3.5.4　知識創成コミュニケーション研究センター　音声コミュニケーショングループ

グループリーダー　　中村　哲　ほか 43 名

ナチュラル言語コミュニケーション技術に関する研究開発

概　要

誰が、いつ、どこで、どのような表現で、何語で話そうとも、音声や身振り・手振りなどの人間にとって自然な言語・非言語表現によって情報を補いながら、息の合ったコミュニケーションを実現するナチュラル言語コミュニケーションの構成技術を推進している。本年度は、研究の基盤となる対話コーパスの整備、非言語情報を利用するための画像処理技術の開発、音声認識・合成技術の改良および多言語化など、対話システムを構成する要素技術の開発を進めた。さらに、これらの技術を統合して対話処理のプロトタイプシステムを開発し、実証実験を行って、問題点を抽出・整理した。

平成 21 年度の成果

(1)音声対話システムの研究開発

コーパス整備・観光スポットデータベース整備

- コンピュータによる自然な対話を実現するために、人と人の対面対話の分析を継続して進め、発話行為タグ・コンテンツタグ付与によるデータ整備を継続して行った（10 対話）。対面対話より制限された状況での人間の自然な振る舞いを調べるために、非対面対話データに対しても同様にタグ付与を行なった（20 対話）。また、統計的な対話制御によって、より人間らしく応答するシステムを実現するために、非対面対話データに対する専用のタグを付与した（20 対話）。
- 大規模実証実験を前提として京都市内の主要な観光スポット 100 か所に関する詳細な対話システム用データベースを整備した。既存の簡易情報とあわせて合計 840 件のデータベースを構築した。

音声対話処理の構成技術開発

- 人間の胸部の高さに設置したカラーステレオカメラの画像を入力として、複数の人物が融合した 3 次元点群のクラスタを分割して個人領域に切り分けることにより、頭部位置を高精度（混雑環境で 93%、疎な環境で 99.7%）に推定するシステムを開発した。また、複数台カメラの誤差を最急降下法で最小化することにより、特別な照明なしにユーザの顔向きを高精度（誤差は約 5 度）に推定することに成功した。
- 多言語音声対話システムの実現に向けて、英語話者による対話音声コーパスの収録、ソフトウェアプラットフォームの多言語対応化、対話シナリオの英語化を行った。
- 評価グリッド法及び Web 調査により抽出した選好構造※1 に基づいて、選択式質問の繰り返しにより京都の観光スポットを推薦するシステムを開発した。

図 1　スマートフォン型音声対話システム

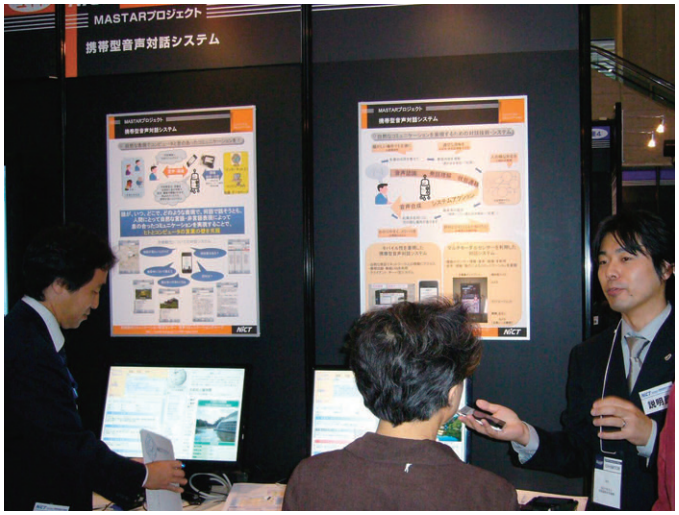

図 2　CEATEC Japan 2009 での展示

プロトタイプシステム開発

- これまでに収集し整備した対話コーパス（256 対話、約 130 時間）から音声認識のための言語モデル、音声認識・音声合成のための固有名詞辞書、音声認識用試験データを作成し、統計的対話制御機能を用いたスマートフォン型音声対話プロトタイプシステムを開発した（図 1）。
- 音声情報に加え、非言語情報として人物の立ち位置と顔向きの情報を利用した大画面対話システムのプロトタイプを構築し、注視領域に基づきシステム側から自発的に提案発話を行う対話制御を実現した。

