2007－SLP－67 (15)
2007／7／21

効率的なクロスバリデーション尤度評価に基づく混合ガウス分布の最適化

篠崎 隆宏　河原 達也

京都大学学術情報メディアセンター
〒606–8501 京都市左京区吉田二本松町

あらまし　従来の自己尤度に代えてクロスバリデーション尤度を用いる新しい混合分布最適化アルゴリズムの提案を行い、HMM を用いた音声認識への応用を行う。混合分布の最適化の目的は、過剰な要素を削減することでモデルの一般性を高めることであり、最適化は混合分布要素の対を尤度に従い順次選択・併合することで行う。クロスバリデーション尤度はモデルパラメタのオーバーフィッティングを避ける上で従来の尤度よりも有効であり、また十分統計量を活用することで高速に評価することができる。これにより、従来よりも優れた分布要素対選択を行うことができるとともに、経験的な閾値に頼らない併合停止基準が与えられる利点がある。日本語話し言葉コーパスを用いた大語彙連続音声認識をタスクとし、HMM の学習に対して本手法を適用した実験結果において、本手法が従来手法よりも高い認識率を与えることを示す。

Gaussian Mixture Optimization based on Efficient Cross-validation

Takahiro SHINOZAKI　Tatsuya KAWAHARA

Academic Center for Computing and Media Studies
Kyoto University
Yoshida Nihonmatsu-cho, Sakyo-ku, Kyoto, 606–8501 Japan

Abstract　A Gaussian mixture optimization method is explored using cross-validation likelihood as an objective function instead of the conventional training set likelihood. The optimization is based on reducing the number of mixture components by selecting and merging a pair of Gaussians step by step based on the objective function so as to remove redundant components and improve the generality of the model. Cross-validation likelihood is more appropriate for avoiding over-fitting than the conventional likelihood and can be efficiently computed using sufficient statistics. It results in a better Gaussian pair selection and provides a termination criterion that does not rely on empirical thresholds. Large-vocabulary speech recognition experiments on oral presentations show that the cross-validation method gives a smaller word error rate with an automatically determined model size than a baseline using the conventional training procedure.

1 はじめに

混合ガウス分布や混合ガウス分布 HMM の学習において適切に混合要素数を決定し、また適切に要素分布を配置することは、高い性能のモデルを得る上で重要である。これまでに MDL などの情報量基準を用いた混合数最適化手法が提案されているが [1]、理論的なペナルティ項において実際には経験的な補正項を必要とする問題が指摘されている。さらに、混合ガウス分布や HMM のような隠れ変数を含んだモデルの学習では有限のパラメタ数においても特定の学習サンプルに過度に依存して尤度が無限大に発散する不安定性の問題が存在するが、そのような状況では情報量基準は本質的に意味のあるスコアを与えることができない欠点がある。これは、尤度とパラメタ数に応じたペナルティ項の和として定式化される情報量基準は尤度とともに発散してしまうためである。不安定性を回避するための対処法としては分散フロアリングが挙げられるが、閾値をどう設定するかが問題となる [2]。

これらの問題は、従来の自己尤度にかえてクロスバリデーション尤度 (CV 尤度) を用いることで解決することができる。クロスバリデーション (CV) はデータ駆動のモデル選択手法であり、理論的な仮定や近似を用いる情報量基準と比べて頑健な動作が期待できる。また、モデルの推定と尤度の評価でデータが分離されるため、特定の学習サンプルに依存した尤度の発散は起こり得なくなる。しかしながら、従来、CV を学習プロセス内部に組み込もうとすると計算量が非現実的に増大してしまう問題があり、CV の応用は学習プロセス外側での少数のモデルの比較に留まっていた。これは、大規模なモデルの学習を効率的に行うために従来用いられてきた、尤度を効率的に評価するための工夫がそのままでは使えなくなってしまうためである。

本研究では、十分統計量を活用することで混合ガウス分布に対する CV 尤度が効率的に評価可能であることを示し、さらに提案アルゴリズムを応用した混合ガウス分布最適化手法の提案を行う。これまでに著者等は CV 尤度を用いた HMM 決定木状態クラスタリング [3] (CV-DTC) 手法の提案を行っているが、本提案手法は CV-DTC で用いたガウス分布に対する CV 尤度評価アルゴリズムの、混合ガウス分布への応用とみなすことができる。また、本手法は異るデータセットを用いたモデルの推定と尤度の評価を十分統計量を用いて行うという点で、文献 [4] や [5] に見られる尤度評価手法の応用と見ることができる。

