4S-6

混合方言言語モデルと混合比推定による方言音声認識システム

平山 直樹 † 吉野 幸一郎 † 糸山 克寿 † 森 信介 †‡ 奥乃 博 †

† 京都大学 大学院情報学研究科 知能情報学専攻 ‡ 京都大学 学術情報メディアセンター

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，国会議事録作成 [1] など実世界応用が盛んになっている．しかし，話者ごとの言葉遣いの差異，特に方言の混合使用を考慮した音声認識システムは少ない．不特定多数の話者を想定する音声認識システムは，単一のシステムで多様な方言，しかも混合使用に対しても機能することが不可欠である．本稿では，日本語方言の語彙変化と発音 (使用される音素) 変化を対象とする，音声認識と方言混合比推定の同時実行手法を開発する．

方言音声認識には以下の 3 点の課題がある．

1) 方言言語資源の不足 方言は元来話し言葉であり，方言言語資源の不足が言語モデルの構築を困難にする．

2) 話者方言推定 地域間の人々の移動により，方言は互いに影響し合い,「純粋な」方言は仮定できない．話者方言は各方言が混合して使われるという前提の下で，話者方言に適した認識モデル選択を行う．

3) 実時間音声認識 音声認識を用いたアプリケーションでは，実時間応答のために実時間音声認識が必要である．

上記 1) については，共通語-方言対訳コーパスによる方言コーパス模倣を行って方言言語モデルを学習し，それらを混合する手法を開発した [2]．本稿では，上記の手法を 2) 3) に拡張する手法を中心に述べる．

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．そこで使用される単語発音辞書 [3] や方言間対応ルール [4] は，人手で構築されていた．また，音響特徴による方言同定 [5, 6, 7] では，方言特徴の大部分を占める [8] 言語特徴を利用できないという問題があった．そこで，我々は単一方言音声認識において，共通語-方言対訳コーパスを用いた統計的手法により上記の自動化を行った [2]．本稿では，既開発手法を入力方言未知の場合に拡張し，さらに実時間動作させる手法を報告する．

3. 提案手法

本稿における方言音声認識 (図 1) では，方言未知の発話音声を入力に取り，共通語単語列を出力する．このような出力は，音声認識を用いたアプリケーションへの応用のために，認識結果の方言による差異を吸収する目的がある．後の音声認識精度評価も共通語を基準に行う．

本章では，以下の流れで本手法について述べる．まず，単一方言言語モデルのアイデアについて述べる．次に，各単一方言言語モデルの混合による話者方言推定について述べる．最後に，方言言語モデル自体は混合せずに各方言言語モデルによる認識結果を統合することで，実時間音声認識に対応させる手法について述べる．

Dialect Speech Recognition System by Mixed Dialect Language Models and Estimation of the Mixing Ratio: *Naoki Hirayama, Koichiro Yoshino, Katsutoshi Itoyama, Shinsuke Mori, and Hiroshi G. Okuno (Kyoto Univ.)*

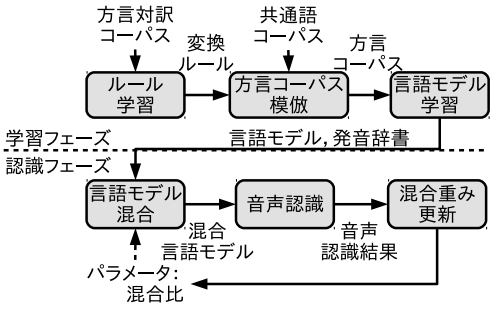

図 1: 方言音声認識の流れ．

3.1 単一方言言語モデル

単一方言言語モデルは，大規模共通語言語コーパスの各単語に方言発音を付与して得られる大規模方言言語コーパスから学習する [2]．方言発音付与には，小規模な共通語-方言対訳コーパス [9] から学習した重み付き有限状態トランスデューサ (WFST) を利用する．共通語単語ごとに方言発音を集計して，単語発音辞書に各発音と出現確率を記述する．これは共通語単語をクラスとするクラス n-gram モデルに相当する．

3.2 混合方言言語モデル

各方言 d の混合比 r_d ($r_d \geq 0, \sum_d r_d = 1$) に対応する混合方言言語モデルを単一方言言語モデルから構築する．

単語 (クラス) n-gram モデルは，各言語モデルの単語 n-gram 確率の重み付き平均で混合する．混合方言言語モデルの単語 n-gram 確率 $P_{\mathrm{mix}}(w_i|w_{i-n+1}^{i-1})$ を，方言 d の言語モデルにおける単語 n-gram 確率 $P_d(w_i|w_{i-n+1}^{i-1})$ と混合比 r_d を用いて

$$P_{\mathrm{mix}}(w_i|w_{i-n+1}^{i-1}) = \sum_d r_d P_d(w_i|w_{i-n+1}^{i-1}) \quad (1)$$

