論文 / 著書情報
Article / Book Information

論題	音声認識における転移学習：話者適応
著者	篠田浩一
掲載誌/書名	人工知能学会誌, vol. 27, no. 4, pp. 359-364
発行日	2012, 7

特　集「知識の転移」

音声認識における転移学習：話者適応

Transfer Learning in Speech Recognition: Speaker Adaptation

篠田　浩一　東京工業大学大学院情報理工学研究科
Koichi Shinoda　School of Information Science and Technology, Tokyo Institute of Technology.
shinoda@cs.titech.ac.jp, http://www.ks.cs.titech.ac.jp

Keywords: speaker adaptation, speech recognition, transfer learning.

1. はじめに

統計的パターン認識システムを構築する際に，その対象となるドメインにおける手持ちのデータが極めて少ないことが問題になることが多い．そのような場合に，しばしば，そのドメインに類似した別のドメインでは，利用できるデータが豊富にあり，したがって高性能な認識モデルがすでに存在している，というケースがある．統計的パターン認識における適応技術とは，あるドメインAにおける認識モデルを，別のドメインBにおける少量のデータを用いて更新することにより，ドメインBにおける認識モデルを構築する技術のことである．容易にわかるように，適応と転移学習は同義である．

そして，音声認識における話者適応とは，上述の「パターン」を音声に，ドメインを「話者」に限定したときの適応であり，したがって，音声認識における，話者による違いに対する転移学習である．ここで話者による音声特徴の違いとして，音響的な違い（声の音色の違いなど）と言語的な違い（口癖など）があるが，ここでは前者に焦点を絞る．

2. 話者適応の歩み

話者適応（Speaker adaptation）は1980年代にその研究が開始され，90年代にその研究の最盛期を迎え，90年代半ばには実用化され一般に用いられている．つまり，話者適応は，機械学習の分野で転移学習の研究が始まる前に，すでに実用化されている．音声認識の対象となる音声メディアは，その他の画像や映像と比べデータ量が少ないために必要な計算資源も少なく，したがって音声認識技術自体の実用化が早かったことがその背景にある．

音声認識では，90年代に，二つの方向に大きく進歩した．まず，それまでは十数字など比較的小語彙を認識の対象としていたが，あらゆる用途に使用可能な，大語彙（2万単語以上）の認識が可能となった．また，それまでの単語発声のみを受け付ける孤立単語認識，あるいは，あらかじめ桁数の決まった数字など簡単な文法に沿った発声を受け付ける連続単語認識に対し，任意の単語の連続，すなわち一般の文発声を受け付ける，文認識が実用化された．以下，話者適応の対象となる音声認識としては，このような大語彙連続音声認識に限ることにする．

音声認識には，ある特定の話者の音声のみを受け付ける特定話者認識（speaker-dependent recognition）と，誰の声でも受け付ける不特定話者認識（speaker-independent recognition）がある．特定話者認識では，その話者の音声をあらかじめ登録する必要がある．大語彙音声認識の実用化は，特定話者認識から始まった．しかし，そこでは事前登録の手間が問題となっていた．例えば，90年代に実用化されたディクテーション（口述筆記）ソフトでは，事前登録のために1時間以上の発声が必要であった．

実用においては，事前の登録を必要としない不特定話者認識が望ましい．不特定話者認識に用いる認識モデル（不特定話者モデル）のパラメータは，多くの話者の音声を用いて最尤推定（Maximum-likelihood estimation）を用いて事前に学習される．不特定話者モデルは，いわばすべての話者の平均的な特徴をもつ話者のモデルといえる．90年代後半から，数百～数千名の規模の話者の音声データを収集し，用いることにより，その実用化が盛んになってきた．

