「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

H23 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. 研究実施体制

(1)京大グループ

①研究代表者: 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

(2)奈良先端大グループ

①主たる共同研究者: 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

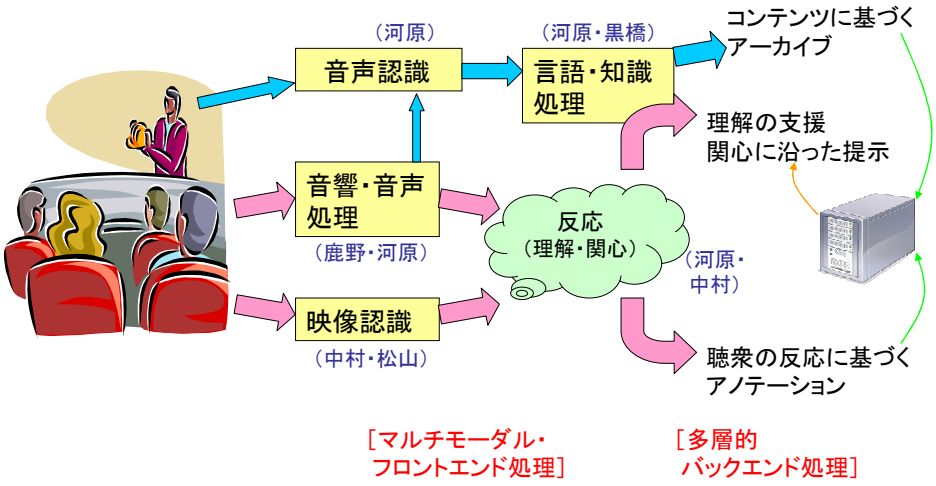
§2. 研究実施内容

(文中に番号がある場合は(3－1)に対応する)

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、実際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、衆議院の新会議録作成システムを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



平成 23 年度は、セミナー室とポスターボードといった共通的な基盤(プラットフォーム)を想定しながら、要素技術の本格的な実装・評価に注力した。特にポスター発表をマルチモーダルにアーカイブ・分析するシステム(スマートポスターボード)については、要素技術を統合したプロトタイプを構築した。また音声認識については、衆議院のシステムの本格運用後も引き続き評価を行い、講演などへの適用についても研究を進めた。

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**

(1) **実験室環境の整備**：マイクロフォンアレイやカメラ群などのセンサ群を備えた電子掲示板（スマートポスターボード）のプロトタイプを構築した。

(2) **コーパスの収集とアノテーション**：実験用データとしてポスター会話を新たに収集し、評価に必要なアノテーションを行った。具体的には、日本語・英語同数の計12セッション収録し、音声の書き起こしに加えて、視線やうなずきなどのアノテーションを行った。

(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整理を行った。

○ **マルチモーダル認識及び情報支援に関する研究開発**

(4) **音声認識**：セミナーや講演などの音声認識のために、話者や話題に対してモデルを適応する方法を研究した。特に、予稿集のテキストに基づいて、音声認識の言語モデルを効率的に構築する方法を実現した。

(5) **書き起こしの整形及び構造抽出**：音声認識結果を整形し、句読点を挿入し[文献 10]、字幕を生成する方法を研究した。

(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、遠隔話者の発話を分離・強調し、各話者の発話区間を検出する方法を研究した[文献12-14]。前記のスマートポスターボードを対象として実装を行った。

(7) **聴衆を対象とした映像処理**：複数のカメラを用いて、電子掲示板との位置関係をキャリブレーションしながら、人物の顔と視線（顔の向き）などを検出する方法を研究した[文献 26]。前記のスマートポスターボードを対象として実装を行った。

(8) **聴衆の反応の認識**：視線やあいづち・うなずきなどの情報に基づいて、聴衆の反応を予測するモデルを研究した[文献3]。

(9) **質問応答及び情報推薦**：セミナーや講演・講義で話される専門用語を抽出する方法を研究した。特に、複数の単語からなる用語や未知語を扱える手法を研究した[文献 30]。

