2004－SLP－53　(3)
2004／10／22

『日本語話し言葉コーパス』を用いた汎用的な発音変動モデルの統計的学習

秋田 祐哉 †‡　　河原 達也 †‡

† 京都大学大学院情報学研究科
‡ 科学技術振興機構さきがけ研究 21

あらまし　　話し言葉音声の認識において，発音変動のモデル化は認識性能に深く関わる課題である．通常，音声認識に用いる発音辞書は形態素解析器が出力する標準的な読みに基づいて生成されるが，これでは話し言葉に多く含まれる発音変動をカバーできない．本研究では，まず『日本語話し言葉コーパス』（CSJ）を用いて発音変動のパターンを汎用的な音素系列のレベルで統計的に学習した．コーパスから自動的に獲得された音素列の変動パターンは 265 種類であり，音韻論的に妥当なものに加えて人手による規則化が困難なものを頻度統計とあわせて抽出することができた．これらのパターンに対して，バックオフ手法により可変長の音素文脈を扱える確率つき音素書き換え規則を構築する．これらの規則を適用することで，任意の語いに対して標準的な読み（baseform）から話し言葉特有の変動を含んだ発音（surface form）を生起確率とともに生成することができる．本手法を CSJ とは異なるドメインのための発音辞書に適用したところ，エントリ数が 21%増加した．さらに，この発音辞書を用いた音声認識により有意な単語誤り率の改善を得ることができた．

Generalized Statistical Modeling of Pronunciation Variations using the Corpus of Spontaneous Japanese

Yuya AKITA†‡　　Tatsuya KAWAHARA†‡

†School of Informatics, Kyoto University
‡PRESTO, Japan Science and Technology Agency (JST)

Abstract　　Pronunciation variation modeling is one of major issues in automatic transcription of spontaneous speech. We present statistical modeling of subword-based mapping between baseforms and surface forms using a large-scale spontaneous speech corpus (CSJ). Variation patterns of phone sequences are automatically extracted together with their contexts of up to two preceding and following phones, which are decided by their occurrence statistics. Then, we derive a set of rewrite rules with their probabilities and variable-length phone contexts. The model effectively predicts pronunciation variations depending on the phone context using a back-off scheme. Since it is based on phone sequences, the model is applicable to any lexicon to generate appropriate surface forms. The proposed method was evaluated on a transcription task whose domain is different from the training corpus (CSJ), and significant reduction of word error rate was achieved.

1　まえがき

近年，大語彙連続音声認識の研究対象は，講演・講義や討論・会議のような自然な自発音声（話し言葉）に移行しつつある．このような話し言葉の音声認識は，音声の記録としての用途のほか，リアルタイムの字幕生成や書き起こしに基づくインデキシング・要約処理などへの応用も期待されている．しかし，話し言葉音声で観測される音響的・言語的現象は読み上げ音声や放送ニュース音声と比べて多様であることから，同等の認識精度を実現するに至っていない．

話し言葉の多様性の 1 つとして，言語的に同一の単語が異なって発音される，発音変動あるいは言語変異と呼ばれる現象がある [1]．発音変動における音響的な変動はさまざまであるが，単語内の音節や音素のレベルで変動を捉えることが可能なものは，認識時に利用する発音辞書（単語辞書）でカバーされる．すなわち，標準的な発音（baseform）に加えて実際にあり得る発音（surface form）が発音辞書に登録される．ただし，発音変動の抽出はテストセットとタスクドメインが合致したデータを用いて行われているのがほとんどである．

一方，近年整備の進んでいる大規模な話し言葉コーパスを用いて，広範かつ精密に発音変動をモデル化するアプローチが考えられる．日本語では，話し言葉音声の諸相を包含した『日本語話し言葉コーパス』（CSJ）[2] が構築されている．発音辞書の問題についても先行研究 [3, 4] で扱われているが，CSJ の語いに特化したモデル化になっており，CSJ のテストセットにしか事実上適用できないものである．

これに対して本研究では，CSJ を用いて音素系列レベルで発音変動を抽出し，任意の発音辞書へ反映させる手法を提案する．一般的な日本語の発音辞書は形態素解析器が出力した標準的な読みに基づいて作成されるが，これでは話し言葉に含まれる発音変動はカバーされていない．本研究では，CSJ のテキストを用いて，形態素解析器が出力する読みに対して変動の発生する音素列パターンとその頻度統計を学習する．これに基づき，可変長の音素文脈を用いた確率つき音素書き換え規則を生成する．この規則を用いることで，同様の音素列パターンからなる任意の単語に対して，あり得る発音を確率つきで生成することが可能になる．頻度の小さいパターンについては確率は十分に推定できないが，より短い文脈にバックオフすることで，頑健な音素列マッチングによる変動形の生成と確率の推定を実現する．本研究では提案手法を CSJ とは異なるドメインにおける音声認識に適用し評価を行う．