音声は，話者によりその特徴が異なる．したがって不特定話者認識の性能は，当然，特定話者認識の性能に及ばない．また，しばしば，不特定話者モデルを用いた場合の認識率が極めて低くなる話者（特異話者と呼ぶ）が存在する．そこで不特定話者認識において，話者の違いに対する頑健性を高める必要が生じた．そこでは，使用者の少量の発声を用いて認識性能を高める技術，言い換えれば，特定話者認識において，高い認識性能を保ちつつ話者の発声の量を減らす技術が求められた．これが話者適応技術である．使用者の発声が少ないときに，それを用いて最尤推定を行うと，データ不足のため，しばしば何もしない（すなわち不特定話者認識）よりも性能が低くなってしまう．何らかの別のパラメータ推定手法が求められる．図1に，理想的な話者適応の性能曲線を示す．

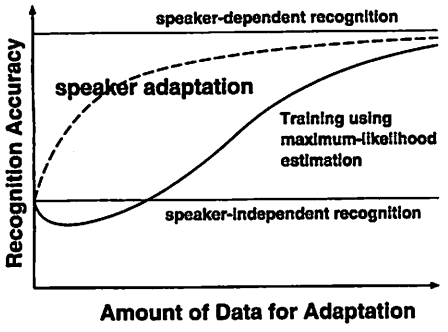

図1 話者適応に求められる性能曲線を点線で示す．
単に最尤推定（maximum-likelihood estimation）を行うと，データが少量のときは何もしないよりもかえって性能が低くなってしまう

3. 転移学習としての話者適応

転移学習は，転移元データと転移先データのおのおのにおける教師ラベルの有無により4種類に分類される[神嶌 10]．話者適応における教師ラベルとは，音声に対するtranscription（書起こし）のことであり，話者適応においても，その4種類それぞれが考え得る．しかしながら，現段階では，実用のための基本技術は，両方のデータに教師ラベルがある帰納転移学習であり，その応用として，転移元データにのみ教師ラベルがあるTransductive転移学習がある．それぞれ，「教師あり適応」，「教師なし適応」と呼ばれる．また，機械学習における多くの応用と同じく，転移先のデータが一度すべて与えられ，それを用いてモデルを更新する「バッチ適応」と，1サンプル（1文，1単語などに相当）が与えられる都度，モデルを更新する「逐次適応」がある．性能向上の観点からは教師ありバッチ適応が，使用者の利便性の観点からは教師なし逐次適応が望ましく，アプリケーションにより使い分けられている．以下では，特に断らない限り，教師ありバッチ適応について述べる．多くの場合，その技術を教師なし適応や逐次適応に応用することは容易である．

また，話者適応には，事前知識として他の多数の話者から得られた知見を用いるものが多く，その場合は話者適応はマルチタスク学習となる．マルチタスク学習の範疇に入らない話者適応は，現在では音声認識に用いられることは少ない．例えば，ある話者の特定話者モデルを別の話者に適応する技術が，音声合成のための声質変換に用いられている．

4. 話者適応の理解のための音声認識の解説

ほかの転移学習に対して話者適応がもつ特色は，その基盤となる認識モデルが生成モデルであることであろう．この生成モデルは音声に関する先験的な知識を利用して時系列パターンをモデル化したものであり，分野外の研究者・技術者にとっては若干敷居が高いと思われる．ここでは，話者適応の理解のために最低限必要と思われる知識についてごく簡単に解説する．

まず，入力された音声に対して，8kHz以上の十分高い周波数でサンプリングを行い，ディジタル信号に変換する．次に，10msごとに20～30ms程度の窓幅で信号を切り取り，それに対し短時間フーリエ変換を行う．ここで，窓幅のスケールでは音声は定常であると仮定している．そして，フーリエ変換で得られたパワースペクトルを対数に変換し，さらに逆フーリエ変換を行い，ケプストラム特徴量を得る．音声特徴量としては，このケプストラム特徴量と全帯域対数パワー，およびそれらの微分量が一般に用いられる．結果として，音声特徴量は10ms間隔で得られる数十程度の次元をもつ音響特徴ベクトルの時系列となる．

今，Xを入力特徴量（音響特徴ベクトル列），Yをラベル（音素，単語，文など）としたとき，解くべき問題は，$P(Y|X)$を最大にするYを求めることである．音声認識では，これを直接解く代わりに，ベイズの定理