実験では、提案手法を音声認識において用いられる混合ガウス分布 HMM の学習へ応用する。『日本語話し言葉コーパス』(CSJ) [6] を用いた音声認識実験において、本手法とガウス分布分割法を組み合わせることで、モデルサイズを自動決定しつつ局所最適解問題を改善し、従来よりも高い認識性能が得られることを示す。

2 混合分布併合アルゴリズム

本節ではまず従来の尤度を用いた混合分布の最適化について簡単に述べた後、提案手法である CV 尤度を用いた最適化手法を示す。混合分布の最適化は、要素分布の対を尤度を基準として順次選択・併合することにより行う。CV 尤度との区別のため、従来の学習セットに対する尤度を自己尤度 (Self-test likelihood) と呼ぶ。ここではボトムアップクラスタリング手法について示すが、十分統計量を用いた効率的な CV 尤度評価はトップダウンクラスタリングにおいても同様に適用可能である。

2.1 自己尤度による要素併合

混合ガウス分布の要素対を自己尤度を目的関数として繰り返し選択・併合することにより、混合分布の最適化を図る手法である。入力となる M 混合の混合ガウス分布のパラメタ集合を θ、そのどれか一つの要素対を併合してできる $M-1$ 混合の分布のパラメタ集合を $\bar{\theta}$ とする。対角共分散行列を持つガウス分布の十分統計量は、観測カウントおよび観測事後確率により重み付けされた一次および二次の特徴量の和である。したがって、混合ガウス分布の m 番目の要素分布の十分統計量は以下のように表すことができる。

$$A^0(m) = \sum_{t\in T} \gamma_m(t), \quad (1)$$

$$\boldsymbol{A}^1(m) = \sum_{t\in T} \boldsymbol{x_t}\gamma_m(t), \quad (2)$$

$$\boldsymbol{A}^2(m) = \sum_{t\in T} \boldsymbol{x_t}^2\gamma_m(t). \quad (3)$$

ここで、T は学習セット、t は特徴量フレームのインデックスとしての時刻、m は混合分布要素のインデックス、$\boldsymbol{x_t} = (x_1(t), x_2(t), \cdots, x_d(t))^T$ は時刻 t における d 次元特徴量ベクトルであり $\boldsymbol{x}^2 =$

$\left(x_1^2, x_2^2, \cdots, x_d^2\right)^T$ はその各要素を2乗したベクトル、$\gamma_m(t) = P(m(t) = m|T, \theta_0)$ は初期モデル θ_0 が与えられた下での時刻 t における m 番目の混合要素の事後確率すなわち観測カウントである。混合ガウス分布の m 番目の要素の平均 $\boldsymbol{\mu}(m)$ と分散 $\boldsymbol{v}(m)$ は、これらの十分統計量を用いて以下のように求めることができる。

$$\boldsymbol{\mu}(m) = \frac{\boldsymbol{A}^1(m)}{A^0(m)}, \tag{4}$$

$$\boldsymbol{v}(m) = \frac{\boldsymbol{A}^2(m)}{A^0(m)} - \boldsymbol{\mu}(m)^2. \tag{5}$$

最適化プロセスの間アライメントが固定されていると仮定すると、混合ガウス分布 $\bar{\theta}$ の十分統計量は θ の十分統計量から容易に求めることができる。すなわち、要素対併合に関係しない要素の十分統計量はそのまま $\bar{\theta}$ に引き継がれ、併合により作成された混合要素の十分統計量は、併合前の要素対の十分統計量の和として与えられる。さらに、HMM の決定木状態クラスタリングと同様にモデルの学習セットに対する尤度を事後確率で重み付された対数尤度として定式化すると [7]、混合ガウス分布の自己尤度は以下のように表すことができる。

$$L_{self}(\theta) \approx \sum_{m=1}^{M} \sum_{t \in T} \{\log P(x_t|m, \theta)\} \gamma_m(t) \tag{6}$$

$$= -\frac{1}{2} \sum_m \left\{ \left(\log\left((2\pi)^d |\boldsymbol{\Sigma}(m)| \right) + d \right) \cdot A^0(m) \right\}. \tag{7}$$