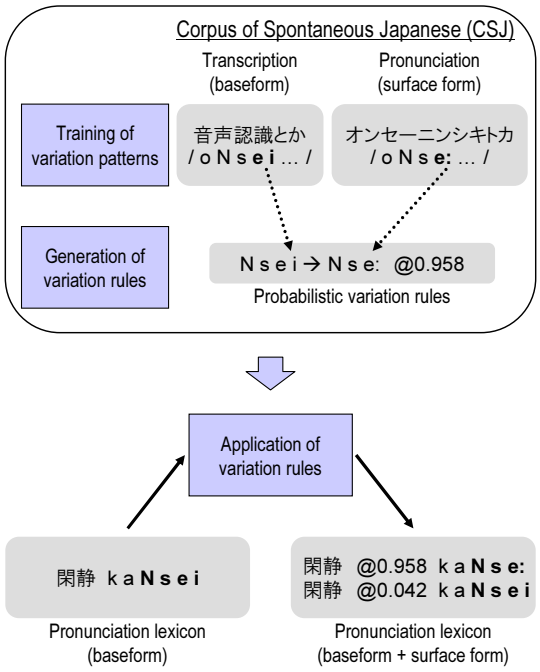

図 1: 提案手法の概要

2　CSJ を用いた発音変動モデルの学習と適用

提案手法の処理の概要を図 1 に示す．まず，CSJ のテキストを用いて発音変動の検出とパターンの学習を行う．次に，この学習に基づいて確率的変動規則を生成する．そして，標準的な読みによる発音辞書に対して，この変動規則を適用して変動形を追加し，新たな発音辞書を生成する．以下，処理の各ステップについて詳細に述べる．

2.1　学習データ

本研究では CSJ の学会講演及び模擬講演を用いる．講演数の合計は 2,540 である．CSJ では，これらの音声の書き起こし（基本形）とその実際の発音（発音形）が併記されている．基本形は「形態的な分析を申し上げます」のように正書法に基づいて書き起こされているのに対し，発音形では「ケータイテキナブンセキオモーシアゲマス」のように実際の発音が忠実に記述されている．したがって，基本形に対

する標準的な読みと発音形の対応づけをとることで発音変動を抽出することが可能である．ただし，発音形の表記は仮名（すなわち音節）を単位としているため，母音の脱落のように通常の日本語音節（子音＋母音）を構成できなくなる変動は完全に捉えることはできない．音素や音節に加えて機能語モデルを導入することにより対処することも考えられるが，本研究ではこのような変動は対象としない．

2.2 発音変動の抽出と頻度の算出

第 1 段階として，CSJ の書き起こしにおける発音変動箇所を同定し，音素列パターンごとの頻度を求める．処理の流れと具体例を図 2 に示す．

まず，書き起こしに対して形態素解析を行い，単語境界と読みを付与する．解析器としては茶筌 Ver. 2.2.3 を，形態素辞書には IPADIC 2.4.4 を用いた．これにより得られた単語の総数は約 630 万語で，語いのサイズは 51,720 である．なお,「日本語ディクテーション基本ソフトウェア」[5] で開発された IPADIC においては，NHK 日本語発音アクセント辞典（新版）に基づいて,「東京（トーキョー）」のように読みが付与されているが，話し言葉の発音を包含するものではない．

次にこの読みと発音形表記との間で DP マッチングによる単語単位のアライメントを行い，発音形表記に対しても単語境界を挿入する．これと同時に，複数の読みが与えられた単語については，発音形表記と最も近い読みを選択する．そして音素単位でのアライメントを行い，変動箇所を同定する．これによる変動前と変動後の音素列の組について，その前後それぞれ最長 2 音素までの音素文脈を含んだパターンを抽出し，それらの頻度をカウントする．この際，単語境界を音素と同様に文脈として扱うこととする．

