情報処理学会論文誌 Vol.55 No.7 1681–1694 (July 2014)

擬似生成した複数方言言語モデル混合による混合方言音声認識

平山 直樹[1,a)] 吉野 幸一郎[1] 糸山 克寿[1] 森 信介[1,2] 奥乃 博[1]

受付日 2013年11月2日, 採録日 2014年4月4日

概要：本論文では，様々な方言の混合に対応する音声認識システムを構築する．まず，単一方言音声認識の言語モデルを，大規模共通語言語コーパスから擬似生成した方言言語コーパスで学習する．擬似生成には，共通語–方言対訳コーパスから WFST（重み付き有限状態トランスデューサ）によって学習されたルールを用いる．次に，構築された各方言言語モデルを混合し，発話ごとに最適な混合比を推定しながら認識を行う．これは，実際に話される方言が純粋な単一方言ではなく，人の移動やテレビ，ラジオなどの放送の影響を受けた様々な方言の混合であると考えられるからである．この推定には，音声認識用言語モデルにおける対数尤度の値を用いる．実験により，方言音声認識用言語モデルを用いて方言音声の認識精度が向上することを確認した．また，対数尤度と音声認識精度に強い相関があること，対数尤度を最大化する混合比を発話ごとに選択することで，固定混合比の場合と比較して音声認識精度が向上することを確認した．

キーワード：方言音声認識，混合比推定，方言対訳コーパス

Dialect-mixed Speech Recognition by Mixing Simulated Multiple Dialect Language Models

Naoki Hirayama[1,a)] Koichiro Yoshino[1] Katsutoshi Itoyama[1] Shinsuke Mori[1,2] Hiroshi G. Okuno[1]

Received: November 2, 2013, Accepted: April 4, 2014

Abstract: This paper designs and implements an automatic speech recognition (ASR) system that accepts a mixture of various kinds of dialects. The language model for a particular dialect is trained on a dialect language corpus simulated from a large common language corpus. The simulation is carried out with a weighted finite-state transducer (WFST) trained on a parallel corpus of a dialect and common language. The resulting system recognizes dialect utterances with a mixture of dialect language models by estimating the optimal dialect mixing proportion for each utterance. Since actually-spoken dialect is not a purely single dialect but a mixture of various dialects, influenced by communication in daily lives and broadcasting such as television and radio, estimating optimal dialect mixing proportion, that is, what maximizes the value of log-likelihood for the input utterance, is critical in ASR. Experiments showed that recognition accuracy improves by using the dialect language model, that log-likelihood and recognition accuracy are highly correlated, and that recognition accuracy improves by choosing the dialect mixing proportion that maximizes log-likelihood for each utterance, compared to a fixed dialect mixing proportion.

Keywords: dialect speech recognition, mixing proportion estimation, dialect parallel corpus

1. はじめに

近年，計算機による自動音声認識技術は実時間での音声認識精度が劇的に向上しており，実世界応用が盛んになっている．国会議事録作成 [1] や生放送のニュース番組におけるリアルタイム字幕放送 [2] などの応用，さらにはスマー

[1] 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[2] 京都大学学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University, Kyoto 606–8501, Japan
[a)] hirayama@kuis.kyoto-u.ac.jp

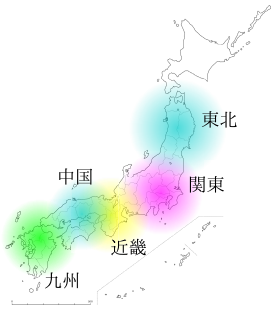


図 1 方言混合のイメージ
Fig. 1 Schematic usage of dialect mixtures.

トフォンとそれに付随する音声認識アプリケーションの普及により，音声認識技術は人々に身近な存在となった．しかし，不特定多数の話者を対象とするものを含め，話者ごとの言葉遣いの差異，とりわけ方言を考慮した音声認識システムは少ない．たとえば，裁判所での自動調書作成システム開発では，証人や裁判員の方言音声認識が大きな課題となっている [3]．様々な方言の差異を持つ不特定多数の話者の利用を想定する音声認識システムでは，各方言に個別に対応するのではなく，単一のシステムで様々な方言を対象とすることが不可欠である．

本論文では，話者が実際に話す方言は様々な方言の混合であると見なし，方言の混合比を推定しながら音声認識を行うシステムを構築する．方言は，言葉の地理的変異 [4] (pp.2–3)，すなわち地域ごとに異なる特徴を持つ言葉である．しかし，地域間の人の移動・交流があることから，方言は地理的境界で分離できるものではない [5]（p.71）．実際に話者が使用する方言は，各地で話される方言が混合したもの（**図 1**）と考えられ，近接地域であっても混合の度合いは少しずつ異なる．共通語 [6] と呼ばれる，異なる方言話者同士で理解できる言葉を指す概念があるが，言語調査において，元来の方言と異なるが東京方言とも一致しない言葉を表したことが起源であり [7]（p.204），方言の一部を構成するものと考えられる．本論文では，共通語も方言の 1 つとして混合対象になるものとする．

方言の特徴は，発音，語彙，語順の 3 タイプに大別される．発音の例として，英語における *“marry”*，*“merry”*，*“Mary”* という 3 単語の発音があげられる [8], [9]．3 単語はすべて同じように発音される地域もあれば，すべての発音が区別される地域もある．語彙の例として，駅のプラットフォームで乗客に段差への注意を促す *“watch your step”* や *“mind the gap”* という表現があげられる [10]．アメリカでは主に前者が使用され，イギリスでは主に後者が使用される．語順の例として，アメリカで *“next Tuesday”* と表現されるものが，カナダでは *“Tuesday next”* と表現されることがあげられる [11]．本論文では，方言の特徴として発音および語彙を対象とする．また，方言間で音素集合および音響的特徴は同一であると仮定し，発音と語彙の変化をいずれも使用語彙の変化ととらえる．

方言の音声認識と解析を行うにあたり，以下の 3 課題について解決する必要がある．

課題 1：方言言語モデル学習コーパス
課題 2：混合方言発話の音声認識
課題 3：方言混合比を推定する手法および目的関数

課題 1（方言言語モデル学習コーパス）は，方言音声認識には不可欠な大規模言語コーパス収集において，共通語と異なり新聞記事や Web 文書の利用が困難であることを指す．本論文では，大規模な共通語コーパスから方言コーパスを擬似生成する手法を開発し，この課題を解決する．課題 2（混合方言発話の音声認識）は，特定地域の方言の言語モデルを構築しても，対応する純粋な方言だけが音声認識対象となり，実際に話される混合した方言をうまく扱えないことを指す．本論文では，各方言に対応する言語モデルを重み付き混合し，様々な方言をカバーする言語モデルを構築する．課題 3（方言混合比推定）は，前述の言語モデル混合比を，音声認識精度を最大化するよう適切に推定することを指す．音声認識タスクにおいて一般に正解文は未知であり，音声認識結果から音声認識精度を知るのは不可能である．そこで，音声認識精度に代わる混合比決定の目的関数を定める必要がある．特に課題 1，課題 2 の解決方法は，これまでの方言音声認識手法では取り組まれていない新規手法である．

本論文は以下のように構成される．2 章で，方言音声認識に関する関連研究についてまとめる．3 章で，本論文で扱う方言音声認識システムの大枠について述べる．4 章で，言語コーパスの方言変換の詳細について述べる．5 章で，音声認識結果からの方言混合比推定について述べる．6 章で本手法の有効性を評価する．最後に，7 章で本論文の結論を述べ，今後の課題をあげる．

2. 関連研究

方言音声認識は，従来から取り組まれてきた課題であるが，その多くは様々な点で実用性，汎用性を欠いていた．

Lyu ら [12] は，中国語の方言に対応した音声認識システムを，単語発音辞書に付与する発音を変化させることで構築している．この手法では，方言ごとに漢字と発音との対応を変えて，各単語に対して方言ごとに異なる発音を付与する．しかし，この対応辞書は人手で構築したものであり，方言ごとに同様の作業を必要とするうえに，語彙が増える度に方言母語話者の協力を得て辞書を更新する必要がある．そこで本論文では，少量の共通語–方言対訳コーパスを用いて，共通語単語から統計的に対応する方言発音を推定する手法を提案する．

また，多種多様な方言認識を行うために，話者方言を推定してそれに適合したモデルを選択するという戦略が考えられる．Ching ら [13] は，中国語 2 方言（北京語・広東語）

