

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

篠崎 隆宏 久保田 雄 古井 貞照

東京工業大学 大学院情報理工学研究科 計算工学専攻

あらまし 教師無し適応における適応化性能を向上させる目的で、クロスバリデーションおよびバギングに似た手法を教師無しバッチ適応の枠組に組み込んだ教師無しクロスバリデーション適応法および教師無し集合適応法の提案を行なう。提案アルゴリズムは MLLR 音響モデル適応化手法のような一般的な適応化技術の適用方法に関するものであり、それら元とする適応手法の詳細には基本的に依存しない一般性がある。また提案する適応化アルゴリズムは、これまでに提案を行なった教師あり学習法である CV-EM や Ag-EM の拡張と見ることができる。CV-EM や Ag-EM では過学習問題のみを対象としていたが、今回提案する適応化手法はそれに加えて、教師無しバッチ適応において用いられる認識仮説中の認識誤りの影響を低減させ、適応化性能を向上させる効果がある。実験では、提案する教師無しクロスバリデーション適応法および教師無し集合適応法を日本語話し言葉コーパスの学会講演音声認識に適用する。従来のバッチ型教師無し適応による相対的な単語誤り率の削減率が 12% であるのに対し、教師無しクロスバリデーション適応法を用いた場合の削減率は 17% であった。

Unsupervised Cross-validation and Aggregated Adaptations for Improved Speech Recognition

Takahiro SHINOZAKI Yu KUBOTA Sadaaki FURUI

Department of Computer Science, Tokyo Institute of Technology

Abstract Unsupervised cross-validation and aggregated adaptation algorithms are proposed that introduce cross-validation and bagging-like ideas in the unsupervised adaptation framework to reduce the over-training problems and to improve the recognition performance for speech recognition. The algorithms are constructed on a general adaptation technique such as the MLLR acoustic model adaptation method and basically independent from the details of the underlying adaptation method. These algorithms are extensions of our previously proposed CV and Ag-EM methods. The proposed algorithms are also useful to suppress the negative effects of unsupervised adaptation which reinforce the errors included in the hypotheses used for the adaptation. In the experiment, unsupervised cross-validation and aggregated MLLR adaptation have been applied to presentation speech recognition. The relative word error rate reduction by the cross-validation adaptation was 17% whereas the reduction by the batch-mode baseline adaptation was 12%.

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る。特に、話し言葉音声は話者の違いや話者のおかれた状況など様々な要因により大きく影響を受け、また周囲の環境をシステム側で制御することが困難な場合が多い。このため、話し言葉音声認識において高い認識精度を得るためには、一般的な不特定話者モデルを対象音声に対して教師無しで効果的に適応させることが極めて重要となる。適応手法には様々なものがあるが、音響モデルなどの適応化の枠組として広く用いられているものに教師無しバッチ適応がある。音響モデルにおけ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て教師あり適応を行なうのが一般的である。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、より高い認識性能が得られる [1]。

しかしながら、教師無し話者適応において常に問題となるのが、適応用のデータは限られており、また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。一般に適応化技術はモデル適応における自由変数の数を少なく抑えることで適応化の汎化性能を高めるように設計されている。しかしながら、自由変数の数を減らせばモデルの自由度が減少するため、汎化性能と精密な適応化を行なうことの間にはトレードオフがある。このため自由変数の数の制御は有効ではあるが問題は完全には解決されず、さらなる改良の余地があると考えられる。

本研究ではバッチ型教師無し適応の汎化性能を向上させることで効果的な適応を可能とする新しい教師無し適応の枠組を提案する。提案手法では、クロスバリデーションおよびバギング [2] に似た手法をバッチ型教師無し適応の繰り返しループの中に組み込む。クロスバリデーションを繰り返し型推定法の内部に組み込むという点で、提案法は MMI 学習における勾配の推定にクロスバリデーションを使用する方法 [3] や、我々がこれまでに提案した教師あり学習手法である CV-EM や Ag-EM [4] と類似点がある。

提案する適応化手法はモデルのパラメタを適応化する既存の適応化技術の上に組み立てられるので、基本的にそれら元にするアルゴリズムの詳細には依存しない。したがって、提案法は潜在的には音声認識に限らず繰り返し手法による教師無し適応一般に応用が可能である。本研究ではしかしながら、提案手法を MLLR [5] 手法と組み合わせて音響モデルの適応に用いる場合にしばって実験を行なう。実験での提案手法の性能評価は、日本語話し言葉コーパス