ここで、$\boldsymbol{\Sigma}(m)$ は対角成分が $\boldsymbol{v}(m)$ となる対角共分散行列である。混合分布重みはアライメント固定の仮定の下では混合分布最適化と独立であり、省略してある。式 (7) は学習セット全体へのアクセスを意味する時刻 t に関する全ての和が十分統計量の中に押し込まれた形となっている。このため最適化の対象となる混合ガウス分布の十分統計量を事前に一度だけ計算しておくことで、任意の混合要素対について併合前後の自己尤度を高速に評価することが可能である。

混合数 M の混合ガウス分布 θ においてその要素対の可能な組み合わせ数は、M 個の要素から重複を避けて2つの要素を選択する組合せであるから $\frac{M(M-1)}{2}$ である。混合ガウス分布要素の併合最適化はこの組合せの数だけ存在する、混合数が元の分布よりも1だけ小さい混合ガウス分布モデルの集合 $\bar{\Theta}$ から最適なモデルを選択するモデル選択問題として以下のように定式化できる。

$$\hat{\theta} = \underset{\bar{\theta} \in \bar{\Theta}}{\operatorname{argmax}} L_{self}(\bar{\theta}). \tag{8}$$

このプロセスを繰り返すことで、混合ガウス分布の要素数を順に1つずつ削減することができる。また、混合ガウス分布を状態毎の観測確率分布とする HMM の場合は、アライメント固定の仮定の元に、各状態ごとに独立に最適化を適用することができる。

自己尤度を用いた混合分布の最適化においては尤度が学習セットに対して計算されるため正のバイアスを含み、データサンプル数が少ないときにはとくに不正確となる。このため、モデルの汎化性の点で最適ではない要素対が選択されることになる。さらにこのバイアスが原因で自己尤度が要素数の削減に対して単調減少するため、いつ要素併合を停止すべきか停止基準が与えられない問題がある。

2.2 クロスバリデーション (CV) 尤度による要素併合

提案法である K-fold の CV を用いた混合要素最適化では、まず学習セットを K 個のサブセットに分割する。分割は各サブセットが統計的にできるだけ均一となるようにすることが望ましい。

$$T = \bigcup_{k=1}^{K} T_k, \qquad T_i \bigcap T_j = \phi \quad (i \neq j). \tag{9}$$

そして、各サブセットごとに十分統計量を計算する。ここでは、k 番目のサブセットに対して計算された十分統計量を $\boldsymbol{A}_k = \left\{A_k^0, \boldsymbol{A}_k^1, \boldsymbol{A}_k^2\right\}$ と表すことにする。これら統計量は各サンプルごとに計算された事後確率に基づく重み和であるので、学習セット全体を用いるモデル θ の十分統計量はこれらサブセットごとの統計量の和 $\sum_{i=1}^{K} \boldsymbol{A}_i$ として容易に求めることができる。同様に、k 番目の CV モデル θ_k の十分統計量は k 番目のサブセットを除いた和 $\sum_{i \neq k}^{K} \boldsymbol{A}_i$ として求めることができる。

これらの十分統計量を用いることで、θ の CV 尤度は従来の自己尤度法と同じ仮定の下、以下のように表すことができる。

$$L_{cv}(\theta) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{t \in T_k} \{\log P(x_t|m, \theta_k)\} \gamma_m(t). \tag{10}$$

この式では、k 番目の CV モデル θ_k を k 番目のデータサブセット T_k に対する尤度を計算するために使

用しており、これによりモデルの推定と尤度の評価におけるデータの重複をなくしている。確率分布関数に混合ガウス分布を代入し、時刻 t に関する和をできるだけ式内部 (右側) に寄せるように変形することで、式 (10) は式 (12) のように表すことができる。式 (12) は従来の自己尤度の評価式 (7) と同様に、時刻 t に関する和が全て十分統計量の中に押し込まれているため、高速な評価が可能である。これによりCV 尤度を現実的な計算量で混合分布最適化に応用することが可能となる。このことが本研究の主要な貢献である。計算量のオーダーは K-fold CV に対して $O(K)$ となる。

CV 尤度を用いた混合分布の最適化は、式 (12) を従来の自己尤度に代えて目的関数として用いることで得られる。言い替えれば、CV 尤度の評価式 (12) は自己尤度の評価式 (7) の CV に対する拡張と見ることができる。実際、式 (12) において CV のインデックス k を無視すると、式を簡略化することができ、式 (7) に一致する。

