

タスク依存音響モデルのための発話レベルでの選択学習法

ツインツアレク・トビアス[†] 戸田 智基[†] 猿渡 洋[†] 鹿野 清宏[†]

† 奈良先端科学技術大学院大学

情報科学研究科音情報処理学講座

〒 630-0192 奈良県生駒市高山町 8916-5

E-mail: †{cincar-t,tomoki,sawatari,shikano}@is.naist.jp

あらまし 高性能音響モデルを構築するために、音声データが大量に必要である。音響モデルの認識性能が対象タスクによるので、タスク別に音響モデルを準備する必要がある。しかし、音声データの収集と書き起こしにおけるコストが膨大であり、任意タスクのために十分の音声データを用意するのが困難である。本稿では、コスト削減を目的にした発話単位の選択学習法を検討する。提案手法は、既存の音声データベースを利用し、対象タスク用の開発データに対する尤度が上昇するように、学習発話を選択する。十分統計量を用いることで、尤度計算は高速に可能である。評価実験において、小学生の音声データで幼児モデル、大人の音声データで高齢者モデルを構築する選択学習を適用した。選択学習は10発話程度の開発データの場合にも有効であった。又、選別した発話で再学習した音響モデルの認識性能は、開発データに基づくMAPとMLLR適応で得られたモデルより優位であった。

キーワード 音響モデル、タスク依存性、コスト削減、選択学習、十分統計量

Utterance-based Selective Training for Task-Dependent Acoustic Modeling

Tobias CINCAREK[†], Tomoki TODA[†], Hiroshi SARUWATARI[†], and Kiyohiro SHIKANO[†]

† Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, Nara-ken, 630-0192 Japan

E-mail: †{cincar-t,tomoki,sawatari,shikano}@is.naist.jp

Abstract Large amounts of speech data are necessary to construct high performance acoustic models. Since speech recognition performance is task-dependent and the effort and costs for speech data collection and transcription are very high, it is infeasable to prepare enough data for every new application which makes use of speech recognition technology. In this paper an algorithm for utterance-based selective training is proposed, which enables the automatic and cost-effective construction of task-dependent acoustic models. Training utterances are selected from existing speech data resources so that the likelihood of an independent development data set is maximized. Fast calculation of the likelihood is possible with sufficient statistics. The algorithm is evaluated for constructing an infant-dependent model with speech from elementary school children and an elderly-dependent model with adult speech data. Selective training is already effective with only ten development utterances. Furthermore, a higher word accuracy than with the standard adaptation methods MAP and MLLR was achieved.

Key words Acoustic Modeling, Task-Dependency, Cost Reduction, Selective Training, Sufficient Statistics

1. まえがき

ディクテーションシステム、音声対話システム、音声翻訳システム、ロボットなどを始め、音声認識技術を活用するアプリケーションは数多く存在する。アプリケーション毎に音響的かつ言語的な状況が異なるため、全てに対して同様の認識システムを使用することは極めて困難である。そのため、発話内容に

適した言語モデル、又は話者特徴や環境状況に合わせた音響モデルの構築が欠かせない。

通常高性能音響モデルのパラメータ数は膨大であるため、頑健なモデル学習には音声データが大量に必要である。しかし、音声データの収集及び整備におけるコストが非常に高い[1]。また、あらゆるタスクのために十分の学習データを用意するのは現実的ではない。それらが、音声認識技術を活用するアプリ

ケーションをより広く普及させるまでの障害となっている。

収集した音声データをより効率的に活用する手法として教師なし学習[2], [3]と能動学習[4], [5]が使われている。能動学習では、全てのデータを書き起こすより、認識信頼度を用いることで認識しにくい発話のみを書き起こす。準備した学習データに対する更新モデルの尤度の上昇値が止まるまで、または、経済的な手段が尽くされない限り、データ収集とラベルづけを続ける。全データより、選別したデータのみで構築したモデルの方が性能が高いという結果が報告されている。教師なし学習では、収集した音声データを既存の音響モデルと言語モデルで認識し、自動的に書き起こしたデータでモデルを再学習する。いずれの方法も、書き起こしに必要な時間とコストが削減されるが、大量に音声データを収録しなければならないという欠点は克服できない。

