論文 / 著書情報
Article / Book Information

論題(和文)	コミッティに基づく能動学習・半教師付き学習を用いた音声モデル
Title(English)	Speech modeling using committee-based active and semi-supervised learning.
著者(和文)	蔦岡拓也, 篠田浩一
Authors(English)	Takuya Tsutaoka, Koichi Shinoda
出典(和文)	日本音響学会2012年春季研究発表会講演論文集, Vol. , No. , pp. 55-56
Citation(English)	, Vol. , No. , pp. 55-56
発行日 / Pub. date	2012, 3

コミッティに基づく能動学習・半教師付き学習を用いた音声モデル*

☆蔦岡拓也, 篠田浩一 (東工大)

1 はじめに

母語話者の人口が少ない言語のための音声認識システム開発では，需要の低さから大規模なコーパスの構築に掛けられるコストが限られてしまう．書き起こしコストを削減しつつ十分な認識精度を得るため，能動学習や半教師付き学習，その両者の組み合わせる学習の研究が行われている [1, 2, 3, 4, 5, 6].

本稿では，コミッティに基づく能動学習法 [2] に基づいて，能動学習と半教師付き学習を組み合わせる学習法を提案する．また，半教師付き学習における発話選択手法の検討も行う．

2 コミッティに基づく能動学習と半教師付き学習を組み合わせる学習法

重複が無いように分割した書き起こし付き発話データから，それぞれ相異なる複数の認識器を学習してコミッティを構成し，各認識器による書き起こし無し発話の認識結果の不一致度をもとに発話選択を行う．不一致度の高い発話は人手で書き起こして能動学習を行い，不一致度の低い発話は認識結果を書き起こし文として用いて半教師付き学習を行う．具体的なアルゴリズムを以下に示す (Fig.1).

まず，書き起こし付き学習データを T，書き起こし無し学習データを U とし，コミッティを構成する認識器の数を K，一度に選択する発話の時間量を N とし，能動学習のために選択する発話セットを T_a，半教師付き学習のために選択する発話セットを T_s とする．次に以下の手続きを行う．

(1) データ T をランダムに等分割し，データセット $T_k(k = 1, \ldots, K)$ を作成する．

(2) T_k を用いて認識器 M_k を学習する．

(3) データ U の全ての発話を M_k で認識し，各認識結果の不一致度 D を求める．D の計算方法は (i) から (iii) に示す.

(4-1) D の低い順に N 時間分の発話を選択する (T_a).

(4-2) D の高い順に N 時間分の発話を選択する (T_s).

(5-1) T_a を U から取り除き，人手による書き起こしを行って T に追加する．(1) に戻る．

(5-2) T_s を U から取り除き，認識結果文を書き起こしに用いて T に追加する．認識結果文は T を分割せずに学習した認識器から得る．(1) に戻る．

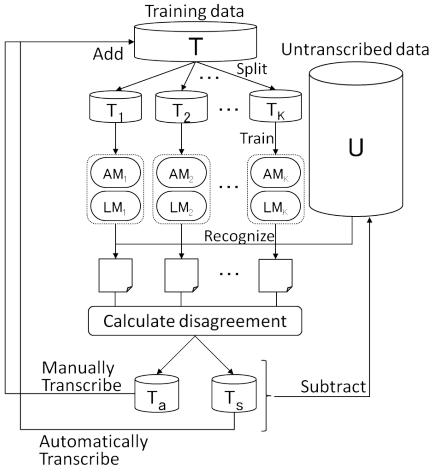

Fig. 1 Scheme of the proposed method.

不一致度 D の計算方法は以下の通りである．

(i) K 個の異なる認識結果文に対し，プログレッシブ法によってマルチプルアライメントを行い，結果文の単語列数を揃える．

(ii) 列 c に P 種類の単語 $w_p(p = 1, \ldots, P)$ が存在するとする．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 Vote Entropy を次のように定義する．

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

(iii) 発話内全ての列 $c(1 \leq c \leq C)$ に渡る $VE(c)$ の平均をその発話の不一致度 D とする．

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

3 半教師付き学習における発話選択の改良

信頼度に基づく半教師付き学習では，信頼度の高い発話のみを学習に用いると学習データに偏った発話選択が起こり，認識精度の低下を引き起こす場合があることが報告されている [4]．この問題はコミッティに基づく半教師付き学習においても生じる可能性がある．

そこで本研究では，もとの学習データに偏った発話選択を防ぐため，信頼度や不一致度が平均値付近の発話を選択する手法を試みた．単語事後確率を用いた信頼度に基づく学習 (WPP) では，信頼度が平均以上の発話を信頼度が低い順に選択する (WPP(avg)). また，不一致度に基づく場合 (Com) は，不一致度が平均値以下の発話を不一致度が高い順に選択する (Com(avg)).

