コードブック適応を用いた離散混合分布型HMMによる
講演音声認識

山本明祥  熊倉拓哉  加藤正治  小坂哲夫  好田正紀
山形大学大学院理工学研究科
E-mail: eny73423@dip.yz.yamagata-u.ac.jp

あらまし これまで我々は、離散混合分布型HMM(DMHMM:Discrete-Mixture HMM)による雑音のない
クリーンな環境での性能評価を行ってきた。さらなる性能向上を目指し、本研究では、パラメータ数の検
討やヒストグラム同等化によるコードブック適応の検討を行う。評価に当たっては、他機関との比較ができ
る共通コーパスであり、困難なタスクでの性能が評価できる「日本語話し言葉コーパス」(CSJ)を用い
る。また、認識性能向上のために、認識結果がどのようなところで誤っているかの分析を行った結果につ
いても報告する。

キーワード 講演音声認識、離散混合分布型HMM、ヒストグラム同等化、
日本語話し言葉コーパス、音響モデル

Lecture Speech Recognition by Using Codebook Adaptation of
Discrete-Mixture HMMs

Akiyoshi YAMAMOTO, Takuya KUMAKURA, Masaharu KATOH, Tetsuo KOSAKA,
and Masaki KOHDA

Graduate School of Science and Engineering, Yamagata University
E-mail: eny73423@dip.yz.yamagata-u.ac.jp

Abstract  We have investigated speech recognition on clean data by using discrete-mixture HMMs
(DMHHMs). Aiming to further improvement in performance, we investigate codebook adaptation by
a histogram equalization and the number of model parameters. In evaluation, we used the "Corpus of
Spontaneous Japanese" (CSJ) because we want to compare the performance of our system with that of
other recognition systems with common speech corpus. Furthermore, for the improvement in recognition
performance, we analyzed that what kind of errors were shown in recognition results.

Key words  lecture speech recognition, discrete-mixture HMM, histogram equalization,
corpus of spontaneous Japanese, acoustic model

1. はじめに

これまで我々は離散混合分布型HMM(DMHHM: Discrete-Mixture HMM)による「日本語話し言葉コーパ
ス」(CSJ)を用いた雑音のないクリーンな環境での性能評
を目的とする。HEQ は一般に入力特徴の正規化に用いられていた手法であるが、この手法をコードブック適応に用いる。評価において「日本語話者音声コーパス」（CSJ）を用いた。これは、他機関との比較ができる共通コーパスを用いるのが望ましいこと、及び、なるべく困難なタスクでの性能を明らかにしたいためである。また、今後の認識率向上のために、認識結果についてどのようにところで誤っているのか分析を行った。現状では単に単語誤り率を算出することにより評価を行っている。この評価には、通常は必要のないフィラーや言い誤りなどの認識誤りも含まれている。このような認識誤りを考慮しない評価をする必要がある。

本文論の構成は以下のようになっている。第 2 章で、DMHMM の概要、パラメータ推定について述べる。第 3 章で、ヒストグラム同等化、コードブック適応について述べる。第 4 章では、本研究で使用した発音形変を考慮した形態素解析モデルに基づく語音モデルについて述べる。第 5 章では、各種実験の実施条件について述べる。第 6 章で、6000 節点のモデルを用いた認識実験、CHMM と比較した結果および、ヒストグラム同等化を行った場合の結果について述べる。第 7 章では、認識結果についての分析の方法とその結果について述べる。最後に第 8 章でまとめと今後の課題について述べる。

