社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59（５）
2005／12／21

日本語話し言葉コーパスを用いた発音変形依存モデルによる講演音声認識の性能評価

阿部拓也　草間隆　武田千春　加藤正治　小坂哲夫
好田正紀

山形大学
E-mail：esk23329@dip.yz.yamagata-u.ac.jp

あらまし　近年，自然発話に近い大規模な音声データベースを用いて，大語彙連続音声認識が研究されている．本論文では，2004 年 6 月に公開された「『日本語話し言葉コーパス：CSJ』最終公開版」の書き起こしテキストを用いて，音声に忠実な読みを持つ形態素解析データを作成し，その解析データから構築した言語モデル（これを，発音変形依存モデルと呼ぶ）の性能評価を行う．さらに，音響モデル，言語モデルに対して評価セットの認識結果を用いて教師なし適応を繰り返し行い，その性能について評価する．公開版 CSJ の testset1 により評価した結果，発音変形依存モデルを用いて平均 19.96 %の単語誤り率を達成し，教師なし適応を繰り返し行うことで，最終的に平均 15.41 %の単語誤り率を達成した．

キーワード　講演音声認識，発音変形依存モデル，N-gram，教師なし適応，日本語話し言葉コーパス

Performance evaluation of lecture speech recognition by pronunciation variant model using Corpus of Spontaneous Japanese

Takuya ABE, Takashi KUSAMA, Chiharu TAKEDA, Masaharu KATOH, Tetsuo KOSAKA, and Masaki KOHDA

Yamagata University
E-mail：esk23329@dip.yz.yamagata-u.ac.jp

Abstract　In recent years, many aggressive approaches for large vocabulary continuous speech recognition system trained on large-scale spontaneous speech database have been investigated. In this paper, we introduce a method of language modeling based on morphological analysis data designed for pronunciation variant. The method was evaluated on the Corpus of Spontaneous Japanese (CSJ) and reduced the word error rate (WER) by about 5 % absolute. In addition, unsupervised adaptation of both acoustic and language models was introduced to improve the recognition performance further. The results showed the decrease in WER from 19.96 % without adaptation to 15.41 % with unsupervised adaptation.

Key words　lecture speech recognition, pronunciation variant modeling, N-gram, unsupervised adaptation, Corpus of Spontaneous Japanese

1. はじめに

大語彙連続音声認識（Large Vocabulary Continuous Speech Recognition:LVCSR）の中でも，新聞記事等のあらかじめ用意されたテキストの読み上げ音声ならば，単語誤り率（WER）５%以下の高い認識性能がすでに達成されている．しかしながら，言い直し，言い淀み，繰り返し，間投詞や未知語，不正確な発音等の現象を多く含んだ自然な話し言葉を認識しようとすると，認識性能が大幅に低下してしまうのが現状である．

音声認識の要素技術が向上し，読み上げ文の音声認識の次の研究目標は，話し言葉の音声認識である．話し言葉の音声認識は，実用性が高いと同時に，その実現には多くの未解決の課題があり，研究課題として挑戦的で学術的意義が大きい．

近年，LVCSR システムの研究が進展している．DARPA プログラムの参加機関では，2000 時間以上の大量の音声データが利用可能になり，音素文脈の拡大（triphone → quinphone → septaphone）や最小音素誤り（MPE）基準に基づく識別学習による音響モデルの精密化等が進展して，会話電話音声で数年前の WER30 %程度から現在は WER20 %を切るまでに認識性能が上がっている [8] [9] [10].

国内では，自然発話に近い大規模な音声データベースとして，2001 年に「『日本語話し言葉コーパス（CSJ）』モニター版 2001」（以後，モニター版 CSJ），2004 年 6 月に「『日本語話し言葉コーパス：CSJ』最終公開版」（以後、公開版 CSJ）が公開されてから，それを利用した LVCSR の研究が盛んに行われている [3] [4] [5]. 文献 [11] では，WFST（重み付き有限状態トランスデューサ）を用いてフル共分散ガウス分布，quinphone 音

響モデルを実現し，講演音声で WER20 %を切る認識性能を得ている．

我々は，モニター版 CSJ の書き起こしテキストを用いて音声に忠実な読みを持つ形態素解析データを作成し，その解析データから構築した言語モデル（これを，発音変形依存モデルと呼ぶ）を提案し，その有効性を確認した [7]．しかしながら，モニター版 CSJ で利用できる学習テキストは約 95 万語と少なく，十分な精度のモデルが作成できているとは言えなかった．本研究の目的は，公開版 CSJ を用いて，文献 [7] で提案した発音変形依存モデルの性能評価を行うことである．

