2008－SLP－72 (14)
2008／7／19

適応学習を用いた音響モデルの言語間適応手法の開発

酒井 優　木田 祐介　河村 聡典

東芝 研究開発センター
〒212-8582 川崎市幸区小向東芝町1
e-mail: masaru4.sakai@toshiba.co.jp

あらまし　音声認識の性能は，一般に，音響モデルを学習するための音声コーパス量に依存する．しかし，ある言語の大規模な音声コーパスを収集するためには膨大なコストが必要となり，このコストは対応する言語数に応じて増大する．この問題を解決する手法として，複数の既存言語の大規模音声コーパスを用いて学習した音響モデルを「種」モデルとし，種モデルに対して，目的言語の小規模音声コーパスを用いた適応を実施する「言語間適応 (Cross Language Adaptation)」手法が提案されている．そこで我々は，言語間適応手法のさらなる効率化を検討する．具体的には，種モデルの学習に適応学習を導入し，目的言語と既存言語との音響的特徴の差異を補正することで，従来の言語間適応手法に比べて適応後の音響モデルの性能をさらに改善することを試みる．本稿では，タイ語の音響モデル学習に提案手法を適用した実験について報告し，小規模音声コーパスに対するタイ語音響モデルの性能を提案手法によりさらに改善できることを示す．

Cross Language Adaptation using Adaptive Training on Acoustic Modeling

Masaru Sakai　Yusuke Kida　Akinori Kawamura

Toshiba R&D Center
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582, Japan

Abstract　Performance of an automatic speech recognition (ASR) generally depends on the amount of training corpus used for acoustic modeling. However, it takes so much cost to collect large amount of corpus for a new language, and the cost discourages multilingualization of ASR applications. To solve this problem, a method called as "cross language adaptation" has been proposed. On the method, a "seed" model is trained by large corpora of source languages, and post-adaptation with small amount of the target language's corpus is arranged. In this paper, we propose more effective method of cross language adaptation. We apply adaptive traininig technique on "seed" model training, to transform acoustic feature of source languages to that of target language. Experimental results on Thai acoustic modeling with proposed method show futher performance gain for any amount of training corpus.

1 はじめに

音声認識で利用される音響モデルは，ある言語における音声の様々な変動，例えば，音韻環境 (コンテキスト) や話者性の違いによる音声の多様性を十分にカバーするために，できるだけ大量の音声コーパスで学習されることが望ましい．しかし，ある言語の大規模な音声コーパスを収集するためには，膨大なコストが必要となる．音響モデル学習用の音声コーパスは，数十人から数百人規模の話者が，それぞれ数十から数百程度の語彙を発話した音声から構成されており，その収集・整備にかかる工数は非常に大きい．また，多言語の音声認識を実現する場合，このコストは対応する言語数に応じて増大する．このため，音声コーパスの収集・整備にかかるコストが，音声認識を利用したアプリケーションを多言語展開する際のボトルネックとなっている．

この問題を解決するためには，少量の音声コーパスしか利用できない言語の音響モデルを効率的に開発できる手法が必要となる．そのような手法として，言語間適応 (Cross Language Adaptation) と呼ばれる手法が提案されている [1] [2]．言語間適応手法では，複数の既存言語の大規模音声コーパスで学習した音響モデルを「種」モデルとし，種モデルに対して，目的言語の小規模音声コーパスによる適応を実施する．この手法により，小規模な音声コーパスしか利用できない目的言語に対して，既存言語の大規模音声コーパスを利用することで，目的言語の音声コーパスのみを利用する場合に比べて高精度な音響モデルを学習できる．この手法の応用例として，音声コーパスがほとんど利用できない言語の音響モデルを高速かつ低コストに試作するシステム [3] が提案されている．

そこで我々は，音声認識を利用した実アプリケーション開発における言語間適応手法の応用を目指して，言語間適応手法のさらなる効率化を検討する．具体的には，種モデルの学習において，話者適応の分野で応用されている適応学習 (Adaptive Training) を導入し，学習の目的とする言語と既存言語との音響的特徴の差異を補正することで，従来の言語間適応手法に比べて適応後の音響モデルの性能をさらに改善させることを試みる．本稿ではその検討結果を報告する．

以下，本稿の構成を説明する．まず，2章において，言語間適応手法の概要を説明するとともに，適応学習を用いた言語間適応の提案手法について説明する．次に，3章において，タイ語を目的言語とした音響モデルの学習・評価実験を実施し，学習用の音声コーパス量に対する性能変動曲線の推移を見ることで，提案手法の有効性を確認する．最後に4章でまとめとする．

