様式 C-19

科学研究費補助金研究成果報告書

平成２３年　６月　１日現在

機関番号：１２６０８
研究種目：基盤研究（Ｂ）
研究期間：２００８～２０１０
課題番号：２０３０００６３
研究課題名（和文）　ヒューマンコミュニケーション検索・要約のためのマルチモーダル認識の研究
研究課題名（英文）　A study of multimodal recognition for human communication search and summarization
研究代表者
篠田　浩一（SHINODA　KOICHI）
東京工業大学・大学院情報理工学研究科・准教授
研究者番号：10343097

研究成果の概要（和文）：
ヒューマンコミュニケーション理解のために、音声・動画から構成されるマルチメディアデータに対するマルチモーダルパターン認識技術を開発した。まず映像におけるイベントの抽出では混合ガウス分布とサポートベクターマシンを用いた統計的手法を提案し、世界40機関が参加して開催されたTRECVID2010ワークショップで世界4位(日本からの参加者中では1位)の成果を得た。また、音声モデルの能動学習・能動適応、耐雑音音声認識、ミーティング音声認識のための信号処理、マルチモーダル認識アルゴリズム、話者認識・ジェスチャー認識、発話スタイル解析、映像要約の手法をそれぞれ開発した。

研究成果の概要（英文）：
We developed multimodal pattern recognition techniques for human communication using speech and video. We proposed a statistical technique using Gaussian mixture models and support vector machines for event extraction. We participated in TRECVID2010 workshop, where our method achieved the 4-th performance among 40 participants from all over the world. We also developed new methods for active learning for speech modeling and adaptation, noise robust speech recognition, signal processing for meeting speech recognition, multimodal pattern recognition, speaker/gesture recognition, speech style analysis and video summarization.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	5,900,000	1,770,000	7,670,000
２００９年度	4,300,000	1,290,000	5,590,000
２０１０年度	3,600,000	1,080,000	4,680,000
年度			
年度			
総　計	13,800,000	4,140,000	17,940,000

研究分野：総合領域
科研費の分科・細目：情報学　・　知覚情報処理・ロボティックス
キーワード：音声認識、動画像認識、マルチモーダル認識、ヒューマンコミュニケーション理解、情報検索

１．研究開始当初の背景

近年、インターネット上のテキストデータからのマイニング技術が急速に進展し、人間の知的生産活動の支援に大いに役立っている。そのおかげで、人間の知的生産技術自体が劇的に変化している。しかし、組織内(イントラネット)・家庭内情報のマイニングについては、まだ十分に検討されていない。人間の

知的生産活動ではインターネット上の情報のみならず、このような、小規模なコミュニティで対面コミュニケーションにより得られる情報が重要である。

このようなヒューマンコミュニケーションは、2 つのモード、言語モードと非言語モードに分けられる。前者は、音声の書き起こしに対応する言語メッセージを伝搬するもので、音声認識技術の対象となるものである。後者は、それ以外の感情などの非言語メッセージを伝搬するもので、さらに、顔の表情、ジェスチャー(手振り、身振り)、音声中のパラ言語(抑揚、発声速度など)などのいくつかのサブモードに分かれる。過去の研究により、ヒューマンコミュニケーションではこれらの２つのモードが同程度に重要であることがわかっている。

言語モードの自動インデクシング技術は音声認識技術である。現在、読み上げ調の発声で 95%, 講演などのモノローグで 80%の認識率を達成しており、マイニングの用途に十分な性能をもつ。しかし、対話音声に対しては 60%以下の認識性能にとどまっており、未だ実用には不十分である。一方、非言語モードの自動インデクシングにも課題が多く、それぞれ単独での実用化は困難である。

２．研究の目的

現在、音声や映像に対して統計的パターン認識や協調的タグ付けを用いた自動インデクシング(タグ付け)技術の研究開発が進んでいる。この技術を応用し、ヒューマンコミュニケーションから有用な情報を自動抽出して、知的生産活動の支援に役立てることは重要な将来目標である。

