氏　　名	澤田　直輝
博士の専攻分野の名称	博士（工学）
学 位 記 番 号	医工農博甲第１２号
学 位 授 与 年 月 日	平成３１年３月２０日
学 位 授 与 の 要 件	学位規則第４条第１項該当
専　　攻　　名	情報機能システム工学専攻
学 位 論 文 題 目	高精度な音声認識結果の推定技術とその応用に関する研究
論 文 審 査 委 員	主査　教　授　鈴木　良弥
	教　授　大渕　竜太郎
	教　授　宗久　知男
	教　授　福本　文代
	准教授　渡辺　喜道
	准教授　丹沢　勉
	准教授　西﨑　博光

学位論文内容の要旨

近年，動画コンテンツに代表されるマルチメディアコンテンツが充実してきた．その理由に，Graphics Processing Unit（GPU）の発達による個人のマルチメディアデータの生成・編集の普及，Hard Disk Drive（HDD）や Solid State Drive（SSD）などのストレージの大容量化, YouTube や Twitch などに代表されるマルチメディアコンテンツの配信サイトの増加などが挙げられる．また，ビジネスにおいても，会議や講演や，病院のカルテの作成などに，映像の録画や音声の録音を用いることも増えてきている．このようなコンテンツが増加してきている背景には, ネットワークインフラの充実に加えて, タブレット PC, スマートフォンの普及により，容易にマルチメディアコンテンツにアクセスすることができるようになったことが挙げられる．

また近年では，深層学習技術が発展してきており，画像処理の分野や音声処理の分野で様々な手法が提案されている．このような深層学習技術の発展には，ハードウェアの発展と対応したコンテンツが充実してきたことが挙げられる．

音声処理の分野では，大量の音データを効率良く扱うために音声の内容から人間が求めているコンテンツを検索する技術や，人間が話している内容を理解して会話を行う音声対話システムなどの技術が提案されている．これらの技術の基盤技術として音声認識技術が存在する．音声認識技術は，音声データに対して発話内容を文字列として書き起こす技術である．音声認識技術を活用することで，発話内容を理解することや特定の内容を探すことが容易になる．

このような音声認識技術を活用する場合に，正しく音声認識できていない認識誤り箇所が様々な悪影響を及ぼす．このような認識誤りに対して，様々な認識誤り対策が考えられている．例えば音声検索では，単語より小さい音の単位であるサブワード単位で検索する方法や，認識結果の出現確率を用いて検索する方法などが用いられている．

また，音声認識技術においても，様々な新しい手法が提案されており，音声認識性能の改善が図られている．例えば音声認識システムにおいて，深層学習を用いて高精度な音素識別を導入することで認識性能の改善が試みられている．

しかし，音声認識システムは通常単語を認識する必要があるため，音声認識システムで学習されていない単語（未知語）を正しく認識することは困難である．また，認識誤りした結果は音声認識を用いた技術において性能低下の要因となってしまう．そこで，後処理で認識誤りを含む認識結果を正しいサブワード系列に変換することが有用であることが明らかとなっている．さらに，先行研究では，文字列の表現を変更した複数の音声認識システムを用いることが，検索技術において認識誤りに頑健であることが示された．そこで，複数の認識システムを用いることで認識誤りを修正することができる高精度な正解音識別が可能であると考えられる．

本研究での目的は，音声からの検索などの応用技術に利用することが可能であり，様々な音声認識技術にも適用することが可能な方法で音声認識性能を改善することである．具体的には，音声認識結果のサブワード系列を入力情報として，正しいサブワード系列を推定することによりそのまま応用技術に入力することが可能な出力を獲得することが実現する．本研究では，このような音声認識結果から正しいサブワード系列を獲得する高精度な認識結果推定器を提案する．さらに，この認識結果推定器が応用技術に有用であるか検証するために，検索実験と単語変換実験を行い性能の調査を行った．

本研究ではサブワード単位として音声の最小単位である音素を用いて高精度な認識結果を推定する．認識結果を推定するために深層学習技術を利用して正しい音素列の推定を行う．具体的には，複数の音声認識結果を音素列に変換し，各音素列を時間情報に基づいてアライメントを行う．このアライメントされた結果に対して深層学習技術を適用して各アライメント区間に対して正しい音素列の推定を行う．このような正解音素推定器を用いて，正しい音素列を推定することにより，元の音声認識列や複数の認識結果を多数決により統合した方法よりも，高い性能が得られた．このような結果から，正解音素推定器により高精度な認識結果を生成できる可能性が示された．

しかし，音声は，時系列を持ったデータであり，音と音が時間の軸で組み合わさることで意味が付いてくる．また，時間方向において制約も存在する．例えば，日本語において子音の後に母音が現れることは絶対であり，子音の後に子音が現れることはない．しかし，時間情報を用いなければこのような制約を考慮することができない．そのため，単純な正解音素推定器では，このような間違いが出現する可能性がある．このことから，正解音素推定器に対して時間情報を付与することは高精度な推定には必要である．

そこで，時間情報を考慮することができる深層学習技術を利用する．時間情報を考慮した正解音素推定器は，考慮していない正解音素推定器と比較して高い性能が示された．このことから，正解音素推定には時間情報が有用であることが示された．

