論文要旨

Adaptation of Acoustic Models for Speech Recognition with A Focus on Intra-speaker Variation
（話者内変動に着目した音声認識のための音響モデルの適応）

氏名　　李　宝潔

一、背景

音声認識の研究はこの二十年間に大きな進歩を遂げてきた。その技術を生かして、高い認識性能を持つ音声認識システムも数々開発され、市販ソフトにもなっている。大概に言えば、音声認識システムは音響モデル、言語モデルとディコーダの三つの部分から構成されている。言語モデルは新聞、雑誌などの文書から単語の出現確率を推定して、確率論の立場から、出力した認識結果を“文章らしい”文章にする。一方、音響モデルは大量の発話から推定され、主に隠れマルコフモデル（Hidden Markov Model（HMM））で記述される。音響モデルと言語モデルの使用によって、現在の音声認識システムは朗読音声に対して、９５％以上の認識結果が得られる。本研究は音響モデルに焦点を当て、よりよい認識結果を目指している。ここで、言語モデルの説明を省略する。

音響モデルの訓練には一般論として、発話の量が多ければ多いほど、より精密なモデルが得られる。ある話者一人のデータを使用して訓練したモデルは話者依存音響モデルと言い、大量話者の発話を使用して訓練したモデルは話者独立音響モデルと言う。音声認識システムの性能を低下させる主な原因は音素の音響的な変動である。この変動は話者内と話者間のものに分類できるが、話者間の変動は話者内の変動より非常に大きく、話者依存認識システムと話者独立認識システムの性能の差の主因となっている。従って、話者独立音響モデル（のパラメータ）をある話者のモデルに近付ける話者適応が盛んに研究されている。そのうち Maximum a posteriori Adaptation (MAP)や Maximum Likelihood Linear Regression(MLLR)などの手法が開発され、大きな成果をあげている。

MAP 手法は適応データの量の増加に連れて、話者独立モデルを話者依存モデルに漸近的に近づくような性質を持つ。適応データの量が多いときに有利である。一方、MLLR 手法は話者間の音素の音響空間上での相対位置は線形回帰の関係を有することを仮定し、その回帰係数（Transformation Matrix）を適応データの出現尤度を最大化するようにして求め、求めた Transformation Matrix を利用して、話者独立音響モデルのパラメータを目標話者モデルのパラメータに変換する。この手法は音素の音響的パラメータを利用して、Regression Tree を求めておいて、それを利用して、適応データサンプルに出現していない音素でも、Regression Tree の中で同じノードに属するほかの音素（音響的にはこの音素に近い）の Transformation Matrix を利用して適応することもできる。適応データは少量の時にこの手法の効果は著しい。

人間が音声を発声する場合、各音素の特徴空間上での位置は大きく変化するが、音素

間の相対的な位置関係は比較的安定していると予想される（MLLR 手法はある程度この関係を利用している)。特に、母音ではそのフォルマントの位置関係が安定していることが知られている。そこで、このような音素間の位置関係をモデルに反映させることができれば、話者間、話者内の変動に有効に対処できると考えられる。このような観点から、音素間の相関情報を音声認識に利用する手法として Extended MAP が報告されている。この手法では、すべての音素モデルの平均ベクトル間の相関情報を学習して、適応データが少量の時にも有効である。しかし、音素の相関行列を学習するには大量の話者のデータが必要という問題点がある。

以上に紹介した手法が有効であることは（特に MAP と MLLR 手法)、多数の実用音声認識システムによって証明された。しかしながら、今までの研究や応用システムが扱う対象はほとんど“朗読音声”である。“朗読音声”の大きな特徴としては、話者内の音響的な変動は小さいという点にある。この特徴こそ、上述手法の有効性の基盤である。しかし、自由会話とか感情音声になると、話者内の音素の音響的な特徴が発話ごとに大きく揺れ、音声認識システムの性能を大幅に低下させる。このような観点から、我々は、話者内音素の音響的な変動に焦点をあて、それを確率的な表現に適した統計的手法によって捕らえることを試した。今までに二つの手法を提案した：

その一は音素間の関係を表現することを考え、新しく音素ペアモデル（PPM と略す）を開発した。認識時に少量適応データと事前に得た音素ペアモデルを利用して、入力音声を再スコアことによって認識精度を上げる。

その二は適応手法の改善である。先ず従来の適応手法によって話者依存モデルを獲得する。その後適応データをクラスタリングして、カテゴリごとに適応する。ゆえに話者間、話者内の二段階適応を行う。ここで二段階適応と称す。