ヒューマンコミュニケーションは本質的にマルチモーダルであり、言語モードと非言語モードを同時に密接に関連づけながら用いていることで情報を交換している。自動インデクシングにおいても、言語モード/非言語モードを組み合わせて用いることで、それぞれを単独で用いる場合より認識性能が向上するのではという期待がある。本研究期間内では、言語・非言語情報に対するマルチモーダル認識によるヒューマンコミュニケーションを理解する技術を開発することを目的とした。

３．研究の方法

本研究では、音声と動画によって与えられる情報を対象とし、言語としては日本語を用いた。まず、既存の音声データベースを用い、音声情報中の言語モードと非言語モード(パラ言語)について、それぞれの要素技術を開発した。言語モードにおいては、音声認識のモデル学習技術、耐雑音技術に重点をおいて研究開発を行った。また、人間の動作のデータベースを構築し、そこからの非言語情報を取得する技術を開発した。これらと並行して、マルチメディア会議データベースを収録し、そこからの音声・動作認識とその周辺技術の開発を行った。これらの要素技術を統合し、マルチモーダル認識技術を開発した。

４．研究成果

(1)映像からのマルチモーダル特徴検出

映像(音声と動画)から、特徴的なイベント・シーンの検出を行う手法を開発した。この手法では、まず、予め大量のタグ付きのビデオデータを用いて、検出対象ごとに混合ガウス分布モデル（検出モデル）を構築しておく。そして、入力された映像に対し、やはり混合ガウス分布モデルを作成し、それと検出モデルとの距離を用いて検出を行う。検出には、混合ガウス分布モデルから作成したガウス・スーパーベクトルを入力とするサポートベクターマシンを用いた。この手法は、特に、動画情報と音声情報を組み合わせてモデル化・検出を行う点に特色がある。2009 年にアメリカ国立標準技術研究所(NIST)主催の映像検索・評価ワークショップ TRECVID に参加し、20 種類の物体・イベント・シーンなどの検出で、参加 40 期間中 4 位、日本からの参加機関中では 1 位の検出性能となった。この手法は、「歌っている」、「話している」など、ヒューマンコミュニケーションに関連したイベントについての検出率が高く、特にマルチモーダルなヒューマンコミュニケーション理解のためのイベント・シーン検出手法として、効果的であることを示した。

(2)音声モデルの能動学習・能動適応

ヒューマンコミュニケーションにおける言語情報取得のために、音声モデルの高性能化を図った。音声モデルの学習にはその書き起こしの情報が付与された大量の音声データが必要であり、その作成には多くのコストがかかる。なるべく少ないコストで高性能な音声モデルを構築したい。この目的のために音声データから学習効果が高いと予想される音声を選択して学習する能動学習を行った。書き起こしのない音声データから書き起こすべきデータを選択する手法、学習に効果的と予想されるテキストを生成、あるいは、選択して、それを読み上げた音声データを収集する方法、の 2 通りの方法を試みた。また、後者については、すでに存在する不特定話者の音声モデルを特定の話者の音声に適応させる、話者適応の枠組みにも適用した。前者については、選択方法として、複数の認識器で構成されるコミッティにおいて多数決をとる手法を用い、従来の半分の量の学習データで、同等の性能をもつ音声モデルを構築可

能であることを示した。後者については、音素誤り率の高い音素を多く含んだテキストを選択/生成する方法を用い、話者適応において、同等の性能を得るのに必要なテキストの量を25%削減できることを示した。

(3)耐雑音音声認識
ヒューマンコミュニケーションにおける音声には、周囲雑音が多く含まれており、その認識のためにはその影響を軽減する必要がある。この研究期間内には2つの方法を開発した。まず、従来の音声認識では音声の基本周波数(F0)情報は通常利用されていなかったが、雑音下ではその利用が効果的であることが期待できる。そこで、音声のスペクトログラム(時間-周波数空間におけるパワー分布)に対するハフ変換により基本周波数を抽出し、それを従来の音韻特徴量と組み合わせて用いることにより、音声認識率の向上を図った。大語彙音声認識で評価を行い、10dBのSN比の雑音下で2.6ポイントの性能向上を得た。また、雑音下で、特徴量空間における音声特徴量空間の縮小(スペクトル縮小)が起きることを発見し、その縮小の度合いに合わせて、モデル空間の縮小を行う手法を開発した。これにより最大でやはり2.6ポイントの性能向上を得た。