について，パワー，ピッチ，話速といった音響的特徴に基づいた識別手法を提案している．英語についても同様に，アメリカ南北の方言（Miller ら [14]）やアメリカ，イギリス，オーストラリアの 3 カ国の方言（Chitturi ら [15]）で音響的特徴での識別が試みられている．しかし，音響的特徴に基づき識別器を学習するには，大量の音声データが必要であり，特に話者人口が少ない方言に対して適用することが難しい．また，これらの手法では語彙そのものの変化をとらえられない．さらに，方言を択一問題として扱っており，複数の方言の特徴をあわせ持つ話者に対していずれか 1 つの方言のみを選択してしまう．そこで本論文では，方言の語彙変化を各語彙の発音（音素表記）の変化ととらえ，共通語言語コーパスを方言発音に変換することで方言を擬似生成する手法を導入する．また，前述の各方言に対応した言語モデルを適切な比率で混合し，複数の方言が混合した場合に対応する言語モデルを構築する．そのうえで，適切な比率の自動推定により，話者方言の推定を可能とする．

3. 手法の概略

3.1 処理手順

本論文で扱う方言音声認識，およびその結果を用いた方言混合比推定の流れを**図 2** に示す．本手法は大きく学習フェーズと認識フェーズに分かれる．学習フェーズ（図 2 上）では，共通語–方言対訳コーパスを用いて大量に利用できる共通語コーパスを方言発音へと変換し，各方言に対応する言語モデルを学習する．認識フェーズ（図 2 下）では，方言言語モデルを混合して音声認識を行い，その結果を用いて方言混合比を推定する．

学習フェーズは以下の 3 手順で実現される．

手順 1：方言変換ルール学習

手順 2：方言言語コーパスの擬似生成

手順 3：方言言語モデルの作成

手順 1（方言変換ルール学習）では，共通語–方言対訳コーパス [16] から単語単位での発音対応について統計的ルールを学習する．本論文では方言間で語順変化がないと仮定し，

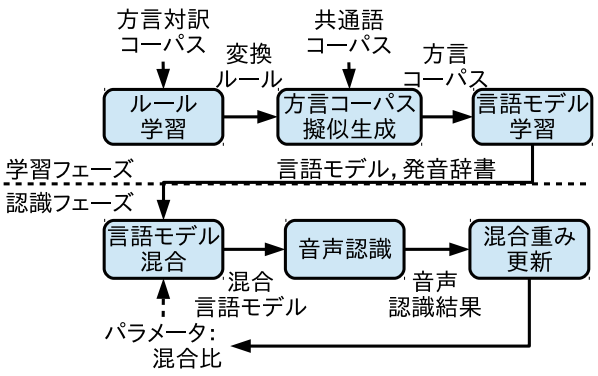


図 2 方言音声認識における学習・認識の流れ

Fig. 2 Training and recognition phases in our dialect speech recognition.

共通語–方言間単語単位の対応のみを考慮したルールを作成する．このルールは，重み付き有限状態トランスデューサ（Weighted Finite-State Transducer，WFST）[17] による音素列対 n-gram モデルで表現する．音素列対とは，共通語単語単位で，共通語と実際の方言の発音を表す音素列を対にしたものである．音素列対 n-gram モデルにより，当該単語だけでなく，前後の単語の発音対応にも依存した発音変換が行える．WFST を利用する理由として，発音のバリエーションを考慮して複数候補を確率付きで出力できることがあげられる．たとえば，近畿方言で「行けない」に対応する表現は「行けん」「行けへん」「行かれへん」など複数あり，これら複数の候補を認識できる必要がある．入力された共通語音素列に対して，単語単位の方言音素列の候補と，各候補の文全体に対する対数尤度を出力するように WFST を設計する．方言の混合を考える場合は，対象となるそれぞれの方言に対して，同様に共通語–方言対訳コーパスを用いて WFST の学習を行う．本論文では，WFST を扱うライブラリとして OpenFst [*1] を利用した．

手順 2（方言言語コーパスの擬似生成）では，学習したルールを用いて共通語言語コーパスを処理し，各単語に方言発音を付与する．ここで，方言発音は共通語言語コーパスの各単語にタグとして付与する．これにより，認識結果を利用するアプリケーションでは，共通語のみを前提として言語理解部を作成することができる．

手順 3（方言言語モデルの作成）では，各方言に対して n-gram モデルと発音辞書（以下，両者を総称して言語モデルと呼ぶ）を出力する．ここで，方言では 1 つの単語に対して複数の発音がなされる場合を考慮して，共通語単語を 1 つのクラスとするクラス n-gram モデルを導入する．発音辞書は，言語コーパスに付与された方言発音を集計し，各単語に対する発音と確率（クラス内確率）の対を求めて出力する．

認識フェーズ（図 2 下）では，学習した方言言語モデルを適当な比率で混合して方言発話を音声認識し，その結果を用いて音声認識精度を最大化する比率（発話の方言混合比）を求める．この混合比と音声認識精度には互いに依存関係がある．適切な混合比で言語モデルを話者の方言に適応できれば，音声認識精度が向上する．また，認識精度が高くなれば，その認識結果を用いてより正確な混合比推定が可能となる．そこで，音声認識結果を用いて最適な混合比を推定する枠組みを提案する．本論文では，各混合比に対する認識結果が得られた場合に，認識精度を向上させる混合比推定の目的関数を提案する．4 章で学習フェーズについて，5 章で認識フェーズについて述べる．

[*1] http://www.openfst.org/

3.2 変換元共通語言語コーパス

音声認識精度を向上させるには，発話のドメインに適合した言語モデル学習コーパスの利用が重要である．本論文では，日常発話の認識を対象とするため，新聞記事など書き言葉中心のコーパスではなく，話し言葉の文体に近く，くだけた表現も含まれる「Yahoo! 知恵袋」コーパス[*2]を用いた．Yahoo! 知恵袋コーパスは Web テキストの一種であり，各文書に与えられたカテゴリ・サブカテゴリの分類情報を用いることで，認識対象ドメインに近いカテゴリの文書のみを選択して言語モデル学習を行うことが可能である．しかし Web テキストには，インターネットスラングやアスキーアートなど，日常の話し言葉とは異なる特有の表現も含まれる．そこで，パープレキシティを用いたコーパスフィルタリング手法 [18] を適用し，このような言語モデル学習に不適な文を取り除く．この手法は，基準となる言語モデルを用いて各文の単語当たりパープレキシティの平均値を計算し，その値の小さい文から順にコーパスに採用するものである．本論文では，基準となる言語モデルの学習データとして，日本語書き言葉均衡コーパス（BCCWJ）[19] コアデータのうち，ブログデータを利用した．

4. 方言言語モデルの構築と 1 方言音声認識

本章では，方言言語モデル構築のための，共通語–方言対訳コーパスからの方言変換ルールの構築手法，およびそれを用いた方言言語コーパスの擬似生成手法について述べる．

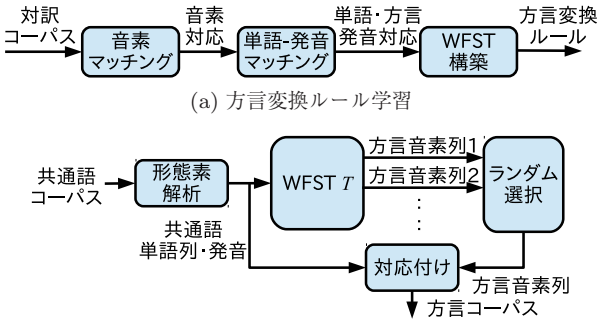
4.1 方言変換ルール学習

方言発話の書き起こしと共通語訳が対になった共通語–方言対訳コーパスから，各共通語単語に対応する方言発音を自動で求める手法を提案する．処理の流れを**図 3** (a) に示す．ここでは，実際に利用できる対訳コーパスの規模が小さいので，コーパス擬似生成時にルールが過剰に適用されて，実際の方言では使用されない表現が付与される可能性がある．そこで，これを抑制するために，コーパス擬似生成に利用する共通語–方言変換ルールを音素や音節単位ではなく単語単位で抽出する．

まず，共通語文と方言発音の対からなる対訳コーパス [16]（**図 4** (a)）を音素表記に統一し，動的計画法に基づくマッチング（DP マッチング）により対応を取る（図 4 (b)）．その際，共通語音素列と方言音素列の Levenshtein 距離（挿入・削除・置換の各コストは 1）を最小化 [20] する対応を求める．対訳コーパスの共通語文はあらかじめ形態素解析ツール（KyTea [*3][21]）を用いて単語分割してから利用する．

続いて，音素単位での音素列対から，共通語単語単位で

[*2] Yahoo! JAPAN，国立情報学研究所（NII）提供．
[*3] http://www.phontron.com/kytea/



(a) 方言変換ルール学習

(b) 方言言語コーパス擬似生成

図 3 学習フェーズ（図 2 上）の詳細な流れ
Fig. 3 Details of the training phase.

```
あなた は どこ に 住ん で いる の
アンタドコスンデルン
```

(a) 共通語–方言対訳コーパスの例．上段が共通語（単語境界あり），下段が方言（発音を表すカナ列のみ）

```
a n a t a | w a | d o k o | n i | ...
|  /  |  |     |  |  |  |
a N   t a        d o k o       ...
```

(b) 共通語–方言間の音素列対．共通語の単語境界を記号 | で示す

```
a_n_a_t_a+a_N_t_a  w_a+NULL  ...
```

(c) 単語単位の音素列対．各音素を記号 _ で区切り，共通語音素列と方言音素列を記号 + で区切って示す．NULL は長さ 0 の音素列

図 4 単語単位方言変換ルール構築の手順
Fig. 4 Three steps for constructing word-wise rules of dialect transformation.

