様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　25 年　5 月 23 日現在

機関番号：１１５０１
研究種目：基盤研究(C)
研究期間：2010　～ 2012
課題番号：22500144
研究課題名（和文）　高精度な話し言葉認識技術の開発

研究課題名（英文）　Development of high-accuracy system for recognizing spontaneous speech

研究代表者
小坂　哲夫（KOSAKA TETSUO）
山形大学・大学院理工学研究科・教授
研究者番号：50359569

研究成果の概要（和文）：
本研究では，書き言葉の音声認識と比較し，認識が困難と考えられる話し言葉音声認識の性能向上を目指す．本研究では研究課題として，(1) 音響・言語モデルの高性能化，(2) システム統合，(3) 話者インデキシング，の 3 点について重点的に取り組む．音響モデルの高精度化に関して離散分布モデルの識別学習，話者クラスモデル，quinphone，残響クラスモデルなどについて検討を行った．システム統合については，連続と離散分布モデルの統合．多種の quinphone の統合，残響クラスモデルの統合について検討し有効性を示した．言語モデルに関してはクロス適応やクロスバリデーション適応の有効性を示した．さらに話者適応時に必要となる話者ベクトルを用いた話者インデキシングの性能向上について検討した．

研究成果の概要（英文）：
In our research, we aimed to improve the system performance for recognizing spontaneous speech, which was considered to be more difficult than recognizing read speech. We focused on three technical issues: (1) acoustic and language models, (2) system combination techniques, and (3) speaker indexing. For improving the performance of acoustic models, we investigated a discrete-mixture hidden Markov model based on discriminative training, a speaker-class model, a quinphone, and a reverberation-class model. Some system combination techniques were investigated, such as the combination of continuous and discrete models, the combination of various quinphones, and the combination of reverberation-class models. For the issues of language models, we proposed the cross adaptation and cross-validation adaptation techniques. In addition, we improved the performance of speaker indexing techniques based on speaker vectors required during the execution of speaker adaptation.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,400,000	420,000	1,820,000
2011 年度	1,100,000	330,000	1,430,000
2012 年度	500,000	150,000	650,000
年度			
年度			
総　計	3,000,000	900,000	3,900,000

研究分野：音声情報処理
科研費の分科・細目：情報学・知覚情報処理・知能ロボティクス
キーワード：音声認識，話し言葉，音響モデル，言語モデル，話者適応

１．研究開始当初の背景

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，WER5%以下の高い認識性能が既に達成されている．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．音声は本来人と人とのコミュニケーションの中で使用されるのが基本であり，コンピュータあるいはネットワーク上の音声コンテンツを利用するためには，話し言葉音声認識の実用化が不可欠である．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って，研究課題として挑戦的で学術的意義が大きい．

海外の状況としては，DARPA EARS プログラムの参加機関では 2100～2300 時間の音声データと 9～14 億語の言語テキスト(書き起こし 2500 ～2800 万語を含む) が利用可能になった．IBM(米)，BBN 社(米)，ケンブリッジ大(英)，LIMSI(仏) 等ではこれら大量の音声データ・言語テキストを用いて，音素文脈の拡大，ガウス分布形の改善，識別学習のモデル推定法改善等による音響モデルの精密化，単語連鎖の拡大，言語モデルの精密化，クロス適応やシステム統合等による性能改善策の研究が進展している．国内では 2004 年 6 月に公開された「日本語話し言葉コーパス：CSJ」を利用した研究が盛んに行われている．以上のように，話し言葉音声認識の現状の技術レベルは，大量の音声コーパス，テキストコーパスの使用により，音響モデルおよび言語モデルの精度が向上した結果，講演音声など発話が明確な音声に対しては単語誤り率 20% を切るまでに認識性能が向上した．しかし，実用を考えればまだ認識性能は不十分であり，より高い性能を持つ認識システムの開発が望まれる．