(4)ミーティング音声認識のための信号処理
ヒューマンコミュニケーション、特に、ミーティングに対する音声認識においては、話者の装着したマイクに別の話者の音声が重畳され、そのために認識性能が劣化する。このような回り込みの影響を低減するために、従来、音声から雑音を取り除く技術として用いられてきたスペクトルサブトラクション処理を用いる方法を開発した。マイク間の伝達関数の推定とスペクトルサブトラクション処理とを繰り返すことにより、より正確に他者の音声を差し引くことが可能になる。会議音声にノイズを人為的に重畳して作成した疑似的な雑音下音声データを用いて評価を行い、音声認識率が 66.5%から 77.7%へと顕著に改善した。

(5)マルチモーダル認識アルゴリズム
ヒューマンコミュニケーションにおける言語情報と非言語情報は一般に同期していない。我々は、音声とジェスチャーの同時入力における、準同期的なマルチモーダル・アルゴリズムの高度化を行い、そのヒューマンコミュニケーションへの応用を検討した。発話者への適応手法などを開発した。

(6)話者認識
複数人の音声コミュニケーションの理解のためには、周辺技術として発話者を同定する話者認識の技術が重要となる。従来音声認識で効果的であることが知られている構造的事後確率最大化法(SMAP)を話者認識に適用することによりその性能改善を試みた。SMAPを用いてガウス混合分布を適応し、そのガウス・スーパーベクトルを入力としたサポートベクターマシンを用いる方法、SMAPにおけるガウス分布木構造を複数用意し、それらの間の投票によって検出を行う方法、の2つの方法を開発し、いずれも話者認識性能の顕著な向上を得た。

(7)ジェスチャー認識
ヒューマンコミュニケーション理解の要素技術として、人間の動作の認識に関する各種研究を行った。まず、対象への距離を測ることも可能な ToF(Time-of-Flight)カメラを用いて収録した3次元データを用いた、人間の身振りの認識手法を検討した。隠れマルコフモデルなどの確率統計手法を用いた。手話の認識を用いて評価を行い、従来の 2D 画像の認識に比べ顕著な性能改善を得た。なお、これに関連し、多数の人物が存在する映像から特定の動作を検出する手法、歩行動作から人物の同定を行う手法などについて研究開発を行い、一定の成果を得ている。

(8)映像要約
ヒューマンコミュニケーションの要約を行うための要素技術として、映像要約の研究を行った。この研究では、映像全体を類似した音声・動画特徴をもついくつかのクラスタに分類し、そのクラスタの代表シーンをつなげることで映像要約を作成する。ヒューマンコミュニケーションの要約の前処理として利用が可能である。

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4件）

1. Koichi Shinoda、Acoustic Model Adaptation for Speech Recognition、IEICE Transactions on Information and Systems、Vol. E93-D、No. 9、pp. 2348-2362、2010、査読有
2. 井上中順、斉藤辰彦、篠田浩一、古井貞熙、大規模映像資源のためのマルチモーダル高次特徴検出、電子情報通信学会論文誌、Vol. J93-D, No. 12、pp. 2633-2644、2010、査読有

3. Koichi Shinoda、Yasushi Watanabe、Kenji Iwata、Yuan Liang、Ryuta Nakagawa、Sadaoki Furui、Semi-synchronous speech and pen input for mobile user interfaces、Speech communication、Vol. 53、pp. 283-291、2010、査読有
4. Nazrul Effendy、Koichi Shinoda、Sadaoki Furui、Somchai Jitapunkul、Automatic recognition of Indonesian declarative questions and statements using polynomial coefficients of the pitch contours、2009 The Acoustical Society of Japan、 Accoust. Sci. & Tech.、No. 30、pp. 249-256、2009、査読有

〔学会発表〕(計 42 件)

