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

2007－ＳＬＰ－69 (11)
2007／12／20

言語モデルと発音辞書の統計的話し言葉変換に基づく国会音声認識

秋田 祐哉† 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 我々は，国会などの話し言葉音声認識のために，言語モデルと発音辞書を話し言葉スタイルに変換する手法を提案している．提案法では，発話の忠実な書き起こしとこれに対応する正書体のデータの相違点が統計的に抽出され，これをもとに確率的な変換パターンからなる変換モデルが言語モデルと発音辞書のそれぞれに対して構成される．このモデルに基づき，言語モデルに対しては話し言葉の N-gram の予測と統計頻度の推定を行う．一方発音辞書に対しては，話し言葉特有の発音変動の予測および発音確率の推定を行う．生成された話し言葉スタイルの言語モデルと発音辞書を衆議院の委員会会議音声において評価したところ，従来のモデルと比較して単語誤り率が絶対値でそれぞれ0.6%～0.7%・1.0%改善し，これらの併用により合計で1.7%の削減を得ることができた．

キーワード 話し言葉，音声認識，言語モデル，発音辞書，発話スタイル

Automatic Speech Recognition of Congressional Speech Based on Statistical Style Transformation of Language Model and Pronunciation Model

Yuya AKITA† and Tatsuya KAWAHARA†

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

Abstract For automatic speech recognition (ASR) of spontaneous speech such as congressional meetings, we have been proposing statistical transformation methods of language model and pronunciation model. In these methods, differences between faithful transcripts and orthographical transcripts are statistically extracted. Then, transformation models which consist of probabilistic transformation patterns are derived from the statistics for language model and pronunciation model. For language model, the transformation model predicts spoken-style N-gram entries with estimated occurrence counts. For pronunciation model, pronunciation variants and their probabilities are predicted by the transformation model. The language model and pronunciation model generated by the proposed methods were evaluated on ASR of committee meetings of Japanese National Congress (Diet), and realized absolute reduction of word error rates by 0.6–0.7% and 1.0%, respectively, compared with models produced by conventional methods. Finally, total reduction of 1.7% was obtained by combining both models.

Key words Spontaneous speech, Speech recognition, Language model, Lexicon, Speaking style

1. はじめに

1990 年代に読み上げ音声 [1]～[3] や放送ニュース音声 [4]～[6] を対象として研究が進められた大語彙連続音声認識は，2000 年代に入り講義や講演，会議や討論などの「話し言葉」音声 (Spontaneous speech) に対象を移して研究が行われている．話し言葉音声認識の代表的な研究プロジェクトとしては，Switchboard コーパス [7] などの電話会話音声を扱った，米国 DARPA による Hub-5 および EARS があり，多くの研究を通じて現在の音声認識の基盤となる重要な技術が確立された [8]～[11]．米国では NIST により会議音声を対象とする Rich Transcription プロジェクトも進められており [12]～[14]，会議の実環境における音声認識の取り組みが進められている．一方国内では，学術講演と一般のスピーチを中心とした「日本語話し言葉コーパ

ス」(CSJ) が構築され [15], 音声・言語研究に活用されている. 音声認識に関する研究もさかんに行われており, 80%を越える単語認識精度が実現されている [16], [17].

このような中, 音声認識の対象として近年注目されている分野に議会の音声がある. 代表的な研究プロジェクトとしては欧州議会 (European Parliament) を対象とする TC-STAR があり, 複数の公用語を持つ欧州議会における音声認識と機械翻訳の研究が多くの研究機関によって進められている [18]〜[20]. 議会では原則として全ての発言を会議録として記録することから音声認識技術のマーケットとして有望と考えられており, 実際に国内でもいくつかの地方議会において音声認識に基づく会議録作成が導入されている. 我々もこれまで衆議院の会議を対象とした音声認識の研究に取り組んできている [21], [22]. 会議録は正確性がきわめて重視されることから, これまでのインデキシング等を目的とした場合と比べて高い認識精度が求められることが議会の音声認識の特徴といえる.

