自然言語処理　143－8
音声言語情報処理　36－3
（２００１．６．１）

話し言葉音声認識のための言語モデルとデコーダの改善

河原 達也　　加藤 一臣　　南條 浩輝　　李 晃伸*

京都大学 情報学研究科
〒 606-8501 京都市左京区吉田本町

*奈良先端科学技術大学院大学 情報科学研究科
〒 630-0101 生駒市高山 8916-5

あらまし

講演音声のような話し言葉を認識するための言語モデルとデコーダ (音声認識プログラム) の改善法について述べる。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考察する。また話し言葉では、ポーズで区切られる単位が言語的な単位と一致しないため、間投詞とあわせて扱いの検討が必要である。特に発話のセグメンテーションが容易でないため、ショートポーズを検出しながら逐次確定していく方式をデコーダ Julius に実装した。10 名の講演音声の認識で評価を行い、平均 65.3%の認識精度を得た。

Language Model and Decoder for Lecture Speech Recogntion

Tatsuya Kawahara, Kazuomi Kato, Hiroaki Nanjo and Akinobu Lee*

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

*Nara Institute of Science and Technology
Ikoma, Nara 630-0101, Japan

Abstract

Language modeling and decoding algorithms for spontaneous lecture speech recognition are addressed. In order to complement training data for the language model, incorporation of other text corpora is explored. Optimization of the text mixture weights is proposed. Handling of pauses as well as fillers is another problem in spontaneous speech recognition. We revised our decoder Julius so that speech segmentation is performed along the recognition process. We have achieved word accuracy of 65.3% on the automatic transcription of real oral presentations.

1 はじめに

話し言葉の認識においては、ディクテーションシステムの方法がそのまま適用できるわけではなく、音響モデル・発音モデル・言語モデル・デコーダの各々において新たな方法論が必要である [1]。実際に、開放的融合研究「話し言葉工学」で構築されている講演音声を中心とした日本語話し言葉コーパス (CSJ: Corpus of Spontaneous Japanese)[2][3] を用いた初期的実験評価 [4] [5] においては、書き言葉・読み上げ音声で作成されたモデルではきわめて低い性能しか得られず、話し言葉のデータでモデル化することの重要性を示したが、それでも 60%程度の認識精度しか得られなかった。この結果は、発音や音響的な変動に対する抜本的なモデル化の見直しを示唆しているが、本稿では言語モデルとデコーダに焦点をおいて、それらの改善について述べる。

言語モデルの構築に際しては、話し言葉に固有のフィラーや言い淀みなどの現象に対処する必要がある。また話し言葉に特徴的な表現・言い回しを含めた統計モデルを学習するためには、話し言葉のコーパスが必要であるが、話し言葉の収集や書き起こしは容易でないため、新聞記事などの書き言葉に比べると、学習テキストは十分ではない。大規模な構築が進められている CSJ コーパスでも、そのテキストサイズは新聞記事数ヶ月分程度であり、現時点では 1ヶ月分程度である。そのため、他の言語資源の有効利用が必要であると考えられ、その際には、関係あるテキストや語彙を効率よく選別し、効果的に混合することが不可欠である。

また、話し言葉においては文やフレーズの区切りが明確になされず、複数の文が続けて発話されたり、逆に文中に頻繁にポーズが置かれたりすることもあるため、ポーズで区切られた入力が一文であるという仮定の下で認識を行うことはできない。実際に 15 分とか 1 時間といった音声データを適切な単位に区分化する作業は容易でないため、デコーダ (認識プログラム) において、認識とセグメンテーションを並行して行うことが望ましい。

本稿では、上記の問題に関して検討を行い、2001 年 2 月時点の CSJ コーパスを用いた認識実験結果について報告する。

表 1: 言語モデル用学習セットの概要

	set-1	set-2	set-3
	2000 年 10 月	2001 年 2 月	2001 年 2 月+
講演数	186	316	612
単語総数	466842	759512	1480834
異なり語数	17172	21381	29939