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)} \tag{1}$$

を用いて，$P(X|Y)P(Y)$を最大にするYを求める問題に置き換える[*1]．

式（1）の意味は，音声知覚の運動説（Moter theory）に則って考えると理解しやすい．この説では，人間は音声の生成過程を模擬することによって音声を認識しているとする．$P(Y)$のモデルは，言語モデルと呼ばれ，脳内で，運動指令，すなわち発声しようとしているシンボルとしての「ことば」を出力する過程をモデル化している[*2]．一方，$P(X|Y)$のモデルは，音響モデルと呼ばれ，その運動指令を音声信号（シグナル）に変換する過程をモデル化している．運動指令の音声信号への変換は，人間の調音運動により実現される．この調音運動では，人間の調音器官（肺，声帯，声道，口舌など）が用いられる．大語彙連続音声認識では，多くの場合，音響モデルとして，状態空間モデルの一種である隠れマルコフモデル（Hidden Markov Model：HMM）が用いられる．HMMは，そのパラメータとして，最初の状態を規定する初期確率，状態間の遷移確率，各状態における出力確率の3種類をもつ．また，出力確率分布として連続密度分布をもつHMMは連続密度分布（Continuous density：CD）HMMと呼ばれ，その場合の出力確率分布としては混合ガウス分布が用いられることが多い．混合ガウス分布は，そのパラメータとして，それを構成する各ガウス分布の平均ベクトルと共分散行列，およびガウス分布間の混合重みをもつ．

大語彙連続音声認識では，出現可能なすべて単語・文

*1 このようなアプローチをとる理由は，音声はその特徴量Xの次元が変動する時系列パターンで，識別すべきカテゴリー数が膨大（人間が発声し得る文全体の数）であり，このような場合に，$P(Y|X)$を直接解く有効な枠組みがまだないからである．

*2 言語モデルの適応技術について本特集に解説がある[森 12]．

について別個にHMMを用意することは，データ量，学習コストの点で現実的でない．そこで，音素などのより小さい認識単位ごとにHMMを用意し，学習・認識のプロセスではそれらを連結して用いる．これらのHMMのパラメータは，Expectation-Maximization（EM）アルゴリズムのHMM版であるBaum-Welchアルゴリズムにより，その局所解が推定される．そこでは，まず，入力音声を各認識単位に対応付けて各認識単位HMMに対応する学習データの十分統計量を蓄積し（Eステップ），次に，それを用いてHMMのパラメータを更新する（Mステップ）．この手続きをある程度収束するまで繰り返し行う．

5. 代表的な話者適応技術

話者により，その調音器官の物理的性質が違うため，同じ運動指令に対する音響信号が異なる．これが音響特徴における話者の違いの原因である．話者適応は，音響信号Xと運動指令Yの対の少数のサンプルを用いて，音響モデル$P(X|Y)$を更新する技術である．前述したように音響モデルとしてはしばしばCDHMMが用いられる．

CDHMMのパラメータはすべて話者適応の対象となり得るが，最も効果的なのは，ガウス分布の平均ベクトルμの適応であることが経験的に知られている．そこで，ここでは話者適応の対象としてはガウス分布の平均に限る．そして，適応の対象とならない，初期確率分布，遷移確率分布，混合重み係数，共分散行列は，適応に用いる初期モデルのものを適応後もそのまま用いる．

平均ベクトルの適応では，EMアルゴリズムのEステップと同様の手続きで，特徴ベクトルとHMMのガウス分布との間の対応付けを行い，その後，対応付けられた特徴ベクトルとガウス分布の対の集合を用いて平均ベクトルを更新する．その場合，適応の初期モデル（適応前の平均ベクトル）としてどのようなものを用いるかが問題となる．適応に求められる性質として，推定されるパラメータは，使用者の発声が全く得られない場合は不特定話者HMMのものと一致し，発声が増えるに従い最尤推定量に近づくことが望ましい．そこで，適応の初期モデルとしては，あらかじめ用意された多数話者の音声データを用いて学習された不特定話者HMMの平均ベクトルを用いる．