CSJ [6] の講演音声データを用いた大語彙音声認識をタスクとして行う。

本論文の構成は以下の通りである。第 2 章ではまず従来のバッチ型教師無し適応の枠組の説明を行ない、ついで提案する教師無しクロスバリデーション適応法および集合適応法を説明する。第 3 章で実験条件について説明し、第 4 章で実験結果を示す。最後に第 5 章でまとめと今後の課題を示す。

2 教師無しクロスバリデーション (CV) 適応法および集合適応法

本章ではまず従来のバッチ型教師無し適応の枠組について簡単に説明を行なう。ついで、提案する教師無しクロスバリデーション (Cross-validation: CV) 適応法および集合 (Aggregated: Ag) 適応法について説明する。提案適応アルゴリズムは一般的なものであるが、簡単のため音声認識システムにおける話者適応を仮定した説明を行なう。

2.1 バッチ型教師無し適応

図 1 に典型的な教師無し繰り返しバッチ型話者適応のプロセスを示す。まず最初のステップは適応対象の話者から与えられた発話音声全体に対し音声認識を行なうことである。この際用いる音響モデルは、適応ループの一番はじめでは不特定話者モデル、それ以降ではそのループのひとつ前で話者適応されたモデルである。これにより、適応対象音声の認識仮説を得ることができる。次のステップではその認識仮説を適応用の書き起こしとすることで、教師あり適応法によりモデルパラメタの更新を行なう。モデルパラメタを実際にどのように更新するかは MLLR など用いる適応化手法に依存する。この認識処理とモデル更新処理を一ループとし、より高い適応化の効果を得るためにこのプロセスは何度か繰り返される。最終的な適応モデルを用いた認識結果は、ループの最後の認識処理に得られた認識仮説を出力することにより得られる。

このバッチ型繰り返し教師無し適応の問題点として、過学習が顕著であるという点が挙げられる。これは一度モデルパラメタがどれか特定の適応データに特化しだすと、認識処理とモデル更新処理を同じデータを用いて繰り返す為、適応の繰り返しとともにその偏りが強化されてしまうためである。さらに、認識器を用いた音声認識では認識誤りが避けられないが、同様な理由により認識誤りも適応ループの中で強化されてしまうことも大きな問題である。これ

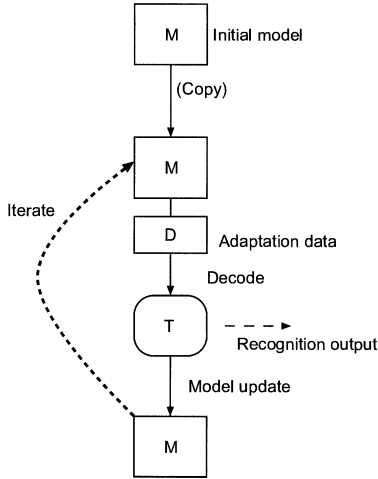


図 1: Batch-mode unsupervised adaptation. (M: model, T: transcript, and D: target data)

らの問題は最終的に適応化性能を低下させ、認識性能を制限する要因となっている。

2.2 教師無し CV 適応

提案する教師無し CV 適応法では、従来のバッチ型適応における問題を効果的に抑制するために、繰り返しループ中に K -fold CV 手法を導入することで認識ステップとモデル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。図 2 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、最終的な話者適応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。

提案アルゴリズムは MMI 学習における勾配の推定に CV を用いる手法 [3] および我々がこれまでに提案した CV-EM [4] 手法と、CV を繰り返し推定法の内部に導入するという点が類似している。違いとしては、提案手法は教師無し学習手法であり、CV 手法を MMI 学習における勾配や CV-EM 学習における人手による書き起こしに対する十分統計量を推定するためではなく、デコーダによる認識仮説を得るために用いる点が挙げられる。

