「共生社会に向けた人間調和型情報技術の構築」
平成21年度採択研究代表者

H22 年度
実績報告

河原　達也

京都大学・学術情報メディアセンター・教授

マルチモーダルな場の認識に基づくセミナー・会議の多層的支援環境

§1. 研究実施の概要

本研究では、人間の知的活動の源泉ともいえる音声コミュニケーションをマルチモーダルな観点で分析・モデル化した上で、セミナー・ポスター発表及び会議を対象として、リアルタイムに支援したり、効果的なアーカイブ化を行うための情報環境を構築する。主な話者の発話内容を音声認識して言語解析を試みるという従来のアプローチ（コンテンツに基づく処理）だけでなく、視線やあいづち・うなずきなどの聴衆の反応に着目した新たなアプローチ（インタラクションに基づく処理）を導入する。

第2年次の平成 22 年度では、ポスターボードとセミナー室を想定した共通的な基盤（プラットフォーム）について設計を行い、マルチモーダルなデータの収集・アノテーションを進めながら、上記の目標を実現するための要素技術に関する研究を本格的に開始した。具体的には、知能化ポスターボードを想定したカメラ群とマイクロフォンアレイを設計・試作し、参加者の発話を検出・強調したり、行動を検出・分類する方法を研究した。また、あいづちのパターンなどから、聴衆が関心を持った箇所を推定し、ポスター会話において重要な区間（ホットスポット）を抽出できることを明らかにした。さらに、衆議院の新会議録作成システムの試験運用において音声認識の評価を行った。

§2. 研究実施体制

(1)京大グループ

①研究分担グループ長：河原達也　（京都大学・学術情報メディアセンター・教授）

②研究項目

マルチモーダルな場の認識に基づくセミナー・会議の多層的支援環境

具体的には、音声認識、映像処理、反応の認識、質問応答・情報推薦

(2)奈良先端大グループ

①研究分担グループ長：鹿野清宏　（奈良先端科学技術大学院大学・情報科学研究科・教授）

②研究項目

セミナー・会議のための音響・音声処理

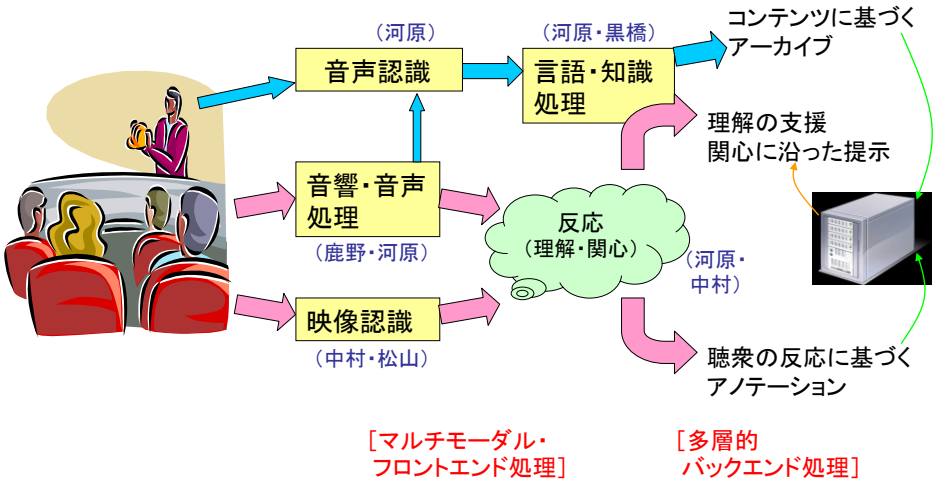
§3. 研究実施内容

（文中の文献番号は（4－1）に対応）

本研究では、人間の知的活動の源泉ともいえる音声コミュニケーションをマルチモーダルな観点で分析・モデル化した上で、セミナー・ポスター発表及び会議を対象として、リアルタイムに支援したり、効果的なアーカイブ化を行うための情報環境を構築する。主な話者の発話内容を音声認識して言語解析を試みるという従来のアプローチ（コンテンツに基づく処理）だけでなく、視線やあいづち・うなずきなどの聴衆の反応に着目した新たなアプローチ（インタラクションに基づく処理）を導入する。知能化したセミナー室やポスターボードを構築し、定例のセミナーやポスター発表会で実証実験を行う。さらに、音声認識システムについては、衆議院の次期会議録作成システムで運用して得られる大規模なデータ・知見をフィードバックすることで、音響モデル・言語モデル（辞書）の高精度化を行い、幅広い話し言葉音声の書き起こしに供することができるようにする。

