論文 / 著書情報
Article / Book Information

論題(和文)	教師なしクロスバリデーション適応法の諸条件における評価
Title(English)	
著者(和文)	久保田 雄, 篠崎 隆宏, 古井 貞熙
Authors(English)	Yu Kubota, Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	情報処理学会研究報告, IPSJ SIG Technical Report, Vol. 2009-SLP-77, No. 7,
Citation(English)	, Vol. 2009-SLP-77, No. 7,
発行日 / Pub. date	2009, 7
権利情報 / Copyright	

Vol.2009-SLP-77 No.7
2009/7/17

教師なしクロスバリデーション適応法の諸条件における評価

久保田 雄　篠崎隆宏　古井貞熙

音響モデルの効率的な教師なし適応を目的として提案し話者適応においてその効果を示した教師なしクロスバリデーション適応法[1]について，さらに諸条件において評価を行なう．提案手法はこれまでに 15 分程度の音声データに対する話者適応ではその効果が示されたが，数分程度の比較的短い音声データではその効果は評価がされていなかった．また話者適応以外のケースにおける提案手法の応用についても確かめることが課題となっていた．そのため本実験では短時間の発話データを用いた話者適応，およびタスク適応へ提案手法を応用して単語誤り率や演算時間について評価を行う．結果，話者適応では短時間の適応音声に対しても提案法は従来法と比べ低い単語誤り率であり，タスク適応においても全ての実験条件において提案法が従来法を上回った．また話者適応では 20-fold 教師なし CV 適応法提案法の計算時間は従来法のおよそ 3 倍であった．

Unsupervised Cross-validation Adaptations in Various Conditions

Yu KUBOTA, Takahiro SHINOZAKI
and Sadaoki FURUI

Unsupervised cross-validation adaptation has previously been proposed for effective acoustic model adaptation, and has been shown to work well for speaker adaptation on voice data of around 15 minutes, but the effect on voice data of shorter duration, or the application on other adaption tasks, has not been previously evaluated. In this paper we apply this method to speaker adaptation using voice data of a few minutes and task adaptation. The results show that the word error rate of unsupervised cross-validation is smaller than that of the batch-mode baseline, both for speaker adaptation using voice data of every length, as well as for task adaptation. The computing time for unsupervised 20-fold cross-validation was three times longer than baseline in the case of speaker adaptation.

1. はじめに

音声認識において話者性や周囲環境の違いは認識性能に影響を与える大きな要因であり，連続音声認識システムにおいて高い認識精度を得るためには認識対象音声データと話者特徴やタスク特徴が類似した学習用音声データを用いて音響モデルを学習する必要がある．しかし音響モデルの学習は膨大な量の学習データを必要とするため，認識対象音声データ毎に学習データベースを用意することはコストがかかりすぎ非現実的である．そのため既存の適当なデータベースを用いて学習を行い不特定話者モデルを作成し，そのモデルを話者特徴，タスク特徴に対して適応させ，認識対象音声データに対して認識性能の高いモデルを推定する手法が実用上非常に重要である．

適応手法には様々な手法が存在するが，音響モデルの適応化の枠組みとして広く用いられているものにバッチ型教師なし適応法がある．バッチ型教師なし適応法は書き起こしを必要としない点が大きなメリットであり，適応対象音声データの認識とそこで得られた認識仮説を教師信号として用いる適応の繰り返しを行うのが一般的である．しかし認識仮説には認識誤りが避けられず，音響モデルが認識誤りに対しても適応してしまうことで適応化性能の低下が起こる問題がある．

そこで我々は適応化性能の低下を軽減する手法として教師なし CV 適応法[1]を提案し，15 分程度の講演音声データに対する話者適応においてその成果を示した．しかし数分程度の比較的短い発話時間の話者音声データに対してはその効果は確かめられておらず，また話者適応以外のケースにおける提案手法の応用についても確かめることが課題であった．

そのため本研究では比較的短時間の話者音声データに対する適応やタスク適応およびタスク適応と話者適応を組み合わせたケースにおいて教師なし CV 適応法を適用することでその有効性を示す．また同時に計算時間の比較や教師あり適応との適応化性能の比較を行う．

