非負値行列因子分解を用いた言語識別技術、音声入力・制御に用いられる音声認識技術

髙木 研究室

髙木 一幸
Kazuyuki TAKAGI

研究概要

音声認識技術の発達と普及

当研究室では、機械による音声言語の処理と認識について研究している。音声認識とは、人間が発した音声を、文字化したり、何を話しているかを判断することを言う。

音声認識技術は、ここ数年で私たちの生活に組み込まれることが多くなった。例えば、検索エンジンのGoogleでは音声認識機能を搭載し、スマートフォンなどでも音声入力ができるようになった。また、ルームエアコンやLED照明、テレビなど家電製品を音声でコントロールできる音声リモコンも市販されている。音声はもはや人間同士のコミュニケーションだけでなく、人間と機械のコミュニケーションにも利用されるようになってきた。

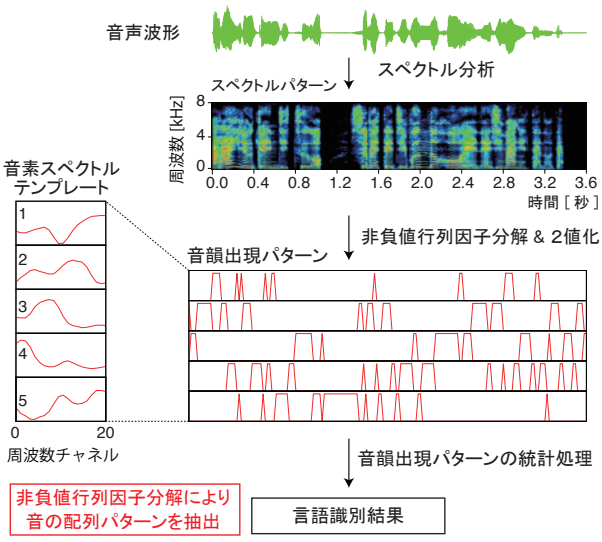
言語識別

特に力を入れているのは言語識別だ。これは未知の話者がどこの国の言葉で話しているかをコンピュータが自動判別するものだ。

アメリカなどの多民族国家では、この言語認識技術の利用ニーズが高い。例えば消防署や警察などへ緊急電話する場合、通常は英語を使ってしまうケースがある。そのため、アメリカではいろいろな言語を話せるスタッフを用意しているが、電話で何語かを判断することは難しく、その判断の遅れが大きな問題になることもある。そのとき言語認識技術が利用できれば、的確な担当者へ迅速に電話を回すことができるのだ。

このように、自動言語認識技術を応用したサービスへのニーズが高まっている。近い将来、アジア地域を含め、全世界的にこの技術が必要になるはずだ。

言語の違いは、音素の種類、単語の音素配列パターン、韻律パターン、語彙などにあらわれるので、これらの違いに関する音声の情報を利用し判断を行う。例えば日本語なら、母音、子音、母音、子音、母音、子音という順番に並ぶ特徴があり、子音、子音、子音、母音という配列はまずありえない。この音素配列情報を機械的に取り込んで、統計を取って、その言語の特徴をモデル化する。そのモデルと入力される音声を比較し言語を特定する。



非負値行列因子分解(NMF)による特徴抽出を用いた言語識別

キーワード

言語識別、音素配列情報、非負値行列因子分解、音声認識、耐雑音性、マルチバンド・マルチSNR・マルチパス音声認識、パターン認識、機械学習

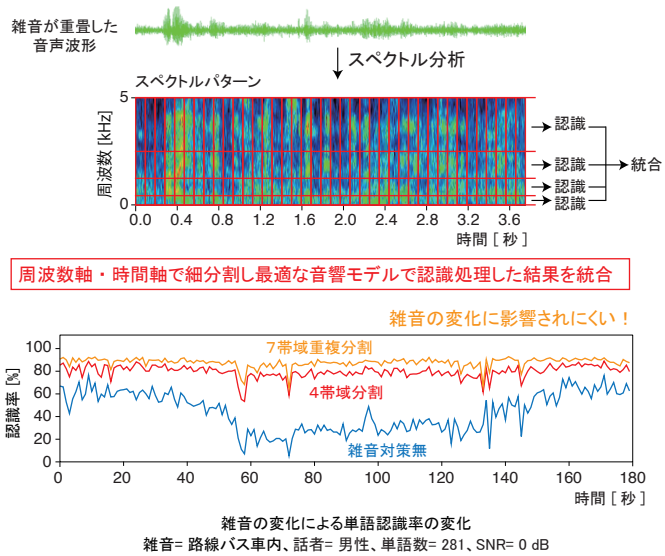
所　属	大学院情報理工学研究科 情報学専攻
メンバー	髙木 一幸 助教
所属学会	電子情報通信学会、日本音響学会、情報処理学会、人工知能学会、日本音声学会、IEEE、ISCA
E-mail	takagi@uec.ac.jp
研究設備	音声言語メディア情報処理システム