2. 日本語話し言葉コーパス（CSJ）

日本語話し言葉コーパス（CSJ）[2] は，現代日本語の自発音声を種々の研究用付加情報とともに大量に格納したデータベースである．

2001 年 8 月に約 86 時間の音声と書き起こしテキストからなるモニター版 CSJ が公開され，講演音声認識において様々な研究がなされてきたが，モニター版 CSJ の総単語数は約 95 万語で，話し言葉の学習に十分なデータベースとは言えなかった．

2004 年 6 月に約 660 時間の音声と書き起こしテキストからなり，総単語数が約 750 万語の公開版 CSJ が公開された．公開版 CSJ はモニター版 CSJ の約 8 倍のデータ量であり，話し言葉の学習に十分なデータベースである．本研究では，公開版 CSJ を用いる．

2.1 収録音声

公開版 CSJ には，学会講演・模擬講演・朗読・対話・その他の 5 つのタイプの音声が収録されている．そのうち学会講演と模擬講演が約 90 %を占める．本研究では，学会講演と模擬講演のデータを用いる．

学会講演は，理工学，人文，社会の 3 領域の学会における研究発表のライブ録音である．講演時間は 10 分から 25 分程度が大半であるが，1 時間前後に及ぶ特別講演も少数含まれる．学会講演の多くをしめる理工学系の学会では，男性の大学院生の発表が多いので，学会講演の話者は，年齢と性別に偏りがある．発話スタイルは概してあらたまり度が高い．

模擬講演は，できるだけ年齢と性別のバランスをとった一般話者による，日常的話題についての講演である．話者の大部分は人材派遣会社からであり，あらかじめ指定された一般的テーマ（例えば「人生で一番嬉しかったこと」「人生で一番悲しかったこと」「私の住んでいる街」等）に基づいて，具体的な講演内容を決めた，1 講演 10〜15 分程度のスピーチである．発話スタイルは概して学会講演よりもくだけたものとなっている．

2.2 書き起こしテキスト

書き起こしテキストの一部を表 1 に示す．それぞれの行は以下のように区別される．

(1) 情報部

通し番号 (4 桁の数字) や時間情報 (秒単位) などが記されている．情報部の後には，その単位の発話内容が表記される．笑い声や雑音やベルなどの言語音以外の音の場合は，その情報が情報部の末尾に表記される．表 1 の 19,20 行目にある＜咳＞や＜雑音＞がこれに相当する．

(2) 音声部

音声データを書き起こしたものが表記されている．基本形と発音形からなる．

(a) 基本形（&の左側）

漢字・仮名を中心に表記したもの．発音形に比べて可読性が高い．同一の語句の表記が揺れないよう，基準を構築している．

(b) 発音形（&の右側）

片仮名を用いて実際に発音された音を忠実に記録したもの．自発音声は，発音の怠けや言い間違いなどが頻繁に生じる．発音形では，実際の発音をできるだけ忠実に記録する．

また，CSJ では情報部と音声部の一まとまりを転記基本単位と呼び，以下のように定義している．

転記基本単位

情報部と音声部の一つの組．原則として，言語音が 200 ミリ秒以上の途切れがなく連続して生じている区間を転記基本単位とする．ただし，言語的な文末形式 (述語の終止形や終助詞など) が存在している場合には，50 ミリ秒以上 200 ミリ秒未満の途切れであってもその文末形式の後までを転記基本単位とする．

表 1 書き起こしテキストの例 (一部抜粋)

0001 00000.467-00001.368 L:	
(R ××) です	& (R ××××) デス
0067 00132.520-00134.291 L:	
＜ FV ＞一つの	& ＜ FV ＞ヒトツノ
要因かと	& ヨーインカト
思います	& オモイマス
0412 01005.930-01010.681 L:	
(F えー)	& (F エー)
紀元三世紀ぐらいまで	& キゲンサンセーキグライ (W マエ; マデ)
作られたという	& (W (? ツクラエアエッ); ツクラレ) タト (W ユ; ユウ)
(D ごろっ)	& (D ゴロッ)
%TYPE=O_HYO「五六百」の言い淀みの可能性あり．	
0415 01013.619-01015.621 L:	
大体	& ダイタイ
(F あのー)(F おー)	& (F アノー)(F オー)
0434 01053.629-01053.914 L:＜咳＞	
0435 01054.450-01054.574 L:＜雑音＞	

