様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ５ 月 １６ 日現在

機関番号：１１５０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００２２７

研究課題名（和文）ディープラーニングに基づく音声認識の音響モデル適応の研究

研究課題名（英文）A study on acoustic model adaptation for deep-learning-based speech recognition

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

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：近年ディープラーニングにもとづく音声認識が大きな成果を挙げているが，話し言葉についてはまだ十分な結果は得られていない．認識性能の低下の大きな原因として話者の個人性，多様な音響環境，多様な発話スタイルなどが挙げられる．これらを解決するために音響モデル適応を中心とした技術を検討し，認識性能の向上を図った．結果として話し言葉音声や感情音声の認識精度の向上，雑音下における音声区間検出の性能向上を達成した．

研究成果の学術的意義や社会的意義
本研究により，1)話し言葉音声認識における適応精度の向上，2)雑音下音声区間検出の精度向上，3)感情音声認識の性能向上を達成した．1)は話し言葉音声認識に限らず，異なる分野においても応用可能な適応手法で汎用性の高い技術である．2)の成果を利用してマルチモーダル対話コーパスが整備されており，当該分野の研究者にとって有益と考えられる．また3)についてもロボットと人間との会話など様々な分野に利用が可能である．以上，本研究で開発した技術は波及効果が高く，学術的，社会的意義が高いと考えられる．

研究成果の概要（英文）：Although the deep-learning-based speech recognition technology has made great achievements in recent years, the spontaneous-speech-recognition technology has not yet obtained sufficient results. As major factors of performance degradation in speech recognition, a variety of speaker characteristics, acoustic environments, and speaking styles can be mentioned. To solve these problems, I developed techniques centered around acoustic-model adaptation to improve the speech-recognition performance. Consequently, performance improvement was achieved with regard to spontaneous and emotional speech. Additionally, the performance of voice-activity detection was also improved.

研究分野：音声情報処理

キーワード：音声認識　音響モデル　ディープニューラルネットワーク　適応技術　話し言葉　感情音声　音声区間検出

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

近年, 人工知能の分野でディープラーニングが注目されている. 音声認識の分野では 2011 年にディープニューラルネットワーク(DNN)に基づく隠れマルコフモデル(DNN-HMM) を用いた大語彙音声認識手法が発表され，その高い認識精度が注目された(文献①)．日本においても 2012 年以降, この分野の研究が活発化している. 図 1 に近年の音声認識の性能の変遷を示す. 図では各研究機関が共通の評価基盤として用いている日本語話し言葉コーパス(CSJ) のテストセット 1 の単語誤り率を示した．コーパスが構築された 2000 年代中頃は研究が活発化し，様々な成果が得られたが，その後ブレークスルーとなる技術が見つからず，研究は一時停滞した．しかし DNN の技術を導入することにより大幅な性能向上が可能となったのが研究開始当初の状況である．平均的な認識性能は向上したが，話者性の問題については依然解決していない．従来法である混合ガウス分布に基づく HMM (GMM-HMM) と DNN-HMM の性能比較を比較すると，全体的には性能向上しているものの, 従来法で認識が困難な話者については, DNN でも困難で話者性の問題の解決が望まれていた．また音響環境の変動や発話スタイルの変動についても十分な頑健性が得られていないという問題があった.

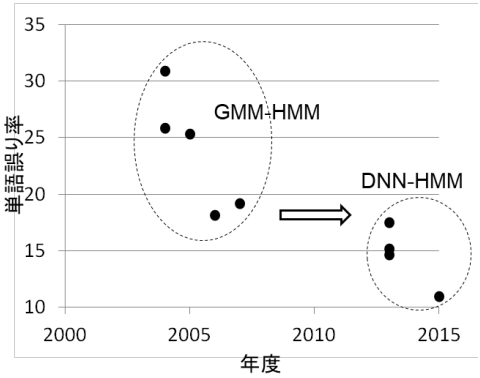

図 1: CSJ における単語誤り率の変遷

２．研究の目的

本研究は DNN に基づく音声認識において，音響モデル適応技術を中心とした手法により，話し言葉音声認識，雑音下音声認識，多様な発話スタイルの音声認識等の認識性能の向上を図るのが目的である．DNN の適応について最も単純な方法は，使用可能な適応データを用いて DNN のすべてのパラメータを更新する方法であるが，この方法では適応データが大量に必要になり，効率的な適応はできない．効率的な適応法として話者間の相関関係を事前知識として得ておき，それを利用する方法が考えられる(文献②)．この方法のほかクロス適応など効率的に適応を進める方法について種々検討し音声認識の高精度化を図る．まずは話し言葉音声認識による検討を行うが，さらに雑音や発話スタイルといった各種変動に頑健な認識手法についても検討を行う.

