様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ８ 日現在

機関番号：１１５０１
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３３０１８３

研究課題名（和文）大規模コーパスを利用した音声・音響信号の自動分類と音声認識への応用

研究課題名（英文）Automatic classification of speech and audio signals using large-scale corpus and its application to speech recognition

研究代表者

小坂　哲夫（Kosaka, Tetsuo）

山形大学・理工学研究科・教授

研究者番号：５０３５９５６９

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：音声コーパスの拡大および計算機の性能向上により，音声認識の高性能化が図られている．しかし音声や音響信号には話者性や背景雑音など多様性があり，それが認識率低下の原因ともなっている．本研究ではクラスタリング技術を用い，音声・音響信号の多様性による音声認識の性能劣化の問題に取り組んだ．
本研究では日本語大規模音声コーパスを用い，性質の類似した信号をクラス別にまとめ，クラスごとにモデルを構築し音声認識の性能向上を図った．研究ではガウス混合分布(GMM)ベースのモデルのみならず近年注目されているディープニューラルネットワーク(DNN)も用い検討した．

研究成果の概要（英文）：Nowadays, due to the expansion of speech corpus and advancement of computational performance, performance of speech recognition is improving. However, speech and audio signals are highly variable in terms of their features such as speaker characteristics and background noise. This variability sometimes causes the degradation of recognition performance. In this study, we investigate this problem by using clustering techniques.
We attempt to improve recognition performance by using class models trained with categorized data based on acoustic features. The training of models was carried out using the large-scale Japanese speech corpus. In this study, we utilize not only Gaussian mixture models (GMMs) but also deep neural networks (DNNs) as acoustic models.

研究分野：音声情報処理

キーワード：音声認識　音響モデル　クラスタリング　隠れマルコフモデル　ディープニューラルネットワーク

１．研究開始当初の背景

音声認識の分野では，隠れマルコフモデル(HMM)などの統計的確率モデルの利用により研究が進展してきた．確率モデルではパラメータ推定のための学習データの量が問題となるが，近年大規模な音声データの使用が可能となったため，その性能が大きく向上している．既にスマートフォンなどの小型携帯端末では，検索語の入力や電子秘書への問い合わせなどに音声認識技術が利用されている．しかし，一般にパターン認識では，学習データと同一の性質を持つ入力に対しては良好な認識性能を示すが，異なる場合は性能が低下するという問題がある．例えば，雑音のない環境で収録された音声データを用いて音響モデルを学習した場合，雑音下の音声に対しては低い認識精度を示す．この問題に対しては，あらかじめ多様な雑音下の音声を収録して学習に利用する，マルチコンディション学習が提案されている．この方法では種々の雑音環境を含むため，ある程度の認識性能を示すが，学習と認識が同一条件の場合と比較すると十分な性能は得られない．一方学習と認識を同一の条件に近づけるためクラス分類に基づく方法が種々提案されている．この方法では，学習時には学習データを分類して類似したデータごとに音響モデルを作成する．次に認識時には音響的な条件が近いモデルを自動選択し使用することにより，学習と認識における音響的ミスマッチの低減を図る．クラス分類としては一般的には性別や年齢などの先験的な知識が使用されている．

２．研究の目的

従来，クラスモデルに基づく音声認識においては，性別や年齢層別など 2～数クラス程度のモデルを構築して使用するのが一般的であった．しかし近年大規模な音声コーパスが使用可能となったため，多種類のクラスモデルを作成するこが可能となってきている．本研究では従来行なわれてこなかった 100 クラス以上の多種類のクラスモデルを構築し音声認識の性能向上を目指す．

以上説明したクラスモデルに基づく方法では，(1)学習時，どのような方法で類似したデータをまとめるか，および，(2)認識時に条件が類似するモデルをどのように自動選択するか，の 2 点が特に重要となる．この 2 点について検討を行ない音声認識性能の向上を目指すのが本研究の目的である．また話者の特性による分類だけではなく，雑音特徴の分類についても検討し，それを音声区間検出へと応用する．