の音素列対を求める．共通語単語ごとに，共通語発音の各音素に対応する方言音素列を連結したものを，その共通語単語に対応する方言音素列とし，音素列対として出力する（図 4 (c)）．この手法では，

(1) 方言間の部分的な音の変化であれば，前後の音はほぼ共通語と一致する，
(2) 単語そのものが異なる場合でも，前後の単語はほぼ共通語と同一である，

の 2 点を仮定し，音素単位のアライメントを手がかりとして単語単位のアライメントを求める．ただし，方言の音節が共通語文の単語境界をまたいで対応する場合には，またいだ複数の共通語単語列を連結して 1 語と扱う．たとえば，共通語の「これ は」と方言の「こら」が対応するときには，この場合，「これ-は」の 2 語に方言「こら」が対応していると見なす．

最後に，単語単位の音素列対を利用して，共通語音素列を方言音素列に変換する WFST を構築する．この WFST は，3 つの WFST L, T_1, T_2 から構築される．L は，音素列対の文脈依存性を音素列対 3-gram モデルで表現したもの [22] である．対訳コーパスの各文を図 4 (c) で示す単語単位の音素列対表記に変換し，音素列対 3-gram モデルを

学習することで，音素列対が与えられると音素列対の文全体での対数尤度を出力する WFST を構築する．ある文に対する単語単位の音素列対表記 $p_1, p_2, \ldots, p_N$（p_i は音素列対）の尤度は

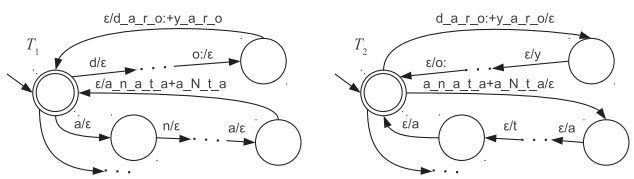
$$P(p_1 p_2 \ldots p_N) = P(p_1)P(p_2|p_1)P(p_3|p_1 p_2) \ldots P(p_N|p_{N-2}p_{N-1}) \quad (1)$$

と n-gram 確率の積で表される（実際にはバックオフも考慮する [23]）ので，対数尤度は

$$\log P(p_1 p_2 \ldots p_N) = \log P(p_1) + \log P(p_2|p_1) + \ldots + \log P(p_N|p_{N-2}p_{N-1}) \quad (2)$$

と表される．WFST は，遷移した枝に付与された重みを加算していき，全入力記号を読み終わった際の最終的な重みを出力する．そこで，n-gram モデルを有限状態機械で表現し，各遷移枝に重みとして対数 n-gram 確率を付与しておくことで，WFST がすべての音素列対を読み終わった際に文全体に対する対数尤度を出力する設計としている．

T_1 は，共通語音素列を単語単位の共通語と方言の音素列対に変換する WFST である．対訳コーパスに現れた共通語音素列と方言音素列の対ごとに，共通語音素列を 1 音素ずつ入力すると音素列対が 1 つ出力されるように遷移を加える（**図 5** (a)）．これにより，T_1 は入力された共通語音素列に対して可能な方言音素列との対応を列挙する．T_2 は，共通語と方言の音素列対を方言音素列に変換する．つまり，音素列対の入力に対してその方言音素列部分を 1 音素ずつ出力する遷移を持つ WFST である（図 5 (b)）．これらを用いて，WFST $T = T_1 \circ L \circ T_2$ で発音変換と各変換候補の対数尤度計算を行う．ここで $\circ$ は WFST の合成演算を表す．L は音素列対に対する方言変換を定義しているが，共通語・方言発音は音素列として WFST に入出力される設計であり，入出力音素列を音素列対表現に変換するのが T_1, T_2 である．共通語文の単語単位の発音をそのまま WFST で変換する設計にしないのは，対訳コーパスに含まれない未知語が入力された場合に対応できないからである．単語は事実上無限にあるが，音素セットは有限であるので，音素列の変換に加えて，各音素を変換せずにそのまま出力するというルールを追加することで，未知語が入力された場合でも変換せずに出力することが可能となる．



(a) 共通語音素列から音素列対　(b) 音素列対から方言音素列

図 5 方言変換 WFST の構造．ε は入出力記号なしを示す

Fig. 5 Structure of WFSTs for dialect transformation. Symbol ε denotes an empty input or output symbol.

4.2 方言言語コーパスの擬似生成

先に構築した WFST を用いて，方言言語コーパスの擬似生成を行う．共通語コーパスから方言コーパスを出力する手順を図 3 (b) に示す．まず，共通語コーパスの各文を形態素解析して単語列と各単語の読みを得る．読みを音素列として WFST T に入力し方言変換を行うと，単語単位の方言音素列候補および各候補の文全体での対数尤度の対が複数得られる．続いて，得られた変換候補から 1 つをランダム選択する．これにより，尤度の大きい変換候補のみが出力されるのを防ぎ，様々な表現をカバーする．ここでは，各変換候補の選択確率は尤度に比例する確率とする．すなわち，共通語音素列 x の入力に対して，方言音素列候補 $y_1, y_2, \ldots$ と対応する尤度 $L(y_1|x), L(y_2|x), \ldots$（$i < j$ ならば $L(y_i|x) \geq L(y_j|x)$ となるようにソートする）が得られたとき，各方言音素列 y_i の選択確率は

$$P(y_i|x) = \frac{L(y_i|x)}{\sum_j L(y_j|x)} \quad (3)$$

となる．しかし，可能な音素列対の組合せは一般に無限個になるので，すべての候補に対する尤度計算は不可能である．そこで，尤度の上位 n 個の候補のみを選択対象とする近似を行う．式 (3) の選択確率を，式 (4) で近似する．

$$P(y_i|x) \approx \begin{cases} \frac{L(y_i|x)}{\sum_{j=1}^{n} L(y_j|x)} & \text{if } i = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

本論文では，すべての方言に共通で $n = 5$ とした．最後に，方言音素列を共通語単語列に対応付けてタグとして付与したものが，方言コーパスとなる．WFST により複数単語にまたがって対応する方言音節が出力された場合には，それらを区切り文字を介して連結し 1 語として方言コーパスに出力し，音声認識時にこの単語が認識されればこれらの複数単語が認識されたものと見なす．前節の「これ は」と「こら」の例であれば，方言コーパス出力時に「これ-は」という共通語単語に「こら」を表す方言音素列をタグとして付与し，この単語が認識されれば「これ は」の 2 語が認識されたものとする．

上記のコーパス擬似生成手法を用いて得られるコーパスの例を示す．共通語言語コーパスに含まれる文

これ/コレ は/ワ そう/ソー いう/ユー 意味/イミ で/デ は/ワ ない/ナイ と/ト 思い/オモイ ます/マス 。/NULL

に対して，各単語の読みを音素列に変換し，各方言の対訳コーパスから学習した WFST で得られた方言音素列を元の単語に割り当てる．近畿方言対訳コーパスを用いると

これ-は/k_o_r_a そう/s_o: いう/y_u: 意味/i_m_i で-は/y_a ない/n_a_i と/t_o 思い/o_m_o_i ます/m_a_s_u 。/NULL

という文が生成された．肥筑方言対訳コーパスを用いると

これ-は/k_o:_r_a そう-いう/s_u_gy_a_y_u: 意味/i_m_i で-は/j_a ない/n_a_k_a と/t_o 思い/o_m_o_i ます/m_a_s_u 。/NULL

という文が生成された．実際には，文脈にそぐわない変換が起こることもある．たとえば，

これ/コレ は/ワ ない/ナイ と/ト 思い/オモイ ます/マス 。/NULL

という共通語文章に対して，近畿方言対訳コーパスによる変換結果は

これ/k_o_r_e は/w_a **ない/N** と/t_o 思い/o_m_o_i ます/m_a_s_u 。/NULL

となった．近畿方言で助動詞としての「ない」は「ン」といわれる（例：「知らない」→「知らん」）が，形容詞の「ない」は「ン」とはならないので，太字部分の変換は誤りである．このような変換誤りについては，次節の言語モデル学習時に対策を行う．

