Vol.2009-SLP-79 No.4
2009/12/21

音声認識のための複数の認識器を利用した能動学習

濱中悠三[†1] 江森正[†2] 越仲孝文[†1,†3]
篠田浩一[†1] 古井貞熙[†1]

大語彙連続音声認識器の学習データに対する書き起こしコスト削減のための複数の認識器を利用した能動学習手法を提案する．この手法では複数の認識器から得られた複数の異なる認識結果文を用いて発話の選択を行う．認識結果文をアラインメントするためのプログレッシブ法と Voting Entropy を発話選択に用いている．提案手法を日本語話し言葉コーパスの 190 時間の音声データを使い評価し，能動学習を行わないランダムな発話選択より顕著に良い結果を得た．74%の単語正解精度を得るために必要な書き起こし付きデータ量はランダム選択では 97 時間，単語事後確率を用いた従来手法では 72 時間であるが，提案手法では 60 時間で済むという結果になった．

Active learning using multiple recognizers for speech recognition

Yuzo Hamanaka ,[†1] Tadashi Emori ,[†2]
Takafumi Koshinaka ,[†1,†3] Koichi Shinoda [†1]
and Sadaoki Furui[†1]

We propose an active learning method with multiple recognizers for large vocabulary continuous speech recognition. In this approach, the recognition results obtained from recognizers are used for selecting utterances. Here, a progressive search method is used for aligning sentences, and voting entropy is used as a measure for selecting utterances. Our method was evaluated by using 190-hour speech data in the Corpus of Spontaneous Japanese. It proved to be significantly better than random selection. It only required 60 h of data to achieve a word accuracy of 74%, while standard training (i.e., random selection) required 97 h of data. The recognition accuracy of our proposed method was also better than that of the conventional uncertainty sampling method using word posterior probabilities as the confidence measures for selecting sentences.

1. はじめに

統計的音声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．人手によるデータのラベル付けは多くのコストが掛かるため，コストを減らすために様々な研究がなされている．能動学習もその 1 つであり，決められたデータ選択基準で選ばれた少量のラベルなしデータを書き起こし，学習に使用する．能動学習の主な研究課題はすぐれた選択基準を考案することであり，研究目標は一定の認識精度をより少ないラベル付き学習データで得ることである．

音声認識のための能動学習の研究は数多くあり[1)–4)]，多くが信頼度による Uncertainty Sampling を行っている[1)–3)]．Uncertainty Sampling では最初に少量のラベル付きデータを使って初期認識器を学習し，認識器を用いて全てのラベルなしデータを認識する．認識したデータの中から認識結果の信頼度の低い発話が選択される．信頼度としては発話に含まれる単語の事後確率 (word posterior probabilities; WPPs) がしばしば使用される．他の例では Varadarajan ら[3)] は認識器から得られたそれぞれの発話の単語ラティスのエントロピーを利用した．

本稿では大語彙連続音声認識のための Query by Committee(QBC)[5)] に基づく新しい能動学習手法を提案する．この手法では複数の音声認識器を作成し，それらによる認識結果文の不一致度が高い発話を書き起こす．Dagan ら[6)] は QBC に基づく能動学習の有効性を品詞タグ付け問題で確認し，Tur ら[7)] はその有効性をコールタイプ分類問題に適用し確認した．我々はこの手法を音声認識に適用し，その有効性を従来手法と比較することで示す．

2. アルゴリズム概要

QBC に基づく音声認識のための能動学習アルゴリズムの概略図を図 1 に示す．書き起こし付き学習データを T，書き起こされていない学習データを U とする．音声認識器の数を K，アルゴリズムの繰り返しの 1 サイクルで選択するデータの時間量を N(h) とする．アルゴリズムは以下の 5 ステップから成る．

(1) 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 $(k = 1, ..., K)$.

