

情報量基準を用いたトライフォン HMM の状態共有に関する一考察

知識情報処理研究

3603R004-6 岩田錦弥

指 導 松嶋敏泰教授

A Note on States Sharing of Triphone Hidden Markov Models by Information Criterion

by Kinya Iwata

1 はじめに

近年、音声認識技術がカーナビ、ゲーム、PC 上のソフトウェアなど広く用いられるようになってきた。人がある単語を発声した時、その周波数を解析し、どの単語だったかを認識する単語認識問題は、実用上重要な問題である。単語認識問題の確率モデルとしては、音声のパワースペクトルを信号処理した特徴ベクトルが隠れ状態に依存すると仮定した Hidden Markov Model (HMM) が広く使われている。しかし認識する単語数が多い場合、単語毎の HMM (単語 HMM) を用意するのは、学習データが多く必要となり現実的ではない。そこで単語を構成する音素ごとに HMM を用意し、それらを連結して単語を表現することで 1 つの HMM 当たり十分な学習データを確保している。

一般に音素の音響的特徴は、その前後の音素の影響を受け変化することが知られている (調音結合)。そのため、前後の音素を考慮した 3 つ組みの認識単位 (トライフォン) が提案され、トライフォン毎の HMM (トライフォン HMM) を使うことの有効性が確認されている [1]。しかしトライフォン HMM の総数は音素数に対して指数オーダーで増えてしまう。そこで、複数のトライフォン HMM を同一中心音素、同一状態ごとに共有することで、この問題を解決する手法が提案されている [2] [3]。

中心音素を固定した下でのトライフォン HMM の状態共有に対する従来研究では、重み付けユークリッド距離 [2]、エントロピー [3]などを基準に用いて共有を行うかどうか決めていた。しかし、これらの方法は状態共有の停止条件としてあらかじめ定めた閾値を利用するため、統計学的な視点から見ると真の状態共有に対する保証がない。さらにこの閾値の決定に予備的な繰り返し実験が必要となるため、多くのデータ量や計算量が必要となる問題点もあった。

そこで本研究では、トライフォン HMM の状態共有

問題をモデル選択問題ととらえ、情報量基準 Akaike's Information Criterion (AIC) [4] を利用することで、真の状態共有のモデルと Kullback-Leibler 距離の意味で近いモデルを選択することを試みた。この結果、従来法で必要であった予備的な繰り返し実験を行う必要なくなると考えられる。このアプローチの性能を検証するため、単語認識実験を行なった。

2 準備

2.1 トライフォン

トライフォンとは、前後の音素を考慮した 3 つ組み音素の認識単位である。トライフォンの種類数は最大で音素数の 3 乗 (音素数が 43 個なら 79507 個) となる。3 つ組み音素のうち中央の音素のことを中心音素と呼び、中心音素に対して前の音素を-、後の音素を+で連結トライフォンを表わす。図 1 でトライフォン系列の例を示す。

単語	: わせだ
音素系列	: w a s e d a
トライフォン系列	: w+a w-a+s a-s+e s-e+d e-d+a d-a

図 1: トライフォン系列

2.2 トライフォン HMM

トライフォン $u \in V$, $V = \{V_1, V_2, \dots, V_{|V|}\}$ と対応する D 次元特徴ベクトル系列 $\mathbf{x} = (x_1, x_2, \dots, x_t, \dots, x_T)$, $x_t \in \mathbb{R}^D$ を left-to-right 型 HMM で確率モデル化し、確率分布 $f(\mathbf{x}|u, \theta_{m_k}^u, m_k)$ をトライフォン HMM とする。ここで k を負でない整数として m_k は状態共有のパターンを表わすモデル (3.1.2 節を参照), $\theta_{m_k}^u = \{A_{m_k}^u, B_{m_k}^u\}$ はパラメータ集合である。図 2 は $S = \{S_i | 0 \leq i \leq 4\}$ とした時の

トライフォン HMM の例である．状態 S_i の確率分布を $b_{i,m_k}^u(\cdot)$ とし，これに式 (1) で表わされる正規分布を仮定する．

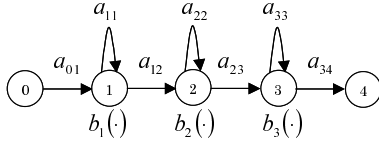


図 2: トライフォン HMM

$$b_{i,m_k}^u(x_t) = (2\pi)^{-\frac{1}{2}} |\Sigma_{i,m_k}^u|^{-\frac{1}{2}} \times \exp \left\{ -\frac{1}{2} (x_t - \mu_{i,m_k}^u)^T \Sigma_{i,m_k}^{u-1} (x_t - \mu_{i,m_k}^u) \right\}. \quad (1)$$