音声認識では，上述したように，通常10 ms程度のフレーム間隔で数十次元の特徴ベクトルxが入力される．例えば3秒の発声は，この特徴ベクトル300個からなる時系列となる．一方，音響モデルにおけるガウス分布数は1万～10万である．当然，対応する特徴ベクトルがない分布や，対応する特徴ベクトルがごく少数の分布が圧倒的多数を占める．

話者適応は，このようにデータが極端に不足している状況で，事前知識を活用してガウス分布の平均ベクトルを頑健に推定することを目的としている．もちろん「ない袖は振れない」ので，データ不足を補うために，事前知識を活用することが前提となる．どのような事前知識を，どのように用いるかが重要である．

この観点から話者適応技術は以下の3種類に大別できる．

- ●事前分布を仮定した確率モデル学習の一般的な枠組み．事後確率最大化法，ベイズ推定法など．
- ●マルチタスク学習ではない帰納転移学習．二話者の特定話者モデルの間の写像関数を推定する．写像関数として自由パラメータ数がHMMの全平均ベクトルのパラメータ数よりも小さいものを選ぶ．最尤線形回帰法，区分線形シフト法など．
- ●マルチタスク学習．使用者以外の多数の話者の特定話者モデルがあらかじめ存在していることを前提とし，それを利用する枠組み．固有声法，話者クラスタリング法など．

以下に，上記の3カテゴリの代表的な手法，事後確率最大化手法，最尤線形回帰法，固有声法について順に解説する．

5·1 事後確率最大化適応法

事前知識を利用する方法として，パラメータの事前分布を仮定するベイズ的な枠組みを用いるのは自然な考え方であろう．この種の手法は話者適応手法として，しばしば用いられる．ここではその中でも代表的な手法として，事後確率最大化（Maximum a posteriori：MAP）推定を用いたMAP適応 [Gauvan 94] について述べる．

事後確率最大化推定では事前分布として自然共役事前分布を用いることが多い．しかし，HMMは隠れ変数が存在するために指数分布族に属さず，したがって自然共役事前分布が存在しない．したがって，解析的に解を求めることができない．そこで，ここでは，HMMを構成する，遷移確率分布，出力確率分布などのパラメータが互いに独立であると仮定する．その仮定のもと，平均ベクトルのパラメータの推定においてその自然共役事前分布を利用することが可能になる．最尤推定の場合と同様にEMアルゴリズムを用いて，事後確率を最大化するパラメータを求める．

今，HMMを構成する各ガウス分布の平均ベクトルのうち，ある一つの平均ベクトルμを推定する場合を考える．MAP適応では，使用者の適応のための発声（音響特徴ベクトル列）$\chi = x_1, \cdots, x_T$が入力されたときの事後確率を最大とする値$\hat{\mu}$を推定する．すなわち

$$\hat{\mu} = \arg\max_{\mu} p(\mu|\chi) \propto \arg\max_{\mu} p(\chi|\mu)\, p(\mu) \quad (2)$$

ここで，平均ベクトルμの事前分布$p(\mu)$として自然共役分布であるガウス分布を仮定し，その平均ベクトルとしては，不特定話者HMMにおいてμに対応する平均ベクトルμ_0を用いる．そして，EMアルゴリズムで，以

下の平均ベクトルのMAP推定値$\hat{\mu}$を繰り返し更新する．

$$\hat{\mu}=\frac{\tau\mu_0+\sum_{t=1}^{T}c_t x_t}{\tau+\sum_{t=1}^{T}c_t} \tag{3}$$

ここで，μ_0は事前分布の平均ベクトル，τは事前分布の精度（分散の逆数）に比例する制御パラメータである．c_tは時刻tの特徴ベクトルx_tがこの分布から発生する事後確率であり，時刻tにおける混合重み係数の推定値に相当する値である．そして，c_tのtに関する和$\sum_{t=1}^{T}c_t$は，その分布に対応付けられる特徴ベクトルxの数に相当する．なお，実用においては，EMアルゴリズムを繰り返してもさほど性能が向上しないため繰返しをしないことが多い．