本研究の概要を図1に示す。

図1： 本研究の処理の概要



第2年次の平成 22 年度では、ポスターボードとセミナー室を想定した共通的な基盤（プラットフォーム）について設計を行い、マルチモーダルなデータの収集・アノテーションを進めながら、上記の目標を実現するための要素技術に関する研究を本格的に開始した。さらに、衆議院の新会議録作成システムの試験運用において音声認識の評価を行った。

具体的には以下の通りである。

○ **実験環境の整備とコーパス収集分析**

前年度に引き続き、実験環境の整備とコーパスの収集を進めた。

(1) **実験室環境の整備**：セミナー・ミーティング、電子掲示板によるポスター発表を行える環境を整備した。特に、知能化ポスターボードについて、カメラなどのセンサ群を京大の研究室に設置するとともに、奈良先端大でマイクロフォンアレイの設計と試作を行った。

(2) **コーパスの収集とアノテーション**：前年度に収録したポスター会話について、あいづち・うなずき・視線などの様々なアノテーションを引き続き行った。また、セミナーや講演等の収録も多数行い、発話の書き起こしやスライドとの対応付けなどを行った。

(3) **マルチモーダル情報の分析・抽出と認知状態との関連付け**：マルチモーダルな非言語情報と認知状態について、特にポスター会話を対象として分析を行った。

○ **マルチモーダル認識及び情報支援に関する研究開発**

以下の要素技術に関する研究開発を行った。図1に研究全体の処理における位置づけと分担を示している。

(4) **音声認識**：セミナーや講演・講義などの音声認識[文献 7]のために、話者や話題に対してモデルを適応する方法を研究した。具体的には、講演の予稿やスライドのテキストから発話の言語モデルを予測する方法を検討した。

(5) **書き起こしの整形及び構造抽出**：話し言葉の整形を行う方法について引き続き研究した。国会の会議録だけでなく、講演録や字幕の作成を想定して実装・評価を行った。

(6) **聴衆を対象とした音響・音声処理**：マイクロフォンアレイを用いて、ポスター会話中の発話を分離・強調する方法を研究した。また、入力音声から発話区間を頑健に検出する方法[文献 2]、及び残響を抑圧する方法[文献 3]について論文発表を行った。

(7) **聴衆を対象とした映像処理**：複数のカメラを用いて、聴衆の視線や反応（うなずき、ジェスチャーなど）を検出・認識する方法を研究した。また、ポスター会話の参加者間のインタラクションに関して、様々な行動を検出・分類し、ハイブリッドダイナミカルシステムを用いて分析・モデル化した。

(8) **聴衆の反応の認識**：特定のパターンのあいづちを検出することにより、聴衆が関心を持った箇所を推定し、ポスター会話中の重要な区間（ホットスポット）を抽出することができることを明らかにした[文献 6]。

(9) **質問応答及び情報推薦**：話題や文脈に応じて関連した情報を検索し、提示する方法を研究した。

○ **セミナー・会議の支援システムの構築と運用**

(10) **衆議院での音声認識システム運用**：衆議院の新会議録作成システムの試験運用において、音声認識の評価を行った[文献1,5]。多数の審議について公式の会議録と照合した結果、文字正解率で 89%を実現した。

§4. 成果発表等

（4－1） 原著論文発表

●論文詳細情報

[1] *三村正人, 秋田祐哉, 河原達也.
統計的言語モデル変換を用いた音響モデルの準教師つき学習.
電子情報通信学会論文誌, Vol.J94-D, No.2, pp.460--468, 2011.

[2] *D.Cournapeau, S.Watanabe, A.Nakamura, and T.Kawahara.
Online unsupervised classification with model comparison in the Variational Bayes framework for voice activity detection.
IEEE J. Selected Topics in Signal Processing, Vol.4, No.6, pp.1071--1083, 2010. (DOI:10.1109/JSTSP.2010.2080821)

[3] *R.Gomez and T.Kawahara.
Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood.
IEEE Trans. Audio, Speech & Language Processing, Vol.18, No.7, pp.1708--1716, 2010. (DOI:10.1109/TASL.2010.2052610)

[4] R.Gomez and T.Kawahara.
Optimizing wavelet parameters for dereverberation in automatic speech recognition.
Proc. APSIPA ASC, pp.446--449, 2010.

