『日本語話し言葉コーパス』を用いた音声認識の進展

河原 達也

京都大学 学術情報メディアセンター
〒 606-8501 京都市左京区吉田二本松町

あらまし

『日本語話し言葉コーパス』(CSJ) を用いた音声認識の進展について報告する。話し言葉音声においては、言語的／音響的な特徴が書き言葉／読み上げ音声と大きく異なり、形態素や発音の変動も問題となるが、そのような点を考慮した話し言葉の音響・言語モデルを構築する上で、この大規模なコーパスがきわめて有用であった。さらに、これらのモデルを話者に教師なしで適応し、また発話速度に応じたデコーディングを行うことで認識精度が大きく改善し、単語認識率で 78.0%を達成した。

Progress of Speech Recognition using the Corpus of Spontaneous Japanese (CSJ)

Tatsuya Kawahara

Kyoto University
Academic Center for Computing and Media Studies
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

The report gives an overview of the current state of spontaneous speech recognition using the “Corpus of Spontaneous Japanese (CSJ)”. It is shown that the large-scale corpus had strong impact in training acoustic and language models considering morphological and pronunciation variations which are characteristic to spontaneous Japanese. Unsupervised adaptation of these models and the speaking rate is also effective, and we have achieved word accuracy of 78.0%, which is a significant improvement over a couple of years.

1　はじめに

大語彙連続音声認識は、ディクテーションシステムに代表されるように、書き言葉の読上げ音声に対しては 90%を上回る認識精度を達成している。しかしながら、音響的・言語的な変動が大きな話し言葉の音声認識は、依然として大きな未解決の問題である。そのため、科学技術振興調整費 開放的融合研究「話し言葉工学」プロジェクト [1][2] において、話し言葉の大規模なコーパスの構築と本格的な話し言葉の音声認識・理解の研究を推進してきた。

本プロジェクトの主要な成果である『日本語話し言葉コーパス』(CSJ)[3][4] は、おおそ 7 百万形態素、500 時間の講演音声とその書き起こしから構成され、話し言葉の音声データベースとしては世界最大規模のものである。

講演や口頭発表は、米国の ARPA プロジェクトで主な対象とされている放送ニュースと電話での会話の中間として捉えることができる。すなわち、放送ニュースのように、原稿を読み上げているわけではないが、ある程度リハーサルや事前の準備を行った上で話している。また、アナウンサーのようにプロの話し手ではないが、公共の場で聴衆に向かって話すので、通常の会話に比べれば発話スタイルも丁寧である。また電話での会話に比べて、SN 比はよいものの、話題が専門的で語彙やパープレキシティが大きいという特徴がある。

多くの従来研究で指摘されているように、話し言葉音声においては、発話速度が速く、発音が不明瞭であるために音響的な変動が大きい。また、口語的な表現や言い淀みなどの現象により言語的な変動も大きい。これらの問題は、音声認識における音響モデル、発音モデル及び言語モデルのすべてにおいて考慮する必要がある [5]。

大規模な CSJ によって、これらの精密な統計モデルを構築することが可能になった。本稿では、主に京都大学で CSJ を用いて進めてきた音声認識 [6] の進展について報告する。

2　逐次デコーディング

まず、認識エンジン Julius [1] を講演のような長い話し言葉音声に対応できるように改良した [7][8]。

話し言葉においては、ポーズが任意の場所に挿入されるため、発話が言語的な文と必ずしも一致しない。また、文末等で発声が明瞭でなかったり、発話速度が変動するために、発話の切出し自体が容易でない。

そこで、認識前に発話の切出しを必要としない逐次デコーディングアルゴリズムを実装した。これは、音響モデルと言語モデルを用いてポーズの判定を行い、認識と切出しを同時並行的に行うものである。ポーズモデルが最尤となるフレームが継続すれば、前向きの探索を打ち切り、後ろ向きの探索を実行してその時点までの認識結果を確定する。そして、その履歴を利用して以降の認識処理を続行する。

ショートポーズの検出を音響・言語モデルを用いて行うために、パワーや零交差数に基づいて判定するより信頼性が高いと考えられる。また、モデルのオンラインの適応も可能である。

実際にこの逐次デコーディングにより、講演全体を事前に区分化することなく認識処理することができ、事前に発話切出しを行う場合と比べて同等以上の認識精度を実現した。

3　音響モデル

CSJ の大部分は学会講演と模擬講演の 2 種類から構成される。この講演種や性別を考慮して、種々のベースライン音響モデルを構築した [9]。学会講演の方が発話速度が速いなど、講演種によって発話スタイルに違いがみられ、実際に学会講演に対しては学会講演のみで学習した音響モデルが最も高い認識率を得た。また学会講演の大多数が男性であったので、本稿では男性の学会講演を対象とする。テストセットは表 1 に示す 15 講演、学習データは 781 講演、106 時間である。