複数の音声コーパス使用による、タスク非依存音響モデル構築に関する研究も行われている[6]。タスク非依存とタスク依存モデルの性能差異が小さく、自発話音声の認識精度が向上した反面、数字認識精度が低下した結果が報告されている。

本稿では、既存の音声データを利用して、タスク依存音響モデルを発話単位選択学習で構築する方法を検討する。提案手法では、従来EM学習時に用いる学習データに対する最尤基準ではなく、対象タスク用の少量の開発データに対する尤度が最大になるように、音響モデルを構築する。尤度の計算は高速に十分統計量に基づいて可能である。幼児音響モデルを小学生音声データ、高齢者音響モデルを大人音声データで構築し、提案法とMAP推定、MLLR適応の比較を行う。以下、節2で発話単位選択学習法、節3で実験条件、節4で評価実験の結果について述べる。

2. 提案アプローチ

通常は対象タスクに対して大量の音声データの収録と書き起こしを行い、音響モデルを構築する。既存の音声データを活用すれば、データ収集及び準備における膨大なコストを抑えられるが、対象タスクとの音響的な特徴があまりにも異なっていれば、高性能な音声認識の実現が困難になる。そこで、既存の音声データを全部用いるより、学習データを選別し、対象タスクに近いデータのみで音響モデルを構築するアプローチが考えられる。こういう選択学習法を具体的に実現するために、選択単位と選択基準と選択アルゴリズムを定めなければならない。

過去に提案された音響モデル用選択学習及び適応法のうち、選択単位として主に話者単位[7], [8]が使われている。公共の場所に設置されている音声対話システム、例えば「たけまるくん」[9]で収録した音声データの場合、各話者が発声している発話数が少数に過ぎなく、各発話の話者ラベルも不明である。従って、話者単位の選択は不可能である。話者よりも小さい単位は発話である。発話単位を使用することで、話者と比べより細かいデータ選択が可能になる。

図1は提案する選択学習の枠組を示す。提案法の目的は、対象タスクに対応できる音響モデル構築なので、学習データに対する尤度や認識信頼度は選択基準として適していないことが明

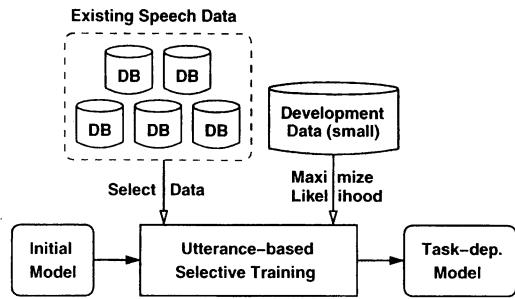


図1 選択学習の枠組

らかである。提案法では開発データに対する尤度最大化基準を用いる。モデルパラメータの値は学習データから推定されるが、尤度が学習データと完全に異なる開発データに対して計算される。コスト削減のため、対象タスクの音響的特徴を、1,000発話以下の少量の開発データのみで最低限カバーする。

学習データは、発話数が少くとも、部分集合の選択肢が膨大で、一部しか検討できない。本稿では、選択アルゴリズムとして、学習データ全部が選択された状態で探索を開始し、逐次的にそれぞれの発話を検討する。ある発話を抜くことで、尤度が上昇した場合は、該当する発話を破棄し、減少した場合は、該当する発話を最終的な音響モデル構築に用いる。一度破棄した発話をまた加える戦略も考えられる。また、この選別処理を数回繰り返すことも可能である。高速可能である尤度計算を節2.1で詳しく説明し、詳細な発話選択アルゴリズムを節2.2で解説する。

2.1 開発データに対する高速尤度計算

通常の音響モデルとして、それぞれの音素が状態毎に混合正規分布を持つHMMが用いられる。HMMのパラメータは大量の学習データに基づいてEMアルゴリズム[10]で推定できる。最適化基準は次の補助関数である。

$$Q(\Theta, \hat{\Theta}) = \sum_{\vec{s}} P(\vec{s}|\mathcal{D}, \Theta) \log P(\vec{s}, \mathcal{D}|\hat{\Theta}) \quad (1)$$