$A_{m_k}^u = \{a_{ij,m_k}^u | 0 \leq i, j \leq 4\}$ は状態遷移確率集合， $B_{m_k}^u = \{\mu_{m_k}^u, \Sigma_{m_k}^u\}$ ， $\mu_{m_k}^u = \{\mu_{i,m_k}^u | 1 \leq i \leq 3\}$ は平均ベクトル集合， $\Sigma_{m_k}^u = \{\Sigma_{i,m_k}^u | 1 \leq i \leq 3\}$ は対角分散共分散行列集合， $\mathbf{z} = (z_1, z_2, \dots, z_T)$ ， $z_t \in \{i | 0 \leq i \leq 4\}$ は状態系列とする．この時，トライフォンを入力とし，特徴ベクトル系列を出力するトライフォン HMM は次のようになる．

$$f(\mathbf{x} | u, \theta_{m_k}^u, m_k) = \sum_{\mathbf{z}} a_{0z_1,m_k}^u \times \left\{ \prod_{t=1}^{T-1} b_{z_t,m_k}^u(x_t) a_{z_t z_{t+1},m_k}^u \right\} b_{z_T,m_k}^u(x_T) a_{z_T 4,m_k}^u. \quad (2)$$

2.3 単語 HMM

単語 $y \in W$ ， $W = \{W_1, W_2, \dots, W_{|W|}\}$ と対応する特徴ベクトルを M 状態 left-to-right 型 HMM で確率モデル化し，確率分布 $f(\mathbf{x} | y, \theta_{m_k}^y, m_k)$ を単語 HMM とする．

$S = \{S_i | 0 \leq i \leq M-1\}$ を状態とした時の状態 S_i の確率分布に正規分布を仮定し $b_{i,m_k}^y(\cdot)$ とする．単語 y を入力とし，特徴ベクトル系列を出力する単語 HMM は式 (2) のトライフォン HMM と同じく以下で表わす．

$$f(\mathbf{x} | y, \theta_{m_k}^y, m_k) = \sum_{\mathbf{z}} a_{0z_1,m_k}^y \times \left\{ \prod_{t=1}^{T-1} b_{z_t,m_k}^y(x_t) a_{z_t z_{t+1},m_k}^y \right\} b_{z_T,m_k}^y(x_T) a_{z_T M,m_k}^y. \quad (3)$$

ここで単語 y に対する HMM のパラメータ集合を前節の表記にならい， $\theta_{m_k}^y$ のように表わすこととする．

3 単語認識問題

本研究で扱う単語認識問題では，単語を $y_i \in W$ ， $W = \{W_1, W_2, \dots, W_{|W|}\}$ ，対応する D 次元特徴ベクトル系列を $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iT_i})$ として確率モデル $f(\mathbf{x} | y, \theta_{m_k}^y, m_k)$ を仮定する．ここで m_k は状態共有のモデル， $\theta_{m_k}^y$ はモデル m_k のパラメータを表す．いま， n 組の学習データ $\mathbf{x}^n = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ ， $y^n = y_1, y_2, \dots, y_n$ と $n+1$ 番目の特徴ベクトル $\mathbf{x}_{n+1}$ が与えられた時に，未知の単語 y_{n+1} を推定する問題のことを単語認識問題と定義する．

この問題は，確率モデル $f(\mathbf{x} | y, \theta_{m_k}^y, m_k)$ の $\theta_{m_k}^y$ と m_k を推定する「学習」と y_{n+1} を推定する「認識」に分けられる．以下，これらについて詳しく述べる．

3.1 学習

学習データ $\mathbf{x}^n, y^n$ から $\theta_{m_k}^y$ と m_k を推定する．「連結学習」，「トライフォン HMM の状態共有」，「単語 HMM の推定」の流れで行う．

3.1.1 連結学習

学習データ $\mathbf{x}^n, y^n$ から共有を全く行わないモデル m_0 に対するトライフォン HMM のパラメータ $\theta_{m_0}^u$ の推定を行う．この時，与えられた学習データは単語に対するものである．よって，トライフォン HMM を連結することで単語 HMM とし，トライフォン HMM のパラメータ $\hat{\theta}_{m_0}^u$ を近似最尤法で推定する．これを連結学習と呼ぶ．

3.1.2 トライフォン HMM の状態共有

トライフォン HMM の状態共有 (= モデルの選択 $\hat{m}_k$) とパラメータの推定 $\hat{\theta}_{\hat{m}_k}$ を行う．状態共有を行うメリットとしては，限られた学習データの中で高精度よくパラメータの推定を行えることが挙げられる．ここで，トライフォン HMM の状態共有とは，同一中心音素を持つトライフォン HMM の中で同一状態のものを共有することである．図 3 では，中心音素 a を持つトライフォンの状態共有の様子を示している．

