様式 C-19

# 科学研究費補助金研究成果報告書

平成 21 年 6 月 1 日現在

研究種目：若手研究(B)
研究期間：2006～2008
課題番号：18700175
研究課題名（和文）音声・画像のマルチモーダル情報協調・情報統合を用いた音声認識の高度化
研究課題名（英文）Improvement of audio-visual speech recognition using multi-modal cooperation and integration techniques
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

研究成果の概要：本研究では、音声と発声時の口唇動画像を用いたマルチモーダル音声認識において、音声と画像それぞれの情報を相互利用する情報協調手法、および、音声と画像の情報を効果的にまとめる情報統合方法に関するさまざまな検討を通じて、マルチモーダル音声認識の認識性能の向上を試みた。その結果、認識性能の向上を達成しただけでなく、情報協調や情報統合に関する多くの新しい知識を得ることができた。

交付額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2006 年度 | 2,400,000 | 0 | 2,400,000 |
| 2007 年度 | 600,000 | 0 | 600,000 |
| 2008 年度 | 500,000 | 150,000 | 650,000 |
| 年度 | | | |
| 年度 | | | |
| 総　計 | 3,500,000 | 150,000 | 3,650,000 |

研究分野：音声情報処理、マルチモーダル情報処理
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：マルチモーダル音声認識、情報統合、情報協調、マイクロフォンアレー、マルチモーダル VAD

## １．研究開始当初の背景

（1.～4.の各項目は、同じ括弧付きインデックスで対応付けられている）

(1) 従前の音声認識技術は、雑音下での認識性能低下の問題を抱えていた。この問題を解決する方法の一つとして、音声信号に加えて、発話時の口唇動画像を用いる、マルチモーダル音声認識が注目され、国内外で研究が行われるようになってきた。

(2) マルチモーダル音声認識では、音声から

得られる音響特徴量と、動画像から得られる画像特徴量を連結・統合し、音声と画像の重みづけが可能なマルチストリーム HMM（Hidden Markov Model）によって音声認識を行う初期統合法が、基本的手法として用いられてきた。これとは対照的に、音声、画像それぞれのモダリティで認識を行い、その出力を融合する結果統合法が存在する。後者の研究事例は世界的にも希少であったが、より高性能な音声認識のためには、これら「情報統合」手法の調査･検討が不可欠である。

(3) 従来のマルチモーダル音声認識では、カメラから得られる 2 次元画像情報を基に画像特徴量を算出していた。一方で、複数のカメラを用いるなどして口腔内や頬周辺の筋肉の動きを取得し発話に関する情報を計算したり、ないしは画像のキャプチャの方法（カメラの個数、入力画像の撮影条件など）によらない画像特徴量を新たに規定したりすることは、認識性能向上のみならず、実システムの構築にも有利であると考えられる。

(4) また、音声を受音する段階で、マイクロフォンアレーを用いて、雑音を低減させ音声認識性能を向上させる方法が古くより行われてきた。雑音除去には音源方向（音声到来方向）情報を用いることが効果的であるが、雑音が重畳した受音データから音源方向を推定することは困難であった。ここで、画像情報を用いて音源方向を推定しその結果を用いて雑音低減を行うなど、音声と画像の「情報協調」手法が利用できるのであれば、マイクロフォンアレーによる性能改善がより期待できると考えられる。

## 2．研究の目的

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、情報協調および情報統合に関する以下の検討を行い、もって音声認識性能の向上、さらには視覚と聴覚の情報処理手法の工学的を目指す。

(1) マルチモーダル音声認識の情報統合の一手法として、結果統合による手法の構築と評価を行う。

(2) 情報協調で得られる画像情報を音声認識で有効に利用するため、汎用性の高い画像特徴量・モデリングの開発を行う。

(3) 情報協調システムの一環として、画像側で得られる音声到来方向の情報を用いたマイクロフォンアレーの雑音抑制手法を検討する。

(4) 上記の成果を用いたリアルタイム・マルチモーダル音声認識の実現に向けた調査を行う。

(5) 以上で得られた情報統合・情報協調の知見を他分野に適用し、本研究の成果の普遍性や発展性について議論する。

## 3．研究の方法

(1) マルチモーダル音声認識の情報統合の一手法として、結果統合による認識手法を検討する。具体的には、音声と画像の認識結果を各々コンフュージョンネットワーク（Confusion Network、CN）の形式で求め、これを統合する Confusion Network Combination（CNC）の手法を構築し、あわせて認識性能を調査した。