抽出された発音変動の例として，変動頻度の大きなものを表 1 に示す．抽出された発音変動には，音韻論（例えば [6]）的な予測が可能なものが含まれている．表 1 では「e-i → e:」などの母音の長音化が顕著にみられるが，これらは音素の調音における特徴（音声素性）の点で規則性があるものである．また,「k-u → q」のような促音化は音韻論では無声子音に挟まれた母音（/u/）の消失と考えられているが，本手法でもこのような文脈における変動であることが確認された．このほか，子音に関しては「k → g」などの濁音化が観測されており，その多くが単語境界直後に発生していることから複合語の連濁に起因すると考えられる．これらの音韻論的予測の可能な変動に対しては，本手法によって変動を抽出するだけでなく，発生確率の推定まで行うことができた．

一方,「n-i → N」や「e-r-e → e:」,「o:→ o」などの発音の怠けによる変動は，特定の文脈で発生するために個別の検討が必要であり，必ずしも音韻論的な予測ができるわけではない．このような変動に対しては，本手法のような大規模なコーパスを用いた発音変動の抽出が特に有効であるといえる．

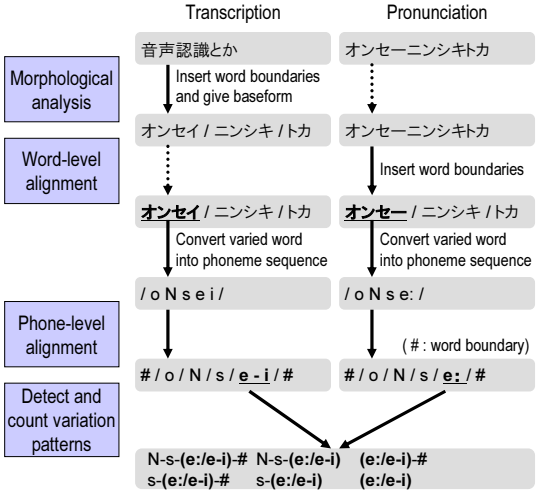

図 2: 発音変動の学習

表 1: CSJ から抽出された発音変動の例

パターン	種類	例
e-i → e:	長音化	音声（オンセイ→オンセー）
u-u → u:	長音化	いう（ユウ→ユー）
i-i → i:	長音化	用い（モチイ→モチー）
o: → o	短音化	本当に（ホントーニ→ホントニ）
a: → a	短音化	データー（データー→データ）
u: → u	短音化	ふう（フー→フ）
k → g	濁音化	会社（カイシャ→ガイシャ）
k-u → q	促音化	百（ヒャク→ヒャッ）
n-i → N	撥音化	毎日（マイニチ→マインチ）
u →	脱落	いう（ユウ→ユ）
r →	脱落	それ（ソレ→ソエ）
i →	脱落	帯域（タイイキ→タイキ）
e-r-e → e:	その他	けれども（ケレドモ→ケードモ）
i → u	その他	エキスポ（エキスポ→エクスポ）
u → i	その他	出場（シュツジョーシツジョー）

2.3 確率付き変動規則の生成

次に，変動のパターンと頻度から変動の発生確率を推定し，確率付き変動規則とする．同一の発音変動

においては，文脈の長いパターンから規則として採用し，得られないときは短い文脈のパターンを採用する．提案手法で採用するバックオフ手法は N-gram 言語モデルにおけるバックオフスムージングと同様の考え方であるが，文脈が前後両方向であるためグッド・チューリング法やウィッテン・ベル法などの単純な適用は困難である．

例として，ある音素（列）q が q' に変化する場合を考える．文脈 c において，q が出現した頻度を $C(q|c)$，その中で q' に変動した頻度を $C(q \to q'|c)$ とする．頻度の小さな変動パターンは信頼できないと考えられるため，頻度のしきい値 θ_1 を導入し，$C(q|c) \geq \theta_1$ であるパターンを変動規則として採用する．このとき，文脈 c において変動 $q \to q'$ が発生する確率は次式で定められる．

$$P(q \to q'|c) = \frac{C(q \to q'|c)}{C(q|c)} \quad (1)$$

本研究では音素文脈として前後それぞれ最大 2 音素を用いている．前方と後方の音素文脈の長さをそれぞれ i，j とし，採用された規則の文脈の中でこの長さをもつものの集合を R_{ij} で表すと，長さ 4 の文脈集合 R_{22}，長さ 3 の文脈集合 R_{21}，R_{12}，長さ 2 の文脈集合 R_{20}，R_{11}，R_{02}，長さ 1 の文脈集合 R_{10}，R_{01}，長さ 0 の文脈集合 R_{00}（文脈なし）が考えられる．したがって，長さ 4 の文脈から降順に，しきい値 θ_1 と式（1）により規則を採用し R_{ij} を定める．ただし，異なる文脈長の規則で頻度を重複して用いないように，頻度を補正する必要がある．例えば，前方の文脈が $a\ b$，後方の文脈が d である長さ 3 の文脈の規則については，式（2）のように長さ 4 の文脈の中で規則として採用されたもの（すなわち R_{22} の要素）の頻度を減じる．