これらの話し言葉音声は人と人の間の (human-to-human) コミュニケーションであり, 一方的な伝達となっている読み上げ音声や放送ニュース音声, あるいは機械との (human-to-machine) インターフェースである音声対話システムと比べて多様な音響的・言語的現象が観測される. 話し言葉では話者の心的状態や感情, 個人性のために発声が一様ではなく, 観測される音響的特徴は読み上げ音声等とは明らかに異なる. また, 文法や正書法に基づく書き言葉に準じている読み上げ音声に対して, 話し言葉では断片的な文や冗長表現, フィラーなどの非流暢現象が頻繁に観測される. 特に日本語では文末表現の置換や丁寧語の多用なども発生し, 書き言葉と話し言葉の違いが大きい. 高い認識精度を達成するためには, これらを効果的・効率的にカバーすることが求められる.

この問題に関して, 電話会話音声 (Switchboard) や講演 (CSJ), 会議 (NIST-RT) のタスクでは, それぞれ十分な量の学習データを収集し音声認識システムの構築に利用している. しかし一般のタスクではこのような大規模な収集は現実的ではなく, TC-STAR をはじめ多くの他のタスクでは, タスクに適合しているが話し言葉ではないデータと, タスクに無関係な話し言葉のコーパスを組み合わせて利用している. これは, たとえば言語モデルにおける線形補間のように, タスクへの適合と話し言葉の性質を特に区別することなく, データを一緒に学習に用いる手法が主流である. このような手法は単純ながら効果的であり, 一定の性能が得られることがこれまでの報告で示されているが, タスクに適合しない不適切なデータまで学習に利用することから不都合が生じる可能性も内包している.

これに対して本研究では, 言語モデルと発音辞書に関して話し言葉の性質をコーパスから統計的に抽出し, 話し言葉スタイルへの変換モデルとして構成する手法を提案している [23], [24]. 本手法は話し言葉の特徴のみをコーパスから学習するため, 構築された変換モデルのタスクへの依存度が比較的小さい. このため広範なタスクへの応用が期待でき, かつタスクに不適合なモデルの生成を防げるという利点がある. 本稿では, 本手法を用いて行った国会音声認識について報告する.

表 1 国会音声で観測される話し言葉表現の例

Table 1 Examples of spoken-style expressions in congressional speech

挿入		脱落		置換			
え一	5,623	を	586	てる	→	て いる	2,209
ですね	2,728	は	535	ていう	→	という	406
お一	2,434	が	233	やっぱり	→	やはり	246
あの	2,097	に	150	ん	→	の	240
あー	2,008	と	120	けども	→	けれども	235
まあ	1,953	いる	52	いろんな	→	いろいろ な	181
あのー	1,438	よう	40	けど も	→	けれども	163
その	1,171	です	34	けど	→	けれども	161
で	1,078	の	30	て	→	という	138
と	1,063	か	29	もん	→	もの	106

2. 話し言葉のための言語モデルと発音辞書

本節では, 話し言葉音声において実際に観察された現象について述べ, 次にこのような現象に対する既存のアプローチについて検討する. 言語表現に関する分析に用いた音声は, 2003・2004 年における, 衆議院予算委員会を主とする一部の会議の書き起こしと対応する会議録であり, 総単語数は書き起こしで 737K である. 会議録ではフィラーや口語表現・文末表現などの典型的な話し言葉表現が修正されているため, 文書に近いスタイルとなっている. 一方, 発音に関する分析には CSJ を利用した [24].

2.1 言語モデル

表 1 は, 会議録に対して実際の音声で挿入・脱落・置換されている表現のうち, 高頻度の表現とその出現回数である. 表 1 から明らかなように, フィラーや文末表現 (ですね, と) の挿入が圧倒的に多い. 脱落した表現は助詞が中心であり, 係助詞「は」・「が」や格助詞「を」には脱落が比較的多くみられるのに対して格助詞「の」の脱落が少ないなどの特徴がみられるが, 発生頻度はいずれも少ない. 置換された表現には口語表現 (ていう, やっぱり, いろんな) や発音の怠け (てる, ん) があり, やはり発生頻度は少ない.

