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　　７　日現在

機関番号：　１３７０１
研究種目：　若手研究(B)
研究期間：　2013～2015
課題番号：　２５７３０１０９

研究課題名（和文）環境に応じたマルチモーダル音声認識の構成最適化手法の研究

研究課題名（英文）Investigation of method optimization for multi-modal speech recognition

研究代表者
田村　哲嗣（Tamura, Satoshi）

岐阜大学・工学部・助教

研究者番号：１０４０２２１５

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：本研究では、音声と画像を用いるマルチモーダル音声認識において、その要素技術や処理を環境やタスクに応じて最適化する手法の確立を目指した。異なる基本特徴を組み合わせ深層学習を適用することの有効性、音声・画像統合法の最適形態と認識モデルの確率的統合法の有用性、および個人・環境適応の改善による性能改善を確認し、これらにより、頑健かつ高性能なマルチモーダル認識アルゴリズムを構築した。本研究によるマルチモーダル音声認識を多種多様なタスク・環境において試用し、性能を確認するとともに、今後に向けた課題を明らかにした。

研究成果の概要（英文）：For multi-modal speech recognition that uses speech signals and lip images, this research aimed at development of method optimization according to tasks and environments. Effectiveness of incorporating several basic features and applying deep-learning techniques, the optimal architecture of audio-visual integration in addition to effectiveness of stochastic model combination, and improvement of model adaptation were clarified. A robust and high-performance multi-modal speech recognition method was thus developed. The method was applied in various tasks and environments, then recognition improvement was observed and future works were also found.

研究分野：　音声情報処理・マルチモーダル情報処理

キーワード：　音声認識　マルチモーダル情報処理　読唇　最適化　実環境

１．研究開始当初の背景

雑音環境下での音声認識性能改善手法として、音声信号と口唇動画像を用いるマルチモーダル音声認識（Multi-Modal Speech Recognition、以下 MMSR）が注目され、研究が行われている。MMSR の実現には、さまざまな要素技術の研究が必要である。例として、雑音に対する音声信号の前処理や画像外乱に頑健な顔検出、音声信号や口唇動画像からの有効な特徴量の抽出、音声と画像の統合法、および音声区間検出やモデル適応など関連技術の開発などが挙げられる。

これまで本研究代表者は、雑音環境を考慮した MMSR に関する認識実験を行ってきた。その結果、上述した MMSR に関するさまざまな要素技術は、おのおの特性や利点が異なることが分かってきた。例えば、音声と画像の統合にはいくつかの方法があるが、それぞれ音声と画像の結合度が異なり、このため雑音や外乱に対する挙動や性能に違いが生じる。また、特徴量抽出やモデル構築においては、特定環境に強い手法もあれば汎用的に性能発揮する手法もあり、各々使い分ける必要があると考えられる。

上記のことは、環境やタスクによって最適な要素技術の組み合わせが異なる可能性を示唆している。言い換えれば、環境やタスクにより最適な認識手法を構成できれば、さらなる認識率の向上が見込めると考えられる。

２．研究の目的

本研究は、(1) MMSR において、要素技術の最適な活用を行う「構成要素最適化手法」について検討する。また、(2)異なる認識環境において本研究による MMSR を適用し、その有効性を明らかにする。

(1)に関して、本研究では主に以下の 3 点について取り組み、これらを通じて要素技術の最適利用法を検討する。

① 複数の特徴量の特性や性能を調べ、それらを適宜組み合わせることの有効性を明らかにする。また、さらに深層学習を適用することによる性能改善を示す。

② MMSR における音声と画像の統合方法について調査・検討を行う。また、最適化の一つとして、複数の認識モデルを確率的に統合する方法を構築しその有効性を示す。

③ モデル適応の改良による MMSR のための個人・環境適応手法について検討する。

また(2)に関しては、本研究では主として次の２点に取り組む。

④ 小語彙タスクやさまざまな雑音環境において上記の手法を適用し、MMSR の性能改善を目指す。

⑤ 大語彙タスクやモバイル端末環境など研究事例の少ない環境において、本研究による手法の評価を行う。

３．研究の方法

(1)「構成要素最適化手法」の検討

① 特徴量の組み合わせでは、特に認識性能がこれまで不十分であった画像特徴量について検討する。具体的には(ⅰ)複数の話者において、これまで読唇で用いられてきた特徴など以下の 5 種類を求め、話者全体・話者個別の性能を調査する。

- 主成分分析（PCA）
- 離散コサイン変換（DCT）
- 線形判別分析（LDA）
- 識別的特徴量（GIF）
- 口唇特徴点情報（COORD）

次に、(ⅱ)これら特徴量を組み合わせた場合の認識性能について、同様に性能を調査する。これに加え、(ⅲ)音声特徴量や上記の画像特徴量を入力として深層学習による特徴抽出を行い、認識実験を通じてその性能を調査する。