提案法においてモデルパラメタの更新自体は任意の教師あり適応化手法を用いることが可能であり、MLLR など特定の適応法に限られない一般性がある。認識ステップにおける計算コストは、複数のモデルを読み込むオーバーヘッドを除けば K に対して一定である。これは K が大きくなれば使用するモデルの数は増えるが、各モデルが処理すべきデータ量はその分減るためである。モデルパラメタ更新にかかる計算量は、 K 個のモデルを作成するため一般的には単一のモデルを推定する場合と比較して K 倍となる。ただし、用いる適応化手法によっては十分統計量を利用するなどにより、これより減らすことも可能である。

教師無し CV 適応法はもし $K = 2$ とすると、クロスアダプテーション [7] とも類似していると言える。違いは、クロスアダプテーションでは特徴量を変えるなどして同じ音声データを異なるシステムを用いて認識した認識仮説を用いるのに対し、教師無し CV 適応では同じ特徴量を用いた異なるデータを用いたモデルを使用することである。すなわち、教師無し CV 適応は単一の認識システム上で動作するが、クロスアダプテーションでは 2 つの認識システムが必要となる。

2.3 教師無し集合適応

教師無し CV 適応法と異なり、教師無し集合適応法では認識ステップとモデル更新ステップでデータの重複を図 3 に示すように許す。その代わりに、教師無し集合適応法において汎化能力はバギングと同様に N 個のモデルを集合的に用いることで向上さ

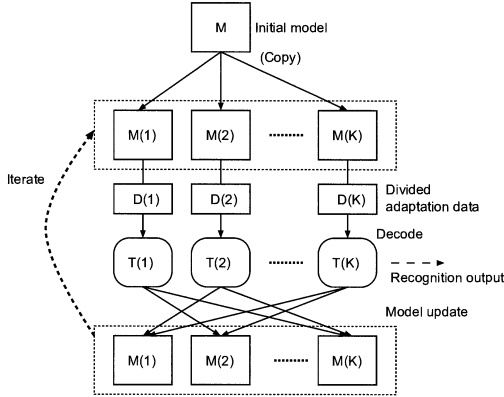


図 2: Cross-validation (CV) adaptation.

せている。より具体的には、適応の発話データ集合をまず K 個の排他的な部分集合に区分化する。そして、各データ部分集合に対し N 個のモデルを用いて繰り返し認識処理を行なう。適応化の一番はじめのステップでは、これら N 個のモデルは単に初期モデルを個数分コピーすることにより用意する。次のモデルパラメタ更新ステップでは、ランダムに選択した K' 個のデータサブセットから得られた NK' 個の認識仮説からモデルを推定することを N 回繰り返すことで、 N 個のモデルを作成する。各モデルの推定において同じデータから得られた N 個の認識仮説を用いることになるため、モデル更新に用いる具体的な適応化アルゴリズムによっては、必要に応じてその適応アルゴリズム内で用いるイベント観測回数などを N で正規化する。作成した N 個のモデルは次の認識ステップにおいて各サブセット毎に N 個の認識仮説を生成するために用いる。

教師無し集合適応はこれまでに提案した教師あり学習法である Ag-EM [4] の拡張であり、教師無し CV 適応法と CV-EM の比較と同様の類似点と相異点がある。それ以外の違いとしては、教師無し集合適応では最終的に単一の、適応を行なった認識仮説が出力される必要があるという点が挙げられる。このためには、最後の認識ステップで得られた N 個の異なる認識仮説を ROVER [8] や CNC [9] 手法を用いて統合することが考えられる。あるいは、 N 個の集合モデルとともに、全ての適応データサブセットから得られた全ての認識仮説を用いたモデルを作成し、そのモデルを用いて認識を行なった結果を出力することも考えられる。本研究では後者の方法を採用し、 N 個の集合モデルとともに作成する、全ての認識仮説を用いて作成したモデルをグローバル集合モデルと呼ぶことにする。十分統計量を利用する

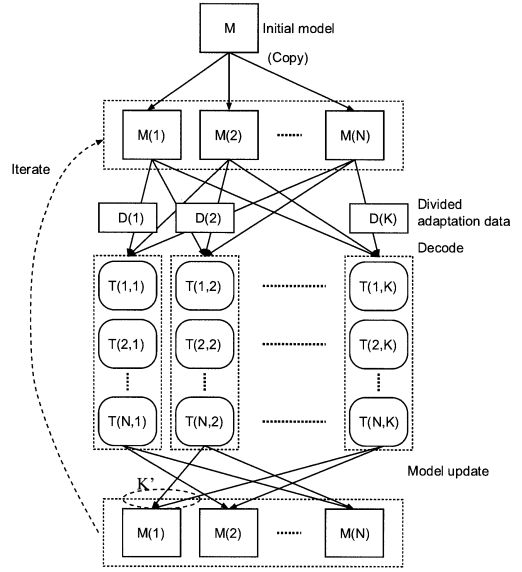


図 3: Aggregated (Ag) adaptation. ($T(n,k)$: recognition hypothesis of the k -th subset made by using n -th Ag model)

