1L-05 感性的マルチモーダル対話のための音声からの感情抽出

松岡文吾　　土肥浩　　伊庭斉志　　石塚満

東京大学工学部電子情報工学科

1 はじめに

人間とコンピュータのインタラクションに用いられている技術としてマウス、キーボード、音声認識、画像認識などが挙げられる。これらは確かに論理的で正確であると言えるが、一方では機械的であり冷たいというイメージも同時に人々に与えている。このことは、人間同士のコミュニケーションでは必ずやりとりされる感性が欠けていることの現れである。

専門家の分析によると、我々の日常会話で交わされるメッセージのおよそ65%は「言葉」以外のノンバーバルな要素（動き・雰囲気・声質・表情・感性など）が担っているとしている。この評価は、ノンバーバル要素が我々の日常的コミュニケーションを図る上で、いかに重要であるかを示している。

人間同士でのコミュニケーションと同様に、人間とコンピュータとのインタラクションに関しても、このノンバーバルな要素を的確に利用することでより良いコミュニケーションを図ることが可能となる。本研究では、ノンバーバルな要素の中でもとりわけ「感情」の要素に焦点をしぼり、「コンピュータに人間の感情を分析・理解させ、その分析結果に応じた対応を取らせることによって、人間とコンピュータのマルチモーダルな対話の実現する」ことを目的とする。

2 音声と感情

音声情報から感情を分析する方法としては「韻律情報の解析」が挙げられる。音声情報の持つ「韻律」は以下に挙げる3要素に分類が可能である。ストレスに代表される「振幅構造」、単語単位の継続時間やリズム・ポーズを構成する「時間構造」、イントネーションやアクセントなどを生む「ピッチ構造」の3つである。

Extraction of Emotion from Voice for Application with Emotional Multimodal Interaction

Bungo Matsuoka

Department of Information and Communication Engineering, Faculty of Engineering, University of Tokyo

7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

「振幅構造」は「音声パワー」、「時間構造」は「音声全体の長さ」、「ピッチ構造」は「音声周波数」とそれぞれ深い関わりをもっており、複合的に絡み合うことで人間の言語音声受容に重要な役割を果たす。近年の研究により、この「振幅構造」、「時間構造」、「ピッチ構造」の3つの要素は「感情」を特徴づける韻律規則のパラメータとして有効であることが証明されている。具体的に言うと「怒り」は音声パワーとピッチの変動が大きく、全体を通じて音声パワーが大きい。「歓喜」は音声パワーにこそ大きな特徴は現れないが、ピッチは全体的に高く、その変動も大きい。一方「悲哀」は音声パワー、ピッチともに全体的に低く、また変動も小さいなどの特徴がある。

3 感情の抽出過程

音声情報の入力環境に関しては、背景雑音が入力音声に比べて無視できるほど十分に小さな環境で行った。音声を入力する際も、マイクに出来るだけ息がかからないように、また入力が終わるまではマイクに近づいたり遠ざかったりしないよう注意して行った。図1はその音声データの1例である。なお、解析にはアニモ社の Voice Base II と、Visual C++を用いた。