2. 離散混合出力分布型 HMM

2.1 概要

離散分布型 HMM では正規分布の仮定がないため、任意の分布形状を表現できる一方、パラメータ推定精度の問題がある。この問題に対し、量子化サイズが小さくなるような HMM の構造が提案されている。特徴パラメータの個々の空間の次元数を落とす方法として離散混合出力分布型 HMM（DMHMM）が提案されている。このタイプの HMM では、量子化法の違いで 2 つの方法が提案されている。文献[2]では特徴空間をサブベクトル分割し、サブベクトルごと量子化する方法が提案されている。例えば特徴ベクトルの隣接する 2 次元を 1 つのサブベクトルとして量子化を行う、またスカラー量子化に基づく方法[3]では、サブベクトルではなく特徴の各次元のスカラー量子化を行っている。このように特徴ベクトルを分割し量子化することにより、個々の量子化サイズを十分小さく抑えることができる。本研究では前者のサブベクトルごとに量子化する手法を用いる。

以下に DMHMM の概要を示す。まず入力特徴ベクトル \( \mathbf{o}_t \) をサブベクトル \( S \) に分割し \( \mathbf{o}_t = [o_{t1}, \ldots, o_{ts}, \ldots, o_{tT}] \) とする。その時、特徴量の隣接するものを一まとめるのに \n
を対応させる。コードブック作成用データを用いて、各サブベクトルごとのコードブックを作成する。\( q_s(o_{ts}) \) をサブベクトル \( s \) における入力 \( o_t \) に対するセントロイドとすると、入力 \( o_t \) は以下のように量子化される

\[
q(o_t) = [q_1(o_{ts1}), \ldots, q_s(o_{ts}), \ldots, q_S(o_{ts})]
\]  \( (1) \)

この時、DMHMM の出力確率 \( b(o_t) \) は以下のようになる。

\[
b(o_t) = \sum_m w_m \prod_s \tilde{p}_{sim}(q_s(o_{ts}))
\]  \( (2) \)

但し、

\[
\sum_m w_m = 1.0
\]  \( (3) \)

- \( \tilde{p}_{sim} \):サブベクトル \( s \), 状態 \( i \), 混合要素 \( m \) における
離散確率

- \( w_m \):混合分布の重み係数

上式は、混合内で異なるサブベクトル間の離散確率は互いに独立だが、状態内の異なるサブベクトル間の従属性は、混合要素でモデル化されるという仮定にとどく。

2.2 パラメータ推定

本研究ではパラメータ推定法として、尤度推定法（ML 推定法）と最大事後確率推定法（MAP 推定法）の 2 種類の検討を行った。MAP 推定では事前分布を考慮に入ることにより、より少ないデータでパラメータ推定ができることが知られている。 \( k \) をコードブックのインデックス、\( \gamma_{im} \) を時刻 \( t \) で状態 \( i \), 混合要素 \( m \) に存在する確率とすると Baum-Welch アルゴリズムにより、離散出力確率の ML 推定値は以下のよう求められる。

\[
p_{sim}(k) = \frac{\sum_{t=1}^{T} \gamma_{im} \delta(q_s(o_{ts}), k)}{\sum_{k=1}^{S} \gamma_{im}}
\]  \( (4) \)

\[
\delta(q_s(o_{ts}), k) = \begin{cases} 1 & q_s(o_{ts}) = k \\ 0 & \text{otherwise} \end{cases}
\]  \( (5) \)

また事前分布を考慮した離散出力確率の MAP 推定値を以下のよう求めめる。

\[
r_{im} = \sum_{t=1}^{T} \gamma_{im}
\]  \( (6) \)

このとき、離散出力確率の MAP 推定値 \( \tilde{p}_{sim}(k) \) は事前分布をディレクタ分布とした場合、

\[
\tilde{p}_{sim}(k) = \frac{\tau \cdot p_{0}^{\tilde{p}_{sim}}(k) + r \cdot n_{im} \cdot p_{sim}(k)}{\tau + n_{im}}
\]  \( (7) \)

となる。ここで \( p_{0}^{\tilde{p}_{sim}} \) は事前分布のパラメータ、\( \tau \) は事前知識の確からしきに関する係数である。今回の実験では
3. ヒストグラム同等化を用いたコードブック適応

3.1 ヒストグラム同等化 (HEQ)

