音声言語情報処理 19－8
(1997. 12. 11)

# 状態クラスタリングによるHM-Netの構造決定法の検討

堀 貴明　加藤正治　伊藤彰則　好田正紀

山形大学工学部
〒992 米沢市城南4丁目3-16
E-mail : hori@ei5sun.yz.yamagata-u.ac.jp

**あらまし** 隠れマルコフ網(HM-Net)は，音素環境依存HMMの状態共有関係をネットワークとして表現する手法であり，高精度かつ頑健な音声認識を実現する音響モデルとして有効であることが示されている[5]. HM-Netを生成する方法として，これまで逐次状態分割法(Successive State Splitting:SSS)やその改良[6][7]が提案されているが，これらの手法は学習サンプルを用いて状態分割とパラメータ推定を繰り返す手法であるため，大量の学習サンプルを利用する場合には計算量が膨大になるという問題がある. 通常，HM-Netは少量のサンプルによって構造決定され，各状態の分布のみが多量のサンプルによって再推定される. しかし，少量のサンプルでは大語彙連続音声認識などに用いる大規模なHM-Netの構造が得られない. 本報告では大語彙連続音声認識のための高精度なHM-Netを作成する手法として，状態クラスタリングに基づく高速なHM-Netの構造決定法を提案する. 同じ学習サンプルセットを用いた場合に，本手法が高速かつ従来と同等のHM-Netを生成可能であることを連続音素認識実験により示す.
**キーワード** 音声認識，隠れマルコフモデル，音素環境依存モデル，隠れマルコフ網，状態クラスタリング

# A Study on A State Clustering-Based Topology Design Method for HM-Nets

Takaaki HORI, Masaharu KATOH, Akinori ITO, and Masaki KOHDA

Faculty of Engineering, Yamagata University
4-3-16 Johnan,Yonezawa-shi, 992 Japan

**Abstract** A Hidden Markov Network (HM-Net) is a highly accurate and robust acoustic model which represents a tied-state structure of context dependent Hidden Markov Models as a network[5]. A Successive State Splitting (SSS) method and its improved ones[6][7] have been already proposed to generate HM-Nets. However, there is a common problem in these algorithms. The problem is that much amount of computation is required when large amount of training data is used, because state splitting and parameter estimation are repeated using the training data. Although topologies of HM-Nets are usually designed with a part of training data and then only their output density distributions are estimated with all of the data, HM-Nets with large-scale topologies for large vocabulary continuous speech recognition (LVCSR) cannot be derived. In this paper, we propose a state clustering-based rapid topology design method to generate high accuracy HM-Nets for LVCSR. In continuous phoneme recognition experiments, it is shown that the proposed method is a fast algorithm and can generate HM-Nets equivalent to ones designed by conventional methods when the same training data is used.
**key words** speech recognition, Hidden Markov Model, context dependent model, Hidden Markov Network, state clustering

# 1 はじめに

音素環境依存 HMM は音素の平均的な特徴だけでなく前後の音素に依存した調音結合や異音化をも表現できる精密な音響モデルである．しかし，一般に精密なモデルは多くのパラメータを必要とするので，音声データが有限である限り，パラメータ推定精度とモデルの精密さの間にはトレードオフが存在する．有限な音声データを有効に利用して音素環境依存 HMM を作成するためにはパラメータの共有化が不可欠である．

パラメータの共有化を行なう方法は，始めに最も精密なモデルを作成してパラメータをクラスタリングする手法 [1]〜[4] とパラメータを逐次増加させる手法 [5]〜[7] とに分けられる．前者で代表的なのは音素決定木に基づく状態クラスタリング，後者で代表的なのは逐次状態分割法 (Successive State Splitting:SSS) [5] である．

SSS は状態分割とパラメータ推定を繰り返すことによってモデルを段階的に精密にする手法であり，状態共有関係を隠れマルコフ網 (Hidden Markov Network:HM-Net) という状態のネットワークとして表現する．時間方向の状態分割を含んでいるので，音素環境に依存した適切なモデルの状態長を自動的に設定できるという特徴がある．また，SSS の改良 (DT-SSS[6], ML-SSS[7]) も提案されており，性能改善がなされている．