4.3 方言言語モデル学習

前節で生成された方言言語コーパスを用いて，統計的言語モデル（単語 n-gram モデル）と単語発音辞書を生成する．方言言語コーパスにおいては，同じ共通語単語に対して多くの種類の方言音素列が付与される可能性がある．それらをすべて別々の単語エントリと考えた場合，出現回数の少ないエントリが大量に生成され，語彙のカットオフを行った際にこれらのエントリが除外されてしまう．さらに，4.2 節で述べた変換誤りが発生する場合でも，他の文の同じ単語には正しい発音変換結果が付与されている場合が多い．そこで，同じ共通語単語を方言発音にかかわらず同一視し，発音確率を持つようにすることで誤りをカバーすることを考える．本論文では共通語単語をクラスとし，各クラスが複数種類の方言発音を含むクラス n-gram モデル [24] を構築する．各方言発音のクラス内確率は，方言言語コーパスの擬似生成により出現した方言発音の回数を共通語単語ごとに集計し，総和が 1 になるよう正規化して求める．すなわち，共通語単語 w の方言言語コーパスにおける出現回数を $\#(w)$，そのうち方言発音 y が付与された回数を $\#(y|w)$ とすると，x に対する y のクラス内確率 $P_c(y|w)$ は

$$P_c(y|w) = \frac{\#(y|w)}{\#(w)} = \frac{\#(y|w)}{\sum_y \#(y|w)}. \tag{5}$$

で定義される．前節の方言音素列選択確率と同様に，確率の小さい発音まですべてを発音辞書に記述するのは非効率なので，各共通語単語で出現回数が上位 m 個に入る方言発音のみを残す．すなわち，付与された方言発音を出現回数の大きい順に $y_1, y_2, \ldots$ とし，式 (5) を式 (6) で近似する．本論文の以下の実験では $m = 5$ とした．

$$P_c(y_j|w) \approx \begin{cases} \frac{\#(y_j|w)}{\sum_{j=1}^{m} \#(y_j|w)} & \text{if } j = 1, 2, \ldots, m \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

以上によりクラス n-gram モデルと単語発音辞書が生成された．

ここまでの手順で，方言音声認識に必要な言語モデルが構築できた．音声認識エンジンの言語モデルを差し替えれば，共通語の場合と同様にして方言音声認識が可能となる．クラス n-gram モデルは各単語に付与された方言発音を外した文章から生成するので，複数単語が連結されている可能性があることを除けば，共通語言語コーパスから構築したものと同一である．単語発音辞書は，各単語に対して以下のようなエントリを追加したものである．ここに示したのは，共通語「ない」に対して，確率が上位 $m = 5$ 個の発音を出力した近畿方言の Julius 用単語発音辞書のエントリである．各行が発音を表し，空白区切りの 2 列目が対数発音確率（常用対数），5 列目以降が音素列である．

```
ない/ナイ @-1.434385 ない/ナイ [ない/ナイ/n_a:_e:] n a: e:
ない/ナイ @-0.468007 ない/ナイ [ない/ナイ/n_a_i] n a i
ない/ナイ @-0.828831 ない/ナイ [ない/ナイ/n_e:] n e:
ない/ナイ @-2.903801 ない/ナイ [ない/ナイ/n_a:] n a:
ない/ナイ @-0.176948 ない/ナイ [ない/ナイ/N] N
```

5. 方言混合比の推定

4 章で構築した単一方言言語モデルを混合し，発話に現れる様々な方言やその混合状態を扱えるようにする．認識フェーズ（図 2 下）において，話者の方言混在比が未知の状態から，音声認識・方言混合比推定を繰り返し行うことによって方言音声認識を行う．

本論文では，話者方言をいくつかの方言の重み付き混合と仮定し，各方言に対する単一方言言語モデルを混合して話者方言を認識する．同じ地域に在住する人でも，自身や両親・親類の出身地などにより，各地の方言が話し言葉に影響する程度は異なると考えられるからである．

言語モデルの適切な混合比を推定するためには，以下の 2 つの課題を解決する必要がある．

(1) 音声認識結果と混合比の相互依存の解決
(2) 混合比推定のための目的関数の設計

(1) 音声認識結果と混合比の相互依存の解決のために，音声認識と方言混合比推定を同時に行う枠組みを導入する．(2) 混合比推定のための目的関数の設計は，音声認識精度が計算できない実際のアプリケーションで，音声認識精度最大化を行うためのものである．本論文では，音声認識器が出力する対数尤度の利用を提案する．

各方言言語モデルの混合比率が既知の場合に，混合方言言語モデルを作成する手法について述べる．

単語（クラス）n-gram モデルは，複数方言に対応させるために，各言語モデルの n-gram 確率を重み付き平均する

ことにより言語モデルの混合を行う．方言 d に対する単語 n-gram 確率 $P_{c,d}(w_i|w_{i-n+1}^{i-1})$ とモデル混合比 r_d を用いて

$$P_{c,\mathrm{mix}}(w_i|w_{i-n+1}^{i-1}) = \sum_d r_d P_{c,d}(w_i|w_{i-n+1}^{i-1}), \quad (7)$$
$$\text{s.t.} \quad \sum_d r_d = 1,\ r_d \geq 0$$

として，与えられたモデル混合比に対応する言語モデル（n-gram 確率 $P_{c,\mathrm{mix}}(w_i|w_{i-n+1}^{i-1})$）を生成する．式 (7) にバックオフ確率も考慮して計算を行う [23]．

また，単語発音辞書に付与する共通語単語ごとの発音確率（クラス内確率）についても混合を行う．言語モデルの重み付き混合とは，元の言語モデル学習コーパスにおける出現頻度の重み付き和である [23]．これをふまえて，式 (5) に基づいてクラス内確率を計算する．方言 d において，共通語単語 w，および w に対する方言発音 y の各出現回数をそれぞれ $\#_d(w), \#_d(y|w)$ とすると，式 (5) は

$$P_c(y|w) = \frac{\#(y|w)}{\#(w)} = \frac{\sum_d r_d \#_d(y|w)}{\sum_d r_d \#_d(w)} \quad (8)$$

となる．ここで，同一の共通語言語コーパスを用いて各方言コーパスを擬似生成する場合，$\#_d(w)$ は w のみの関数 $C(w)$ となる．この結果，式 (8) は

$$P_c(y|w) = \frac{\sum_d r_d \#_d(y|w)}{\sum_d r_d C(w)} = \frac{\sum_d r_d \#_d(y|w)}{C(w)}$$
$$= \sum_d r_d \frac{\#_d(y|w)}{C(w)} = \sum_d r_d P_{c,d}(y|w) \quad (9)$$

となる ($\sum_d r_d = 1$)．混合後の言語モデルの単語発音辞書に付与する発音確率は，元の発音確率の r_d による重み付き平均となる．

このように作成された混合方言言語モデルは，各方言言語モデルに含まれる方言特有の語彙を含んでおり，様々な方言による発話を認識可能にする．ただし，適切な比率で混合しなければ，実際に話されている方言との乖離が生じるので，認識精度が向上するとは限らない．そこで，実際に方言発話を高い精度で認識するには，混合比率 r_d の組をうまく推定することが主要な課題となる．

本論文では，文献 [25], [26] での考え方に基づき，同じ発話を複数の混合方言言語モデルで認識し，そのときの対数尤度の値が最大となる認識結果を出力する手法をとる．前述のように，各方言言語コーパスを同一の共通語言語コーパスから擬似生成し，かつ音響モデルを方言混合比にかかわらず共通とすれば，尤度に影響を及ぼす要素は各共通語単語に付与された方言発音の差異のみとなる．さらに，方言発音には各共通語単語内で和が 1 となるように確率を与えているので，共通語言語モデルと混合方言言語モデル間での尤度比較が可能であり，尤度比較を利用した単語信頼度の計算も試みられている [27]．以上から，同一発話を認識する場合において，対数尤度の大小が混合方言言語モデルの発話への適応度の指標となると考えられる．実験では，まず対数尤度の値と認識精度の相関を調べ，本手法の適用が妥当であることを確認したうえで，対数尤度から推定した混合比で作成した混合方言言語モデルによる音声認識精度を単一方言・共通語言語モデルの場合と比較する．

6. 実験

本章では，方言言語モデルの導入と，その混合による音声認識精度の向上を確認する．また，方言混合比の推定による方言音声認識精度に関して以下の 2 点を確認する．

(1) 音声認識エンジンの出力する対数尤度は音声認識精度と相関があり，対数尤度が方言混合比推定に利用できること
(2) 実際に発話ごとに対数尤度による方言混合比推定を行い，音声認識精度が向上すること

6.1 実験条件

6.1.1 方言・話者

対象とする話者の方言 [28] は以下の 5 種類とし，方言ごとに 5 名の話者を選定した．括弧内は本実験における話者の出身都道府県（18 歳まで在住していた都道府県）である．

- 東京方言（東京都・埼玉県）
  本実験では，東京方言は共通語と同一視し，東京方言言語モデルとして通常の共通語言語モデルを用いた．
- 近畿方言（大阪府・兵庫県）
- 肥筑方言（福岡県・熊本県）
- 北奥羽方言（青森県・山形県）
- 東山陽方言（広島県・岡山県）