+: 未チェックの書き起こし (trans1) を含む

表 2: テストセットの概要

講演名 (略称)	単語数	時間 (分)
AS99SEP022 (AS22)	6294	28
AS99SEP023 (AS23)	4391	30
AS99SEP097 (AS97)	2508	12
PS99SEP025 (PS25)	5372	27
JL99OCT001 (JL01)	9833	57
KK99DEC005 (KK05)	6527	42
NL00MAR007 (NL07)	2644	15
SG00MAR005 (SG05)	4460	23
YG99JUN001 (YG01)	2759	14
YG99MAY005 (YG05)	3108	15

(注) 最初の 4 講演は前回の報告 [4] と同一

2 学習セットとテストセット

まず、本稿の実験に使用した学習セットと評価セットについて説明する。言語モデル学習用に用いたデータを表 1 に示す。CSJ コーパスは構築が進行中であり、2000 年 10 月時点に比べて 1.6 倍程度までデータ量が増加した。また前回の報告 [4] では、言語モデル学習にはチェック済のテキスト (trans4) しか用いていなかったが、今回は一通りの書き起こしテキスト (trans1) も利用する。これにより、さらにデータ量が 2 倍程度増強される。

音響モデルの学習には、未チェックの書き起こしテキスト (trans1) も用いている。ただし現在、男性依存モデルしか作成していないため、女性話者のデータは用いていない。また前回の報告 [4] で、一般の人が特定の話題について独話した模擬講演のデータを加えても認識精度に効果がなかったため、音響モデルの学習にはそれらを含めていない。このような条件の下では、音響モデルの学習データはそれほど前回から増えていなかったため、本稿の大半の実験に

表 3: カバレージ (%)

	set-1	set-2	set-3
語彙サイズ	10346	13314	19158
AS22	94.5	95.2	95.8
AS23	95.7	96.6	97.1
AS97	95.9	96.3	97.1
PS25	95.7	96.4	97.6
平均	95.3	96.0	96.8

表 4: テストセットパープレキシティ

	set-1	set-2	set-3
語彙サイズ	10346	13314	19158
AS22	141.2	141.1	133.5
AS23	127.3	122.5	107.5
AS97	140.6	131.1	117.2
PS25	193.0	168.3	164.4
平均	152.8	143.2	134.1

用いた音響モデルは前回の報告のものと同じである。すなわちベースライン音響モデルは、195 講演 (35.3 時間) から学習された男性依存の PTM triphone モデル (64 混合 x129 コードブック; 2000 状態) である。

次に、テストセットを表 2 に示す。このうち最初の 4 講演は前回の報告で用いたものであり、本稿の大半の実験においてもこれらを用いている。いずれも講演に熟練した話者による講演であり、原稿を用いずに話している。また、YG は学会でなく融合研究プロジェクトの会合での発表である。

なお本研究においては、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よるものである。また、単語総数の計数においてポーズを含めていない。

3 学習テキスト量の影響

最初に、単語辞書と言語モデルの学習テキスト量が大きく増えたので、この影響について報告する。

各々の学習セット中で、2 回以上出現した単語で語彙を構成した (カットオフ 1)。これによるカバレージを表 3 に示す。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 万になったが、カバレージも改善されていることがわかる。

表 5: 単語認識精度 (%)

	set-1	set-2	set-3
語彙サイズ	10346	13314	19158
AS22	55.5	57.2	58.4
AS23	68.1	69.3	71.0
AS97	67.8	70.6	72.1
PS25	60.3	64.0	64.9
平均	61.5	63.8	65.1

次に、単語 3-gram モデルを学習した。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。テストセットパープレキシティを表 4 に示す。言語モデルにはポーズ<sil>, <sp> も含められているが、パープレキシティの計算ではそれらは除外している。

語彙が一致していないとパープレキシティの公正な比較は困難であるが、学習テキストの増加に伴って、語彙サイズが大きくなっているにもかかわらず、パープレキシティが大きく (約 12%) 改善されている。

