論文 / 著書情報
Article / Book Information

論題(和文)	音声認識のためのコミッティを用いた能動学習
Title(English)	
著者(和文)	濱中 悠三, 江森 正, 越仲 孝文, 篠田 浩一, 古井 貞熙
Authors(English)	yuzo hamanaka, Tadashi Emori, Takafumi Koshinaka, Koichi Shinoda, SADAOKI FURUI
出典(和文)	日本音響学会2009年秋季講演論文集, Vol. , No. 1-1-5, pp. 15-18
Citation(English)	, Vol. , No. 1-1-5, pp. 15-18
発行日 / Pub. date	2009, 9

音声認識のためのコミッティを用いた能動学習*

◎濱中悠三 (東工大), 江森正 (NEC 情報システムズ), 越仲孝文 (NEC / 東工大),
篠田浩一 (東工大), 古井貞熙 (東工大)

1 はじめに

現在の統計的パターン認識器の教師あり学習には大量のデータとその正解ラベルが必要である。音声認識を含むパターン認識ではラベルなしのデータを収集することは比較的容易だが、正解のラベル付け作業は人手で行わなければならず多くの時間と費用のコストが掛かる。大量のラベルなしデータからの認識性能向上に役立つデータの選択を認識器自身が行う学習方法を能動学習と呼び、これによってラベル付けコストを減らすことができる。

音声認識のための能動学習の研究 [1], [2], [3] では何らかの選択基準に基づいてラベルなしデータを選択してラベル付けを行う。Hakkani ら [1] は認識器から出力された単語事後確率の平均を発話の信頼度とし、これが小さい発話から選択をしている。Varadarajan ら [3] は認識器から生成されたラティスのエントロピーを利用して発話の選択をしている。

本稿では新しい音声認識のための能動学習手法を提案する。この手法では少量のラベル付きデータから複数の異なる認識器を作成する。その後、ラベルなしデータを認識し、複数の認識器による認識結果文が最も一致しないデータを選択する。

この複数の異なる認識器 (コミッティ) に基づく手法は機械学習 [4] の課題において提案され、その有効性は Dagan ら [6] が HMM を用いた品詞タグ付けの課題で実証した。我々はこの手法を音声認識に適用する。この手法では

- 複数の認識器を学習データからどのように作成するか
- 認識結果文の不一致度をどのように定義するか

を決める必要がある。これらについては後の章で述べる。

今まで能動学習の研究は、学習データの発話数が 2 万から 3 万のデータを用いて行われている (例えば [2],[3])。ここでは話し言葉音声認識に対する実用的な能動学習の手法の効果を検証するため 22.4 万発話の学習データを使用する。この大規模データを使った実験において我々は音響モデルだけでなく言語モデルに対しても能動学習を行う。

2 アルゴリズム概要

提案する能動学習アルゴリズムの概略図を Fig.1 に示す。書き起こし (ラベル) 付き学習データを T、書き起こしされていない学習データを U、認識器の個数を K、一度に選択される発話の時間量を N とする。能動学習は以下の 5 ステップで実行される。

1. 学習データ T をランダムに等分割し、データセット T_k $(k = 1, ..., K)$ を作成する
2. T_k を用いて認識器 M_k を学習する $(k = 1, ..., K)$
3. データ U の全ての発話を認識器 M_k $(k = 1, ..., K)$ を用いて認識し、K 個の異なる認識結果文を出す
4. U の発話の中から、認識結果文の不一致度が高い発話を N 時間分選択する
5. 選択した発話を U から取り除き、書き起こして、T に追加し、1. に戻る

以上を書き起こしコストが尽きるまで繰り返す。最終的に全ての書き起こされたデータを用いて認識器を作成し音声認識に使用する。ステップ 4 の発話選択の詳細については 4 章で述べる。

3 Query by Committee

コミッティに基づく手法では Query by Committee(QBC)[4] の理論を用いる。QBC は

*Committee-based active learning for speech recognition. by Yuzo Hamanaka (Tokyo Institute of Technology), Tadashi Emori (NEC Informatec Systems), Takafumi Koshinaka (NEC / Tokyo Institute of Technology), Koichi Shinoda (Tokyo Institute of Technology), and Sadaoki Furui (Tokyo Institute of Technology)