正解音素推定器は，認識結果の誤りを減少させた高精度な認識結果を生成することができるため，音声認識を用いた様々なアプリケーションに適用可能である．そこで，正解音素推定の性能が音声認識を用いたアプリケーションにおいてどのような影響を与えるか調査した．

まず，音声の検索技術である音声中の検索語検出の技術に適用した．音声中の検索語検出は，音声中に存在する目的の単語を探し出す技術である．この技術は一般的に音声認識結果を利用するため，認識精度が検索精度に影響されやすい．そこで，正解音素推定結果から検索することで，高精度な音素推定性能が応用技術に好影響を与えるか調査した．結果として，推定性能が高い結果から検索するほど検索性能も改善の傾向が見られた．このことから，正解音素推定器が応用技術に適用することで，応用技術の性能を改善できる技術であることが示された．

しかし，実際に応用技術に音声認識結果を利用する場合，音声認識結果として単語列を入力することが一般的である．そのため，正解音素推定した結果を単語列に変換して認識結果の単語列の精度を改善することで，どのような応用技術にも適用可能にすることができる．そこで，認識結果の単語列の誤っている単語を，正解音素推定器で変換した単語で置換することで誤りが少ない単語列を生成する．これは，認識結果において誤り単語箇所を検出し，誤り単語箇所に正解音素推定結果から変換して得られた単語を置換することで性能を改善させる．認識結果に対して適応することで，単語の認識結果が改善することが示された．このことから，正解音素推定器を様々な応用技術に適用することが可能であると考えられる．

審査結果の要旨

本論文で解決しようとしている課題は，音声からの検索などの応用技術に利用することが可能であり，様々な音声認識技術にも適用することが可能な方法で音声認識性能を改善することである．

近年，動画コンテンツに代表されるマルチメディアコンテンツが充実してきた．その理由に，個人のマルチメディアデータの生成・編集の普及，データ記憶装置の大容量化，マルチメディアコンテンツの配信サイトの増加などが挙げられる．また，ビジネスにおいても，会議や講演や，病院のカルテの作成などに，映像の録画や音声の録音を用いることも増えてきている．このようなコンテンツが増加してきている背景には，ネットワークインフラの充実に加えて，タブレット PC，スマートフォンの普及により，容易にマルチメディ

アコンテンツにアクセスすることができるようになったことが挙げられる．このようなマルチメディアコンテンツが増えることにより，大量のコンテンツから必要な部分を検索する技術が不可欠になってきており，その問題を解決することは社会的・学術的意義が大きい．

音声処理の分野では，大量の音データを効率良く扱うために音声の内容から人間が求めているコンテンツを検索する技術や，人間が話している内容を理解して会話を行う音声対話システムなどの技術が提案されている．これらの技術の基盤技術として音声認識技術が存在する．音声認識技術は，音声データに対して発話内容を文字列として書き起こす技術である．音声認識技術を活用することで，発話内容を理解することや特定の内容を探すことが容易になる．

音声認識技術を活用する場合に，正しく音声認識できていない認識誤り箇所が様々な悪影響を及ぼす．このような認識誤りに対して，様々な認識誤り対策が考えられている．例えば音声検索では，単語より小さい音の単位であるサブワード単位で検索する方法や，認識結果の出現確率を用いて検索する方法などが用いられている．また，音声認識技術においても，様々な新しい手法が提案されており，音声認識性能の改善が図られている．例えば音声認識システムにおいて，深層学習を用いて高精度な音素識別を導入することで認識性能の改善が試みられている．

本研究ではサブワード単位として音声の最小単位である音素を用いて高精度な認識結果を推定している．認識結果を推定するために深層学習技術を利用して正しい音素列の推定を行っている．具体的には，複数の音声認識結果を音素列に変換し，各音素列を時間情報に基づいてアライメントを行い，このアライメントされた結果に対して深層学習技術を適用して各アライメント区間に対して正しい音素列の推定を行う．このような正解音素推定器を用いて，正しい音素列を推定することにより，元の音声認識列や複数の認識結果を多数決により統合した方法よりも，高い性能が得られた．このような結果から，正解音素推定器により高精度な認識結果を生成できる可能性を示すことができた．

音声は，時系列を持ったデータであり，音と音が時間の軸で組み合わさることで意味が付いてくる．また，時間方向において制約も存在する．そのため，時間情報を用いない単純な正解音素推定器では，このような間違いが出現する可能性がある．このことから，正解音素推定器に対して時間情報を付与することは高精度な推定には必要である．

そこで，本研究では時間情報を考慮することができる深層学習技術を利用している．時間情報を考慮した正解音素推定器は，考慮していない正解音素推定器と比較して高い性能であることを示すことができ，正解音素推定には時間情報が有用であることを示すことができた．

以上より，本論文は今後爆発的に増加することが見込まれるマルチメディアコンテンツなどの音声検索に利用可能な高性能な正解音素推定器の提案し，その優位性を音素認識実験により示している．また提案した正解音素推定器を検索語検出に適用する実験を行い，

提案手法の優位性を示している.

博士論文で取り組んだ課題の重要性，提案した手法の斬新さ，手法の有用性を示す緻密な実験と考察，最終試験でのプレゼンテーションの構成，質疑応答の的確さを鑑み，本論文は，博士論文としてふさわしい研究であると判断した.