２．研究の目的

本研究では，書き言葉の音声認識と比較し，認識が困難と考えられる話し言葉音声認識の性能向上を目指す．LVCSR システムに現在最も求められていることは，認識エンジン自体の高性能化である．そこで，本研究では，話し言葉音声認識の基本性能の向上に注力する．性能向上のためには音響モデルと言語モデルの性能向上が特に重要となる．このための各種手法について検討を行う．また適応手法も性能向上に有効である．本研究では音響モデルや言語モデルの適応，さらには両者を組み合わせた適応を行う．特に発話内容が未知の状態で行う教師なし適応が実用上有用のため，本研究では教師なし適応を扱う．

講演音声は話者が 1 名であるが，会議音声や会話音声では複数の話者を扱うため，話者適応を行う場合，話者インデキシングを行うなどの工夫が必要となる．この手法の開発も行う．さらに実際の使用を考えると，静粛な環境のみならず雑音環境下での性能向上が必要である．このため雑音に頑健な手法の開発を目指す．

話し言葉音声認識の性能向上が実現できれば，コンピュータとの対話，講演や会議の書き起こし，要約，音声検索など様々な用途に利用できると期待される．

３．研究の方法

本研究では，上記目的に示したように認識が困難と考えられる話し言葉音声認識の基本性能の向上を目指す．このための方策として本研究では，(1) 音響・言語モデルの高性能化，(2) システム統合，(3) 話者インデキシング，の 3 点について重点的に取り組む．

(1)のうち音響モデルについては，(2)のシステム統合とも関連するが，認識誤り傾向の異なる高精度な音響モデルの構築を検討する．言語モデルについては，発話の自然性に応じたモデル適応について検討する．

(2)では単語グラフ統合による性能改善を図る．我々はこれまで，一般的な最尤推定による連続分布 HMM のほか，離散混合分布 HMM，識別学習による HMM など種々の音響モデルの検討を行ってきている．各モデルは誤り傾向が異なるため，統合して利用することにより，相補的な情報を得ることができる．以上によりさらなる性能向上が期待できる．

上記(3)は会議音声など複数の話者が発話を行う場合の話者適応に必要な技術であり，各発話の話者を同定することにより話者適応が可能となる．これまでの話者認識の検討で，発話区間が極めて短い場合も高精度に話者を認識する手法を開発しており，この方法を応用する．

４．研究成果

(1) 音響モデルの精度向上

近年の音声コーパスの整備により，数百時間規模の大量の音声データを音響モデルの学習に使用することが可能となった．これに伴い，従来学習データの不足により充分な性能が発揮できなかった手法が利用可能となりつつある．論文[6]では，従来困難と考えられていた離散分布 HMM による話し言葉音声認識の検討を行い，混合分布やサブベクトル量子化の利用により高精度の認識結果が得られることを示した．また近年最尤推定に代わり識別学習が注目されている．論文[3]では

識別学習の一種である最大相互情報量(MMI)に基づくパラメータ推定手法を離散分布 HMM に応用し，更なる性能向上が得られることを示した．次に話者クラスモデルによる性能向上について述べる．これまで，認識対象話者の声質に類似した学習話者を選択し作成した話者クラスモデルの有効性が示されている．しかし，学習話者を何人選べばよいかについては実験的に定める場合が多く適切な設定法が十分には検討されていなかった．これに対し論文[7]では，学習話者の異なる複数の話者クラスモデルを作成し，発話毎，尤度基準でモデルを選択することにより，性能が向上することを示した．

(2) 単語グラフ統合による精度向上
従来，複数認識システムの出力を統合することによる認識性能の向上が種々検討されている．代表的な手法として ROVER やコンフュージョンネットワーク統合などが挙げられる．従来は異なる研究機関の認識システムを用いその出力を統合することが主に行われてきた．それに対し本研究では音響モデルに焦点を当て，性質の異なる音響モデルを複数用意し，それを使った複数の認識結果を統合することにより認識性能の向上を図った．出力統合に当たっては単語グラフ統合法を利用した．この手法において重要なのは，どのような音響モデルを用いたかという点である．以下本研究で行った各種の単語グラフ統合を音響モデルごとにまとめて記述する．

① 話者クラス音響モデル
論文[7]では複数の音響モデルから，尤度基準で最適な話者クラスモデル 1 つを選択することにより性能向上を図った．これに対し論文[10]では，単語グラフ統合により複数の話者クラスモデルの出力を統合することにより性能向上を図った．この結果更なる性能向上が得られることが分かった．