1. 村上博子、篠田浩一、古井貞熙、音響モデル学習のための相対エントロピーを用いた学習文選択手法、日本音響学会 2011 年春季講演発表会、2011 年 3 月 9 日、東京
2. Sangeeta Biswas、Marc Ferras、Koichi Shinoda、Sadaoki Furui、Voting Approach in SMAP Adaptation for Speaker Verification、日本音響学会 2011 年春季研究発表会、2011 年 3 月 9 日、東京
3. 別府真由美、篠田浩一、古井貞熙、雑音下音声におけるスペクトル縮小の分析とその対雑音音声認識への利用、電子情報通信学会 SP 研究会、2011 年 3 月 4 日、東京
4. 井上中順、上嶋勇祐、篠田浩一、マルチモーダル・マルチフレームな手法を用いた TTECVID セマンティックインデクシング、電子情報通信学会 PRMU 研究会、2011 年 2 月 17 日、さいたま市
5. 村上博子、篠田浩一、古井貞熙、音響モデル学習のための相対エントロピーを用いた学習文選択、情報処理学会　音声言語情報処理学会、2011 年 2 月 4 日、福山市
6. Marc Ferras、Koichi Shinoda、 Sadaoki Furui 、 Inter-speaker weighted MAP adaptation for GMM-supervector speaker recognition、情報処理学会　音声言語情報処理学会、2010 年 12 月 20 日、東京
7. Sangeeta Biswas、Marc Ferras、Koichi Shinoda、Sadaoki Frui、Optimal use of trees in structural MAP adaptation for speaker verification、報処理学会　音声言語情報処理学会、2010 年 12 月 20 日、東京
8. Nakamasa Inoue、Toshiya Wada、Yusuke Kamishima、Koichi Shinoda、Ilseo Kim、Byungki Byun and Chin-Hui Lee、TT+GT at TRECVID 2010 Workshop、TRECVID 2010 workshop 、 2010 年 11 月 15 日 、 Gaithersburg
9. Muhammad Rasyid Aqmar、Koichi Shinoda、Sadaoki Furui 、 Gait-based Person Identification Robust against Speed Variation using CHLAC features and HMMs、電子情報通信学会　PRMU 研究会、2010 年 10 月 8 日、千葉市
10. 那須悠、篠田浩一、古井貞熙、会議音声認識のためのスペクトル減算に基づく音源分離、日本音響学会 2010 年秋季研究発表会、2010 年 9 月 14 日、大阪
11. 井上中順、上嶋勇祐、篠田浩一、古井貞熙、SIFT 混合ガウス分布を用いた一般物体認識のためのマルチカーネル学習、電子情報通信学会　PRMU 研究会、2010 年 9 月 5 日、福岡市
12. Muhammad Rasyid Aqmar、Koichi Shinoda、Sadaoki Furui、Robust Gait Recognition