しかしながら，SSS に基づく手法は学習サンプルを用いて状態分割とパラメータ推定を繰り返すため，計算量が膨大になるという問題がある．特に大量の学習サンプルを利用する場合には不向きであり，たいてい構造決定は少量のサンプルで行われ，各状態の分布のみを全サンプルによって再推定するのが一般的である．しかし，構造決定に用いるサンプルが少なければ大語彙連続音声認識などに用いる大規模な音響モデルを構成することが難しく，全学習サンプルを反映した HM-Net の構造が得られない．

本報告では大語彙連続音声認識のための高精度な HM-Net を作成する手法として，SSS を状態クラスタリングに基づいて改良した高速な HM-Net の構造決定法を提案する．同じ学習サンプルセットを用いた場合に，本手法が高速かつ従来と同等の HM-Net を生成可能であることを連続音素認識実験により示す．

# 2 隠れマルコフ網（HM-Net）

HM-Net は，HMM の状態をノードとするネットワークとして表され，音素環境依存 HMM がそれらの状態を共有している．個々の状態は，それぞれ以下の情報を保有している．

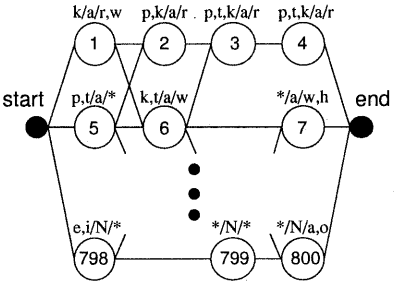



図 1: Hidden Markov Network

(1) 状態番号

(2) 受理可能なコンテキストクラス

(3) 先行状態・後続状態のリスト

(4) 自己遷移確率と後続状態への遷移確率

(5) 出力確率分布パラメータ

HM-Net では，コンテキスト情報が与えられたとき，そのコンテキストを受理することができる状態を先行状態・後続状態のリストの制約内で連結することにより，そのコンテキストに対するモデルを一意に決定することができる．このモデルは自己ループと隣の状態への遷移のみを許す left-to-right 型 HMM と等価であるため，通常の HMM と同様にパラメータ推定を行うことができる．

HM-Net の例を図 1 に示す．図の例は前後 1 音素に依存する triphone モデルの場合で，各状態は受理可能なコンテキストクラスとして先行／中心／後続音素の集合を持っている．例えば，コンテキスト k/a/w の HMM は状態列 1,6,7 によって与えられる．

# 3 状態クラスタリングによる HM-Net の生成

## 3.1 アルゴリズム

HM-Net の生成は，基本的に小さな HM-Net から状態分割を繰り返して大きく精密な HM-Net へ成長させることによって行なわれる．状態の分割にはコンテキスト方向と時間方向があり，各段階で尤度を基準に適切な方向が選択される．SSS は状態分割時の尤度の計算と分割後のパラメータ推定に要する処理量が非常に大きく，その処理量は学習サンプルの量に大きく依存する．これに対し，状態クラスタリングは始めに最も詳細な音素環

境依存モデルを学習し，その後でそれらの状態をクラスタリングするという枠組なので，その処理量はモデル数に依存し，SSS よりも遥かに少ない．

提案法は，基本的に SSS の枠組を保ちながら事前に作成された音素環境依存モデルの状態クラスタリングを行なう手法である．コンテキスト方向の分割は音素決定木に基づく質問によって，時間方向の分割は状態のコピー後にパラメータを更新することによって行われる．

アルゴリズムを次に示す．

(1) 3状態の単一ガウス分布の音素環境独立 HMM を学習する．

(2) 音素環境独立 HMM をコピーすることで音素環境依存 HMM を作成し，再学習する．

(3) 状態を時間方向にコピーすることで各モデルを時間方向に伸ばし，再学習する．

(4) HM-Net の初期構造を定義し，それに従って状態を共有化する．

(5) コンテキストクラスを分割する質問群を用意し，任意の状態数になるまで次の状態分割を繰り返す．

- 分割可能なすべての状態に対して以下の処理を行なう．
  - (a) コンテキスト方向：尤度比が最大となる質問によって状態を分割する．
  - (b) 時間方向：時間方向に状態をコピーし，パラメータを更新した後で尤度比を求める．
- 各状態のコンテキスト方向，時間方向の分割の中から尤度比最大のものを採用し，分割の影響が及ぶ範囲で HM-Net のパラメータを更新する．