３．研究の方法

前記クラスモデルに基づく方法における 2 つの重要な点について，以下の方法で検討を行なう．まず(1)の点について，クラス数が増加した場合１クラスに属するデータ数の減少によりモデル精度の低下が懸念される．しかしデータの複数クラスへの重複を許すことにより，この問題の回避が可能である．大規模な話者クラスモデルの構築時に，どのような重複の方法が有効かの検討を行なう．次に(2)の選択法について，一般的には尤度などの情報により入力音声に適したモデルを一つ選び認識に使用するのが一般的である．しかし必ずしも入力話者に良くマッチするクラスモデルが存在するとは限らない．その場合は複数のモデルを選択し，併用する方法が有効となる．本研究ではその方法をさらに発展させ，最尤推定により複数モデルに重みを付与し使用する方法について検討する．また(1)や(2)以外の検討事項として，クラスモデルを音声認識に組み込む場合の検討やクラスモデルのモデル化手法の比較なども行なう．さらに話者のみならず雑音の種類ごとのモデル化についても検討する．

４．研究成果

以下，本研究課題における主要な研究成果を項目ごとに説明する．

(1) 大規模な話者クラスモデル音響モデルを用い，音声認識の精度向上の検討を行なった．まず話者クラス音響モデル作成の際の類似話者の選択方法にハードクラスタリングとソフトクラスタリングを用い，それぞれの性能の比較を行った．その結果，クラスタ数が大きく増加した場合に（数百以上のクラス数），１人の学習話者が複数のクラスタに重複して属することのできるソフトクラスタリングが非常に有効であることが分かった．話者クラス音響モデル作成の際の類似話者の選択方法には種々あるが，我々は手法として話者クラスタリング法を用いた．話者クラスタリング法とは，学習話者に対して事前にクラスタリングを行ってモデルを作成し，その中から認識対象話者に近いモデルを選択する方法である．しかし，以上のような方法によって作成された話者クラス音響モデルは，2 パス音声認識システムにおいて，従来は単語グラフの音響リスコアにのみ用いられていた．認識過程における 1st パスでは不特定話者（SI）モデルを用いているため，アライメントは SI モデルで行われるが，2nd パスでの音響リスコアは話者クラス音響モデルで行うため，この際にミスマッチが生じ，認識精度に悪影響を与えている可能性があった．そこで学会発表文献⑪では 2nd パスで選択された話者クラス音響モデルを，1st パスから用いて再認識を行い，上記のミスマッチを改善することで，認識精度の向上を目指した．実験の結果，従来法や性別による話者クラスモデルを使用した場合と比較し性能向上が得られることが分かった．また最適なモデルを選択した場合の性能上限の検討の結果から，より性能を上げるためには，話者クラスモデル自体の改良よりも，選択方法自体の検討が重要であることが分かった．

(2) これまで GMM（ガウス混合分布モデル）に基づく HMM，すなわち GMM-HMM を音響モデルとして話者クラスモデルの検討を行なってきた．しかし近年深層学習による Deep Neural Network(DNN)による HMM，すなわち DNN-HMM が，その性能により注目を浴びている．そこで，学会発表文献④では話者モデルの構造として DNN-HMM を用いた場合の検討を行なった．基本的に話者をクラスタリングする方法については従来と同様な方法を用いた．また複数話者クラスモデルを選択するに当たっては 2 つの手法の比較をした．ひとつは多数の話者クラスモデルから尤度により１つのモデルを選択する方法である．もう一方は，選択した複数のモデルに対し，最尤推定に基づいて重みをつける方法である．まず，音響モデルとして GMM-HMM から DNN-HMM に変更することにより，全体的に性能向上が得られた．またモデル選択法の比較としては，最尤推定による重みづけ法においてより高い性能を得ることができた．