② 本研究では音声と画像の統合法として、まず(ⅰ)深層学習の活用法とモデル化の異なる以下の６つの手法の性能を調査する。なお、従来特徴量として、音声は MFCC、画像は PCA を用いることとする。

- 深層学習によらない従来特徴量＋GMM-HMM（Baseline(l)）
- 深層学習によらない従来特徴量＋マルチストリーム GMM-HMM（Baseline(m)）
- 深層学習によらない従来特徴量＋DNN-HMM（Hybrid）
- 音声特徴量と画像特徴量を統合し深層学習を適用して得られる特徴量(DBAVF)＋GMM-HMM（Tandem(i)）
- 音声特徴量に深層学習を適用して得られる特徴量(DBAF)と画像特徴量に深層学習を適用して得られる特徴量(DBVF)を統合して得られる特徴量＋GMM-HMM（Tandem(l)）
- 音声特徴量に深層学習を適用して得られる特徴量(DBAF)と画像特徴量に深層学習を適用して得られる特徴量(DBVF)を統合して得られる特徴量＋マルチストリーム GMM-HMM（Tandem(m)）

マルチストリーム GMM-HMM においては、音声と画像を重み付けする必要がある。この重み付けを何らかの基準で決める従来のアプローチに代えて、(ⅱ)複数の重みを設定した認識モデルを確率的に統合する手法について検討する。

③ モデル適応について、本研究では実用面を考慮して、教師なし適応に着目する。従来は、音声認識→モデル適応→再認識というプロセスであったが、本研究では MMSR の特性に着目し、音声認識とモデル適応を繰り返し適応する手法について検討する。

(2) 各種タスク・環境での MMSR の適用

④ 本研究では、乗用車での利用を想定した条件で、数字認識タスクにおいて、上述の手法を評価する。具体的には、市街地走行雑音(cityroad)、高速道路走行雑音(expressway)、カーステレオを想定した音楽データ(music)を用意し、さらに music と走行雑音を両方重畳した雑音も用意する。これを 6 種類の SNR で音声に重畳する。そしてテストデータに対して MMSR を行い、本研究の有効性と性能改善を明らかにする。

⑤ 上記に加えて、以下の 2 種類のタスク・環境において本研究による成果を適用し、基礎的検討と性能評価を行う。まず、(ⅰ)大語彙タスクにおいて MMSR を適用する。また、(ⅱ)モバイル端末で音声と口唇動画像の録音・録画を行うアプリを作成し、これを用いてさまざまな実環境でデータを収録する。このデータを用い音声認識を行う。

4．研究成果

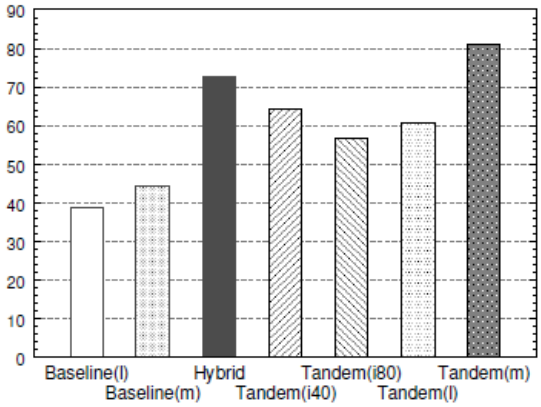
(1)「構成要素最適化手法」の検討

①（ⅰ）研究用データベース CENSREC-1-AV 構築に用いた顔画像を用いて読唇実験を行った。連続数字タスクの口唇動画像で、学習に 42 名・3,234 発話、テストに 51 名・1,963 発話を用いた。結果の認識率を次表に示す。5 種類の特徴量は 40%前後とほぼ同様だが、話者により最適・最良な特徴量が異なることが判明した。

PCA	DCT	LDA	GIF	COORD
42.52	33.06	41.70	39.76	39.78

(ⅱ)性質の異なる上記5種類の特徴量を統合した結果、認識率は 60.21%となった。ほとんどの話者において統合後の特徴量が最も高い性能を示しており、性質の異なる特徴量を統合することで、多くの話者において最適な性能を得ることができた。さらに(ⅲ)深層学習を適用した結果、73.66%の認識率が得られた。従前の特徴量と比べ大幅な性能改善に成功した。

