博 士（情報科学） 大 貫 和 永

学 位 論 文 題 名

ロバスト音響モデルと連続音声認識システムに関する研究

学位論文内容の要旨

近年インターネットを利用したe-learningシステムで，学習の場を提供することが一般的になっている.OCW-MITではMITで正規に提供された講義とその関連情報を学生のみならず，一般に無償公開している．大学で行なわれた講義がそのままビデオ情報として提供されており，このような講義ビデオによる授業の視聴は，臨場感をもった講義が受けられ意義深い．

しかし，視聴者が講義ビデオを復習のために利用する際を考えると，再び最初からビデオを視聴することはなく，一度の視聴で理解しきれなかった特定箇所の視聴を希望するはずである．また，何かの理由で特定の知識が必要になったとき，辞書や百科事典のような文字情報を調べるときのように，必要な部分にだけ，ビデオにランダムアクセスしたいと考える．そこで問題になるのが，ビデオが基本的にシーケンシャルにアクセスするメディアであり，ランダムアクセスに適さないため，手軽ではないという難点である．

そこで音声認識技術により，ビデオに含まれる講義音声をテキスト化し，その発話された時間情報とともに蓄え，解説が必要な言葉を，テキストから探し出し，ビデオの該当箇所にランダムアクセスできるシステムを構築することができれば現在公開されている講義情報をさらに便利に利用することができる．このようなシステムを構築することが本研究の目的である．

しかし，現在の大語彙連続音声認識技術では，ノイズ等の影響があると認識性能が大幅に劣化する．そのため研究目的のシステムを実現するには雑音に対するロバスト性を向上させる必要がある．そこで，大語彙連続音声認識に用いる音響モデルのロバスト性の向上を本論文の目的とした．

まず，これまでに実現された孤立フレーズ音声認識の実用的なロバスト化に寄与しているRSF法とDRA法を紹介した.RSF法は変調スペクトル上の低域と高域部分に存在する音声認識に悪影響を与える成分をランニングスペクトル上のフィルタリングにより削減し，雑音を除去する．そこで，大語彙連続音声認識に利用する音素単位の音響モデルに，孤立フレーズ単位の音響モデルと同様に変調スペクトル上のスペクトルを操作することによりロバスト性を高めることを目指した．

音素単位音響モデルを構築するための学習データとして，日本音響学会から提供されている新聞記事読み上げコーパスJNASがある．このデータにより，雑音の少ない場合は90パーセント以上の精度で認識可能な音響モデルを構築できる．そこで，変調スペクトル上の音声認識に悪影響を与える成分の低減を，学習データに施して音響モデルを学習し，ロバスト音響モデルの構築を試みた．

コーパスで提供されている発話データに対する変調スペクトル操作は，通常の音声認識時と違い，発声される音声信号をリアルタイムに処理を行なう必要性は無く，既に蓄積されている発話データに対して，一括処理を行なって変調スペクトル上のスペクトル操作を行なえる．そのため，フィルタによらず離散フーリエ変換に基づくランニングスペクトルアナリシス,RSA法を提案した．即ちランニングスペクトルにDFTを掛け得られるスペクトルの任意部分を強制的に減衰し,IDFTにて元の信号にもどす．これにより，柔軟な変調スペクトル操作を可能となる．

JNASを使って音響モデルを構築する際，音声を収録した場所による音響系の違いを吸収するために乗法性雑音除去の目的でCMSの処理を行なうのが標準である．これは変調スペクトル上の直流成分除去処理になっている.RSAは変調スペクトル上の任意箇所に柔軟な操作を行えるが，スタンダードな処理との比較のため低域

はCMSを活かし,高域側にRSAを適用した.

発話データの低域にはCMS,高域にRSAを適用し音響モデルを学習した.この音響モデルをロバスト音響モデルと呼ぶ.ロバスト音響モデルを使って,雑音の混じった音声を認識すると,標準の音響モデルを利用する場合より,3% 程度高い精度で認識できることを実験で確認した.SNR=20dB の時に,男声特定話者の単語正解率が標準音響モデルを使った場合に 76.71% であるのに対して,ロバスト音響モデルを使った場合は 79.19%,SNR=10dB のときは,37.07% に対して 40.15% という性能の向上があった.しかも被認識信号に対して,RSA による雑音成分除去が不要であり,認識速度に影響を与えない利点が明らかとなった.