2. 教師なしクロスバリデーション (CV) 適応法

本章ではまず教師なしクロスバリデーション (Cross-validation:CV) 適応法の概要とそのバリエーションについて説明し，その後教師なし適応を用いた話者適応とタスク適応の概要について述べる．

2.1 教師なし CV 適応法

図 1 に提案手法である教師なし CV 適応法[1]のプロセスを示す．教師なし CV 適応法は

Vol.2009-SLP-77 No.7
2009/7/17

バッチ型教師なし適応における適応の繰り返しにおいて問題となる認識誤りの強化の悪循環を防ぐための手法である．従来のバッチ型教師なし適応ではまず初期モデルを用いて適応用音声データを認識し，得られた認識仮説を教師ラベルとして用いることでモデルのパラメタ更新を行う．そして以後得られたモデルで再び適応用音声データの認識と認識仮説によるモデルのパラメタ更新を繰り返す．しかし教師なし適応では認識仮説内の認識誤りが避けられず，従来のバッチ型教師なし適応ではパラメタ更新時にこの認識誤りに対してもモデルが適応してしまう．そして再び同じ音声データを認識することにより繰り返しの過程で同じ認識誤りが強化されてしまうという問題がある．これは音声データの認識ステップとパラメタ更新ステップで同じデータを使っているためである．

教師なし CV 適応法では繰り返しループ中に K-fold CV 手法を導入し認識ステップとパラメタ更新ステップにおいて使用される適応データを分離することにより，従来のバッチ型適応の問題を効果的に抑制する．具体的にはまず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 D(1)〜D(K) に分割し，初期モデルを用いてそれぞれの部分集合に対し認識処理を行うことで K 個の認識仮説 T(1)〜T(K) を得る．そして従来のバッチ型教師なし適応では次のパラメタ更新ステップで全ての認識仮説を用いてただ一つのモデルを作成するのに対し，教師なし CV 適応法では K 個の部分集合のうち i 番目を除いた $K-1$ 個の部分集合を用いてモデルを作成する．これを全ての部分集合について行うことで，図 1 の M(1) から M(K) に該当する K 個の CV モデルを作成する．(i 番目の CV モデルを作成する際の初期モデルとしては，前のステップでの i 番目の CV モデルを用いた．) そして次のループの認識ステップでは，i 番目の発話部分集合の認識にはパラメタ更新ステップで i 番目の部分集合を除いてパラメタ更新を行った CV モデルを用いる．これにより認識ステップとパラメタ更新ステップにおけるデータの重複を防ぐことができ，このような認識ステップとパラメタ更新ステップを従来のバッチ型適応と同様に何度か繰り返すことでより高い適応効果が期待できる．

さらに教師なし CV 適応法では認識ステップで得られた K 個の認識仮説を全て用いてパラメタ更新を行うことで，ループ毎に全認識仮説を統合したモデルを得ることもできる．これは図 1 の M(0) に該当し，CV 統合モデルと呼ぶことにする．(各ループでのパラメタ更新ステップでは，初期モデルとして前のループで得られた CV 統合モデルを用いた．) 適応後のモデルを一つにまとめたものを得る手段として CV 統合モデルを用いることができる．また教師なし CV 適応法では認識ステップでの計算量はバッチ型教師なし適応法と等しくなり，パラメタ更新ステップでの計算量はモデルの数 K に比例するため CV モデルを用いる場合の計算量は $O(K)$ となる．CV 統合モデルを用いる場合はさらにもう一つ分のモデルのパラメタ更新をする計算量が必要となる．