[5] *T.Kawahara.
Automatic transcription of parliamentary meetings and classroom lectures -- a sustainable approach and real system evaluations --.
Proc. Int'l Sympo. Chinese Spoken Language Processing (ISCSLP), pp.1--6 (keynote speech), 2010. (DOI:10.1109/ISCSLP.2010.5684907)

[6] *T.Kawahara, K.Sumi, Z.Q.Chang, and K.Takanashi.
Detection of hot spots in poster conversations based on reactive tokens of audience.
Proc. INTERSPEECH, pp.3042--3045, 2010.

[7] *T.Kawahara, N.Katsumaru, Y.Akita, and S.Mori.
Classroom note-taking system for hearing impaired students using automatic speech recognition adapted to lectures.
Proc. INTERSPEECH, pp.626--629, 2010.

[8] R.Gomez and T.Kawahara.
An improved wavelet-based dereverberation for robust automatic speech

recognition.
Proc. INTERSPEECH, pp.578--581, 2010.

[9] Y.Akita, M.Mimura, G.Neubig, and T.Kawahara.
Semi-automated update of automatic transcription system for the Japanese national congress.
Proc. INTERSPEECH, pp.338--341, 2010.

[10]T.Kawahara, Z.Q.Chang, and K.Takanashi.
Analysis on prosodic features of Japanese reactive tokens in poster conversations.
Proc. Int'l Conf. Speech Prosody, 2010.

[11]T.Inoue, H.Saruwatari, K.Shikano, and K.Kondo.
Theoretical analysis of musical noise in Wiener filter via higher-order statistics.
Proc. of APSIPA ASC, pp.121-124, 2010.

[12]T.Inoue, H.Saruwatari, Y.Takahashi, K.Shikano, and K.Kondo.
Theoretical analysis of musical noise in generalized spectral subtraction based on higher-order statistics.
IEEE Trans. Audio, Speech & Language Processing (in press).
(DOI:10.1109/TASL.2010.2098871)

[13]T.Inoue, H.Saruwatari, K.Shikano, and K.Kondo.
Theoretical analysis of musical noise in Wiener filtering family via higherorder statistics.
Proc. IEEE-ICASSP (accepted).

[14]R.Miyazaki, H.Saruwatari, R.Wakisaka, K.Shikano, and T.Takatani.
Theoretical analysis of parametric blind spatial subtraction array and its application to speech recognition performance prediction.
Proc. Workshop Hands-free Speech Communication & Microphone Arrays (HSCMA2011) (accepted).

[15]Y.Murawaki and S.Kurohashi.
Online Japanese Unknown Morpheme Detection using Orthographic Variation.
Proc. Int'l Conf. Language Resources & Evaluation (LREC), pp. 832-839, 2010.

[16]Y.Murawaki and S.Kurohashi
Semantic Classification of Automatically Acquired Nouns using Lexico-Syntactic Clues.
Proc. COLING, Poster Volume, pp. 876-884, 2010.

[17]T.Tung.
3D Video Understanding using a Topology Dictionary.
Dagstuhl Seminar Proc. Computational Video, p.21, 2010.

[18]T.Tung and T.Matsuyama.
3D Video Performance Segmentation.
Proc. IEEE-ICIP, pp.25-28, 2010. (DOI:10.1109/ICIP.2010.5652541)
[19]T.Tung and T.Matsuyama.
Dynamic Surface Matching by Geodesic Mapping for 3D Animation Transfer.
Proc. IEEE-CVPR, pp.1402-1409, 2010. (DOI:10.1109/CVPR.2010.5539806)
[20]P.Huang, T.Tung, S. Nobuhara, A.Hilton, and T. Matsuyama.
Comparison of Skeleton and Non-Skeleton Shape Descriptors for 3D Video.
Proc. Int'l Sympo. 3D Data Processing, Visualization & Transmission (3DPVT), 2010.
[21]近藤一晃, 西谷英之, 中村裕一.
協調的物体認識のためのインタラクション設計.
電子情報通信学会論文誌 (採録決定).
[22]K.Kondo, H.Nishitani, and Y.Nakamura
Human-Computer Collaborative Object Recognition for Intelligent Support.
Pacific-Rim Conference on Multimedia (PCM), pp.II-471-482, 2010.

(4－2) 知財出願

① 平成22年度特許出願件数(国内 0 件)

② CREST 研究期間累積件数(国内 0 件)