| | |
|---|---|
| 氏名（本籍地） | 大河（おおかわ） 雄一（ゆういち） |
| 学位の種類 | 博　士（情報科学） |
| 学位記番号 | 情　博　第　343　号 |
| 学位授与年月日 | 平成18年3月24日 |
| 学位授与の要件 | 学位規則第4条第1項該当 |
| 研究科、専攻 | 東北大学大学院情報科学研究科（博士課程）情報基礎科学専攻 |
| 学位論文題目 | 自然発話音声を対象とした高精度音声認識に関する研究 |
| 論文審査委員 | （主査）東北大学教授　牧野　正三（工学研究科）<br>東北大学教授　鈴木　陽一　東北大学教授　木下　哲男<br>東北大学助教授　伊藤　彰則（工学研究科） |

## 論文内容の要旨

### 第1章　序　論

近年の社会・産業の発展の伴い，機械の操作などは複雑になり続けている。そのため，音声言語の優れた特徴をユーザーインターフェースとして用いたいという声は非常に大きい。そのため近年，音声認識技術が注目されている。音声認識を多くの分野に応用するには，自然発話音声又は自由発話音声などと呼ばれる人が人に対して行う自発的な会話を認識する必要があると考えられる。しかし一般に，この自然発話音声を対象とした音声認識の性能は十分なレベルには達していない。本論文は，これに対し2つのアプローチから，自然発話音声認識の精度改善への取り組みを行うものである。

その1つは，発話速度の影響により音素の音響的特徴が変化するという問題に対する取り組みである。発声された音声に含まれる各音素は，その前後の音素の影響を受け，音響的特徴が変化する。また，この影響の強さは音素の持続時間に大きく影響されるものである。過去の研究により自然発話音声では発話速度が速いだけでなく，発話速度の変動も大きいことが明らかになっており，自然発話音声にはこの現象が強く現れる発話と，あまり強く現れない発話の双方が存在することとなる。その結果，従来の音響モデル学習では，自然発話音声をモデル化する際に異なる2つの音響的特徴が1つのモデルに学習されることになっていた。本論文では，この問題に対し，音素の種類や持続時間の長さにより異なる音響的特徴の違いを高精度にモデル化することを目的に検討を行った。もう1つは，自然発話音声認識での認識誤りのうち，極端に持続時間の異なる誤りに対する対処である。自然発話音声は，先に上げた発話速度による音響的特徴の変化以外にも，発話スタイルの違いなど多様な理由により，同じ音素であっても多くのバリエーションが存在することになる。そのため，自然発話音声で音響モデルを学習した場合，音響的特徴の違いに対する許容範囲の大きな分布が得られることになる。このモデルを使用して音声認識を行うと，音響的に少々おかしくとも，間違ってマッチしてしまうことがある。そういった中に，極端に長い区間や短い区間にマッチした認識誤りが含まれる場合があり問題となっていた。そこで本論文では，音声合成の分野における知見なども参考とし，言語的情報を用いた音素の持続時間予測法の検討を行う。その上で，予測された持続時間を用いて音声認識の極端な誤りを排除することで，自然発話音声の高精度な認識を行うことをもう1つの目的とする。

## 第２章　マルチパス HMM による自然発話音声モデル化法

自然発話音声では，発話速度により音素の音響的特徴が変化し，音声認識の性能の悪化が引き起こされていた。そこで本章では，音素の持続時間が短いものを別にモデル化する方法について検討を行った。

本論文では，自然発話音声の音響的特徴をモデル化するためマルチパス HMM を利用した。その上で，従来法で問題とされていた各パスの学習サンプルの決定方法の欠如について改善するため，音素の種類ごとに異なる最適な持続時間（分割しきい値）により音声サンプルを２つに分け，それぞれのサンプルを用いて短時間モデルと長時間モデルを持つマルチパス HMM を学習する方法について検討を行った。この時，各パスの学習サンプル決定する必要が有るが，本論文では準最適な分割しきい値を現実的な計算量で求める方法としてオールスターモデル選択法を提案した。

オールスターモデル選択法では，あらかじめ異なる設定を持つ複数の音素モデルセット（種モデル）を学習する。次に学習された種モデルを用いて，モデルの性能を表す評価値を音素種類ごとに求める。その上で，ある音素についてすべての種モデルの中で最も高い評価値を与えたものをその音素のモデルとする。これらの選択された音素モデルを集めることにより，音素モデルセットを作成することができる。