ここで、

- 初期モデルパラメータ Θ
- 推定したパラメータ $\hat{\Theta} = \{\hat{\mu}_{qm}, \hat{\sigma}_{qm}, \hat{w}_{qm}, \hat{a}_{qq'}\}$
- 対象タスク用の開発データ $\mathcal{D} = \{\vec{x}\}$
- 音声の特徴ベクトル時系列 x_t
- 学習データ \mathcal{T}
- 状態及び混合指數系列 \vec{s}

である。従来のEM学習では、式1の \mathcal{D} が学習データである。提案するEM選択学習では、新パラメータ $\hat{\Theta}$ は学習データで推定するが、尤度は開発データ \mathcal{D} に対して計算する。

次に Q 関数は学習と開発データの十分統計量のみで表せる事を示す。ここでは、公式の簡易化のため、 x_t を一次元と仮定し、HMMの状態遷移確率 $\hat{a}_{qq'}$ を記述しないが、実際の計算では、多次元特微量ベクトルを用い、状態遷移確率も考慮している。 Q 関数は

$$\propto \sum_q \sum_m \sum_t \gamma_{qm}(t) \log \frac{\hat{w}_{qm}}{\sqrt{2\pi\hat{\sigma}_{qm}}} \quad (2)$$

$$- \sum_q \sum_m \sum_t \gamma_{qm}(t) \frac{1}{2} (x_t - \hat{\mu}_{qm})^2 \frac{1}{\hat{\sigma}_{qm}} \quad (3)$$

に比例する。ここで、 $\hat{\mu}_{qm}$ 、 $\hat{\sigma}_{qm}$ 、 \hat{w}_{qm} はそれぞれ混合正規分布の平均ベクトル、分散、重みである。 $\gamma_{qm}(t)$ は音声フレーム x_t に対する状態 q と混合指標 m の状態占有確率である。上式は次のように書き換えられる。

$$\propto \sum_q \sum_m y_{qm} \log \frac{\hat{w}_{qm}}{\sqrt{2\pi\hat{\sigma}_{qm}}} \quad (4)$$

$$- \sum_q \sum_m \frac{z_{qm} - 2\hat{\mu}_{qm}o_{qm} + \hat{\mu}_{qm}^2 y_{qm}}{2\hat{\sigma}_{qm}} \quad (5)$$

ここで、変数 y_{qm} 、 o_{qm} 、 z_{qm} はそれぞれ開発データ D の十分統計量である。新モデルパラメータ $\hat{\mu}_{qm}$ 、 $\hat{\sigma}_{qm}$ 、 \hat{w}_{qm} は学習データ T の十分統計量 μ_{qm} と σ_{qm} と c_{qm} による関数として表せ、学習発話 u_i 毎に分解できる。例えば、平均ベクトルの場合、開発データの十分統計量とモデルパラメータの再構成は次のようになる。

$$o_{qm} = \sum_t \gamma_{qm}(t) x_t \quad (6)$$

$$\hat{\mu}_{qm} = \frac{\mu_{qm}}{c_{qm}} = \frac{\sum_i \mu_{qm}^i}{\sum_i c_{qm}^i} \quad (7)$$

発話毎の選択学習は学習発話の十分統計量を減算か加算することで実現できる。つまり、新パラメータ $\hat{\Theta}$ を変化させ、開発データ D に基づく Q 関数が最大になるように、発話選択を行う。

2.2 発話単位選択学習アルゴリズム

発話単位の学習データ選別戦略として $ST_DelScan$ と ST_DelAdd という二つのアルゴリズムを検討する。 $ST_DelScan$ では各発話を独立に抜き、尤度が上昇したら、該当する発話を棄する。

1. 全ての学習発話の集合を $R \leftarrow \{u_1 \dots u_n\}$ とする。
2. 各学習発話の十分統計量 $\{S_i\}$ を用意する。
3. 開発データの十分統計量 S_D を用意する。
4. 学習データの十分統計量 S_T を用意する。
5. 初期尤度 $q \leftarrow Q(\Theta, f(S_T))$ を計算する。
6. R に含まれている各発話 u_i に対して：
 - a. 尤度 $q' \leftarrow Q(\Theta, f(S_T - S_i))$ を計算する。
 - b. $q' > q$ が成立したら、発話 u_i を棄する。
 発話集合の更新 $R \leftarrow R - \{u_i\}$
7. R に含まれている発話で $\hat{\Theta}$ を復元する。
8. R に含まれている発話で学習を繰り返す。