CV 手法では、学習データをモデルパラメタの推定とデータ尤度の評価のために分割するが、モデル推定における実効学習データ量の減少は、CV のサブセット数 K を大きくとることで無視できる程度に小さくすることができる。すなわち、K-fold の CV において各モデルは学習セット全体の $\frac{K-1}{K}$ を用いて学習されるため、例えば K を 10 とすると各モデルは学習セット全体の 90%を用いて学習されることになる。

CV 尤度ではモデルパラメタ推定と学習データの尤度評価においてデータの重複が存在しないため、データの重複に由来する尤度バイアスが存在せず、モデル性能の公平な評価の点で従来の自己尤度よりも優れている。またその結果 CV 尤度は、あたかもテストセットに対する尤度であるかのように、混合要素数の削減に対して単調減少とならずに最適点を持ち、経験によらない併合停止基準が得られる利点がある。

2.3 混合要素削減の予備実験

図 1 に従来の自己尤度および提案手法による CV 尤度を用いた混合分布最適化を行った際の、学習データにおける尤度変化を示す。最適化の対象としたのは、HMM の一状態として学習された 256 混合の混合ガウス分布である。CV 尤度は 40-fold の CV により計算した。図の横軸が混合要素数であり、併合最適化の進行とともに 256 から減少する方向に変化する。縦軸は学習セットにおける自己尤度および CV 尤度である。学習セットに対する尤度であることから自己尤度は正のバイアスを含み、他方 CV 尤度はそのようなバイアスを含まないため、自己尤度の方が CV 尤度よりも大きな値となっている。また自己尤度は混合数の減少に対して単調減少するのに対して、CV 尤度は最適点を持つことがわかる。CV 尤度の増加は過剰なモデルパラメタの削減によるモデル汎化性能の向上を示し、また低下はパラメタ数の行き過ぎた削減によるモデル精度の低下を示している。したがって CV 尤度を用いる場合は、尤度が最大値をとる点として容易に最適停止点を知ることができる。この例の場合では、要素数 210 が最適停止点である。

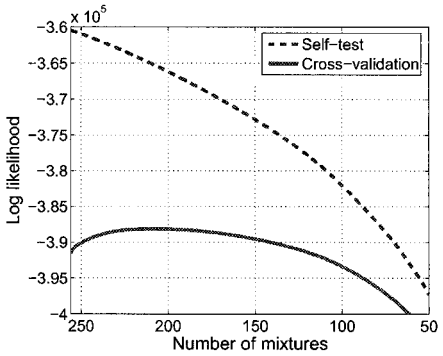

図 1: 混合要素併合と学習セットにおける尤度

3 実験条件

混合分布の最適化を HMM の学習に適用するにあたっては、様々な方法が考えられる。1 つには、まず十分大きな混合数をもつモデルを学習し、ついで混合分布の最適化を適用することである。しかしこの方法では、“十分な大きさの混合数” をどのように決めればよいかが問題となる。もう 1 つの方法は、少ない混合数のモデルから開始し、混合要素の分割と併合を繰り返すことである。この方法では、混合分布最適化の初期モデル混合数を事前に決定しなければならない問題を避けることができる。さらに、要素分布の分割と併合を繰り返し、混合分布を “こね回す” ことから、局所最適解により頑健な学習を行えると期待される。そこでここでは、後者の学習方式を用いることにする。

実験には、音響モデルの学習セットとして『日本

$$L_{cv}\left(\bar{\theta}\right)=\sum_{k}^{K}\sum_{t\in T_k}\sum_{m}^{M}\log\left\{\frac{1}{\sqrt{(2\pi)^{d}\left|\boldsymbol{\Sigma}_k\left(m\right)\right|}}\exp\left(-\frac{1}{2}\left(\boldsymbol{x}_t-\boldsymbol{\mu}_k\left(m\right)\right)^{T}\boldsymbol{\Sigma}_k\left(m\right)^{-1}\left(\boldsymbol{x}_t-\boldsymbol{\mu}_k\left(m\right)\right)\right)\right\}\gamma_m\left(t\right) \tag{11}$$