2 音響モデルの言語間適応手法

2.1 言語間適応を構成する処理

本稿で検討する言語間適応手法では，音響モデルとして，コンテキスト依存音素 (triphone) の単位で学習された HMM(Hidden Markov Model) を用いる．言語間適応手法の概要を図1に示す．言語間適応手法は，以下に挙げるステップで構成される．

S1. 言語間音素マッピングの作成

S2. 制約付き決定木クラスタリングの実施

S3. 言語間適応に基づく音響モデル学習

- **S3-1.** 種言語コーパスを用いた種モデルの学習
- **S3-2.** 目的言語コーパスを用いた目的言語モデルへの適応

言語間適応手法では，音響モデルの学習を，種モデルの学習，および，目的言語モデルへの適応というサブステップの連結として実施する．以下，各ステップの処理を説明する．

S1: 言語間音素マッピングの作成

まず，目的言語と複数の種言語との間で，相互に類似する音素の対応関係 (マッピング) を作成する．本稿では，この対応関係のことを「言語間音素マッピング」と呼び，その作成手法として，IPA (International Phonetic Alphabet) [4] を介して対応関係を作成する規則型の手法を用いる [2]．すなわち，目的言語と種言語の音素を IPA に対応付けた上で，同一あるいは類似する IPA に対応する音素を言語間で相互に対応付けることで，言語間音素マッピングを作成する．

S2: 制約付き決定木クラスタリングの実施

前述の言語間音素マッピングにより相互に対応付けられた目的言語・種言語の音素の集合に対して，HMM の状態位置ごとに，音素コンテキストに関す

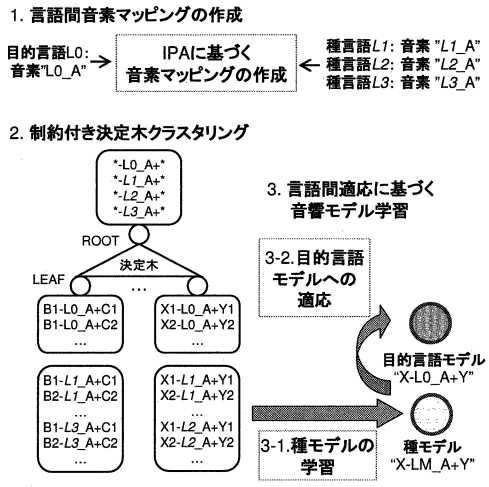


図 1: 言語間適応手法の概要

る質問を用いたコンテキスト依存音素モデルの決定木クラスタリングを行う．このとき，「目的言語のコンテキスト依存音素モデルのみを含むノードを新たに生成しない」という制約を導入する [5]．この制約により，決定木のリーフノードのうち，目的言語のコンテキスト依存音素モデルを含む任意のリーフノードは，種言語のコンテキスト依存音素モデルを含むことが保証される．この結果，目的言語と種言語との間で，相互に類似するコンテキスト依存音素が同じ HMM 状態を共有する音響モデルを取得する．

S3-1: 種言語コーパスを用いた種モデルの学習

状態共有された音響モデルに対して，まず，種言語コーパスを用いた種モデルの学習を行う．ここで種モデルとは，後述する適応の「種」となるパラメタを与える音響モデルであり，複数の種言語を混合した多言語モデルとして学習される．前述の IPA を介した言語間音素マッピング，および，決定木クラスタリングの結果から，ここで学習される種モデルは，目的言語の音響モデルを近似する音響モデルとなることが期待できる．

S3-2: 目的言語コーパスを用いた目的言語モデルへの適応

さらに，種モデルに対して，目的言語コーパスを用いた目的言語モデルへの適応を実施する．すなわち，種モデルが与えるパラメタを初期値として，目的言語コーパスを用いてパラメタを更新する．この結果，少量の目的言語コーパスしか利用できない場合でも，高い精度をもつ目的言語モデルが得られることが期待できる．適応手法には，予備実験において比較的高い性能を与えたブートストラップ法 [6] を用いる．ただし，少量の目的言語コーパスによる過適応を避けるために，更新するパラメタをガウス分布の平均ベクトルおよび分岐確率のみに制限する．

2.2 適応学習を用いた言語間適応手法

2.2.1 種モデル学習における問題点

図 1 で示した手法では，IPA を介した言語間音素マッピングを利用し，目的言語のある音素に対して，当該音素の音響的特徴を近似する種モデルを学習する．しかし，目的言語と種言語のある音素が同一の IPA に対応する場合でも，それらの音素が同一の音響的特徴を有するとは限らない．例えば，図 2 に例示するように，同一の IPA であっても，言語ごとに母音四辺形中の位置は異なる [7]．すなわち，同一の IPA に対応付けられる音素であっても，言語間で音響的特徴の差異が存在する．