ST_DelAdd では、各発話を操作する度に学習データの十分統計量と尤度を即時更新する。それぞれの発話が数回検討される上、一度抜いた発話を再び学習データに加える処理が行われる。発話を抜く処理と発話を加える処理が交代で繰り返される。

前記の $ST_DelScan$ のステップ (6.) を以下のステップ置き換えると ST_DelAdd アルゴリズムになる。

6. 以下の処理を予め指定した回数に繰り返す

- I. R に含まれている各発話 u_i に対して：
 - a. $q' \leftarrow Q(\Theta, f(S_T - S_i))$ 。
 - b. $q' > q$ が成立したら、発話 u_i を被棄する。
 発話集合の更新 $R \leftarrow R - \{u_i\}$
 十分統計量更新 $S_T \leftarrow S_T - S_i$
 尤度更新 $q \leftarrow q'$
- II. R に含まれていない各発話 u_i に対して：
 - a. $q' \leftarrow Q(\Theta, f(S_T + S_i))$ 。
 - b. $q' > q$ が成立したら、発話 u_i を使用する。
 発話集合の更新 $R \leftarrow R \cup \{u_i\}$
 十分統計量更新 $S_T \leftarrow S_T + S_i$
 尤度更新 $q \leftarrow q'$

両方のアルゴリズムに関して図2を参照されたい。 ST_DelAdd は $ST_DelScan$ より計算量が多い。 $ST_DelScan$ は並列処理可能であるが、 ST_DelAdd は尤度 q と選択した学習データに対する十分統計量が更新されるので、並列処理が不可能という欠点がある。

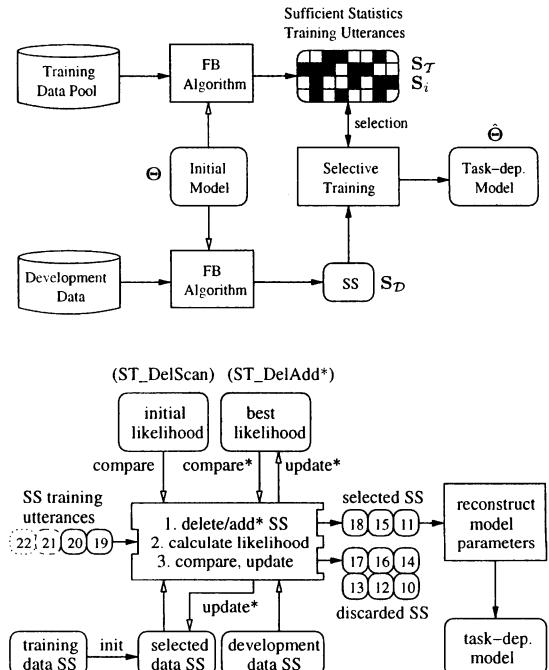


図2 選択学習の詳細図。処理 (*) は ST_DelAdd のみ。十分統計量は SS として記述されている。

過学習を緩和するために、音素毎の最低サンプル数、又は、 Q 関数の変動閾値を設定する必要がある。

3. 実験条件

ST_DelScan もしくは *ST_DelAdd* により選択された、選択されたデータで EM 学習を数回繰り返す。過学習を避けるため、各音素の最低サンプル数を 200 に設定する。

3.1 音声データ

実験において使用する音声データは公共施設に設置してある音声対話・音声案内システム「たけまるくん」[9]で収録したものである。2002 年 11 月のシステム稼働開始から今までの入力数が 30 万を越え、今現在 2 年分が完全に書き起こされている。また、発話全部が主観的に五つの話者グループ（幼児、小学生、中学生、大人、高齢者）に分類されている。表 1 に本実験において使用したデータを示す。実験 A では小学生のデータで幼児音響モデルを、実験 B では大人のデータで高齢者モデルを構築する。

