Vol.2014-NL-216 No.14
Vol.2014-SLP-101 No.14
2014/5/23

DNN-HMMを用いた音響モデルおよび言語モデルのクロス適応

高木 瑛[1,a)]　今野 和樹[1]　加藤 正治[1]　小坂 哲夫[1]

概要：近年，深層学習によるニューラルネットを用いることにより，音声認識システムの大幅な性能向上が得られることが示されている．本研究では deep neural netowrk(DNN) と隠れマルコフモデル (HMM) のハイブリッド型の音響モデル (DNN-HMM) を使用した日本語講演音声認識システムの更なる性能向上を目指し，モデル適応の検討を行った．検討する適応手法としては教師なしのバッチ適応を対象とする．教師なし適応において適応用ラベルの作成に認識結果を用いるが，誤り傾向の異なる複数の認識システムを使うことで誤りの影響を低減するクロス適応が提案されている．本研究ではこの考えに基づき DNN-HMM，GMM-HMM の 2 種類の音響モデルおよび言語モデルを加え計 3 種類のモデル適応を併用するクロス適応を提案する．提案手法を日本語話し言葉コーパス (CSJ) の評価セットを用いて評価し，その有効性を示す．

1. はじめに

近年深層学習によるニューラルネットを使用した音声認識システムが高い性能を示し，注目を集めている．音声データ量の増加および GPGPU を用いた計算性能の向上などにより，大規模な音響モデルを構築することも可能となっている．国内においてはニューラルネットワークにより得られた事後確率を HMM の状態確率として使用する DNN と HMM のハイブリッド型の音響モデル (DNN-HMM) を使用した日本語音声認識の評価も進んでいる [1][2][3]．本研究では更なる性能向上を目指し，教師なしのバッチ適応について検討を行う．Gaussian mixture ベースの HMM(GMM-HMM) においては，MAP や MLLR など有効な適応手法が種々提案されている．しかしこれら平均や分散などの統計値を用いる適応手法は DNN-HMM では利用できないため，新たな適応手法を検討する必要がある．

DNN-HMM 用の適応手法としては適応データによる再学習が検討されている [4] [3]．しかし一般に教師なし適応においては，誤りを含む教師信号に従った学習を行うため，DNN の識別器としての性能が高いほど，誤りを忠実に再現してしまうという問題が存在する．この問題に対処する方法として，モーメンタムや正則化などを用いて，過度な学習を抑制する手法が検討されている [5]．また学習データについては事前に話者が既知であることを利用した適応手法も検討されている．三村らは学習データ中から評価話者に近い話者を選択し適応する手法を提案している [3]．また落合らは話者正規化学習の DNN への応用を検討している [6]．

本研究ではクロス適応を利用した教師なし適応について検討する．教師なし適応において適応用ラベルの作成に認識結果を用いるが，誤り傾向の異なる複数の認識システムを使うことで誤りの影響を低減するクロス適応が提案されている [7]．我々は文献 [8] において，DNN-HMM と GMM-HMM の誤り傾向の違いを利用してクロス適応を行う手法を提案した．これらは音響モデル適応の組み合わせとなるが，音響モデル以外の適応法として言語モデル適応が存在する．言語モデル適応の場合は音響的特徴ではなく言語的な単語出現頻度の偏りを用いるため，音響モデル適応とは異なる誤り傾向を示す．以上よりクロス適応を行う場合に言語モデル適応も利用することにより，更なる性能向上を目指す．

2. 認識手法

本研究で用いる認識システムの構成図を図 1 に示す．本研究で用いる認識システムは，第 1 パスで triphone と bigram を用いてビームサーチを行い，単語グラフを生成し，第 2 パスでは生成した単語グラフを trigram でリスコアし認識結果を得る構成となっている．また本研究で用いる DNN-HMM の構成を図 2 に示す．入力層は特徴ベクトルの次元数と同数のノード数を持つ．一般的に DNN-HMM を用いた音声認識では複数フレームの特徴ベクトルをひと