各話者に全話者共通の共通語文章 100 文を提示し，話者は方言訳を作成してテキストに起こした上で，それを読み上げた．ただし，東京方言話者は提示した文章をそのまま読み上げた．音声認識エンジン Julius[*4] [29] での認識結果を元の共通語文章と比較して，音声認識精度を計算した．

6.1.2 モデル学習データ

以下に方言変換ルール，方言言語モデルおよび音響モデルの学習に用いたデータについて述べる．

方言変換ルール学習には，国立国語研究所が刊行している共通語–方言対訳コーパス [16] を利用した．各方言の対訳コーパスとして，選定話者の出身都道府県に対応する 2 府県のデータを利用（共通語換算で 0.9–1.3 万語）して変換ルール学習を行った．

言語モデル学習は以下のデータやパラメータを用いて行った．方言変換元の共通語言語コーパスは，Yahoo!知恵袋コーパス（第 2 版）のうち，「日常生活」カテゴリに属する質問および対応する回答から選択（3.2 節参照）した 300

[*4] http://julius.sourceforge.jp/

万文（71.2 百万語）を用い，ここから擬似生成した方言言語コーパスを言語モデルの学習に用いた．クラス n-gram モデルは $n = 3$[*5]とし，11 回以上出現した単語を語彙に採用した（語彙サイズ，すなわちクラス数は 42,845 であった）．2-gram，3-gram のカットオフは行わなかった．n-gram 出現頻度のスムージングには Kneser-Ney 法 [30] を用いた．

音響モデル学習は以下のデータやパラメータを用いて行った．音響モデルは triphone HMM（Hidden Markov Model）とし，状態数は 2,000，出力確率分布は 64 混合 GMM（Gaussian Mixture Model）とした．特徴量としては MFCC（Mel-Frequency Cepstrum Coefficient）をベースとして，MFCC 12 次元，ΔMFCC 12 次元，ΔPower 1 次元の計 25 次元を用いた．特徴量計算のフレーム幅は 25 msec，フレーム間隔は 10 msec とした．音素セットは，文献 [31] で用いられているものと同じ 42 種類（文頭・文末無音を除く）からなるものを用いた．モデル学習用音声データとしては，話し言葉認識という観点から認識精度を上げるために，日本語話し言葉コーパス（CSJ）[32] の 500 話者の講演音声（70.2 時間）を用いた．ただし，CSJ だけでは出現しない triphone が生じることから，新聞記事読み上げ音声コーパス（JNAS）[33] の 308 話者の音素バランス文読み上げ音声（23.3 時間）を併用し，すべての triphone のサンプルが確保できるようにした．この音響モデルはある方言に特化したものではなく，すべての方言話者について同一の音響モデルを用いている．

6.1.3 比較対象言語モデル

実験では，従来の共通語言語モデルに加え，単一方言言語モデル（4 章，東京方言を除く），混合方言言語モデル（5 章）を用いて音声認識を行った．混合方言言語モデルについて，混合比は連続量であり，すべての混合比を試すことはできない．そこで，各方言の混合比を 100/M%刻み（$0 < M \leq 100$, M は 100 の約数）の離散値とし，比較する言語モデルの個数を絞る．K 個の方言 $D = \{d_1, d_2, \ldots, d_K\}$ をそれぞれ $r_1, r_2, \ldots, r_K$ の混合比で混合するとすると，作成すべき混合比の組合せの集合 P は

$$P = \left\{ (r_1, r_2, \ldots, r_K) \middle| \begin{array}{c} \sum_{k=1}^{K} r_k = 1, \\ r_k \in \{m/M \,|\, m = 0, 1, \ldots, M\} \\ (k = 1, 2, \ldots, K) \end{array} \right\} \tag{10}$$

と定式化できる．それぞれの混合比に対して混合方言言語モデルを作成する必要があり，その総数は

$$|P| = \binom{M+K-1}{M} = \frac{(M+K-1)!}{M!(K-1)!} \tag{11}$$

で表される．たとえば，$K = 3$，$M = 5$（3 方言，20%刻み）の場合に作成すべき言語モデルの混合比を図示すると図 6 のようになり，言語モデルの総数は

$$|P| = \frac{(5+3-1)!}{5!(3-1)!} = \frac{7!}{5!2!} = 21 \tag{12}$$

となる．本論文では，6.1.1 項で述べた 5 方言を扱い，混合比を 20%単位で扱う．すなわち，$K = 5$，$M = 5$ となり，

$$|P| = \frac{(M+K-1)!}{M!(K-1)!} = \frac{9!}{5!4!} = 126 \tag{13}$$

だけの混合方言言語モデルを同じ発話に対して試し，対数尤度最大の結果を選択する．

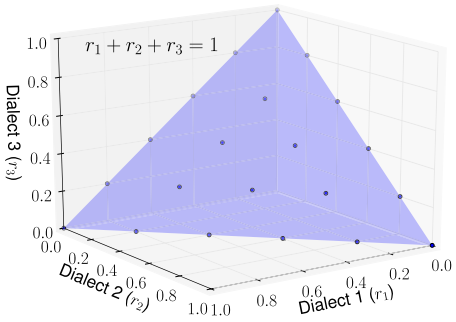


図 6 作成すべき混合方言言語モデルの混合比（3 方言，20%刻みの場合）．青の点が混合比の組合せを示す

Fig. 6 Mixing proportions of three dialects represented by a blue plane and those selected in units of 20% for building dialect-mixed language models represented by blue points.

6.2 対数尤度による方言混合比推定の妥当性

対数尤度が方言混合比推定の目的関数として利用できることを示すために，話者ごとに方言混合比を固定して全発話を認識した場合の，1 発話当たりの対数尤度（対数音響尤度・対数言語尤度の和）と平均音声認識精度の相関を求めた（図 7）．5 方言の混合比率を 20%単位で変化させた言語モデルで認識を行い，それぞれのモデルでの対数尤度を横軸に，認識精度を縦軸にプロットしている．同時に，各点は話者方言混合比で色分けしている．このグラフに対して相関係数を計算すると，北奥羽方言話者 3 名を除いて 0.8 以上，最小でも 0.739 となっており，全体として強い正の相関があることが分かり，対数尤度を目的関数として音声認識精度が向上する方言混合比を推定できるといえる．また，音声認識精度と話者方言混合比の関係には以下の傾向がみられる．

- 話者方言混合比が小さい場合には音声認識精度が低い．
- 東京方言話者に対しては，話者方言（東京方言）の比率に比例して音声認識精度が向上する．
- 共通語以外の方言話者に対しては，話者方言混合比が 100%の場合ではなく，20–60%の場合に音声認識精度が最大になる．