で計算する．実際にはバックオフ確率も考慮する [10]．

次に，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) の計算を行う．共通語単語 w に対応する方言発音 y の出現頻度を，単一方言言語モデルにおける出現頻度の r_d による重み付き和とする．共通語単語 w，w に対する方言発音 y の出現回数をそれぞれ $\#_d(w), \#_d(y|w)$ とすると，発音確率 $P_{\mathrm{class,mix}}(y|w)$ は

$$P_{\mathrm{class,mix}}(y|w) = \frac{\#(y|w)}{\#(w)} = \frac{\sum_d r_d \#_d(y|w)}{\sum_d r_d \#_d(w)} \quad (2)$$

となる．共通語言語コーパスを同一とすると $\#_d(w)$ は d によらず一定の値 $\#(w)$ となるので，

$$P_{\mathrm{class,mix}}(y|w) = \frac{\sum_d r_d \#_d(y|w)}{\#(w)} = \sum_d r_d P_{\mathrm{class},d}(y|w) \quad (3)$$

となる ($\sum_d r_d = 1$)．$P_{\mathrm{class},d}(y|w)$ は方言 d の言語モデルにおける発音確率であり，混合後の言語モデルの発音確率は，これらの重み付き平均となる．

混合比 r_d は，実際の方言発話が高い精度で認識できるように，適切に指定する必要がある．本稿では，各方

表 1: 単語認識精度 [%] と，単語頻度と単語信頼度の最適な重み付け α の値 (実時間のみ)．#1–5 は話者番号を示す．

(a) 東京方言話者

言語モデル	#1	#2	#3	#4	#5
実時間	82.7	77.2	82.8	80.9	78.3
α	0.7	0.7	0.6	0.6	0.7
発話単位尤度最大化	84.6	79.2	84.2	81.7	79.9
話者単位尤度最大化	84.7	78.1	84.7	82.4	80.0
等比混合	77.8	74.1	80.0	75.5	72.5
共通語	84.7	78.1	84.7	82.4	80.0

(b) 近畿方言話者

言語モデル	#1	#2	#3	#4	#5
実時間	57.4	56.9	64.7	56.9	55.8
α	0.8	0.8	0.5	0.9	1.0
発話単位尤度最大化	61.4	60.1	67.3	60.3	60.0
話者単位尤度最大化	57.2	55.9	63.1	56.5	55.6
等比混合	55.8	57.0	62.1	57.3	52.3
近畿方言	57.2	56.6	62.8	55.2	53.3
共通語	51.6	49.4	61.2	50.9	50.1

(c) 肥筑方言話者

言語モデル	#1	#2	#3	#4	#5
実時間	48.7	52.9	45.5	63.2	57.5
α	0.5	0.7	0.7	0.6	0.9
発話単位尤度最大化	49.4	57.5	47.2	66.6	59.9
話者単位尤度最大化	47.4	56.6	45.3	62.8	58.4
等比混合	47.3	54.3	44.5	61.5	57.3
肥筑方言	43.5	51.9	39.9	54.4	50.6
共通語	44.6	46.0	41.2	57.5	50.4

(d) 北奥羽方言話者

言語モデル	#1	#2	#3	#4	#5
実時間	46.1	40.3	36.3	41.8	62.4
α	0.8	0.6	0.6	0.6	0.8
発話単位尤度最大化	49.7	42.7	37.9	42.8	67.9
話者単位尤度最大化	49.7	37.6	32.9	43.0	64.8
等比混合	45.3	37.3	35.1	39.2	65.2
北奥羽方言	38.8	26.5	28.5	37.2	53.2
共通語	44.5	33.0	28.9	33.3	58.8

(e) 東山陽方言話者

言語モデル	#1	#2	#3	#4	#5
実時間	77.4	73.4	61.4	61.6	73.6
α	0.6	0.9	1.0	1.0	0.9
発話単位尤度最大化	81.8	76.2	65.2	66.0	76.1
話者単位尤度最大化	78.6	73.7	61.7	61.9	76.5
等比混合	76.9	72.0	61.8	60.9	74.1
東山陽方言	75.6	71.3	58.8	61.2	73.6
共通語	66.1	65.5	51.7	54.4	66.3

表 2: 実験に用いたデータセット．語数は共通語ベースでカウント．

用途	データセット	規模
言語モデル学習	Yahoo! 知恵袋[2]	300 万文，6,030 万語
音響モデル学習	CSJ [12, 13]	500 講演，70.2 時間
	JNAS [14]	308 話者，23.3 時間
共通語-方言対訳コーパス	方言談話 DB [9]	各方言 0.9–1.3 万語

