サウンドメディア論最終課題

K19093　福本光重

* 近年，注目される音声/音楽/音響/聴覚情報処理技術の応用について、具体例を挙げて説明しなさい。またその例について今後の期待できる点、課題となる点について述べなさい。

・概要  
近年、人工知能の進歩によって、インターネットの検索エンジンやソフトウェアはもちろんのこと、家電製品や自動車などにもAIが活用され、我々を取り巻く世界は劇的に変化した。中でも特に注目を集めているのが、人工知能の進歩によって精度を増した音声認識の分野である。人の話し言葉などを聞き取って解釈し、それに対してさまざまなサービスを行う、ソフトウェアや製品を開発するため、さまざまな研究が進められている。

・音声認識とは  
音声認識とは、入力された音声データをコンピュータに認識させ、テキストデータに変換させる技術のことである。音声認識の技術を用いた、身近な機器といえば、AppleのSiri、Android OSのGoogle Nowなどのスマホに搭載されているパーソナルアシスタント機能やAmazon EchoやGoogle Homeなどの音声認識操作による家電、ソフトバンクが開発したPaperなどの対話型のロボットのなどが挙げられる。

・音声認識を行うための代表的な技術  
音声認識を行うためには、主に「音響分析」「音響モデル」「発音辞書」「言語モデル」の４つの技術・工程が存在する。  
　音声認識のAIは、生の録音データからそのまま認識できるわけではない。人間は耳に入ってきた音から言葉を瞬時に認識し、無意識のうちに意味を理解することができるが、コンピュータは音声も環境音も同じ一つの音の波形としか認識できない。そのため音響分析の技術を用いて、録音データを分析し、その音から音声認識に必要な情報を抽出してコンピュータが認識できるデータに変換する。音声認識で抽出する情報のことを、特徴量という。特徴量は、その名の通り、データ内にある特徴を抽出したもので、AIによる分析では欠かせない要素である。音響分析では、音の周波数、強弱、時間情報などが特徴量として挙げられる。例えば、周波数のパターンを抽出することで、それが人間の音声なのか環境音なのかをコンピュータが認識できるようになる。  
　音響モデルでは、コンピュータがあらかじめ学習していた音や単語の情報と照らし合わせ、抽出した特徴量がどのパターンと整合するかを計算する技術である。例えば「おはよう」という音声があったとき、人であれば「お」という音声を誰が発しようとも「お」だと認識できるが、実際には声帯や音の高低の差、前後の単語とのつながりにより音は変わっていく。AIでは「お」の音のパターンを学習することで、入力された音声「おはよう」の「お」であることを認識できるようになる。  
　発音辞書では、膨大な情報データベースの中から、音の組み合わせを抽出し、単語として認識する技術である。「お・は・よ・う」のそれぞれの音を認識したら、それらを組み合わせて「おはよう」という単語として認識する。  
　言語モデルでは、音響モデルと発音辞書で認識した単語を、大量のデータを基に、単語のつながりを予測し、意味のある文章として認識する技術である。ここで、主に使われるのが「隠れマルコフモデル」というモデル化手法である。隠れマルコフモデルとは、ある文字列に続く直後の文字の、出現のしやすさをパターン化し、確率を定義するものである。  
これらの4つの工程を経て、人が発した音声が、文字データとして認識される。  
・音声認識の今後の期待  
上記した音声認識の技術を活用した、サービスや業務改善はますます進んでいくと考えられる。近年、音声認識の技術を用いた、ニュースなどの音声を文字に変換し、トピックを表す言葉や語句を自動的に抽出したり、ほしいニュースを音声で検索したりする研究が行われている。また、音声、音楽、雑音などを自動的に判別する研究も行われており、この技術と、自然言語理解、画像処理、映像処理、情報検索などのあらゆる分野との技術の融合によって、さらなるサービスの活性化が期待できる。

・音声認識の今後の課題  
以上で上げたとおり、音声認識の活躍の場は広がりつつあるが、今現在では、あらゆる人が、あらゆる環境で、任意の話題をコンピュータに認識してもらうのは、実現できているとは言い難い。あらゆる人という点においては、複数人のユーザや、多言語の問題や、声の大小高低などが挙げられる。あらゆる環境という点においては、雑音やノイズが多く入ってしまうと、認識の精度は急激に落ちてしまう。この課題に対しては、周囲の雑音、電話音声などの歪み、入力機器との距離などさまざまな環境に対応する自動適応機能の研究が必要となってくる。任意の話題という点においては、方言やスラング、話し言葉などが挙げられる。この課題に対しては、大規模な方言、スラング、話し言葉などのそれぞれのデータベースを構築し、これらに適する言語モデルを組み立てる必要がある。