3. LVCSR システム

本研究で用いる LVCSR システムの構成を図 1 に示す [1]．

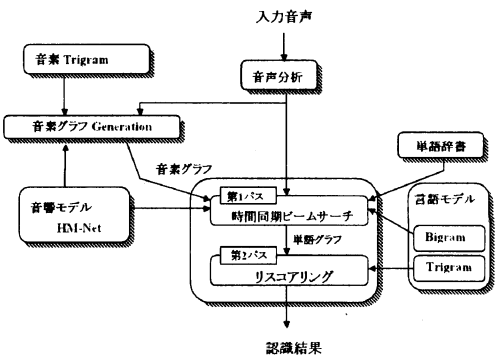

図 1 LVCSR システム

3.1 音声分析

音声分析条件を表 2 に示す．

表 2 音声分析条件

標本化周波数	16kHz、量子化：16bit
分析窓	ハミング窓
フレーム長	25msec
フレーム周期	8msec
高域強調	$1-0.97z^{-1}$
特徴ベクトル	1〜12 次の MFCC と対数パワー、及びその 1 次と 2 次の回帰係数（計 39 次元）
⊿窓幅	7 フレーム
正規化	発話毎のケプストラム平均正規化（CMN）

3.2 音響モデル

音響モデルの作成には，前後の音素文脈に依存する triphone 音素モデルを用いる．音素モデルは隠れマルコフモデル (Hidden Markov Model : HMM) で表現される．

隠れマルコフ網 (HMnet) は 1 つの状態を複数の HMM 間で共有させることで，1 状態あたりの学習データ数を増加させ，統計的に信頼性の高いモデルとしたものである．

あとの実験では，モニター版 CSJ で作成した HMnet の構造を利用して，各状態の出力確率分布などのパラメータのみを公開版 CSJ の学習データで再推定する．

3.3 言語モデル

言語モデルは発音変形依存の単語 bigram 及び trigram である．バックオフは Witten-Bell discounting を使用する．言語モデルの作成には palmkit [6] を用いた．

3.4 単語辞書

学会講演から上位 2 万語，模擬講演から出現回数 2 回以上の単語をそれぞれ選択したものを合わせて語彙リストとし、単語辞書を作成する．

3.5 デコーダ

デコーダでは，第 1 パスで triphoneHMnet 及び単語 bigram を用いて単語グラフを生成し，第 2 パスで単語 trigram を用いて単語グラフをリスコアする，2-Pass サーチの手法を用いる．

第 1 パスでは，音素グラフに基づく仮説制限を行って処理時間の削減を図っている [1]．また，第 1 パスの中で単語間の音素環境を考慮する．

なお，評価セットの講演は 1 文が長いため，あらかじめ適当な単位に分割して認識を行う．ここでは「5 秒経過直後の 200msec 以上の無音区間」で分割する．分割された単位の実際の時間は 6 秒〜8 秒程度で，転記基本単位が約 3 個含まれる．これは，文の単位と必ずしも一致しない．

4. 発音変形依存モデル

4.1 発音変形を考慮した形態素解析データ

話し言葉音声では，文法的に曖昧で，口語的な表現が多彩に出現し，間投詞や言いよどみ，発音の怠けなど，認識を困難にさせる要因が多い．さらに，一つの単語に対して読みが一意に定まらないこともあるため，一般的な形態素解析システムから得られる読み付き形態素テキストでは不十分である．そこで，言語モデルと単語辞書に対して，実際の話し言葉音声に含まれる発音変形などの様々な情報を正確に学習させることで，言語的情報に音声情報を組み合わせてモデリングする．本研究ではこの**発音変形依存モデリング** [7] を用いる．

一般に朗読調の音声の認識では，単語増加が許されるなら複合語など長い単語を形態素として用いた方が認識性能が向上する．しかし，話し言葉においては複合語のような長い単語では，単語内で発音の怠けやショートポーズが起こる確率が高くなる．また，ショートポーズは複合語内の単語の切れ目に偏在すると予想される．そこで，長い単語は分割し，可能な限り短い単語で語彙を構成する．