†1 東京工業大学 (Tokyo Institute of Technology)
†2 株式会社 NEC 情報システムズ (NEC Informatec Systems, Ltd.)
†3 日本電気株式会社 (NEC Corporation)

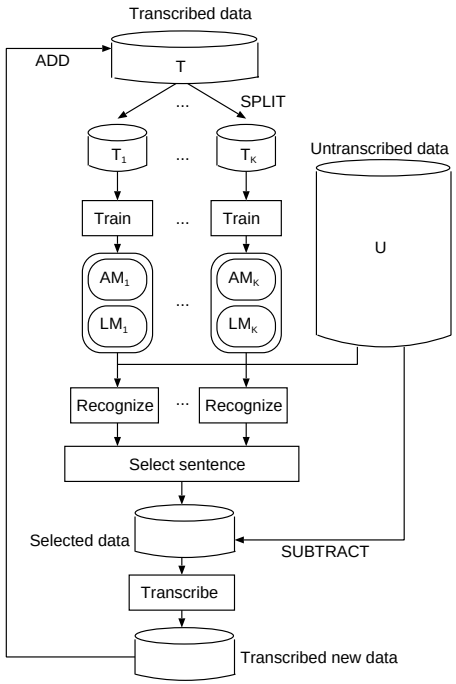

図 1 Active learning scheme using query-by-committee based approach for speech recognition.

(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，書き起こす．T に追加して，1. に戻る．

以上の繰り返しを書き起こしコストが尽きるまで行う．最後に，書き起こした全てのデータを用いて認識器を作成して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 3 章で述べる．

3. Query by Committee に基づく発話選択

3.1 Query by Committee

Query by Committee[5] は汎化誤差を減らすためのデータ選択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ありの学習データから複数の分類器 (コミッティ) を作成し，コミッティによる分類結果が最も一致しないデータを学習データとして使用することで，線形分離可能な場合のパーセプトロンなどの 2 クラス分類問題に対して，汎化誤差を指数的に減らせることを理論的に証明している．

線形分離可能や 2 クラス分類問題などの制約のため QBC の理論を音声認識の確率的手法にそのまま適用することはできない．しかし，このような制約に従わない場合においても認識 (分類) 結果が一致しないデータを選択することで汎化誤差を減らせることが実験的に示されている[6]．ここでは認識結果文の不一致度が高い発話を選択する．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．

3.2 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，アライメントを行う必要がある．ペアワイズアライメントを 3 本以上の配列に拡張したマルチプルアライメントの計算量は対象とする文の数が増えるに従い指数的に増加するため，現実的な時間では計算不可能なことが多い．そのため，それを近似したより計算量の少ない手法が数多く研究されている．複数の音声認識器から得られた認識文をアライメントし，それらを統合することで認識精度を改善する ROVER の研究[8] では，ベースとなる文と残りの文を 1 つずつペアワイズアライメントしていくことでアライメントを行っている．この方法はアライメント精度自体には注力しておらず，アライメントの順番によって結果が異なるという問題がある．そこで，ここではより高精度なアライメントが期待できる方法として，バイオインフォマティクスの分野でしばしば用いられるプログレッシブ法を用いる．プログレッシブ法のアルゴリズムを以下に示す．

(1) アライメント対象の全ペア間の類似度を用いて案内木を作成する．
(2) 案内木中で最初に作られた節点から最後に作られた節点の順番に，全ての文がアライメントされるまで節点間のアライメントを行う．節点間のアライメントは文対文，文対アライメント結果文 (以下結果文)，結果文対結果文のアライメントの 3 つの場合がある．

案内木の作成は UPGMA 法を用いて行う．以下，DNA の塩基配列のアライメントを例に取り説明する．

初期化: 各文 s_i に対して，その文のみから成るクラスタ C_i を作る．全ての文のペア s_i, s_j

Vol.2009-SLP-79 No.4
2009/12/21

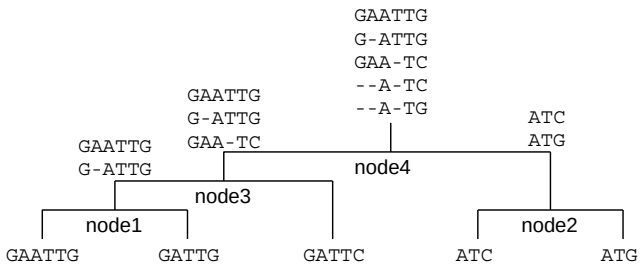

図 2 An example of guide trees for base sequences. The symbol "-" indicates a gap.