式（3）から，事後確率最大化適応で得られる平均ベクトルの値は最尤推定値と初期値（不特定者モデルの値）を内挿したものであること，また，データが増加するについて最尤推定値に近づくことがわかる．これはデータ量が少なく最尤推定値の精度が低い場合にも頑健な推定が行われることを意味する．一方で，データが少量の場合，それに対応する平均ベクトルしか更新されず，その他の平均ベクトルは不特定話者モデルのものがそのまま使われることになる．したがって，データ量の増加に伴う認識性能の改善の度合は比較的緩やかである．

5・2 最尤線形回帰法

話者適応においては，特徴量空間において，ある話者から別の話者への写像を推定する方法が80年代から研究されてきた（図2）．転移学習の素朴なイメージに最も近い方法であろう．その中でも，最尤線形回帰（Maximum Likelihood Linear Regression：MLLR）法 [Leggetter 95] が．その取り扱いやすさと性能の高さにより広く用いられている．

最尤線形回帰法では，写像元の特定話者HMMのガウス分布の平均ベクトル$\mu=(\mu_1,\cdots,\mu_n)'$を以下の式により変換する．ここで，$n$は特徴ベクトルの次元数である．

$$\hat{\mu}=A\mu+b \tag{4}$$

ここで，Aは$n\times n$の行列，bは次元数nのベクトルであり，HMMのすべてのガウス分布で共通である．このAとbを，話者の少量の発声を用いてEMアルゴリズムにより推定する．この手法は，写像元の話者の空間で互いに近い平均ベクトルは，写像先の話者の空間においてもやはり互いに近いであろう，という類推（実証はされていない）に基づいている．

この手法は前述の事後確率最大化法に比べ，話者の発声が少ないときの性能向上が大きい．しかし，使用者の発声量が増えても，特定話者認識の性能に達することはない．これは，HMMの全平均ベクトルに対して一組の線形変換と並行移動では，それらの写像関数としては粗すぎるからである．

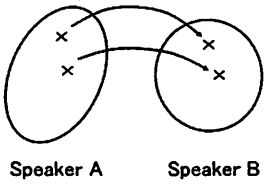

図2 話者間写像の概念図

また，この手法では，写像元の特定話者モデルとして何を用いるかも問題となる．多くの場合に不特定話者HMMが用いられるが，不特定話者HMMの各ガウス分布は，発声の話者内の変動だけでなく，話者間の変動もモデル化している．一方，話者適応の写像元のモデルとしては，話者内の変動のみをモデル化したものが望ましい．そこで，不特定話者HMMから話者間の分散を取り除いた，仮想的な正準話者（Canonical speaker）モデルを作成する，話者正規化学習というモデル学習がしばしば行われる [Anastasakos 96].

話者正規化学習では，まず，不特定話者モデルを正準話者モデルとして，おのおのの学習話者の特定話者モデルからこの正準話者モデルへの写像を作成する．次に，この写像を用いて，おのおのの話者の音声データを正準話者の音声データに変換し，そして，この変換されたデータを用いて正準話者のモデルを再学習する．このプロセスを収束するまで繰り返す．特定話者モデルから正準話者モデルへの写像としては，最尤線形回帰法の写像の逆写像を推定して利用する．この正準話者モデルを初期モデルとして用いることで，最尤線形回帰法による話者適応の性能が若干向上する．

5・3 固有声法

上記二つの手法は，多数話者の発声データから話者間の違いを取り去った平均的な話者のモデル（不特定話者モデル）を作成し，それを適応の初期モデルとして利用している．それに対し，90年代後半頃から多数話者のデータベースの構築が比較的容易になるに伴い，そこにおける話者間の違いを積極的に利用する手法が登場してきた．

その種の手法で最も単純なのは，多数話者の中から最も音響的に使用者と類似した話者を選択し，その話者の特定話者モデルを用いる手法であろう．しかしながら，高性能な特定話者モデルをつくるのに十分な量の発声データを多くの話者から得る必要があり，その実現は現段階では難しい．

