隠れマルコフモデルによる時間領域での音声信号の学習と生成

宮崎大学 工学部 情報システム工学科
76050200 黒田 優士

1 はじめに

通常，音声研究では時間領域の音声波形はいったん短時間フーリエ変換され，周波数領域で特徴量が分析される．一度，周波数領域で表現された音声信号[1]を時間領域に戻すことは難しい．なぜなら，周波数領域では音声の情報が失われているからである．そのため，音声信号を生成するためには別のモデルを考える必要がある．

1995 年，Künsch らにより “Hidden markov random fields” と題する画期的な論文が発表された [1]．いま，$X = X_1, X_2, \cdots$ をマルコフ性をもつ確率過程とする．ある関数 f を使い $Y_t = f(X_t)$ として得られる $Y = Y_1, Y_2, \cdots$ を隠れマルコフ過程という．論文 [1] では，図 1 のような単純な隠れマルコフモデルが，Y に関する任意の確率分布をいくらでも精度良く表現する能力をもつことが証明されている．隠れマルコフモデルは確率的生成モデルである．例えば，Y として音声信号を考える．Y の確率分布を適切にモデル化することができれば，音声認識と音声合成が同一のモデルで実現できる．

本研究では，第一に，Künsch らの実験のアイデアをもとに，音声信号を学習させた．次に，学習と同一のモデルを用いて音声信号を生成することを試みた．隠れマルコフモデルは数学的には任意の確率分布を近似的に表現可能なモデルであっても，実際の表現能力については不明な点が多い．この点を重点的に調べた．その結果，音声信号の特徴を捕らえた学習に成功した．以下では，まずはじめに数理モデルおよび学習アルゴリズムについて述べる．次に，モデルの学習能力および生成能力を音声信号を用い，調べた結果を示す．また，合成についても実験をおこなったのでその結果を紹介する．最後に，まとめを述べる．

2 隠れマルコフモデル

確率的生成モデルとして，隠れマルコフモデル (HMM，Hidden Markov model) を使う．内部データ $X = X_1, X_2, \ldots, X_T$ について，時間 t で X_t が値 i をとるとき，$t+1$ で X_{t+1} がとる値 j は X_t と遷移確率 $\mathrm{Prob}\{X_t = i, X_{t+1} = j\} = p_{ij}, (i, j = 1, \ldots, N)$ のみで決まる．t 以前にどんな値をとってきたのかは関係しない．このような性質をマルコフ性といい，このときの X はマルコフ過程に従うという．ここで，遷移確率 p_{ij} をパラメータと呼ぶ．観測値 $Y = Y_1, Y_2, \ldots, Y_T$ は不可逆関数 f を用いて $Y_t = f(X_t)$, $t = 1, \ldots, T$ で求まる．X_t から Y_t は一つに決まるが，Y_t から X_t は一つに決められない．このような性質を持つ観測値 Y を隠れマルコフ過程という．一般的に Y はマルコフ過程ではない．

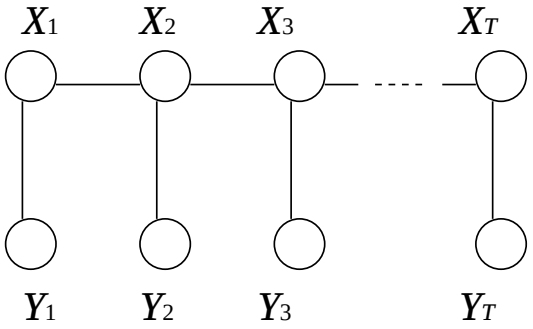

図 1: 隠れマルコフモデル．確率変数の依存性を示すグラフ．

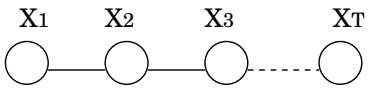

図 2: 確率変数の依存性．$p(X|Y)$

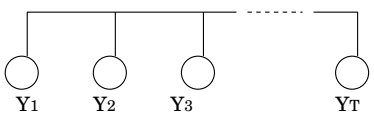

図 3: 確率変数の依存性．$p(Y) = \sum_X p(X, Y)$

図 1 に示すモデルを使う利点は 2 つある．一つは，事後確率 $p(X|Y)$ の構造は単なる線型のグラフで表現できることである（図 2）．この構造があるので，$p(X_2, X_3|Y)$ などが楽に計算できる．もう一つは，Y の周辺確率分布 $p(Y)$ が完全結合型のグラフで表現

[1] 本論文では，時間領域での音声波形を音声信号と呼ぶ．

できることにある（図3）．モデルがどんな複雑な確率分布も表現できることが理解できる．つまり，確率変数間 $Y_1, Y_2, \ldots, Y_T$ に複雑な依存性のどんな同時確率分布 $p(Y_1, Y_2, \ldots, Y_T)$ も，いくらでも精度良くモデル化できる．Y は観測値であるので推定する必要はない．一方，内部データ X については観測値 Y に対する解釈を表現するため X の周辺分布をはじめ，様々な関数を推定したい．図1の HMM は，この目的を満足させてくれる．