言語間適応手法では，目的言語モデルに対する種モデルの近似精度が高まるほど，適応後の目的言語モデルの性能が向上すると考えられる．そこで本稿では，種モデルの学習に適応学習 (Adaptive Training) [8] を導入することで，目的言語に対する種言語の音響的特徴の差異を補正し，より近似精度の高い種モデルを学習する手法を提案する．

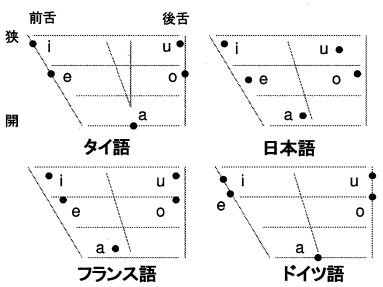


図 2: 母音四辺形中の母音の位置 (一部)

2.2.2 種モデルの適応学習

提案手法では，CMLLR(Constraint MLLR) を利用した種モデルの適応学習を実施する．まず，目的言語のコンテキスト非依存音素 (monophone) の HMM を，少量の目的言語コーパスを用いて学習する．次に，ある音素 p の HMM の平均ベクトル μ_p および共分散行列 Σ_p について，種言語 s の音声コーパスに対して最尤となる平均ベクトル $\hat{\mu}_{(s,p)}$ および共分散行列 $\hat{\Sigma}_{(s,p)}$ を与える線形変換 $(\bar{A}_{(s,p)}, \bar{b}_{(s,p)})$ を推定する．線形変換は，以下の式 (1) および (2) で実施される．

$$\hat{\mu}_{(s,p)} = \bar{A}_{(s,p)}\mu_p + \bar{b}_{(s,p)} \tag{1}$$

$$\hat{\Sigma}_{(s,p)} = \bar{A}_{(s,p)}\Sigma_p\bar{A}^T_{(s,p)} \tag{2}$$

最後に，種モデルの学習において，中心音素 p をもつ任意のコンテキスト依存音素 $x-p+y$ の HMM を学習する際に，種言語 s の音声特徴ベクトル $o_s(t)$ に対して式 (3) による逆変換を実施し，変換後の音声特徴ベクトル $\hat{o}_s(t)$ を用いて HMM を学習する．

$$\hat{o}_s(t) = A_{(s,p)}o_s(t) + b_{(s,p)} \tag{3}$$

ただし，

$$A_{(s,p)} = \bar{A}^{-1}_{(s,p)}, \quad b_{(s,p)} = -\bar{A}^{-1}_{(s,p)}\bar{b}_{(s,p)}$$

以上の処理により，目的言語に対する種言語の音響的特徴の差異を音素ごとに補正した上で，より近似精度の高い種モデルを学習することができる．

3 実験

3.1 実験条件

3.1.1 種言語と目的言語

種言語は，アメリカ英語 (US)，イギリス英語 (UK)，ドイツ語 (DE)，イタリア語 (IT)，オランダ語 (NL)，中国語普通話 (ZH) の 6 言語を用いた．目的言語はタイ語 (TH) とした．タイ語の音素を表1に示す．タイ語の 42 音素 (無音の"sil"を除く) のうち，タイ語のみに存在する 8 音素に対しては，類似する IPA に対応する種言語の音素をそれぞれ手作業で選択した上で言語間音素マッピングを作成した．

表 1: タイ語の音素

	種言語にも存在	タイ語のみ
母音	@, {, a, a:, e, e:, i, i:, o, o:, O, O:, u, u:	1, 1:, @:, {: 1a, ia, ua
子音	?, b, c, d, f, h, j, k, k_h, l m, n, N, p, p_h, r, s, t, t_h, w	c_h

3.1.2 音響モデルの学習と評価

音響モデルは，left-to-right 型の構造をもつ 3 状態/1 音素の連続混合分布 HMM とし，特徴量は MFCC_E_D_A_Z(39 次元) とした．音響モデルの状態数・混合数は，学習用の音声コーパス量に応じて最適値が変動すると考えられるため，状態数を (129,250,500,1000) 状態で可変とし，ある状態数に対して最良の性能を与える混合数を (8,16,32,64) 混合から選択した．なお，状態数 129 の音響モデルは，コンテキスト非依存音素 (monophone) を単位とする音響モデルである．