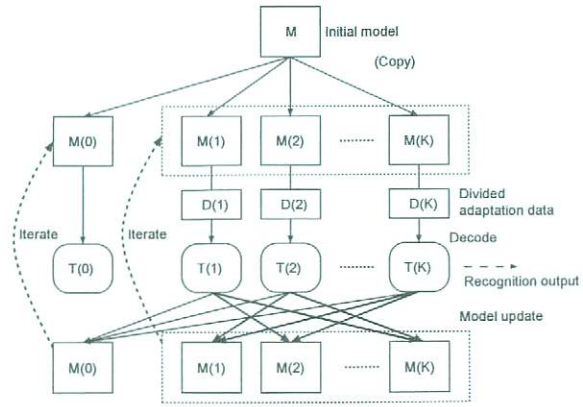

図 1 Unsupervised cross-validation(CV) adaptation

2.2 教師なし話者適応

教師なし話者適応は，書き起こしのない対象話者音声データを適応用データとして用いることでその認識率を向上させるための方法である．教師なし CV 適応法を用いた場合には認識仮説を出力する方法は二通りある．一つ目は適応繰り返し過程の各認識ステップにおいて CV モデルから得られる認識仮説 T(1)〜T(K) を集めて出力する方法である．二つ目は各ステップにおいて得られた CV 統合モデルを用いて認識対象音声データの認識を行い各ステップでの出力 T(0) とする方法である．そのため，本実験では前者の CV モデルと後者の CV 統合モデルの両方について認識結果の評価を行いその結果を比較する．本実験では話し言葉音声コーパスから学習した不特定話者音響モデルを初期モデル，同じく話し言葉音声コーパスのうち学習に用いたものと重複しない話者の音声データを認識対象音声データとして実験を行う．さらに認識対象音声データの発話時間による教師なし CV 適応法の効果の違いを比較するため，発話時間を短く区切って同様の実験を行った場合とも結果を比較し検討する．

2.3 教師なしタスク適応

教師なしタスク適応は対象とするタスクの適応用音声データに書き起こしが存在しない場合に，その音声データのみを用いて音響モデルを目的タスクに適応させることでそのタス

Vol.2009-SLP-77 No.7
2009/7/17

クの音声データの認識率を向上させるための手法である．タスク適応では評価用モデルとして CV 統合モデルを用いた．本実験では読み上げ音声コーパスから学習した不特定話者音響モデルを初期モデル，数時間程度の話し言葉音声をタスク適応用データ，話し言葉音声を目的タスクの評価音声として実験を行い，教師あり適応と比較してその効果を検討する．さらに教師なしタスク適応によって得られた不特定話者モデルに対して教師なし話者適応を適用し，タスク適応と話者適応を組み合わせた場合の効果についても合わせて検討する．

3. 実験条件

本実験では教師なし CV 適応法を MLLR[2] を用いた話者適応，および MAP[3] を用いたタスク適応に応用した．使用した音響モデルは状態共有混合ガウス分布トライフォンモデルであり，日本語話し言葉コーパス CSJ[4] の学会講演音声により EM 学習したものと新聞記事読み上げ音声コーパス JNAS[5] の音声データにより EM 学習したものを用いた．CSJ の学会講演音声を用いて学習した音響モデルの学習データ量は 254 時間であり，HMM の状態数は 3000，各状態の混合数は 32 である．JNAS の音声データを用いて学習した音響モデルの学習データ量は 52 時間であり，HMM の状態数は 2000，各状態の混合数は 32 である．音声認識特徴量は MFCC12 次元と対数エネルギー，およびそれらの Δ 項と $\Delta\Delta$ 項の計 39 次元である．不特定話者モデルの EM 学習，MLLR を用いた平均ベクトルの適応化，MAP を用いた平均および分散ベクトルの適応化には HTK ツールキット[6] を用いた．MLLR 適応は 32 の葉ノードを持つ回帰木を用いてツールキットのデフォルトのパラメタ設定で行った．MAP 適応についてもツールキットのデフォルトのパラメタ設定を用いて行った．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．タスク適応における適応データは CSJ の学会講演音声から 1 講演あたり 5 分を単位として必要な時間分だけランダムに選択したものを用い，適応の繰り返しを 5 回行った．テストセットは男性話者による学会講演音声 10 講演分からなる CSJ 評価セットである．テストセット中の各話者の講演時間はおよそ 10 分から 20 分程度であり，10 講演全体では約 2.3 時間のデータ量がある．話者適応はテストセットの各話者についてそれぞれ独立に行い，それらの単語誤り率を平均化した値を評価値として用いた．また発話時間と適応化性能の比較実験ではテストセットの各話者の音声を一定時間単位 (3 分，5 分) で排他的な区画に分割してそれぞれの区画毎に独立に話者適応を行い，それらの単語誤り率を平均化した値を評価値として用いた．音声認識システムは T^3 WFST 認識器[7] である．