ヒストグラム同等化 (HEQ) の基本的な考え方は、評価データの累積密度関数 (CDF: Cumulative Density Function) が学習データの CDF と一致するように評価データのケプストラムを変換するというものである。本手法では、学習データ、評価データのそれぞれの音声パラメータの分布は、外乱化しなければ同じように分布するという仮定に基づく。この HEQ をコードブック適用に用いる。

この手法では離散化のセントロイドの位置を一点変換するため、分布形状も任意に変換することが可能である。また HEQ では、変換関数は非線形な関数であるため、非線形変換が可能である。CHMM での HEQ による適応も同様な方法で可能であるが、平均値の変換しかできないため HEQ を使う利点は少ないと考えられる。また、CHMM での HEQ ではモデルパラメータ全体を変換する必要があるが、DMHMM ではコードブックのみを変換すればよいため、計算量が大きく変わってくる。モデル空間での正規化することにより、いくつかの利点が得られる。特徴空間での正規化ではないため入力 1 フレームごとに正規化する必要がなくなる。またモデルごとに異なる正規化することも可能になる。例えば、音声と非音声を別々に正規化したり、母音と子音を別々に正規化したりすることも可能である。

HEQ の具体的な方法として、学習データ、評価データそれぞれについて音声パラメータの累積確率分布を求め、これらの分布から同じ確率値を持つ音声パラメータが対応すると考え、音声パラメータの変換関数を求める。この変換関数を用いて、離散分布 HMM のコードブックを変換し、評価データに適応させる。ヒストグラム同等化は以下の式によって行われる。

\[ q'_b(o_{st}) = HEQ(q_b(o_{st})) = C_b^{-1}(C_T(q_{st}(o_{st}))) \] (8)

ここで,

- \( C_T \): 学習データから推定された CDF
- \( C_b \): 評価データの CDF
- \( q'_b(o_{st}) \): 入力 \( o_{st} \) に対するセントロイド

となっている。本手法では、\( q' \) のみが変換され、離散確率値 \( \tilde{p}_{st} \) は変換されない。コードブックのセントロイドが正規化前後で移動する例を図 1 に示す。図は特徴ベクトルの \( C_1, C_2 \) 次元のセントロイドの位置を示している。before が正規化前で、after が正規化後である。この図の場合、セントロイドが内側に移動しているのが分かる。離散確率を正確に推定するためには多くのデータが必要と考えられるが、CDF は比較的少量のデータで推定が可能である。本実験では独立デルタケプストラム同等化 (DCN) の手法を用いる。独立 DCN とは、入力のケプストラムから \( \Delta, \Delta^2 \) を求め、それぞれをヒストグラム同等化する手法である。

3.2 コードブック適応

図 2 に示すように、学習データ、評価データからヒストグラムを作成し、作成したヒストグラムから変換関数を計算する。その変換関数を用いてコードブックを変換し、適応コードブックとして使用する。この適応コードブックを用いることにより、モデル毎の適応が可能となる。
表１ 実験条件

<table>
<thead>
<tr>
<th>データベース</th>
<th>日本語話し音声コーパス</th>
</tr>
</thead>
</table>

学習データ
公開版 CSJ の男性者 795 講演
時間：約 164 時間 単語数：約 266 万語

開発データ
モニター版 CSJ の男性者 4 件講演
時間：約 1.2 時間 単語数：約 2.0 万語

評価データ
公開版 CSJ の男性者 10 講演
時間：約 1.7 時間 単語数：約 2.7 万語

標本化周波数
16kHz

フレーム長（周節）
25msec / 8msec

高域強調
1-0.978^2

特徴ベクトル
MFCC (1-12 次) と対数パワー
及び一次と二次の回帰係数（計 39 次元）

ヒストグラム作成
公開版 CSJ の男性者 795 講演及び
モニター版 CSJ の男性者 4 講演
公開版 CSJ の男性者 10 講演

コードブック作成
公開版 CSJ の男性者 795 講演

表２ コードブックサイズ

<table>
<thead>
<tr>
<th>LogF</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
<th>C8</th>
<th>C9</th>
<th>C10</th>
<th>C11</th>
<th>C12</th>
</tr>
</thead>
<tbody>
<tr>
<td>サイズ</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>64</td>
</tr>
</tbody>
</table>