これらの表現をカバーするためには十分な量の書き起こしテキストを用意することが望ましいが, タスクごとに書き起こしを作成することは現実的ではない. したがって限られた量の書き起こしや, 他の話し言葉コーパスを利用することとなる. これらを組み合わせる最も単純な手法は線形補間であり, 特に書き起こしや話し言葉コーパスのサイズが小さい場合は N-gram 頻度に重みを乗じて利用することが一般的である. これにより話し言葉の N-gram を得ることができるが, 話し言葉特有の表現とそれ以外の部分にまたがる N-gram は得ることが難しいため, これによる言語モデルではこの部分の予測を適切に行えない可能性が大きい. さらに, 他のコーパスを利用した場合はタスクに無関係の N-gram も強化されることとなり, 誤認識を招くなどの問題が生じる.

このような単純な補間手法ではなく, 話し言葉表現のシミュレーションに基づき言語モデルを構築する手法も提案されてい

表 2 CSJ から抽出された発音変動の例

Table 2 Examples of pronunciation variations extracted from CSJ

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）
n-i → N	撥音化	毎日（マイニチ→マインチ）
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）
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）

る [25], [26]. これらは，ランダムなフィラーの挿入 [25]，あるいは Conditional Random Field (CRF) に基づく識別モデルによるフィラーの挿入 [26] により作成した擬似的な話し言葉テキストを用いて言語モデルを学習する．表 1 にあるように，対処が必要な話し言葉表現の多くはフィラーであるから，これらの手法によって言語予測能力が改善することは明らかである．しかし，逆にそれ以外の話し言葉表現に対処することはできず，性能上の限界があることも明らかである．

2.2 発音辞書

表 2 は，文献 [24] にて分析された話し言葉の発音変動の例である．日本語の場合，読みを規則的に変換することにより単語の発音辞書を生成することができるが，このような標準的な発音（baseform）に対して，実際にはさまざまな出現形（surface form）があることがわかる．このような発音のエントリを生成する手法としては，音素認識に基づき音声から発音を獲得する手法（例えば [27]）と，サブワードレベルでの決定木 [28] やニューラルネットワーク [29]，規則 [30], [31] などによるモデル化に基づき発音変動を予測する方法に大別できる．前者は実際の音声の収集と認識が必要であり，コストが大きい．これに対して後者は任意の単語に対して発音を予測できるが，これらの手法ではモデルの柔軟性・頑健性が十分ではなかったり，発音の確率が予測できないなどの問題がある．

3. 話し言葉への統計的変換モデル

本研究では，このような話し言葉現象を言語モデル・発音辞書でカバーするために，話し言葉テキスト（書き起こし）とその正書体（本稿ではこれらをパラレルコーパスと呼ぶ）を用いた，話し言葉現象の統計的抽出に基づく統計的変換モデルを提案している [23], [24]．モデル化の基本的な方針として，話し言葉の表現や発音に関してパラレルコーパス中の書き起こしと正書体で相違点を検出し，その統計頻度をもとに確率的な変換規則を定めて変換モデルを構成する．この変換モデルを用いて正

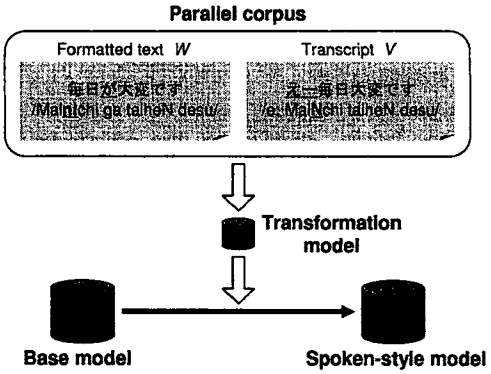

図 1 統計的変換の基本的概念

Fig. 1 Fundamental concept of statistical transformation

書体の言語モデルや発音辞書から話し言葉スタイルのモデルを生成する．以下に言語モデルと発音辞書のそれぞれの変換モデルについて詳しく述べる．

3.1 言語モデルの変換

言語モデルの変換では，正書体テキスト（W）と話し言葉テキスト（V）をそれぞれ別の言語としてとらえる．このとき，ベイズ則によりそれぞれの言語モデル $P(W)$・$P(V)$ について次の関係が成立する．