これらの言語モデルを用いて、認識実験を行った結果を表 5 に示す。ベースラインの音響モデルとデコーダ (Julius 3.1) を用いている。ここでは、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。

この結果から、学習データ量の増加に伴って、カバレージとパープレキシティだけでなく、認識精度も着実に改善されていることがわかる。これは、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 種々のコーパスの重み付き混合

前章の結果から、話し言葉の言語モデルを高精度に構築するには、学習データ量がまだ不足していることが考えられる。

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、そのテキストサイズは一億語 (新聞記事 7 年分) 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 [6]。一方、話し言葉は電子化されたテキストがあらかじめ用意されているわけではなく、収録した上で音声を聞き取りながらの書き起こし作業を伴う。

表 6: 収集した Web 講演録

	講演数	単語総数	語彙数
Web 講演録	81	1692802	37462

話し言葉のコーパスとしてはかなり大規模な CSJ コーパスにおいても、そのテキストサイズは新聞記事数ヶ月分程度であり、現時点 (set-3) においては 1 か月分程度である。したがって、他のコーパス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 [7]。

4.1 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 [8]。ここでは、Web 上で公開されている講演録を収集した。講演数は 81 である。この Web 講演録データの出現単語総数・語彙数を表 6 に示す。CSJ コーパスに比べて、講演数は少ないが、出現単語総数は 169 万語で set-3 を若干上回っている。

ここでは、特定の話題への偏りを避けるために、話題との相互情報量に基づいて話題独立な語彙の選択を適用している [9]。

4.2 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あり効果的である [10]。

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。各テキスト中での単語列 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パラメータを $\{\mu_i \mid (i = 1, \ldots, n)\}$ とすると、出現頻度の重み付けによる混合を行った後の単語 w の合計出現頻度 $C(w)$ は、

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

となり、単語履歴を h とすると $P(w|h) = \dfrac{C(hw)}{C(w)}$ に基づいて、統合された語彙やカットオフ条件のもとで、言語モデルの学習及びバックオフスムージングを行う。この手法では、元の学習コーパスが利用できるという条件が不可欠であるが、カットオフの条件の違いなどを考慮する必要 [11] がなく、出現頻度の加え合わせという単純な操作で構築が可能である。

4.3 削除補間法による混合重みの自動推定

重み付き混合は効果的な手法であるが、混合重みの自動最適化が大きな問題である。従来は、重みの値を種々に変化させて、音声認識に用いるテストセットのパープレキシティなどを用いて事後的に決めるのが一般的である。

これに対して、テストセットを用いない混合重みの推定法を考える。そのために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用いる。

ここでは、実際の講演音声認識に最も合致する CSJ コーパスをターゲットとする。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、学習データ量が大きくなくても最大限に活用することができる。

具体的な手順は以下の通りである。

1. あらかじめ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 $U_1, \ldots, U_n$ に分割し、各部分集合 U_j に対して 6 までを行う
3. μ_i の初期値を設定する。
4. 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。
5. U_j を除く残りのマッチドコーパスと大規模コーパスから併合語彙を用いて単語 N-gram を学習する。
6. 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。
7. μ_{ij} を全ての部分集合 U_j について平均し、最尤の μ_i とする。

手順 6 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最適化を行う。

表 7: Web テキスト混合の効果

	Web のみ	set-1	set-1+Web		set-3	set-3+Web	
			1:1 混合	最適化		1:1 混合	最適化
語彙サイズ	8000	10346	13197		19158	20531	
カバレージ (%)	84.6	95.3	95.8		96.8	96.9	
パープレキシティ	236.9	152.8	186.5	170.9	134.1	146.7	139.9
単語認識精度 (%)	50.4	61.5	61.1	63.0	65.1	64.8	65.6

AS22, AS23, AS97, PS25 に対する平均

4.4 Web 講演録との混合によるモデルの評価

ここでは、CSJ コーパスの学習セット set-1 と set-3 の各々について、Web 講演録テキストとの混合を行った。削除補間法を適用するために、CSJ コーパスのテキストを 7 つの集合に分割した。分割した各集合は、同程度のテキストサイズになっており、学会の分類も均等になっている。