(2) 以下に列挙する画像の取得方式や形式によらない画像特徴量および画像のモデリングの検討を行った。
   ① アクションユニット（Action Unit、AU）と呼ばれる、口の動きを表したモデルにより認識を行う手法
   ② 画像から直接得られる特徴量（一次特徴量）を用いて、各音素モデルの出力尤度を求め、正規化したうえで最終的な画像特徴量（二次特徴量）として用いる手法

(3) マイクロフォンアレーを構築し、同時に、取得する画像情報から音声の到来方向を推定することで得られる効果、および、認識性能の改善について検討した。

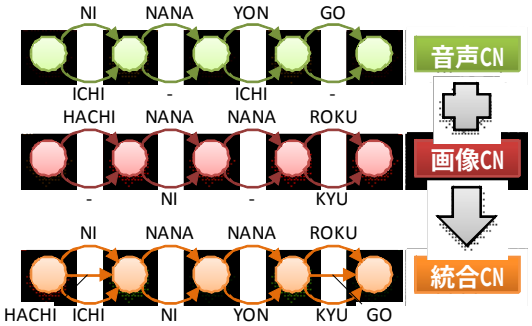
(4) 上記に関連して、マルチモーダル音声認識システムを構築するため、以下の調査およびデータベース作成を行った。
   ① リアルタイム・マルチモーダル音声認識における計算量削減の手法
   ② 音声と画像の同期、および、双方のフレームレートが認識率におよぼす影響
   ③ マルチモーダル音声認識の評価基盤構築のためのコーパスの作成

(5) 本研究の主たる成果である「情報協調」「情報統合」の別分野への適用として、マルチモーダル音声区間検出（Voice Activity Detection、VAD）の検討・実験を行った。

## 4．研究成果

(1) CNC では、はじめに統合元の CN で対応するノードの組を、認識結果の時間情報を用いて求め、それを基にノードを統合する。次に元のノードに付与されていたアーク（認識単語が付与されている）について、信頼度スコアを基準に選択を行い、これを統合後のノードに付与する方法で最終的な統合後の CN を得る（次頁の図）。認識実験によって性能を評価したところ、SNR5dB 白色雑音重畳条件で、

音声単独（49.2%）、画像単独（48.5%）に比べ、CNC による結果統合を行うことで、約 5%認識率が向上した（54.1%）。さらに得られた CN のアークを調べたところ、潜在的な認識率は高いことが判明し、結果統合による性能改善の可能性が認められた。一方、CNC において有益なノードやアークが選択できていないことも明らかとなり、さらなる手法の改善が必要であることが判明した（論文[8,9]）。背景で述べたとおり、結果統合法によるマルチモーダル音声認識の例は世界的に少なく、本研究はその先駆的試みとして注目される。現在は、上で述べた CNC の問題を解決する方法や、初期統合法と組み合わせた手法の検討を行っている。既存の初期統合法の利点を生かしつつ、




CN の持てる性能を十分に発揮することができれば、飛躍的な認識性能の改善が見込まれる。

(2) ① 画像のモデルの単位に、顔面の動きを記述する AU のうち、発話に関する 5 つを用いた手法の検討を行った。その結果、AU 自体の認識精度として、音響雑音がない条件で 58.6%、実環境で 50.3%が得られた。また、マルチモーダル音声認識の精度としては、音響情報のみと比べ、初期統合で約 1%性能が向上した。AU の認識が十分でなく、加えて、ひとつの単語に複数の AU 系列が対応するため、性能の向上が限定的であったと結論づけられる（論文[1]）。