表 1 実験用音声データ。実験 A では小学生のデータで幼児音響モデルを、実験 B では大人のデータで高齢者モデルを構築する。

実験	学習データ		開発データ		評価データ	
	分類	発話数／時間	分類	発話数／時間	発話数／時間	
A	小学生	30k / 17 h	幼児	500 / 17 m	1.5k / 53 m	
B	大人	18k / 9 h	高齢者	53 / 2 m	400 / 12 m	

3.2 音響モデルと言語モデル

モノフォンの音響モデルによりアルゴリズムの評価を行う。初期音響モデルは全学習データで構築する。音響的特徴量として、12 MFCC と 12 Δ MFCC と ΔE を用いる。音響モデルは 38 音素に対する 3 状態の HMM からなる。状態出力確率密度は対角分散行列を持つ混合正規分布によりモデル化する。混合数は最大 16 である。

実験 A では、約 15,000 発話の書き起こしで学習した幼児言語モデル、実験 B ではたけまる音声対話システム用 4 万語認識可能な大人用言語モデルを用いる。デコーダとして Julius[11]を使用する。

4. 結果と考察

4.1 選択学習アルゴリズムの性能評価

表 2 は従来 EM 学習と選択学習を比較した結果を示す。実験 A では全学習データで初期モデルを再学習しても、単語正解精度改善が僅か (0.9%) であるが、選択学習を用いることで認識精度が大幅 (5.4%) に上昇する。実験 B も同じ傾向が見られた。*ST_DelScan* も *ST_DelAdd* も同程度に有効であり、認識性能に関してどちらがより優れているとは言えない。

学習データの選択率は表 3 を参照されたい。*ST_DelAdd* では、破棄処理が 5 回繰り返されているので、最終的な選択率は *ST_DelScan* より低い。今回の評価で用いた音響モデルがモノフォン HMM からのもので、EM 学習における学習データ不足は問題ではないが、より数多くのパラメータを持つ音響モデルの場合、破棄処理を一回のみ行った方がより頑健な再学習行われる傾向がある。

表 2 選択した音声データで数回 EM 学習を行った音響モデルでの単語正解精度。

モノフォン HMM (A) 幼児モデル	学習回数					
	初期	1	2	3	5	8
No Selection	46.4	46.7	47.3	47.4	47.3	47.3
<i>ST_DelScan</i>	46.4	50.4	50.6	50.8	51.3	51.2
<i>ST_DelAdd</i>	46.4	50.7	51.1	51.6	51.6	51.8
(B) 高齢者モデル	初期	1	2	3	5	8
	No Selection	73.5	73.2	73.5	73.3	73.2
<i>ST_DelScan</i>	73.5	75.7	75.4	74.8	74.1	74.7
<i>ST_DelAdd</i>	73.5	74.9	74.5	74.1	74.2	74.5

表 3 提案法の学習発話選択数及び選択率。

実験 A	実験 B	
	<i>ST_DelScan</i>	<i>ST_DelAdd</i>
10,697 (36%)	4,302 (14%)	7,704 (43%)
		3,165 (18%)

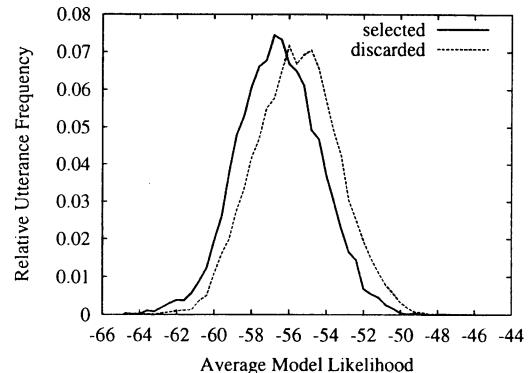


図 3 選択と破棄した発話に対する初期モデルの平均尤度（実験 B、*ST_DelScan*）

実験 A で *ST_DelScan* を用いた際の、開発データ量と選択学習性能の関係を表 4 に示す。実験 B に関してたけまるデータベースに高齢者用の開発データが殆んど存在しないため、検討が困難である。開発データ数が少くても提案手法は有効であることが分かる。選択した発話数が開発データ量とともに多少増える傾向が見られるが、著しい変動はない。