[1] 山形大学大学院理工学研究科
Graduate School of Science and Engineering, Yamagata Uniersity
[a)] tth18357@st.yamagata-u.ac.jp

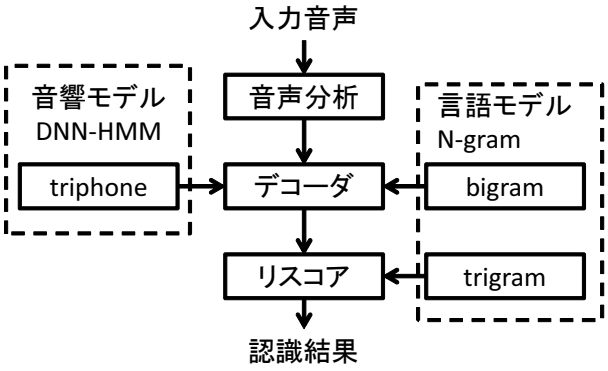

図 1 Structure of recognition system

まとめにしたセグメント特徴量が用いられる．本研究でも 11 フレームの特徴を入力とする．隠れ層の総数については日本語話し言葉コーパス (CSJ) の学習データ量では 5~7 層程度で飽和することが示されているため [3]，本研究では 7 層とした．またノード数は 512~ 2048 程度が使用されるが，本研究では 2048 とした．出力層はハイブリッド型の場合，認識に用いる HMM の総状態数に揃える必要がある．本研究では triphone を用い 3003 ノードとした．

DNN の学習は，適切な初期値を得るための pre-training と呼ばれる教師なし学習のステップと，fine-tuning と呼ばれる教師つき学習の 2 ステップからなる．pre-training は隠れ層を入力層に近い層から 1 層ごとに学習し，それを積み重ねることにより深層構造を得る．各層のモデルとしては Restricted Boltzmann Machine(RBM) を使用する．pre-training により局所最適解へ陥ることが避けられると言われており，実験によりその効果が示されている [9]. fine-tuning では，フレームごとに状態番号ラベルを与え教師つき学習を，確率的勾配降下法 (SGD) による誤差逆伝搬法で行う．損失関数にはクロスエントロピーを用いる．認識時にはベイズ則に基づくスケーリングを行って出力確率を求め HMM を用いた確率計算を行う．

3. クロス適応にもとづく教師なし適応法

教師なしのバッチ適応を行う場合，一般的に一度適応前モデルで認識を行い，その後その認識結果を使用してパラメータの更新を行う．認識結果には誤りが含まれているため教師つき適応と比較して性能が劣化する．この問題に対する対応法の一つとしてクロス適応が提案されている [7]. クロス適応の基本的な考えは誤り傾向の異なる認識システムを組み合わせ，相互に補完することにより誤りの傾向を軽減する．

我々はこれまで DNN-HMM と GMM-HMM を併用するクロス適応法を用いた話者適応について検討を行い，その有効性を示してきた [8]．DNN-HMM と GMM-HMM はいずれも音響モデルであるが，言語的な単語出現頻度の偏りを用いる言語モデル適応は，音響モデル適応とは，また異