② 母音ごとに画像 GMM（Gaussian Mixture Model）を構築しておき、それぞれの出力尤度をベクトル化し、特異値分解を用いて直交化・正規化を行って得られたベクトルを新たな画像特徴量として用いる実験を行った。認識データに白色雑音を重畳したところ、従来の特徴量と比較して、最大で約 7～8%程度性能が改善した（論文[3,7]）。本手法は従来の 1 カメラによる場合も、複数カメラで得られる立体情報を用いる場合でも、尤度に変換することで同様に扱うことができるというメリットがある。

(3) 全方向ステレオカメラ（Stereo Omni-directional Camera）とマイクロフォンアレーを用いた会議記録システムでは、画像認識により話者の方向を推定でき、これを音声到来方向として用いることで、雑音を抑制し、結果として音声認識率を向上させることが可能である（論文[11]）。そこで本研究では、音声到来方向が正しく推定できたという仮定のもとで、新たな雑音抑制手法と認識率の改善を試みた。既存法である CSCC（Complex Spectrum Circle Centroid）法をベースに、音声スペクトルの計算手法を改良したところ、シミュレーション環境において、CSCC 法と比べ NNR（Noise Reduction Ratio）で約 1dB 音質が改善し、音声認識では約 9%誤りを削減した（論文[6]）。この成果を情報協調手法に組み込むことが今後の課題であり、計算量削減など他の課題とあわせて取り組んでいきたい。

(4) ① リアルタイムシステムを念頭に、雑音変動を考慮したマルチストリーム HMM の重みづけ最適化手法を提案した。一定発話ごとにそれまでの認識出力を用いて逐次重みを更新する方式では、重み最適化を行わない場合に比べ約 7%、全体でのみ最適化を行う場合と比べ 0.4%性能が改善した。並行して、計算時間を要する画像特徴量の計算時間削減を試み、約 50fps のスペックを達成した。後述するように、この速度は性能的には十分である（論文[10]）。リアルタイム・マルチモーダル音声認識の実現例は世界的にも報告がなく、近年の PC や携帯電話へのカメラ内蔵の流れを鑑みると、本研究の継続課題である、リアルタイム音声認識システムの構築は大きな意義があるといえる。今後も引き続きシステムの開発を行っていく。

② 音声と画像は別々に取得されるため、統合手法に関わらず、両者の情報統合時には、音声と画像の時間ずれが生じ得る。これが認識性能に与える影響を調査したところ、およそ 100ms 程度、画像が音声に先行した場合に認識率が最大となった。これは口の動きが音声の生成に先行するためと考えられ、興味深い結果となった。さらに先述の実時間システムを念頭に、音声と画像のフレームレートと認識率の関係を調査した。その結果、著しく低いレートでなければ影響は限定的であり、おおよそ 7.5～10fps 以上あれば問題ないことが判明した（論文[4]）。本成果では、特に音声と画像の時間ずれに関して特筆すべき成果が得られ、情報統合システムを構築するうえで重要な知見が得られた。

③ 本研究に関連して、現在、マルチモーダル音声認識の共通評価基盤（コーパスや認識結果のベースラインなど）の構築を行っている（論文[5]）。認識手法同士の比較評価を行うにはそれ以外の環境を統一する必要があり、通常の音声認識では共通評価基盤は広く利用されている。一方、マルチモーダル音声認識に関しては、評価基盤は世界的にも存在しない。ゆえに、このような取り組みは十分に意義がある。のみならず、本研究の成果を統一的に評価するためにも、評価基盤は必要である。今後は本年中の公開を目指して、データベースの整備およびベースラインの作成を行っていく。