(3) 話者をクラス分類し，話者クラスモデルを作成し利用する方法の有効性を(1)および(2)で示したが，他の応用におけるクラスモデルの有用性についても検討した．学会発表文献⑨では，音声区間検出におけるクラス分類について検討した．音声区間検出の代表的な手法として，音声区間用の音響モデルおよび非音声区間用の音響モデルの 2 つを作成し，両者の尤度を比較して，音声区間であるか否かを判定する手法がある．しかし音響特徴には多様性があり，単純に 2 クラスでモデルを作成しても性能に限界がある．特に雑音の種類や，雑音と音声がオーバーラップする場合，しない場合などで様々な音響的な特徴を示す．このため，音響的特徴の種類により複数のクラスを設定し，音声区間検出の精度向上を目指した．その結果，音響的特徴の違いにより 4～6 種類のモデルを作成して使用することにより性能向上が得られることが分かった．更に学会発表文献①では，音響モデルとして GMM の代わりに DNN を使用した実験を行なった．GMM を DNN に変更しただけでも大幅な性能向上が得られるが，さらに複数モデルを使用することにより性能向上が得られることが分かった．

(4) 上記(1)～(3)のクラス分類による音声認識や音声区間検出に関連して，いくつか音声認識の精度向上に関する検討を行なった．学会発表文献⑥では，大語彙連続音声認識の誤り結果の解析を行なった．特に音響モデルの性能が極めて良い場合を仮定し，音響モデルに関係する誤り以外，どのようなものが存在するかについて検討した．分析の結果，発音同一に関する誤りが 55%を占めることが分かった．発音同一とは同音異義語のように音素の系列レベルでの誤りではなく単語レベルで誤る場合である．また話者の話し方により，誤り傾向が大きく異なることが分かった．学会発表文献⑧では，DNN-HMM を用いた話者適応の検討を行なった．異なる種類の話者適応法を順次適応することにより，同一手法の適応を繰り返すよりも適応性能が向上することを示した．これらの成果とクラス分類手法を組み合わせることにより，更なる音声認識の性能向上が見込まれる．

以上の研究成果をまとめ，今後の展望を以下に述べる．研究期間内に従来の GMM-HMM と比較し DNN-HMM を用いることにより大幅な性能向上が得られることが広く認知されるようになった．これにより 1990 年台初頭から続いた GMM-HMM ベースの音声認識の研究は DNN-HMM ベースへの研究へと取って代わられつつある．本課題でもいち早くこの流れに対応し，研究のベースを DNN-HMM にシフトさせている．ここで問題となるのが，DNN-HMM における話者適応の手法についてである．従来 GMM-HMM での話者適応はガウス分布が持つ平均や分散などのパラメータに基づく手法が主流であった．一方 DNN においてはそのような方法は使うことができないため，新たな手法を開発する必要が出てきている．しかし本課題で検討した話者クラスモデルに基づく話者適応においては，そのようなパラメータを必要としないため，GMM-HMM を DNN-HMM に置き換えるだけで，話者適応が実現可能となる．実際 DNN-HMM に置き換えても，GMM-HMM で検討した結果と同様に，話者適応が可能であることが示された．DNN-HMM の話者適応手法は他にもいくつか提案されているが，教師なしでの適応は難しい，あるいは教師なしで適応しようとすると，多くの適応データが必要になるなどの問題が生じる(引用文献①)．一方，本課題の手法に基づく DNN-HMM の話者適応では少量の適応データで教師なし話者適応が可能となる．以上 DNN-HMM における教師なし話者適応の可能性を示した点で，大きな成果が得られたと考えられる．また学会発表文献⑧に示したクロス適応と組み合わせることにより更なる性能の向上が得られると期待できる．以上から今後は DNN-HMM の適応について，教師なしで，より少量のデータで，より高い性能向上を得ることが目標になっていくと考えられる．

＜引用文献＞

① S.Xue, O.A.-Hamid, H.Jiang and L.Dai, Direct adaptation of hybrid DNN/HMM model for fast speaker adaptation in LVCSR based on speaker code, Proceeding of ICASSP2014, 2014, 6339-6343

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 4 件）
① Tetsuo Kosaka, Kazuki Konno, Masaharu Kato, Deep nearal network-based speech recognition with combination of speaker-class models, Proceeding of APSIPA ASC 2015, 査読有, SP2-2.3, 4 pages, DOI: 10.1109/APSIPA.2015.7415464

② Akira Takagi, Kazuki Konno, Masaharu Kato, Tetsuo Kosaka, Unsupervised cross-adaptation using language model and deep learning based acoustic model adaptation, Proceedings of APSIPA ASC 2014, 査読有, WA-P-16, 2014, 4 pages, DOI: 10.1109/APSIPA.2014.7041581