本実験で使用するコードブックの量子化サイズは各次元 64 となっているが、これらのセントロイドをヒストグラムからの変換関数に従って移動させる。本実験ではこのコードブック適応について検討する。

4. 発音変形依存モデル

日本語話し音節コーパス（CSJ）は、日本語の自発音声を種々の研究付加情報とともに大量に格納したデータベースである。収録にはハードウェアを用いるため抜音の混入は少ないが、発音個別の変動、発音形状など話者特性の様々な現象が存在し、音声認識では使用が可能なタスクとなっている。本研究では発音変形に対処するため、発音変形を考慮した形態素解析データに基づく音声モデルを利用した[4]。この音声モデルでは、実際の発音の一部についても発音が異なるものは別音素として扱い、これらの音素を元に形態学的依存学習を行う。また形態素の音変形を考慮した音節間の切れ音と音節内の切れ音を考慮したため、音素の分割は自然な音素単位でモデル求解に適している。ただし発音の音素が複数音素にわたる現象が見られる場合、分割を行わず 1 単語としている。

5. 認識実験条件

実験条件を表１に示す。データベースには「日本語話し音節コーパス」（CSJ）を用いて実験を行う。学習データ
には公開版 CSJ の男性者話者会講演 795 講演を用いた（公開版-all）。発話時間は約 164 時間、単語数は約 266 万語である。次に各種パラメータ調整用の開発データとして学習講演男性者 4 講演を用いた（testset0）。語彙量・生音入ペナルティなどのパラメータの調整は公開版 CSJ の開発データセットを用いた。評価セットとしては学会講演男性者 10 講演を用いる（testset1）。学習データから開発及び評価セットの講演は除外している。

テストセット作成データは、学習データ、testset0、testset1 を用いた。コードブック作成データは、学習データ、ヒストグラム変換関数に用いて話者適応したものを使用して評価を行う。

音声モデルは音声に忠実な読みを付与した発音変形依存モデルを使用している。音声モデルは公開版 CSJ の男性+女性話者の学会講演および模擬講演の全講演 2,668 講演を用いている。ただし、公開版 CSJ の評価データ 30 講演、モニター版 CSJ の評価データ 4 講演は除外している。以上の単語数は 668,3 万語の音声から作成したものである。音声は学会講演、模擬講演とともに 2 以上に達したものを、エントリ数 45199 単語である。

音声モデルは文脈での状態共有した各 3-6 状態、総状態数は 6000 状態、状態数は 2-16 混合のサブベクトル行列化分割分布 HM-Net を用いた。

コードブックの量子化サイズを表 2 に示す。CHMM モデルを DHMM へ変換したモデルを初期モデルとし、MAP 推定による推定を行った。CHMM は、構造決定、統計学習とともに公開版-all を用いて ML 推定で作成されたモデルである。（学習回数は 5 回）

評価用認識システムは第 1 バスで triphone および単語 bigram を用いて単語グラフを生成し、第 2 バスで単語 trigram を用いて単語グラフをリスコーデする 2-pass デコードを用いる。本実験の評価は全て第 2 バスの結果で評価した。

6. 認 識 実 験

6.1 混合数・推定法・学習回数の検討

6000 状態で混合数が 2,4,8,16 混合モデルを用いて学習回数の検討を行った。学習は 5 回以内で性能向上が飽和するまで学習を繰り返し行った。testset0 の結果を図 3 に示す。計算時間のため、測定実験では単語間のビーム幅/単語内のビーム幅を 80/180 に変更して行った。また、2 回学習の時点で ML、MAP 推定の良い方の推定法で学習を繰り返した。結果を見ると、混合数が少ないとき MAP 推定よりも ML 推定の方が好調であった。混合数が多いときは MAP 推定が良い結果となった。
定の方が良く、1 回学習モデルが一番良い結果となった。また、混合数が増加するとほど結果が良くなっている。結果として 6000 状態 16 混合、MAP 推定 1 回学習モデル (WER:17.43%) が最良の結果となった。さらに、この結果について式 7 の τ の値の検討を行ったところ、τ = 100 でわずかながらの改善が見られた (WER:17.40%)。また、ビーム幅を 150/250 に設定すると最終的に単語誤り率 17.07%を得ることができた。

