K-057

テレビ視聴者の操作意図を推定するための マルチモーダルデータベースの枠組み

A Framework for Multi-modal Database to Infer TV Viewer's Intentions

小峯一晃† 森田寿哉† 金淵培† 浦谷則好†
Kazuteru Komine Toshiya Morita Yeun-Bae Kim Noriyoshi Uratani

1．まえがき

テレビ用のユーザインタフェースとして，マルチモーダルインタフェースを検討している．発話による言語情報とそれ以外の非言語情報によって視聴者の操作意図を把握し，その状況に応じた適切な支援を行うことにより，使いやすいインタフェースが実現できると考えられる．

しかしながら，音声対話によるテレビ操作時の非言語情報については収集・分析されている例がほとんど無く，操作時に得られるマルチモーダル情報を操作意図にマッピングするためには，発話に同期してこれらの情報を収集し分析する枠組みを開発する必要がある．そこで，今回，そのような情報を収集・分析する枠組みとしてマルチモーダルデータベースの入力・分析ツールを開発したので報告する．

2．操作意図推定の方法

日常の人対人の対面コミュニケーションにおいて，情報伝達に占める非言語情報の役割は大きく，状況によっては，伝達されるメッセージの６割～９割が非言語情報によるものであるとの説もある(1)．同様に，機器を音声対話で操作する際にも，そのユーザは無意識に非言語情報を用いたコミュニケーションを行っていると考えられる．筆者らは Wizard of OZ 方式による模擬対話実験により，音声対話によるテレビ操作時の発話を分析するとともに操作時の視線や韻律など非言語情報についても観察した(2)．その結果，音声対話によるテレビの操作というコミュニケーション環境においても発話には省略や指示があるだけでなく，韻律や表情の変化も日常の会話と変わらない程度に大きく，非言語情報が豊富に表出されていることが確認された．

これらの検討から，テレビ操作時に発話内容だけでなく非言語情報を利用することにより，高い精度でユーザの操作意図を推定することが可能になると考えられる．

†NHK 放送技術研究所

非言語情報は各モダリティで独立した意味を表現して標識・例示(1)などの役割を果たす場合も見られるが（特に身体動作），言語情報に付随して無意識のうちにある意図を表現している場合が多く，各モダリティ間には相関があり，それぞれが補完・協調して一つの意図を表出していると捉えることができる．発話から得られる言語情報で不明確な部分をこれらの非言語情報で補完し，操作意図を推定するためには，対話システムの状態や言語情報との相関を考慮して意図に変換する必要があり，そのためには同期してそれぞれのコーパスを収集し，一元的に分析する枠組みが必要となる．

今回開発したマルチモーダルデータベースの枠組みは音声対話による操作時に得られるマルチモーダル情報を管理・分析する目的で構築したものである．以下にその内容について述べる．

3．非言語情報による意図推定の例

図１は筆者らが提案している操作意図を推定するシステムの構成である．あらかじめ，個々の操作意図の同定に必要な要素から構成されるスロット群（操作意図スキーマ）を想定される操作意図の数だけ用意する．ユーザが何らかの操作を行う際に，発話からだけでは埋められないスロットを非言語情報や操作に関する知識を用いて補完することにより，ユーザの操作意図を推定しようとするものである．以下に具体例を挙げる．

3.1 視線による補完

音声対話によるテレビ操作時の発話には操作するオブジェクトを省略したり，指示代名詞で置き換えたりすることが頻繁に発生する．その際に操作者の注視点を利用することで操作の対象となるオブジェクトを推定することが可能になる．また，「これは何ですか？」のような発話時に指差しで画面上のオブジェクトを示すこともよく見られる行動であるため，指示動作（ジェスチャ）による情報と併用することにより，推定の精度は高められると思われる．

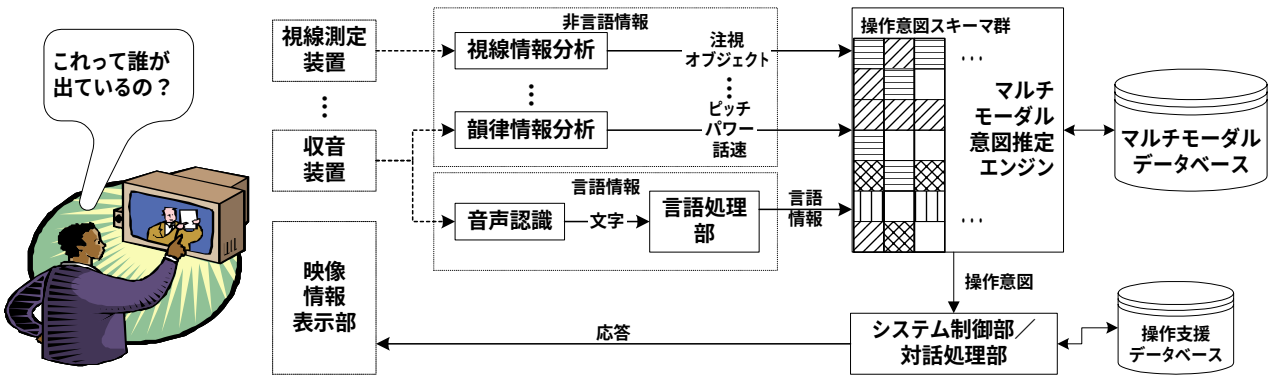

図１　想定するシステム構成

発話内容とこれら非言語情報の共起性から意図を推定するためには，両者の発生するタイミングを調べる必要がある．各モダリティの状態を同期して表示することができ，一元的に分析可能なマルチモーダルデータベースはその有効な手段となり得る．