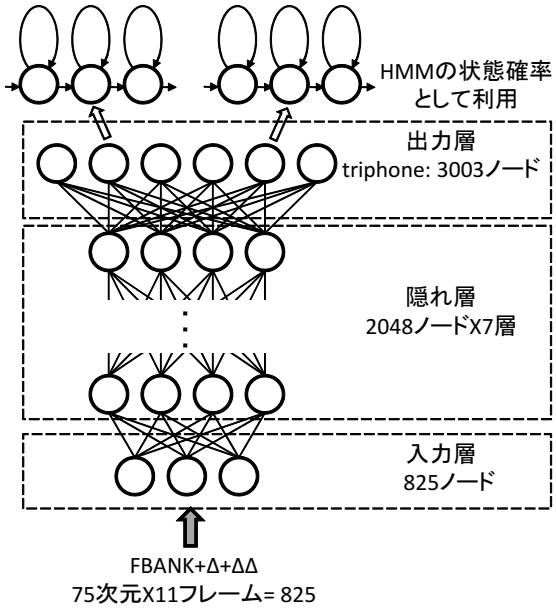

図 2 Structure of DNN-HMM

なる誤り傾向を示す．文献 [10] では GMM-HMM と言語モデル適応を併用しクロス適応することで性能向上が得られることを示している．本研究ではこれらの考えに基づき，GMM-HMM，DNN-HMM，言語モデルの 3 種のモデル適応を組み合わせ，適応性能の向上を目指す．

クロス適応では様々なインプリメントの方法が考えられるが，本研究では適応に使用するラベル生成のための認識に用いるモデルと，パラメータ更新の対象となるモデルに別種のモデルを使用することによりクロス適応の効果を得る手法を採る．

適応の手順の一例を図 3 に示す．まず適応前の DNN-HMM(DNN-HMM base) で認識を行い，認識結果の漢字仮名交じり文を変換して音素系列を得る．これを教師信号として GMM-HMM の適応を行う．本研究で用いる GMM-HMM の共分散はブロック型全共分散で表現する．これは FBANK とデルタ，デルタ・デルタ間の相関は考慮しないが，次元間の相関は考慮したものである．GMM-HMM の適応としては MLLR 法を使用した．適応サンプルから最尤推定による線形回帰係数を求めてパラメータの更新を行う．分散については共分散行列のうち対角要素のみ更新を行った．次に適応で得られたモデル (GMM-HMM adapt1) を用いて再度認識を行い，HMM 状態系列を得る．得られた状態系列を教師信号として DNN-HMM base の適応を行う．更に適応して得られた DNN-HMMadapt1 を用いて認識を行い，その認識結果を利用して適応前言語モデル (LM base) の適応を行う．以上の例では，DNN-HMMbase の認識結果で GMM-HMM の適応，GMM-HMMadapt1 の認識結果で DNN-HMMbase の適応，DNN-HMMadapt1 の認識結果で LMbase の適応と 3 通りのクロス適応が行われることになる．これはあくまで 1 例であり，適応の順番に関

Vol.2014-NL-216 No.14
Vol.2014-SLP-101 No.14
2014/5/23

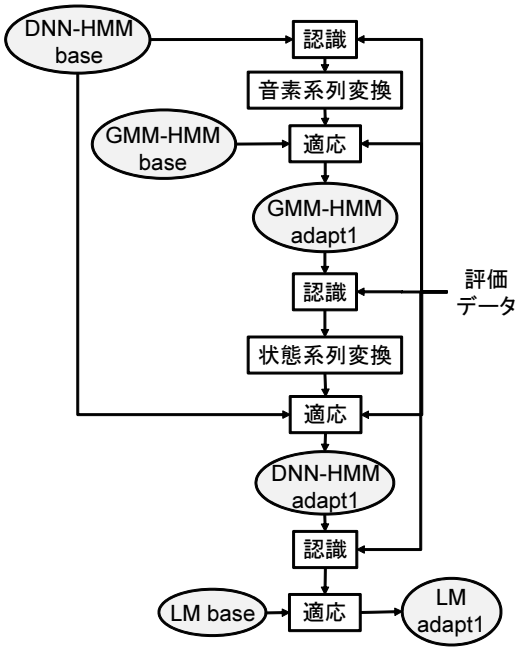

図 3 Procedure diagram of unsupervised adaptation

しては様々な組み合わせが考えられる．

図 3 に示す音素系列変換および状態系列変換の詳細を図 4 に示す．DNN-HMM の適応には GMM-HMM の認識結果，GMM-HMM の適応には DNN-HMM の認識結果を用いる．この認識結果は漢字かな混じり文の形で得られる．これを音素系列に変換するが，その際に各単語間に無音 (sil) の音素記号を候補として挿入する．実際に単語間に無音が挿入されるかは音響モデルでアライメントを取って決定する．そのアライメントの際の音響モデルとして GMM-HMM を使用する場合と DNN-HMM を使用する場合の比較をすると，GMM-HMM の方でより正しい結果が得られたため，実験ではこちらを使用する．なぜ無音の挿入に関して GMM-HMM がより高い性能が得られるかについては今後検討する必要がある．最終的には状態番号の系列あるいは音素系列を出力する．