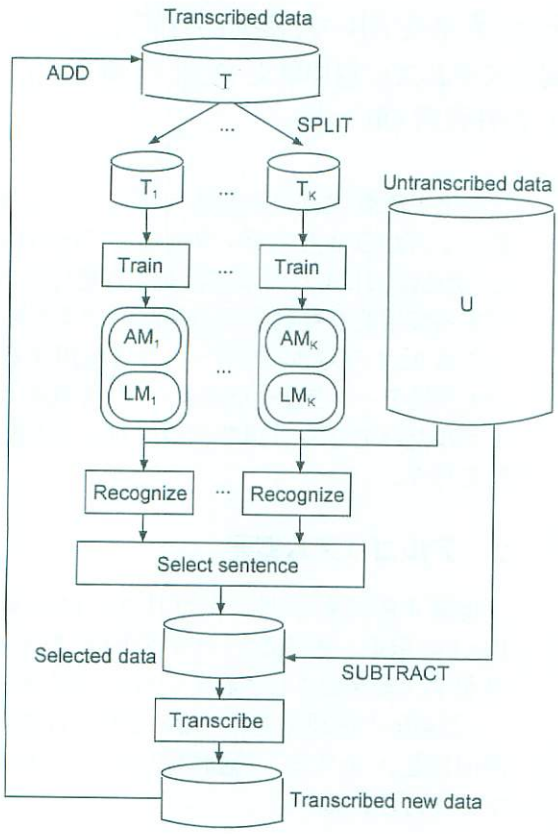

Fig. 1 Active learning scheme using query-by-committee based approach for speech recognition.

線型分離可能な場合におけるパーセプトロンなどの二分類問題において、汎化誤差を減らすためのデータ選択方法に関する理論である。

QBC ではバージョン空間 (学習データに無矛盾な分類器の集合) を効率的に狭めていくアルゴリズムが提案された。このアルゴリズムではラベルなしデータが提示されたとき、バージョン空間からランダムに選ばれた複数の分類器による分類結果によってデータをラベル付けするかどうかを決定する。Seung ら [4] は分類結果がばらつくデータを選択していくことでラベル付けコストを指数的に減らせることを理論的に証明している。

4 発話選択

音声認識では確率的手法が用いられるため、QBC の理論的枠組を音声認識にそのまま適用することは困難である。確率的手法は常に正しいクラスに最も高い確率を割り当てるとは限らない。さらに大語彙連続単語音声認識におけるバージョン空間は二分類問題より遥かに複雑であり、ほとんどの場合バージョン空間の考え方は適用できない。

そこで、ここでは簡単のため認識結果文の不一致度が高い発話を選択することにする。不一致度を定義するために、まず K 個の認識結果文をマルチプルアライメントし、Voting Entropy を計算する。全ての発話に対して不一致度を計算した後、不一致度が高い発話から書き起こしのための選択を行う。

4.1 認識結果文のアライメント

品詞タグ付けの場合と違い、音声認識では認識結果文に含まれる単語数が一定ではないため、認識結果文のアライメントを行う必要がある。ここではアライメントにはヒューリスティックな方法の 1 つであるプログレッシブ法 [7] を用いる。プログレッシブ法はマルチプルアライメントに最も一般的に使われる方法である。ここでアライメント前の 1 つの認識結果文を配列と呼ぶ。また 2 本以上の配列のアライメント結果をクラスタと呼ぶ。プログレッシブ法では配列と配列、配列とクラスタ、またはクラスタとクラスタのアライメントを繰り返すことによりマルチプルアライメントを実現する。

プログレッシブ法のアルゴリズムではまず、案内木と呼ばれる木構造を計算する。次に、案内木の葉ノードから根ノードに向けて順番にアライメントを行う。

案内木の作成には UPGMA 法 [8] において距離の定義を簡単化した方法を用いる。UPGMA 法では全ての入力配列 (N 個) のペアに対して配列間類似度を計算し、$N(N-1)/2$ 個の類似度行列を作成する。配列間類似度の定義に関しては、様々な方法が提案されているが、ここではペアワイズアライメントを計算した際のスコアとする。類似度が近いペアから合併してクラスタを作成し、類似度行列を更新する。クラスタ間の類似度は、UPGMA 法では 2 つのクラスタを構成する全配列ペアの配列間類似度の平均値であるが、ここでは簡単化し、更新前の類似度行列における、それぞれのクラスタとの類似度の平均値とする。