それぞれの混合数の最良の結果に対して testset1 による評価と学習回数をそろえた CHMM との比較の結果を図 4 に示す。このとき、言語量と陣入ペナルティは testset0 で最適化した値を用いている。結果として 16 混合の MAP 推定 1 回学習モデルで単語誤り率 (21.63%) が最良の結果となった。CHMM と比較すると、2 混合の場合は DHHMM の方が結果が良く、4,8 混合はほぼ同等の性能であり、16 混合は CHMM の方が良い結果となっている。どの混合数においても結果の差は 1 ポイント未満となっていて、CHMM と同等の性能であるといえる。

6.2 HEQ による同等化単位の検討

ヒストグラム同等化を用いたコードブック適応において、最も有効な同等化単位を検討する。本実験で検討する
7. 認識結果の分析

7.1 分析方法
現状では単純に単語認識率を算出することにより評価を行っているが、今後の認識性能向上のためには、どのようなことで誤っているのか分析する必要がある。

調査方法として、認識結果評価時に正解文と認識文のアライメントを取ったものを表示し、認識によるどのような種類の誤りがあったかのカウントを行う。誤りは合計 8 種類に分類した。一般的には置換・挿入・脱落の 3 種類に分類されるが、認識結果を用いる場合、通常はフィラーを必要としないため、フィラーに関する誤りは全てカウントした。また、音響モデルでは対応できない表記の異なりや表記上は誤りとはいえない形態素の異なりも別にカウントした。どのような種類に分類したかを表 3 に示す。誤りの例の一部を以下に示す。例 1 では、「やって」の部分が形態素の区切りが異なるだけで表記は一致している。これを形態素の異なりとしてカウントし、「え」をフィラーの挿入としてカウントした。また、例 2 では、「を」と「の」が置換となり、「見た」と「みた」が表記の異なりとなる。

例 1
正解文：アクセント指令 よって INS INS 実測
認識結果：アクセント指令 よって え 実測
例 2
正解文：どこを 基準として 見れば
認識結果：どこの 基準として みれば

7.2 分析結果
6000 状態 16 混合の MAP 推定 1 回学習モデルの同等化前の結果（WER: 21.63%）を用い、認識結果の分析を行った。分析結果を図 7 に示す。図の通常の誤りは、表 3 の置換・挿入・脱落をまとめたもので、フィラーに関する誤りは、f 置換・f 挿入・f 脱落をまとめたものである。

分析した結果を見ると、全体的に通常の誤りが多く占めているが、話者によっては表記・形態素異なりやフィラーに関する誤りが多い話者がある。このフィラーに関する誤りや形態素の異なりによる誤りを考慮しない評価を行うと、認識率は向上すると思われる。

8. まとめと今後の課題
本研究では、クリニック的な環境での離散混合分布型 HMM(DM-HMM) の性能を検討するために、日本語話者言葉コーパスを用いた実験を行った。実験により、一般的な離散混合分布型 HMM(CHMM) と比較してほぼ同等の性能が得られた。さらに、ヒストグラム同等化を用いたコードブック適応を行うことにより性能が向上し、最終的に testset1 で 96.9%の単語認識率が得られ、この手法が有効であることが分かった。また、認識結果の調査では本質的でない誤りや実用上問題ない誤りが含まれていることが分かった。

今後の検討課題としては、フィラーを除いた評価や形態素の異なりを考慮しない評価、音素每に有音・無音部に分けたヒストグラム同等化の検討を行って行く予定である。

参考文献