社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－SLP－62 (5)
2006／7／7

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 (DMHMMs). 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) beause 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(DMHMM : 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価を行ってきた．その結果，3000状態16混合モデルで従来の混合連続分布型HMM(CHMM:Continuous HMM)と比較して遜色ない性能が得られた[1]．本研究では，パラメータ数の検討やヒストグラム同等化法(HEQ)を用いたコードブック適応を提案し，性能向上を目指すこと

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きた手法であるが，この手法をコードブック適応に用いる．評価において「日本語話し言葉コーパス」(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なところで誤っているのか分析を行った．現状では単純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．

本論文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，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変形を考慮した形態素解析データに基づく言語モデルについて述べる．第 5 章では，各種実験の実験条件について述べる．第 6 章で，6000 状態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に分割し，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充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 $\boldsymbol{o}_t$ をサブベクトル S に分割し $\boldsymbol{o}_t = [\boldsymbol{o}_{1t}, \cdots, \boldsymbol{o}_{st}, \cdots, \boldsymbol{o}_{St}]$ とする．この時，特徴量の隣接するものを一まとめにする．次に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．$q_s(\boldsymbol{o}_{st})$ をサブベクトル s における入力 $\boldsymbol{o}_{st}$ に対するセントロイドとすると，入力 $\boldsymbol{o}_t$ は以下のように量子化される

$$q(\boldsymbol{o}_t) = [q_1(\boldsymbol{o}_{1t}), \cdots, q_s(\boldsymbol{o}_{st}), \cdots, q_S(\boldsymbol{o}_{St})] \tag{1}$$

この時,DMHMM の出力確率 $b_i(\boldsymbol{o}_t)$ は以下のようになる．

$$b_i(\boldsymbol{o}_t) = \sum_m w_{im} \prod_s \hat{p}_{sim}(q_s(\boldsymbol{o}_{st})) \tag{2}$$

但し，

$$\sum_m w_{im} = 1.0 \tag{3}$$

- $\hat{p}_{sim}$:サブベクトル s, 状態 i, 混合要素 m における離散確率
- w_{im}:混合分布の重み係数

上式は，混合内での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もとづく．

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，最尤推定法（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れることにより，より少ないデータでパラメータ推定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，γ_{imt} 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に求められる．

$$p_{sim}(k) = \frac{\sum_{t=1}^{T} \gamma_{imt} \delta(q_s(\boldsymbol{o}_{st}), k)}{\sum_{t=1}^{T} \gamma_{imt}} \tag{4}$$

$$\delta(q_s(\boldsymbol{o}_{st}), k) = \begin{cases} 1 & q_s(\boldsymbol{o}_{st}) = k \\ 0 & otherwise \end{cases} \tag{5}$$

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に求める．

$$n_{im} = \sum_{t=1}^{T} \gamma_{imt} \tag{6}$$

このとき，離散出力確率の MAP 推定値 $\hat{p}_{sim}(k)$ は事前分布をディレクレ分布とした場合，

$$\hat{p}_{sim}(k) = \frac{\tau \cdot p^0_{sim}(k) + n_{im} \cdot p_{sim}(k)}{\tau + n_{im}} \tag{7}$$

となる．ここで p^0_{sim} は事前分布のパラメータ，τ は事前知識の確からしさに関する係数である．今回の実験では

断りがなければ $\tau = 10.0$ である．また事前分布のパラメータ設定にはCHMMを利用した．各セントロイドに対応する確率を正規分布から求めることにより，CHMMを離散化して用いる．

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，外乱が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をコードブック適応に用いる．

この手法では離散確率のセントロイドの位置を一点一点変換するため，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，計算量が大きく違ってくる．モデル空間で正規化することにより，いくつかの利点が得られる．まず，特徴空間の正規化ではないため入力を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を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．

$$q'_s(o_{st}) = HEQ(q_s(o_{st})) = C_E^{-1}(C_T(q_s(o_{st}))) \quad (8)$$

ここで，

C_T：学習データから推定されたCDF

C_E：評価データのCDF

$q_s(o_{st})$：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_s のみが変換され，離散確率値 $\hat{p}_{sim}$ は変換は行われない．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C1,C2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 Δ，Δ^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．

