コンテンツの解析からインタラクションの解析へ

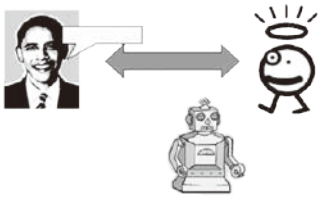
河原 達也 京都大学・学術情報メディアセンター

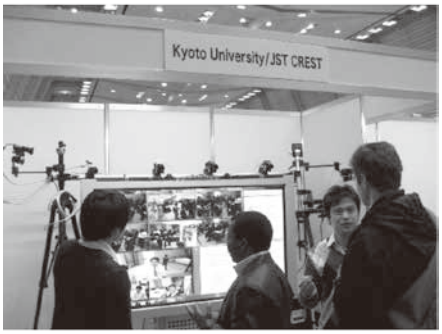
〔受賞論文〕
音声会話コンテンツにおける聴衆の反応に基づく音響イベントとホットスポットの検出
河原達也，須見康平（京都大学情報学研究科），緒方淳，後藤真孝（産業技術総合研究所）
情報処理学会論文誌，Vol.52, No.12, pp.3363-3373（2011）

「情報爆発」や「ビッグデータ」の波が，音声や画像・映像などにも及んできている．現在このようなマルチメディアコンテンツを検索できるようにするには，基本的に人手でタグを付与する必要があり，自動化のためのコンテンツ解析に関する研究が世界的に行われている．これにあわせて，音声認識・文字認識・顔認識などのパターン認識技術が急速に進展している．筆者の専門である音声認識技術についても，近年のスマートフォンアプリや，2011年度喜安記念業績賞をいただいた国会の会議録作成システムなどで実用化が進んでいる．しかしながら，我々がふだん話しているような会話をテキストにするのはまだ遠い夢物語である．

これは例えると，多くの日本人が外国に行って，旅行会話や（自分の分野の）学会講演の理解はできても，トークショーやテレビドラマをあまり聞き取れないのと同様である．皆さんもバンケットのスピーチでジョークが理解できなかった経験はないでしょうか？　たとえ内容が理解できなくても，周囲の聴衆の反応を見ていると，どこが面白くて，どこが話のポイントであったかは推測可能である．このようなアイディアを実現したのが，本論文の内容である．このアイディアは，Webページのランクがそこに書かれている内容よりもリンク数（他の人の反応）に基づいているのと少し類似している．本研究では具体的に，会話における聞き手の笑い声や，「へー」などの特定のパターンの相槌に着目し，それらを検出することで，会話コンテンツの重要個所（＝ホットスポット）を抽出することを試みた．アイディア自体は，従来の音声認識や対話処理とまったく異なる斬新なものであったが，初期的な段階であったので，論文賞をいただけるとは思わなかった．

現在，JST CRESTのプロジェクト「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」において，学会やオープンラボで行われるポスターセッション





を対象として，マルチモーダルな情報処理に展開している．このプロジェクトでは，カメラやマイクアレイを搭載した電子掲示板（大型LCD）で行うポスターセッションで，訪れた聴衆の顔・視線（顔向け）・発話を検出し，興味・理解度を推定することを目指している．講演形式の口頭発表の多くが録音・録画されて，音声認識等の研究も進められているが，ポスターセッションでは発表者と聴衆とのインタラクションが重要であり，このモデル化・アノテーションを行う試みはほかにない．写真は国際会議でのデモ風景である．まださまざまな技術的課題があるが，あと数年でシステムの形にしていきたい．

最後に，本研究の主要部分を実装してくれた当時大学院生の須見康平君（現在ヤマハ（株）），研究協力をいただいた産業技術総合研究所の緒方淳さん・後藤真孝さん，日頃から議論していただく京都大学の研究室の皆さん，ならびにCREST関係者の皆さんに深い感謝の意を表したい．
（2013年5月10日受付）

河原 達也（正会員）　kawahara@media.kyoto-u.ac.jp
1989年京都大学大学院工学研究科情報工学専攻修士課程修了．現在，京都大学学術情報メディアセンター教授．京大博士（工学）．音声言語処理，特に音声認識および対話システムに関する研究に従事．