DNN-HMM の適応手法としては fine-tuning と同じ方法を用いる．適応のパラメータとして遷移確率の更新も考えられるが，今回は DNN のみのパラメータ更新を行った．DNN の教師なし適応を行う場合，過学習が問題となる．この問題に対処する方法として，モーメンタムや正則化などを用いる手法が検討されている [5]．基本的にはモデルの自由度を制限することにより過学習を抑制する．また dropout[12] と呼ばれる学習時の各反復において，一部のノードをランダムに取り除いて学習する方法も過学習に有効と考えられる．文献 [8] において，モーメンタムおよび L2 正則化の有効性について検討したところ後者が有効であったため，本実験でも L2 正則化を利用した．

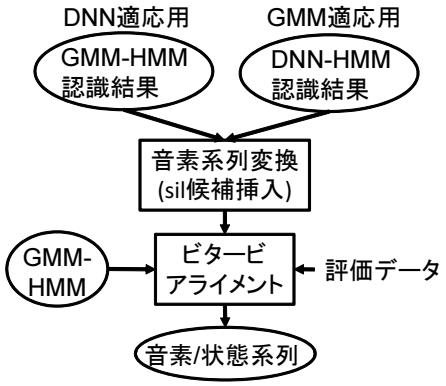

図 4 Procedure diagram of phoneme or state alignment

4. 言語モデル適応法

図 5 に今回用いた言語モデル適応法を図示する。言語モデルの教師なし適応では大量テキストから作成した単語 trigram と，認識結果および大量テキストから作成した品詞 trigram を線形補間することで, 認識に使用する適応 trigram を作成する [11]．まず，大量テキストから単語 trigram を作成し, そのモデルを用いて適応データをデコーディングし認識結果を得る．次に認識結果に含まれる品詞情報を利用して品詞からの単語の出現確率 $P(w_i|c_i)$ を推定する．また大量テキストから推定した品詞列の出現回数を用いて, 品詞連鎖確率を次式で求める．

$$P(c_i|c_{i-2}c_{i-1}) = \frac{N_0(c_{i-2}c_{i-1}c_i)}{N_0(c_{i-2}c_{i-1})} \quad (1)$$

N_0 は大量テキストから推定した品詞列の出現回数である．最後にベースラインの単語 trigram，$P(w_i|w_{i-2}w_{i-1})$ と品詞 trigram を次式のように線形補間して適応 trigram を構築する．

$$\begin{aligned} P'(w_i|w_{i-2}w_{i-1}) &= \lambda P(w_i|w_{i-2}w_{i-1}) \\ &+(1-\lambda)P(w_i|c_i)P(c_i|c_{i-2}c_{i-1}) \end{aligned} \quad (2)$$

右辺第 1 項が単語 trigram の確率，右辺第 2 項が品詞 trigram の確率である．λ は線形補間係数である．予備実験より λ は 0.7 と定めて実験を行った．

5. 実験条件

以下に音声認識実験の条件について記述する．まず DNN の学習のための状態ラベルは GMM-HMM を使用し，強制アライメントを取って作成した．GMM-HMM の音声分析条件は，フレーム長/周期が 25ms/8ms，特徴ベクトルは 12 次元の MFCC と対数パワー，及びその 1 次と 2 次の回帰係数の計 39 次元を CMN により正規化した．CSJ の学会講演および模擬講演 2667 講演を学習データとして用い最尤推定 (ML) を行った．共分散の型はブロック型全共分散で総状態数および混合数は 3003 状態，32 混合であ