②（ⅰ）CENSREC-1-AV を用い、音声と画像の統合について認識実験を行った。なお音声データには④と同じものを利用し、①と異なり口唇画像を利用した。結果を次図に示す。まず、Baseline(l)は MFCC、Tandem(l)は DBAF とおおむね同等の性能となった。Tandem(i)は Baseline(l)より高く、特徴抽出に深層学習を用いることの有効性が明らかとなった。さらに Hybrid の方が性能は高く、DNN-HMM は MMSR でも有用であることが判明した。他方で、これらは音声と画像の重み付けができないという欠点がある。実際、重み付け可能なマルチストリーム HMM を利用した Tandem(m)が最も性能が高くなり、MMSR における音声と画像の統合方法として最適であることがわかった。



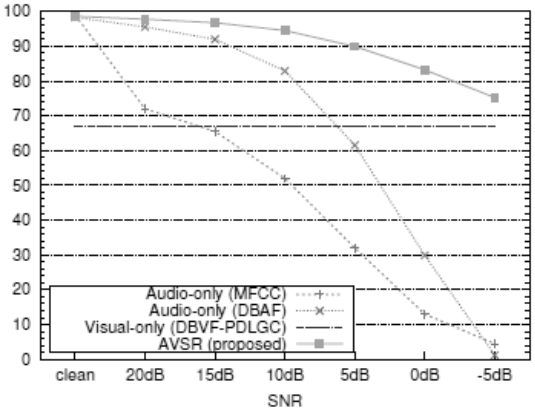
(ⅱ)異なる重み付けによるマルチストリーム GMM-HMM を確率的に統合する手法の評価を行った。上記を踏まえ、実験は後述の④と同じ条件で行った。その結果、事後的に最適な重みを設定する場合に匹敵する性能が得られ、手法の有効性が確かめられた。

③ モデル適応の実験では、音声については CENSREC-1-AV を利用した。ただし雑音はホワイトノイズとし、これを 3 種類の SNR で重畳した。画像は①(ⅱ)の特徴量を用いた。モデル適応は MLLR 法により話者ごとに行った。結果を次表に示す。1 回目の適応でも十分な改善がみられ、さらに2回目の適応でも性能が改善した。以上より、MMSR において適応を繰り返すことの有用性が確認できた。

	0dB	10dB	20dB
適応前	31.01	55.76	80.51
適応 1 回目	79.61	92.30	96.81
適応 2 回目	83.83	93.70	97.42

(2) 各種タスク・環境での MMSR の適用

④ ①による画像特徴量 DBVF と、②で利用した音声特徴量 DBAF、さらに②のマルチストリーム GMM-HMM を用いて認識実験を行った。結果を次図に示す。なお参考として MFCC による結果もあわせて示す。DBAF の MFCC に対する優位性、および①(iii)のとおり DBVF の有効性が判明した。加えて、これらとマルチストリーム GMM-HMM によって、全体平均で 90%と非常に高い認識性能を達成できた。



⑤ (ⅰ)大語彙タスクとして、音素バランス 503 文の音声と口唇動画像を収録した。収録条件は CENSREC-1-AV とほぼ同様であり、収録話者は 4 名、本実験では学習とテストは同一とした。雑音は付与していない。特徴量には MFCC および PCA を用いた。認識モデルとして triphone のマルチストリーム GMM-HMM を構築し利用した。実験を行ったところ、画像特徴のみで 20.06%、音声特徴のみで 76.40%、MMSR で 76.77%となり、音声のみと比較して 1.6%の誤り削減がみられたものの、改善の割合は小さかった。これは用いた画像特徴量や、データ不足による原因が考えられる。以上を踏まえ現在は、①にて得られた DBVF の利用と、追加データの収録を行い、引き続き大語彙タスクでの性能改善を図っているところである。(ⅱ)モバイル環境の実験は、学習データは CENSREC-1-AV、テストデータはアプリで収録した 16 名を用いた。収録は屋外、乗用車内、駅ホーム、夜間幹線道路など 8 環境で行った。特徴量は(ⅰ)と同様である。認識実験の結果、音声特徴のみで 48.72%、画像特徴のみは 16.44%となった。音声に関しては、モデル適応や深層学習などによる性能改善が必要である。画像ではこれらに加え、フレームレート低下による性能劣化がみられたため、これを補償する方法が必要であることが判明した。これらをにより性能改善したうえで、MMSR に取り組む予定である。

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 0 件）

〔学会発表〕（計 17 件）

[1] S. Tamura, T. Kawasaki, K. Miyazaki, K. Ukai and S. Hayamizu, "Visual speech recognition using optical and depth image features," Proc. FCV2016, Takayama (Japan), pp. 17-21 (2016-2-17).

[2] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda and S. Hayamizu, "Audio-visual speech recognition using deep bottleneck features and high-performance lipreading," Proc. APSIPA ASC 2015, Hong Kong (China), pp. 575-582 (2015-12-17).