音響モデルは対角共分散の HMM である。音響特徴量は 12 次までの MFCC とその一次回帰係数 (ΔMFCC)、及びパワーの回帰係数 (Δpower) の計 25 次元である。音素数は 43 であり、主に PTM (Pho-

[1] `http://julius.sourceforge.jp` からダウンロード可能

表 1: 本稿で用いるテストセット講演

講演 ID	単語数	時間 (分)	PP	RR	SR	OOV
A01M0097	2592	14.4	39.6	0.62	8.65	0.21
A04M0051	2581	14.6	52.5	2.02	8.03	0.57
A04M0121	2964	15.4	87.1	3.31	8.77	1.05
A03M0156	3243	16.6	86.3	0.90	9.78	1.02
A03M0112	3254	15.1	53.8	0.52	9.62	0.47
A01M0110	1307	9.6	81.5	2.07	8.25	0.55
A05M0011	2791	16.0	72.8	1.33	6.90	1.27
A03M0106	3091	13.0	72.8	1.07	10.66	2.75
A01M0137	2073	11.1	56.0	1.69	8.64	1.01
A04M0123	2619	12.3	52.1	0.84	9.06	1.28
A01M0056	2364	11.4	41.6	0.85	8.77	0.53
A02M0012	4034	22.2	95.8	0.45	8.45	2.28
A06M0064	2399	12.5	81.2	0.33	7.30	1.05
A01M0141	2334	15.4	72.4	2.37	8.45	0.92
A03M0016	3171	15.7	61.5	1.82	10.20	1.60
合計/平均	40817	215.3	65.8	1.29	8.77	1.18

PP: 単語パープレキシティ，RR: 言い直し率 (%)，
SR: 発話速度 (mora/sec)，OOV: 未知語率 (%)

表 2: 言語モデル学習データ量の効果

	LM1	LM2	LM3	LM4	現在*
講演数	186	316	612	1125	2592
テキストサイズ	0.5M	0.8M	1.5M	2.7M	6.3M
語彙サイズ	10K	13K	19K	21K	24K
OOV (未知語率)	4.7	4.0	3.2	3.0	1.5
PP (パープレキシティ)	152.8	143.2	134.1	115.4	105.6
WER (単語誤り率)	38.5	36.2	34.9	34.5	33.?

音響モデルも最新のものでなく、テストセットも異なっているので、以降の表とは直接的な比較ができない
* 現在のモデルは形態素体系が異なるので、以前のものと直接比較ができないが、数値は推定したものである

netic Tied Mixture) triphone モデル [10] を用いている。これらは基本的に、「日本語ディクテーションソフトウェア」[11] で用いていたものと同じである。ただし学習データ量の増加に伴い、混合分布数は 192 まで大きくし、全体として 25K 個のガウス分布、576K 個の混合重みを有するものになっている。

実際に学習データ量の増加につれて、認識精度においても混合分布数を増やす効果が確認できた。例えば、学習データが 38 時間の際は 64 混合のモデルが最良で、WER(単語誤り率) が 35.8%であったが、60 時間分で学習した 192 混合のモデルにより 34.7%に改善した。参考のため、読上げ音声コーパスで学習した IPA モデルでは 10%以上低い認識精度であった。

4 言語モデルと発音モデル

ベースラインの言語モデルは単語 trigram モデルであり、学会講演と模擬講演すべてを含めた (テストセットの話者の講演を除く)2592 講演から学習した。テキストサイズは 6.67M 形態素である。形態素解析は、CSJ に人手で付与された短単位のものを基にして、最大エントロピ法で学習されたシステム [12] を用いた。

まず学習データ量の効果を表 2 に示す。学習データ量の増加につれて、単語認識精度が着実に改善していることがわかる。当初は Web 上から収集した講演録を学習データに混合する効果があったが、LM3 以降ではみられなかった [7]。この結果は、話し言葉をモデル化する上でこの規模のコーパスが非常に有用であったことを示すものである。

日本語の話し言葉においては発音の変動が大きい

表 3: 発音モデルの効果

手法	WER
単一の発音エントリのみ	31.6
発音変形エントリ登録 (確率なし)	31.4
発音 unigram (pron-unigram)	30.7
発音 trigram (pron-trigram)	30.5