言の混合比を 20% 単位とし，和が 100% となるすべての組合せに対する混合方言言語モデルで音声認識を行って，尤度が最大となる認識結果で認識精度を評価する．

3.3 実時間方言音声認識

実際の対話システムでは，すべての混合比の組合せを比較するのは計算時間の観点から現実的ではない[1]．そこで，各単一方言言語モデルによる音声認識結果の統合により，計算時間を抑えると同時に認識誤りの削減を図る．実験では 5 方言を対象とするが，音声認識器を実時間 (音声の長さより短い時間) で行えるよう設定すれば，並列処理により処理全体も実時間に収まると考えられる．

本稿では，統合手法として ROVER (Recognizer output voting error reduction) 法 [11] を用いる．すなわち，複数の音声認識結果のマッチングにより対応単語組の列を生成し，それぞれから 1 単語を選択する．各対応単語組中で，単語頻度と単語信頼度の平均値の重み付け (比率 $\alpha : 1-\alpha$) で票数を決定し，最も票数の多かった単語を出力する．なお評価実験では，話者ごとに認識精度最大となる α を求め，認識精度とともに示す．

4. 評価実験

Julius[3]による方言音声認識を行い，共通語ベースの単語認識精度により本手法の有効性を評価した．5 方言の話者各 5 名 (出身都道府県は，東京方言: 東京・埼玉，近畿方言: 大阪・兵庫，肥筑方言: 福岡・熊本，北奥羽方言: 青森・山形，東山陽方言: 広島・岡山) に共通語 100 文 (全話者共通) を提示し，東京方言話者はそのまま，方言話者は各自の方言に翻訳した文を読み上げた．実験に用いたデータセットを表 2 にまとめる．

表 1 に各手法における認識精度の結果を示す．尤度最大化の手法では，いずれの話者についても単一方言や共通語言語モデルより認識精度が向上した．発話単位の混合比推定によりさらに高い認識精度が得られ，共通語言語モデルからの上昇幅は最大 15.7 ポイント (東山陽方言話者 1) であった．また，実時間音声認識のための統合手法では，25 話者中 19 話者で各方言を等しく混合した (等比混合) 場合を上回り，東京方言話者を除く 20 話者中 19 話者で話者方言言語モデルをそのまま利用した場合を上回る認識精度が得られた．

5. おわりに

本稿では，複数方言音声認識システムのための言語モデルの構築と，話者方言未知の状況下での混合言語モデルの選択，そして実時間動作のための認識結果統合手法を提案し，実験によりその有効性を確認した．今後の課題として，方言の影響を受けやすい語や文について考察し，認識結果統合に利用することが考えられる．本研究は，科研費基盤研究 (S) No. 24220006 の援助を受けた．

参考文献

[1] 秋田祐哉ほか：会議録作成支援のための国会審議の音声認識システム，信学論，Vol. J93-D, No. 9, 1736–1744 (2010).

[2] Hirayama, N. et al.: Automatic Estimation of Dialect Mixing Ratio for Dialect Speech Recognition, *INTERSPEECH 2013*, 1492–1496 (2013).

[3] Lyu, D. et al.: Speech recognition on code-switching among the Chinese Dialects, *ICASSP 2006*, Vol. 1, 1105–1108 (2006).

[4] Zhang, X.: Dialect MT: a case study between Cantonese and Mandarin, *ACL and COLING 1998*, Vol. 2, 1460–1464 (1998).

[5] Ching, P. et al.: From phonology and acoustic properties to automatic recognition of Cantonese, *Speech, Image Processing and Neural Networks, 1994*, 127–132 (1994).

[6] Miller, D. and Trischitta, J.: Statistical dialect classification based on mean phonetic features, *ICSLP 1996*, Vol. 4, 2025–2027 (1996).

[7] Chitturi, R. and Hansen, J. H. L.: Dialect classification for online podcasts fusing acoustic and language based structural and semantic information, *ACL 2008: Human Language Technologies, Short Papers*, 21–24 (2008).

[8] Wolfram, W.: *Ethnolinguistic Diversity and Literacy Education*, Routledge (2009).

[9] 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).

[10] 長友健太郎ほか：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，情処学論，Vol. 43, No. 9, 2884–2893 (2002).

[11] Fiscus, J. G.: A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER), *ASRU 1997*, 347–354 (1997).

[12] Maekawa, K.: Corpus of Spontaneous Japanese: Its design and evaluation, *SSPR* (2003).

[13] 篠崎隆宏ほか：話し言葉コーパスを用いた音声認識の検討，日本音響学会 2001 年春季研究発表会講演論文集，31–32 (2001).

[14] Itou, K. et al.: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, *Acoustical Society of Japan (English Edition)*, Vol. 20, 199–206 (1999).

[1] 5 方言を 20% 単位で混合する場合の数は 126 通りである．

[2] ヤフー株式会社・国立情報学研究所 (NII) が提供．

[3] http://julius.sourceforge.jp/