4. 実験結果

図 2 から図 4 に話者適応での適応用音声データの発話時間として 3 分，5 分および講演全体を用いた場合の単語誤り率を示す．使用した初期モデルは CSJ の学会講演音声から学習した不特定話者モデルである．図中の CV は CV モデルでの，CV0 は CV 統合モデルでの評価結果を表している．教師なし CV 適応法では従来のバッチ型適応のベースラインと比較して全ての条件において，CV モデルと CV 統合モデルのいずれにおいても低い単語誤り率が得られていることが分かる．特に単位適応データ量の少ない場合ではその差が大きく，単位適応データ量が 3 分の時では適応を 8 回繰り返した後のバッチ型適応での相対的な単語誤り削減率が 5%であったのに対し，CV 適応法の CV モデルおよび CV 統合モデルでの相対的な単語誤り削減率はそれぞれ 13%，および 14%となった．さらに従来のバッチ型適応では 1 回目の適応でほぼ単語誤り率が収束してしまっているのに対し，CV 適応法ではより多い適応回数でも単語誤り率が減少し続けている．CV モデルと CV 統合モデルを比較すると，1 回目の適応では CV 統合モデルとバッチ型適応のモデルは同一となるため単語誤り率は従来法と等しくなり，CV モデルの単語誤り率はそれより低い値となった．しかし 2 回目以降の適応で CV 統合モデルは CV モデルとほぼ等しい単語誤り率となり，その後の CV モデルと CV 統合モデルの単語誤り率は共に収束していく．

また図 5 は各ループでの認識ステップとパラメタ更新ステップでの計算時間の平均値を従来のバッチ型適応，CV 適応法の CV モデルおよび CV 統合モデルについて示したものである．認識ステップでは CV 適応法の計算時間は従来のバッチ型適応の計算時間の 1.2 倍となった．理論的にはほぼ等しい値が得られるはずであり，値に差がある原因としては認識器読み込みによるオーバーヘッドが考えられる．一方パラメタ更新ステップでは CV 適応法の CV モデル，CV 統合モデルを用いた場合ではそれぞれバッチ型適応の 18 倍，19 倍とおよそデータ分割数に比例した計算時間となった．そして全体では CV 適応法はバッチ型適応の 3 倍の計算時間となった．(ただし計算時間の計測は CPU の状態による影響を受けるため厳密な値ではない．)

次に図 6 に JNAS の読み上げ音声データで EM 学習した不特定話者モデルを初期モデルとして用い，CSJ の学会講演音声タスクに対してタスク適応を行った際の適応データ量と単語誤り率の関係を示す．凡例の Init は不特定話者モデル，Batch は従来のバッチ型教師なし適応，CV0(20) は教師なし CV 適応法（CV 統合モデル），S-Batch は教師あり適応の場合をそれぞれ表している．教師あり適応での単語誤り率は教師なし適応で得ることが

Vol.2009-SLP-77 No.7
2009/7/17

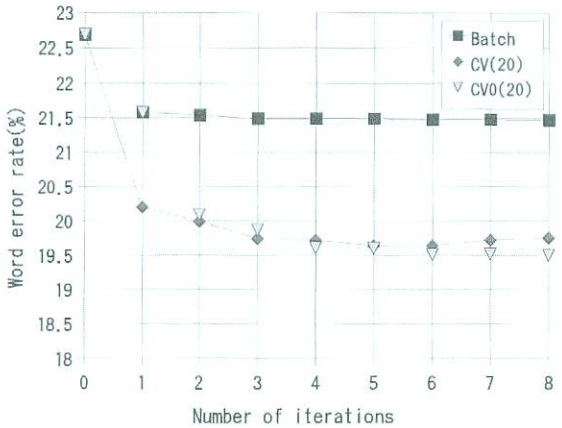

図 2 The word error rate of speaker adaptation. (using 3 minutes of adaptation data.)

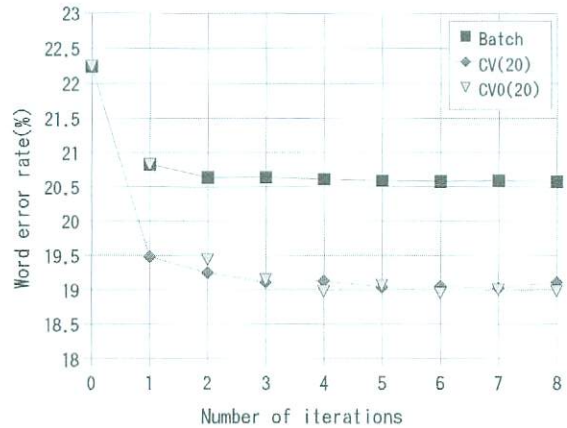

図 3 The word error rate of speaker adaptation. (using 5 minutes of adaptation data.)