(6) 各状態の分布数を任意の数に増やし，再学習する．

分割基準となる尤度比として，分割前と分割後の対数尤度の期待値の差を用いる．HM-Net の状態 $S$における対数尤度の期待値は，

$$L(S) = -\frac{1}{2}\sum_{k=1}^{K}(\log(2\pi\sigma_{Sk}^2)+1) \times \sum_{m\in C(S)}\sum_{n=1}^{N_m}\gamma_S(s_{mn})f(s_{mn}) \quad (1)$$

のようにして求められる．ここで，$\sigma_{Sk}^2$は状態 $S$の分布(無相関)の $k$番目の次元の分散，$C(S)$ は状態 $S$が受理できるコンテキストクラス，$N_m$は音素環境依存モデル $m$ の状態数，$s_{mn}$はモデル $m$ の $n$ 番目の状態，$K$は特徴ベクトルの次元数である．また，$f(s_{mn})$ は状態 $s_{mn}$ のフレーム単位の状態占有時間であり，音素環境依存モデルの学習時に得られる．$\gamma_S(s_{mn})$ は状態 $S$に状態 $s_{mn}$が対応する確率であり，求め方については次節で述べる．

### 3.2 状態対応確率

従来の音素決定木に基づく状態クラスタリングでは，音素環境依存モデルはすべて同じ状態数であり，状態の位置毎に共有化が行われる [3][4]．よって，どの状態がどの葉のグループに対応するかは明確であった．しかし，提案法においては各音素環境依存モデルの長さと HM-Net の中に含まれる状態列の長さが一致しないので，音素環境依存モデルの各状態が HM-Net のどの状態に対応するかは明確にせず，確率的に対応するものとする．

音素環境依存モデル $m$ の状態 $s_{mn}$の学習に用いられたサンプル集合 $X_{mn}$が HM-Net の状態 $S$から出力される対数尤度は

$$\begin{aligned} &L(X_{mn}|S) \\ &\approx f(s_{mn})\int_{-\infty}^{\infty}P(x|s_{mn})\log P(x|S)dx \\ &= -\frac{1}{2}\sum_{k=1}^{K}\left\{\log(2\pi\sigma_{Sk}^2)+\frac{(\mu_{mnk}-\mu_{Sk})^2+\sigma_{mnk}^2}{\sigma_{Sk}^2}\right\} \\ &\qquad \times f(s_{mn}) \end{aligned} \quad (2)$$

のように近似でき，$P(s_{mn}|S)=e^{L(X_{mn}|S)}$を $S$から $s_{mn}$が生じる確率として定義する．ここで，$\mu_{mnk}$，$\sigma_{mnk}^2$，$\mu_{Sk}$，$\sigma_{Sk}^2$は音素環境依存モデルの状態と HM-Net の状態の分布の平均と分散である．始めと終りの状態は必ず対応するという制約の下で，音素環境依存モデル $m$ の状態列 $s_{mn}$がそれを受理する HM-Net の状態列 $S_i$ $(1 \le i \le H_m)$ から生起する時の Forward 確率と Backward 確率は

$$\alpha_m(1,i) = \begin{cases} P(s_{m1}|S_i) & i=1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$\alpha_m(n,i) = \sum_{k=i,i-1}\alpha_m(n-1,k)P(s_{mn}|S_i) \quad (4)$$

$$\beta_m(N_m,i) = \begin{cases} 1.0 & i=H_m \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$\beta_m(n,i) = \sum_{k=i,i+1}\beta_m(n+1,k)P(s_{m(n+1)}|S_k) \quad (6)$$

のように計算される (図2)．ここで，$H_m$は音素環境依存モデル $m$ を受理する HM-Net の状態列の長さを表す．状態遷移確率は状態の対応関係にほとんど影響を与え

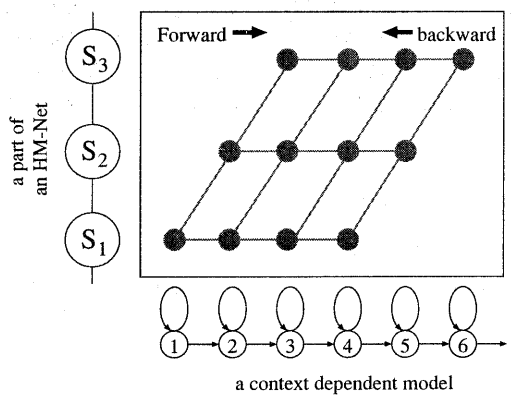



図 2: 音素環境依存 HMM と HM-Net の状態対応確率の計算

ないと考えられるため，本手法では考慮されていない．

この Forward 確率と Backward 確率より，$s_{mn}$が $S_i$ に対応する確率は

$$\gamma_{S_i}(s_{mn}) = \frac{\alpha_m(n,i)\beta_m(n,i)}{\sum_n \alpha_m(n,i)\beta_m(n,i)} \tag{7}$$