案内木に沿ったアライメントでは、まず配列と配列がアライメントされる。その後の配列とクラスタまたはクラスタとクラスタのアライメントでは、クラスタを構成する配列間のアライメ

Table 1 An example sentence-alignment result. The symbol "-" indicates a gap.

		i				
		1	2	3	4	5
	1	-	英語	文字	以下	と
	2	-	英語	用い	か	と
h	3	-	英語	文字	か	と
	4	えー	この	字	以下	と
	5	えー	も	ジー	か	と
	6	えー	この	ジー	か	と
	7	えー	この	ジー	か	と
	8	えー	この	ジー	か	と

ント後の位置関係を固定し、ギャップ "-" を挿入するときはクラスタを構成する全ての配列の同じ列に挿入する。配列とクラスタまたはクラスタとクラスタのアライメントはクラスタの要素間の位置関係を固定することで、配列と配列のペアワイズアライメントと同様に DP マッチングを行い、最適なアライメントを探索する。

以下の SP スコア $S(m_i)$ の各列 i に渡る合計が最大になる様にギャップが挿入された複数の認識結果文をマルチプルアライメント結果とする。

$$\sum_{i=1}^{I} S(m_i) = \sum_{i=1}^{I} \sum_{h=1}^{K} \sum_{l=h+1}^{K} s(m_i^h, m_i^l) \qquad (1)$$

マルチプルアライメント結果の例を Table 1 に示す。ここで式 (1) の i はアライメント結果の先頭からの列番号である。I は i の要素数を表し、Table 1 の例では 5 である。$S(m_i)$ は列 i における SP スコアである。m_i^h はアライメント結果を構成する h 番目の認識結果文の列 i にある単語である。コスト $s(a,b)$ の値は以下を用いた。

$$s(a,b) = \begin{cases} 2 & (a = b \text{ and } (\mathrm{a} \neq - \text{ and } \mathrm{b} \neq -)), \\ -1 & (a = b \text{ and } (\mathrm{a} = - \text{ and } \mathrm{b} = -)), \\ -1 & (a \neq b \text{ and } (\mathrm{a} = - \text{ or } \mathrm{b} = -)), \\ -1 & (a \neq b). \end{cases}$$

この値はいくつかの値で予備評価した中で最も良い結果の値である。

4.2 認識結果文の不一致度

不一致度を計算するために、マルチプルアライメント結果の列ごとに Voting Entropy を計算する。列 i に存在する単語の種類数を P、それぞれの単語を w_p $(p = 1, ..., P)$ とし、列 i に単語 w_p が何個存在するかを $V(w_p, i)$ で表す。列 i における Voting Entropy $\mathrm{VE}(i)$ を以下のように定義する。

$$\mathrm{VE}(i) = -\sum_{p=1}^{P} \frac{V(w_p,i)}{K} \log \frac{V(w_p,i)}{K} \qquad (2)$$

アライメントで生じたギャップは 1 つの単語として扱う。全ての列 i に渡る $\mathrm{VE}(i)$ の平均を発話の認識結果文の不一致度 D

$$D = \frac{\sum_{i=1}^{I} \mathrm{VE}(i)}{I} \qquad (3)$$

と定義する。発話選択では D が大きい発話から順に選択を行う。

5 実験

5.1 実験条件

データベースとして、日本語話し言葉コーパス (CSJ) [9] を使用した。その中の男性話者による学会講演音声を実験用データに用いた。全実験データの内、224,434 発話 (666 話者、190.8 時間) を学習データとし、2328 発話 (10 話者、1.95 時間) をテストセットとした。

音声認識に使う特徴量は MFCC12 次元とパワー、及びその一次微分成分と二次微分成分の計 39 次元を用いた。分析周期は 10ms、分析窓幅は 25ms とし、発話単位ごとに CMS 処理を行った。音響モデルは 16 混合 3000 状態 triphone HMM を用いた。認識は 2 パスサーチを行い、言語モデルは 1 パス目に 2 gram、2 パス目に 4 gram を用いた。実験には HTK [10] を用いた。