表 4 開発データ量と選択学習性能の関係（実験 A、*ST_DelScan*）

開発データ量	5	10	20	50	100	500
1 回 EM 再学習	49.1	49.3	50.4	50.4	50.5	50.4
5 回 EM 再学習	49.6	50.3	50.2	50.8	51.2	51.3
選択した発話数	9,633	8,715	9,393	9,617	10,287	10,697

4.2 選択学習過程の分析

図 3 は選択または破棄した初期音響モデルに対する平均尤度の相対的頻度を示す。より低い平均尤度を持つ発話の方が選択される傾向が見られるが、重なる部分の方が多い、「ある閾値より平均尤度が低い発話を選択する」という選択基準が提案法ほど有効ではないといえる。実験 B も同様の傾向が見られた。

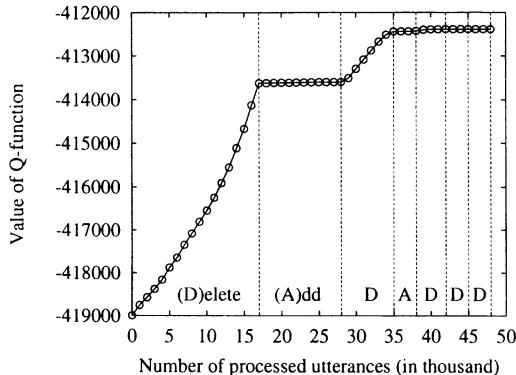


図 4 発話選択過程中的 Q 関数値の変動（実験B、*ST_DelAdd*）

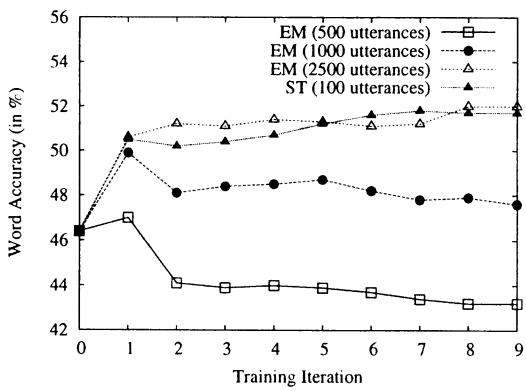


図 5 従来の EM 学習と選択学習の性能比較（実験A、*ST_DelScan*）

ST_DelAdd の場合、 Q 関数は発話選択過程中にどの程度変動するかを調査する。結果を図 4 に示す。最初と二回目の破棄処理段階の間に、 Q 関数が上昇するが、それ以後はほとんど変化しない。これは最低サンプル数閾値で拘束がかかるからである。一度破棄した発話を加えることによる Q 関数の上昇は僅かであり、加える発話数も少数に過ぎない。それぞれの段階に破棄した (-)、又は、加えた (+) 発話数は、-11,715, +203; -2,870, +44; -341, +21; -42, +1; -10, +0 であった。

4.3 大量の開発データでの従来 EM 学習

少量の開発データに基づく選択学習と同じ性能をもたらす音響モデルを従来の EM 学習で構築するのに、どの程度の学習データが必要かを調査する。結果を図 5 に示す。幼児の 500 あるいは 1,000 発話で初期モデルを再学習しても、認識性能はほとんど上昇しないが、2,500 発話を用いると、改善がみられる。選択学習では、幼児の 100 発話でも、既存の音声データを使用することで同性能の音響モデルを構築できる。

4.4 従来音響モデル適応技術との比較

選択学習と MLLR [12] と MAP [13] による音響モデル適応と比較する。評価実験では、HTK が提供する平均ベクトルの MAP 適応と、平均ベクトルと共分散行列を適応する MLLR を用いた。MLLR の場合、最も高い認識精度を得るために、回帰クラス数 2,4,8,16,32 と変え、繰り返し適応を行った。適応データ