混合重みの最適化の結果、set-1 については、CSJ コーパス: 0.93, Web 講演録: 0.16 となった。また set-3 については、CSJ コーパス: 0.95, Web 講演録: 0.066 となった。このように CSJ コーパス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のカバレージ・パープレキシティ・単語認識精度による評価を表 7 に示す。

Web 講演録単独で作成したモデルは、大きく他より劣っている。これは、忠実な書き起こしテキストでないためである。しかし、CSJ コーパスに Web 講演録モデルを混合するとカバレージが若干向上した。パープレキシティが一様に上昇しているが、語彙サイズが大きくなったためである。

両者の混合により作成した言語モデルにおいては、混合重みを変化させても同じ語彙を持つので、パープレキシティは公正に比較できる。混合重みを最適化することで、いずれの話者においてもパープレキシティが低下した。このことから、混合重みの最適化手法の有効性が示された。

認識精度に関しても、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認識精度が低下したが、混合重みの最適化を行うことによりおおむね改善された。特に、set-1 については平均で単語認識精度が 1.5%向上した。set-3 の場合は混合による改善の度合いは小さいが、混合重みが小さくなったこととあわせると、この程度のテキスト量が集まれば、異なる性質のテキストを混合することの意義が薄れていくものと考えられる。

逆に、テキスト量が十分でない場合には、提案手法を用いて、複数のコーパスを利用することは有効である。現実にはタスクやドメインを限定した場合は、CSJ コーパスのように大量のデータを集めるのは困難であり、本手法の応用範囲は広いと考えられる。

さらに大規模なテキストコーパスとして新聞記事の利用も考え、話し言葉調の文を選択して混合することも試みたが、推定された重みはきわめて小さく、全く効果がなかった [12]。

なお、Web 講演録には通常のテキストのように句読点が用いられているので、言語モデル作成に Web 講演録を混合することにより、認識結果に句読点の挿入が可能になるという副次的な効果がある。

5 ポーズと間投詞の透過単語化

話し言葉においては文やフレーズの区切りが明確になされず、複数の文が続けて発話されたり、逆に文中に頻繁にポーズが置かれたりする。

ここでは予備実験の結果 [12]、しきい値 1000msec 以上のロングポーズを始終端<sil>に対応させ、しきい値未満をショートポーズ<sp>に対応させて、言語モデルを構築している。間投詞については、書き起こしに明示的に含まれているので、通常の単語と同様のモデル化が可能である。

しかし、間投詞・ショートポーズが頻出すると、前後の単語の連鎖が分断され、認識精度に影響することが考えられる [13]。そこで、間投詞やショートポーズによる影響を除くため、これらは予測するが単語履歴には含めない透過単語化の処理 [14][15] を導入

表 8: 透過単語処理による単語認識精度 (%)

sp 透過	なし	あり	なし	あり
間投詞透過	なし	なし	あり	あり
AS22	57.9	56.4	56.5	55.6
AS23	67.9	67.3	68.0	67.3
AS97	70.4	68.6	69.4	68.9
PS25	61.4	60.7	59.7	58.6
平均	63.0	61.9	61.9	61.0

言語モデル： set-1+Web (最適混合)

する。単語列 $w_1 w_2 w_{trans} w_3$ が与えられると、通常 w_3 の生起確率は $P(w_3|w_2 w_{trans})$ で計算するのに対して、w_{trans} を透過単語とする場合は $P(w_3|w_1 w_2)$ で与えることで、透過単語 w_{trans} による w_2 と w_3 の単語連鎖の分断を回避することができる。

ここでは、言語モデルの学習の際にはこれまでと同様に間投詞とショートポーズを通常単語として学習し、認識時に間投詞・ショートポーズがコンテキストに含まれる時に、一つ前の単語履歴を用いて透過単語化する手法を試みた。