全学習データからランダムに選択した 30,000 発話 (25.3 時間) を書き起こし付き学習データ T として、初期の認識器作成に使用し、残りの学習データを書き起こしなし学習データ U とした。発話選択 (ステップ 4) において一度に選択される音声時間量 N は 25 時間、コミッティを構成する認識器の数 K は 8 とした。

提案手法を 2 つの手法と比較した。1 つはランダム選択であり、発話選択をランダムに行った。もう 1 つは単語事後確率 (WPP) に基づく選択 [1] であり、発話文中の単語事後確率の平均が低い発話から順に選択した。ラティスからのコンフュージョンネットワーク作成には SRILM [11] を用いた。

5.2 実験結果

Fig. 2 に 3 つの能動学習手法による学習効率を示す。提案したコミッティに基づく手法はラン

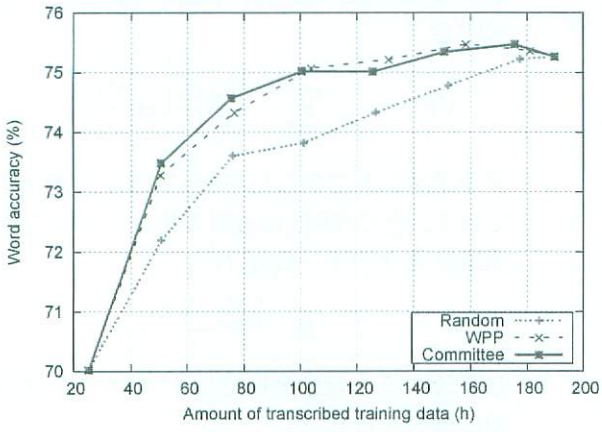

Fig. 2 Recognition results for proposed committee-based method of active learning (Committee) compared with random selection (Random) and selection using the WPP-based confidence measure (WPP).

ダム選択と比べて顕著に良い結果となった。例えば、74%の単語正解精度を達成するためにランダム選択では 110 時間の書き起こしが必要だが、提案手法では 60 時間で同じ精度を達成できた。単語事後確率を用いた手法と比べても 100 時間まではより良い結果を得ることができた。

6 まとめ

本稿では音声認識における書き起こしコスト削減を目的として、音声認識のためのコミッティ基準の能動学習法を提案した。プログレッシブ法を用いて複数の認識結果文のアライメントを行い、それらの間の不一致度を発話選択の基準とした。提案手法を CSJ を用いて評価し、ランダム選択より顕著に書き起こしを減らせることを実証した。100 時間より多い書き起こしを用いても単語正解精度は大きく向上しないことから 100 時間までの範囲で単語事後確率を用いた従来法よりも良い結果を得ることができたことは意義があると言える。

今後の課題としては、現在の複数のモデル作成方法 (等分割) よりすぐれた方法や、信頼度に基づく手法と組み合わせる方法を検討することが挙げられる。

謝辞 本研究は文部科学省研究費補助金基盤研究 (B) 20300063 による支援を受けた。

参考文献

[1] D. Hakkani-Tur, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in Proc. ICASSP, pp.3904-3907, 2002.

[2] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," IEEE Trans. Speech Audio Process., vol. 13, no. 4, pp.504-511, 2005.

[3] B. Varadarajan, D. Yu, L. Deng, and A. Acero, "Maximizing global entropy reduction for active learning in speech recognition," in Proc. ICASSP, pp.4721-4724, 2009.

[4] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in Proc. Workshop on Comput. Learning Theory, pp.287-294, 1992.

[5] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," in Machine Learning, vol. 28, pp.133-168, 1997.

[6] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," in Proc. ICML, pp.150-157, 1995.

[7] D.-F. Feng and R. F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees," Molecular Biology and Evolution, vol. 13, pp. 93-104, 1996.

[8] Sokal R. R. and Michener C. D., "A statistical method for evaluating systematic relationships," University of Kansas Science Bulletin, vol. 28, pp. 1409-1438, 1958.

[9] K. Maekawa, H. Koiso, S.Furui, and H. Isahara, "Spontaneous speech corpus of japanese," in Proc. LREC, vol. 2, pp.947-952, 2000.

[10] "Hidden Markov Model Toolkit (HTK)." [Online]. Available: http://htk.eng.cam.ac.uk/

[11] "The SRI Language Modeling Toolkit." [Online]. Available: http://www.speech.sri.com/projects/srilm/