3.2 韻律情報による補完

発話時の音声にはテキスト情報以外にパワー・ピッチ・話速などの韻律情報が含まれており，これらは感情や強調など，発話者の心的状況を表出していることが多いとされている[(3)]．実際に筆者らが行った模擬対話実験においても，操作方法がわからずに迷っているような状況でピッチや話速などの韻律情報に変化が観察されており，ユーザはシステムに対して操作支援を要求していると捉えることができる．このように韻律情報は心的状況や操作意図の推定に利用できる．

4．マルチモーダルデータベース入力・分析ツール

今回開発したマルチモーダルデータベースの枠組みでは表１の言語情報，非言語情報を取り扱うことを想定している．

これらの各種モダリティの情報を同期して表示し，ある意図を表出した対話におけるモダリティ間の相関やタイミングを分析するために，データベースの入力・分析ツールを開発した（図２参照）．本ツールの構成を以下に示す．

- **操作映像表示部**：ユーザの操作状況を撮影したバストショット映像とタイムコードの表示．映像の制御．
- **非言語情報表示部**：各種非言語情報の波形および位置に関する生データ，分析結果の表示．コメントの入力・表示．
- **視線画像・実験情報表示部**：ユーザの視線情報を操作画面に重畳した映像の表示．実験条件，ユーザのプロファイルなどの入力・表示．
- **言語情報表示部**：発話内容，発話行為タグ，タイムコードの入力・表示．発話中に現れる単語の品詞・意味等の入力・表示．

4．まとめ

音声対話でテレビを操作する際のマルチモーダル情報を一元的に管理・分析するマルチモーダルデータベースの枠組みを構築した．

今後は同期して各種モダリティの情報を収集できる実験システムを構築し，マルチモーダルコーパスを収集するとともに，各種モダリティ間の相関を分析する予定である．

参考文献

(1) 黒川隆夫：“ ノンバーバルインタフェース ”，pp.30-31，pp.41-68，オーム社（1994）
(2) 小峯一晃他：“ 音声インターフェースによる番組選択操作時の発話内容分析 ”，ヒューマンインタフェースシンポジウム 2002，3231，pp.631-634（2002）
(3) 田村博：“ ヒューマンインタフェース ”,pp.264-265，オーム社（1998）

表１　想定している各種モダリティ

言語／非言語	情報の種類	内容
言語情報	発話内容	操作時の発話内容
	意味マーカ	発話中に現れる単語の意味・概念
非言語情報	韻律情報	発話を音響分析して得られるパワー・ピッチ・話速
	視線情報	操作時に注視している画面上の位置，オブジェクト
	表情	顔面上の設定された特徴点の位置，推測される感情
	動作	肢体に設定されたマーカの位置，推測される意味
	生体情報	顔表面温度，発汗量，血流量，呼吸量，筋電，脳波など

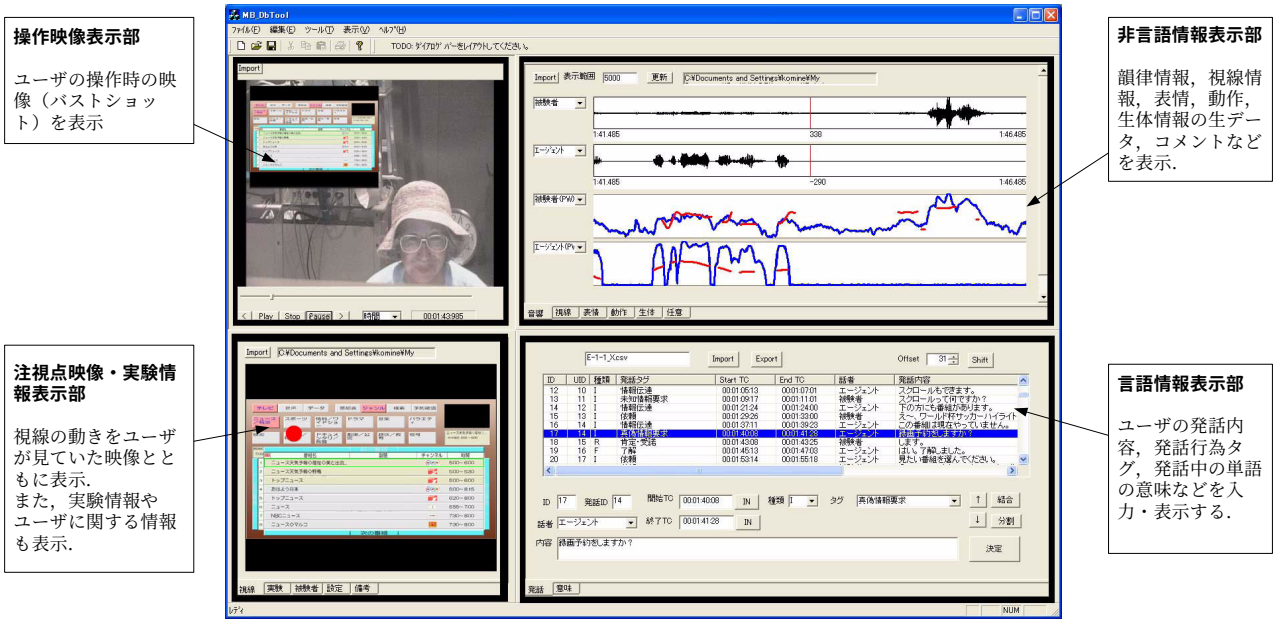

図２　マルチモーダルデータベースの入力・分析ツール画面