ので、発音辞書に多くの変形エントリを記述する必要がある。CSJ の書き起こしはかな漢字まじりの正書法と発音に忠実なかな表記の両方で行われており、形態素単位で両者の対応付けを行うことにより、発音変形エントリを自動抽出することができる。ただし、解析誤りを含めて頻度の少ないものもすべて登録するとかえってモデルの精度が悪くなるので、一定の頻度以下の発音エントリは削除している。その結果、語彙エントリ総数 24437 に対して、発音エントリ総数は 30820 となった。

発音変動を統計的にモデル化するのに、単語毎に発音エントリの unigram 生起確率を求める方法 (pron-unigram) と、発音エントリの単位で trigram を構成する方法 (pron-trigram) が考えられる。両者も含めて、発音モデルの比較を表 3 に示す。この結果から、発音変形を統計的にモデル化することの効果がわかる。また、発音エントリの unigram より trigram の方が、若干ではあるが高い認識精度を実現した [13]。

5 モデルの適応と発話速度依存デコーディング

次に、音響モデルと言語モデルの話者適応を行った。講演音声は話者毎に比較的長時間のデータであるので、教師なし適応の枠組みも有効である。

まず、ベースラインの話者独立のモデルによる音声認識結果から初期的な書き起こしを生成する。音響モデルについては、この音素ラベルからガウス分布の MLLR 適応を行い、話者適応モデルを作成する。

言語モデルについても、話者や話題に適応する方式を検討した [13]。まず初期の認識結果から言語モデルを作成し、このモデルによるパープレキシティに基づいて学習コーパスから当該講演に類似したテキストを選択する。同様に tf-idf 尺度に基づいて類似テキストを選択する。前者は主に話者の言い回しを考慮し、後者は主に講演の話題に着目したものになっている。これらの類似テキストに重みをつける形で言語モデルを再構築する。さらに、初期認識結果から作成した言語モデルとも補間混合することにより、言語モデルを適応する。表 4 に示すように、これらのすべての適応手法及び組合せの効果が確認された [14]。

表 4: 言語モデル適応の効果

手法	WER	PP
(0) ベースライン	30.5	74.9
(1) パープレキシティに基づくテキスト選択	29.7	68.7
(2) tf-idf に基づくテキスト選択	29.1	70.2
(1)+(2) テキスト選択の統合	28.8	65.1
(3) 認識結果の直接利用	28.8	51.8
(1)+(2)+(3) テキスト選択+認識結果	27.6	46.7

WER: 単語誤り率, PP: パープレキシティ

表 5: モデル適応と発話速度への適応の効果

手法	WER
ベースライン	30.9
+ 音響モデル適応	26.0
+ 言語モデル適応	23.9
+ 発話速度依存デコーディング	22.0

さらに、発話速度に対しても適応的にデコーディングを行う方式を提案した [15][16]。話し言葉、特に講演音声では、発話速度が全般に速く、また変動も大きい。発話速度が速い区間と遅い区間では、認識誤りの傾向も異なっている。そこで、発話速度を自動推定し、それに応じて音響分析、音響モデル、デコーディングパラメータを切り替える。具体的には、話速が速い場合には、分析フレーム長を短くし、音節単位のモデルを併用する。一方、話速が特に遅い発話については、デコーディング時の挿入ペナルティ値を大きくする。この発話速度依存デコーディングにより、さらに認識精度が改善した。

以上の手法の統合効果を表 5 に示す。音響モデルの教師なし適応により 4.9%もの改善が得られた。言語モデルの教師なし適応と、発話速度に適応したデコーディングも相乗的な効果が得られ、それぞれ 2.1%と 1.9% の認識精度の向上を実現した。最終的に、単語認識精度は 78.0%(WER:22.0%) となった。

