T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	高精度音声認識のための教師なしクロスバリデーション適応法の提案
Title(English)	
著者(和文)	篠崎 隆宏, 久保田 雄, 古井貞熙
Authors(English)	Takahiro Shinozaki, Yu Kubota, SADAOKI FURUI
出典(和文)	日本音響学会2009年春季講演論文集, , No. 1-5-10, pp. 27-28
Citation(English)	, , No. 1-5-10, pp. 27-28
発行日 / Pub. date	2009, 3

Powered by T2R2 (Science Tokyo Research Repository)

高精度音声認識のための教師なしクロスバリデーション適応法の提案*

○篠崎 隆宏, 久保田 雄, 古井 貞熙 (東工大)

1 はじめに

音声認識において高い認識性能を得るためには，認識対象とする音声と録音環境や発話スタイルの出来るだけ近い大量の音声データとその書き起こしを収集し，音響モデルを学習することが有効である．さらに，認識対象話者の音声データと書き起こしが事前に得られれば，その話者に特化したモデルを作成することで，より高い認識性能が得られる．しかし，大量のデータを収集しその書き起こしを作成するためには非常に大きな手間とコストがかかるため，アプリケーション毎に音声コーパスを用意することは非現実的である．また，話者毎にデータを収集することは音声認識の利便性を低下させたり，そもそも不可能であることも多い．その為，多くの場合においては既存のコーパスより学習した不特定話者モデルを用いて音声認識を行うことが必要となる．

そこで音声認識技術の実用化の観点からは，認識タスクや話者に依存した音声データを事前に必要とせずに一般的な不特定話者モデルと認識対象の音声のみを用いて高い認識性能を実現するための，効果的な教師なし適応技術が非常に重要となる．教師無し適応技術には様々なものがあるが，音声認識において広く用いられている枠組として，教師無しバッチ適応が挙げられる．教師無しバッチ適応では，まず音声認識器を適応対象音声に適用して認識仮説を得，ついでその認識仮説を適応用の書き起こしとした教師あり適応を行なう．教師あり適応手法としてはMLLRやMAPなどが用いられる．更に，得られた適応モデルを元に認識を行ない同様の適応化処理を繰り返すことで，より高い認識性能を得ることができる．

しかし，教師無し適応において常に問題となるのが，認識器により生成した認識仮説には認識誤りが避けられないことである．更に，バッチ型適応では認識処理とモデル更新処理が同じデータを用いて繰り返されるため，適応の繰り返しとともに認識誤りも強化されてしまい，適応後の認識性能を制限する要因となっている．

本研究では，バッチ型教師無し適応における認識誤りの影響を低減させ効果的な適応を可能とする新しい教師無し適応の枠組として，教師無しクロスバリデーション（CV）適応法を提案する．

2 教師無しクロスバリデーション（CV）適応法

従来のバッチ型教師無し適応の問題点として，認識処理とモデル更新処理で同じデータを用いるために認識誤りが適応ループ中で繰り返し強化されてしまうことが挙げられる．そこで，認識ステップとモデル更新ステップで使用するデータを分離することで繰り返しループにおける認識誤りの影響を低減するための手法として，教師無しクロスバリデーション（CV）適応法を提案する．

図1に教師無しCV適応法の適応プロセスを示す．教師無しCV適応法では，適応対象の発話セット全体をほぼ同じサイズの K 個の排他的な部分集合に分割する．最初の認識ステップは基本的に従来のバッチ型適応と同じであり，K 個の部分集合に対して単に同じ初期モデルを用いて認識処理が行なわれる．次のモデルパラメタ更新ステップでは，従来のバッチ型適応が全ての適応データを用いてただ一つのモデルを作成するのに対し，教師無しCV適応法では K 個の部分集合のどれか一つを取り除いた K 個のCVモデルを作成する．(k 番目のCVモデルを作成する際の適応初期モデルとしては，その前のループの k 番目のCVモデルを使用した．) そして，次回以降の認識ステップでは各発話部分集合に対し，直前のモデル更新ステップで作成されたCVモデルのうちで，その部分集合を除いて作成したモデルを用いて認識処理を行なう．認識ステップとモデル更新ステップを従来のバッチ適応と同様に何度か繰り返し，最終的な適応化モデルによる認識仮説は最後の認識ステップにおいて生成される K 個の部分集合の認識仮説を1つに集めることにより得られる．

このようなプロセスに従うことで，認識ステップとモデル更新ステップにおけるデータの重畳を効果的に避けることができる．更に，データの分割はCV手法により行なわれるため，各CVモデルは適応データ全体の $(K-1)/K$ を使って推定され，K をある程度大きくとれば推定に使われる実質的なデータ量の減少は無視できる程度に小さくすることができる．CV適応法の認識ステップの計算量はモデル読み込みのオーバーヘッドを除けば K に対し一定であり，モデル更新ステップの計算量は K に比例する．