CSJ 書き起こしテキストは，基本形 (漢字・仮名で綴った表記に揺れがない文) と発音形 (音声を忠実に片仮名で綴った正解文) で成り立っている．

ChaSen と CSJ 書き起こしテキストの基本形及び発音形を利用し，音声に忠実な読みを持つ言語テキストを作成する．CSJ 書き起こしテキストの基本形と発音形を利用する形態素解析データ作成の流れ図を図 2 に示す．

言語テキスト作成の具体的な手順を以下に示す．

＜言語テキスト作成手順＞

（Ⅰ）基本形を ChaSen で形態素解析．

ここで簡単な形態素解析誤りを自動修正．

e.g. 表+ヒョー+2　われる+ワレル+47/6/1
→ 表れる+アラワレル+47/6/1

（Ⅱ）発音変形を考慮した読みを持つ解析データを以下の手順で作成．

ⅰ）読みを複数持つ形態素は発音形に基づいて一意に決定．これにより実際に発声された読みに変換．

e.g. 一日+イチニチ/ツイタチ+2
→ 一日+ツイタチ+2

ⅱ）次のような発音変形を持つ形態素の読み部は（Ⅰ）の形態素解析データと発音形のマッチングに基づいて決定．

a）母音の長音化・短音化・脱落・挿入．

e.g. 音声+オンセイ+2
→ 音声+オンセー+2

b）子音の脱落・濁音化・清音化．

e.g. 開発+カイハツ+2
→ 開発+カイアツ+2

c）音節の促音化・拗音化・撥音化．

e.g. 脱落+ダツラク+2
→ 脱落+ダツラッ+2

d）発音の怠けなどの変化．

e.g. けれども+ケレドモ+58
→ けれども+ケードモ+58

e.g. そういう+ソーユウ+57
→ そういう+ソユ+57

ⅲ）上記以外の発音変形をもつ形態素の読みは，文字単位で存在し得る全てのパターンのマッチングに基づいて決定．

e.g. 声質+セイシツ+2 →声質+コワタチ+2

この場合では，[声] と [質] が持つ全ての読みの組み合わせに関してマッチング．

[セイシツ][セータチ][コエシツ][コワタチ]...

（Ⅲ）複合語のような長い単語を可能な限り短い単語に分割．

e.g.「という」→「と　いう」

e.g.「だとすれば」→「だ　と　すれ　ば」

e.g.「バリ島」→「バリ　島」

e.g.「国勢調査」→「国勢　調査」

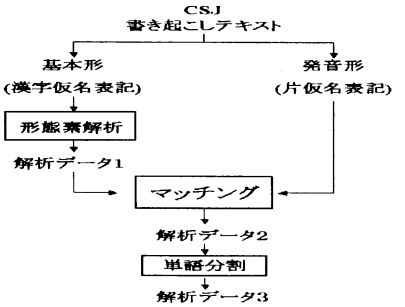

図 2　発音変形を考慮した形態素解析

4.2 言語モデリング

学会講演と模擬講演は異なるタスクとみなして言語モデリングする．具体的には，学会講演の N-gram 出現回数のみ数倍し，模擬講演の N-gram 出現回数と加算することで，全体の N-gram 出現回数を求める．この重み付き言語モデリングの確率推定を次式に示す．

$$P(\omega_i|\omega_{i-2}\omega_{i-1}) = \frac{V * N_A(\omega_{i-2}\omega_{i-1}\omega_i) + N_S(\omega_{i-2}\omega_{i-1}\omega_i)}{V * N_A(\omega_{i-2}\omega_{i-1}) + N_S(\omega_{i-2}\omega_{i-1})}$$

ここで，$N_A(\ )$，$N_S(\ )$ はそれぞれ，学会講演，模擬講演の出現回数，V は重み係数である．あとの実験では，$V=3$ とした．

このようなモデリングにより，模擬講演のデータを利用しながら学会講演の認識により有効な言語モデルを作成できる．

4.3 N-gram 作成手順

言語モデルは以下の手順で作成する．

① まず，学習テキストを学会講演と模擬講演に分け，それぞれの単語頻度リスト，語彙リストを作成．
② 学会講演，模擬講演の語彙リストから，全講演の語彙リストを作成．
③ 全講演の語彙リストを用いて，学会講演と模擬講演に重み付けをした bigram，trigram それぞれの言語モデルを作成．
⑤ bigram には 1，trigram には 3 のカットオフを施す．

