〈学位論文の紹介〉

統計的音素セグメントモデルによる日本語音声認識に関する研究

相澤　桂

古市 千枝子（監修）
桐蔭横浜大学工学部制御システム工学科

相澤桂君は、本学大学院工学研究科制御システム工学専攻博士後期課程を修了、表記題目の研究により本年3月工学博士の学位を得た。以下はその博士論文の内容を紹介するものである。

人間は、お互いに言葉を交わすことによってコミュニケーションをとることができる。特に言葉を音声で表現する場合は、「話す」、「聞く」という動作がコミュニケーションの手段になる。我々はこれらの能力を、生まれてからの学習や経験の積み重ねによって自然に獲得することができる。一方、機械や電子機器とのコミュニケーションでは、ボタンやスイッチなどの入力装置を用いて機械・電子機器へ命令を伝え、ランプやディスプレイなどの出力装置を用いて結果や情報を得る。このような手段には人間的な温かみが感じられない。また、命令を入力するために複雑な操作を強いられたり、結果表示が分かり難かったりすることも多く、操作になかなか馴染めない人々も少なくない。

そこで、機械や電子機器と人間の接点に、音声を用いたヒューマン・インターフェースを用いることが考えられる。近年、音声認識機能や音声出力機能を持った電子機器やパソコンソフトが少しずつ実用化されてきており、操作に慣れていない人でも、機器が簡単に扱えるようになってきた。音声出力は、決まった単語や文章の読み上げであれば、あらかじめサンプリングした人間の音声波形データを再生して用いる。この方法は音質も良く、声質もナチュラルである。また、任意のテキスト文章の読み上げには合成音声が用いられるが、こちらも技術が進歩し、より人間の音声に近くなってきた。一方音声認識は、用途によってはまだ十分な性能を持っているとは言えない状況である。音声認識とは、あらかじめ学習用音声信号から音素や単語などの音響モデルを作成し、入力音声信号の音響的特徴との類似度を計算して認識候補を得るものである。しかし、言葉を構成している音素の違いや、それを発声する話者の声質の違いによって、音声の音響的な特徴が多様に変化するため、ひとつの音響モデルで表す音素や単語の発声話者数が多くなると、より多くのモデルパラメータを必要とし、認識精度の低下と認識演算量の増加を招く。従って、特定の話者が発声する小語いの単語音声を対象とする場合は、比較的高い認識率が得られるが、不特定の話者による大語いの連続音声を認識する場合は、実用的といえる認識性能がなかなか得られない。

現在、連続音声の認識を行うパソコン用のソフトウェアがいくつか実用化されている。このようなソフトウェアは、あらかじめ大量の音声資料を学習に用いて、不特定話者の大語い連続音声に対応するように作られている。しかし、使用者の音声は学習されていないので、一般にそのままでは十分な認識精度が得られない。そこで、あらかじめ使用者の音声を用いて、話者適応という作業を行う。あるソフトウェアでは、あらかじめ決められた300の文章を正確に、かつ自然に使用者に発声してもらい、この音声データをもとに、システムを使用者の音声に適応させる処理を行う。この作業には、音声の入力だけで1時間近くを要し、さらにその後のシステムの話者適応の演算に数十分から1時間近くを要する。適応後のシステムでは90%近い認識率を得ることができるが、このような適応作業は、使用者にとって

かなりの負担である。

話者適応を行わなくても良い認識率が得られるような不特定話者連続音声認識システムの実現を目指すには、音響モデルの精度向上が重要である。すなわち、多数話者の音声信号を、なるべく少ないパラメータで効率的に学習できることが望ましい。現在、音声の音響的特徴は、静的な特徴だけでなく動的な特徴も考慮してモデル化する方が、高い認識率が得られることが分かってきており、なるべく少ないパラメータで音声の動的・静的特徴をモデル化する必要がある。

本研究では、多数話者音声を効率的に学習できる音素モデルとして、音素の音響的特徴の確率密度分布に混合ガウス分布を仮定した統計的音素セグメントモデル（Stochastic Phonemic Segment Model; SPSM）を用いる日本語音声認識システムを提案している。セグメントモデルとは、音声の音響的特徴を、ある時間区間から複数サンプル抽出して、まとめて学習したモデルのことである。すなわち、音声の静的な音響的特徴とその時間変化の軌跡をモデル化したものである。