についてペアワイズアライメントを行い，match,mismatch,gap のコストの平均をクラスタペア C_i, C_j の類似度 d_{ij} とする．ここではコストはそれぞれ 1,0,0 とする．

繰り返し: d_{ij} が最大となるペア C_i, C_j をクラスタリングし，クラスタ C_k を作る．他のクラスタ C_l 全てに対して類似度 d_{kl} を計算する．クラスタ C_k に含まれる文のインデックスの集合を X，クラスタ C_l に含まれる文のインデックスの集合を Y とするとき，d_{kl} を式 (1) のように計算する．

$$d_{kl} = \frac{1}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} d_{s_x s_y} \tag{1}$$

終了: クラスタが 2 つだけとなったら終了する．クラスタリングしていく順番を使って案内木を作成する．

次に，節点間のアライメントを行う．作成された案内木の例を図 2 に示す．"-" はギャップを表す． 文対文 (図 2 の例では節点 1 と節点 2) のアライメントは通常のペアワイズアライメントを行う．文対結果文，または結果文対結果文 (図 2 の例では節点 3 と節点 4) のアライメントを行うときは，表 1 のようにペアワイズアライメントの DP マッチングを拡張してアライメントを行う．このとき，結果文の元々のアライメント関係は維持して，つまり結果文にギャップを挿入する際は結果文を構成する全ての文の同じ列にギャップを挿入して，最適なアライメントを探索する．例えば図 2 の節点 3 の例で，"GAATTG" の 3 列目と 4 列目の間にギャップを挿入するときは "G-ATTG" の 3 列目と 4 列目の間にも必ずギャップを挿入する．結果として，それぞれの節点間のアライメントは以下の SP スコア $S(m_c)$ の各列 c $(c = 1, ..., C)$ に渡る合計が最大となるように行われる．

表 1 An example DP-matrix. The symbol "-" indicates a gap.

			G	A	A	T	C
		0	0	0	0	0	0
G	G	0	6	3	0	0	0
A	-	0	3	6	3	0	-3
A	A	0	3	9	12	9	6
T	T	0	3	9	12	18	15
T	T	0	3	9	12	18	18
G	G	0	6	9	12	18	18

表 2 An example alignment result of sentence vs. alignemnt-result ($C = 6, H = 3$).

		c					
		1	2	3	4	5	6
h	1	G	A	A	T	T	G
	2	G	-	A	T	T	G
	3	G	A	A	-	T	C

$$\sum_{c=1}^{C} S(m_c) = \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{h'=h+1}^{H} s(m_c^h, m_c^{h'}) \tag{2}$$

ここで式 (2) の c はアライメント結果の先頭からの列番号である．m_c^h はアライメント結果を構成する h 番目 $(h = 1, ..., H)$ の文の列 c にある単語である．コスト $s(a, b)$ の値は以下を用いる．

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

この値はいくつかの値で予備実験した中で最も良い結果の値である．表 2 にアライメント結果の例を示す．

3.3 Voting Entropy

マルチプルアライメント結果の列ごとに Voting Entropy を計算し，不一致度を定義する．列 c にある単語の種類数を P，それぞれの単語を w_p $(p = 1, ..., P)$ とする．K は認識器の数である．列 c に単語 w_p が出現する回数を $V(w_p, c)$ とするとき，列 c における Voting Entropy VE(c) を以下のように定義する．

$$\mathrm{VE}(c) = -\sum_{p=1}^{P} \frac{V(w_p, c)}{K} \log \frac{V(w_p, c)}{K} \tag{3}$$

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．マルチプルアライメント結果の全ての列 c $(1 \leq c \leq C)$ に渡る VE(c) の平均をその発話の認識結果文の不一致度 D と定義し，D が大きい発話から発話選択を行う．

$$D = \frac{1}{C} \sum_{c=1}^{C} \mathrm{VE}(c) \tag{4}$$

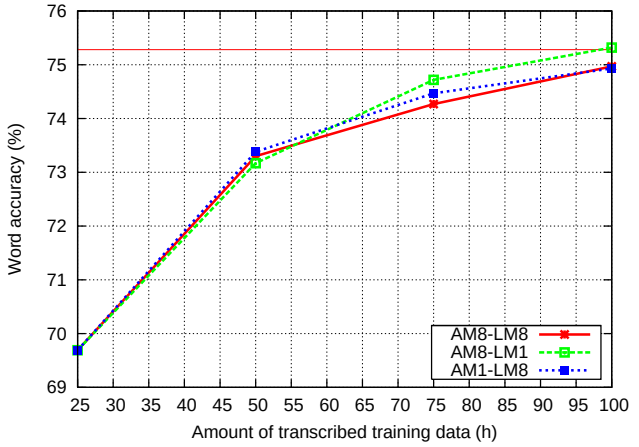

図 3 Recognition results with different model combinations. The horizontal solid line showed the recognition result (75.2%) obtained by using all the training data (190h) we prepared for the experiment.