$$=-\frac{1}{2}\sum_{k}^{K}\sum_{m}^{M}\left\{\log\left((2\pi)^{d}\left|\boldsymbol{\Sigma}_k\left(m\right)\right|\right)A_k^0\left(m\right)+\left(\boldsymbol{v}_k\left(m\right)^{-1}\right)^{T}\boldsymbol{A}_k^2\left(m\right)-2\left(\boldsymbol{\Sigma}_k\left(m\right)^{-1}\boldsymbol{\mu}_k\left(m\right)\right)^{T}\boldsymbol{A}_k^1\left(m\right)+\left(\boldsymbol{v}_k\left(m\right)^{-1}\right)^{T}\boldsymbol{\mu}_k\left(m\right)^{2}A_k^0\left(m\right)\right\}. \tag{12}$$

語話し言葉コーパス』(CSJ) 学会講演から無作為に抽出した 30 時間の音声を用いた。1000 状態に状態共有された 1 混合トライフォンモデルを初期モデルとして、混合ガウス分布 HMM の学習を以下の手順により行った。出力される HMM は、各状態ごとに最適化された混合数をもつ。

1. 1 混合 HMM を入力
2. 学習セットを発話を単位としてランダムに区分化。並列 EM を 5 回繰り返す
3. 並列 EM で得られる各区画ごとの十分統計量を入力として、CV 尤度基準混合分布併合法により混合数を最適化。繰り返しが一定回数に達していたら HMM を出力して終了
4. 要素分布を分割して混合数を 2 倍に増やし、ステップ 2 へ

この学習プロセスでは、繰り返し初期には併合される要素分布が少ないため混合数は指数関数的に増加するが、混合数が増えるに従い併合される要素分布が増加するため、モデルサイズは次第に適当な値へと収束していく。以下では EM 学習 5 回と要素分割・併合を学習の 1 ループとして数える。

HMM の EM 学習には HTK [8]、認識エンジンには Julius [9] を用いた。言語モデルは CSJ 学会講演と模擬講演約 6.8M 形態素から学習されたトライグラムモデルである。テストセットは CSJ 学会講演評価セット 10 講演である。

4 実験結果

図 2 に EM 学習および 40-fold CV を用いた混合分布最適化にかかる計算コストを示す。40-fold と比較的大きな K の値を使用しているにも関わらず、混合分布最適化にかかる計算コストは EM を含めた全計算コストのおよそ 7%ほどであり、提案した CV 尤度評価アルゴリズムが実用上十分高速であることがわかる。

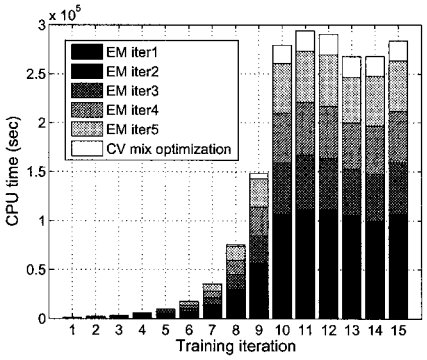

図 2: *EM* 学習および混合分布最適化の計算コスト

図 3 に要素分布併合を行わず単純に混合数を増加させたベースラインと、CV 尤度基準により要素併合を行った場合の混合数の推移を示す。ベースラインでは混合数が指数関数的に増大するのに対して、CV 尤度基準による要素併合を組み合わせることで、混合数が次第に一定値に収束していく様子がわかる。

図 4 にベースラインと CV 尤度基準により要素併合を行った場合、および MDL 基準により同様に要素併合を行った場合の認識率を示す。ベースラインよりも MDL 基準法の方が、MDL 基準法よりも CV 尤度法の方が、高い認識率が得られた。これはモデルサイズの最適化とともに混合分布の分割と適切な併合を繰り返す効果により、よりよい局所最適解が得られたためと考えられる。また、併合最適化前後のモデルを観察すると、MDL 基準では少数の学習サンプルのみに依存して高い自己尤度を得ている不