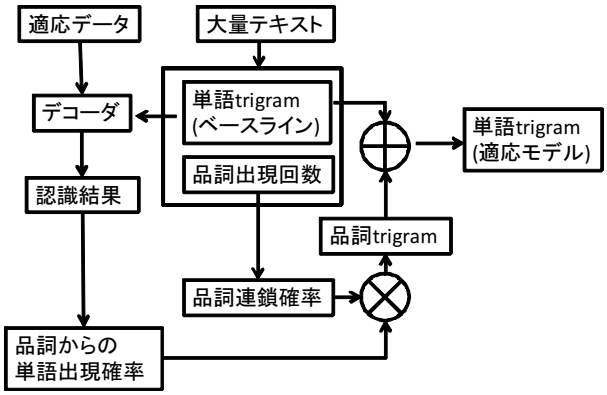

図 5 Procedure diagram of language model adaptation

表 1 Conditions for DNN training

pre-training	
初期学習係数	0.4 (1 層目のみ 0.01)
エポック数	10 (1 層目のみ 20)
ミニバッチサイズ	1024
モメンタム	0.9 (最初の 50 時間データのみ 0.5~ 0.9 へ増加)
L2 正則化係数	0.0002
fine-tuning	
初期学習係数	0.008
エポック数	交差検定によりフレーム認識率向上が 0.1%未満の場合停止
ミニバッチサイズ	512

る．次に DNN-HMM の学習について述べる．入力特徴量は 24 次対数メルフィルタバンクと対数パワー，及びその 1 次と 2 次の回帰係数の計 75 次で，これを計 11 フレームのセグメント特徴 ($75 \times 11 = 825$ 次元) として使用する．また平均分散正規化を行う．また学習は CSJ の学会男性女性話者 963 講演 (203 時間) を用いる．学習のための諸条件を表 1 に示す．これらの設定はミニバッチサイズ以外は文献 [13][14] とほぼ同様であり，細かな検討は行っていない．fine-tuning では学習データから 1/10 のデータをランダムに取り出しヘルドアウトデータとして交差検定を行いフレーム認識率向上が 0.1%未満で学習の繰り返しを停止する．言語モデルの語彙セットは学会講演及び模擬講演から出現回数 2 回以上の単語を合わせた 47,099 語とする．言語モデルは第 1 パスでバイグラム，第 2 パスでトライグラムを用い，総単語数約 6.68M の CSJ の学習データより生成する．評価データは CSJ の testset1，学会男性 10 講演を用いる．DNN の学習には Kaldi tool kit[13] を用いた．また認識には研究室独自の 2 パスデコーダを用いる．

教師なし適応について，モーメンタム，L2 正則化係数，学習係数，ミニバッチサイズについて複数の値を用いて比較検討を行った．この結果モーメンタムは 0，即ち使用せず，L2 正則化係数は 0.0002，学習係数は 0.0001，ミニバッチサイズは 2048 と設定した．

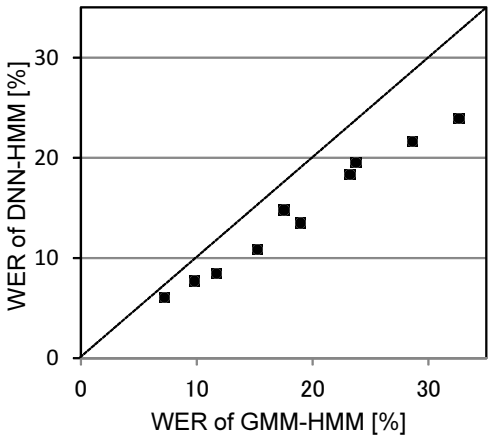

図 6 Word error rate for each speaker

6. 認識実験結果

まずベースラインとなる適応前の音声認識結果を示す．学習用状態ラベル作成のための GMM-HMM の単語誤り率 (WER) は 19.75% であるのに対し，DNN-HMM の WER は 15.12%と向上した (DNN-HMMbase)．このときの両者の話者ごとの WER を図 6 に示す．図から分かるように，認識精度の低い話者ほど改善率が高くなっている．しかし，両者は高い相関を示しており，認識し易い話者，認識が難しい話者については変わりがなく，依然として話者性の問題が存在することが分かる．