これらのことから，方言音声認識を行う際には，共通語言

[*5] Julius で使用するために，実際には前向き 2-gram モデルと逆向き 3-gram を学習する．

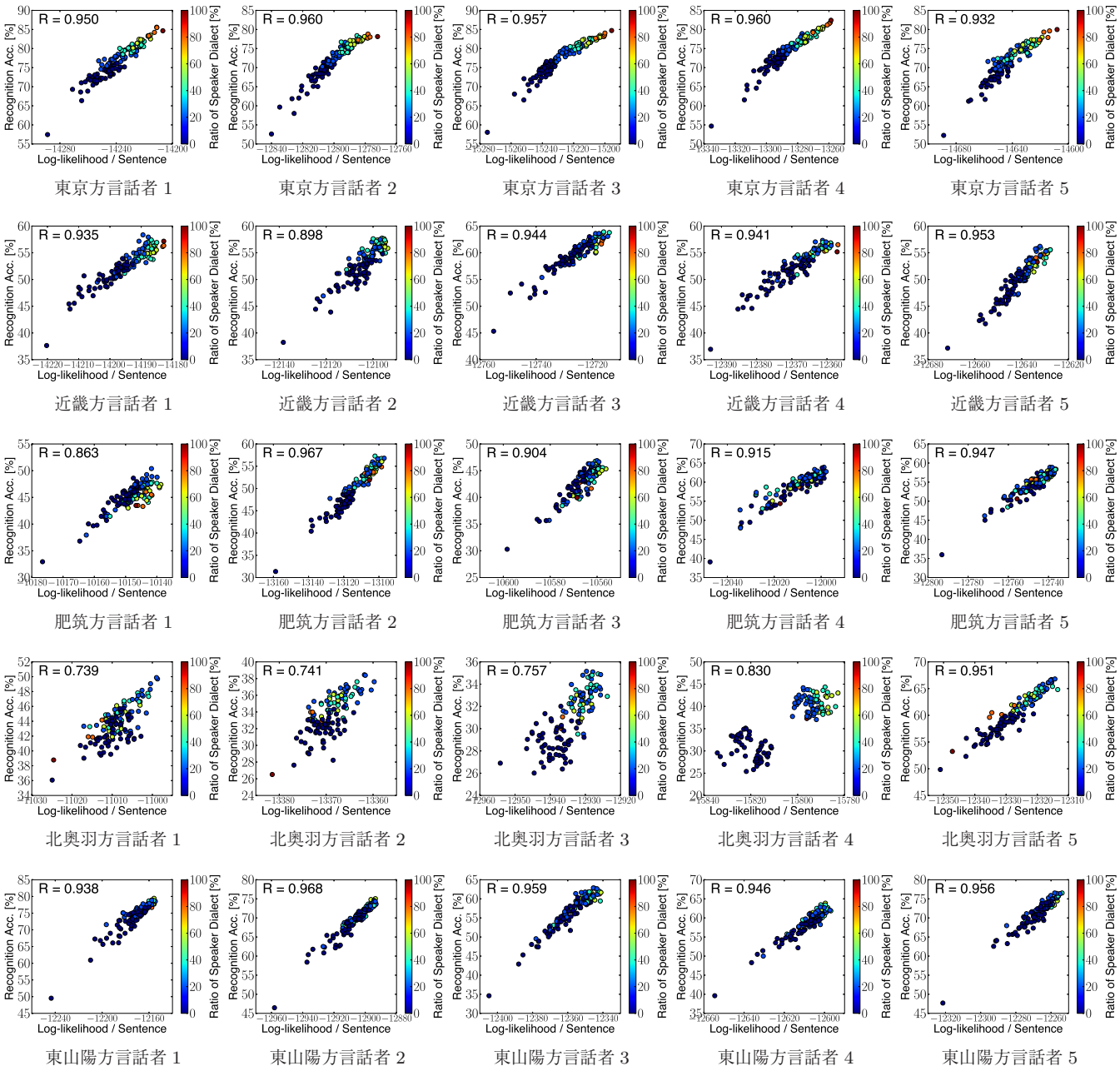


図 7 音声認識エンジンが出力した 1 文あたりの平均対数尤度（横軸）と単語認識精度（縦軸）．R は相関係数．各データ点は 1 組の方言混合比に対応しており，話者方言の比率に従って青（0%）から赤（100%）の色分けを行っている

Fig. 7 Mean log-likelihood (x-axis) and word recognition accuracy (y-axis) per sentence output by the speech recognition engine. Each data point corresponds to a combination of dialect mixing proportions and is colored blue (0% of the speaker's dialect) to red (100%).

語モデルの混合が精度向上に効果的であるといえる．単一方言言語モデルの学習時には，対訳コーパスが小規模であることから，低頻度の方言特有の語彙が方言コーパス擬似生成時に全体にわたって反映されていた．しかし，実際の方言発話においては，方言に特有の語彙ばかりでなく共通語と同様の語彙を用いることも多く，共通語言語モデルの混合によりこうした現象が正しく認識できるようになったと考えられる．

6.3 方言混合言語モデルの効果

共通語言語モデル，単一方言言語モデル，混合方言言語モデルを用いて音声認識を行い，単語認識精度を比較した．単一方言言語モデルは，話者の方言と同じものを選択した．混合方言言語モデルは，前述した混合比の中から，各話者ごとに 100 発話の対数尤度の総和が最大となるものを選び，そのときの単語認識精度を計算した．

音声認識精度の比較を**表 1** に示し，混合方言言語モデル

表 1 共通語・単一方言・方言混合言語モデルの単語認識精度 [%]（LM：言語モデル）

Table 1 Word recognition accuracies [%] for common language, single dialect and mixed dialect language models (LM: language model).

(a) 東京方言

話者	#1	#2	#3	#4	#5	平均
混合 LM	84.7	78.1	84.7	82.4	80.0	82.0
共通語 LM	84.7	78.1	84.7	82.4	80.0	82.0

(b) 近畿方言

話者	#1	#2	#3	#4	#5	平均
混合 LM	**58.4**	**57.7**	**63.9**	**57.3**	**56.2**	**58.7**
1 方言 LM	57.2	56.6	62.8	55.2	53.3	57.0
共通語 LM	51.6	49.4	61.2	50.9	50.1	52.6

(c) 肥筑方言

話者	#1	#2	#3	#4	#5	平均
混合 LM	**47.4**	**56.8**	**45.3**	**62.8**	**58.4**	**54.1**
1 方言 LM	43.5	51.9	39.9	54.4	50.6	48.1
共通語 LM	44.6	46.0	41.2	57.5	50.4	47.9

(d) 北奥羽方言

話者	#1	#2	#3	#4	#5	平均
混合 LM	**49.7**	**37.6**	**32.9**	**43.0**	**64.8**	**45.6**
1 方言 LM	38.8	26.5	28.5	37.3	53.2	36.9
共通語 LM	44.5	33.0	28.9	33.3	58.8	39.7

(e) 東山陽方言

話者	#1	#2	#3	#4	#5	平均
混合 LM	**78.6**	**73.7**	**61.7**	**61.9**	**76.5**	**70.5**
1 方言 LM	75.6	71.3	58.6	61.2	73.6	68.1
共通語 LM	66.1	65.5	51.7	54.4	66.3	60.8

で対数尤度が最大となった混合比を**表 2** に示す．東京方言以外のすべての方言および話者で，混合方言言語モデルで単語認識精度が最大となった．東京方言話者はすべて共通語言語モデル 100%の場合に対数尤度最大となり，認識精度は共通語言語モデル単独と同じ値であった．肥筑方言と北奥羽方言では，単一方言言語モデルを用いると共通語言語モデルに比べて認識精度が低下した話者がそれぞれ 3 名，4 名いたが，混合方言言語モデルではすべての話者に対して認識精度が向上した．これは，個人の方言発話が異なる複数の方言の混合であるという我々の仮定を裏付けるものである．

言語モデル自体の性能向上を確認するために，共通語言語モデルを使った場合と，提案する混合方言言語モデル（表 2 に示した対数尤度最大となる混合比）を使った場合の，各方言話者 1 名ずつ（#1）の方言書き起こし 100 文を用いて，1 音素あたりのテストセットパープレキシティを計算した．未知語モデルは音素 0-gram とし，文頭・文末無音を除く 42 種類の音素が前後の音素にかかわらず等確率で出現するとした．結果を**表 3** に示す．共通語言語モデルと混合方言言語モデルが同一となる東京方言話者を除いてパープレキシティの値が改善しており，各方言話者の発話内容を予測しやすいモデルとなっていることが分かる．

表 2 混合言語モデルによる認識時に，対数尤度最大となった混合比（混合比固定）．東京方言話者はすべて東京方言 100%と推定されたので省略

Table 2 Mixing proportions that maximized log-likelihood in recognition with dialect-mixed language models (fixed mixing proportions). All Tokyo dialect speakers were estimated to have 100% of Tokyo dialect.

(a) 近畿方言話者

話者	#1	#2	#3	#4	#5
東京方言	60	20	40	20	40
近畿方言	**20**	**40**	**40**	**20**	**20**
肥筑方言	0	20	20	20	0
北奥羽方言	0	0	0	20	0
東山陽方言	20	20	0	20	40

(b) 肥筑方言話者

話者	#1	#2	#3	#4	#5
東京方言	60	20	20	40	0
近畿方言	0	0	40	20	60
肥筑方言	**20**	**40**	**20**	**20**	**20**
北奥羽方言	0	20	0	0	0
東山陽方言	20	20	20	20	20

(c) 北奥羽方言話者

話者	#1	#2	#3	#4	#5
東京方言	80	60	0	40	80
近畿方言	0	0	0	0	0
肥筑方言	0	0	40	0	0
北奥羽方言	**20**	**20**	**20**	**40**	**20**
東山陽方言	0	20	40	20	0

(d) 東山陽方言話者

話者	#1	#2	#3	#4	#5
東京方言	20	40	40	20	80
近畿方言	0	0	20	20	0
肥筑方言	20	0	20	0	0
北奥羽方言	0	0	0	0	0
東山陽方言	**60**	**60**	**20**	**60**	**20**

表 3 方言書き起こし文の音素当たりテストパープレキシティ

Table 3 Test perplexity per phoneme of transcription of dialect utterances.

話者	東京 1	近畿 1	肥筑 1	北奥羽 1	東山陽 1	全体
共通語 LM	4.530	6.270	6.748	7.280	6.947	6.259
混合方言 LM	4.530	**5.686**	**5.672**	**6.201**	**5.603**	**5.504**

6.4 方言混合比の発話単位推定の効果

発話ごとに対数尤度最大となる方言混合比を選択し認識に用いることで，話者の全発話を同じ混合比として認識する場合に比べて音声認識精度が向上することを実験で確認した．この結果を**表 4** に示す．対数尤度を用い，混合比を発話ごとに推定して変更した場合に，近畿方言と肥筑方言

表 4 混合方言言語モデルで，混合比を全発話で固定した場合と発話ごとに変更した場合の単語認識精度 [%] の比較

Table 4 Comparison of word recognition accuracies [%] between fixed mixing proportions and changed ones by each utterance.