5. 教師なし適応

講演音声認識の主な目的の一つとして，講演の書き起こしがある．一般には人手で行う作業であるが，大変手間がかかり自動化することが望まれる．このような用途の場合はリアルタイムに認識する必要がないため，バッチ処理で教師なし適応を行い，性能の向上を図ることができる．ここでは，音響モデルおよび言語モデルの教師なし適応で，どの程度認識性能が向上するかを検討する．

5.1 音響モデルの教師なし適応

音響モデルの教師なし適応として MLLR 法を用いる．音素木を用いたクラスタリングを行い（最大クラスタ数は音素数の 33），回帰クラスタ数は適応データ量に依存して自動的に決まる．1 つの回帰クラスタに対して 16 秒以上の適応データを保証している．適応パラメータはガウス分布の平均，分散，混合分布の重みで，平均はフル変換行列で，分散は対角変換行列で更新した．

5.2 言語モデルの教師なし適応

言語モデルの教師なし適応では，頑健性の向上を目的としてクラスモデルを用いる．大量テキスト（学習テキスト）から作成した単語 N-gram と認識結果から作成したクラス N-gram を線形補間することで，認識に使用する適応 N-gram モデルを作成する．

trigram 言語モデル適応の流れ図を図 3 に示す．大量テキスト（学習テキスト）から単語 trigram モデルを作成し，そのモデルを用いて適応データ（評価データ）をデコーディングし認識結果を獲得する．次いで，認識結果に含まれる品詞情報を利用して，品詞列の出現回数と品詞からの単語の出現確率を推定する．さらに，大量テキストから推定した品詞列の出現回数も用いて，品詞 trigram で用いる品詞連鎖確率を次式で求める．

$$P(c_i|c_{i-2}c_{i-1}) = \frac{N_0(c_{i-2}c_{i-1}c_i) + W * N(c_{i-2}c_{i-1}c_i)}{N_0(c_{i-2}c_{i-1}) + W * N(c_{i-2}c_{i-1})}$$

$N_0()$，$N()$ はそれぞれ，大量テキスト，認識結果から推定した品詞列の出現回数であり，W は重みである．あとの実験では，W=1 とした．最後に，ベースラインの単語 trigram と品詞 trigram を次式のように線形補間して，適応 trigram を構築する．

$$P'(\omega_i|\omega_{i-2}\omega_{i-1}) = \lambda P(\omega_i|\omega_{i-2}\omega_{i-1}) + (1-\lambda)P(\omega_i|c_i)P(c_i|c_{i-2}c_{i-1})$$

右辺第 1 項が単語 trigram の確率，右辺第 2 項が品詞 trigram の確率である．$P(\omega_i|c_i)$ はある品詞カテゴリに対する各単語の出現確率であり，認識結果を使って求めている．認識結果を利用して品詞クラスモデルを作成しているため，品詞からの単語の出現確率 $P(\omega_i|c_i)$ で話題の情報を表現し，品詞の連鎖確率 $P(c_i|c_{i-2}c_{i-1})$ で言い回しを捉えていると考えられる．線形補間係数 λ は，0.7 とした．品詞クラス数は活用型と活用形も考慮して 316 である．

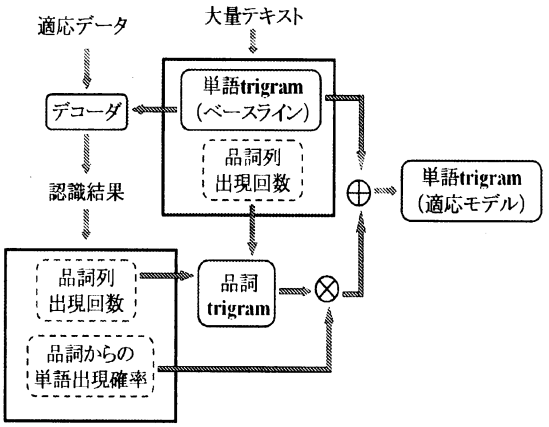

図 3 言語モデル適応の流れ図

6. 実験条件

6.1 学習セット

言語モデル作成に用いる学習セットの講演数，単語数を表 3 に示す．表 3 の学習セットは，公開版 CSJ の全講演のおおよそ 1/8，1/4，1/2，all からなる 4 種類で，モニター版の testset と公開版の testset 1 ～ 3 (計 34 講演) は除いてある．公開版-1/8 はモニター版から上の 34 講演を除いたものである．公開版-1/4，1/2 は，男性・女性の学会・模擬それぞれの講演を 1/4，1/2 ずつ含む．