次にクロス適応の結果を図 7 に示す．この図では様々な順番でモデル適応した場合の WER を示している．また下線で示す値は音素ミスマッチ率 (PMR:Phoneme mismatch rate) であり，2 つの認識結果の誤り傾向の違いを示す指標として使用している．2 つの認識結果を音素系列に変換し，片方を正解，片方を認識結果と見立てて置換，脱落，挿入を考慮した誤り率を求めることにより算出する．値が大きいと 2 つの誤り傾向が異なると判断できる．ただし両者の WER に差があるとその影響も入るので解釈には注意が必要である．

まず DNN-HMMbase の認識結果を利用して DNN-HMM の適応を行った (DNN-HMMadapt1a)．この場合は同種モデルで認識および適応を行っているので，クロス適応とはならない．図における epo はエポック数 (適応繰り返し回数) を表しており，文献 [8] の検討結果より 100 とした．このときの WER は 14.72%となった．

次にクロス適応の場合として，DNN-HMMbase の認識結果を用いて言語モデル適応や GMM-HMM の適応を行った場合の結果を述べる．言語モデル適応を行った場合 (LMadapt1b) では WER が 14.73%, GMM-HMM の適応を行った場合 (GMM-HMMadapt1) では WER が 14.53%となり，3 種の中で最良の結果が得られた．PMR を比較す

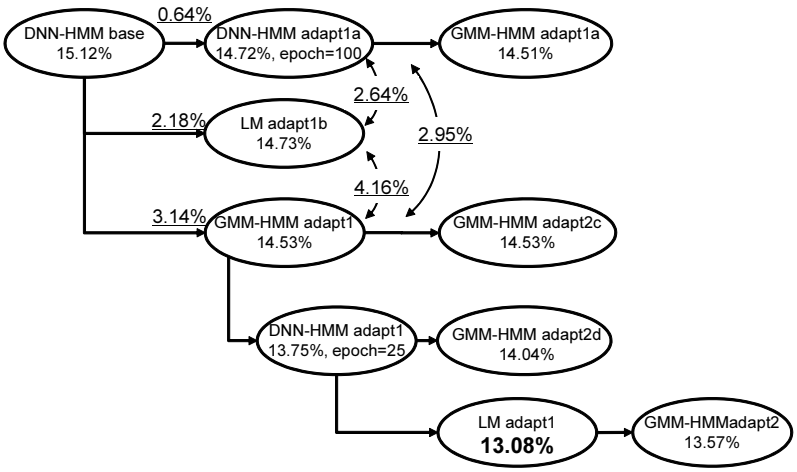

図 7 Word accuracy using cross adaptation

表 2 Comparisons of substitution, insertion and deletion errors (%)

Type of errors	DNN-HMMbase	DNN-HMMadapt1a	LMadapt1b	GMM-HMMadapt1
Sub	9.57	9.35	8.99	9.30
Ins	2.65	2.39	2.40	1.96
Del	2.89	2.98	3.34	3.27
WER	15.12	14.72	14.73	14.53