このことから,音素単位の音響モデル構築時に,JNAS のような新聞記事読み上げ文を利用する場合は,発話データ単位で RSA による変調スペクトル操作を行なうことが,孤立フレーズ音声認識と同様に音声認識のロバスト性を高めるために役立つことを確認できた.

また,雑音を含まない音声に対しては,男声不特定話者,女声特定話者についてはロバスト音響モデルよりも,標準音響モデルのほうが高い性能を示した.これは孤立フレーズ音声認識の場合と共通な結果となる.この結果とRSA 音響モデルで RSA を掛けた MFCC 特徴ベクトルを認識するよりも,RSA を掛けない標準の MFCC 特徴ベクトルを認識する方が,性能が良くなるという現象を説明するために,モノフォンのロバスト音響モデルと標準音響モデル内の音素間マハラノビス距離を算出した結果,ロバスト音響モデルの距離が明確に大きいことが分かった.

ロバスト音響モデルの音素間マハラノビス距離が大きくなる理由は,変調スペクトルの高域成分を除去することにより,雑音成分が押さえられるため音素データのバラツキが減り,分散ベクトルが小さくなることが原因である.このため,RSA を掛けない MFCC 特徴ベクトルを認識する際にロバスト性が出てくることを明らかにすることができ,それを定性的に説明した.

しかし,現時点のロバスト性では,例えば講義室内の音声のように加法性の雑音が付加している音声にたいして,正確な認識を行えるほどのロバスト性は実現できていない.今後は柔軟なスペクトル操作が可能な RSA に最適なスペクトル操作法を検討する.さらに孤立フレーズ音声認識に適用されていて,まだ連続音声認識に試していない DRA 法の適用を行い,連続音声認識のロバスト性を上げることが必要である.

更に,ロバスト認識を実現している孤立フレーズ音声認識とのハイブリッド化により,高雑音下でも認識できるフレーズを聞き分け,そのフィードバックにより,連続音声の認識精度を上げる方策を考える.これは,他のことに気を取られた人間が,周囲の音声を認識していないのに,名前を呼ばれた場合に,それに気がつく,音素単位と単語単位の認識を行える人間のようなシステムの構築につながる.

学位論文審査の要旨

主　査　教　授　宮　永　喜　一
副　査　教　授　小　柴　正　則
副　査　教　授　野　島　俊　雄
副　査　教　授　小　川　恭　孝

学位論文題名

ロバスト音響モデルと連続音声認識システムに関する研究

現在の大語彙連続音声認識技術では、ノイズ等の影響があると認識性能が大幅に劣化する。そのため研究目的のシステムを実現するには雑音に対するロバスト性を向上させる必要がある。そこで、大語彙連続音声認識に用いる音響モデルのロバスト性の向上が本論文の目的である。

まず、これまでに実現された孤立フレーズ音声認識の実用的なロバスト化に寄与している RSF 法と DRA 法を紹介した。RSF 法は変調スペクトル上の低域と高域部分に存在する音声認識に悪影響を与える成分をランニングスペクトル上のフィルタリングにより削減し、雑音を除去する。そこで、大語彙連続音声認識に利用する音素単位の音響モデルに、孤立フレーズ単位の音響モデルと同様に変調スペクトル上のスペクトルを操作することによりロバスト性を高めることを目指した。

音素単位音響モデルを構築するための学習データとして、日本音響学会から提供されている新聞記事読み上げコーパス JNAS がある。このデータにより、雑音の少ない場合は 90 パーセント以上の精度で認識可能な音響モデルを構築できる。そこで、変調スペクトル上の音声認識に悪影響を与える成分の低減を、学習データに施して音響モデルを学習し、ロバスト音響モデルの構築を試みた。