$$P(V) = P(W)\frac{P(V|W)}{P(W|V)} \tag{1}$$

ここで $P(V|W)/P(W|V)$ が変換モデルに相当する．したがって，変換モデルの学習では，V と W の表現の対応（変換パターン），およびその確率をパラレルコーパスを用いて求めることとなる．

本手法では，話し言葉変換に際して前後の単語文脈が一致することを適用の条件とし，文脈を含めた変換パターンと確率を学習する．さらに，単語文脈の一致に基づくパターンでは適用範囲が学習データに出現したものに限られるため，これに加えて前後の単語の品詞を条件とした変換パターンと確率も求める．観測された話し言葉表現を v，これに対応する正書体テキスト中の表現を w とし，頻度を N とすると，単語文脈の場合の条件付き確率 $P_{\mathrm{word}}(v|w)$ は (2) 式により定められる．

$$P_{\mathrm{word}}(v|w) = \frac{N(v|w)}{N(w)} \tag{2}$$

さらに，文脈の単語を品詞ごとにまとめて品詞文脈ごとの頻度を同様に求め，条件付き確率 $P_{\mathrm{POS}}(v|w)$ を推定する．なお，条件付き確率 $P_{\mathrm{word}}(w|v)\cdot P_{\mathrm{POS}}(w|v)$ についても，これらの手順において w と v を入れ替えることで求めることができる．

変換モデルを適用する際は，入力 N-gram 単語列について，まず単語文脈のパターンの適用を試みる．パターンが適合した場合は話し言葉の N-gram 単語列を生成するとともに，変換確率 $P_{\mathrm{word}}(v|w)/P_{\mathrm{word}}(w|v)$ をもとの N-gram 単語列の頻度に乗じることで，生成された N-gram の頻度を推定する．単語ベースのパターンが適合しない場合は品詞ベースのパターンの適用を試み，適合した場合は同様に話し言葉の N-gram 単語列を生

成して，変換確率 $P_{\mathrm{POS}}(v|w)/P_{\mathrm{POS}}(w|v)$ を用いてその頻度を推定する．

3.2 発音辞書の変換

発音辞書の変換では，入力された発音エントリの音素列に対して，音素列の書き換え規則からなる変換モデルを適用し，発音変動を含むエントリを生成してその確率を推定し話し言葉の発音辞書を出力する．

変換規則の学習には，話し言葉の発音を含む書き起こし（発音形）と，それに対応する正書法に基づく標準的な発音（読み）を利用する．本手法では形態素解析器を用いて読みと単語境界を書き起こしに付与している．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．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．

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．同一の変換にあたってはより長い文脈の規則を選択し，この規則の確率に基づいて発音エントリの確率を定める．

4. 評価実験

4.1 テストセットと実験条件

これまでに述べた手法により変換を行ったモデルについて，国会音声の認識タスクで実験的評価を行った．評価に用いたデータは，衆議院において2006年の第165臨時国会にて収録された4種類の会議の音声である．各会議の仕様を表3に示す．なお，表3における単語数にはポーズ単語も含まれている．本実験では，言語モデルに対してはこのテストセットにおけるパープレキシティと単語誤り率を，発音辞書に対しては単語誤り率を評価の指標として用いる．

言語モデル用変換モデルの学習には，2.1節で分析に利用したパラレルコーパスを用い，総数6,339のパターンを得た．うち2,310（36%）は品詞文脈のパターンである．発音辞書用変換モデルの学習にはCSJの学会・模擬講演（合計2,540講演）の書き起こしを用い，1,381個の変動規則を得た．なお，本研究で利用している形態素解析器は茶筌 Ver.2.2.3+IPADIC-2.4.4である[32]．

