論文内容の要旨

論文題目　音声の構造的表象に基づく単語音声認識に関する研究

氏名　　朝川 智

音声認識システムは進歩を遂げ，現在では携帯電話やカーナビゲーションシステムなどに搭載されるまでに至った．しかしながら，システムの想定する環境下や非常に整った条件の下では非常に高い認識率を示すが，多様な認識タスク，多様な環境の下では認識性能は劣化する．そして，人間による音声認識能力と比較した場合，その性能には未だ遠く及ばないのが現状である．仮に人間が生涯耳にするデータを認識システムの学習に用いたとしても，現在の音声認識技術では人間の認識性能には遠く及ばないであろうとの予測もなされている．その要因の一つとして挙げられるのが，音声に混入する非言語的特徴の存在である．音声は様々な話者により様々な環境下で発声され，様々な伝送経路を通り，様々な音響機器により収録される．人間が聴取をする場合は，様々な聴覚特性を持つ話者により聴取される．これらのプロセスの中で，音声の物理的実体は様々な形で変形され，例え同じ言語的な情報を持った発話であったとしても，非言語的特徴による変動によって様々に変化する．これらの変動は，人間同士がコミュニケーションを行う際においても，音声認識システムがその発話を認識する際にも，不可避的に混入する変動である．人間はこれらの多様な変動が含まれた音声に対しても非常に頑健に言語的な情報の抽出を行うことができる一方で，音声認識システムの性能はこれら変動に多大な影響を受ける．

従来の音声認識技術は，この問題を解決するために多くの話者の多様な環境による音声データを集め，これらのデータを用いて統計的手法を駆使することにより，不特定話者音響モデルを構築して音声認識に用いてきた．しかしながら，この不特定話者音響モデルを用いても性能を劣化させる話者が必ず存在する．そこで，学習したモデルを入力話者に近づける適応，あるいは入力音声の話者性を変形することにより一定に近づける正規化に関する研究がこれまで様々な形で行われてきた．しかし，これらの何れの技術も，あくまでモデルと入力音声との特性を近づけるだけに過ぎない．そもそも言語的な情報と同時に非言語的特徴までもが含まれる物理的実体に対して，その絶対量を直接的に用いて音声を表現していることがミスマッチを生む根本的な原因であるといえる．その一方で，人間は非常に偏った音声提示環境の下で，様々な音響的な変動に対する対処法を獲得する．例えば，幼児の聞く音声の大部分は両親の声である．更には，対話が自分と相手との音声コミュニケーションで成立することを考えると，人の聞く声の約半分は自分の声であり，偏った音

声提示環境が一生続くことが分かる．これは不特定話者音響モデルが構築される何百，何千という学習話者環境とは相反するものである．それにも関わらず音声は人間にとって一番楽なコミュニケーションメディアの一つである．

近年，上記の非言語的特徴を表現する次元そのものを保有しない音響的普遍構造が提案された．これは音声の物理的実体を捨象し，相対関係のみをとらえることによって得られる音声の構造的表象である．このような非言語的特徴による影響を受けることのない音声の物理的表象に基づいて人間がコミュニケーションを行っていることが知覚実験によって示唆されている．そして，音声の構造的表象を用いることで非言語的特徴の違いに対して頑健な音声アプリケーションが可能となると考えられる．本論文では，この音声の構造的表象を用いて，非言語的特徴の違いに対して頑健な単語音声認識の枠組みを構築し，種々の認識実験によりその有効性を検証した．

本論文では，単語音声認識を対象として，一般的で実用的な認識タスクに適用可能な構造的表象に基づく単語音声認識の枠組みを提案した．構造的表象を単語音声認識へと適用することを考えた場合，話者性を効果的に消失させる一方で，その強すぎる不変性のために全く異なる単語が同一と見なされてしまう可能性がある．単語発話を表現する構造間で比較を行う際に不適切な構造間照合を抑制するような制約が必要となる．そこで，特徴量を分割することにより，許容される変換に対して制約を施す手法であるマルチストリーム構造化を提案し，このマルチストリーム構造化に基づく構造的単語音声認識の枠組みを提案した．そして，日本語 5 母音連続発声系列と子音を含む単語音声データベースである東北大・松下単語音声データベースの 2 種類の認識タスクに対して単語認識実験を行い，提案手法の有効性を実験的に検証した．本手法は，音そのものはすべて捨てているため，各音素を個別に認識することは不可能である．しかし，単語発話全体をとらえたときに，音のスカラー差，言い換えると発話内での音の動きの情報のみからその単語を識別することが可能であることが実験的に示された．

前述の構造的単語音声認識の枠組みにおいて，最終的な認識結果を出力する識別器の部分は非常に単純であり，性能向上の余地が大いに残されていた．そこで，線形判別分析に基づく高精度な識別器を提案し，更なる性能向上を図った．マルチストリーム構造化を導入することにより，各発話を表現する特徴量ベクトルは非常に高次元なものとなるが，特徴量の高次元性は計算コストの問題だけでなく，識別性能にも多大な影響を与える．提案手法では，線形判別分析を段階的に適用することにより，低次元かつ識別的な特徴量へと変換することが可能となる．日本語 5 母音系列及び子音を含む単語に対して認識実験を行い，線形判別分析に基づく識別器を用いることで大幅な認識率向上が得られることを確認した．また，これまでの認識実験においては冗長なパラメータとして用いていなかったデルタケプストラムに対して，線形判別分析を適用することによって冗長性を削除し，識別的な特徴量を抽出することが可能となる．認識実験の結果から，線形判別分析を適用することにより，単語認識においてデルタパラメータを有効に利用することが可能であること

が確認できた．更に，特徴量を分割した時点で別個のものとして扱っていた各ストリームに対して，その相互関係を考慮するためにストリーム間距離という特徴量を導入し，更なる認識率の向上を図った．最終的に，日本語 5 母音系列においては従来手法である単語 HMM による認識率を超える性能を示し，子音を含む単語音声の認識では，従来手法には若干及ばないもののほぼ同等の認識性能を示した．

音声の構造的表象は非言語的特徴の違いを消失させた音声の表象手法であり，それらの違いに頑健な音響的照合が可能であることが従来手法に対して優位性を持つ点である．しかし，ここまで行ってきた実験では，日本人成人男女複数名による学習データを用いて，同様の特性の話者による評価データで認識を行っていた．このような認識タスクでは非言語的特徴のミスマッチは小さく，従来の手法でもかなり高い認識性能を示し，構造的表象の有効性を完全には検証できていない．そこで，ケプストラムドメインでの線形変換により多様な話者性を人工的に生成し，意図的に非言語的特徴のミスマッチを生じさせた条件下での認識実験を行い，提案手法の話者性の違いに対する頑健性を実験的に検証した．結果として，従来の音響的実体に基づく手法では声道長が極端に異なる話者の音声に対しては認識率が大きく下がるのに対して，提案手法では非常に幅広い話者性に対して頑健な認識が可能であることが実験的に確認された．

以上論じたように，音声の構造的表象を用いた単語音声認識の枠組みを提案し，認識実験の結果より非言語的特徴に対する頑健性が確かめられた．本表象は，音そのものをとらえる従来の方法論とは全く異なる観点から，音の差に基づいて音声を記述することにより得られる話者不変の音響モデリングである．本論文により，従来とは全く異なる新しい音響モデリングでの音声認識が実現可能であることを確認できた．