音響モデルの学習セットは，学会講演・男性話者（795 講演，164 時間）を用いる．本研究では，モニター版 CSJ の学会講演・男性話者から作成された HMnet の構造を利用し，公開版 CSJ で連結学習を 5 回行って各状態の出力確率分布などのパラメータのみを再推定した，3000 状態 16 混合のモデルを使用する。

6.2 評価セット

評価セットにはモニター版 CSJ の testset（4 講演，約 2 万語）と公開版 CSJ の testset1（10 講演，約 2.7 万語）を用いる．

表 3 言語モデルの学習セット

		公開版 CSJ			
		1/8	1/4	1/2	all
学会講演	男性	103	203	400	795
	女性	31	38	82	168
模擬講演	男性	80	203	401	800
	女性	143	223	451	905
合計	講演数	357	667	1334	2668
	単語数 (万)	88	173	342	686

7. 評価実験

7.1 発音変形依存モデルの評価

図 2 の解析データ 1，解析データ 2，解析データ 3 から作成

の言語モデルをそれぞれ，LM 0，LM 1，LM 2 とする．
LM 0 は発音変形を考慮しないモデル，LM 1 は単語分割処理をしていない発音変形依存モデル，LM 2 は単語分割処理をした発音変形依存モデルである．言語モデルの学習データ量を公開版-1/8. 1/4. 1/2, all と増やして，言語モデル LM 0，LM 1，LM 2 の性能を比較する．単語辞書もその都度作成する．

それぞれの単語辞書のエントリー数（表記の数，発音変形を含む数）を表 4 に示す．カッコ内の数値が表記のエントリー数である．

表 4　単語辞書のエントリー数

言語モデル	公開版 CSJ			
	1/8	1/4	1/2	all
LM 0	18528 (15827)	26240 (22518)	33203 (28517)	40045 (34338)
LM 1	20091 (15280)	28580 (21368)	34415 (25752)	42991 (31626)
LM 2	19968 (14971)	28429 (20831)	34158 (25015)	42548 (30542)

モニター版 CSJ の testset，公開版 CSJ の testset 1 に対する実験結果をそれぞれ図 4，図 5 に示す．それぞれの図の左のグラフが補正 perplexity，右のグラフが WER である．ただし、これらのグラフにおいては、それぞれの言語モデルの語彙の内容自体が変わるため、WER や perplexity の正確な比較とはなっていないが，おおよその傾向はわかる．

図 4，図 5 より，言語モデルの学習データ量が増えるに従って perplexity，WER ともに徐々に改善されている．また，言語モデル LM 1，LM 2 を用いた場合の WER は，LM 0 と比べて，モニター版 CSJ の testset で約 7 ポイント，公開版 CSJ の testset 1 で約 5 ポイント良い結果が得られ，発音変形依存モデルの有効性が示された．しかし，発音変形依存言語テキストで単語分割処理を行うことによる性能の改善はほとんど見られなかった．この理由としては，語彙選択を学会講演から上位 2 万語，模擬講演から出現回数 2 回以上の単語と設定したために，語彙サイズが一定でなく，Cover 率がほとんど変わらないことによると考えられる．最終的には公開版 all の学習セットから作成の LM 2 を用いた場合に，モニター版 CSJ の testset で 17.78 %，公開版 CSJ の testset 1 で 19.96 %の WER が得られた．

7.2　教師なし適応実験

教師なし適応は，まず音響モデルを 2 回適応し，次いで言語モデルを 2 回適応した．これを 2 回繰り返し，音響モデル，言語モデルそれぞれに計 4 回ずつの適応を施した．

モニター版 CSJ の testset，公開版 CSJ の testset 1 に対する実験結果をそれぞれ図 6，図 7 に示す．横軸の「AM」は音響モデルの適応，「LM」は言語モデルの適応を表す．また，LM 2 を用いた場合の適応前と適応後の第 1，第 2 パス WER，GER，WGD を表 5，表 6 に，各講演の第 2 パス WER を表 7，表 8 に示す．GER は単語グラフ内で正解単語列に最も近い候補の WER で，第 2 パス WER の下限値を示す．WGD は正解 1 単語当りの仮説数を表し，単語グラフの大きさを示す．