のように求めることができる．

### 3.3 状態分割と分布の推定

状態対応確率$\gamma_{S_i}(s_{mn})$を用いれば，HM-Net の状態の分布パラメータを次式のように最尤推定することができる．

$$\hat{\mu}_{S_i k} = \frac{\sum_{m\in C(S_i)} \sum_{n=1}^{N_m} \gamma_{S_i}(s_{mn}) f(s_{mn}) \mu_{mnk}}{\sum_{m\in C(S_i)} \sum_{n=1}^{N_m} \gamma_{S_i}(s_{mn}) f(s_{mn})} \tag{8}$$

$$\hat{\sigma}^2_{S_i k} = \frac{\sum_{m\in C(S_i)} \sum_{n=1}^{N_m} \gamma_{S_i}(s_{mn}) f(s_{mn}) (\mu^2_{mnk} + \sigma^2_{mnk})}{\sum_{m\in C(S_i)} \sum_{n=1}^{N_m} \gamma_{S_i}(s_{mn}) f(s_{mn})} - \hat{\mu}^2_{S_i k} \tag{9}$$

これらの再推定式によって，状態分割時の分布の推定，および，分割後のパラメータの更新が行われる．

コンテキスト方向の状態分割における yes 側と no 側の分布は，先にコンテキストクラスを分割してから式 (8)(9) を適用して求める．このときの状態対応確率は分割前に求めた値を用いるので，ここには状態分割の前後で状態の対応関係が変化しないという仮定がある．最適なコンテキスト方向の分割を選択するために，分割処理はおおよそ (質問数×状態数) 回必要になるが，この仮定によって処理量は大幅に削減される．

一方，時間方向の場合は，状態の対応関係は被分割状態の前後で大きく変化すると考えられるので，被分割状態の分布パラメータを新しい状態にコピーしてからその分割が影響するすべての状態の分布を再推定する．そのため，分割の処理量はかなり多くなるが，最適な分割を選択するのに必要な分割処理は (状態数) 回であり，かつ，前回の分割において採用されなかった分割結果を保持して計算の重複を避ければ，さほど問題にはならない．

状態分割後，状態対応確率と各状態の分布パラメータを，その分割の影響が及ぶ範囲で更新する．

表 1: 音声分析条件

| | |
|---|---|
| 標本化周波数 | 12 kHz |
| 量子化 | 16 bit |
| フレーム長 | 32 msec |
| フレーム周期 | 8 msec |
| 分析窓 | ハミング窓 |
| 高域強調 | $1 - z^{-1}$ |
| 特徴ベクトル | 1〜12 次の LPC メルケプストラム係数と対数パワー，およびその一次と二次の回帰係数（計 39 次元） |
| 正規化 | 発話毎のケプストラム平均正規化 |

## 4 連続音素認識実験

### 4.1 音声資料・音声分析

音声資料は，日本音響学会連続音声データベースの男性 26 名が発声した音素バランス文を用いる．A〜J セット（1 セット 50 文，J セットのみ 53 文）の中から A〜I セットを発声している男性 20 名の計 3000 文を学習用，残りの 6 名の J セット計 318 文を評価用とする．音声分析条件を表 1 に示す．

### 4.2 HM-Net の生成

音素カテゴリーは 27 音素＋無音 (sil) とする．状態クラスタリングのための音素環境依存 HMM は先行音素と後続音素を考慮する triphone モデルとする．始めに 3 状態，単一ガウス分布の triphone モデルを学習し，状態占有時間 (state occupancey) を求める．その後，平均占有時間の長い状態から順に，各モデルが 4〜6 状態，

全状態数が1.5倍になるまで時間方向に状態のコピーを繰り返すことによって長いtriphoneモデルを作成する．

初期HM-Netは，各音素3状態の音素環境独立HMMを並列に接続した81状態のHM-Netを用いる．コンテキストクラスを分割する質問群は，文献[8]に基づいて音素群47種類を選定し，これらを音素環境要因（先行音素，後続音素）と対にして作成した．