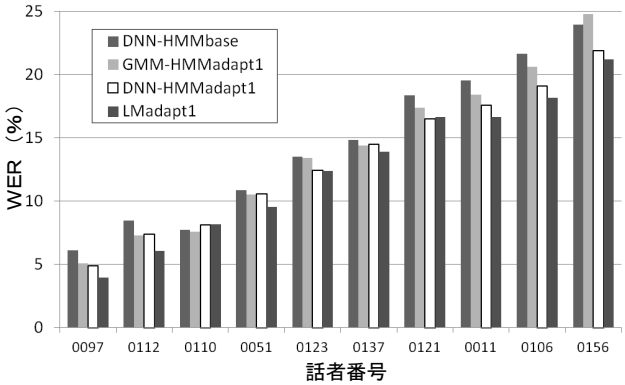

図 8 Results of adaptation for each speaker

ると GMM-HMM 適応で一番 PMR が大きくなっており，ベースラインと比較して誤り傾向の違いが大きいことが分かる．一方 DNN-HMM の適応を繰り返した場合の PMR は一番小さくなっており (0.64%)，誤り傾向がベースラインと類似していることが分かる．表 2 に以上の 3 者の単語誤りの内訳を，置換，挿入，脱落に分けて示した．DNN-HMMadapt1a と GMM-HMMadapt1 を比較すると，挿入誤りと脱落誤りの割合が異なり，GMM-HMMadapt1 では挿入誤りが減少し，脱落誤りが増加していることが分かる．実際の認識結果を確認するとフィラー等の挿入誤りが減少している傾向が見られる．一方 LMadapt1b では置換誤りが減少しているのが特徴的である．実際の認識結果では同音異義語の改善が目につくが，これは置換誤りの減少として現れる．以上のように適応ごと誤りの傾向がそれぞれ異なり，これによりクロス適応の効果が得られていると考えられる．

さらに一番結果の良かった GMM-HMMadapt1 の後に様々な適応をした結果も図に示している．GMM-HMM 適応を繰り返して行った場合 (GMM-HMMadapt2c) は性能の向上は見られず認識性能は飽和した．一方クロス適応と言える DNN-HMM の適応を行った場合は，更に認識性能が向上し 13.75%が得られた．その後に言語モデルを適応することにより (LMadapt1) 今回の適応実験の最良値 13.08%を得た．このように GMM-HMM→DNN-HMM→LM と異なる種類の適応を順次行うことにより，高い適応性能が得られることが分かった．以上を繰り返して行うことにより更なる性能向上が得られることも予想されたため，更に GMM-HMM の適応を行ったが (GMM-HMMadapt2) 性能は逆に低下し 13.57%となった．この場合の誤り傾向を分析すると，脱落誤りの増加が認められた．GMM-HMM をクロス適応に使用した場合，今回の実験全体を通じて脱落誤りが増加する傾向があることが分かった．

最良の結果 13.08%が得られた条件における各話者の認識性能の推移を図 8 に示す．多くの話者では適応ごとに順次性能が向上するが，いくつか例外も存在する．話者 0110 はいずれの適応もあまり効果が無い．また 0156 のように GMM-HMM の適応で性能が劣化する場合や，0123 や 0121 のように LM 適応が効果的ではない話者も存在する．話者による適応の効果の出方の違いについては今後検証が必要である．

以上より GMM-HMM，DNN-HMM および LM の 3 種の適応を組み合わせることによりクロス適応の効果が得られ良い性能が得られることが分かった．一方適応の順序については網羅的な実験は行っていないため，この順番が良いかどうかは今後の検討が必要である．図 9 に各種適応実