図 6，図 7 より，教師なし適応が有効であることが示され，音響モデルと言語モデル（LM2）に適応を施した場合に，最終的にモニター版 CSJ の testset で 15.16 %，公開版 CSJ の testset 1 で 15.41 %の WER が得られた．

表 5，表 6 より，GER 6〜7 %以下の高性能の単語グラフが得られていることがわかる．これは，第 2 パスのリスコアリングをうまく行えば WER を大きく改善できる可能性があることを示している．また，教師なし適応後の単語グラフは，GER が適応前とほとんど変わらないにもかかわらず，大きさが 1/2〜1/3 以下になっている．

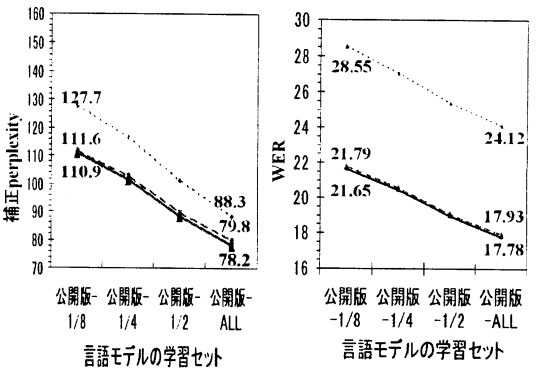

図 4　モニター版 CSJ の testset による言語モデルの性能評価
（…:LM 0，－－－:LM 1，――:LM 2）

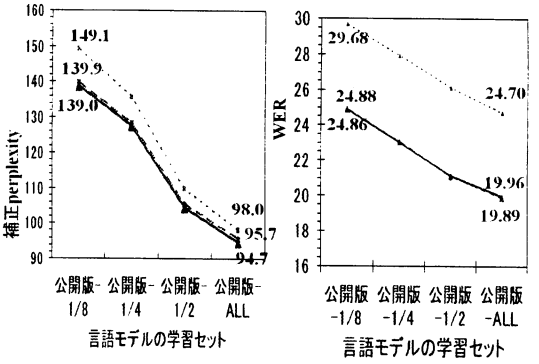

図 5　公開版 CSJ の testset 1 による言語モデルの性能評価
（…:LM 0，－－－:LM 1，――:LM 2）

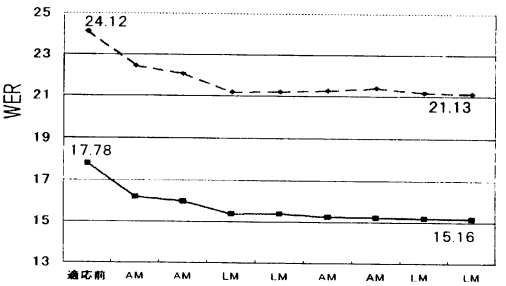

図 6　モニター版 CSJ の testset による教師なし適応結果
（- - -:LM 0，—:LM 2）

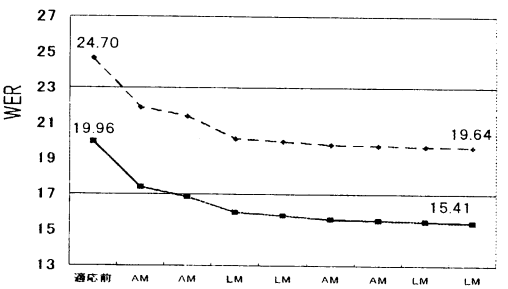

図 7　公開版 CSJ の testset1 による教師なし適応結果
（- - -:LM 0，—:LM 2）

表 5 モニター版 CSJ testset の WER，GER，WGD

	適応前	適応後
WER(第 1 パス)	19.88	15.93
WER(第 2 パス)	17.78	15.16
GER	5.59	5.74
WGD	60.18	25.88

表 6 公開版 CSJ testset1 の WER，GER，WGD

	適応前	適応後
WER(第 1 パス)	22.03	15.94
WER(第 2 パス)	19.96	15.41
GER	6.69	6.89
WGD	54.82	16.33

表 7 モニター版 CSJ testset の各講演の WER

	適応前	適応後
A01M0007	14.76	12.05
0035	22.93	19.99
0074	11.34	9.62
A05M0031	17.27	14.66
平均	17.78	15.16