そこで，これまでにクラスタリングや多変量解析を用いた手法がいくつか提案されている．ここでは，その代表的なものとして，固有声法（Eigenvoice）[Kuhn 00] を取り上げる．固有声の名前はもともと顔画像認識で用いられている主成分分析手法，固有顔（Eigenface）に由来する．

今，多数話者データベース中の話者数をS，各話者のHMMのすべての認識単位のすべての状態にわたる総分布数をMとし，入力特徴ベクトルの次元数をKとする．学習段階では，まず，多数話者データベースのおのおのの話者について，特定話者HMMを作成する．次に，お

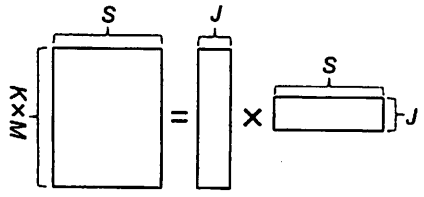

図3　多数話者の話者ベクトルに対する主成分分析．
Kは特徴ベクトルの次元数，MはHMMの総分布数，
Sは話者数，Jは固有話者ベクトルの数である

のおのの話者について，特定話者HMMのおのおのの分布の平均ベクトルを一列に連結して，$K\times M$の次元をもつ話者ベクトルを作成する．この手続きを多数話者データベース中のすべてのS人の話者に対し行い，S個の話者ベクトルを作成する．

次に，この話者ベクトルの集合に対し，主成分分析を行い，固有値の大きいほうから順に互いに直交する固有話者ベクトル$e(j)$ $(j=1,\cdots,J)$を選択する（図3）．ここで，Jは固有話者ベクトルの個数であり，総話者数Sに比べて極めて小さく，10から50前後である．固有話者ベクトルは話者ベクトルと同じ次元$K\times M$をもつ．このJ個の固有話者ベクトル群$\{e_j\}$を固有声(Eigenvoice)と呼ぶ．

適応時には，使用者の話者ベクトルは，この固有話者ベクトル群の張るJ次元の部分空間における一点として表されると仮定し，使用者の少数の発声を用いて，各軸への射影の長さ（重み係数）を求める．すなわち，使用者の話者ベクトルを以下のように固有話者ベクトルの重み付け和で近似する．

$$\hat{\mu}=\sum_{j=1}^{J}w(j)e(j) \qquad (5)$$

重み$w(j)$ $(j=1,\cdots,J)$は，使用者の少数の発声を用いて推定される．データ量が少なく，話者ベクトルの多くの要素（個々の平均ベクトル）には対応するデータサンプルがないので，この重みの推定には，上記の二つの手法と同様に，EMアルゴリズムを用いる．すなわち，入力音声とHMMとの対応付けを行って重みwの推定に必要な統計量を求め，それを用いて重みを更新してそれをHMMに反映させる，という処理を繰り返す．最後に，推定された話者ベクトルを分解して得られる平均ベクトルをその話者のHMMの平均ベクトルとする．

この手法は，少量の発声での性能向上が認められるが，発声数が多くなると，性能が頭打ちになる傾向がある．また，おのおのの特定話者HMMにおける平均ベクトルの数が多く，そのまま連結した話者ベクトルは極めて高次元になるので，そのまま主成分分析を行うのは難しい．そこで，しばしば，重み推定には，認識単位数や混合ガウス数を小さくした小さいサイズのHMMが用いられる．

6. 話者適応における課題

前章で代表的な話者適応手法について紹介した．話者適応は，ごく少量しか使用者の発声が得られない場合でも音声認識性能が向上し，また，十分な量の発声が得られる場合には特定話者モデルと同等の性能を示すことが望ましい．しかしながら，これまで述べて来た手法はこれらを同時に満たさない．