実証実験

2 つのプロトタイプシステム（スマートフォン型（A）と大画面型（B））の有効性、実稼働性を検証するために、実験環境下での 100 名に対する実証実験を実施した。その結果、各システムの音声認識精度（単語誤り率。値が小さいほど精度が高い。）は、（A）4,400 発話に対して 20%、（B）5,300 発話に対して 40%であり、応答成功率は（A）65%、(B）50%であった。システム（B）の画像処理による頭部検出率は 99.7%であり、顔向き推定によるシステムからの提案発話の受け入れ率は 40%であった。事後的分析によれば、システム（A）と（B）の音声認識精度の差は、主として周囲雑音の影響によるものである。また、応答記録、被験者アンケートの分析より、周辺検索などバックエンド処理機能の不足、紹介可能スポットの不足などが問題点として抽出された。

展示等

- Pervasive 2009（5 月）、CEATEC Japan 2009（10 月）、けいはんな情報通信研究フェア 2009（11 月）、IUCS 2009（12 月）、情報処理学会創立 50 周年記念全国大会（3 月）において、対話システムの展示を行った（図 2）。
- ロボットに磨く、捨てるなどの動作を学習させる技術（動作学習技術）と、物体の形状・名前をその場で学習する技術（未登録語学習技術）など家庭で役立つ機能を搭載したロボットを玉川大学、電気通信大学と共同で開発し、ロボカップ 2009 世界大会（6 月 29 日～ 7 月 5 日）において準優勝した。

⑵音声認識技術・音声合成技術の研究開発

音声認識技術の頑健化

「誰が」「どこで」話しても認識できるように、話者や環境に頑健な音声認識手法の開発を進めた。具体的には、話者に適応して音声認識精度を高める技術の開発、英語音声認識における非母語話者への対応、雑音や残響に強い音声認識手法の開発を行った。話者への適応では、音素を安定して識別する学習手法を導入し、10 ～ 20 文程度の少量の発声で従来法を上回る高い適応精度を実現した。非母語話者への対応では、適応技術を利用して、日本人、韓国人、フランス人などの英語非母語話者用の音響モデルを開発し、単語誤り率を約 30%から約 15%に改善した。雑音・残響対策では、残響等の影響を受けにくい特徴量を導入し、さらに正規化処理により残響の影響を低減することで、話者がマイクから 1m 以上離れた場合の認識率の低下を抑えることに成功した（図 3）。

音声認識および音声合成の多言語化

音声認識では、韓国語、ポルトガル語の認識エンジンの構築に着手した。学習用の音声コーパス、テキストコーパスを収集・整備して音響モデル、言語モデルを作成し、基本的な認識動作を確認した。

音声合成では、韓国語 HMM[※2] 音声合成システムを開発した。また、Web ページから音声データを自動収集し、HMM を自動作成するツールを開発し、利用可能性を確認した。

対話音声合成

対話システムのガイド音声として自然で聞き取りやすい合成音声を実現するため、声優による演技 12 対話分、及び旅行ガイドによる模擬対話 42 回分を高音質で収録し、コーパスとして整備した。

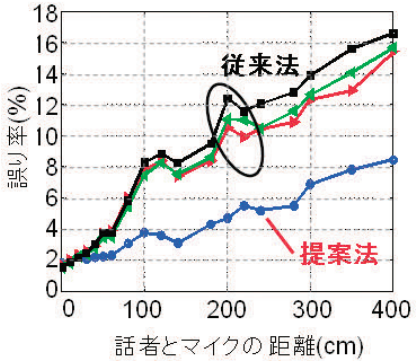

図 3　残響に頑健な音声認識手法

※ 1 選好構造: 嗜好間の優先順位、因果関係など表現する構造。
※ 2 HMM: Hidden Markov Model。音声の認識・合成技術で一般的に用いられる確率モデルの一種。