表 8 公開版 CSJ testset1 の各講演の WER

	適応前	適応後
A01M0097	7.12	5.01
0110	16.84	10.57
0137	18.43	16.11
A03M0106	28.53	20.07
0112	11.30	7.75
0156	30.94	24.60
A04M0051	15.84	12.05
0121	21.74	17.06
0123	20.04	16.10
A05M0011	23.87	20.58
平均	19.96	15.41

8. むすび

日本語話し言葉コーパス (CSJ) 公開版を用いて，発音変形依存モデリングの有効性について検討した．

言語モデルの学習データ量が増えるに従って perplexity，WER ともに徐々に改善されることが確認された．言語モデルの学習データ量がモニター版 CSJ と比べ約 8 倍になったことにより，perplexity では約 30〜40，WER では約 4〜5 ポイント改善された．また，発音変形を考慮することにより，モニター版 CSJ の testset では，perplexity で約 10〜17，WER で約 6〜7 ポイント，公開版 CSJ の testset1 では，perplexity で約 3〜10，WER で約 5 ポイント性能が改善しており，これにより発音変形依存モデルの有効性が示された．最終的には，単語分割処理後の発音変形依存言語テキストから作成した言語モデルを用いて，モニター版 CSJ の testset で WER17.78 %，公開版 CSJ の testset1 で WER19.96 %の結果を得た．さらにその結果に音響モデル，言語モデルそれぞれについて計 4 回の教師なし適応を繰り返し施すことにより，モニター版 CSJ の testset で WER15.16 %，公開版 CSJ の testset1 で WER15.41 %の結果を得た．

参考文献

[1] 堀貴明，岡直生，加藤正治，伊藤彰則，好田正紀："大語彙連続音声認識のための音素グラフに基づく仮説制限法の検討"，情報処理学会論文誌，Vol.40，No.4，pp.1365-1373，1999.

[2] 小磯花絵，前川喜久雄，"『日本語話し言葉コーパス』の概要と書き起し基準について"，情処学研報，SLP-36-1，pp.1-8，Jun. 2001.

[3] 河原達也，"『日本語話し言葉コーパス』を用いた音声認識の進展"，第 3 回話し言葉の科学と工学ワークショップ講演予稿集，pp.61-66，Feb. 2004.

[4] 篠崎隆宏，古井貞熙，"超並列デコーダによる話し言葉音声認識"，第 3 回話し言葉の科学と工学ワークショップ講演予稿集，pp.67-72，Feb. 2004.

[5] 堀 貴明，渡部晋二，エリック・マクダーモット，南 泰浩，中村 篤，"音声認識システム SOLON の日本語話し言葉コーパスによる評価"，第 3 回話し言葉の科学と工学ワークショップ講演予稿集，pp.85-91，Feb. 2004.

[6] 伊藤彰則，好田正紀，"単語およびクラス n-gram 作成のためのツールキット"，信学技報，SP2000-95，pp.67-72，http://palmkit.sourceforge.net/，Dec. 2000.

[7] 堤 怜介，加藤正治，小坂哲夫，好田正紀，"発音変形依存と教師なし適応による講演音声認識の性能改善"，第 3 回話し言葉の科学と工学ワークショップ講演予稿集，pp.93-98，Feb. 2004.

[8] E.Evermann, H.Y.Chan, M.J.F.Gales, T.Hain, X.liu, D.Mrva, L.Wang, P.C.Woodland:"Development of the 2003 CU-HTK conversational telephone speech transcription system", Proc. of ICASSP2004, Vol.1, pp. 249-252 (2004).

[9] E.Evermann, H.Y.Chan, M.J.F.Gales, B.Jia, D.Mrva, L.Wang, P.C.Woodland, K.Yu:"Training LVCSR systems on thousands of hours of data", Proc. of ICASSP2005, Vol.1, pp. 209-212 (2005).

[10] H.Soltau, B.kingsbury, L.Mangu, D.Povey, G.Saon, G.Zweig:"The IBM 2004 conversational telephony system for rich transcription", Proc. of ICASSP2005, Vol.1, pp. 205-208 (2005).

[11] M.Schuster, T.Hori, A.Nakamura:"Mixtures of probabilistic principal component analyzers in speech recognition", Technical Report of IPSJ, 2004-SLP-54, pp. 67-71 (2004).