$$\begin{aligned} &C'(q|ab:d) \\ &\quad = \quad C(q|ab:d) - \sum_{\substack{(ab:dz) \\ \in R_{22}}} C(q|ab:dz) \end{aligned} \quad (2)$$

以上より得られた R_{ij} $(0 \leq i, j \leq 2)$ が変動 $q \to q'$ に関する変動規則であり，生起確率 $P(q \to q'|c)$ がそれぞれに付与されている．最後に，この確率についてもしきい値 θ_2 を導入し，$P(q \to q'|c) \geq \theta_2$ の場合に規則として採用する．本研究ではしきい値 θ_1・θ_2 に関して予備的な実験を行って調査し，事後的に $\theta_1 = 20$，$\theta_2 = 0.1$ と定めた．このとき抽出された発音変動は 265 種類，変動規則の総数は 1,381 である．変動規則の例を表 2 に示す．

表 2: 抽出された発音変動規則の例

パターン			種類	確率
N s **e-i** #	→	N s **e:** #	長音化	0.9713
o y **u-u** #	→	o y **u:** #	長音化	0.9564
y u **i-i** ts u	→	y u **i:** ts u	長音化	0.4167
N t **o:** n i	→	N t **o** n i	短音化	0.8680
e: t **a:**	→	e: t **a**	短音化	0.3563
# **s** u sh	→	# **z** u sh	濁音化	0.3475
# **f** u s	→	# **b** u s	濁音化	0.1238
a **k-u** k	→	a **q** k	促音化	0.1818
ts-u t a	→	**q** t a	促音化	0.2162
ch i **n-i** ch i	→	ch i **N** ch i	撥音化	0.3891
t a **i** i k	→	t a i k	脱落	0.4782
s o **r** e d	→	s o e d	脱落	0.1051
a g **a-w-a** #	→	a g **a:** #	その他	0.1379
# **sh** i ch	→	# **h** i ch	その他	0.1072

※「#」は単語境界を示す．

2.4 発音辞書への変動規則の適用

音声認識用の発音辞書に対しては，これらの変動規則を用いて新たな発音エントリ（surface form）を追加する．規則の適用にあたっては，文脈が最も長くなるように c を選択する．同一の長さの文脈が複数ある場合は，より信頼できると考えられる頻度のより大きなものから適用する．複数の変動があり得るエントリについては，それぞれ規則を適用してエントリを追加する．このとき得られたエントリの確率は，それぞれの変動に関する確率を乗じて求める．

以上をまとめると，単語 w の発音エントリ p に対して変動規則 $q \to q'$ が適用可能で，これにより新しい発音エントリ p' が得られる場合，もとの発音確率 $P(p|w)$ に対して式（3）により p' の確率が設定される．また，p についても式（4）により確率が更新される．

$$P(p'|w) \quad \leftarrow \quad P(p|w)P(q \to q'|c) \quad (3)$$

$$P(p|w) \quad \leftarrow \quad P(p|w)\{1 - P(q \to q'|c)\} \quad (4)$$

$P(p|w)$ の初期値は，w にあらかじめ与えられた読み（baseform）の数で 1 を除した値である．なお，この確率がしきい値 θ_2 以下となった場合は，そのエントリは登録しない．このようにして，任意の語彙に対して話し言葉特有の発音変動をカバーする単語辞書を構成することができる．

2.5 音声認識デコーダにおける発音確率の利用

一般的な統計的音声認識の枠組みは，入力音声（特徴量）x，文 w に対して，式（5）のように定式化される．

$$w' = \arg\max_{w} P(x|w)P(w) \tag{5}$$

$P(x|w)$ は w に対する x の音響的なゆう度であり，$P(w)$ は w の言語的なゆう度である．ここで，文の発音 p が複数あり得ることを考慮し発音モデルを導入すると，式（5）は式（6）に改められる．

$$w' = \arg\max_{w,p} P(x|p)P(p|w)P(w) \tag{6}$$

$P(x|p)$ は発音 p に対する音響ゆう度で，$P(p|w)$ は w が p と発音される確率である．なお，ここでは最ゆうの発音のみを考慮することとする．