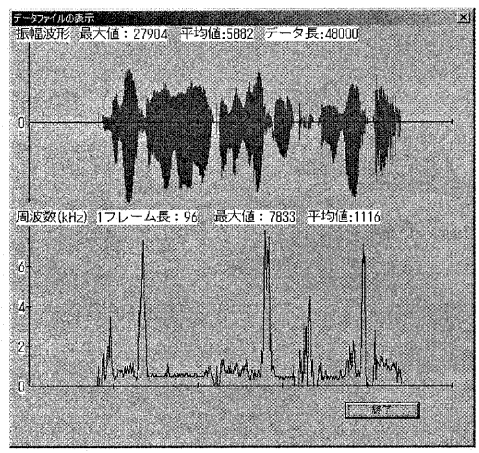

図1．音声データの例

次に図1の中である一定の閾値を設定し、閾値よりも振幅の値（音声パワー）が小さな部分を雑音区間、大きな部分を音声区間とし、雑音を除去した音声区間

のみのデータから感情解析を行った。解析では、感情を決定するパラメータとして振幅最大値、振幅平均値、振幅の分散、振幅の最頻値（上位５つ）、周波数最大値、周波数平均値、周波数の分散、周波数の最頻値（上位５つ）などを使用した。前述の通り「怒り」の場合は音声パワーとピッチの変動が激しく、全体的を通じて音声パワーが大きいため、振幅平均値や振幅の分散、周波数の分散は共に大きい値を取る。逆に「悲哀」の場合は音声パワーとピッチが全体的に小さくまた変動も小さいため、振幅平均値と振幅の分散、周波数平均値と周波数の分散が全て小さい値を取る。「歓喜」は、全体的にピッチが高くて変動も激しいので、周波数平均値と周波数の分散が大きくなる。このように実験で取り扱う感情の特徴として、あらかじめ基本値となるデータを複数の被験者の平均として採取し、その基本値と実際のデータを照らし合わせることで感情の解析を行った。

4　システムの実装

今回の研究では、入力した音声データに対するコンピュータの応答を、Microsoft 社から配布されている Microsoft Agent（以下 MSAgent とする）を用いて実現させた。MSAgent を使用することにより、キャラクタに動作を持たせることや、喋る言葉の韻律に変化をつけることが容易に実現できる。操作は音声の解析時と同様に Visual C++を用いて行った。

マルチモーダル対話の実現として、まずは比較的感情としてはっきり特徴に現れやすい「怒り」に重点をしぼり、キャラクタとの対話を行った。図２は実際に怒っている状態の音声を入力させた場合のキャラクタの対応である。この例では、キャラクタは入力データに「怒り」が感じられた場合、ユーザを落ち着かせる態度をとるように設計している。

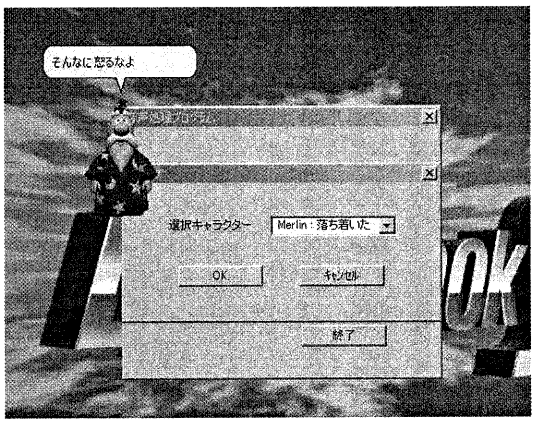

図２．MSAgent の対応例

今後の展開としては現在扱っている「怒り」、「歓喜」、「悲哀」の３つの感情以外にも取り扱う感情の種類を増やし、入力データに応じて多種多様な対応ができるようにする、また、MSAgent のキャラクタにそれぞれ基本的な個性を持たせ、全く同じ音声データに対してもキャラクタ毎に対応が異なるといった汎用性の高いシステムの設計を目指すつもりである。

5　おわりに

このように、音声には感情を含ませることができ、コミュニケーションにおけるノンバーバル要素の中で重要な役割を担っている。そしてこの事実は、確かに音声の基本周波数や音声波形の振幅における最高値や平均値、分散、最頻値の比較などからある程度の感情を推測し、それに伴ったコンピュータとの感性的対話が可能であることを示している。しかし、正確な感情を解析するためには、これらの情報のみでは明らかに不十分である。近年では音声情報のみから出来るだけ正確に感情を判断するために、今回使用した振幅波形と基本周波数の数値的データ解析以外にも、感情別の基本周波数パターンや発声速度の変化過程、語尾周辺の振幅波形変動といった「文章全体」ではなく「文を構成する個々の単語レベル」、さらには「モーラ」自体についても検討するべきだと考え方が強くなっている。本研究の目的はマルチモーダルな対話の実現であり、そのためには出来るだけ正確な感情の解析が必要となる。今後はこれらの解析アプローチも取り入れ、より自然で快適なインタラクションが可能となるシステムの実現に努力したいと思う。

参考文献

[1] 北原、東倉「音声の韻律情報と感情表現」、音声研究会資料 SP88-158、pp. 27-32（1989）

[2] 広瀬，高橋、藤崎、大野「音声の基本周波数パターンにおける話者の意図、感情の表現」、信学技法、HC94-41,（1994-09）

[3] 重永、小川，中尾、「単語音声による感情表現について」、信学技法、sp95-15, pp. 39-46,（1995-05）