などの特別な工夫を行わない場合には、モデル更新ステップの計算量は単一のモデルを作成する場合と比較して $N^2 K' / K$ 倍となる。

3 実験条件

提案手法を MLLR を用いた音声認識における音響モデルの話者適応に応用した。音声認識システムは T^3 WFST 認識器である [10]。使用した音響モデルは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 [6] の学会講演音声より学習した。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。音声認識特徴量は MFCC12 次元と対数エネルギー、およびそれらのデルタ項とデルタデルタ項の計 39 次元である。不特定話者モデルの EM 学習及び MLLR を用いた平均ベクトルの適応化には HTK ツールキット [11] を用いた。MLLR 適応はツールキットのデフォルトのパラメタ設定で回帰木を用いて行なった。例外として、教師無し集合適応では同じデータに対し N 通りの認識仮説を生成し使用することになるため、適応クラス数を決める観測回数の閾値を N 倍とすることで、正規化を行なった。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。

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。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 分程であり、10 講演全体では約 2.3 時間のデータ量がある。適応化は各話者ごとに行ない、それらの単語誤り率を平均した値を各手法の評価に用いた。

本実験においては、MLLR 適応化は既存のツールをそのまま用い、十分統計量を用いた高速化などは行なわなかった。参考として、この条件において教師無し 10-fold CV 適応において認識ステップに要した計算時間と MLLR によるモデル更新ステップに要した計算時間は同程度であった。

4 実験結果

図 4 に教師無し CV 適応におけるクロスバリデーションのサブセット分割数 K と単語誤り率の関係を示す。教師無し CV 適応では全ての条件において従来のバッチ型適応のベースラインと比較して低い単語誤り率が得られていることが分かる。特に分割数 K がおよそ 10 よりも大きいと安定してほぼ最良の結果が得られることが分かる。 K の値が 2 など小さいときに適応性能の向上が小さい理由は、モデルパラメタの推定に実際に使用されるデータ量が減少してしまうためである。反対に K の値がある程度以上大きくなれば、各モデルの推定に実際に使われるデータ量は $(K - 1) / K$ なので実質的にほぼ全データがパラメタ推定に使用されるようになり、安定した性能が得られるようになる。

図 5 に教師無し CV 適応と集合適応における、適応繰り返し数と単語誤り率の関係を示す。教師無し CV 適応は $K = 40$ 、教師無し集合適応は $K = 10, K' = 6, N = 8$ の条件で行なった。教師無し集合適応の繰り返し第一回目の単語誤り率は従来のバッチ型適応と同じである。これは、教師無し集合適応の認識仮説をグローバル集合モデルから得ている為である。繰り返し 2 回目以降では教師無し集合適応は一貫して従来のバッチ型適応よりも低い単語誤り率を与えている。教師無し CV 適応に関しては、繰り返し一回目から従来法よりも低い単語誤り率となっている。教師無し CV 適応および教師無し集合適応のどちらも従来法と比べて高い適応性能を示しているが、本実験条件においては教師無し CV 適応の方が教師無し集合適応よりも適応性能および計算コスト両面において優れた結果となった。不特定話者モデルによる単語誤り率 22.5% に対し、適応を 8 回繰り返した後の従来のバッチ型適応、教師無し CV 適応、およ

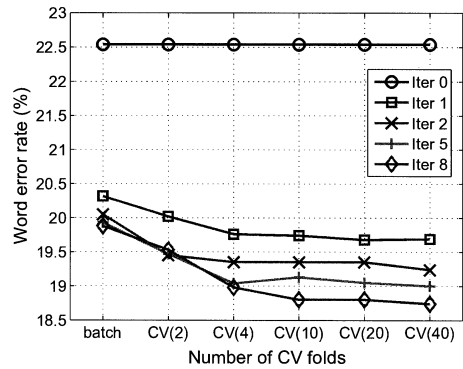


図 4: Number of cross-validation folds (K) of CV adaptation and recognition performance. The zero-th iteration is the result by the speaker independent model. (batch: batch-mode baseline adaptation result)