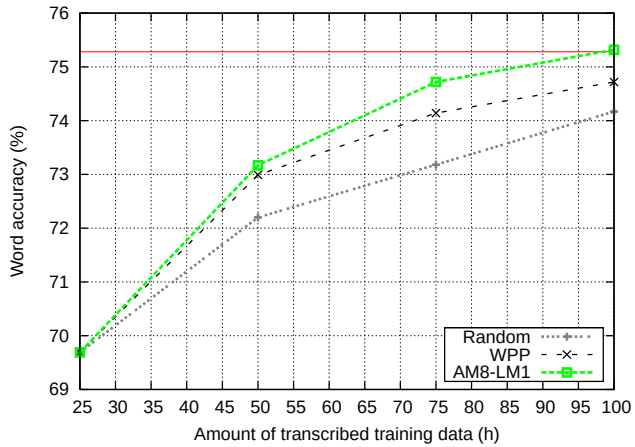

図 4 Recognition results with different selection methods: random selection (Random), WPPs-based confidence measure (WPP), and proposed method (AM8-LM1).

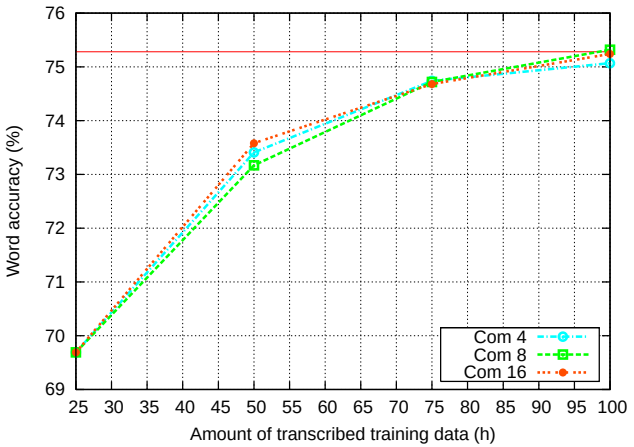

図 5 Recognition results with different numbers (4,8,16) of recognizers. Here the same language model was shared among the recognizers.

4. 実　　験

4.1 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) 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．実験用データの内，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．

特徴量は MFCC12 次元とパワー，及びその 1 次微分成分と 2 次微分成分の計 39 次元，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 triphoneHMM を用いた．言語モデルは 1 パス目に 2gram，2 パス目に 4gram を用いた．実験には HTK を使用した[9]．

全学習データからランダムに選択された 29,461 発話 (25 時間) を書き起こし付きデータ T として初期の音響モデルと言語モデル学習に使用し，残りの学習データを書き起こしなしデータ U とし，1 サイクルで選択するデータの時間量 N は 25 時間とした．

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 (WPPs) に基づく選択[1] であり，発話文中の単語事後確率平均が低い発話から順に選択した．

4.2 実験結果

複数の認識器を作成するための方法として以下の 3 つを用いた．

- 音響モデル (AM)，言語モデル (LM) 共に K 分割したデータ T を用いて学習し，$\mathrm{AM}_k, \mathrm{LM}_k (k = 1, ..., K)$ のペアを用いて認識器を作成する (AM8-LM8)．
- K 分割したデータ T を用いて音響モデルを学習し，全データ T を用いて言語モデルを学習する (これを $\mathrm{LM}_{\mathrm{all}}$ とする)．$\mathrm{AM}_k, \mathrm{LM}_{\mathrm{all}} (k = 1, ..., K)$ のペアを用いて認識器を作成する (AM8-LM1)．
- 全データ T を用いて音響モデルを学習し (これを $\mathrm{AM}_{\mathrm{all}}$ とする)，K 分割したデータ T を用いて言語モデルを学習する．$\mathrm{AM}_{\mathrm{all}}, \mathrm{LM}_k (k = 1, ..., K)$ のペアを用いて認識器を作成する (AM1-LM8)．

K を 8 とし，上記 3 つの方法で複数の認識器を作成したときの実験結果を図 3 に示す．結果は似ているが，言語モデルを共有して認識器を作成した場合が最も良い結果となった．これは言語モデルの学習は音響モデルの学習より多くの学習データが必要なためと思われる．

図 4 に，ランダム選択と単語事後確率に基づく選択と図 3 で最も結果が良かった複数の

Vol.2009-SLP-79 No.4
2009/12/21

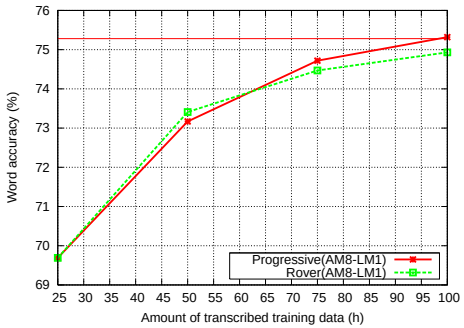

図 6 Recognition results with different sentence alignment method, progressive search and ROVER. The number of recognizers is 8.