○ **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が正式運用になった以降も継続的に音声認識の評価とモデルの更新を行った。文字正解率で 90%を達成した。

(11) **電子掲示板を用いたポスター発表**：上記の要素技術を統合し、ポスター発表のアーカイブ・分析を行うスマートポスターボードのプロトタイプシステムを構築し、国際会議 ICASSP2012 においてデモ展示を行った。

(12) **セミナーの支援とアーカイブ**：「聴覚障害者のための字幕付与技術」シンポジウムにおいて、音声認識を用いた講演の字幕付与のデモを行った。

§3. 成果発表等

（3－1） 原著論文発表

●論文詳細情報

[1] G.Neubig, M.Mimura, S.Mori, and T.Kawahara.
Bayesian learning of a language model from continuous speech.
IEICE Trans., Vol.E95-D, No.2, pp.614--625, 2012.

[2] 吉野幸一郎, 森信介, 河原達也.
述語項の類似度に基づく情報抽出・推薦を行う音声対話システム.
情報処理学会論文誌, Vol.52, No.12, pp.3386--3397, 2011.

[3] 河原達也, 須見康平, 緒方淳, 後藤真孝.
音声会話コンテンツにおける聴衆の反応に基づく 音響イベントとホットスポットの検出.
情報処理学会論文誌, Vol.52, No.12, pp.3363--3373, 2011.

[4] G.Neubig, Y.Akita, S.Mori, and T.Kawahara.
A monotonic statistical machine translation approach to speaking style transformation.
Computer Speech and Language, (accepted for publication).

[5] M.Ablimit, T.Kawahara, and A.Hamdulla.
Discriminative approach to lexical entry selection for automatic speech recognition of agglutinative language.
Proc. IEEE-ICASSP, pp.5009--5012, 2012.

[6] R.Gomez and T.Kawahara.
Optimized wavelet-based speech enhancement for speech recognition in noisy and reverberant conditions.
Proc. APSIPA ASC, 2011.

[7] M.Mimura and T.Kawahara.
Fast speaker normalization and adaptation based on BIC for meeting speech recognition.
Proc. APSIPA ASC, 2011.

[8] M.Ablimit, T.Kawahara, and A.Hamdulla.
Lexicon optimization for automatic speech recognition based on discriminative learning.
Proc. APSIPA ASC, 2011.

[9] T.Hirayama, Y.Sumi, T.Kawahara, and T.Matsuyama.
Info-concierge: Proactive multi-modal interaction using mind probing.
Proc. APSIPA ASC, 2011.

[10]Y.Akita and T.Kawahara.
Automatic comma insertion of lecture transcripts based on multiple annotations.
Proc. INTERSPEECH, pp.2889--2892, 2011.

[11]R.Gomez and T.Kawahara.
Denoising using optimized wavelet filtering for automatic speech recognition.
Proc. INTERSPEECH, pp.1673--1676, 2011.

[12]T.Inoue, H. Saruwatari, Y.Takahashi, K.Shikano,K.Kondo.
Theoretical analysis of musical noise in generalized spectral subtraction based on higher-order statistics.
IEEE Trans. Audio, Speech & Language Processing, vol.19, no.6, pp.1770-1779, 2011. (DOI:10.1109/TASL.2010.2098871)

[13]H.Saruwatari, Y.Ishikawa, Y.Takahashi, T.Inoue, K.Shikano, K.Kondo.
Musical noise controllable algorithm of channelwise spectral subtraction and adaptive beamforming based on higher-order statistics.
IEEE Trans. Audio, Speech & Language Processing, vol.19, no.6, pp.1457-1466, 2011. (DOI: 10.1109/TASL.2010.2091636)

[14]R.Miyazaki, H.Saruwatari, K.Shikano.
Theoretical analysis of amounts of musical noise and speech distortion in structure-generalized parametric spatial subtraction array.
IEICE Trans. vol.95-A, no.2, pp.586-590, 2012.

[15]R.Wakisaka, H.Saruwatari, K.Shikano, T.Takatani.
Speech prior estimation for generalized minimum mean-square error short-time spectral amplitude estimator.
IEICE Trans. vol.95-A, no.2, pp.591-595, 2012.