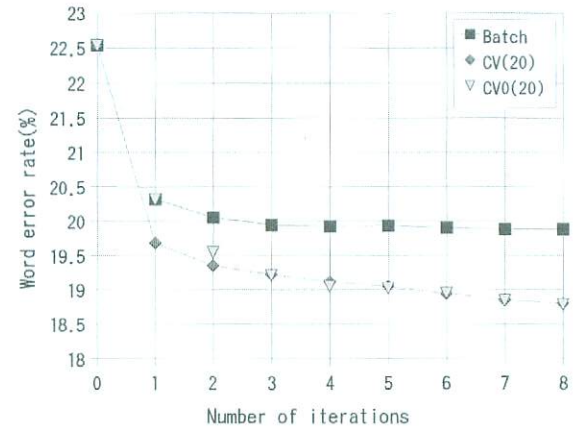

図 4 The word error rate of speaker adaptation. (using an entire lecture.)

できる値の限界値として捉えることができる．教師なし CV 適応では適応データ量を変えた全ての場合において従来のバッチ型適応と比較し低い単語誤り率となっている．特に適応セットが 2 時間分の場合に提案法と従来法とでの単語誤り率の差が最大になった．その場合不特定話者モデルでの単語誤り率 34.7%に対し，従来法での単語誤り率は 31.2%，提案法を用いた場合は 30.4%であり従来法を基準とした相対的な単語誤り削減率は 2.6%であった．また教師あり適応と比較した場合，適応データが 2 時間の場合に教師あり適応での単語誤り率と従来のバッチ型適応での単語誤り率の中間程度の単語誤り率を得ることができたが，それ以外の適応データ量では教師あり適応との差が大きくなっている．

図 7 はタスク適応時に 10 時間分の適応データを用いてバッチ型教師なし適応法，教師なし CV 適応法でパラメタ更新をした音響モデルをそれぞれ初期モデルとして話者適応を行った場合の単語誤り率の推移である．凡例の Batch-CV0 はタスク適応でバッチ型適応，話者適応では CV 適応法をそれぞれ用いたことを意味する．タスク適応にバッチ型適応を用いた場合と CV 適応法を用いた場合を比較すると，話者適応で用いた手法がバッチ型適応か CV 適応法かに関わらず話者適応を 8 回繰り返した後でもタスク適応時の CV 適応法の効果が残り，より低い単語誤り率が得られたことが分かる．また単語誤り率の最低値はタスク適応と話者適応の両方で CV 適応法を用いた場合に得られており，話者適応を 8 回繰り返した後での単語誤り率は 23.6%であった．一方タスク適応と話者適応の両方で従来のバッチ型適応を用いた場合の話者適応を 8 回繰り返した後での単語誤り率は 25.5%であり，タスク適応と話者適応の両方で CV 適応法を用いた場合の、両方で従来法を用いた場合を基準とした相対的な単語誤り削減率は 7.6%となった．

5. ま と め

音響モデルの効率的な教師なし適応を目的として提案し話者適応においてその効果を示した教師なしクロスバリデーション適応法[1] について，さらに諸条件において評価を行なった．その結果話者適応において 3 分，5 分程度と比較的少量の適応音声データに対しても，提案手法は従来のバッチ型適応よりも低い単語誤り率を得ることが分かった．特に適応データの量が 3 分ではその差が大きく，適応を 8 回繰り返した後のバッチ型適応での相対的な単語誤り削減率が 5%であったのに対し，CV 適応法の CV モデルおよび CV 統合モデルでの相対的な単語誤り削減率はそれぞれ 13%，および 14%となった．また提案手法でのデータ分割数を 20 とした場合，提案手法に要する計算時間は従来手法の 3 倍程度であることが分かった．さらに提案手法をタスク適応にも適用したところ，適応データ量に関わらず常に提案手法の単語誤り率が従来法を下回った．特に適応データが 2 時間の場合従来のバッチ型