ここで，モデル m_k ， $k = 0, 1, \dots, K$ は全中心音素，全状態に対する状態共有の 1 パターンに対応する．よって，状態共有を決定することとモデル m_k を選択することは等価となる．従来研究では，状態共有を行うかどうかの基準として，重み付きユークリッド距離 [2] やエントロピー [3] を利用している．

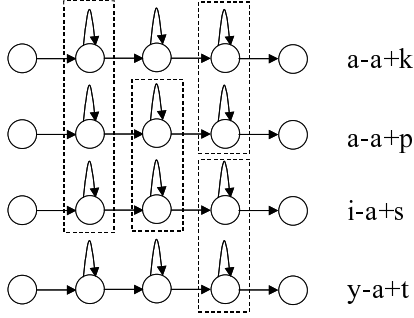


図 3: 状態の共有

また、状態の共有を行った時、すなわちモデル $\hat{m}_k$ を選択した時のパラメータ $\hat{\theta}_{\hat{m}_k}^u$ は最尤推定を利用し求める。

3.1.3 単語 HMM の推定

トライフォン HMM と単語 HMM の連結の関係を利用して、モデル $\hat{m}_k$ に対する単語 HMM を求める。トライフォン HMM のパラメータ $\hat{\theta}_{\hat{m}_k}^u = \{\hat{A}_{\hat{m}_k}^u, \hat{B}_{\hat{m}_k}^u\}$ から単語 HMM のパラメータ $\hat{\theta}_{\hat{m}_k}^y = \{\hat{A}_{\hat{m}_k}^y, \hat{B}_{\hat{m}_k}^y\}$ を以下のように求めることができる。

$$\hat{A}_{\hat{m}_k}^y = \{\hat{A}_{\hat{m}_k}^{u_1}, \hat{A}_{\hat{m}_k}^{u_2}, \hat{A}_{\hat{m}_k}^{u_3}\}, \quad (4)$$

ただし、 $\hat{a}_{01, \hat{m}_k}^{u_2}, \hat{a}_{01, \hat{m}_k}^{u_3}$ は除く。

$$\hat{B}_{\hat{m}_k}^y = \{\hat{B}_{\hat{m}_k}^{u_1}, \hat{B}_{\hat{m}_k}^{u_2}, \hat{B}_{\hat{m}_k}^{u_3}\}. \quad (5)$$

式 (4)(5) よりモデル $\hat{m}_k$ に対する単語 HMM の確率分布 $f(x|y, \hat{\theta}_{\hat{m}_k}^y, \hat{m}_k)$ を求めることができる。

3.2 認識

特徴ベクトル x_{n+1} が与えられた時、3.1 節までで学習した単語 HMM の確率分布 $f(x|y, \hat{\theta}_{\hat{m}_k}^y, \hat{m}_k)$ を利用して式 (6) で表わされる尤度最大基準で認識する。

$$\hat{y}_{n+1} = \arg \max_{y_{n+1}} f(x_{n+1}|y_{n+1}, \hat{\theta}_{\hat{m}_k}^y, \hat{m}_k). \quad (6)$$

4 従来研究

トライフォン HMM の状態共有の従来研究 [2] [3] では、同一中心音素、同一状態の集合 $\hat{S} = \{S^i | 1 \leq i \leq I\}$ をある基準に基づいてグルーピングを行う。ここで、 $C = \{C^i | 1 \leq i \leq I\}$ をクラスター集合とする。

状態共有を行うに当たり、以下の条件を仮定する。

- 状態遷移確率の尤度への寄与を無視する
- 状態共有の操作の間、学習データの各状態への割当てを固定する

これらの仮定により反復計算を用いることなく最尤推定値を求めることができ、計算量を削減できる。

4.1 Odell の研究 [2]

基準として、状態 S^i と S^j の平均ベクトルの重み付けユークリッド距離を利用する。

$$d(i, j) = \left[\frac{1}{D} \sum_{d=1}^D \frac{(\mu_d^i - \mu_d^j)^2}{\sigma_d^i \sigma_d^j} \right]^{\frac{1}{2}}. \quad (7)$$

ここで、 $\{\mu^i, \Sigma^i\}$ は状態 S^i に対するパラメータ、 $\mu^i = (\mu_1^i, \mu_2^i, \dots, \mu_D^i)$ は平均ベクトル、 $\Sigma^i = \text{diag}(\sigma_1^{i2}, \sigma_2^{i2}, \dots, \sigma_D^{i2})$ は分散共分散行列とする。クラスター間の距離には最長距離法を採用する。また、状態共有の停止条件には、クラスター内の状態間距離の最大値に対してあらかじめ定められた閾値を利用する。

アルゴリズムは以下ようになる。

1. 各状態 S^i , $i = 1, 2, \dots, I$ を 1 つのクラスター C^i , $i = 1, 2, \dots, I$ とみなす。
2. $d(i, j)$ を最小にするクラスターの組 C^i, C^j を結合 (共有) し、新しいクラスター C^k とする。
3. クラスター内の状態間距離の最大値が閾値より小さいならば、ステップ 2. を繰り返し、大きいならば結合 (共有) を停止する。