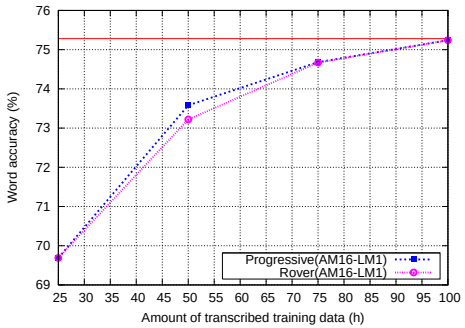

図 7 Recognition results with different sentence alignment method, progressive search and ROVER. The number of recognizers is 16.

認識器方法 (AM8-LM1) の手法の比較結果を示す．提案手法はランダム選択と比べて顕著に良い結果となっている．74%の単語正解精度を達成するために提案手法では 63 時間で済むがランダム選択では 97 時間掛かる．事後確率に基づく選択と比較してもより良い結果となっている．また提案手法ではデータ T (190 時間) 全てを書き起こして学習に用いたときの単語正解精度を 100 時間のデータで達成できている．

図 5 に K を変えたときの認識精度を示す．K を 4,8,16 と変化させ，$\mathrm{AM}_k, \mathrm{LM}_{\mathrm{all}} (k = 1, ..., K)$ のペアを用いて認識器を作成した．K を変化させても結果にそれほど大きな違いは見られなかった．1 サイクルの計算量は認識器の数に比例して増加することを考えると K は 4 で十分である．

図 6,7 に 2 つのアライメント方法 (プログレッシブ法と ROVER) による認識精度の違いを示す．全体的にはプログレッシブ法の方が ROVER より精度が良い結果となった．これはプログレッシブ法の方がアライメントの精度が良いため認識結果の不一致度を正確に測定できたためと思われる．

5. ま と め

Query by Committee に基づく音声認識のための複数の認識器を利用した能動学習の手法を提案した．書き起こし付き学習データからランダムに選択したデータを用いて複数の認識器を作成した．プログレッシブ法を認識結果文のアライメントに用いて，VE によって定義された不一致度を発話選択に用いた．提案手法を CSJ を使って評価し，ランダム選択や事後確率を用いた従来手法より良い結果を得た．データをランダムに等分割する複数の認識器の作成方法では，言語モデルは分割しない方が良いことがわかった．また認識結果文のアライメント手法としては ROVER よりもプログレッシブ法が優れていることを確認した．

今後の課題としては，現在の複数の認識器の作成方法よりすぐれた方法や信頼度に基づく手法と組み合わせた方法の考案が挙げられる．

謝辞 本研究は，科学研究費補助金基盤研究 (B) 2030063 の援助を受けた．

参 考 文 献

1) D.Hakkani-Tur, G.Riccardi, and A. Gorin: Active learning for automatic speech recognition, Proc. ICASSP, pp.3904-3907, (2002).
2) G.Riccardi and D.Hakkani-Tur: Active learning: Theory and applications to automatic speech recognition, Trans. IEEE, 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, *Proc. ICASSP*, pp.4721-4724, (2009).
4) H.Lin, and J.Bilmes: How to select a good training-data subset for transcription: submodular active selection for sequences, *Proc. Interspeech*, pp.2859-2862, (2009).
5) H.S.Seung, M.Opper, and H.Sompolinsky: Query by committee, *Proc. Workshop on Comput. Learning Theory*, pp.287-294, (1992).
6) I. Dagan and S.P.Engelson: Committee-based sampling for training probabilistic classifiers, *Proc. ICML*, pp150-157, (1995).
7) G.Tur, R.Schapire, and D.Hakkani-Tur: Active learning for spoken language understanding, *Proc. ICASSP*, Vol.1, (2003).
8) J.G.Fiscus: A post-processing system to yield reduced error word rates: Recognizer output voting error reduction (ROVER), *Proc. IEEE Workshop on Automatic Recognition and Understanding*, pp.347-354, (1997).
9) The Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/.