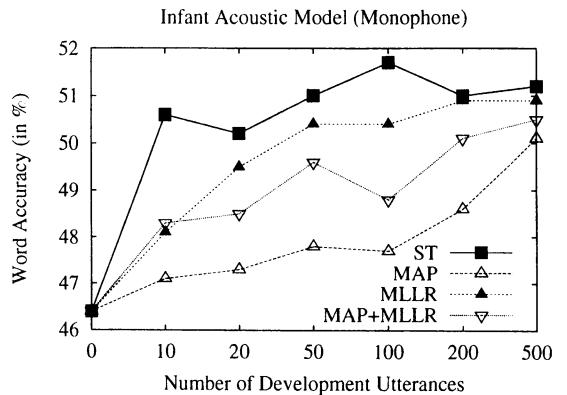


図 6 開発データに基づく MAP 適応と MLLR 適応と選択学習の性能比較（実験A、*ST_DelScan*、選択したデータで 8 回再学習）

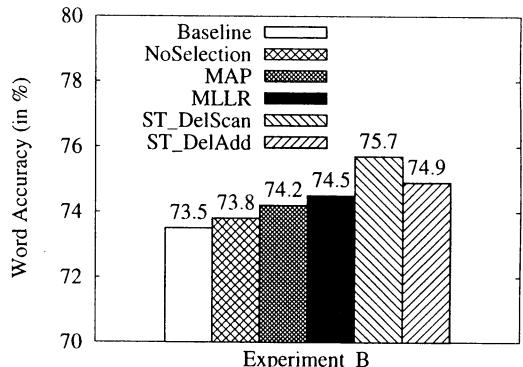
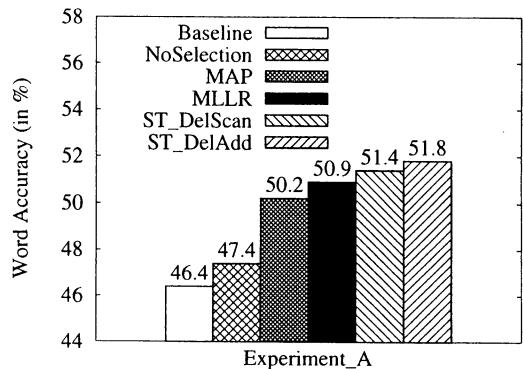


図 7 開発データに基づく MAP 適応と MLLR 適応と選択学習の性能比較（実験Aと実験B）

として、開発データを用いた。図 6 に結果を示す。選択学習の性能が最も高いことが分かる。適応データが少量であるため、MAP あるいは MAP と MLLR を同時に適用した場合の性能改善が MLLR 適応ほどではない。開発データ数 500 に至っても、選択学習の性能は MLLR の性能を上回る。

実験A（開発データ数 500）と実験B（開発データ数 53）において各手法で得られた認識精度をまとめて図 7 に示す。

5. 計算量分析

学習データと開発データの十分統計量を Q 関数計算に導入することで、再学習と尤度計算に必要な計算時間を大幅に削減した。発話選別のためにどの程度の計算量が必要かを分析する。各発話の十分統計量抽出の計算量は EM 学習と同等であるため、総計算量は発話選別処理を行うことによる時間延長分のみを表す。十分統計量の保存に必要なディスク容量は主にモデルパラメータ数に依存し、フレームレベル特徴ベクトル系列を保存する場合より大きい。しかし、各発話に全音素が出現せず、数多くの十分統計量の値がゼロになるため、十分統計量の大幅な圧縮が可能である。実験 A の場合、十分統計量用の容量を 5 分の 1 程度に抑えられた。

具体的な結果を表 5 に示す。3.2GHz の CPU を搭載したパソコンの場合、学習データ数が 3 万程度である実験 A の総実行時間は 20 分に過ぎない。また、2.5GB 程度のディスク容量は問題無く用意可能である。

表 5 全十分統計保存用ディスク容量と発話選択に必要な計算時間

実験	A、ST_DelScan	B、ST_DelAdd
総実行時間 (run time)	約 20 分	約 27 分
総計算時間 (CPU time)	約 10 分	約 18 分
計算時間 Q 関数	216 秒	366 秒
音響モデルサイズ	1,300 KB	1,300 KB
全開発データ十分統計量	368 KB	313 KB
全学習データ十分統計量	400 KB	379 KB
一発話十分統計量	78 KB	84 KB
総ディスク容量	2.5 GB	1.4 GB