*Unsupervised cross-validation adaptations for improved speech recognition. by Takahiro Shinozaki, Yu Kubota, Sadaoki Furui (Tokyo Institute of Technology)

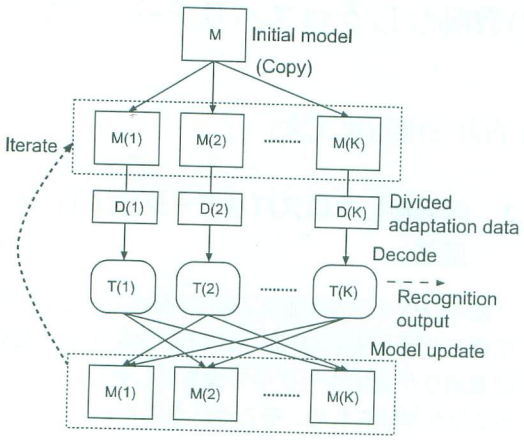

Fig. 1 Cross-validation (CV) adaptation.

3 効率化に関する検討

前章で示した教師無し CV 適応法ではモデル更新ステップにおいて K 個のモデルを作成するため，モデル更新ステップの計算量は K の値に比例する．教師無し CV 適応法は MLLR など様々なモデル更新手法と組み合わせて用いることが出来るが，特定のモデル更新手法を仮定し，若干の近似を導入することで，モデル更新ステップの計算量を削減することも可能である．例えば，MLLR との組み合わせの場合であれば，認識仮説を CV 手法により交換する代わりに認識仮説から求めた MLLR 十分統計量を交換することにより，モデル更新ステップの計算量を減らすことができる．以下では，本手法を効率的な (efficient)CV 手法という意味で教師無し ECV 適応法と呼ぶ．

4 実験条件

提案手法を MLLR を用いた教師無し話者適応に応用した．評価セットは 10 名の異なる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．各講演はおよそ 10 分から 20 分程である．音響モデルは日本語話し言葉コーパス CSJ [1] の学会講演音声 254 時間より EM 学習した 3000 状態 32 混合の状態共有トライフォンである．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．言語モデルは CSJ の学会講演および模擬講演 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．音声認識デコーダは T^3 WFST 認識器を用いた [2]．MLLR 適応には，32 の葉ノードをもつ回帰木を用いた．CV の分割数 K は予備実験より 20 とした．およそ 10 から 20 程度より K を大きくとれば，安定した性能が得られる．

Table 1 Word error rates. (Batch: batch mode adaptation, CV: CV adaptation, ECV: efficient CV adaptation. Zero-th iteration indicates speaker independent initial model)

	# of iterations					
Adpt	0	1	2	3	5	8
Base	22.5	20.3	20.1	19.9	19.9	19.9
CV	22.5	19.7	19.4	19.2	19.1	18.8
ECV	22.5	19.7	19.3	19.2	19.1	19.0

5 実験結果

表 1 に従来のバッチ型適応法，提案 CV 適応法，および ECV 適応法を用いた MLLR 教師無し話者適応の認識結果を示す．バッチ型適応法と比べて CV 法および ECV 法どちらも繰り返し 1 回目から高い認識性能が得られ，また適応処理の繰り返しによる性能の向上効果も大きい．認識ステップを含めた計算量は 1 繰り返しループ当たり，バッチ型適応法が 4.4 時間，CV 適応法が 12.4 時間，ECV 適応法が 7.4 時間であった．CV 法と ECV 法で若干単語誤り率が異なるのは，ECV 法において導入した近似や，ECV 法の現在の実装では回帰木を全ての CV モデルで共通としているなどの違いによるものである．不特定話者モデルを基準とした相対的な単語誤り率の削減率はバッチ型適応法が 12%，CV 適応法が 17%，ECV 適応法が 16%であり，提案法を用いることで従来法と比較して大幅に高い適応効果が得られた．

6 まとめ

様々なモデル更新法と組み合わせることのできる教師無し CV 適応法および MLLR を仮定することで計算量を削減した教師無し ECV 適応法の提案を行った．大語彙連続音声認識実験により，提案法により従来のバッチ型適応法よりも大幅に高い適応効果が得られることを示した．今後の課題としては音声認識に限らず様々な教師無し適応への応用が挙げられる．

謝辞 本研究は科研費 (19700167) の助成を受けたものである。

参考文献

[1] Kawahara, Nanjo, Shinozaki, Furui, *Proc. SSPR*, 135–138, 2003.

[2] Dixon, Caseiro, Oonishi, Furui, *Proc. ASRU*, 443–448, 2007.