３．研究の方法

相関を使う方法，クロス適応による適応法についてその効果を，話し言葉音声認識を対象として検討する．相関を使用する方法としては文献③の話者ベクトルを利用し，入力特徴に話者特徴を加えることにより DNN を学習する．クロス適応としては音響モデル適応と言語モデル適応を交互に繰り返し性能向上を図る．この方法では２つの適応法の誤り傾向が異なることを利用する．誤り傾向が異なれば，相補的効果により同一の適応を繰り返す場合に比べ認識性能が向上することが期待できる．話し言葉音声認識のデータとしては CSJ を使用する．このコーパスは学会講演などを対象としているが，講演者にヘッドセットをつけて収録しており雑音の影響は少ない．また講演音声の場合，発話スタイルのバリエーションは多くない．よって CSJ 以外のコーパスを利用して，雑音下音声認識の性能向上を目指し雑音下の音声検出について検討する．さらに多様な発話スタイルに対する音声認識の性能向上を目指し，感情音声認識の検討を行う．いずれにおいても近年その有効性が知られているフィードフォワード型の DNN を利用する.

４．研究成果

（1）話者間相関を利用する話者適応の検討を行った．多数話者の相関関係を表す方法として文献③で提案した話者ベクトルが挙げあられる．これは話者認識のために使用された特徴であるがこれを話者適応に利用する．方法としてはメルフィルタバンクの特徴量 825 次元に話者ベクトルを追加し学習を行う．認識時には入力話者の話者ベクトルを求め，同様にメルフィルタバンク特徴に追加し入力特徴とする．以上により学習話者の相関関係を利用して，入力話者の話者特徴をベクトル化する．本方式では学習話者数が話者ベクトルの次元数となるが，実験では学習話者が 963 話者と多く特徴次元も過多となる．このため主成分分析により次元圧縮して利用する．話者ベクトルが 100 次元のとき，ベースラインの単語誤り率 14.86%に対し，適応後の単語誤り率が 14.58%となり性能向上が見られた．このように話者適応としては有効であるが, 計算量に対し向上率が少ないという問題点が残った．今後は計算量削減ついて検討する必要がある.

（2）クロス適応の高精度化について検討を行った．教師なし話者適応の方法として，誤り傾向の異なる認識結果を利用して相互に補完することを目的としたクロス適応が提案されている. この手法はクロスシステム適応とクロスバリデーション適応の 2 種類に分類される．前者においては基本的には同一のデータで異なるシステムを組み合わせて使用する．一方後者ではシス

テムは同一であるが異なる適応データを使用する．本研究では前者のクロスシステム適応に関する検討を行った．この場合異なる認識結果を出力する複数のシステムが必要になるが，本研究では GMM-HMM，N-gram 言語モデルのほか DNN-HMM を用い 3 種類の教師なし適応をクロスすることにより性能向上を目指した．特に検討時点で CSJ に関しほぼ最高性能を達成している認識システム(文献④)から更に性能向上が出来るかを検討した．2 種類のベースラインから提案法，DNN-HMM 音響モデルの繰り返し適応，DNN-HMM+N-gram 言語モデルの並列繰り返し適応，の 3 種の結果(単語誤り率)を表 1 に示す．以上より，いずれのベースラインでも提案法が最高性能を示すことが分かった．一方単純な繰り返し適応では性能向上が少ないか，逆に性能が劣化する場合もあることが分かった．その他有益な知見として，誤り傾向の差が大きい適応法を優先してクロスさせることが有効であることが分かった．またクロス適応を行うことにより従来性能が低いと思われた GMM-HMM でも高い性能(誤り率 11.35%)が得られることが分かった．本手法は異なる教師無し適応を交互に繰り返すというシンプルな方法ながら効果が高く，音声認識以外多くの分野で利用できる可能性があり，波及効果が高いと考えられる．

