融合研究『話し言葉工学』プロジェクトの一環として行われた．アドバイスを頂きました東京工業大学の古井貞熙教授をはじめとして，御協力を頂いた関係各位に感謝致します．

文　献

[1] 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）の性能評価，”情処学研報，SLP-31-2，NL-137-7，2000.

[2] 前川喜久雄，“言語研究における自発音声，”音講論，1-3-10，春季 2001.

[3] 小磯花絵，前川喜久雄，“『日本語話し言葉コーパス』の概要と書き起こし基準について，”情処学研報，SLP-36-1，2001.

[4] 伊藤克亘，伊藤彰則，宇津呂武仁，河原達也，小林哲則，清水　徹，田本真詞，荒井和博，峯松信明，山本幹雄，竹沢寿幸，武田一哉，松岡達雄，鹿野清宏，“大語彙日本語連続音声認識研究基盤の整備—学習・評価用テキストコーパスの作成，”情処学研報，97-SLP-18-2，1997.

[5] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, The HTK BOOK, 1995.

[6] 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，”信学論（D-II），vol.J83-D-II, no.12, pp.2517–2525, Dec. 2000.

[7] 小磯花絵，土屋菜穂子，間淵洋子，斉藤美紀，龍宮隆之，菊池英明，前川喜久雄，“「日本語話し言葉コーパス」の書き起こし基準について，”信学技報，SP2000-104，2000.

[8] 李　晃伸，河原達也，堂下修司，“単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識，”信学論（D-II），vol.J82-DII, no.1, pp.1–9, Jan. 1999.

[9] 龍宮隆之，菊池英明，小磯花絵，前川喜久雄，“大規模話し言葉コーパスにおける発話スタイルの諸相—書き起こしテキストの分析から，”音講論，2-Q-9，秋季 2000.

[10] P.R. Clarkson and R. Rosenfeld, “Statistical language modeling using the CMU-Cambridge toolkit.” Proc. ESCA Eurospeech, 1997.

[11] 西村雅史，伊東伸泰，“講義コーパスを用いた自由発話の大語彙連続音声認識，”信学論（D-II），vol.J83-D-II, no.11, pp.2473–2480, 2000.

[12] 西村竜一，長友健太郎，小松久美子，黒田由香，李　晃伸，猿渡　洋，鹿野清宏，“Web からの音声認識用言語モデル自動生成ツールの開発，”情処学研報，SLP-35-8，2001.

[13] 加藤一臣，李　晃伸，河原達也，“講演ディクテーションのための話題独立言語モデルと話題適応，”情処学研報，SLP-26-2，1999.

[14] K. Kato, H. Nanjo, and T. Kawahara, “Automatic transcription of lecture speech using topic-independent language modeling,” Proc. ICSLP, vol.1, pp.162–165, 2000.

[15] 伊藤彰則，好田正紀，“N-gram 出現回数の混合によるタスク適応の性能解析，”信学論（D-II），vol.J83-D-II, no.11, pp.2418–2427, Nov. 2000.

[16] 伊東伸泰，西村雅史，森　信介，“日本語における口語体言語モデル，”言語処理学会年次大会発表論文集，pp.280–283，2000.

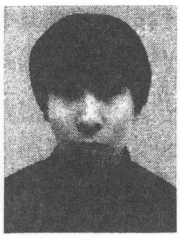
[17] 小林彰夫，今井　亨，安藤彰男，中林克己，“ニュース音声認識のための時期依存言語モデル，”情処学論，vol.40, no.4, pp.1421–1429, 1999.

[18] 加藤一臣，河原達也，“種々のコーパスの重み付き混合に基づく講演音声認識のための言語モデル，”話し言葉の科学と工学ワークショップ，pp.85–92，2001.

[19] 今井　亨，小林彰夫，佐藤庄衛，安藤彰男，“逐次2パスデコーダを用いたニュース音声認識システム，”情処学研報，SLP-29-37，1999.

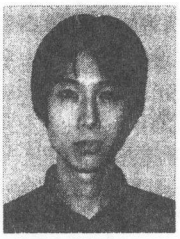
[20] 瀬川　修，武田一哉，板倉文忠，“端点検出を行わない連続音声認識手法，”情処学研報，SLP-34-18，2000.

（平成14年5月10日受付，9月17日再受付）



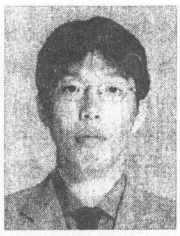
南條　浩輝　（学生員）

1999 京大・工・情報卒．2001 同大大学院情報学研究科修士課程了．現在，同博士後期課程在学中．音声認識・理解の研究に従事．情報処理学会，日本音響学会各会員．



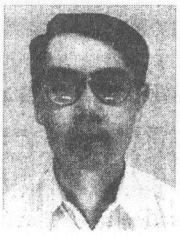
加藤　一臣

1999 京大・工・情報卒．2001 同大大学院情報学研究科修士課程了．現在，松下電器産業株式会社に勤務．



李　　晃伸　（正員）

1996 京大・工・情報卒．1998 同大大学院修士課程了．2000 同大大学院情報学研究科博士後期課程了．現在，奈良先端科学技術大学院大学情報科学研究科助手．音声認識・理解の研究に従事．2001 年度日本音響学会粟屋賞受賞．情報処理学会，日本音響学会各会員．



河原　達也　（正員）

1987 京大・工・情報卒．1989 同大大学院修士課程了．1990 同大博士後期課程退学．同年京大工学部助手．1995 同助教授．1998 同大学情報学研究科助教授．現在に至る．この間，1995〜1996 まで 米国ベル研究所客員研究員．1998 から ATR 客員研究員．1999 から国立国語研究所非常勤研究員．2001 から科学技術振興事業団さきがけ研究 21 研究者．音声認識・理解の研究に従事．京大博士（工学）．1997 年度日本音響学会粟屋賞受賞．2000 年度情報処理学会坂井記念特別賞受賞．情報処理学会連続音声認識コンソーシアム代表．情報処理学会，日本音響学会，人工知能学会，言語処理学会，IEEE 各会員．