この認識結果を表 8 に示す。なお、ここで用いた言語モデルは、set-1 と Web 講演録の混合を行ったテキストで学習したものである。

この結果、通常単語と同様に扱う場合と比較して、一様に認識精度が低下した。これは、間投詞やショートポーズに出現するパターンがあり、また後続する単語にも一定のパターンがあることを示している [15]。特に言語モデルの構築に、ショートポーズや間投詞が十分多く含まれている話し言葉コーパスを主に用いた場合には、それらを含めた出現連鎖統計が学習できるので、コンテキストとして無視するよりも利用する方が効果的であることがわかった。

6 セグメンテーションを並行する逐次デコーディング

話し言葉においては、発声の区切れは必ずしも文の単位と一致しない。認識を意識しながら発声がなされるディクテーションシステムと異なり、発話が極端に長くなったり、逆に細切れになったりする。そのため言語モデルだけでなく、デコーダにおいてもポーズに対する対処が必要となる。

実際に、話し言葉音声をデコーダが扱えるようにセグメンテーションするのは容易でない。単純に一定のポーズ長で区分化すると、極端に長い入力とフィラーだけのような短い入力が多数生成される。特に、Julius[16] を含むマルチパスデコーディングの場合、第 1 パスが終了するまで第 2 パスが実行できないため、認識対象の入力音声が長いと認識処理に長い時間がかかるだけでなく、第 2 パスの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もなる。逆に入力が短すぎると、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。2 パス探索における逐次デコーディングについては、履歴が変化しなくなる場合に確定していく方式 [17] や、固定入力長に基づく方式 [18] などがあるが、ここではショートポーズの検出による区分化を用いる。具体的には、デコーダにおいて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。これはセグメンテーションと単語列の認識を同時並行して行うものである。

ショートポーズの検出には、第 1 パスの認識における sp 単語モデルの尤度を用いる。第 1 パス処理中に各フレームの最尤単語仮説を調べて、sp 単語が最尤となるフレームが一定数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 1 に示す。第 1 パスは sp 区間を開始点までさ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 単語を、初期の単語履歴として直前の第 2 パスで確定した sp 以外の単語を割り当てる。

これにより、候補の確定が早くなると同時に、第 2 パスの探索範囲を小さくすることにより、全体的な認識精度の向上が期待できる。ショート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。また、前向き後ろ向き探索に基づくデコーダ Julius においては、sp の前後で 3-gram が適用されず、2-gram しか適用されないことになるが、実際にはショートポーズの前後の 3-gram

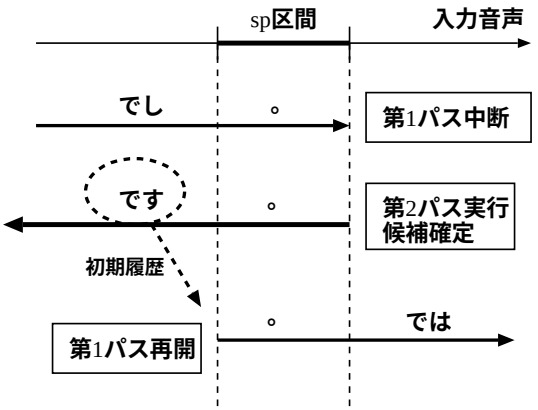

図 1: 逐次デコーディング

エントリは多くないと予想される。

この逐次デコーディングアルゴリズムを Julius-3.2pre において実装し、認識実験を行った。ショートポーズ (sp) の検出のしきい値は、25 フレーム (250ms) に設定している。テストセットのすべての講演に対して評価を行った。従来のセグメンテーションを行ってから 1 つずつ認識を行う方法との比較を表 9 に示す。なおこの際に、音響モデルも 2001 年 2 月時点のデータ (224 講演) で作成し直した。ただし表 5 と同一の講演について比較するとわかるように、音響モデルの改善の効果はわずかである。