[16]R.Miyazaki, H.Saruwatari, T.Inoue, Y.Takahashi, K.Shikano, K.Kondo.
Musical-noise-free speech enhancement based on optimized iterative spectral subtraction.
IEEE Trans. Audio, Speech & Language Processing (accepted for publication).

[17]T.Inoue, H.Saruwatari, K.Shikano, K.Kondo.
Theoretical analysis of musical noise in Wiener filtering family via higherorder statistics.
Proc. IEEE-ICASSP, pp.5076-5079, 2011.

[18]R.Miyazaki, H.Saruwatari, K.Shikano.
Theoretical analysis of musical noise and speech distortion in structure-generalized parametric blind spatial subtraction array.
Proc. INTERSPEECH, pp.341-344, 2011.

[19]R.Wakisaka, H.Saruwatari, K.Shikano, T.Takatani.
Blind speech prior estimaion for generalized minimum mean-square error short-time spectral amplitude estimator.
Proc. of INTERSPEECH, pp.361-364, 2011.

[20]K.Kubo, H.Kawanami, H.Saruwatari, K.Shikano.
Unconstrained many-to-many alignment for automatic pronunciation annotation.
Proc. APSIPA ASC, 2011.

[21]Y.Fujita, S.Takeuchi, H.Kawanami, T.Matsui, H.Saruwatari, K.Shikano.
Out-of-task utterance detection based on bag-of-words using automatic speech recognition results.
Proc. APSIPA ASC, 2011.

[22]K.Nishimura, H.Kawanami, H.Saruwatari, K.Shikano.
Investigation of statistical machine translation applied to answer generation for a speech-oriented guidance system.
Proc. APSIPA ASC, 2011.

[23]H.Saruwatari, N.Hirata, T.Hatta, R.Wakisaka, K.Shikano, T.Takatani.
Semi-blind speech extraction for robot using visual information and noise statistics.
Proc. IEEE ISSPIT, pp.238-243, 2011.

[24]R.Miyazaki, H.Saruwatari, T.Inoue, K.Shikano, K.Kondo.
Musical-noise-free speech enhancement: theory and evaluation.
Proc. IEEE-ICASSP, pp.4565-4568, 2012.

[25]R.Wakisaka, H.Saruwatari, K.Shikano, T.Takatani.
Speech kurtosis estimation from observed noisy signal based on generalized Gaussian distribution prior and additivity of cumulants.
Proc. IEEE-ICASSP, pp.4049-4052, 2012.

[26]T.Tung and T.Matsuyama.
Topology Dictionary for 3D Video Understanding.
IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI). (accepted for publication)

[27]近藤一晃, 西谷英之, 中村裕一.
協調的物体認識のためのマンマシンインタラクション設計.
電子情報通信学会論文誌, Vol. J94-D, No.8, pp. 1206-1215, 2011.
[28]Z.Yu, Z.Yu, X.Zhou, C.Becker, and Y.Nakamura.
Tree-based Mining for Discovering Patterns of Human Interaction in Meetings,
IEEE Trans. Knowledge and Data Engineering,
[29]Z.Yu, X.Zhou, Z.Yu, C.Becker, and Y.Nakamura.
Social Interaction Mining in Small Group Discussion Using a Smart Meeting System.
Proc. UIC, Springer-Verlag LNCS, 2011.
[30]Y.Murawaki and S.Kurohashi.
Non-parametric Bayesian Segmentation of Japanese Noun Phrases.
Proc. EMNLP, pp. 605-615, 2011.

(3－2)その他

IEEE ICASSP 2012（2012 年 3 月 27-29 日；京都国際会館）において、スマートポスターボードシステムのデモ展示を 3 日間にわたって行った。マイクロフォンアレイとカメラ群を備え、ポスターセッションにおける音声・映像・視線などの情報を検出して記録することができる。多数の方が見学し、興味を持たれていた。

http://www.ar.media.kyoto-u.ac.jp/crest/