提案法の特徴は、音声信号を音素モデルで認識する前に、音素セグメンテーションという、発声されている音素の境界検出を行う点と、単語の音響モデルを用いずに、スコア付きの音素認識候補と音素記号列で表された単語辞書項目との記号列マッチングによる単語認識を行う点にある。音素の認識問題は、音声信号のどこで音素が発声されているかを調べる探索問題と、何の音素を発声しているかを調べる識別問題を含むが、音素セグメンテーションを用いることによって、認識部における音素の不要な探索を行わずに済む利点を持つ。さらに、音素セグメントが求まることによって、各音素に最も適した音素特徴量の抽出を行うことができる。また、単語の音響モデルを使わない方法の利点は、語いの増加に対して音素記号列で表した単語辞書項目を追加するだけで対応でき、新しく音響モデルを学習する必要がないという点にある。

提案法では、音素の種類に応じて8種類の型の音素特徴量を抽出しモデル化する。音素特徴量は、5つの特徴ベクトルで構成される。図1は、男性の/ oreseNgurafu /という単語音声を音素セグメンテーションし、特徴抽出を行った例である。図中の縦線が音素境界である。1段目が音声波形を表し、その上部の英字が音素の種類（母音、子音など）を表す記号、下部の英字が音素記号である。2段目のパラメータdは、音声信号の周波数スペクトルの時間変化の度合いを表している。以下、8種類の型の音素特徴量について、その抽出位置を五叉上記号で表している。このようにして抽出された音素特徴量を各型、各音素毎にまとめ、音素特徴量を構成する各特徴ベクトルの番号毎に、その特徴分布に混合ガウス分布を仮定してパラメータを推定したものがSPSMである。図2はSPSMの構造を示している。

従来のテンプレートマッチング法による音声認識システムでは、音素特徴量を音素標準パターンとして全て保存するため、学習データ量を増加するとテンプレート数が膨大になり、認識時の演算量の増加を招いてしまう欠点があったが、提案法では、認識率を低下させることなく、モデルの記憶量、および認識時の演算量とも従来法より大幅に圧縮できることが確認されている。

本論文では、最後にシステム全体の評価実験として、男性話者10名による音韻バランス単語セットを一括学習したSPSMを用いて、非学習男性話者63名の未学習語い6,708単語を認識する実験を行い、従来法と同程度の93.5%の認識率が得られたことを示している。本手法は、多数話者音声の効率的なモデル化手法として有効であり、今後の不特定話者連続音声認識システムの性能向上に大きく貢献するものと思われる。

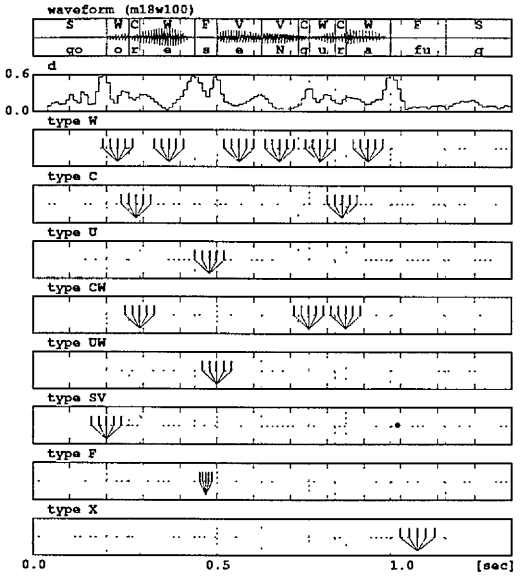

図１：音素特徴量の抽出例：男性の音声／oreseNgurafu／の音素セグメンテーション結果、メル対数スペクトル包絡の時間変化量 di と、各型の音素特徴量の抽出位置

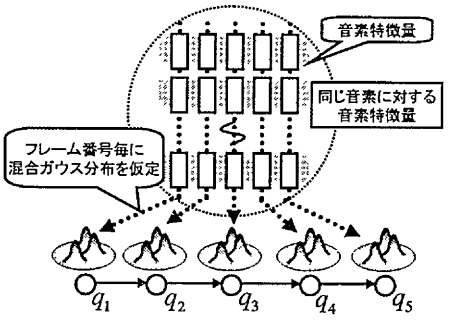

図２：統計的音素セグメントモデル(SPSM)の構造