*Speech modeling using committee-based active and semi-supervised learning. by Takuya Tsutaoka and Koichi Shinoda (Tokyo Institute of Technology)

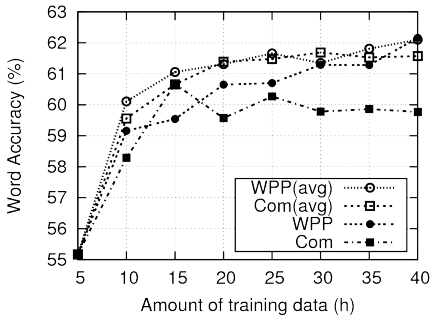

Fig. 2 Recognition results of semi-supervised learning.

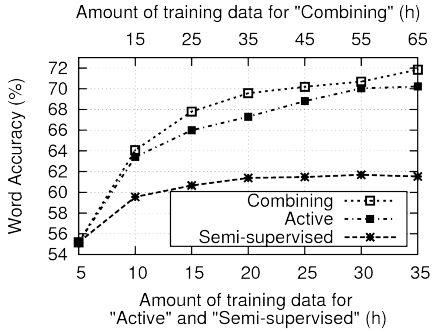

Fig. 3 Recognition results of committee-based learnings.

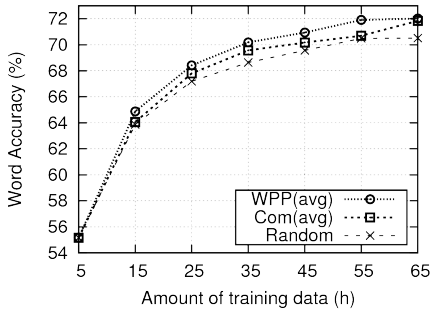

Fig. 4 Recognition results of active and semi-supervised learning.

4 評価実験

4.1 実験条件

日本語話し言葉コーパス [7] の摸擬講演音声を用いた．285 時間を学習データ，1.52 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 次微分と 2 次微分の計 39 次元を用いた．音響モデルは 16 混合のトライフォン HMM を用い，状態数は MDL[8] に基づいて決定した．T は 5 時間分の発話，U は 280 時間分の発話，分割数 K は 4，一度に選択する発話の量 N は 5 時間分とした．認識は 2 パスサーチを行い，言語モデルは 1 パス目に 2gram，2 パス目に 4gram を用いた．

4.2 実験結果

Fig.2 に，半教師付き学習において，信頼度に基づく手法 (WPP) とコミッティに基づく手法 (Com) に対し，発話選択の改良を行った結果 (WPP(avg)，Com(avg)) を示す．信頼度に基づく手法，コミッティに基づく手法ともに，発話選択の改良によって認識精度が向上した．改良した手法によって偏った発話選択を防ぎ，効率良く学習が行えていることが分かった．

Fig.3 に，コミッティに基づく能動学習 [2](Active)，コミッティに基づく半教師付き学習 (Semi-supervised)，それらを組み合わせる提案手法 (Combining) による比較結果を示す．ここで，提案手法では，一度に追加される 10 時間分の発話のうち人手で書き起こす発話は 5 時間分である．提案手法と能動学習について，人手による書き起こしコストが等しい時点 (例えば組み合わせ学習の 25 時間と能動学習の 15 時間) で比較した場合，能動学習と半教師付き学習を組み合わせることで，能動学習を単独で行うよりも少ない書き起こしコストで高い認識精度が得られた．また，提案手法は半教師付き学習を単独で行うよりも良い結果となった．

Fig.4 に，能動学習と半教師付き学習を組み合わせる学習において，ランダム選択 (Random)，信頼度に基づく手法 [5](WPP(avg))，提案手法 (Com(avg)) の比較結果を示す．提案手法はランダム選択よりも良い結果だったが，信頼度に基づく手法を下回った．

5 まとめ

能動学習と半教師付き学習を組み合わせる新たな学習法を提案した．2 つの学習を組み合わせることで，能動学習を単独で行うよりも良い結果となった．提案手法はランダム選択より良い結果を示したが，単語事後確率を用いた信頼度に基づく手法を下回った．また，半教師付き学習における発話選択の改良によって認識精度が向上した．

提案手法が信頼度に基づく手法を下回った原因を分析し，半教師付き学習におけるより高性能な発話選択手法の考案を行うことが今後の課題である．

参考文献

[1] D. Hakkani-Tür *et al.*, Proc. ICASSP, pp.3904-3907, 2002.

[2] Y. Hamanaka *et al.*, Trans. IEICE, vol.E94-D, No.10, pp.2015-2023, 2011.

[3] T. Kemp *et al.*, Proc. Eurospeech, pp.2725-2728, 1999.

[4] R. Zhang, Proc. ICASSP, pp.421-424, 2006.

[5] G. Tur *et al.*, Journal of Speech Communication 45 (2), pp.171-186, 2005.

[6] D. Yu *et al.*, Journal of Computer Speech and Language 24 (3), pp.433-444, 2010.

[7] M. Maekawa *et al.*, Proc. LREC, vol.2, pp.947-952, 2000.

[8] K. Shinoda *et al.*, J. Acoust. Soc. Jpn. (E), vol. 21, no. 2, 2002.