本実験では，言語モデルの学習データとして1999年から2005年までの7年分の衆議院会議録を用いる．会議録ではフィラーや口語表現・文末表現などの典型的な話し言葉表現が修正されているため，ベースラインモデルは文書に近いスタイルとなっている．この会議録から提案法の変換モデルにより生成された言語モデル（"Proposed"）のほかに，比較のために日本語話し言葉コーパスを重み付き線形補間（コーパス混合）したモデル（"CSJ"），およびパラレルコーパス中の書き起こしテキストを重み付き線形補間したモデル（"Transcript"）も構築し，あわせて評価を行った．これらのモデルは全てtrigramモデルである．構築したモデルの仕様を表4に示す．7年分の会議録に対して，CSJモデルでは学会・模擬講演の書き起こしの単語頻度を2倍に，Transcriptモデルではパラレルコーパスの書き起こしテキストの単語頻度を10倍にして補間を行っている．これらの重みは予備実験により選択された最適な値である．3種類の言語モデルの語彙は同一であり，そのサイズは54,321である．これによるテストセットの未知語率は0.15%となった．

表3 国会音声テストセット（2006年）の仕様
Table 3 Specifications of the 2006 congressional speech test-set

会議名	開催日	単語数	未知語率
予算委員会（Budget）	10月10日	48,966	0.13%
党首討論（Question time）	10月18日	7,884	0.05%
安全保障委員会（Security）	10月17日	46,486	0.17%
外務委員会（Foreign affairs）	10月17日	40,416	0.18%
合計・平均		143,752	0.15%

発音辞書は，言語モデル中の各単語に対して，形態素解析器（茶筌）により付与された読みから構成したものをベースライン（baseform）とする．ベースライン発音辞書のエントリ数は57,462である．これに対して，提案法による発音変動エントリ（surface form）の推定を行い，合計64,857エントリからなる発音辞書を得た．

音響モデルは，国会音声（134時間）から音素誤り最小化（MPE）学習により構築された，3,000状態・16混合のtriphone HMMである[33]．特徴量にはMFCCおよびその1次・2次差分各12次元とエネルギーの1次・2次差分の計38次元を用い，ケプストラム平均・分散正規化（CMN・CVN）および声道長正規化（VTLN）を行っている．デコーダはJulius rev.3.5.3である．長時間の発話に対処するために逐次デコーディング[34]を適用している．

4.2 言語モデル変換の評価

図2に各会議におけるパープレキシティを示す．4種類の会議のいずれも同様の傾向が観測され，平均のパープレキシティはCSJモデルが52.0，Transcriptモデルが49.1，提案法によるモデルは41.2であった．CSJモデルは，3種類の言語モデル中で際だってtrigramエントリ数が多いにも関わらず最も大きなパープレキシティとなっている．これは，無関係のコーパスを単純に補間するだけでは有効なN-gramエントリを十分に補うことができず，性能が限られることを示している．一方，TranscriptモデルではCSJモデルよりパープレキシティが削減されたが，その幅は小さい．Transcriptモデルのtrigramエントリ数（5.45M）は会議録のみの場合（5.32M）からわずかしか増えておらず，少量の書き起こしテキストの混合では話し言葉表現のN-gramのカバレージが限定的であるといえる．これに対して，提案法による変換では同一の話し言葉テキストからTranscriptモデルよりも多くのエントリを生成することがで

表 4　言語モデルの仕様
Table 4　Specifications of language models

モデル	CSJ 補間モデル ("CSJ")	書き起こし補間モデル ("Transcript")	統計的変換モデル ("Proposed")
学習コーパス	衆議院会議録 +CSJ	衆議院会議録 +書き起こし	衆議院会議録 +パラレル
総単語数	119M + 7.3M × 2	119M + 0.7M × 10	119M+0.7M
語彙サイズ	54,321		
Trigram エントリ数	7.29M	5.45M	5.86M