② 連続分布 HMM と離散分布 HMM
論文[3]などで提案した離散分布 HMM と従来一般的に用いられる連続分布 HMM では尤度の出力傾向が異なる．このため両者の出力を統合することにより，相補的な効果が得られる可能性がある．論文[11]では ML 推定や MMI 推定した離散分布 HMM と連続分布 HMM の出力を単語グラフ統合法で統合することにより，性能向上が得られることが分かった．また統合するモデルの性質が異なる方がより高い性能が得られることを示した．さらに論文[9]では，この手法が雑音の影響下において，極めて有効であることを示した．

③ quinphone モデル
音響モデルの単位として，音素の前後関係を考慮した音素環境依存モデルで高い性能が得られることが分かっている．従来は当該音素の前後 1 音素までを考慮した triphone が一般的であった．これに対し学会講演[13]および[16]では前後 2 音素までを考慮した quinphone の検討を行った．この結果状態数の異なる quinphone の出力を統合することにより高い認識性能が得られることが分かった．

④ 残響クラスモデル
学会[6]では，単語グラフ統合の対象として残響環境別のモデルを使用した場合，様々な残響環境下での音声認識に有効であることを示した．

(3) 言語モデル適応
言語モデルの性能向上を図るため，言語モデルの教師なし適応について検討を行った．少量の適応データでの効果を得るため，品詞を用いた単語クラスを利用して適応を行った．繰り返し教師なし適応では，誤認識の情報を繰り返し使用することによる弊害が問題となるが，主に音響モデルで用いられている，クロス適応やクロスバリデーション適応を利用することにより教師なし適応の性能向上が得られることが分かった．また教師なしモデル適応について，音響モデルに対する適応と言語モデルに対する適応の 2 つが考えられるが，両者を併用する場合の効果的な方法について検討した．単純に 2 つの適応を繰り返す場合と比較し，音響および言語適応の両方について，クロス適応やクロスバリデーション適応を利用することにより性能向上が図れることが分った(論文[4]，学会発表[11])．繰り返し適応において 1 回の繰り返しごと音響モデルと言語モデルが更新される．この場合，適応回数ごとに言語重みや挿入ペナルティを最適化すると，更に性能向上が得られることが分った(学会発表[1]，[7])．

(4) 話者インデキシング
話者インデキシングとは，複数人が発声する音声において話者ごとに音声を分類する技術であり話者認識技術が用いられる．この技術は話者適応に利用できる．例えば会議音声など複数話者が発声している状況において話者適応を行う場合，話者の分類が必要であり，分類が正確なほど適応性能が向上すると考えられる．論文[5]では話者ベクトルを使用した話者照合についての提案を行い，音素情報を利用した話者ベクトルが話者分類に有効であることを示した．また実際の話し言葉では静粛な環境での発声よりも雑音下での利用が想定される．このため雑音下での検討を行った．この結果，雑音が混入する場合，話者ベクトルの軸として雑音を表現する軸

を追加することが有効であることが分かった.

(5) その他
話し言葉音声認識の性能向上に関し，静粛な環境以外で発声される音声の認識も重要な課題である．この問題に対しヒストグラム同等化(HEQ)を用いた音声正規化法の検討を行った．従来これを正確に行うためには有る程度の量の入力音声が必要であり，音声認識に組み込むと音声入力から認識結果の算出までに正規化を行うためのタイムラグが生じる．この問題について少量の入力音声で HEQ を行う手法を提案した(論文[1])．さらに時間同期処理を行うことにより，正規化によるタイムラグを減少させる手法について検討した．以上により，わずかのタイムラグでの正規化処理を可能とした.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 11 件: 全て査読付）

[1] Fumiya Takahashi, Masaharu Kato and Tetsuo Kosaka, “A time-synchronous histogram equalization for noise robust speech recognition,” Proc. of ICA 2013, (5 pages) (採録決定)

[2] Kei Sato, Masaharu Kato and Tetsuo Kosaka, “An investigation of vowel substitution rules in the automatic evaluation system of English pronunciation,” Proc. of ICA 2013, (5 pages) (採録決定)