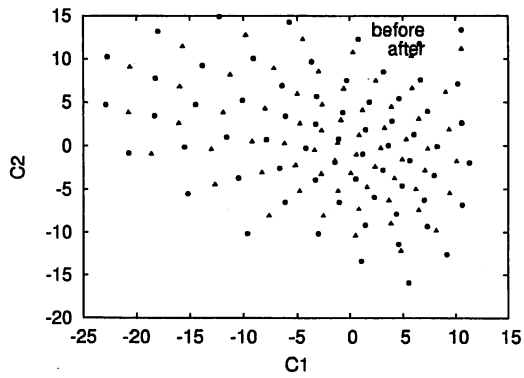

図1 正規化の例

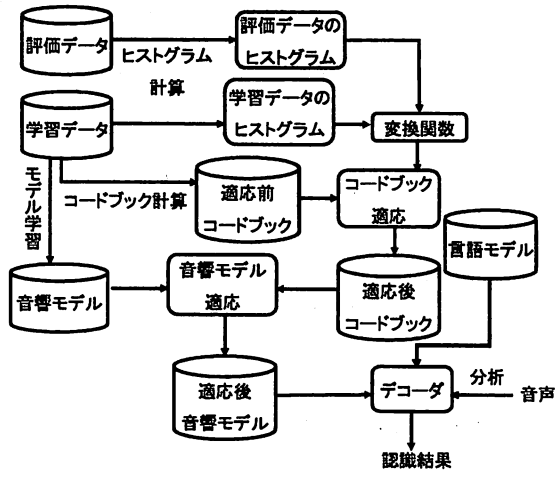

図2 ヒストグラム同等化を用いたコードブック適応

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

データベース	日本語話し言葉コーパス
学習データ	公開版 CSJ の男性話者 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 の男性話者 4 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 の男性話者 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/周期	25msec / 8msec
高域強調	$1\text{-}0.97z^{-1}$
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 の男性話者 795 講演及び モニター版 CSJ の男性話者 4 講演， 公開版 CSJ の男性話者 10 講演
コードブック作成	公開版 CSJ の男性話者 795 講演

表 2　コードブックサイズ

	LogP	C_1 C_2	C_3 C_4	C_5 C_6	C_7 C_8	C_9 C_{10}	C_{11} C_{12}
サイズ	64	64	64	64	64	64	64

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

日本語話し言葉コーパス (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．収録にはヘッドセットを用いているため雑音の混入は少ないが，発話速度の変動，発音変形など話し言葉特有の様々な現象が存在し，認識困難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言語モデルを利用した [4]．この言語モデルでは，実際の発音に即して 1 つの単語でも発音が異なるものは別単語として扱い，これらの単語を元に言語モデルの学習を行う．また形態素の分割も発音変形を考慮した．言語的な切れ目と音響的な切れ目が一致しない現象を回避するため，長い単語は分割し，可能な限り短い単語で語彙を構成した．ただし発音の転訛が複数単語にわたる現象が見られる場合は分割を行わず 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「日本語話し言葉コーパス」(CSJ) を用いて実験を行う．学習データには公開版 CSJ の男性話者学会講演，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セットとして学会講演男性話者 4 講演を用いた (testset0)．言語重みや挿入ペナルティなどのパラメータの調整はこの開発セットを用いた．評価セットとしては学会講演男性話者 10 講演を用いる (testset1)．学習データからは開発及び評価セットの講演は除いている．

ヒストグラム作成データは，学習データ，testset0, testset1 を用いた．コードブック作成データは，学習データを用いた．このコードブックを変換関数を用いて話者適応したものを使用して評価を行う．

言語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．学習テキストは公開版 CSJ の男性 + 女性話者の学会講演 + 模擬講演の全講演，2668 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 講演は除いてある．以上の総単語数約 686.3 万語のテキストから作成したものを用いる．語彙エントリは学会講演，模擬講演ともに 2 回以上出現したもので，エントリ数 45199 単語である．

