論 文 の 内 容 の 要 旨

論文題目　　携帯電話向けHMM音声認識の高精度化と高速化に関する研究

氏　　名　　加 藤 恒 夫

1. はじめに

携帯電話は一般コンシューマ市場への登場から 20 年足らずの間に，移動時用・個人用の電話から個人の携帯情報端末へと急速に進化した．1998 年にメール機能が加わり，1999 年には携帯インターネットの端末となった．2001 年に登場した第 3 世代携帯電話においては，CPU の処理能力が上がり，アプリケーションの実行環境が徐々に整備されていった．2008 年以降はスマートフォンが急速に普及している．

1999 年以前，携帯電話は主に音声通話に用いられた．音声認識は，コールセンターの省力化，自動化や，電話オペレータの補助を目的とする応用について検討された．電話音声認識では，不特定話者音響モデルの高精度化，特に携帯電話に用いられていた低ビットレート音声コーデックによる音声歪みに対する認識精度の向上が課題であった．

1999 年，携帯インターネットが出現すると，急速に普及した．携帯電話のテンキーは電話番号の入力だけでなく文字入力にも多く用いられるようになった．テンキーによる文字入力は慣れを必要としたが，当時テンキーに替わる音声入力機能は存在しなかった．そこで携帯電話とサーバ型音声認識装置が連携して認識処理を行う分散型音声認識システムを開発した．分散型音声認識システムでは，認識結果を画面に表示できるため，ユーザは即座に認識結果を確認することができ，部分的に修正することも容易になった．

2001 年，第 3 世代携帯電話が登場し，CPU の高速化，搭載メモリの増大とともにアプリケーション実行環境が徐々に整った．携帯電話には様々な機能が加わり，機能の呼び出し方も複雑化していった．携帯電話機能の呼び出しやアドレス帳の検索に利用できるローカル音声認識エンジンが必要とされた．ここでは認識精度の劣化なく処理時間を短縮する方式が課題となった．

本論文は，以上の音声認識の実用化に必要となった音声認識の高精度化と高速化に関する提案を纏めている．

2. 混合分布 HMM における決定木に基づく状態クラスタリング

音響モデルの学習において常に問題となるのは，モデルの複雑さと，モデルパラメータの推定精度のバランスである．音素の連鎖の種類などにより学習データ量に多寡があることは避けられないため，HMM 状態などの間でモデルパラメータの共有化を行うことが多い．決定木に基づく状態クラスタリングは，数万種類にも及ぶ音素コンテキスト依存モデルの HMM 状態をまず音素毎に集め，音素コンテキストに関する二者択一の質問を繰り返してトップダウンにクラスタリングすることで，あらゆる音素コンテキストに対して HMM 状態を割り当てる状態共有化手法である．学習データ中に現れない未知の音素コンテキスト依存モデルに対しても，HMM 状態が割り当てられるという優れた特徴を有している．

音素コンテキスト依存モデルの標準的な学習プロセスにも採用されている決定木に基づく状態クラスタリ

ングは，従来，単一の正規分布で表現される状態共有なしモデルを対象としていた．しかし，音声認識に実際に用いられるのは混合分布 HMM であり，表現能力が不足する単一分布 HMM は用いられない．単一分布 HMM に対して決定木に基づく状態クラスタリングを行い，その後 Baum-Welch 再推定と分布数の倍増操作とを繰り返す必要があるため，最終的な混合分布状態共有モデルを獲得するまでの学習のステップが多く，時間がかかるという問題もあった．

そこで，決定木に基づく状態クラスタリングも，混合分布 HMM を取り扱えるように拡張した．クラスタリング途中のノードもすべて混合分布で表現することで，共有構造の改良を図っている．音節タイプライタと連続単語認識で評価したところ，提案手法による音響モデルは，決定木に基づく状態クラスタリングに引き続き Baum-Welch 再推定を繰り返した後でも，従来手法による音響モデルよりも認識精度が優れていた．また，提案手法は状態クラスタリングの結果として混合分布 HMM を出力するため，従来手法に比べて学習時間を大幅に削減することができる．

3. コーデック適応音響モデルおよび雑音モデルによる認識精度の改善

電話音声認識システムでは，2000 年当時，携帯電話サービス毎に異なる低ビットレート音声コーデックに特有の音声歪みと，屋外使用によって多く混入する非定常な雑音は，認識精度の大きな要因となっていた．回線特性のばらつきを抑える代表的な手法として，ケプストラム平均値正規化(CMN)があるが，線形時不変を前提とするケプストラム平均値正規化では，これらの問題に対処することはできない．そこで，音声コーデック毎に不特定話者音声モデルと非定常雑音モデルを用意し，これらを選択的に用いる手法を提案した．