ここで，いくつかの異なる分割しきい値（初期しきい値）を設定し，それぞれのしきい値に対応したマルチパス HMM を学習しておいて，それらを種モデルとしてオールスターモデル選択法を実行すれば，音素ごとに妥当な分割しきい値を選ぶことができる。このモデルを特に，OMD (Optimized Multi Duration) モデルと呼ぶ。なお本論文では，評価値としてはフレームあたりの平均対数ゆう度を採用した。

一方，オールスターモデル選択法により選択された OMD モデルは，多数の種モデルの中から音素ごとにモデルを選択することにより得られたものである。しかし一般に，選択された後のモデルは学習用サンプルに対して最ゆうとはならない。このため，たとえ選択された OMD モデルが種モデルに比べて妥当な分割しきい値が選択されていたとしても，良い認識性能が得られない可能性が高い。そこで，選択された OMD モデルを再学習することで，オールスターモデル選択法により選択された音素ごとの分割しきい値を持つ最ゆうな音響モデルを学習した。選択後のモデルを再学習した結果，１パスの HMM に比べて 1.2%の音素認識精度の改善が得られた。また，全音素一定の固定しきい値を与えたマルチパス HMM である種モデルに対しても，0.3%～0.8%の音素認識精度の改善傾向が得られた。

## 第３章　反復による分割しきい値決定の高精度化

本章では，前章でオールスターモデル選択法により得られた音素ごとの分割しきい値が，他の音素に独立に決定できるという仮定付きのものであったことに着目した。実際には音響モデルの学習において，音素間の影響は避けられないものであり，持続時間しきい値の決定にも少なからず影響を与えると考えられる。そこでオールスターモデル選択法で得られた音素ごとの分割しきい値を出発点とし，これを初期しきい値として再度オールスターモデル選択法を適用することにより，他の音素からの影響を減らし，準最適なしきい値を得る方法の検討を行なった。この方法は，初期しきい値として与える音素ごとの分割しきい値が，最適なものか十分に最適値に近いものであるならば，それらが各音素の分割しきい値決定に与える影響も小さなものになるとの考えに基づく。本手法では，オールスターモデル選択法で得られた分割しきい値を，初期しきい値として再びオールスターモデル選択法による分割しきい値の決定を行うことで，分割しきい値の高精度化を目指す方法である。

ATR 多数話者音声データベースの模擬対話音声を対象に音素認識実験を行い，本提案法を評価した結果，6 回の繰り返しによって，１パスの HMM に比べて最大 5.4%の音素認識誤りの削減が得られた。この結果は，

従来の全音素一定の分割しきい値を与えたマルチパスHMMに対しても2.0%と，5%水準で有意に改善が得られ，本提案法の有効性が確かめられた。

また，日本語話し言葉コーパスの学会講演音声を対象とした音素認識実験においても，3回の繰り返しで，1パスのHMMに比べて最大9.2%の音素認識誤りの削減が得られた。また，同条件で従来法である奥田らのマルチパスHMMを学習し比較したところ，音素認識誤り削減率3.7%と1%有意で改善が得られた。

### 第4章　音素持続時間のモデルを用いた認識高精度化

自然発話音声の認識誤りには，音素の持続時間が本来の長さに比べ極端に異なったものが含まれる場合がある。本論文では，この音素の持続時間に関する知識を従来の認識法と合わせて利用することにより，持続時間が本来の長さに比べて極端に異なる認識誤りを削減することを目指した。また，高精度な持続時間のモデルとして，音素の持続時間に影響を与える2つの特徴をともに考慮した持続時間モデル化の方法を検討した。その目的のため本論文では，音声認識によりあらかじめ認識結果のN-best仮説を求め，リスコアリングを行うことにより持続時間モデルを利用した。一般に発話速度としては，単位時間あたりのモーラ数が用いられることが多い。しかし，モーラの単位に影響を及ぼす長母音・短母音の判別や促音の認識は比較的認識誤りが多いため，認識結果から求めるパラメータとしては安定性に欠ける。そのため本論文では，発話速度を表すパラメータとして平均モーラ長との相関が比較的高い母音音素の平均持続時間を用いた。この際，文全体での平均を行うのではなく，当該音素と母音との位置により重みを与え，文中での発話速度の変動を考慮した，局所平均母音長を用いた。また，発話速度による影響と他の要因による誤差を分けてモデル化するため，音素持続時間と局所平均母音長の2次元正規分布としてモデル化を行った。