例えば，事後確率最大化法では，異なるガウス分布の平均ベクトルは互いに独立であり，おのおのの平均ベクトルはそれに対応する発声データが得られたときにのみ適応される．したがって，データ量が著しく少ない場合には改善はあまり大きくない．また，最尤線形回帰法，固有声法では，発声量がある程度多くなると，それ以上発声量が増えても認識率が向上しなくなる．これは，これらの方法で推定する自由パラメータの数が比較的少なく，データに内在する特徴を十分に活用することができないためである．それぞれの手法のおおまかな適用可能範囲を図4に示す．

もし，使用者のデータ量の増加に合わせて，適応手法を切り換えることができれば，理想的な話者適応手法になると期待できる．しかし，切換えのタイミングを発声語彙や話者の違いに対して頑健に設定することは難しい．使用者の発声量に依存しない，シームレスな適応手法が望まれる．

この目的を達成するための工夫として，まず，上述した3種類のカテゴリの手法を組み合わせることは容易に思いつくであろう．実際，さまざまな組合せ法が提案され，その効果が報告されている．代表的なものとしては，最尤線形回帰法における写像を最尤推定ではなく事後確率最大化推定で推定する方法，固有声法における話者ベクトルとして，最尤線形回帰法の変換行列をベクトル化したものを用いる方法などがある．

図4にあるように適応手法によってその適用範囲が異なるのは，それぞれの適応手法で推定する（実効的な）自由パラメータの数がある値に固定されており，またその値が手法間で異なっているためである．発声量に応じて自由パラメータ数を変化させる仕組みがあれば，シームレスな手法が容易に実現できる．このような手法の一つに構造的事後確率最大化（Structural MAP：SMAP）法[Shinoda 01]がある．この手法では，図5に示すような，HMMのすべての分布をリーフノードとしてもつ音響木構造を構築し，発声量が少ないときは主に上方のノード

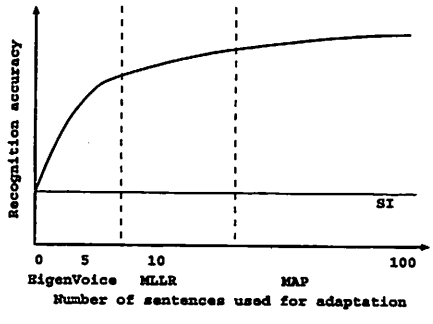

図4　話者の発声量と話者適応の性能の関係

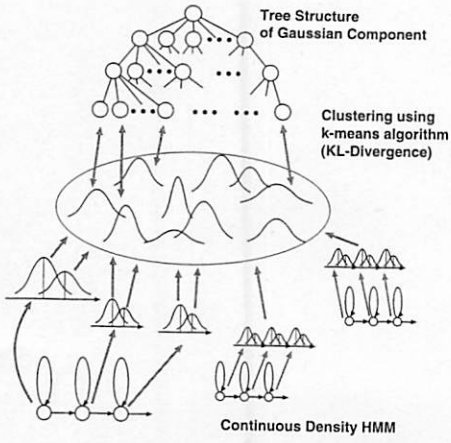

図5　構造的事後確率最大化手法で用いられるHMMのガウス分布の木構造

の分布から得られる情報を適応に用い，多くなるに従い，下層のノードをより重視することで，シームレスな適応を実現している．

7. おわりに

転移学習の一つとしての話者適応技術の概要とその代表的な手法を解説した．ここで述べた以外にも，例えば平均ベクトル以外のほかのHMMパラメータに対する適応法，パラメータの識別学習を用いた適応法など，多くの話者適応手法がある．それらについて，詳しく知りたい方はサーベイ論文 [篠田 04] を参照されたい．また，話者適応を深く理解するためにはHMMを用いた音声認識をまず十分に理解する必要がある．機械学習についてある程度知識があり，音声認識についても勉強したいという方には，Bishopの教科書 [Bishop 06] が最適である．

話者適応とは，音声認識における音響モデルの自由パラメータ数を制限し，それを少ない発声で推定する技術である．今まで述べてきたように，制限方法として，さまざまなヒューリスティックが用いられているが，現段階では残念ながらそれらの理論的根拠は明らかではない．例えば，話者間の写像が一般には線形でないことは明らかである．この問題に対しては，話者によりどのように特徴が異なるのか，また，特異話者にはどのような特徴があるのか，といった現象解析をより掘り下げる必要がある．近年，核磁気共鳴装置（MRI）の進歩により，音声の調音運動を実時間で観測することが可能になってきている．そこで得られるデータを解析することで，話者適応に役立つ知見が得られることが期待される．