携帯電話がよく用いられる雑音環境で収録した携帯電話音声を用いた 3,000 単語の孤立単語認識タスクにおいて，コーデック適応音声モデルと雑音モデルの導入により音声区間の境界推定精度が大幅に改善され，単語誤り率はコーデック適応音声モデルにより約 10%，コーデック適応雑音モデルの導入によってさらに約 15%削減された．

4. 木構造辞書における到達可能単語数を利用した探索高速化

サーバ型の音声認識ではより多くの語彙を含む大規模な言語モデルで認識を行うため，ローカル型の音声認識では限られたリソースでできるだけ早く認識結果を提示するために，高速な音声認識アルゴリズムは常に必要とされている．

HMM に基づく音声認識では，探索空間は，文法や確率的言語モデルで規定される単語レベルのネットワークと，文法や確率的言語モデルを構成する単語を HMM の状態系列で表す HMM 状態系列のネットワークの 2 階層で表現される．後者は，探索を効率化するために異なる単語間で単語の先頭から共通する HMM 状態系列をマージして木構造辞書とする．認識処理中は，様々な単語の系列を検証するために，多数の仮説が並行して木構造辞書上を探索する．探索中の仮説の総数が際限なく増えないようにするために，毎時刻フレームに仮説の枝刈りを行う．

確率的言語モデルに基づくディクテーションタスクの場合，確率的言語モデルの言語確率を木構造辞書の探索にできるだけ早く反映させる言語モデル先読みの効果が非常に大きく，認識精度を最大化するのに必要な仮説数を大幅に削減し，高速な探索を可能にしている．文法に基づく認識はディクテーションに比べて小規模なタスクになることが多いが，それでも語彙が増大すると認識精度を最大化するのに必要な仮説数は増大する．言語確率を用いないため，言語モデル先読みは適用できない．

そこで，木構造辞書のルートに近い HMM 状態にある少数の仮説は様々な単語に発展する可能性があるので，リーフに近い HMM 状態にある多数の仮説よりも重要度が高く，仮説枝刈りにおいて優遇されるべきと

いう考えに基づき，従来のすべての仮説を平等に取り扱う仮説枝刈りに替わり，仮説毎に木構造辞書における到達可能単語数を考慮して枝刈りの厳しさを連続的に変化させる方法を提案した．すなわち，ルートに近い HMM 状態にある少数の仮説に対しては枝刈りの条件を甘く，リーフに近い状態にある多数の仮説に対しては枝刈りの条件を厳しくすることで，認識精度を落とさずに枝刈りの効率を上げる．

提案手法を，孤立単語認識タスク，文法に基づく短文認識タスク，確率的言語モデルに基づく連続音声認識タスクで評価したところ，すべてのタスクにおいて認識精度の最大値を悪化させることなく必要な仮説数を削減し，認識処理を高速化した．特に，文法に基づく短文認識タスクでは確率的言語モデルの先読みを適用できないため，提案手法の高速化効果は大きく，従来の 1/5 以下の処理時間で従来の認識精度の最大値を超える認識精度に到達している．

5. 音声区間検出の情報と木構造辞書における深さを利用した探索高速化

前章の提案では，木構造辞書上のビーム探索において，仮説の尤度に到達可能単語数に応じた時不変の報酬を加算することで探索効率を改善した．この報酬を時変にして制御することでさらなる探索の効率化が可能になるか検証するために仮説数の時間変動を調査した．

仮説数の時間変動を調査すると，音声の始端が検出される前の無音区間において仮説数が爆発的に増大することがわかった．このとき仮説は様々な単語の深い位置まで拡がっていた．そこで，音声始端検出前の仮説の拡がりを抑えるために，音声区間検出の情報を用いて木構造辞書の深さに応じた一時的なペナルティを仮説の累積尤度に加算してビーム探索を行う手法を提案した．探索対象の全区間に対する音声区間検出前の無音区間の割合が大きい孤立単語認識タスクでは，認識処理時間を 7～10%短縮することができた．

6. おわりに

筆者は，音声認識技術を携帯電話向けに実用化するために，電話音声認識，分散型音声認識システム，ローカル音声認識エンジンの開発に携わり，取り組みの中で必要となった音響モデルの精度改善手法と音声認識エンジンの高速化手法を考案した．

提案手法を適用した音声認識サーバをベースにして，第 3 世代携帯電話向けの分散型音声認識システムを開発した．2006 年に同システムは携帯電話を端末とする世界初の分散型音声認識システムとして実用化を果たした．現在まで携帯電話の主要なアプリケーションである乗換検索や目的地検索に利用されている．また，その後に開発したローカル音声認識エンジンはクロック周波数 100MHz 程度の第 3 世代携帯電話上で，約 1 万語までの文法に基づく音声認識をリアルタイムで実行できる．同エンジンは第 3 世代携帯電話の高齢者向けモデル 4 機種に搭載されている．