[3] 田村哲嗣, 二宮宏史, 北岡教英, 大須賀晋, 入部百合絵, 武田一哉, 速水悟, 「深層学習によるボトルネック特徴量を用いたマルチモーダル音声認識」電子情報通信学会 技術研究報告, 神戸大学(兵庫県神戸市), SP2015-69, vol. 115, no. 253, pp. 57-62 (2015-10-16).

[4] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda and S. Hayamizu, "Audio-visual processing toward robust speech recognition in cars," Proc. DSP in Vehicle 2015, San Francisco (U.S.A.) (2015-10-14).

[5] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda and S. Hayamizu, "Investigation of DNN-based modeling for audio-visual speech recognition," Proc. MLSLP2015, Aizu (Japan), PaperID 1 (2015-9-19).

[6] 田村哲嗣, 二宮宏史, 北岡教英, 大須賀晋, 入部百合絵, 武田一哉, 速水悟, 「深層学習による音響・画像特徴量を用いたマルチモーダル音声認識」日本音響学会 2015 年秋季講演論文集, 会津大学(福島県会津若松市), 3-2-5, pp. 65-66 (2015-9-18).

[7] K. Ukai, S. Tamura and S. Hayamizu, "Stream weight estimation using higher order statistics in multi-modal speech recognition," Proc. FAAVSP2015, Vienna (Austria), PaperID 33 (2015-9-11).

[8] H. Ninomiya, N. Kitaoka, S. Tamura, Y. Iribe and K. Takeda, "Integration of deep bottleneck features for audio-visual speech recognition," Proc. INTERSPEECH2015, Dresden (Germany), pp. 563-567 (2015-9-7).

[9] S.Tamura, T.Seko and S.Hayamizu, ″Data collection for mobile audio-visual speech recognition in various environments,″ Proc. Oriental COCOSDA 2014, Phuket (Thailand), pp.134-139 (2014-9-11).
[10] 絹田卓也, 田村哲嗣, 速水悟,「マルチモーダル音声認識における音声と画像の協調によるモデル適応法の検討」第13回情報科学技術フォーラム FIT2014, 筑波大学(茨城県つくば市), E-022, no.2, pp.257-260 (2014-9-5).
[11] 田村哲嗣, 世古拓海, 速水悟,「マルチモーダル音声インターフェースの開発」電子情報通信学会 2014 年総合大会, 新潟大学(新潟県新潟市), D-14-4 (2014-3-19).
[12] 世古拓海, 河﨑卓也, 田村哲嗣, 速水悟,「実環境におけるマルチモーダル音声インターフェースの適用」電子情報通信学会 技術研究報告, 早稲田大学(東京都新宿区), PRMU2013-199, vol.113, no.493, pp.185-190 (2014-3-6).
[13] 鵜飼直弥, 田村哲嗣, 速水悟,「距離画像を用いたマルチモーダル音声認識」電子情報通信学会 技術研究報告, 早稲田大学(東京都新宿区), PRMU2013-198, vol.113, no.493, pp.179-184 (2014-3-6).
[14] T.Kawasaki, N.Ukai, T.Seko, S.Tamura and S.Hayamizu, ″Improvement of lip reading performance in real environments using speaker and environmental adaptation,″ Proc. ACPR2013, Okinawa (Japan), pp.346-350 (2013-11-6).
[15] T.Kawasaki, N.Ukai, T.Seko, S.Tamura, S.Hayamizu, C.Miyajima, N.Kitaoka and K.Takeda, ″An audio-visual in-car corpus ″CENSREC-2-AV″ for robust bimodal speech recognition,″ Proc. DSP in Vehicle 2013, Seoul (South Korea), PaperID 25 (2013-10-1).
[16] T.Seko, N.Ukai, S.Tamura and S.Hayamizu, ″Improvement of lipreading performance using discriminative feature and speaker adaptation,″ Proc. AVSP2013, Annecy (France), pp.221-226 (2013-8-31).
[17] P.Shen, S.Tamura and S.Hayamizu, ″Audio-visual interaction in sparse representation features for noise robust audio-visual speech recognition,″ Proc. AVSP2013, Annecy (France), pp.43-48 (2013-8-30).

〔図書〕(計1件)

[1] 田村哲嗣, ほか「音響キーワードブック」, 日本音響学会 編集, コロナ社 出版, 総494ページ, pp.406-407 (2016).

〔産業財産権〕
○出願状況 (計0件)

○取得状況 (計0件)

〔その他〕
ホームページ等
http://www.asr.info.gifu-u.ac.jp/

６．研究組織
(1)研究代表者
田村 哲嗣 (TAMURA Satoshi)
岐阜大学 工学部・助教
研究者番号：10402215

(2)研究分担者
なし

(3)連携研究者
なし