[3] 小坂哲夫, 加藤正治,「識別学習を用いた離散混合分布 HMM による音声認識」, 情報処理学会論文誌, Vol. 54 No. 2, pp. 436-442 (2013.2),
https://ipsj.ixsq.nii.ac.jp/ej/index.php?active_action=repository_view_main_item_detail&item_id=90262&item_no=1&page_id=13&block_id=8

[4] Tetsuo Kosaka, Taro Miyamoto and Masaharu Kato, “Unsupervised Cross-Adaptation Approach for Speech Recognition by Combined Language Model and Acoustic Model Adaptation,” Proc. of APSIPA ASC 2011, Thu-PM.PS02 (4 pages) (2011.10) ,
http://www.apsipa.org/proceedings_2011/pdf/APSIPA177.pdf

[5] Tetsuo Kosaka, Naoki Tadokoro, Masaharu Kato and Masaki Kohda, “Speaker Vector-Based Verification by Phonetic Class-Based Modeling,” Journal of Information Assurance and Security, Vol. 6, No.3, pp.186-194 (2011.3)

[6] Tetsuo Kosaka, Akiyoshi Yamamoto, Takuya Kumakura, Masaharu Kato and Masaki Kohda, “Lecture Speech Recognition Using Discrete-Mixture HMMs,” IEEJ Transactions on Electrical and Electronic Engineering, Vol. 6 No. 1, pp. 23-29 (2011.1), 10.1002/tee.20602

[7] Tetsuo Kosaka, Yuui Takeda, Takashi Ito, Masaharu Kato and Masaki Kohda, “Unsupervised Speaker Adaptation Using Speaker-Class Models for Lecture Speech Recognition,” IEICE Transactions on Information and Systems, Vol. E93-D, No.9, pp. 2363-2369 (2010.9), 10.1002/tee.20602

[8] Masaru Kusumi, Masaharu Kato, Tetsuo Kosaka and Itaru Matsunaga, “Performance Improvement in Automatic Evaluation System of English Pronunciation by Using Various Normalization Methods,” Proc. of International Congress on Acoustics 2010, 257 (6 pages) (2010.8), http://www.acoustics.asn.au/conference_proceedings/ICA2010/cdrom-ICA2010/papers/p257.pdf

[9] Shunsuke Kuramata, Masaharu Kato and Tetsuo Kosaka, “Speech Recognition in Noise by Using Word Graph Combinations,” Proc. of International Congress on Acoustics 2010, 341 (6 pages) (2010.8), http://www.acoustics.asn.au/conference_proceedings/ICA2010/cdrom-ICA2010/papers/p341.pdf

[10] Tetsuo Kosaka, Takashi Ito, Masaharu Kato and Masaki Kohda, “Speaker Adaptation Based on System Combination Using Speaker-Class Models,” Proc. of Interspeech2010, pp.546-549 (2010.9) , http://www.isca-speech.org/archive/interspeech_2010/i10_0546.html

[11] Tetsuo Kosaka, Keisuke Goto, Takashi Ito and Masaharu Kato, “Lecture Speech Recognition by Combining Word Graphs of Various Acoustic Models,” Proc. of Interspeech2010, pp.2978-2981 (2010.9), http://www.isca-speech.org/archive/interspeech_2010/i10_2978.html

〔学会発表〕(計 16 件)

[1] 高木瑛, 加藤正治, 小坂哲夫,「クロスバリデーションによる教師なし言語適応における各種パラメータの最適化」,
情報処理学会東北支部研究会, 12-6-B3-5 (2013.3.11), 山形大学工学部

[2] 栗原大樹, 加藤正治, 小坂哲夫,「入力音声の韻律情報を用いた HMM 音声合成」, 情報処理学会東北支部研究会, 12-6-B3-4 (2013.3.11), 山形大学工学部

[3] 今野和樹, 大山拓也, 加藤正治, 小坂哲夫,「話者クラス音響モデルを用いた講演音声認識におけるクラスタリング手法の各種検討」, 音声言語情報処理研究報告, 2012-SLP-94-22, pp.1-6 (2012.12.21), 東京工業大学