(5) 本研究で得られた情報協調・情報統合に関する成果の他分野への適用を試みた。音声と動画像を用いるマルチモーダル VAD において、モデルの有無、統合方法など、さまざまな条件で識別実験を行った（論文[1,2]）。初期統合と結果統合の比較では、雑音レベルの低い環境では、初期統合（モデル有）が高い性能を示した。一方、雑音レベルが高い状況ではモデル無の結果が比較的良好な性能を示した。このとき、初期統合と結果統合の性能差はあまり見られなかった。これらの結果は、現在までに得られているマルチモーダル音声認識での傾向と類似していた。すなわち、理想環境に近い状況では初期統合が勝るが、音響的・映像的に実環境に近づくにつれて両者の差は小さくなる現象がみられ、さらに環境悪化した状況では結果統合が優位になると推測される。また、初期統合（モデル有）は音声区間の取りこぼしが最も少なく、結果統合（モデル無）は非音声区間同定に優れていた。これらから、それぞれの統合方法が得意とする識別を組み合わせることで VAD の性能を大きく改善できることが推測され、同様にマルチモーダル音声認識でも検討する必要があると判明した。まとめると、上記の結果は、本研究で得られた成果の普遍性を示唆しているのみならず、VAD への適用で得られた新たな成果を音声認識にフィードバックし、性能向上させることが可能であることも暗示している。このような成果は他に類を見ないものであると確信している。今後の展開としては、以上の知見を用いて、マルチモーダル音声認識や VAD を発展させる研究を行うとともに、音声認識と VAD を組み合わせ、より高度なシステムとすることを目指していく。

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 12 件）


[1] 竹内伸一,羽柴隆志,田村哲嗣,速水悟
「実環境における口唇動画像を用いたマルチモーダル音声区間検出」
日本音響学会 2009 年春季講演論文集,3-5-8, pp.119-120,2009 年 3 月 19 日.

[2] 羽柴隆志,竹内伸一,田村哲嗣,速水悟
「マルチストリーム HMM を用いた音声と画像による音声区間検出」
日本音響学会 2009 年春季講演論文集,1-P-5, pp.131-132,2009 年 3 月 17 日.

[3] 石川雅人,田村哲嗣,速水悟,
「画像特徴量の正規化によるマルチモーダル音声認識の改善」
電子情報通信学会技術研究報告,SP2008-71, vol.108,no.312,pp.7-12,2008 年 11 月 20 日.

[4] 田村哲嗣,石川雅人,速水悟,
「マルチモーダル音声認識における音声と画像の同期に関する調査」
電子情報通信学会技術研究報告,SP2008-70, vol.108,no.312,pp.1-6,2008 年 11 月 20 日.

[5] 田村哲嗣,宮島千代美,北岡教英,速水悟,武田一哉,
”CENSREC-AV: Evaluation frameworks for audio-visual speech recognition,”
Proc.AVSP2008,Morton,Australia,pp.51-54, 2008 年 9 月 27 日.

[6] 菱川恵利子,田村哲嗣,速水悟,
「マイクロフォンアレイによる目的信号スペクトル抽出法の検討」
日本音響学会 2008 年秋季講演論文集,2-8-15, pp.665-666,2008 年 9 月 11 日

[7] 石川雅人,田村哲嗣,速水悟,
「画像 HMM による尤度情報を利用したマルチモーダル音声認識の検討」
日本音響学会 2008 年秋季講演論文集,1-1-24, pp.57-58,2008 年 9 月 10 日

[8] 上澤泰,石川雅人,田村哲嗣,速水悟,
「音声と画像の confusion network 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,SP2007-92, vol.107,no.356,pp.37-42,2007 年 11 月 28 日.

[9] 上澤泰,石川雅人,田村哲嗣,速水悟,
「音声と画像の CNC によるマルチモーダル音

声認識の検討」
日本音響学会 2007 年秋季講演論文集,2-8-2,
pp.111-112,2007 年 9 月 20 日.

[10] 田村哲嗣,速水悟,
「リアルタイムマルチモーダル音声認識の構築に関する検討」
日本音響学会 2007 年春季講演論文集,2-9-14,
pp.63-64,2007 年 3 月 14 日.

[11] 木村文彦,近藤功一,田村哲嗣,速水悟,山本和彦,
「SOS とマイクロフォンアレイの統合による会議記録システムの開発」
情報処理学会研究報告,2006-SLP-63-2,
vol.2006,no.107,pp.7-12,2006 年 10 月 20 日.

[12] 上澤泰,田村哲嗣,速水悟,
「マルチモーダル音声認識のためのアクションユニットによる画像情報の改善」
日本音響学会 2006 年秋季講演論文集,1-2-25,
pp.49-50,2006 年 9 月 13 日.

〔図書〕（計 0 件）

〔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://hym.info.gifu-u.ac.jp/~tamura/multimodal.html

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