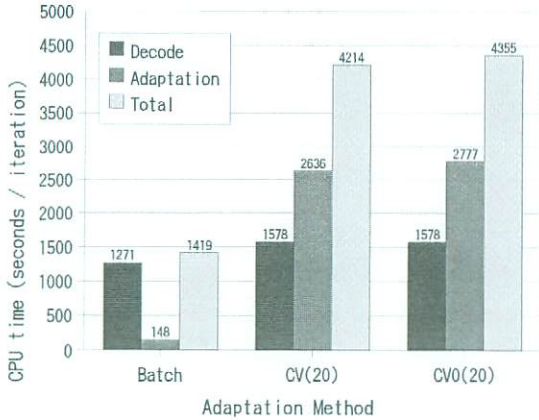

図 5 Calculational cost

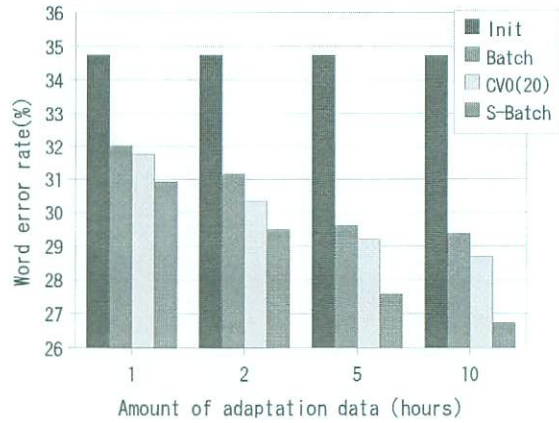

図 6 Unsupervised and supervised task adaptation

教師なし適応と教師あり適応の中間程度の単語誤り率となり，提案手法の有用性を示した．またタスク適応後の音響モデルに対し話者適応を行ったところ話者適応後も提案法によるタスク適応時の効果が残る事が分かった．タスク適応と話者適応のいずれも従来法を用いた場合を基準とすると，いずれも CV 適応法を用いた場合の相対的な単語誤り削減率は 7.6%となった．

参 考 文 献

1) 篠崎 隆宏, 久保田 雄, 古井 貞熙,「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および集合適応法の提案」, 情報処理学会 研究報告, 2009-SLP-75, pp.1-6 (2009-2).
2) C. Leggetter and P. Woodland, "Speaker adaptation of continuous density HMM's using linear regression," in *Int. Conf. Speech Language Processing '94*, vol. 2, Yokohama, Japan, 1994, pp. 451-454.
3) J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing.*, vol. 2, pp. 291-298, Apr. 1994.
4) CSJ(日本語話し言葉コーパス) http://www.kokken.go.jp/katsudo/seika/corpus/.
5) JNAS(新聞記事読み上げ音声コーパス), http://www.milab.is.tsukuba.ac.jp/jnas/.
6) S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, The HTK book (for HTK Version 3.2), 2002.
7) 大西 翼, ディクソン ポール, 古井 貞熙, "WFST 音声認識デコーダの開発とその性能評価," 情報処理学会研究報告, vol.2007, no.103, pp.1-6, 2007.

Vol.2009-SLP-77 No.7
2009/7/17

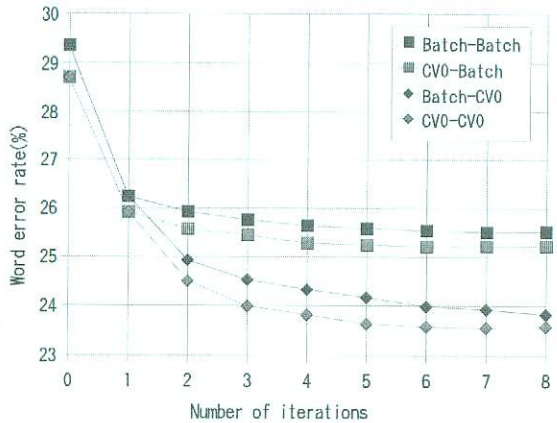

図 7 Speaker adaptation after task adaptation. (The task adaptation used 10 hours of speaker independent data.)