against Speed Variation、ICPR2010、2010年8月23日、Istanbul

13. Nakamasa Inoue、Tatsuhiko Saito、Koichi Shinoda、Sadaoki Furui、High-Level Feature Extraction Using SIFT GMMs and Audio Models、ICPR2010、2010年8月23日、Istanbul
14. 佐藤新、篠田浩一、古井貞熙、ToF カメラによる3D手話認識、画像の認識・理解シンポジウム 、2010年7月27日、釧路
15. Marc Ferras、Sangeeta Biswas、Koichi Shinoda、Sadaoki Furui、NIST SRE 2010:Tokyo Tech Speaker Recognition、NIST 2010 Speaker recognition evaluation workshop、2010年6月24日、Brno
16. 那須悠、篠田浩一、古井貞熙、会議音声認識のためのスペクトル減算に基づくオンライン音源分離、電子情報通信学会SP研究会、2010年5月26日、神戸市
17. Yuzo Hamanaka、Koichi Shinoda、Sadaoki Furui、Tadashi Emori、Takafumi Koshinaka、Speech Modeling Based on Committee-Based Active Learning、ICASSP2010、2010年3月14日、Dallas, U.S.A
18. 斉藤辰彦、井上中順、篠田浩一、古井貞熙、音響特徴を用いた映像からのイベント検出の研究、日本音響学会2010年春季研究発表会、2010年3月8日、東京
19. 濱中悠三、江森 正、越中 孝文、篠田浩一、古井貞熙、音声認識のための複数の認識器を利用した能動学習、情報処理学会 音声言語情報処理学会、2009 年 12月21日、東京
20. 井上中順、斉藤 辰彦、篠田浩一、古井貞熙、SIFT混合ガウス分布と音響特徴を用いた映像からの高次特徴検出、電子情報通信学会 PRMU研究会、2009年11月26日、金沢市
21. Nakamasa Inoue、Shanshan Hao、Tatsuhiko Saito、Koichi Shinoda、Ilseo Kim、Chin-Hui LeeiTITGT at TRECVID 2009 Workshop、TRECVID Workshop (TRECVID 2009)、2009年11月16日、Gaithersburg
22. Hideki Yasui、Koichi Shinoda、Sadaoki Furui、Koji Iwano、Noise robust speech recognition using spectral subtraction and F0 information extracted by Hough transform 、Asia-Pacific Signal and Information Processing Association 2009 Annual Summit and Conference、2009年10月5日、Sapporo、 Japan
23. Hiroko Murakami、 Koichi Shinoda、Sadaoki Furui、Speaker Adaptation Based on Two-Step Active Learning、INTERSPEECH 2009 BRIGHTON、2009 年 9月6日、Brighton UK
24. 濱中悠三、江森 正、越仲孝文、篠田浩一、古井貞熙、音声認識のためのコミッティを用いた能動学習、日本音響学会秋季研究発表会 、2009年9月15日、郡山市
25. 安井 英己、篠田 浩一、古井 貞熙、岩野公司、ハフ変換による基本周波数情報を用いた耐雑音音声認識の高性能化の検討、日本音響学会 2009 年 春季研究発表会、2009年3月17日、東京
26. 村上 博子、篠田 浩一、古井 貞熙、能動的な適応文選択に基づく話者適応化、日本音響学会 2009 年 春季研究発表会、2009年3月17日、東京
27. 山崎航史、篠田 浩一、古井 貞熙、統計的モデル選択によるシーン数の自動推定を用いた動画要約、電子情報通信学会 技

術研究報告、2009 年 2 月 19 日、東京
28. M.-R. Aqmar、K. Shinoda and S. Furui、Gait Recognition Using CHLAC Features and Hidden Markov Models、電子情報通信学会 技術研究報告、2009 年 2 月 19 日、東京
29. 安井 英己、篠田 浩一、古井 貞熙、岩野公司、耐雑音音声認識のためハフ変換による基本周波数情報抽出の高速化、電子情報通信学会 技術研究報告、2009 年 1 月 12 日、奈良
30. S. Hao、 Y. Yoshizawa、K. Yamasaki、 K. Shinoda and S. Furui、Tokyo Tech at TRECVID 2008、TRECVID 2008 workshop、2008 年 11 月 17 日、 Washington D.C.、USA
31. Koji Yamasaki 、 Kochi Shinoda and S. Furui 、 Automatically Estimating Number of Scenes for Rushes Summarization、In Proceedings of the TRECVID BBC Rushes Summarization Workshop (TVS 2008)、2008 年 10 月 31 日、ACM Multimedia、 New York、 USA
32. Yasushi Watanabe、Koichi Shinoda and Sadaoki Furui、Time-lag Adaptation for Semi-synchronous Speech and Pen Input、INTERSPEECH 2008、2008 年 9 月 22 日、Brisbane、Australia
33. 安井 英己、岩野 公司、篠田 浩一、古井 貞熙、スペクトルサブトラクションとハフ変換による基本周波数情報を用いた耐雑音音声認識、日本音響学会、2008 年 9 月 10 日、 九州

〔その他〕
ホームページ等
http://www.ks.cs.titech.ac.jp

６．研究組織
(1)研究代表者
篠田　浩一（SHINODA　KOICHI）
研究者番号：１０３４３０９７
東京工業大学・大学院情報理工学研究科・准教授

(2)研究分担者
古井　貞熙（FURUI　SADAOKI）
東京工業大学・大学院・情報理工学研究科・教授
研究者番号：９０２９３０７６