論文 / 著書情報
Article / Book Information

論題(和文)	識別学習モデルと教師なしCV適応を用いたCSJ講演音声認識
Title(English)	
著者(和文)	篠崎 隆宏, 久保田 雄, ディクソン・ポール, 古井 貞熙
Authors(English)	Takahiro Shinozaki, Yu Kubota, Paul Dixon, SADAOKI FURUI
出典(和文)	日本音響学会2010年春季講演論文集, , No. 1-6-14, pp. 37-38
Citation(English)	, , No. 1-6-14, pp. 37-38
発行日 / Pub. date	2010, 3

識別学習モデルと教師なし CV 適応を用いた CSJ 講演音声認識*

○篠崎隆宏, 久保田雄, ディクソン・ポール, 古井貞熙 (東工大)

1 はじめに

我々はこれまでの研究で，教師無しバッチ適応の繰り返し推定の枠組内部にクロスバリデーション (CV) 手法を組み込んだ教師無し CV 適応を提案し，従来の教師無しバッチ適応と比較して高い適応効果が得られることを示した [1]．しかし，初期モデルとして使用した不特定話者モデルは EM 法による最尤推定モデルであり，より高い認識性能が得られる識別学習モデルを初期モデルとした場合の効果については検討が行われていなかった．そこで，本研究では教師無し適応の初期モデルとして識別学習モデルを用いた際の CV 適応の効果について評価を行う．さらに，システム内部で複数の認識処理を行うという点で類似性のあるクロスアダプテーション [2] との比較も行う．評価実験は，日本語話し言葉コーパス CSJ を用いて行った．最小の単語誤り率は，9 万語の大語彙辞書と識別学習モデルを用いた WFST ベースの認識システムにおいて，CV 適応を用いた際に得られた．以下では，まず CV 適応とクロスアダプテーションについて簡単に説明した後，実験条件と実験結果について示す．

2 CV 適応およびクロスアダプテーション

従来の教師無しバッチ適応では，初期モデルを用いて認識対象音声の認識を行い，それにより得られた認識仮説を用いて MLLR 法などによりモデルパラメタの更新を行う操作を繰り返す．このプロセスでは認識誤りを含めて再推定されたモデルが次の認識ステップで同じデータの認識に使用されるために，同じ認識誤りが適応ループ中で繰り返され，モデルが誤った方向に適応されてしまう問題が考えられる．

これに対し CV 適応では，Fig. 1 に示すように教師無し適応の認識およびモデル更新ステップで用いるデータを CV により分離する．即ち，CV 適応では適応音声発話集合を K 個の排他的な部分集合に分割する．そしてモデル更新ステップでは K 個のうちの 1 つを除いてモデルを更新する操作を，取り除く部分集合を変えながら K 回行い，K 個のモデルを作成する．それに続く認識ステップでは，k 番目の部分集合を認識するのに k 番目のモデルを使用する．これにより，

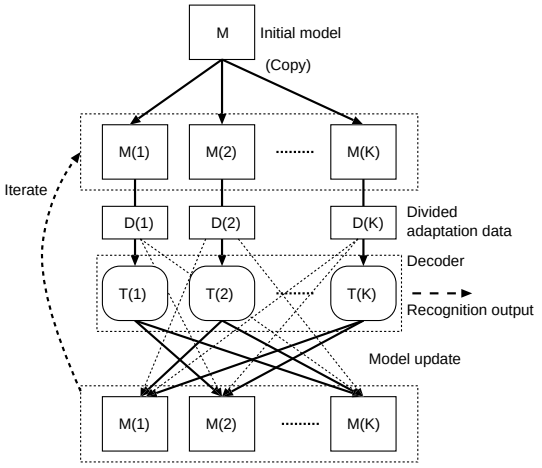

Fig. 1 CV adaptation

モデル更新ステップと認識ステップにおいて使用されるデータが独立となり，同じ認識誤りが繰り返されることを防ぎ，モデル適応化性能の向上を図ることが出来る．各モデルの推定に直接使用するデータは 1 サブセット分少なくなるが，K を大きくとることでデータ量の減少は最小限に抑えることが出来る．

クロスアダプテーションでは，Fig. 2 に示すように 2 つの認識システムを用いる．適応に用いる音声データは 2 つのシステムで共通であるが，波形データから抽出する特徴量や探索アルゴリズムを変えることなどにより，2 つのシステム間での認識誤りが出来るだけ独立となるようにする．そして，モデル更新の際にお互いの認識結果を交換して用いることで，適応化性能を向上させる．この手法では，2 つの認識システムが同一では認識率向上効果はまったく得られ無い．また片方の認識システムの性能がもう片方と比べて極端に低い場合，全体の性能が低下する場合もある．

3 実験条件

従来のバッチ適応，CV 適応およびクロスアダプテーションを MLLR を用いた話者適応に応用し，評価を行った．使用した音声特徴量は MFCC12 次元と対数エネルギー，およびそれらのデルタ項とデルタデルタ項の計 39 次元である．音響モデルは 3000 状態状態共有混合ガウス分布トライホンモデルであり，日本語話し言葉コーパス CSJ [3] の学会講演音声より