評価タスクは，人名・地名からなる 1937 単語の認識と，それぞれ 5〜50 音素からなる単語・短文発声に対する音素列認識を選択し，それぞれ単語認識率と音素認識精度で評価した．音素列認識の実施時には，挿入ペナルティを音響モデルごとに最適値に調整した．

3.1.3 音声コーパス

音響モデル学習用の音声コーパスは，音素バランス文や短文発声で構成されており，話者数と発話時間はそれぞれ 200 名〜300 名・1000 分〜5000 分である．タイ語に関して，後述する評価実験では，音声コーパスを構成する全発話からランダムに選択した発話を削除することで，異なる音声コーパス量の条件を設定した．また，評価用の音声コーパスの話者数と発話数は，単語認識が 10 名・963 発声であり，音素列認識が 198 名・242 発話である．なお，音声コーパスは，静かな室内でヘッドセットマイクで収録されたクリーンな音声であり，サンプリング周波数は 12kHz である．

3.2 ベースライン性能の評価

言語間適応手法を実施せず，目的言語であるタイ語の音声コーパスのみを用いて音響モデルを学習・評価した結果を図3に示す．図3の折れ線グラフは，状態数を (129,250,500,1000) 状態とした音響モデルの学習用音声コーパス量に対する性能変動を示しており，実線が単語認識率，破線が音素認識精度である．図3から明らかなように，音声コーパス量の減少に応じて，単語認識率・音素認識精度はともに劣化した．単語認識は，音声コーパス量が 343 分を超えると性能が収束し，状態数の異なる音響モデル間での性能差はほとんど見られなかった．その一方，音声コーパス量が 171 分を下回ると大きな性能劣化を示し，特に状態数が大きいほどその傾向が顕著であった．これは，音響モデルの状態数が大きい場合，音響モデル学習において推定すべきパラメタ数に対する音声コーパス量が相対的に減少するためだと考えられる．また，音素列認識は，単語認識に比べてゆるやかな性能変動を示したが，音声コーパス量が 171 分を下回ると，単語認識と同様に大きな性能劣化を示した．

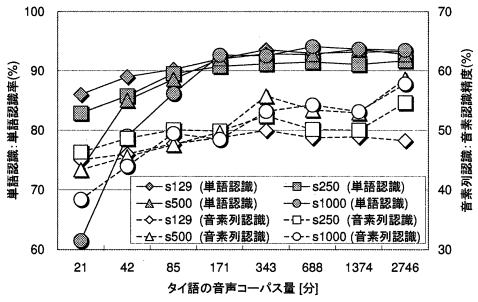


図 3: タイ語音響モデルのベースライン性能

3.3 適応学習の有無に関する言語間適応手法の比較

本稿で提案する言語間適応手法の有効性を確認するために，特に性能劣化が大きい音声コーパス量 171 分以下の領域に着目し，比較的良好な性能を示す状態数 250 の音響モデルを対象として，以下の 3 条件で学習した音響モデルの性能を比較した．

BASE: 言語間適応なし

CLA: 言語間適応を実施

CLA-AT: 適応学習を用いた言語間適応を実施

比較結果を図4に示す．図4は，図3と同様に，実線は単語認識率を，破線は音素認識精度を示す．

言語間適応なしの条件 (BASE) と言語間適応を実施した条件 (CLA) とを比較すると，言語間適応の実施により，音声コーパス量が 21 分の場合に性能が大きく改善した．ただし，単語認識率はどの音声コーパス量に対しても性能改善を示すが，音素認識精度は音声コーパス量が 42 分以上の場合にはむしろ劣化する場合もあった．このことから，言語間適応の実施により，特に音声コーパス量が極めて少ない場合に性能改善を得られるものの，その効果はやや限定的であると言える．

これに対して，適応学習を用いた言語間適応を実施した条件 (CLA-AT) は，どの音声コーパス量に対しても，比較した 3 条件の中で最良の性能を示した (一部の例外を除く)．特に，単語認識率だけでなく，音素認識精度も大きな改善を示した．これは，適応学習を用いた言語間適応により，任意の音素モデルの精度をさらに向上できたためと考えられる．以上の結果から，適応学習を用いることで，言語間適応手法の効果をさらに改善できることを確認できた．