び教師無し集合適応による相対的な単語誤り削減率はそれぞれ、12%、17%、および 16% であった。従来のバッチ型の話者適応結果の単語誤り率を基準とすると、教師無し CV 適応および教師無し集合適応の相対的な誤り削減率は 6% および 4% であり、どちらも統計的に有意であった。

5 まとめと課題

教師無し適応の適応性能を従来のバッチ型適応法よりも向上させる教師無し CV 適応法および教師無し集合適応法の提案を行なった。適応化の汎化性能を高め、また適応時に用いる認識仮説に含まれる認識誤りに対する頑健性を高めるため、教師無し CV 適応法では CV 手法、教師無し集合適応法ではバギングに似た手法を教師無し適応の繰り返しループの中に組み込む。

実験により、教師無し CV 適応法および教師無し集合適応法のどちらも従来のバッチ型適応と比べてより低い単語誤り率を与えることを示した。教師無し CV 適応法および教師無し集合適応法を比べると、教師無し CV 適応法の方が適応性能および計算コストの両面で優れた結果となった。従来のバッチ型適応による相対的な誤り削減率が 12% であったのに対し、教師無し CV 適応法を用いた場合の単語誤り削減率は 17% であった。

今後の課題としては、教師無し集合適応法において仮説の統合に ROVER などの手法を用いること、

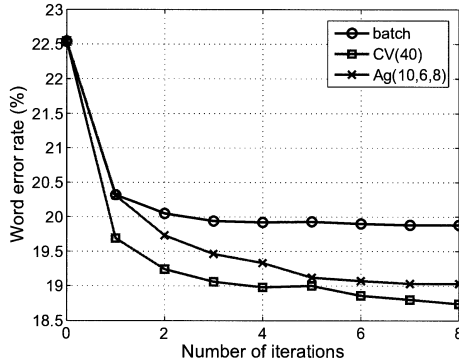


図 5: Number of adaptation iterations and recognition performance of CV and Ag adaptations. The zero-th iteration is the result by the speaker independent model. (batch: batch-mode baseline adaptation, CV(40): CV adaptation ($K = 40$), and Ag(10,6,8): Ag adaptation ($K = 10, K' = 6, N = 8$))

十分統計量を活用することで MLLR と教師無し CV 適応法および教師無し集合適応法を組み合わせた場合の計算量を削減することなどが挙げられる。また、MLLR 以外の適応手法との組み合わせや、音響モデル以外の教師無し適応への応用なども課題としてあげられる。

6 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] M. Gales and S. Young. The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2008.
- [2] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [3] N. S. Kim and C. K. Un. Deleted strategy for MMI-based HMM training. *IEEE Transactions on Speech and Audio Processing*, 6(3):299–303, 1998.
- [4] T. Shinozaki and M. Ostendorf. Cross-validation and aggregated EM training for robust parameter estimation. *Computer speech and language*, 22(2):185–195, 2008.
- [5] C. J. Leggetter and P. C. Woodland. Flexible speaker adaptation using maximum likelihood linear regression. In *Proc. Eurospeech*, pages 1155–1158, 1995.
- [6] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui. Benchmark test for speech recognition using the Corpus of Spontaneous Japanese. In *Proc. SSPR2003*, pages 135–138, 2003.
- [7] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig. The IBM 2004 conversational telephony system for rich transcription. In *Proc. ICASSP*, volume I, pages 205–208, 2005.
- [8] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Proc. IEEE ASRU*, pages 347–352, 1997.
- [9] G. Evermann and P. C. Woodland. Posterior probability decoding, confidence estimation and system combination. In *Proc. Speech Transcription Workshop*, 2000.
- [10] P. R. Dixon, D. A. Caseiro, T. Oonishi, and S. Furui. The titech large vocabulary wfst speech recognition system. In *Proc. IEEE ASRU*, pages 443–448, 2007.
- [11] S. Young *et al.* *The HTK Book*. Cambridge University Engineering Department, 2005.