状態分割は800状態まで行い，時間方向の分割は4状態までとする．分割後に各状態の出力確率分布を4混合に再構成して連結学習を5回行う．

比較のために，次の手法によるHM-Netも作成する．

(1) 逐次状態分割法 (SSS)[5]

(2) 音素決定木に基づく逐次状態分割法 (DT-SSS)[6]

(3) 時間方向の分割を含まないDT-SSS

(4) 被分割状態を一つ選択して分割する状態クラスタリング

(5) 状態位置別の音素決定木による状態クラスタリング[4]

SSSは不特定話者による構造決定が困難であるため，特定話者の音声データによって決定した構造を用い，各状態の出力確率分布のみを不特定話者の音声データによって推定する．特定話者のHM-NetはATR重要語5240単語，発声者MHTの音声データを用いて作成されたものを用いる．

SSSによるHM-Netには受理できないコンテキストが生じる．また，特定話者の構造を利用する場合，学習データの違いから不特定話者の学習サンプルがほとんど割当たらない状態も存在する．受理できないコンテキストや過度に学習サンプルが少ない状態を含むモデルは3状態8混合分布の音素環境独立HMMで代用する．

(4)の被分割状態を一つに絞る基準は，SSS/DT-SSSと同じ基準

$$d_i = n_i \sum_{k=1}^{K} \frac{\sigma_{S_i k}^2}{\sigma_{Tk}^2} \tag{10}$$

を用いる．ここで，$n_i$は状態$S_i$の学習に用いられたサンプルの数，$\sigma_{Tk}^2$は全学習サンプルの分散を表す．

## 4.3 連続音素認識

HM-Netを評価する認識実験として連続音素認識実験を行う．日本語の音節の制約に基づく音素ネットワークを構成し，最適音素系列を求めたときの正解音素系列に対する音素正解率（Phoneme Accuracy）：

$$\text{音素正解率} = \frac{M - S - I - D}{M} \times 100(\%) \tag{11}$$

によって評価する．ここで，$M$は評価サンプルの総音素数，$S$は置換誤り，$I$は挿入誤り，$D$は脱落誤りの数を表す．

## 4.4 実験結果と考察

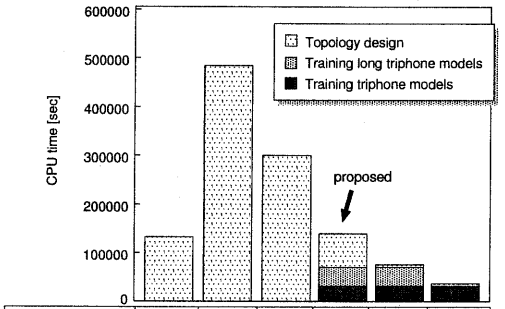
提案する状態クラスタリング，および，比較用の各手法によってHM-Netの構造決定を行い，連続音素認識実験を行った．800状態のHM-Netの構造決定に要する処理量を図3に示す．図中の“split check”の欄は，状態分割時に式(10)に基づいて一つの状態を選択して分割するか(one)，全状態の分割を行ってから最良の分割を選択するか(all)の違いを，“temporal split”の欄は時間方向の分割を行なうか(with)，行なわないか(without)の違いを表す．“PDT Clustering”は，3状態のtriphoneモデルを状態位置依存の音素決定木によってクラスタリングしたときの結果を表す．CPUタイムは汎用ワークステーションHP735によって測定した結果であり，状態クラスタリングの処理量にはtriphoneモデルの学習に要した時間も含まれ，時間方向の分割を行う場合にはtriphoneモデルを長く伸ばした後の再学習も含まれている．また，SSSの場合は特定話者のデータによる構造決定の処理時間を示している．

提案した状態クラスタリングとDT-SSSの構造決定に要する処理量を比較すると，時間方向の分割を含む場合にはDT-SSSの処理量がかなり大きく，全状態をチェックするクラスタリングの3.5倍程度かかることが分かる．