(a) 東京方言

話者	#1	#2	#3	#4	#5	平均
混合比変更	84.6	**79.2**	84.2	81.7	79.9	81.9
混合比固定	**84.7**	78.1	**84.7**	**82.4**	**80.0**	**82.0**

(b) 近畿方言

話者	#1	#2	#3	#4	#5	平均
混合比変更	**61.4**	**60.1**	**67.3**	**60.3**	**60.0**	**61.8**
混合比固定	58.4	57.7	63.9	57.3	56.2	58.7

(c) 肥筑方言

話者	#1	#2	#3	#4	#5	平均
混合比変更	**49.4**	**57.5**	47.2	**66.6**	**59.9**	**56.1**
混合比固定	47.4	56.8	45.3	62.8	58.4	54.1

(d) 北奥羽方言

話者	#1	#2	#3	#4	#5	平均
混合比変更	49.7	**42.7**	**37.9**	42.8	**67.9**	**48.2**
混合比固定	49.7	37.6	32.9	**43.0**	64.8	45.6

(e) 東山陽方言

話者	#1	#2	#3	#4	#5	平均
混合比変更	**81.8**	**76.1**	**65.2**	**66.0**	76.1	**73.0**
混合比固定	78.6	73.7	61.7	61.9	**76.5**	70.5

では 5 話者すべてで，東山陽方言では 4 話者，北奥羽方言では 3 話者（1 話者で変化なし）で認識精度向上がみられ，向上幅は最大 5.1 ポイントであった．また，認識精度が低下した話者でも下落幅は最大で 0.7 ポイントにとどまった．このことから，混合比を発話ごとに推定することが音声認識精度向上に有効であるといえる．この結果は，実際の発話における方言混合比は話者固有のものではなく，同じ話者でも発話ごとに変化することを示唆している．

6.5 認識結果改善例

一連の実験結果から，方言言語モデルやその混合モデルを利用することで，実際に認識精度が向上することが分かった．以下に，近畿方言話者の同じ発話を共通語言語モデル，単一（近畿）方言言語モデル，混合方言言語モデルでそれぞれ認識した例を示す．

読み上げ文 それが人間の心理ってもんやんね
（共通語訳：それが人間の心理ってものだよね）

共通語言語モデル それ/ソレ が/ガ 人間/ニンゲン の/ノ 芯/シン に/ニ 行っ/イッ て/テ も/モ いい/イー よ/ヨ ね/ネ

単一（近畿）方言言語モデル それ/ソレ が/ガ 人間/ニンゲン の/ノ 芯/シン に/ニ 行っ/イッ て/テ もん/モン だ/ヤ ね/ネ

混合方言言語モデル それ/ソレ が/ガ 人間/ニンゲン の/ノ 心理/シンリ っ/ッ て/テ もの/モン だ/ヤ ね/ネ

この例では，文の前半は共通語と同じであるが，後半は方言表現が中心となっている．共通語言語モデルと比較して，単一方言言語モデルでは共通語の「だ」に相当する「ヤ」は認識できている．混合方言言語モデルでは，文の他の部分もほぼ正しく認識できるようになっている．

話者方言以外の言語モデル混合で認識精度が向上するケースもあった．これらは以下の 3 種類に分類できる．

(1) 育った地域の隣接地域の方言を混合することで性能が向上する．図 1 はこのケースに対応する．
(2) 過去に居住してきた地方の方言に現在の居住地の方言が影響し混合していることで，現在の居住地の方言を混合すると性能が向上する．
(3) 柳田國男の方言周圏論 [34], [35] で述べられるように，京都からの距離が育った地域と同じである別の方言を混合することで性能が向上する．

それぞれのケースで認識精度（認識された共通語単語列と共通語訳との比較）が向上した例を以下に示す．

1 番目のケースについて，近畿方言話者の認識精度が隣接する東山陽方言言語モデルの混合により向上した例を示す[*6]．

読み上げ文 やっぱり何でも普段から練習してへんと出来んくなってまうね
（共通語訳：やはり何でも普段から練習していないと出来なくなってしまうね）

単一（近畿）方言言語モデル やはり/ヤッパリ どう-し-て/ナンデ も/モ 普段/フダン から/カラ 練習/レンシュー し/シ て/テ 変動/ヘンドー で/デ 菌/キン が/ガ なっ/ナッ て/テ しまう/マウ ん-だ/ネン

混合方言言語モデル やはり/ヤッパリ どう-し-て/ナンデ も/モ 普段/フダン から/カラ 練習/レンシュー し/シ て/テ ない/ヘン と/ト 出来/デキ ない/ン が/ガ なっ/ナッ て/テ しまう/マウ ん/ン です/ネン

ここで，混合方言言語モデルの混合比は近畿方言 60%，東山陽方言 40%であった．

2 番目のケースでは，多くの話者で東京方言言語モデルの混合により認識精度が向上した．北奥羽方言話者にみられた一例を示す．

読み上げ文 次に何言うのか、何をしたいのかわがるようさなる
（共通語訳：次に何を言うのか、何をしたいのかわかるようになる）

単一（北奥羽）方言言語モデル 次/ツギ に/ニ 何/ナニ 湯/ユー の/ノ 方-は/カ 、/NULL 何/ナニ を/オ

[*6] 「何でも」が「どうしても」と認識されているのは，理由を尋ねる「どうして」を近畿方言で「なんで」ということから発生した誤りと考えられる．

し/シ た/タ 犬/イヌ を/オ 買う/カウ わかる/ワガル よう/ヨー に/サ なる/ナル

混合方言言語モデル 次/ツギ に/ニ 何/ナニ いう-もの/ユーノ か/カ 、/**NULL 何/ナニ を/オ** し/シ たい/タイ の/ノ か/カ 分かる/ワカル よう/ヨー に/サ なる/ナル

混合方言言語モデルの混合比は北奥羽 60%，共通語 40%であった．

3 番目のケースでは，北奥羽方言の一部の話者で，肥筑方言言語モデルの混合により音声認識精度が向上した．

読み上げ文 その人は普通の人じゃねーはんで、注意すると何されるかわがんねじゃ

（共通語訳：その人は普通の人ではないから、注意すると何されるかわからないよ）

北奥羽・共通語混合方言言語モデル その/ソノ 人/ヒト は/ワ 普通/フツ の/ノ 人/シト **し-て-い/シャ** ない-から/ネハンデ 、/NULL 注意/チューイ する/スル と/ド 何/ナニ さ/サ れる/レル か/カ わから/ワガ ない-よ/ネジャ

北奥羽・共通語・肥筑混合方言言語モデル その/ソノ 人/ヒト は/ワ 普通/フツー の/ノ 人/ヒト **で-は/ジャ** ない-から/ネハンデ 、/NULL 注意/チューイ する/スル と/ト 何/ナニ さ/サ れる/レル か/カ わから/ワガ ない-よ/ネジャ

ここで，北奥羽・共通語混合モデルは 2 番目のケースの混合方言言語モデルと同じである．北奥羽・共通語・肥筑混合モデルの混合比は北奥羽 20%，共通語 40%，肥筑 40%であった．

7. おわりに

本論文では，小規模な共通語–方言対訳コーパスを利用して，方言音声認識のための言語モデルを構築する手法，および方言混合比を発話から推定し，これらの方言言語モデルを混合して音声認識を行う手法について述べた．言語モデル学習コーパスは，対訳コーパスから学習した方言変換ルールに基づき，大規模な共通語言語コーパスから擬似生成することで生成した．混合方言言語モデルは，各方言言語モデルの n-gram 確率および単語発音辞書の発音確率を重み付き平均することで生成した．実験では，音声認識エンジンが出力する認識結果の対数尤度が音声認識精度と強い相関を持ち，発話からの方言混合比自動推定に利用できることを示した．また，この結果を用いて発話ごとの方言言語モデルの混合を行うことで，音声認識精度の向上を確認した．

今後の課題としては次の 2 点があげられる．1 点目は，方言による音声認識精度の差を縮小することである．本手法で相対的に音声認識精度は向上したが，肥筑方言，北奥羽方言では，絶対的に十分な音声認識精度が得られていない．本手法では，音素セットを共通語と同一としたので，音韻体系が共通語と比べて大きく異なる方言について十分には対処できない．本手法と合わせて方言ごとの音韻体系の違いを扱うモデルを導入することで，より精度が向上すると考えられる．2 点目は，今回提案した手法を個人の属性推定に活用し，音声認識精度の向上に加えてデータマイニングに応用することである．企業の戦略立案においてデータマイニングの重要性が取り上げられており [36]，顧客の購入・情報閲覧履歴からの関連商品推薦 [37]，電話応対ログ解析の製品開発や顧客満足度向上への活用 [38] などに利用されている．こうしたマイニングを行うにあたり，言葉遣いや話し方からそれらに影響を与える居住地域，職業やコミュニティといった属性が抽出できれば，それらを用いた新たな知識の発見につながると考えられる．