式（6）の右辺のゆう度は実際には対数をとり，式（7）で求める．

$$\log P(x|p) + \log P(p|w) + \log P(w) \tag{7}$$

ここで，言語モデルゆう度 $\log P(w)$ に通常言語モデル重み w_l を乗じるのと同様に，発音モデルのゆう度 $\log P(p|w)$ にも発音モデル重み w_p を導入することを考える．このとき，ゆう度は式（8）で求められる．

$$\log P(x|p) + w_p \log P(p|w) + w_l \log P(w) \tag{8}$$

これは，言語ゆう度 $P(w)$ のダイナミックレンジと発音モデルゆう度 $P(p|w)$ のダイナミックレンジが異なるため，それらを補正して各モデルを効果的に適用するためのものである．

3 評価実験

3.1 テストセット音声

提案手法を CSJ とは異なる話し言葉音声の書き起こしタスクに適用した．本研究では，NHK のテレビ討論番組『日曜討論』を評価用音声として利用する．『日曜討論』は，政治・経済・外交などの分野における時事問題を対象に，政治家や学者，評論家などが 5-8 名程度参加し議論するものである．討論のテーマは毎回異なるため，参加者も毎回異なる．番組は毎回 1 時間である．2001 年 6 月から 2002 年 1 月までの間に放送された中から 10 回分を用いた．討論音声は 400ms の無音により区切って発話単位としている．ただし，相づちなどの短い発話は無視している．1 討論あたりの平均発話数は 550 である．

表 3: 言語モデルの仕様

モデル	国会	講演	混合
学習データ	衆議院会議録 (1999-2002)	日本語話し言葉コーパス（CSJ）（模擬講演のみ）	—
総単語数	70M	2.9M	—
異なり単語数	72K	37K	—
語彙サイズ	29K	5.8K	30K
平均 PP	187.50	111.89	105.62
平均 OOV	4.78%	10.02%	2.13%

3.2 言語モデル・音響モデル・デコーダ

テストセットの音声に含まれる話題と話し言葉表現の特徴をカバーするために，2 つのコーパスを利用して言語モデルを構築した．それぞれのコーパスと，それをもとに構築した言語モデルの仕様を表 3 に示す．国会モデルは話題をカバーするために用い，4 年分の衆議院の全会議録を利用して構築した．これらの会議録では発言の内容が忠実に書き起こされているが，フィラーや言い淀みの除去と文末表現・口語的表現の簡単な修正が行われている．一方，話し言葉の表現をカバーするために講演モデルを用いた．学習には CSJ に含まれる模擬講演のみを利用し，比較的高い頻度の語いに制限している．認識に用いる言語モデルは，これら 2 つの言語モデルを重み付け混合したものである．混合比は予備実験によりあらかじめ国会 0.5・講演 0.5 と定めた．

音声認識に用いる語いは，それぞれの学習コーパスにおいて出現頻度をもとに定めたものを併合しており，サイズは 29,720 である．この語いの中で，発音モデルの学習に用いた CSJ に含まれない単語の割合は 27.7%であった．併合した語いによるベースラインの発音辞書（エントリ数 31,571）に対して提案手法を適用し，エントリ数 38,207 の発音辞書を得た．

音響モデルは，CSJ の発音形表記の書き起こしを用いて学習された triphone HMM[7] に，MLLR による教師なし話者適応 [8] を施したものを利用した．デコーダには Julius[9] rev.3.4.2 を用いた．本実験における言語モデル重みは $w_l = 7.0$ である．