図4は，各手法によるHM-Netの音素正解率(phoneme accuracy)を示している．DT-SSSと提案法の計算量の差が大きいにも関わらず，音素正解率は同等以上の値を示しており，本手法の有効性が示された．状態クラスタリングにおける全状態のチェックや時間方向の分割は処理量をかなり増加させるが，誤り削減が全状態のチェックによって約4%，時間方向の分割によって約7%得られた．

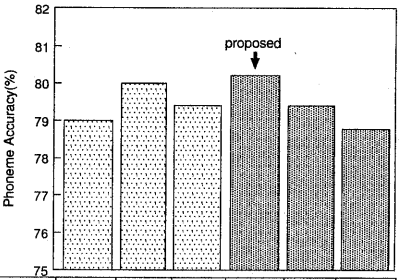
図5は，状態クラスタリングにおける対数尤度の期待値の変化を示している．図中の“All-check”，“1-check”は提案した状態クラスタリングにおける被分割状態の選択方法の違いを表している．これより，全状態のチェックと時間方向の分割がより高い尤度を有するモデルの生成に有効であることが分かる．

同じ量の学習サンプルで性能の差が少ないこと，DT-SSSにおいてこれ以上学習サンプルを増やすことが現実的ではないこと，状態クラスタリングの処理量が学習サンプルの量にほとんど依存しないこと等を考慮すると，大量の学習サンプルを利用する場合に時間方向の分割を含む状態クラスタリングは有効と言える．




| Topology design | sss | DT-SSS | DT-SSS | State Clustering | State Clustering | PDT Clustering |
|---|---|---|---|---|---|---|
| Split check | one | one | one | all | one | all |
| Temporal split | with | with | without | with | with | without |

図 3: DT-SSS と状態クラスタリングによる HM-Net の構造決定に要する処理時間




| Topology design | sss | DT-SSS | DT-SSS | State Clustering | State Clustering | PDT Clustering |
|---|---|---|---|---|---|---|
| Split check | one | one | one | all | one | all |
| Temporal split | with | with | without | with | with | without |

図 4: DT-SSS と状態クラスタリングによる HM-Net の音素正解率

## 5 むすび

状態クラスタリングによる HM-Net の構造決定法を提案し，高速かつ従来と同等の HM-Net を生成可能であることを示した．今後は，分割の基準や適切な状態数を決定する分割停止基準を検討した上で，大量の学習サンプルを用いたより大きな HM-Net を構成し，大語彙連続音声認識システムに実装する予定である．

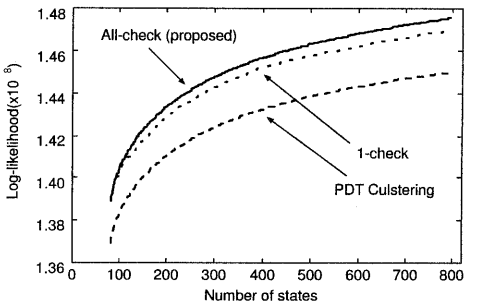



図 5: 状態クラスタリングにおける対数尤度の期待値の変化

## 参考文献

[1] S. J. Young and P. C. Woodland :"State Clustering in hidden Markov model-based continuous speech recognition", *Computer Speech and Language*, Vol.8, No.4, pp.369–383(1994).

[2] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo and M. A. Picheny :"Decision trees for phonological rules in continuous speech", *ICASSP91*, pp.185–188(1991).

[3] M.-Y. Hwang, X.Huang and F.A.Alleva :"Predicting unseen triphones with senones", *IEEE Trans. Speech and Audio Processing*, Vol.4, No.6, pp.412–419 (1996).

[4] S.J.Young, J.J.Odell and P.C.Woodland:"Tree-based state tying for high accuracy acoustics modeling", *Proc. ARPA Human Language Technology Workshop*, pp.307–312 (1994).

[5] 鷹見，嵯峨山:"逐次状態分割法による隠れマルコフ網の自動生成", 信学論 (D-II), J76-D-II, 10, pp.2155–2164(1993-10).

[6] 堀，加藤，伊藤，好田:"音素決定木に基づく逐次状態分割法による HM-Net の検討", 信学論 (D-II), J80-D-II, 10, pp.2645-2654(1997-10).

[7] M.Ostendorf and H.Singer:"HMM topology design using maximum likelihood successive state splitting", *Computer Speech and Language*, Vol.11, No.1, pp.17–41 (1997).

[8] 城生佰太郎 :"音声学，新装増訂版", アポロン音楽工業 (1992).