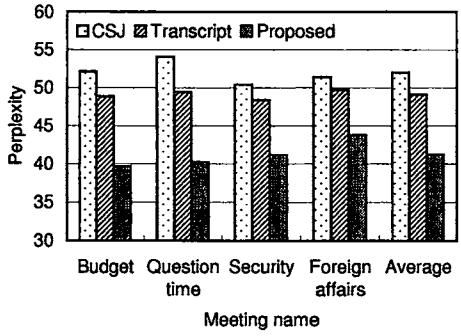

図 2　各言語モデルによるパープレキシティ
Fig. 2　Perplexity by the language models

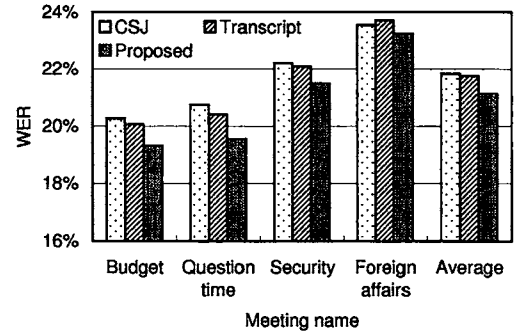

図 3　各言語モデルによる単語誤り率
Fig. 3　Word error rates by the language models

きている．最も小さなパープレキシティを得ていることからも，生成されたエントリが予測の上で有効に機能しているといえる．提案法による改善は，CSJ モデルに対して 20.7%，Transcript モデルに対して 16.0%となり，スタイルの変換が高い予測性能を実現できていることがわかる．

各会議における単語誤り率を図 3 に示す．なお，ここでは発音辞書として話し言葉変換を行っていない（baseform）辞書を利用している．図 3 より，CSJ モデルと Transcript モデルではほとんど同等の単語誤り率であるのに対して，提案法は大きく精度を改善していることがわかる．提案モデルによる改善（絶対値）は，CSJ モデルから 0.70%，Transcript モデルから 0.62%であり，いずれも有意水準 1%で統計的に有意である．提案法では特にフィラー周辺の認識誤りが改善しているが，線形補間による手法ではフィラー前後の予測は観測された N-gram のみしか行えないこともあり，さまざまな N-gram を予測可能な提案法の優位性を裏付けるものといえる．

4.3　発音変動予測の評価

図 4 に，各発音辞書による単語誤り率を示す．ここで利用した言語モデルは提案法により変換されたモデルである．この場合も全ての会議で話し言葉変換により改善の傾向を示しており，平均の改善は 0.97%であった．これは有意水準 1%で統計的に有意である．

言語モデルと発音辞書について，それぞれ話し言葉変換を行ったモデルを利用した場合の単語誤り率は平均で 20.17%となった．CSJ モデルと baseform による発音辞書の場合（21.84%）に対して 1.7%の改善がみられ，それぞれの変換が加算的に機能しているといえる．

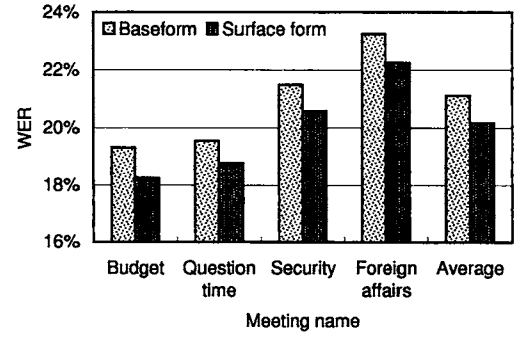

図 4　各発音辞書による単語誤り率
Fig. 4　Word error rates by baseforms and surface forms

5. おわりに

本稿では，我々が提案している言語モデルと発音辞書の統計的話し言葉変換手法について，国会音声認識における評価を行った．提案法では，話し言葉音声の書き起こしと正書体テキストからなるパラレルコーパスから統計的に学習された変換規則（変換モデル）をもとに，正書体の言語モデル・発音辞書から話し言葉スタイルのモデルを生成する．衆議院の会議音声を用いた評価において，提案法による言語モデルではパープレキシティが従来の作成法に対して 16%～21%改善し，また単語誤り率でも絶対値で 0.6%～0.7%の改善が得られた．一方，発音辞書の変換によっても単語誤り率が 1.0%削減され，双方の変換を併用した場合 1.7%の改善を得ることができた．