[4] 高橋郁也, 加藤正治, 小坂哲夫,「雑音下音声認識におけるフレーム重みづけヒストグラム同等化法の検討」, 日本音響学会講演論文集, 1-1-7 (2012.9.19), 信州大学

[5] 佐藤慶, 加藤正治, 小坂哲夫,「日本人英語の自動発音評定における誤り規則の検討」, 日本音響学会講演論文集, 3-Q-2 (2012.9.21), 信州大学

[6] 倉又俊輔, 加藤正治, 小坂哲夫,「単語グラフ統合を用いた残響下音声認識の検討」, 日本音響学会講演論文集, 1-P-22 (2012.3.13), 神奈川大学横浜キャンパス

[7] 今野聡介, 加藤正治, 小坂哲夫,「教師なし話者適応における各種パラメータの最適化」, 情報処理学会東北支部研究会, 11-8-A3-1 (2012.3.9), 山形大学工学部

[8] 佐藤慶, 加藤正治, 小坂哲夫,「自動発音評定における母音置換規則の検討」, 情報処理学会東北支部研究会, 11-8-A3-2 (2012.3.9), 山形大学工学部

[9] 高橋郁也, 加藤正治, 小坂哲夫,「雑音下音声認識におけるヒストグラム同等化法の改良」, 情報処理学会東北支部研究会, 11-8-A3-3 (2012.3.9), 山形大学工学部

[10] 湊竜一, 加藤正治, 小坂哲夫,「少量のデータによるヒストグラム同等化法の検討」, 日本音響学会講演論文集, 1-10-1 (2011.9.20), 島根大学松江キャンパス

[11] 宮本太郎, 加藤正治, 小坂哲夫,「教師なし音響・言語モデル適応の性能改善」, 日本音響学会講演論文集, 2-P-31 (2011.3.10), 早稲田大学

[12] 久住大, 加藤正治, 小坂哲夫,「日本人英語の自動発音評定における精度向上の検討」, 日本音響学会講演論文集, 2-P-45 (2011.3.10), 早稲田大学

[13] 加藤正治, 小坂哲夫 , 伊藤彰則, 牧野正三,「Quinphone HM-Net に基づく講演音声認識」, 日本音響学会講演論文集, 1-9-7 pp. 21-24 (2010.9.14), 関西大学千里山キャンパス

[14] 久住大, 加藤正治, 小坂哲夫,「日本人英語と米国人英語の音素モデル間距離の検討」, 日本音響学会講演論文集, 3-1-1, pp. 253-256 (2010.9.16), 関西大学千里山キャンパス

[15] 倉又俊輔, 加藤正治, 小坂哲夫,「単語グラフ統合を用いた種々の雑音環境下での音声認識」, 電子情報通信学会技術研究報告, SP2010-41, pp.37-42 (2010.7.23), 仙台市秋保温泉

[16] 加藤正治, 小坂哲夫, 伊藤彰則, 牧野正三,「Quinphone HM-net を用いた単語グラフ統合に基づく講演音声認識」, 電子情報通信学会技術研究報告, SP2010-28, pp.37-42 (2010.6.18), 九州大学筑紫キャンパス

〔図書〕(計 2 件)

[1] 原島博 他 編, 電子情報通信学会, 電子情報通信学会知識ベース, 2 群画像・音・言語, 7 編音声認識と合成,「2-4 話者・環境適応」, 小坂哲夫(執筆担当分 3 ページ), (2011)

[2] Tetsuo Kosaka, Takashi Kusama, Masaharu Kato and Masaki Kohda, Information Science Reference, "Improvement of Lecture Speech Recognition by Using Unsupervised Adaptation," E-Activity and Intelligent Web Construction: Effects of Social Design (T.Matsuo and T.Fujimoto ed.) Chapter 16, pp.189-202, ISBN 978-1-61520-871-5 (2011)

６．研究組織

(1)研究代表者

小坂　哲夫　(KOSAKA TETSUO)

山形大学・大学院理工学研究科・教授

研究者番号：50359569

(2)研究分担者

(　　)

研究者番号：

(3)連携研究者

加藤　正治　(KATO MASAHARU)
山形大学・大学院理工学研究科・助教
研究者番号：10250953