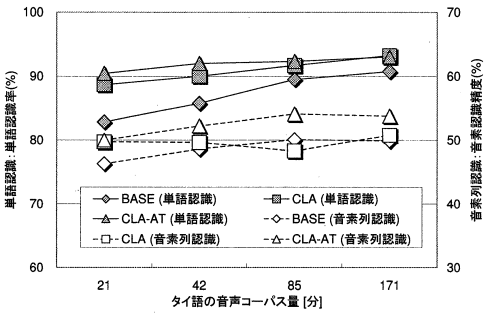
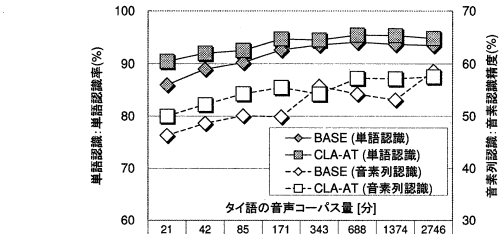


図 4: 言語間適応手法の比較

3.4 提案手法による性能改善

最後に，図3で評価した全ての音声コーパス量の条件に対して，本稿で提案した適応学習を用いた言

語間適応手法を実施した結果を図5に示す．図3から明らかなように，音響モデルの状態数は，学習用音声コーパス量に応じてその最適値が変動する．そこで図5では，ある音声コーパス量に対して最良の性能を与える状態数・混合数の組み合わせを，言語間適応なしの条件(BASE)・適応学習を用いて言語間適応を実施した条件(CLA-AT)ごとにそれぞれひとつ選択して比較した．図5が示すように，提案手法により，音声コーパス量に対する性能がほぼ全ての条件で改善されており，特に音声コーパス量が少ない場合ほどその効果は大きい．その結果，音声コーパス量に対する性能変動曲線はよりフラットな形状となっている．このことから，提案手法は，音声コーパス量が音響モデルの性能に与える影響を軽減し，音響モデル学習の信頼性・安定性を向上するために有効な手法であると言える．



	21	42	85	171	343	688	1374	2746
BASE (単語認識)	86.0	89.0	90.2	92.6	93.6	94.1	93.7	93.5
CLA-AT (単語認識)	90.5	92.0	92.5	94.7	94.5	95.4	95.3	94.8
BASE (音素列認識)	46.3	48.6	50.1	49.9	55.7	54.2	53.1	56.6
CLA-AT (音素列認識)	50.0	52.3	54.3	55.5	54.2	57.2	57.1	57.5

図 5: 適応学習を用いた言語間適応による性能改善

4 おわりに

本稿では，言語間適応手法のさらなる効率化を検討し，適応学習を用いた言語間適応手法を提案した．また，タイ語の音響モデル学習に提案手法を適用し，少量の学習用音声コーパスに対するタイ語音響モデルの性能が，適応学習を用いない言語間適応学習手法に対してさらに改善することを確認した．

本稿で実施した実験では，学習用・評価用の音声コーパスとして，静かな環境で収録されたクリーンな音声を用いた．しかし，音声認識を利用した実アプリケーションを実環境で使用するためには，実環境雑音が重畳された音声コーパスに対しても，提案手法が同様の効果を示すかどうかを確認する必要がある．また，タイ語以外の言語に対しても提案手法を適用し，どのような言語に対しても提案手法が有効であることを確認する必要もある．これらは今後の課題としてさらに検討を進めていく．

謝辞

本研究は，経済産業省「情報家電センサー・ヒューマンインターフェイスデバイス活用技術開発＜音声認識基盤技術の開発＞」プロジェクトの一環として実施されたものである．

参考文献

[1] C.Nieuwoudt, et al., "Cross-Language Use of Acoustic Information for Acoustic Speech Recognition," Speech Communication, Vol.38, pp.101-113, 2002.

[2] T.Schultz, et al., "Language-independent and Language-Adaptive Acoustice Modling for Speech Recognition," Speech Communication, Vol.35, pp.31-51, 2001.

[3] T.Schultz, et al., "SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems," Proc. of Interspeech 2007, pp.2125-2128, 2007.

[4] "http://www.arts.gla.ac.uk/ipa/index.html"

[5] F.Diehl, et al., "CONSTRAINT INDUCTION OF PHONEME-ACOUSTIC DECISION TREES FOR CROSSLINGUAL ACOUSTIC MODELING," Proc. of ICASSP 2007, Vol.4, pp.761-764, 2007.

[6] T.Schultz, et al., "FAST BOOTSTRAPPING OF LVCSR SYSTEMS WITH MULTILINGUAL PHONEME SET," Proc. of EUROSPEECH 1997, pp.371-374, 1997

[7] 国際音声学会編, "国際音声記号ガイドブック," 大修館書店, 2003

[8] T.Anastasakos, et al., "A Compact Model for Speaker-Adaptive Training," Proc. of ICSLP 1996, Vol.2, pp.1137-1140, 1996.