表 1 単語誤り率(%)による各種適応法の比較．CSJ テストセット 1 による評価．

実験の種類	ベースライン	DNN-HMM 適応	DNN-HMM+N-gram 適応	提案法
CSJ-BASE	14.47	13.54	12.69	12.45
Kaldi-BASE	10.97	11.04	10.59	10.38

（3）雑音下で話者適応などを行う場合，正確な音声区間検出(Voice activity detection: VAD)が必要となる．雑音が少ない場合はパワーやゼロクロスなど単純なパラメータで VAD が可能であるが，様々な種類の雑音下では検出精度が大幅に低下する．例えば文献⑤や文献⑥では映画音声中の音声検出を検討しているが，検出における等誤り率(EER)がそれぞれ 33.18%，13.0%となっている．本研究では特徴量抽出法に雑音を考慮したフロントエンドを使用，学習データ量の増大，識別器に DNN を使用といった工夫の他，雑音の性質が異なる音楽や歌について別クラスの雑音モデルを用意するという方法を提案し VAD の高精度化を達成した．代表的な VAD 手法の比較を表 2 に示す．VAD の高精度化は音声認識に限らず雑音下の音声処理一般に広く応用可能で，波及効果が高いと考えられる．また提案した VAD を利用し大規模マルチモーダル映画対話コーパスを構築した(雑誌論文④)．これは同分野の研究に大いに役立つと考えられる．

表 2 映画における各種 VAD 手法の比較

各種手法	特徴量	識別器	学習データの種類	学習データ量(時間)	EER (%)
文献⑤	RASTA-PLP+Δ	LSTM-RNN	音声＋人工付加雑音	34.9	33.18
文献⑥	多種特徴量の混合	SVM	ラジオ放送	4.0	13.0
提案法	ETSI-FE (MFCC+Δ+ΔΔ)	DNN	ホームビデオ	69.5	3.92

（4）多様な発話スタイルに対する適応手法を用いた認識性能向上の検討を行った．多様な発話スタイルの中でも感情音声は通常の発声とは音響的特徴が異なり認識が困難である．まずは基礎的な検討として従来一般的に音響モデルとして用いられてきた GMM-HMM に代わり DNN-HMM を利用することの有効性を示した．次に更なる性能向上を目指し，音響モデル，言語モデルの適応を行った．音響モデルの適応では話者適応，コーパス適応，感情適応，話者＋感情適応の４種類について比較実験を行なった．Japanese Twitter-based Emotional Speech(JTES)を対象とした認識実験の結果を表 3 に示す．いずれも効果があるが，特に話者適応で良好な結果が得られた．更に言語モデル適応の検討を行った．Twitter は口語的表現が多く含まれるが，その中から感情表現を含むテキストを選択し適応に利用した.約 2000 文を使用した適応実験の結果，言語モデル適応の有効性を示した．現状では音響モデル適応，言語モデル適応を別個に行っているが，今後は両者を合わせた実験を行い更なる性能向上を目指す予定である．本研究は感情音声認識の性能向上を図ったものであるが，この結果は音声による感情認識や感情音声の声質変換など様々な分野に応用可能であり，波及効果が大きいと考えられる．

表 3 JTES を対象とした適応実験．単語誤り率(%)を示す．

感情	ベースライン	話者適応	コーパス適応	感情適応
喜び	40.97	27.38	32.62	34.15
怒り	41.23	25.56	28.61	28.61
悲しみ	39.09	22.88	26.52	25.76
平常	23.13	16.39	19.88	19.52
平均	36.11	23.05	26.91	27.01

＜引用文献＞
① F. Seide, et al.: Proc. of Interspeech2011, pp. 437-440 (2011).
② 篠田: 電子情報通信学会誌, Vol.J87-D2, No.2, pp. 371-386 (2004).
③ 小坂他: 電子情報通信学会誌, Vol.J90-D, No.12, pp. 3201-3209 (2007).
④ 森谷他: 音響学会秋季講演論文集, pp. 155-156 (2015).
⑤ F. Eyben, et al.: Proc. of Interspeech2013, pp. 483-487 (2013).
⑥ B. Lehner, et al.: Proc. of Interspeech2015, pp. 2942-2946 (2015).