この結果、逐次デコーディングの方が高い認識精度を得ることができた。特に、新たにテストセットに追加した 6 講演において改善が大きいが、これは最初の 4 講演についてはファイルのセグメンテーションの際にしきい値を適応的かつ事後的に変化させたのに対して、後の 6 講演では固定のしきい値を採用したためである。後の 6 講演では、特に削除誤りが減少していたが、これは単語の途中などで誤ってセグメンテーションされている場合があったことを示唆している。実際にセグメンテーションのしきい値を調整するのは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。

7 おわりに

講演音声のような話し言葉音声を認識するための言語モデルとデコーダの改善法について述べた。

表 9: 逐次デコーディングの効果

	従来の認識法	逐次デコーダ
AS22	58.9	60.1
AS23	72.4	71.9
AS97	72.5	73.8
PS25	64.7	65.2
JL01	58.6	64.8
KK05	62.2	66.8
NL07	65.1	69.0
SG05	54.8	57.4
YG01	59.7	63.3
YG05	64.2	68.0
平均	62.2	65.3

言語モデル: set-3, 新しい音響モデル

言語モデルに関しては、話し言葉コーパスのデータ量不足を補うために、他のコーパスと効果的に混合する手法を提案した。話し言葉のテキストがかなり少ない場合は有効であったが、同程度まで収集されると混合の効果が小さくなることがわかった。また、話し言葉に特徴的なショートポーズや間投詞を透過単語として扱う方法も検討したが、十分な量の話し言葉コーパスがある場合は逆効果であることがわかった。

話し言葉では文の区切り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。本手法は、事前にセグメンテーションを行う必要がなく、また認識精度の改善につながった。

今後は音響モデル・発音モデルの検討 [19] を進めて、音声認識精度の改善を図っていくとともに、話し言葉を理解するという観点からも研究を進めていきたい。

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、ご協力を頂いた関係各位に感謝いたします。

参考文献

[1] 河原達也. (サーベイ) 話し言葉音声認識の概観. 電子情報通信学会技術研究報告, SP2000-95, NLC2000-47 (SLP-34-21), 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, 2001-SLP-36-1, 2001.

[4] 加藤一臣, 南條浩輝, 河原達也. 講演音声認識のための音響・言語モデルの検討. 電子情報通信学会技術研究報告, SP2000-97, NLC2000-49 (SLP-34-23), 2000.

[5] 篠崎隆宏, 斎藤洋平, 堀智織, 古井貞熙. 話し言葉音声の認識を目指して. 情処学研報, 2000-SLP-34-22, 2000.

[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) の性能評価. 情報処理学会研究報告, SLP-31-2, NL-137-7, 2000.

[7] 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, Vol. J83-DII, No. 11, pp. 2473–2480, 2000.

[8] 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, 2001-SLP-35-8, 2001.

[9] 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. 情報処理学会研究報告, SLP-26-2, 1999.

[10] 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, Vol. J83-DII, No. 11, pp. 2418–2427, 2000.

[11] 長友健太郎, 西村竜一, 小松久美子, 黒田由香, 李晃伸, 猿渡洋, 鹿野清宏. 相補的バックオフを用いた言語モデル融合ツールの構築. 情処学研報, 2001-SLP-35-9, 2001.

[12] 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.

[13] 南條浩輝, 加藤一臣, 三村正人, 李晃伸, 河原達也. 種々のタスクにおける大語彙連続音声認識システムの性能評価と診断. 情報処理学会研究報告, SLP-31-11, 2000.

[14] A.Stolcke and E.Shriberg. Statistical language modeling for speech disfluencies. In *Proc. IEEE-ICASSP*, pp. 405–408, 1996.

[15] 伊東伸泰. 自由発話向け言語モデルの現状と課題. 音講論, 1-3-11, 春季 2001.

[16] 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. 電子情報通信学会論文誌, Vol. J82-DII, No. 1, pp. 1–9, 1999.

[17] 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, 99-SLP-29-37, 1999.

[18] 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, 2000-SLP-34-18, 2000.

[19] 南條浩輝, 河原達也. 発話速度に依存した音素・音節モデル化の検討. 話し言葉の科学と工学ワークショップ, pp. 93–100, 2001.