*CSJ lecture speech recognition based on a discriminatively trained acoustic model and CV adaptation. by Takahiro Shinozaki, Yu Kubota, Paul Dixon, and Sadaoki Furui (Tokyo Institute of Technology)

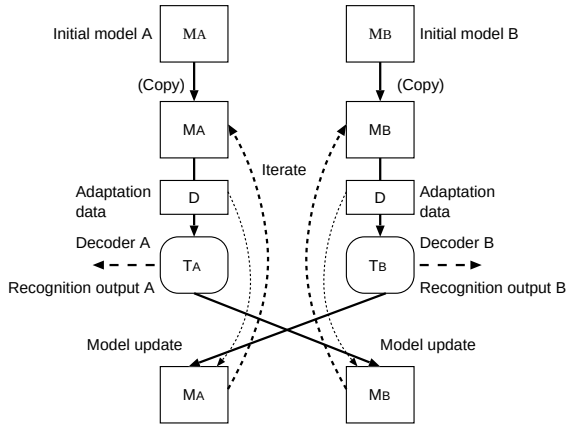

Fig. 2　Cross adaptation

Table 1　Base systems for cross-adaptation

Decoder	Model	Feature	WER
T^3	MPE	MFCC	19.3
T^3	MPE	PLP	19.7
Julius	MPE	PLP	21.1

最小音素誤り基準 (MPE) により学習した．学習データ量は 254 時間であり，各状態の混合数は 32 である．音声認識システムは T^3 WFST 認識器 [4] である．クロスアダプテーションの評価では 2 つのベースシステムが必要となるため，これらに加えて，PLP12 次元と対数エネルギー，およびそれらのデルタ項とデルタデルタ項の計 39 次元を用いた MPE 音響モデル，及び Julius 認識器 [5] を使用した．また比較のため，EM 学習による最尤推定 (ML) モデルも用いた．

言語モデルは CSJ の学会および模擬講演 6.8M 単語から学習したトライグラムモデルであり，辞書サイズは 30k である．テストセットは男性話者による学会講演 10 講演からなる CSJ 評価セットである．なお評価の際，形態素境界の不一致も認識誤りとして扱っている．

4　実験結果

クロスアダプテーションでは，特徴量として MFCC と PLP を用いた 2 つの T^3 認識器を組み合わせたシステム，および MFCC を用いた T^3 認識器と PLP を用いた Julius 認識器を組み合わせたシステムの，2 通りについて評価を行った．Table 1 に各要素システムの単語誤り率を示す．クロスアダプテーションシステムの出力としては，これらの中で誤り率が一番低い MFCC を用いた T^3 側の出力を用いた．

Fig. 2 に，従来のバッチ適応，CV 適応 (20-fold CV) およびクロスアダプテーションを行った場合の単語誤り率を示す．初期モデルとして MPE モデルを用いた

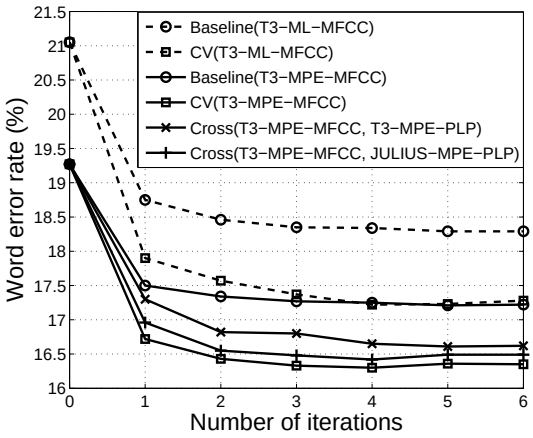

Fig. 3　Number of adaptation iterations and word error rate. Zero-th iteration is the result of the speaker-independent model. Batch-mode baseline adaptation result is denoted as Baseline. CV- and cross-adaptation are denoted as CV and Cross, respectively.

方が ML モデルを用いるよりも低い単語誤り率が得られた．また，CV 適応およびクロスアダプテーションどちらの場合も従来法よりも低い単語誤り率が得られた．クロスアダプテーションでは特徴量のみを変えたシステムを組み合わせるよりも，異なる認識器を組み合わせた方がより低い単語誤り率が得られた．CV 適応では若干ながら，それらよりもさらに低い単語誤り率が得られた．

CV 適応では単一のベースシステムしか必要としないことも利点である．補足実験として，MITLM ツールキットを用いた 90k 語彙の言語モデルを用いたところ T^3 システムではさらに単語誤り率が減少し，不特定話者モデルで 18.1%，従来バッチ適応で 16.1%，CV 適応で 15.5%の単語誤り率が得られた．

5　まとめ

識別学習モデルを初期モデルとした場合においても CV 適応が従来のバッチ型適応と比較して効果的であり，またクロスアダプテーションと比べてもより低い単語誤り率が得られることを示した．

参考文献

[1] T. Shinozaki et al., ICASSP, pp. 4377–4380 2009.
[2] H. Soltau et al., ICASSP, pp. 205–208 2005.
[3] T. Kawahara et al., SSPR2003, pp. 135–138, 2003.
[4] P. R. Dixon et al., IEEE ASRU, pp. 443–448, 2007.
[5] A. Lee et al., ICSLP, pp. 1831–1834, 1998.