6. 結論

本稿では、少量の開発データでの音響モデル構築を目的とした発話単位の選択学習法を提案した。提案法では、既存の音声データベースを利用し、開発データの尤度が上昇するように、学習発話を選択する。開発データと学習データの十分統計量に基づく Q 関数計算により、高速な発話選別処理を実現した。提案した二つのアルゴリズムのうち、独立に学習発話を評価する手法は並列処理可能であるため、更なる処理の高速化を実現できる。

実験的評価においては、小学生の音声データで幼児音響モデル、または、大人の音声データで高齢者音響モデルを提案法を用いて構築した。幼児モデルでは、開発データ数が 10 発話の場合でも認識性能改善が見られ、開発データ数を 500 まで増やしても、選択学習は MLLR と MAP 適応より優れていた。発話数 100 の場合、単語正解精度は全データでの再学習より 4.5%(相対的 9.5%)、開発データでの MLLR 適応より 1.3%(相対的 2.6%) 上昇した。高齢者モデルでも同様の結果が得られた。選択学習と適応技術を同時に適用すると、更なる認識向上に繋がると考えられる。

提案法をモノフォンモデルより実用的な音響モデル、つまり PTM [14] の構築のために評価した場合、同様の認識性能向上

が得られた。今後、開発データと学習データの収集源が異なる場合の選択学習の有効性を検討する必要がある。

謝辞 本研究の一部は文部科学省の COE と e-Society プロジェクトの研究助成金によった。

文 献

- [1] Y. Gao, L. Gu, and H.K.J. Kuo, "Portability Challenges in Developing Interactive Dialogue Systems," International Conference on Acoustics, Speech, and Signal Processing, pp.1017-1020, 2005.
- [2] T. Kemp and A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," European Conference on Speech Communication and Technology, pp.2725-2728, 1999.
- [3] F. Wessel and H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition," IEEE Workshop on Automatic Speech Recognition and Understanding, 2001.
- [4] D. Hakkani-Tür, G. Riccardi, and A. Gorin, "Active Learning for Automatic Speech Recognition," International Conference on Acoustics, Speech, and Signal Processing, pp.3904-3907, 2002.
- [5] T.M. Kamm and G.G.L. Meyer, "Robustness Aspects of Active Learning for Acoustic Modeling," Proceedings of the International Conference on Spoken Language Processing, pp.1095-1098, 2004.
- [6] F. Lefevre, J.-L. Gauvain, and L. Lamel, "Genericity and Portability for Task-dependent Speech Recognition," Computer Speech and Language, vol.19, pp.345-363, 2005.
- [7] C. Huang, T. Chen, and E. Chang, "Transformation and Combination of Hidden Markov Models for Speaker Selection Training," Proceedings of the International Conference on Spoken Language Processing, pp.1001-1004, 2004.
- [8] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, A. Lee, and K. Shikano, "Evaluation on Unsupervised Speaker Adaptation based on Sufficient HMM Statistics of Selected Speakers," European Conference on Speech Communication and Technology, pp.1219-1222, 2001.
- [9] R. Nishimura, Y. Nishihara, R. Tsurumi, A. Lee, H. Saruwatari, and K. Shikano, "Takemaru-kun: Speech-oriented Information System for Real World Research Platform," International Workshop on Language Understanding and Agents for Real World Interaction, pp.70-78, 2003.
- [10] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J.R. Statistical Society, vol.1, no.39, pp.1-38, 1977.
- [11] "Julius, an Open-Source Large Vocabulary CSR Engine - <http://julius.sourceforge.jp/>"
- [12] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," Computer Speech and Language, vol.9, pp.171-185, 1995.
- [13] J.L. Gauvain and C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," IEEE Transactions on Speech and Audio Processing, vol.2, pp.291-298, 1994.
- [14] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A New Phonetic Tied-Mixture Model for Efficient Decoding," International Conference on Acoustics, Speech, and Signal Processing, pp.1269-1272, 2000.