謝辞 本研究の一部は，総務省戦略的情報通信研究開発推進制度(SCOPE)「音声認識技術を用いた会議録及び字幕の作成支援システム」により実施された．

文　献

[1] J.L. Gauvain, L.F. Lamel, G. Adda, and M. Adda-Decker. The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task. In *Proc. ICASSP*, pp. 557–560, 1994.

[2] L.R. Bahl, S. Balakrishnan-Aiyer, J.R. Bellegarda, M. Franz, P.S. Gopalakrishnan, D. Nahamoo, M. Novak, M. Padmanabhan, M.A. Picheny, and S. Roukos. Performance of the IBM Large Vocabulary Continuous Speech Recognition System on the ARPA Wall Street Journal Task. In *Proc. ICASSP*, pp. 41–44, 1995.

[3] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi. The Design of the Newspaper-based Japanese Large Vocabulary Continuous Speech Recognition Corpus. In *Proc. ICSLP*, pp. 3261–3264, 1998.

[4] J.-L. Gauvain, G. Adda, L. Lamel, and M. Adda-Decker. Transcribing Broadcast News Shows. In *Proc. ICASSP*, pp. 715–718, 1997.

[5] P.C. Woodland, M.J.F. Gales, D. Pye, and S.J. Young. Broadcast News Transcription using HTK. In *Proc. ICASSP*, pp. 719–722, 1997.

[6] S.S. Chen, E.M. Eide, M.J.F. Gales, R.A. Gopinath, D. Kanevsky, and P.A. Olsen. Recent Improvements to IBM's Speech Recognition System for Automatic Transcription of Broadcast News. In *Proc. ICASSP*, pp. 37–40, 1999.

[7] J.J. Godfrey, E.C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proc. ICASSP*, pp. 517–520, 1992.

[8] B. Peskin, L. Gillick, N. Liberman, M. Newman, P. van Mulbregt, and S. Wegmann. Progress in Recognizing Conversational Telephone Speech. In *Proc. ICASSP*, pp. 1811–1814, 1997.

[9] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal, and A. Waibel. Recognition of Conversational Telephone Speech using the Janus Speech Engine. In *Proc. ICASSP*, pp. 1815–1818, 1997.

[10] G. Zavaliagkos, J. McDonough, D. Miller, A. El-Jaroudi, J. Billa, F. Richardson, K.W. Ma, M. Siu, and H. Gish. The BBN Byblos 1997 Large Vocabulary Conversational Speech Recognition System. In *Proc. ICASSP*, pp. 905–908, 1998.

[11] T. Hain, P.C. Woodland, T.R. Niesler, and E.W.D. Whittaker. The 1998 HTK System for Transcription of Conversational Telephone Speech. In *Proc. ICASSP*, pp. 57–60, 1999.

[12] J.S. Garofolo, C.D. Laprun, and J.G. Fiscus. The Rich Transcription 2004 Spring Meeting Recognition Evaluation. In *Proc. ICASSP Meeting Recognition Workshop*, 2004.

[13] F. Metze, Q. Jin, C. Fuegen, K. Laskowski, Y. Pan, and T. Schultz. Issues in Meeting Transcription —The ISL Meeting Transcription System. In *Proc. ICSLP*, pp. 1709–1712, 2004.

[14] N. Mirghafori, A. Stolcke, C. Wooters, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf. From Switchboard to Meetings: Development of the 2004 ICSI-SRI-UW Meeting Recognition System. In *Proc. ICSLP*, pp. 1957–1960, 2004.

[15] 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*, pp. 518–521, 2000.

[16] H. Nanjo and T. Kawahara. Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition. *IEEE Trans. Speech & Audio Proc.*, Vol. 12, No. 4, pp. 391–400, 2004.