謝辞 本研究の一部は，科研費 (S)（No.24220006）の支援を受けた．

参考文献

[1] 秋田祐哉，三村正人，河原達也：会議録作成支援のための国会審議の音声認識システム，電子情報通信学会論文誌，Vol.J93-D, No.9, pp.1736–1744 (2010).
[2] 今井 亨：リアルタイム字幕放送のための音声認識，電子情報通信学会技術報告，SP（WIT），Vol.109, No.259, pp.19–24 (2009).
[3] 榎澤幸広：方言話者と法廷，筑波学院大学紀要，Vol.4, pp.83–92 (2009).
[4] 小林 隆，篠崎晃一（編）：ガイドブック方言研究，ひつじ書房 (2003).
[5] Cruse, D.A.: *Lexical Semantics*, Cambridge University Press (1986).
[6] 国立国語研究所：共通語，方言（日本の言語学 第 6 巻），大修館書店，pp.669–685 (1978).
[7] 真田信治：標準語はいかに成立したか — 近代日本語の発展の歴史，創拓社 (1991).
[8] Brinton, L.J. and Fee, M.: *English in North America*, The Cambridge history of the English language, Vol.6, The Press Syndicate of the University of Cambridge (2001).
[9] Thomas, E.R.: Rural Southern white accents, *The Americas and the Caribbean*, Schneider, E.W. (Ed.), Varieties of English, Vol.2, pp.87–114, Mouton de Gruyter (2008).
[10] Ramon, D.: We Are One People Separated By A Common Language, *Viagra, Prozac, and Leeches*, pp.203–206, iUniverse (2006).
[11] Woods, H.B.: A Socio-dialectology Survey of the English Spoken in Ottawa: A Study of Sociological and Stylistic Variation in Canadian English, PhD Thesis, The University of British Columbia (1979).
[12] Lyu, D., Lyu, R., Chiang, Y. and Hsu, C.: Speech Recognition on Code-Switching Among the Chinese Dialects, *Proc. ICASSP 2006*, Vol.1, pp.1105–1108 (2006).
[13] Ching, P.C., Lee, T. and Zee, E.: From phonology and acoustic properties to automatic recognition of Cantonese, *Proc. Speech, Image Processing and Neural Networks 1994*, pp.127–132 (1994).
[14] Miller, D.R. and Trischitta, J.: Statistical dialect classification based on mean phonetic features, *Proc. ICSLP*

1996, Vol.4, pp.2025–2027 (1996).
[15] Chitturi, R. and Hansen, J.H.L.: Dialect Classification for online podcasts fusing Acoustic and Language based Structural and Semantic Information, *Proc. ACL HLT 2008, Short Papers*, pp.21–24 (2008).
[16] 国立国語研究所（編）：全国方言談話データベース 日本のふるさとことば集成（全 20 巻），国書刊行会 (2001–2008).
[17] Allauzen, C., Riley, M., Schalkwyk, J., Skut, W. and Mohri, M.: OpenFst: A General and Efficient Weighted Finite-State Transducer Library, *Proc. CIAA 2007, LNCS*, Vol.4783, pp.11–23 (2007).
[18] Misu, T. and Kawahara, T.: A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts, *Proc. ICSLP 2006*, pp.9–12 (2006).
[19] Maekawa, K.: Balanced Corpus of Contemporary Written Japanese, *Proc. 6th Workshop on Asian Language Resources*, pp.101–102 (2008).
[20] Heeringa, W.J.: Measuring Dialect Pronunciation Differences using Levenshtein Distance, PhD Thesis, University of Groningen (2004).
[21] Neubig, G., Nakata, Y. and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *Proc. ACL HLT 2011*, pp.529–533 (2011).
[22] Chen, S.F.: Conditional and Joint Models for Grapheme-to-Phoneme Conversion, *Proc. EUROSPEECH 2003*, pp.2033–2036 (2003).
[23] 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，情報処理学会論文誌，Vol.43, No.9, pp.2884–2893 (2002).
[24] Brown, P.F., Desouza, P.V., Mercer, R.L., Pietra, V.J.D. and Lai, J.C.: Class-Based *n*-gram Models of Natural Language, *Computational Linguistics*, Vol.18, No.4, pp.467–479 (1992).
[25] 中川竜太，岩野公司，古井貞熙：超並列計算機を用いた入力音声の変動に頑健な音声対話システムの検討，電子情報通信学会技術研究報告，No.105, pp.1–6 (2005).
[26] 安田宜仁，堂坂浩二，相川清明：2 つの認識文法を用いた主導権混合型対話制御，情報処理学会研究報告，Vol.SLP40, No.10, pp.127–132 (2002).
[27] Lee, A., Shikano, K. and Kawahara, T.: Real-Time Word Confidence Scoring using Local Posterior Probabilities on Tree Trellis Search, *Proc. ICASSP 2004*, pp.793–796 (2004).
[28] 加藤正信：方言区画論，方言（岩波講座 日本語 11），岩波書店，pp.41–82 (1977).
[29] Lee, A., Kawahara, T. and Shikano, K.: Julius — An Open Source Real-Time Large Vocabulary Recognition Engine, *Proc. EUROSPEECH 2001*, pp.1691–1694 (2001).
[30] Kneser, R. and Ney, H.: Improved backing-off for M-gram language modeling, *Proc. ICASSP 1995*, Vol.1, pp.181–184 (1995).
[31] 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄（編）：音声認識システム，オーム社 (2001).
[32] Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp.7–12 (2003).
[33] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research, *Acoustical Society of Japan* (*English Edition*), Vol.20, pp.199–206 (1999).
[34] 柳田國男：蝸牛考，岩波文庫 (1980). 初版：1930 年発行（刀江書院）.
[35] 松本　修：全国アホ・バカ分布考 — はるかなる言葉の旅路，太田出版 (1993).
[36] 林田英雄，脇森浩志：テキストマイニング技術とその応用，*UNISYS TECHNOLOGY REVIEW*, No.84, pp.29–44 (2005).
[37] 土方嘉徳：嗜好抽出と情報推薦技術，情報処理，Vol.48, No.9, pp.957–965 (2007).
[38] 佐野紳也，白井康之，東野恒之，二瓶　正：音声解析にもとづくテキスト分析の可能性と今後の展望，三菱総合研究所所報，Vol.2010, No.52, pp.74–88, 三菱総合研究所 (2010).

平山 直樹　（正会員）

2012 年京都大学工学部情報学科卒業．2014 年同大大学院情報学研究科知能情報学専攻修士課程修了．在学中は方言音声認識システムの研究に従事．現在，東芝ソリューション株式会社勤務．

吉野 幸一郎　（学生会員）

2009 年慶應義塾大学環境情報学部卒業．2011 年京都大学大学院情報学研究科修士課程修了．同年同大学院博士後期課程に進学．2013 年より日本学術振興会特別研究員（DC2）．音声言語処理および自然言語処理，特に音声対話システムに関する研究に従事．言語処理学会会員．

糸山 克寿　（正会員）

2006 年京都大学工学部情報学科卒業．2008 年同大学大学院情報学研究科知能情報学専攻修士課程修了．2011 年同大学大学院情報学研究科知能情報学専攻博士後期課程修了．同年より同大学大学院情報学研究科知能情報学専攻助教．京都大学博士（情報学）．音楽情報処理，音楽鑑賞インタフェース等の研究に従事．

森 信介 （正会員）

1998 年京都大学大学院工学研究科電子通信工学専攻博士後期課程修了．同年日本アイ・ビー・エム（株）入社．2007 年より京都大学学術情報メディアセンター准教授．京都大学博士（工学）．音声言語処理および自然言語処理に関する研究に従事．1997 年情報処理学会山下記念研究賞受賞．2010 年，2013 年情報処理学会論文賞受賞．2010 年第 58 回電気科学技術奨励賞．言語処理学会，ACL 各会員．

奥乃 博 （正会員）

1972 年東京大学教養学部基礎科学科卒業．1996 年東京大学博士（工学）．1972 年日本電信電話公社入社．NTT 基礎研究所を 1998 年退職．科学技術振興事業団 ERATO，東京理科大学理工学部情報科学科を経て，2001 年より，京都大学大学院情報学研究科知能情報学専攻教授．2014 年退職，名誉教授．現在，早稲田大学大学院創造理工学研究科教授（任期付）として実体情報学博士プログラムに所属．プログラミング環境，人工知能研究を経て，音環境理解，音楽情報処理，ロボット聴覚の研究に従事．日本ソフトウエア科学会，人工知能学会，本学会各元理事，IEEE Fellow，人工知能学会フェロー，1990 年度人工知能学会論文賞，IROS-2010 NTF Award for Entertainment Robots and Systems，2013 年度科学技術分野の文部科学大臣表彰科学技術賞（研究部門）等受賞．