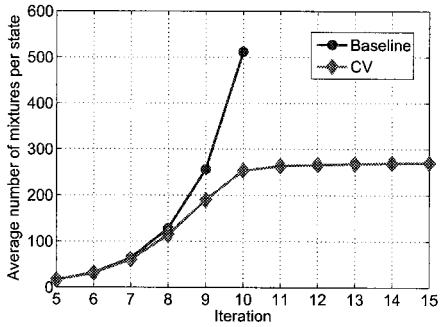

図 3: 学習ループ数と HMM 状態の平均混合数

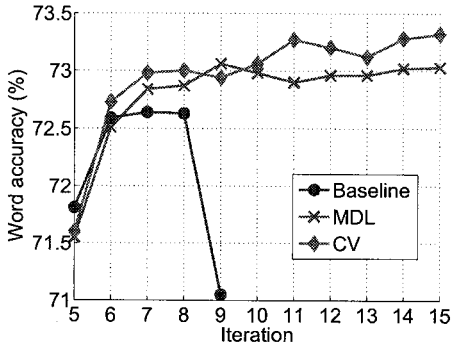

図 4: 学習ループ数と単語正解精度

安定な要素分布が最適化後も残る傾向が強いのに対し、CV 尤度法ではそのような分布は初期の段階で併合対象となり刈り取られる傾向が見られ、CV 手法が従来法よりも効果的に働きよりよい併合対選択が行われることが確かめられた。

5 まとめと今後の課題

CV 尤度を用いた混合ガウス分布の最適化手法の提案を行った。提案法では十分統計量を用いることで、CV 尤度を効率的に評価することが可能である。混合分布の最適化において、CV 尤度は正のバイアスを含む従来の自己尤度に比べてより信頼でき、明確な停止基準を与える利点がある。CSJ を用いた大語彙連続音声認識実験において、提案手法を用いてモデルの学習を行うことで、従来法と比べて高い認識性能が得られることを示した。

今後の課題としては、これまでに提案した CV 学習法である、CV 尤度を用いた HMM 状態クラスタリング (CV-DTC) [3] および CV 尤度を用いた EM アルゴリズム (CV-EM) [10] との組合せが挙げられる。CV-EM は CV を EM の枠組に導入した学習アルゴリズムで、CV をモデル選択ではなく十分統計量のバイアスを取り除く目的で使用するという点が、CV-DTC や今回提案した混合分布最適化法とはやや異なる。これら CV 学習法を組み合わせることで、HMM の基本的な学習プロセスで用いられるすべての尤度を CV 尤度に置き換えることができ、過学習に対してより頑健で、高い汎化性能をもつ学習プロセスを構成できると期待される。

本研究では音声認識をタスクとして評価を行ったが、提案した CV 尤度評価アルゴリズムおよび混合分布最適化手法は汎用性があり、混合ガウス分布を用いる様々な用途に応用可能である。さらに、提案手法はデータ駆動によるため、十分統計量が利用可能であれば目的関数は尤度である必要はなく、尤度以外を目的関数とした最適化手法に対しても応用できると考えている。

参考文献

[1] K. Shinoda and T. Watanabe. Acoustic modeling based on the MDL criterion for speech recognition. In *Proc. EuroSpeech*, volume 1, pages 99–102, 1997.

[2] H. Melin, J. W. Koolwaaij, J. Lindberg, and F. Bimbot. A comparative evaluation of variance flooring techniques in HMM-based speaker verification. In *Proc. ICSLP*, pages 2379–2382, Sydney, 1998.

[3] T. Shinozaki. HMM state clustering based on efficient cross-validation. In *Proc. ICASSP*, volume I, pages 1157–1160, Toulouse, 2006.

[4] M. Ostendorf and H. Singer. HMM topology design using maximum likelihood successive state splitting. *Computer Speech and Language*, 11:17–41, 1997.

[5] T. Cincarek, T. Tomoki, H. Saruwatari, and K. Shikano. Utterance-based selective training for the automatic creation of task-dependent acoustic models. *IEEE Trans. Audio, Speech, and Language Processing*, 15(1):150–161, 2007.

[6] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui. Benchmark test for speech recognition using the Corpus of Spontaneous Japanese. In *Proc. SSPR2003*, pages 135–138, 2003.

[7] S. Young, J. Odell, and P. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proc. ARPA Workshop on Human Language Technology*, pages 307–312, 1994.

[8] S. Young *et al. The HTK Book*. Cambridge University Engineering Department, 2005.

[9] A. Lee, T. Kawahara, and S. Doshita. An efficient two-pass search algorithm using word trellis index. In *Proc. ICSLP*, pages 1831–1834, 1998.

[10] T. Shinozaki and M. Ostendorf. Cross-validation EM training for robust parameter estimation. In *Proc. ICASSP*, pages 437–440, 2007.