4.2 従来法の問題点

従来法は、状態共有の停止条件としてあらかじめ定めた閾値を利用するため、統計学的な視点から見ると真の状態共有に対する保証がないこと、また停止条件の閾値を決定するために予備的な繰り返し実験が必要となるため、多くのデータ量、計算量が必要になることが挙げられる。

5 提案法

本研究では、トライフォン HMM の状態共有問題をモデル選択問題ととらえ、情報量基準 Akaike's Information Criterion (AIC) [4] を利用することで、真の状態共有のモデルと Kullback-Leibler 距離の意味に近いモデルを選択することを試みる。この結果、従来法で必要であった予備的な繰り返し実験を行う必要がなくなると考えられる。

5.1 AIC 基準 [4]

AIC は以下の式のようになり，AIC 値を最小にするモデルを選択することを考える．AIC は，上記のようにモデルの最大対数尤度とモデルの自由パラメータ数のトレードオフの関係を利用して，モデルを選択する．

$$AIC = -2(\text{最大対数尤度}) + 2(\text{自由パラメータ数}). \quad (8)$$

5.2 提案アルゴリズム

提案法では，モデルの総数 K が膨大な数になるため，AIC を最小とするモデルを全数探索することは困難である．そのため，探索アルゴリズムとして，山登り法を利用する．以下の ΔA を最小とする C^i, C^j を結合する．

$$\begin{aligned} \Delta A = & (T(C^k)) \log |\hat{\Sigma}_k| \\ & - (T(C^i)) \log |\hat{\Sigma}_i| - (T(C^j)) \log |\hat{\Sigma}_j| + 4D. \end{aligned} \quad (9)$$

結合停止条件としては全てのクラスター組に対して， $\Delta A < 0$ となる場合とする．

アルゴリズムは以下ようになる．

1. 各状態 $S^i \quad i = 1, 2, \dots, I$ を 1 つのクラスター $C^i \quad i = 1, 2, \dots, I$ とみなす．
2. ΔA を最小にするクラスターの組 C^i, C^j を結合し，新しいクラスター C^k とする．
3. $\Delta A > 0$ ならば，ステップ 2. を繰り返し， $\Delta A < 0$ ならば結合を停止する．

6 シミュレーション

従来法 [2] と提案法に対して単語認識実験による比較を行った．

6.1 シミュレーション条件

入力音声は，標本周波数 16kHz，量子化 16bit，フレームサイズ 25ms，フレームシフト 10ms，特徴ベクトルは 25 次元を用いた．音素の種類数は 43 個，トライフォン HMM として 5 状態 left-to-right 型 HMM を利用した．学習データは日本音響学会研究用連続音声 DB (50 文，100 文，200 文)，テストデータは数字単語データ (10 単語 $\times$ 5 人) を利用した．

6.2 シミュレーション結果

前節で示した条件の下でのシミュレーション結果を表 1 に示す．

表 1: シミュレーション結果

	従来法		提案法
	閾値	認識率	認識率
50 文	2.0	52.0%	62.0%
	4.0	62.0%	
	6.0	62.0%	
100 文	0.5	68.0%	72.0%
	2.5	72.0%	
	4.5	72.0%	
200 文	1.0	64.0%	80.0%
	3.0	84.0%	
	5.0	80.0%	

7 考察

表 1 より，提案法の認識率は従来法 [2] とほぼ等しくなった．しかし，計算量については従来法が適当な閾値の設定に予備実験の繰り返しが必要となるのに対して，提案法では予備実験を行う必要がないため，従来法に比べて約 $1/(\text{予備実験回数})$ の計算量の削減となった．

8 まとめ

本研究ではトライフォン HMM の状態共有問題を扱い，従来法が抱えた共有基準に保証がないという問題をふまえ，真の分布に対して Kullback-Leibler 距離の意味で近いモデルを与える情報量基準 AIC を利用したアルゴリズムを提案した．その結果，認識率をほとんど落とすことなく，計算量を削減することができた．今後の課題としては，探索効率の改善などが挙げられる．

参考文献

- [1] K. F. Lee, *Automatic Speech Recognition. The development of the SPHINX system*, Kluwer Academic Publishers, 1989.
- [2] J. J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. Thesis, Cambridge University, Cambridge, Mar. 1995.
- [3] M.-Y. Hwang, X. Huang, and F. Alleva, "Predicting Unseen Triphones with Senones," *IEEE Trans. Speech Audio Processing*, vol.4, no.6, pp.412-419, Nov. 1996.
- [4] 坂元 慶行, 石黒真木夫, 北川源四郎, "情報量統計学," 共立出版, 1983.