一方，音素の文中での位置や品詞などの言語的特徴により異なる持続時間分布をモデル化する手法として，木構造クラスタリングを用いた。そして，リーフに割り当てられた持続時間分布が最小となるように，あらかじめ用意した質問に従い分割を行い，持続時間分布選択のための決定木を作成した。なお本論文では，クラスタリングによる過学習の問題を回避するため2つの分割停止条件を導入している。1つは，開発用サンプルに対して，対数ゆう度の改善が得られない場合である。もう1つは，学習用サンプルから各リーフに割り当てられたサンプルの数が，事前に設定した定数を下回った場合である。クラスタリングの過程で用いる質問には，146種類の質問を用いた。これらの質問は，様々な言語的特徴のうち，品詞，出現位置，単語長の3つの要素と，音響的特徴と発話速度の言語的特徴以外の2要素を考慮して決定したものである。

提案した持続時間のモデルは，オープンなテストセットの持続時間を予測する実験において，この予測方法は安定した予測性能が得られることを確認した。また，リスコアリングを行い，音素認識精度を求めた結果，持続時間スコアを使用することで有意に認識誤りを削減できることを確認した。

また，**第3章**で求めたマルチパスHMMを用いた実験でも，改善が得られ効果が確かめられた。

### 第5章　結　論

本論文では，自然発話音声を対象とした音声認識を行う上で問題となる，音素の持続時間によって音響的特徴が異なるという問題と，自然発話音響モデルを用いることで生じる，持続時間がおかしな極端な認識誤りの2つの問題に着目し，自然発話音声の認識性能の向上を目指した。そのため，音素ごとに準最適な持続時間の分割しきい値で学習用サンプルを分け学習したマルチパスHMMを用いることとし，準最適な分割しきい値の決定法の提案を行った。また，持続時間が極端に誤った認識結果を排除するため，言語的情報も考慮に入れた持続時間分布予測法並びにその利用法の提案を行った。そこ結果，本論文を通して，自然発話音声の特徴を考慮することにより，音声認識の性能を改善が得られることが確かめられた。

# 論文審査結果の要旨

近年，音声認識技術の発展に伴い，その対象は文章の読み上げ音声から講演や対話などの自然発話音声に変わってきつつある．これら自然発話音声は，これまでの認識対象に比べて認識が非常に困難であるという特徴がある．その一つの理由として，自然に発話された音声では発声速度のばらつきが大きく，従来の音響モデルでは表現しきれない変動が生じていることが挙げられる．著者は，自然発話音声の高精度な認識という困難な課題に対して，発声速度の変動に着目して音響モデルと認識方式を改良するという研究を続けてきた．本論文はこれらの成果を取りまとめたものであり，全編5章からなる．

第1章は序論である．

第2章では，発声速度による音素の特徴変動に着目し，これを音響モデルに反映させるための新しいアルゴリズム「オールスターモデル選択法」を提案している．この方法は，複数の音素HMMを並列に結合するマルチパスHMMを基本とし，音素ごとに最もゆう度の高くなるHMMの組み合わせを選択することによって準最適なモデルを構築するものである．これは，従来困難であった音素HMMでの発声速度による特徴変動のモデル化を可能とするものであり，重要な成果である．

第3章では，第2章で作成した音素HMMをさらに高精度化するために，オールスターモデル選択法によりモデルの構築と再学習を繰り返す手法を提案している．この手法により，第2章で構築した音素HMMおよび従来のマルチパスHMMと比較して有意な改善が得られた．これは実用に向けた音素HMMの高精度化手法として高く評価できる．

第4章では，音素持続時間をモデル化することにより，自然発話の音声認識結果を改善する手法を提案している．この手法は，決定木を使うことにより，ある音素の持続時間の分布に対する隣接音素や言語情報の影響をモデル化できるものである．また，この音素持続時間モデルを使い，音声認識システムから得られた複数候補をリスコアリングし，認識結果を改善した．この手法により，従来法に比べて4.9%の誤り改善が得られた．これは発声速度変動が大きい自然発話の認識において重要な成果である．

第5章は結論である．

以上要するに本論文は，自然発話音声の認識という問題に関して，発声速度変動という点に着目して認識精度の改善を可能としたものであり，音声情報工学ならびに情報科学の発展に寄与するところが少なくない．

よって，本論文は博士(情報科学)の学位論文として合格と認める．