[17] 中村篤, 大庭隆伸, 渡部晋治, 石塚健太郎, 藤本雅清, 堀貴明, エリックマクダーモット, 南泰浩. 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 (2006 年版). 情報処理学会研究報告, 2006-SLP-64-44, 2006.

[18] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu. The LIMSI 2006 TC-STAR EPPS Transcription Systems. In *Proc. ICASSP*, Vol. 4, pp. 997–1000, 2007.

[19] J. Loof, M. Bisani, Ch. Gollan, G. Heigold, Bjorn Hoffmeister, Ch. Plahl, Ralf Schluter, and Hermann Ney. The 2006 RWTH Parliamentary Speeches Transcription System. In *Proc. ICSLP*, pp. 105–108, 2006.

[20] B. Ramabhadran, Olivier Siohan, L. Mangu, G. Zweig, M. Westphal, H. Schulz, and A. Soneiro. The IBM 2006 Speech Transcription System for European Parliamentary Speeches. In *Proc. ICSLP*, pp. 1225–1228, 2006.

[21] 秋田祐哉, 河原達也. 会議音声認識のための発音モデル生成と言語モデル適応. 日本音響学会春季研究発表会講演論文集, 1-5-3, 2005.

[22] 根本雄介, 秋田祐哉, 河原達也. 会議音声の自動話題分割による単語辞書と言語モデルの適応. 情報処理学会研究報告, 2006-SLP-62-12, 2006.

[23] Y. Akita and T. Kawahara. Topic-independent Speaking-style Transformation of Language Model for Spontaneous Speech Recognition. In *Proc. ICASSP*, Vol. 4, pp. 33–36, 2007.

[24] 秋田祐哉, 河原達也. 話し言葉音声認識のための汎用的な統計的発音変動モデル. 電子情報通信学会論文誌, Vol. J88-DII, No. 9, pp. 1780–1789, 2005.

[25] H. Schramm, X.L. Aubert, C. Meyer, and J. Peters. Filled-Pause Modeling for Medical Transcriptions. In *Proc. Workshop on Spontaneous Speech Processing and Recognition*, pp. 143–146, 2003.

[26] 太田健吾, 土屋雅稔, 中川聖一. フィラーの書き起こしのないコーパスからのフィラー付き言語モデルの構築. 情報処理学会研究報告, 2007-SLP-67-1, 2007.

[27] T. Sloboda and A. Waibel. Dictionary Learning for Spontaneous Speech Recognition. In *Proc. ICSLP*, pp. 2328–2331, 1996.

[28] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos. Stochastic Pronunciation Modelling from Hand-labelled Phonetic Corpora. *Speech Communication*, Vol. 29, pp. 209–224, 1999.

[29] T. Fukada, T. Yoshimura, and Y. Sagisaka. Automatic Generation of Multiple Pronunciations based on Neural Networks. *Speech Communication*, Vol. 27, pp. 63–73, 1999.

[30] T. Imai, A. Ando, and E. Miyasaka. A New Method for Automatic Generation of Speaker-dependent Phonological Rules. In *Proc. ICASSP*, pp. 864–867, 1995.

[31] Q. Yang, J.-P. Martens, P.-J. Ghesquiere, and D.V. Compernolle. Pronunciation Variation Modeling for ASR: Large Improvements Are Possible But Small Ones Are Likely to Achieve. In *Porc. ICSLP Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, pp. 123–128, 2002.

[32] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸. 形態素解析システム『茶筌』version 2.2.3 使用説明書. Feb. 2001.

[33] 三村正人, 河原達也. 話し言葉音声認識タスクにおける音素誤り最小化学習 (MPE) の効果. 日本音響学会秋季研究発表会講演論文集, 3-Q-8, 2007.

[34] 李晃伸, 河原達也, 鹿野清宏. 話し言葉の認識のためのデコーダ Julius の改良. 日本音響学会春季研究発表会講演論文集, 1-3-15, 2001.