本稿では，話者の違いに対する適応技術以外の適応技術には触れなかった．容易に想像ができるように，ここで説明した技術は，話者の違いのみならず，例えば，周囲の雑音環境の違い，発声スタイルの違い，など，様々な変動要因に対しても適用が可能であり，また実際に適用されて，その効果が確認されている．さらに，我々は，最近，この技術を，映像認識，ジェスチャー認識，歩容認証，など音声以外の他のメディアにおけるパターン認識の問題にも応用している．特に，適応技術を用いた映像意味インデクシングの手法 [井上 11] は，昨年開催された映像検索・評価のための国際競争型ワークショップTRECVID2011の同タスク部門において，世界トップの性能を示した．そこにおける適応技術の貢献は大きい．今後もさまざまな応用が期待される．

◇ 参考文献 ◇

[Anastasakos 96] Anastasakos, T., McDonough, J., Schwartz, R. and Makhoul, J.: A compact model for speaker-adaptive training, *Proc. ICSLP96*, Vol. 2, FrP2L1.3 (1996)

[Bishop 06] Bishop, C. M.: Pattern Recognition and Machine Learning, Chapter 13, Springer (2006) (邦訳：C. M. ビショップ著，元田 浩，栗田多喜夫，樋口知之，松本裕治，村田 昇 監訳：パターン認識と機械学習下 —ベイズ理論による統計的予測，第13章，シュプリンガー・ジャパン (2008))

[Gauvan 94] Gauvain, J.-L. and Lee, C.-H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298 (April 1994)

[井上 11] 井上中順，篠田浩一：[特別講演] 映像の高性能なセマンティックインデクシングを目指して，信学技報，Vol. 111, No. 353, PRMU2011-140, pp. 89-94 (2011)

[神嶌 10] 神嶌敏弘：転移学習，人工知能学会誌，Vol. 25, No. 4, pp. 572-580 (2010)

[Kuhn00] Kuhn, R., Janqua, J.-C., Nguyen, P. and Niedzielski, N.: Rapid speaker adaptation in Eigenvoice space robust speech recognition, *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 6, pp. 695-707 (2000)

[Leggetter 95] Leggetter, C. J. and Woodland, P. C.: Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models, *Computer Speech and Language*, Vol. 9, pp. 171-185 (1995)

[森 12] 森 信介：自然言語処理における分野適応，人工知能学会誌，Vol. 27, No. 4, pp. 365-372 (2012)

[Shinoda 01] Shinoda, K. and Lee, C.-H.: A structural Bayes approach to speaker adaptation, *IEEE Trans. Speech and Audio Processing*, Vol. 9, No. 3, pp. 276-287 (2001)

[篠田 04] 篠田浩一：確率モデルによる音声認識のための話者適応化技術（サーベイ論文），信学論（D），Vol. J87-D-II, No. 2, pp. 371-386 (Feb. 2004)

2012年4月26日　受理

著者紹介

篠田　浩一（正会員）

1987年東京大学理学部物理学科卒業．1989年同大学院理学系研究科修士課程修了．2001年東京工業大学より博士号取得．1989年日本電気株式会社入社．以後，音声・動画像パターン認識，ヒューマンインタフェースの研究に従事．その間，1997年から1998年にかけて，米国ルーセントテクノロジ・ベル研究所客員研究員．2001年から東京大学大学院情報理工学系研究科助教授，2003年から東京工業大学大学院情報理工学研究科助教授，国立統計数理研究所助教授．現在，東京工業大学大学院情報理工学研究科准教授．1997年日本音響学会粟屋学術奨励賞，1998年電子情報通信学会論文賞．電子情報通信学会，IEEEシニア会員．日本音響学会，ACM各会員．