③ Kazuki Konno, Masaharu Kato and Tetsuo Kosaka, Speech recognition with large-scale speaker-class-based acoustic modeling, Proceedings of APSIPA ASC 2013, 査読有, OS.28-SLA.9, 113, 4 pages, DOI: 10.1109/APSIPA.2013.6694112

④ 小坂哲夫, 伊藤貴, 加藤正治, 好田正紀, 話者クラス音響モデル及び単語グラフ統合を用いた音声認識, 電子情報通信学会論文誌, 査読有, Vol. J96-D, No.11, 2013, 2795-2803, http://search.ieice.org/bin/summary.php?id=j96-d_11_2795

〔学会発表〕（計 12 件）
① 菅郁巳, 安原龍, 井上雅史, 小坂哲夫, ディープニューラルネットワークを用いた映画中の音声区間検出の検討, 日本音響学会講演論文集, 2016.3.9-2016.3.11, 桐蔭横浜大学

② 今野和樹, 加藤正治, 小坂哲夫, ディープニューラルネットによる話者クラス音響モデルを用いた音声認識, 日本音響学会講演論文集, 2015.9.16-2015.9.18, 会津大学

③ 高木瑛, 加藤正治, 小坂哲夫, DNN-HMM を用いた教師なしクロス適応の性能改善の検討, 日本音響学会講演論文集, 2015.3.16 - 2015.3.18, 中央大学後楽園キャンパス

④ 今野和樹, 加藤正治, 小坂哲夫, 最尤推定による話者クラス DNN の出力統合を用いた音声認識, 日本音響学会講演論文集, 2015.3.16 - 2015.3.18, 中央大学後楽園キャンパス

⑤ 小野瑞穂, 加藤正治, 小坂哲夫, DNN-HMM を用いた音声認識におけるパラメータ数の検討, 情報処理学会東北支部研究会, 2015.3.4, 山形大学工学部

⑥ 小野瑞穂, 小関翔太, 加藤正治, 小坂哲夫, Deep Learning による教師つき適応の結果を用いた日本語講演音声認識の誤り解析, 日本音響学会講演論文集, 2014.9.3-2014.9.5, 北海学園大学豊平キャンパス

⑦ 今野和樹, 高木瑛, 加藤正治, 小坂哲夫, 音声認識における DNN を用いた話者クラスモデルの検討, 電気関係学会東北支部連合大会, 2014.8.21-2014.8.22, 山形大学工学部

⑧ 高木瑛, 今野和樹, 加藤正治, 小坂哲夫, DNN-HMM を用いた音響モデルおよび言語モデルのクロス適応, 情報処理学会音声言語情報処理研究報告, 2014.5.22-2014.5.23, 東京工業大学大岡山キャンパス

⑨ 佐々木志貢, 加藤正治, 小坂哲夫, 雑音重複区間のモデル化による音声区間検出の性能向上, 情報処理学会東北支部研究会, 2014.3.14, 山形大学工学部

⑩ 小坂哲夫, 今野和樹, 高木瑛, 加藤正治, DNN-HMM を用いた日本語講演音声認識における話者適応の検討, 日本音響学会講演論文集, 2014.3.10-2014.3.12, 日本大学理工学部

⑪ 今野和樹, 加藤正治, 小坂哲夫, 大規模話者クラス音響モデルを用いた音声認識の精度向上の検討, 日本音響学会講演論文集, 2013.9.25-2013.9.27, 豊橋技術科学大学

⑫ 加藤正治, 小坂哲夫, 単語グラフを用いた音声アライメント, 日本音響学会講演論文集, 2013.9.25-2013.9.27, 豊橋技術科学大学

〔図書〕（計　1 件）
① 小坂哲夫 他, (株)ニッケイ印刷, 進化するヒトと機械の音声コミュニケーション第 1 編第 2 章, 2015, 31-40

〔その他〕
ホームページ等
http://speech-lab.yz.yamagata-u.ac.jp/

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
なし　　（　　）

研究者番号： なし

(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