3.3 実験の結果と考察

評価実験では，ベースラインの発音辞書（Baseline），変動形のエントリのみ追加した場合（+Entry），確率も導入した場合（+Prob）のそれぞれの単語誤り

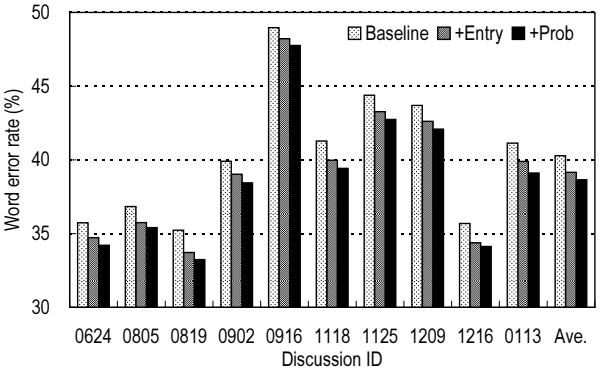

図 3: 各討論データにおける単語誤り率

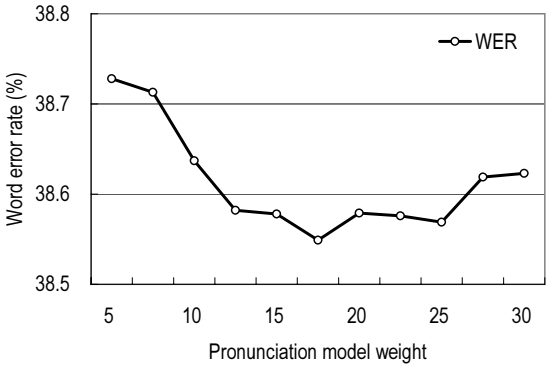

図 4: 発音モデル重みの単語誤り率への影響

率を求めた．ここでは発音モデル重み w_p は言語モデル重み w_l と同一に設定した．10 討論のそれぞれにおける単語誤り率を図 3 に示す．図 3 より，いずれの討論でも提案手法により単語誤り率が改善していることがわかる．10 討論の平均では，ベースラインの単語誤り率 40.3%に対して，エントリのみの追加では 39.2%で，確率の利用により 38.7%となった．したがって，エントリの追加で 1.1%，確率の設定で 0.5%，合計で 1.6%の改善があったといえる．この改善は統計的に有意である．

次に，異なる発音モデル重み w_p ごとに音声認識を行い，単語誤り率を求めた．w_p と平均の単語誤り率の関係を図 4 に示す．図 4 より，w_p が言語モデル重み（$w_l = 0.7$）の 2〜3 倍の場合に認識率が改善されることがわかる．これは，発音モデルの確率が N-gram 言語モデルによる確率よりもダイナミックレンジが小さいことから，発音モデル重みを大きくすることがそれを補正し，モデルが有効に機能することを示している．

4 むすび

本稿では，話し言葉に見られる発音変動を統計的に学習して，任意の語い（単語辞書）に適用できる手法を提案した．CSJ を用いて，標準的な読みに対して発音の変動が生じる典型的な音素列パターンを確率とともに学習した．本手法では，音韻論的な知見に基づいて予測することが可能なパターンに加えて，話し言葉に特有で予測の難しいパターンも抽出することができた．また，このようにパターンの頻度に基づいて意味のある変動確率を推定することができ，この確率は統計的音声認識の枠組みとも適合するものである．これらの変動パターンに基づいて確率つき変動規則が構成され，これを用いて与えられた任意の語いに対して確率つきで発音辞書エントリを生成することができる．本研究では CSJ とは異なるドメインの音声において評価を行い，絶対値で 1.6%の単語誤り率の改善を得ることができた．

参考文献

[1] 前川喜久雄, 小磯花絵, 菊池英明, 間淵洋子, 斎藤美紀. 『日本語話し言葉コーパス』に捉えられた言語変異現象. 国立国語研究所公開研究発表会資料, pp. 41–42, 2003.

[2] S. Furui, K. Maekawa, and H. Isahara. Toward the realization of spontaneous speech recognition – Introduction of a Japanese priority program and preliminary results –. In *Proc. ICSLP*, Vol. 3, pp. 518–521, 2000.

[3] H. Nanjo and T. Kawahara. Language model and speaking rate adaptation for spontaneous presentation speech recognition. *IEEE Trans. Speech & Audio Process.*, Vol. 12, No. 4, pp. 391–400, 2004.

[4] 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. 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, 2004.

[5] 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.

[6] 窪薗晴夫. 日本語の音声. 現代言語学入門, No. 2. 岩波書店, 1999.

[7] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui. Benchmark test for speech recognition using the corpus of spontaneous Japanese. In *Proc. SSPR*, pp. 135–138, 2003.

[8] 秋田祐哉, 河原達也. 多数話者モデルを用いた討論音声の教師なし話者インデキシング. 信学論, Vol. J87-DII, No. 2, pp. 495–503, 2004.

[9] 河原達也, 武田一哉, 伊藤克亘, 李晃伸, 鹿野清宏, 山田篤. 連続音声認識コンソーシアムの活動報告及び最終版ソフトウェアの概要. 情処学研報, 2003-SLP-49-57, 2003.