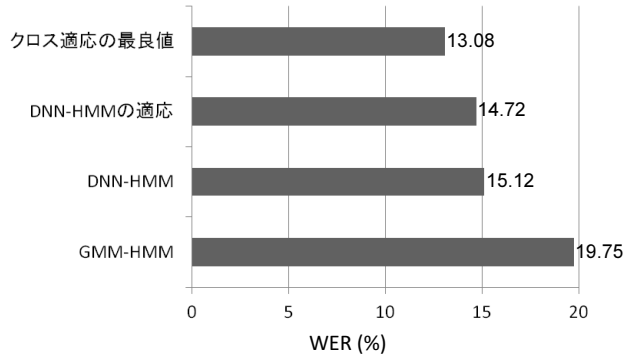

図 9 Summary of recognition results

験のまとめを示した．

7. まとめ

本研究では DNN-HMM を使用した日本語講演音声認識システムの更なる性能向上を目指し，教師なしバッチ適応の検討を行った．教師なし適応において適応用ラベルの作成に認識結果を用いるが，誤り傾向の異なる複数の認識システムを使うことで誤りの影響を低減するクロス適応が提案されている．本研究ではこの考えに基づき DNN-HMM，GMM-HMM の 2 種類の音響モデルおよび言語モデルを加え計 3 種類のモデル適応を併用するクロス適応を提案した．また提案手法を日本語話し言葉コーパス (CSJ) の評価セットを用いて評価を行った．この結果 GMM-HMM，DNN-HMM，言語モデルの 3 種類の適応法を組み合わせるクロス適応で最良の結果が得られた．また分析の結果，適応の種類によって誤り傾向が異なることが分かった．今回は DNN-HMM の教師なし適応法としては単純な再学習を行ったが，ニューラルネットの過学習に考慮した適応手法を導入するなどして [5]，性能向上を図っていく予定である．

謝辞 本研究の一部は科研費（課題番号 25330183）によった．

参考文献

[1] 西野大輔, 篠田浩一, 古井貞熙: “ディープラーニングを用いた日本語大語彙話し言葉音声認識,” 音響講論秋, 2-1-7 pp.71–72 (2012).
[2] 神田直之, 武田徹, 大渕康成: “Deep Neural Network に基づく日本語音声認識の基礎評価,” 情報処理学会研究報告, 2013-SLP-97(8), pp. 1–6 (2013).
[3] 三村正人, 河原達也: “CSJ を用いた日本語講演音声認識への DNN-HMM の適用と話者適応の検討,” 情報処理学会研究報告, 2013-SLP-97(9), pp. 1–6 (2013).
[4] Y. Xiao, et al.: “A initial attempt on task-specific adaptation for deep neural network-based large vocabulary continuous speech recognition,” Proc. of Interspeech2012, (2012).
[5] H. Liao: “Speaker adaptation of context dependent deep neural networks,” Proc. of ICASSP2013, (2013).
[6] 落合翼, 松田繁樹, X. Lu, 堀智織, 片桐滋: “話者正規化学習されたディープニューラルネットワークによる教師なし話者適応,” 日本音響学会春季講演論文集，1-4-18 (2014).
[7] S. Stuker, et al.: “Cross-system adaptation and combination for continuous speech recognition: The influence of phoneme set and acoustic front-end,” Proc. of InterSpeech2006, pp.5212–524, (2006).
[8] 小坂哲夫, 今野和樹, 高木瑛, 加藤正治: “DNN-HMM を用いた日本語講演音声認識における話者適応の検討,” 日本音響学会春季講演論文集，1-4-17 (2014).
[9] A. Mohamed, G. Hinton and G. Penn: “Understanding how deep belief networks perform acoustic modelling,” Proc. of ICASSP2012, (2012).
[10] T. Kosaka, T. Miyamoto and M. Kato: “Unsupervised cross-adaptation approach for speech recognition by combined language model and acoustic model adaptation,” Proc. of APSIPA ASC 2011, (2011).
[11] 堤怜介，加藤正治，小坂哲夫，好田正紀：“発音変形依存モデルを用いた講演音声認識,” 電子情報通信学会論文誌 Vol.J89-D No.2, pp.305-313 (2006).
[12] G.E. Dahl, T.N. Sainath and G.E. Hinton: “Improving deep neural networks for LVCSR using rectified linearunits and dropout,” Proc. of ICASSP2013, (2013).
[13] Kaldi project: “The Kaldi speech recognition toolkit,” http://kaldi.sourceforge.net/index. html
[14] K. Vesely, A. Ghoshal, L. Burget, and D. Povey: “Sequence-discriminative training of deep neural networks,” Proc. of Interspeech2013, (2013).