アドバンテージ

非負値行列因子分解を利用した言語識別技術

当研究室で開発した言語認識技術のアドバンテージは、音素の種類、音素配列の特徴量を抽出する際に、非負値行列因子分解（Non-negative Matrix Factorization：NMF）を利用していることだ。NMFは画像解析や楽器音から自動採譜する際に利用されている手法で、音声にこの手法を利用した場合、あらかじめ音素のスペクトル特性を学習しておけば、時系列でどのような音素が出ているかが分かる。この手法で音を拾っていくと、言語音の特徴を解析することができる。

ここで抽出した音素と出現情報は、言語モデルとして用いられるN-gram（n個の要素の並びの種類と頻度の統計）を利用してモデル化し、最後にモデルのパラメータを特徴ベクトルとしてサポートベクターマシンなどの学習機械を使って言語識別を行う。このような言語分析にNMFを使っている技術はほとんど無く、当研究室のオリジナルな手法と技術だ。



マルチバンド・マルチSNR・マルチパス音声認識方式は雑音の変化に頑健

NMFで言語認識率98・6％を実現

NMFを使うようになったきっかけは、2008～2010年、音声工学や言語学の研究者と多言語音声処理の共同研究をしていた際の経験に基づいている。共同研究で得たさまざまな知見から、この手法が活用できることを考えついた。

この手法を使って日本語と英語の識別を行った結果、音素スペクトルテンプレートの数とN-gramの次数を最適に選んだ場合、98・6％の認識率が実現した。当研究室では、これに加えて約20言語のデータベースを既に所有しているので、今後は順次各言語に適用していきたいと考えている。

環境雑音に強い音声認識

環境雑音に強い音声認識についても研究している。実際に音声認識を利用する場面では、さまざまな雑音があるが、それらの雑音により音声認識の精度が落ちてしまわないようにする必要がある。

当研究室では、電通大の尾関和彦研究室で研究されていたマルチバンド・マルチSNR・マルチパス音声認識方式を用いて、雑音が多い環境でも、高い認識率で単語音声を認識できる方法について研究を続けている。

マルチバンド・マルチSNR・マルチパス音声認識では、周波数の帯域ごとに音声認識を行う。例えば、雑音が低い音の場合、低い音の音声認識は難しいが、高い音は確実に音声認識ができる。周波数帯域ごとの音声認識では、学習用音声と学習用雑音をさまざまな割合（SNR）で混合して作成した音響モデルを用いる。実際の雑音は多種多様で、常に変化している。各周波数帯ごとに、入力音声に最も良くマッチする音響モデルを適宜選択して音声認識を行い、それらを組み合わせることにより、総合的に高い確率で音声認識が可能になる。

音声言語メディア情報処理システムのUNIXサーバ

今後の展開

翻訳ロボットの頭脳を作りたい

単語認識に関しては、長年の研究結果から、より精度を上げるには、周波数帯の分割方法をさらに改良し、音響モデルを作成する際に用いるモデル雑音の特性を工夫すればよいことがわかっているので、その方向で研究を続けていきたい。

言語識別に関しては、2011年に良い結果が上がった。当研究室の言語認識技術は、他に類を見ないユニークな方法を使っているので、世界標準の評価データを用いてどこまで精度が上げられるかが楽しみだ。

また、将来的に映画に出てくる通訳ロボットの頭脳を実際に作ってみたい。その中でも、多言語識別に重きを置き、何語かを判別するだけでなく、各国の地方言語、方言までも識別させることが夢であり目標だ。