音響モデルは文献 [5] の方法で状態共有した各 3〜6 状態，総状態数は 6000 状態，1 状態あたり 2〜16 混合のサブベクトル量子化離散混合分布 HM-Net を使用した．

コードブックの量子化サイズを表 2 に示す．CHMM のモデルを DMHMM へ離散化したモデルを初期モデルとし，MAP 推定や ML 推定を行った．CHMM は，構造決定，連結学習ともに公開版-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 パスで triphone および単語 bigram を用いて単語グラフを生成し，第 2 パスで単語 trigram を用いて単語グラフをリスコアする 2-pass デコーダを用いる．本実験の評価は全て第 2 パ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 混合の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．計算量削減のため，認識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 ML 推定の方が良く，5 回学習モデルが一番良い結果となった．混合数が多いときは MAP 推

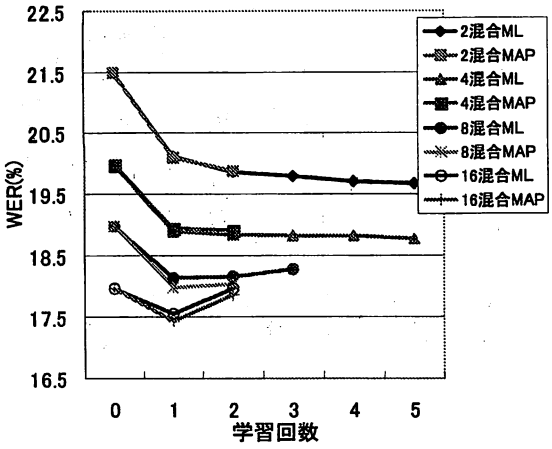

図3 混合数，推定法，学習回数の検討

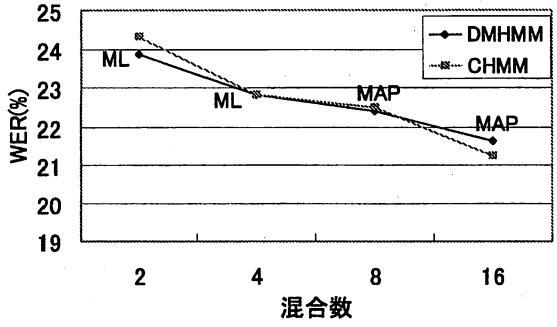

図4 CHMM との比較

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，$\tau=100$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，言語重みと挿入ペナルティは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DMHMM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

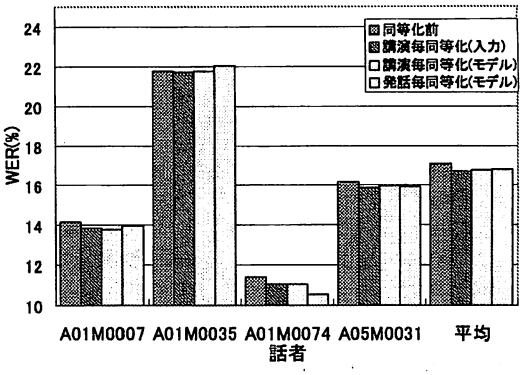

図5 各同等化単位での性能評価 (testset0)

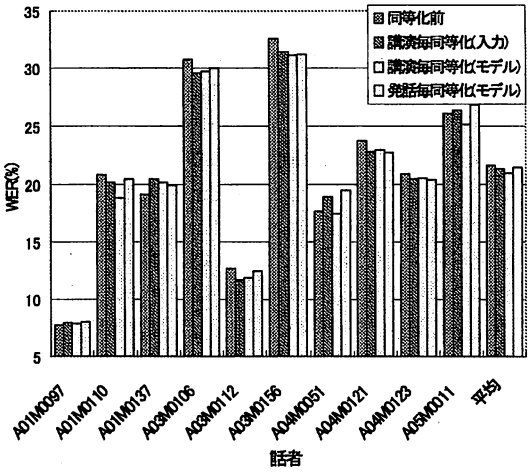

図6 各同等化単位での性能評価 (testset1)