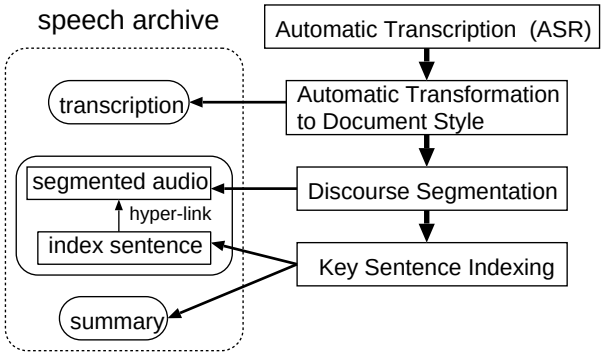

図 1: 講演アーカイブシステムの構成

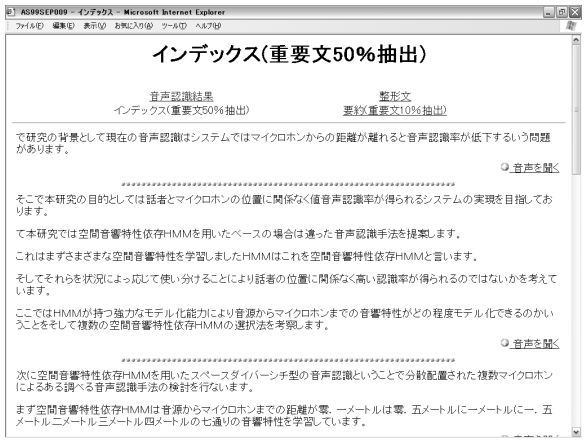

図 2: 講演アーカイブブラウザの概観

6 まとめ

本稿では、『日本語話し言葉コーパス』(CSJ) を用いた講演音声の自動認識の概要について述べた。このような大規模コーパスが話し言葉の音響的・言語的なモデル化において非常に有用であったこと、そしてこれらのモデルを話者適応することの有効性を確認した。

単語認識精度は 80%程度であり、自動書き起こしの精度としてはまだ十分でないかもしれないが、重要文抽出においては正しい書き起こしを用いる場合と同程度の精度が得られており [17][18]、講演の自動インデキシング・アーカイブ化には応用できると考えられる。

そこで、図 1 に示すような講演アーカイブシステムを試作している。これは、(1) 自動書き起こし、(2) 自動整形と文分割 [19]、(3) セクション分割 [20]、(4) 重要文抽出 [17]、の４つのモジュールから構成される。図 2 に示すように、セクション毎に重要文を見ながら、当該区間の音声を聞くことができるようになっている。

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。プロジェクトを主導し、また数々のアドバイスを頂きました古井貞熙先生（東工大)、形態素解析と発音付与に尽力頂きました内元清貴さん（CRL)、山田篤さん（ASTEM）をはじめとして、ご協力を頂いた関係各位に深い感謝の意を表します。

参考文献

[1] 古井貞熙, 前川喜久雄, 井佐原均. 科学技術振興調整費開放的融合研究推進制度 – 大規模コーパスに基づく「話し言葉工学」の構築 –. 音響誌, Vol. 56, No. 11, pp. 752–755, 2000.

[2] S.Furui. Recent advances in spontaneous speech recognition and understanding. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 1–6, 2003.

[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] K.Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12, 2003.

[5] 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.

[6] 南條浩輝, 加藤一臣, 李晃伸, 河原達也. 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. 電子情報通信学会論文誌, Vol. J86-DII, No. 4, pp. 450–459, 2003.

[7] 河原達也, 加藤一臣, 南條浩輝, 李晃伸. 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. 情報処理学会研究報告, SLP-36-3, 2001.

[8] T.Kawahara, H.Nanjo, and S.Furui. Automatic transcription of spontaneous lecture speech. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001.

[9] T.Kawahara, H.Nanjo, T.Shinozaki, and S.Furui. Benchmark test for speech recognition using the Corpus of Spontaneous Japanese. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 135–138, 2003.

[10] 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, 2000.

[11] 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.

[12] K.Uchimoto, C.Nobata, A.Yamada, S.Sekine, and H.Isahara. Morphological analysis of Corpus of Spontaneous Japanese. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 159–162, 2003.

[13] 南條浩輝, 河原達也, 山田篤, 内元清貴. 講演音声認識のための言語モデルの教師なし適応. 電子情報通信学会技術研究報告, SP2002-152, NLC2002-75 (SLP-44-32), 2002.

[14] H.Nanjo and T.Kawahara. Unsupervised language model adaptation for lecture speech recognition. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 75–78, 2003.

[15] 南條浩輝, 河原達也. 発話速度に依存したデコーディングと音響モデルの適応. 電子情報通信学会技術研究報告, SP2001-103, NLC2001-68 (SLP-39-20), 2001.

[16] H.Nanjo and T.Kawahara. Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. In *Proc. IEEE-ICASSP*, pp. 725–728, 2002.

[17] 南條浩輝, 北出祐, 河原達也. 談話標識の統計的選択に基づいた CSJ の講演からの重要文抽出. 電子情報通信学会技術研究報告, SP2003-125, NLC2003-62 (SLP-49-13), 2003.

[18] T.Kawahara, K.Shitaoka, T.Kitade, and H.Nanjo. Automatic indexing of key sentences for lecture archives. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003.

[19] 下岡和也, 南條浩輝, 河原達也. 講演の書き起こしに対する統計的手法を用いた文体の整形. 自然言語処理, (採録決定), , 2004.

[20] 長谷川将宏, 秋田祐哉, 河原達也. 談話標識の抽出に基づいた講演音声の自動インデキシング. 情報処理学会論文誌, Vol. 43, No. 7, pp. 2222–2229, 2002.