二つの手法とも話者内変動を文レベルの単位まで記述することができる。即ち、一つの文内では、各音素の発話環境が同一であることを仮定し、文と文の間に発話環境が常に変わっている（この点では従来の手法は同一話者なら、文間の発話環境が変わらないと仮定している)。以下はこれらの手法について説明する。

二、音素ペアモデル

現在の主流である HMM(隠れマルコフモデル)による音声認識システムでは、一つの音素に対応する音声信号は時系列ベクトルとしてパラメータ化され、HMM によりモデル化されている。ここで各 HMM は 3 つの自己遷移を有する状態を持つとする。二つの時系列ベクトルもしくは二つの HMM の結合確率を求めるとかなり複雑な問題になる。一方、音素の HMM の各状態の平均ベクトルはある程度その音素の特徴を反映できる。但し、最初と最後の状態は前後の音素に影響され易い。ここでは二つの音素の結合確率を音素の HMM の中間状態の平均ベクトル間の結合確率で近似する。入力音声の認識過程において、一つの音素が分かった時、別の音素を両者の相互関係によって推定することができる。

認識エンジン（本研究では HTK を使用）では各単語は音素 HMM の連結より表現さ

れ、すべての単語は同一出現確率で認識ネットワークを構成する（言語モデルを採り入れると、各単語の確率は異なるが、ここで言語モデルを使っていない)。token という概念が導入されており、そこには今まで辿って来たルートとその尤度が保持されている。各 HMM 状態が一つの token をもっていて、入力音声の各フレームが処理される時、各 token が遷移する状態にコピーされ、その際部分スコアが付加される。単語の終点に到着した時点で、その単語までの部分パスの尤度とその単語を構成するすべての音素の境界が決定される。ここまでが従来の手法である。もしこの時点で正解単語のスコアを上げ、逆に別の単語のスコアを下げることができれば、認識率は高くなると考えられる。これを目的として音素ペアモデルにより計算したスコアを付加することを考える。

音素ペアモデルの音声認識への応用は、以下の手順で行なう。

訓練データを用い、不特定話者の音素モデル（SI HMM）を訓練する。訓練データを使用して音素ペアモデルを訓練する。音素ペアモデルは母音－母音ペア 15 種類、母音－長母音 25 種類、母音－子音ペア 150 種類である。なお、訓練上用いる音素ペアのデータは一文内に両音素同時存在するもののみとする。適応時、一般の話者適応手法と同じように、認識する話者の少量データ（以降、適応データ）を用意する。話者適応手法ではそのデータを使って SI HMM を適応するが、ここでは、その話者のデータから 5 つの母音の各一サンプルを獲得する（話しの展開上、少量データを適応データと呼ぶが、提案手法は一般の意味での適応は行なっていない。音素ペアモデルの片方音素として使う。認識時に、各入力ベクトルと取得した母音ベクトルでペアを組み、音素ペアモデルで確率を計算、従来の HMM より得たスコア（対数尤度）に加算する。

この手法の有効性は認識実験より証明された。

三、二段階適応

話者内の音響的な変動が大きいときに、従来の適応手法はこの変動性を取りにくいため、認識率は大幅に下がる。特に、自由会話や感情音声では話者内の音響的な変動が目立ち、今も大きな課題になっている。そこで提案した二段階手法は先ず従来の手法を用いて、新しい話者のすべての適応データを使用して、その話者の SD モデルを獲得する。このステップで新しい話者発話の大まかな特徴を掴む。続いて、適応データを音響特徴に拠って、いくつかのカテゴリにクラスタリングする。各カテゴリのデータを使って、前のステップで得たモデルを再適応して、カテゴリモデルを生成する。このように生成したカテゴリモデルは自分のカテゴリに属する音声に対して、高い認識率が得られる（SD モデルと比較して)。当然ながら、違うカテゴリに属する音声に対して、認識率がかなり劣化している。

仮に認識時に入力音声の属するカテゴリが分かっていれば、それに対応するカテゴリモデルを使えば、従来の SD モデルより高い認識率が期待できる。しかし、入力音声がどのカテゴリ属するかは認識時は知らない。この問題を解決するため、一つの入力音声に対して、すべてのカテゴリモデルを同時に使って認識実験を行う。出力したいくつの結果の中に、尤度の最大のものを選び、最終認識結果とする。

この手法の有効性は感情音声の上で検証した。感情音声を選んだ理由は話者内変動が大きいことにある。

以上紹介したように、二つの手法はともに話者内変動に着目したが、前者は候補の尤度をリスコアし、後者はモデルを適応する。如何にして両者を有機的に結合するかは、新たな一つの興味ある研究課題である。