同等化単位は，講演毎，発話毎である．講演毎は評価データの講演全体のCDFを求めヒストグラム同等化を行い，発話毎は評価データの一発話単位でCDFを求め同等化を行ったものとする．また，比較実験として講演毎の入力の特徴パラメータに対しての同等化も行った．同等化するモデルは，最良の結果(16混合MAP1回学習モデル)を用いてHEQによる評価を行った．testset0の結果を図5に，testset1の結果を図6に示す．

図5，図6より，講演毎と発話毎を比較すると平均で見ると講演毎が良い結果となっているが，話者ごとに見るとばらつきが見られる．入力の同等化とコードブック適応の講演毎の結果を見ると，多少誤差があるもののほぼ同等の結果と言える．また，testset1の平均で見ると同等化前の21.63%から講演毎で20.96%，発話毎で21.5%と改善が見られ，同等化前より性能が向上しているので，ヒストグラム同等化を用いたコードブック適応は有効であるといえる．

表 3　単語誤りの種類

分類	例	内容
置換	の → あの	通常の置換
挿入	→ で	通常の挿入
脱落	いう →	通常の脱落
f 置換	え → えー	フィラーの置換
f 挿入	→ えー	フィラーの挿入
f 脱落	し　下の → 下の	フィラーの脱落
表記異なり	良く → よく	表記の違いによる誤り
形態素異なり	よっ/て → よって	形態素の違いによる誤り

7. 認識結果の分析

7.1 分析方法

現状では単純に単語誤り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，どのようなところ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，目視により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，通常はフィラーは必要としないため，フィラーに関する誤りは分けてカウントした．また音響モデルでは対処の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，「よって」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，「見れ」と「みれ」が表記の異なりとなる．

例 1

正解文：　アクセント 指令 に よって INS INS 実測

認識結果：アクセント 指令 に 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%) を用いて，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いる．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．

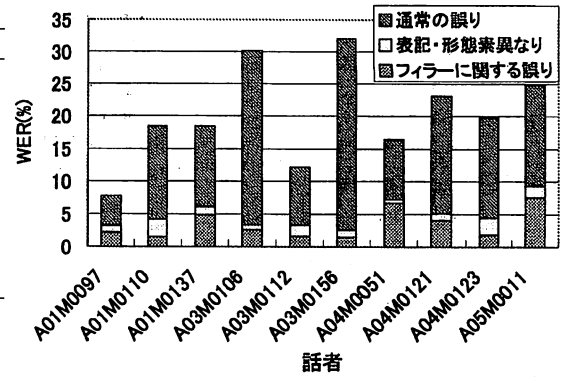

図 7　認識結果の調査 (testset1)

8. まとめと今後の課題

本研究では，クリーンな環境での離散混合分布型 HMM(DMHMM) の性能を検討するために，日本語話し言葉コーパスを用いた実験を行った．実験により，一般的な混合連続分布型 HMM(CHMM) と比べ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 20.96%の単語誤り率が得られ，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，音素毎や有音・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

[1] 小坂哲夫, 山本明祥, 加藤正治, 好田正紀：“離散混合出力分布型 HMM による講演音声認識の検討，” 信学技報, SP2005-25, pp 31-36 (2005).

[2] S. Tsakalidis, V. Digalakis and L. Newmeyer : “Efficient Speech Recognition Using Subvector Quantization and Discrete-Mixture HMMs,” Proc. of ICASSP99, pp.569-572 (1985).

[3] S. Takahashi, K. Aikawa and S. Sagayama : “Discrete Mixture HMM,” Proc. of ICASSP97, pp.971-974 (1997).

[4] 堤怜介, 加藤正治, 好田正紀：“発音変形依存モデルを用いた講演音声認識，” 信学論, Vol. J89-D No.2, pp.305-313 (2006. 2)

[5] 堀貴明, 加藤正治, 伊藤彰則, 好田正紀：“状態クラスタリングによる HM-Net の構造決定法の検討，” 信学論, D-II, J81-D-II, No.10, pp.2239-2248 (1998).