コーパスで提供されている発話データに対する変調スペクトル操作は、通常の音声認識時と違い、発声される音声信号をリアルタイムに処理を行なう必要性は無く、既に蓄積されている発話データに対して、一括処理を行なって変調スペクトル上のスペクトル操作を行なえる。そのため、フィルタによらず離散フーリエ変換に基づくランニングスペクトルアナリシス、RSA 法を提案した。即ちランニングスペクトルに DFT を掛け得られるスペクトルの任意部分を強制的に減衰し、IDFT にて元の信号にもどす。これにより、柔軟な変調スペクトル操作を可能となる。

JNAS を使って音響モデルを構築する際、音声を収録した場所による音響系の違いを吸収するために乗法性雑音除去の目的で CMS の処理を行なうのが標準である。これは変調スペクトル上の直流成分除去処理になっている。RSA は変調スペクトル上の任意箇所に柔軟な操作を行えるが、スタンダードな処理との比較のため低域は CMS を活かし、高域側に RSA を適用した。

発話データの低域には CMS、高域に RSA を適用し音響モデルを学習した。この音響モデルをロバスト音響モデルと呼ぶ。ロバスト音響モデルを使って、雑音の混じった音声を認識すると、標準の音響モデルを利用する場合より、3% 程度高い精度で認識できることを実験で確認した。SNR=20dB の時に、男声特定話者の単語正解率が標準音響モデルを使った場合に 76.71% である

のに対して、ロバスト音響モデルを使った場合は 79.19%、SNR=10dB のときは、37.07% に対して 40.15% という性能の向上があった。しかも被認識信号に対して、RSA による雑音成分除去が不要であり、認識速度に影響を与えない利点が明らかとなった。

このことから、音素単位の音響モデル構築時に、JNAS のような新聞記事読み上げ文を利用する場合は、発話データ単位で RSA による変調スペクトル操作を行なうことが、孤立フレーズ音声認識と同様に音声認識のロバスト性を高めるために役立つことを確認できた。

また、雑音を含まない音声に対しては、男声不特定話者、女声特定話者についてはロバスト音響モデルよりも、標準音響モデルのほうが高い性能を示した。これは孤立フレーズ音声認識の場合と共通な結果となる。この結果と RSA 音響モデルで RSA を掛けた MFCC 特徴ベクトルを認識するよりも、RSA を掛けない標準の MFCC 特徴ベクトルを認識する方が、性能が良くなるという現象を説明するために、モノフォンのロバスト音響モデルと標準音響モデル内の音素間マハラノビス距離を算出した結果、ロバスト音響モデルの距離が明確に大きいことが分かった。

ロバスト音響モデルの音素間マハラノビス距離が大きくなる理由は、変調スペクトルの高域成分を除去することにより、雑音成分が押さえられるため音素データのバラツキが減り、分散ベクトルが小さくなることが原因である。このため、RSA を掛けない MFCC 特徴ベクトルを認識する際にロバスト性が出てくることを明らかにすることができ、それを定性的に説明した。

しかし、現時点のロバスト性では、例えば講義室内の音声のように加法性の雑音が付加している音声にたいして、正確な認識を行えるほどのロバスト性は実現できていない。今後は柔軟なスペクトル操作が可能な RSA に最適なスペクトル操作法を検討する。さらに孤立フレーズ音声認識に適用されていて、まだ連続音声認識に試していない DRA 法の適用を行い、連続音声認識のロバスト性を上げることが必要である。

更に、ロバスト認識を実現している孤立フレーズ音声認識とのハイブリッド化により、高雑音下でも認識できるフレーズを聞き分け、そのフィードバックにより、連続音声の認識精度を上げる方策を考える。これは、他のことに気を取られた人間が、周囲の音声を認識していないのに、名前を呼ばれた場合に、それに気がつく、音素単位と単語単位の認識を行える人間のようなシステムの構築につながる。

以上の点より、本論文の目的である、種々の雑音に対するロバスト連続音声認識システムの研究において、十分な成果を挙げている。

これを要するに、筆者は、新たなロバスト連続音声認識手法の提案とその開発を行い、種々の雑音に有効なロバスト性を持つ新しい信号処理技術を実現し、その有効性を示した。これにより、音声情報処理・音声認識システムの開発・実現に関する多くの有益な知見を得ており、情報科学の分野に貢献するところ大なるものがある。

よって筆者は、北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。