５．主な発表論文等
〔雑誌論文〕（計 4 件）
① Tetsuo Kosaka, Yoshitaka Aizawa, Masaharu Kato, Takashi Nose, Acoustic Model Adaptation for Emotional Speech Recognition Using Twitter-Based Emotional Speech Corpus, Proc. of APSIPA ASC 2018, pp. 1747-1751, 2018, 査読有, DOI: 10.23919/APSIPA.2018.8659756

② Tetsuo Kosaka, Ikumi Suga, Masashi Inoue, Improving Voice Activity Detection for Multimodal Movie Dialogue Corpus, Proc. of IEEE GCCE2018, pp. 453-456, 2018, 査読有, DOI: 10.1109/GCCE.2018.8574730

③ 冨田建斗，高木瑛，加藤正治，小坂哲夫，ディープニューラルネットワークを用いた教師なしクロス適応による音声認識, 電子情報通信学会論文誌, Vol.101-D, No.8, pp. 1190-1199, 2018, 査読有, DOI: 10.14923/transinfj.2017JDP7076

④ Ryu Yasuhara, Masashi Inoue, Ikumi Suga and Tetsuo Kosaka, Large-scale multimodal movie dialogue corpus, Proc. of the 18th ACM International Conference on Multimodal Interaction, pp. 414-415, 2016, 査読有, DOI: 10.1145/2993148.2998523

〔学会発表〕（計 12 件）
① 佐伯和哉，加藤正治，小坂哲夫，言語モデルの改良による感情音声の認識と韻律制御声質変換の性能向上，情報処理学会東北支部研究報告，2019.3.

② 小坂哲夫，相澤佳孝，加藤正治，能勢隆，感情音声認識における音響モデル適応と声質変換への応用，日本音響学会秋季講演論文集，1-R-37，2018.9.

③ 冨田建斗，加藤正治，小坂哲夫，DNN を用いた教師なしクロス適応の性能評価，情報処理学会東北支部研究会，2017-6-B2-2，2018.3.

④ 相澤佳孝，小坂哲夫，加藤正治，能勢隆，感情音声データベース JTES を用いた感情音声認識におけるモデル適応の性能向上の検討，情報処理学会研究報告，Vol.2017-SLP-119 No.7，2017.12.

⑤ 菅郁巳，小坂哲夫，井上雅史，DNN を用いた映画の音声区間検出におけるクラス分類の検討，日本音響学会秋季講演論文集，1-R-2，2017.9.

⑥ 相澤佳孝，小坂哲夫，加藤正治，能勢隆，感情音声データベース JTES を用いた感情音声認識における DNN-HMM 音響モデル適応の検討，日本音響学会秋季講演論文集，1-R-17，2017.9.

⑦ 井上雅史，安原龍，菅郁巳，小坂哲夫，映画からのマルチモーダル対話コーパスの作成，人工知能学会全国大会，2H5-OS-35c-1in1，2017.5.

⑧ 笹田拓臣，相澤佳孝，小坂哲夫，DNN による音声認識を用いた感情音声の声質変換の検討，情報処理学会東北支部研究会，2016-7-A3-4，2017.3.

⑨ I. Suga, R. Yasuhara, M. Inoue and T. Kosaka, Voice activity detection in movies using multi-class deep neural networks, 5th Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan, 2pSC68, 2016.

⑩ Y. Aizawa, M. Kato and T. Kosaka, Many-to-many voice conversion using hidden Markov model-based speech recognition and synthesis, 5th Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan, 1aSC27, 2016.

⑪ 冨田健斗，高木瑛，加藤正治，小坂哲夫，高精度な初期モデルを用いた教師なしクロス適応の評価，日本音響学会春季講演論文集，3-Q-5，2016.9.

⑫ 相澤佳孝，中川由暁，加藤正治，小坂哲夫，HMM 認識・合成による感情音声の声質変換の性能向上，日本音響学会春季講演論文集，3-Q-32，2016.9.

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

６．研究組織
(2)研究協力者
研究協力者氏名：加藤　正治
ローマ字氏名：KATO, Masaharu