情報として与えられるのは観測値の Y である．HMM は尤度が最大となるパラメータを推定する．尤度とは，Y があるパラメータの組 S で作られる尤もらしさである．尤度を最大にするパラメータの組を $\hat{S}$ とすると，

$$\begin{aligned} \hat{S} &= \arg\max_S \mathrm{L}(S, y_1, \ldots, y_T) \\ \mathrm{L}(S, y_1, \ldots, y_T) &= \sum_x \mathrm{Prob}(x_1, \ldots, x_T, y_1, \ldots, y_T; S) \\ &= \mathrm{Prob}(y_1, \ldots, y_T; S) \end{aligned}$$

となる．

3 学習アルゴリズム

尤度を最大にするパラメータの値を求めるためのアルゴリズムを以下に示す．

1. m 回更新したときのパラメータの組を $S_m = \{p_{ij}^m\}$，時間 $t=1$ のときに値 $X_1 = k$ となる確率を $p_m(k)$ とする．$\sum_j p_{ij}^0 = 1$，$\sum_k p_0(k) = 1$ を満たすように，$\{p_{ij}^0\}, p_0(k)$ を初期化する．
2. (E step) S_m が既知．下の式で $X_t = i$ から $X_{t+1} = j$ に遷移する回数 n_{ij} を求める．E_{S_m} は S_m に関する期待値である．
($t = 1, 2, \ldots, T$，$i, j \in S$)

$$\begin{aligned} n_{ij}^{(m)} &= \mathrm{E}_{S_m}[n_{ij}|Y_1, \ldots, Y_T] \\ \mathrm{E}_{S_m}[\sharp n_{ij}|Y] &= \sum_{t=1}^{T-1} p(X_t = i, X_{t+1} = j|Y) \\ &= \sum_{t=1}^{T-1} \frac{p(X_t = i, X_{t+1} = j, Y)}{p(Y)} \end{aligned} \quad (1)$$

3. (M step) 下の式で確率行列を更新する．

$$\begin{aligned} p_{ij}^{m+1} &= \frac{\mathrm{E}[\sharp n_{ij}|Y]}{\sum_{\tilde{j}} \mathrm{E}[\sharp n_{i\tilde{j}}|Y]} \\ &= \frac{\sum_{t=1}^{T-1} p(X_t = i, X_{t+1} = j, Y)}{\sum_{t=1}^{T-1} p(X_t = i, Y)} \end{aligned} \quad (2)$$

また，X_1 がとりうる値についてそれぞれの確率を計算する．特定の学習データに依存しないように状態 $X_1 = k$ となる確率を計算する．

$$\begin{aligned} p_{m+1}(k) &= \frac{\mathrm{E}[\sharp n_k|Y]}{\sum_a \mathrm{E}[\sharp n_a|Y]} \\ &= \frac{\sum_{t=1}^{T} p(X_t = k|Y)}{\sum_a \sum_{t=1}^{T} p(X_t = a|Y)} \end{aligned} \quad (3)$$

4. 全ての p_{ij} の値が収束するまで 2，3 を繰り返す．

i, j, k は内部データ X がとる値である．ここで，このアルゴリズムは尤度の局所的な最大値に収束することに気をつけなければならない．パラメータを更新するための式 (2)，(3) は下のように簡単化できる．

$$p_{ij}^{m+1} = \frac{\sum_{t=1}^{T-1} L_t(i) p_{ij}^m q_{j,y_{t+1}}^m R_{t+1}(j)}{\sum_{t=1}^{T-1} L_t(i) R_t(i)} \quad (4)$$

$$L_1(k) = p_m(k) q_{i,y_1},\ R_T(i) = 1 \quad (5)$$

$$L_{t+1}(i) = \sum_j L_t(j) p_{ji} q_{i,y_{t+1}} \quad (6)$$

$$R_t(i) = \sum_j p_{ij} q_{j,y_{t+1}} R_{t+1}(j) \quad (7)$$

$$p_{m+1}(k) = \frac{\sum_{t=1}^{T} L_t(k) R_t(k)}{\sum_a \sum_{t=1}^{T} L_t(a) R_t(a)} \quad (8)$$

$q_{i,y}$ は，例えば $i \in \{1, 2, \ldots, 20\}$, $y \in \{1, 2, \ldots, 8\}$ の場合，$1 + i \bmod 8 = y$ となれば 1 ならなければ 0 をとる関数である．

以上が動的計画法を用いた計算方法である．これまで説明したモデルとアルゴリズムを使い実験した結果を次の章に示す．

4 音声信号の学習と生成

4.1 データ数の少ない音声信号での学習

音声信号を量子化ビット数 3（8 値），10kHz でサンプリングする．そのうちの 100msec 分を 0.5msec ごとにさらにサンプリングした 200 個のデータを学習データとして使用した（図 4）．

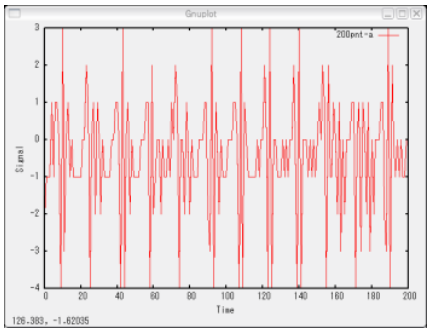

図 4: 100-millisecond segment．(横軸：時間，縦軸：振幅)

$Y_t = X_t \bmod 8$ とする．パラメータの初期値は区間 [0.0, 1.0) の一様乱数で設定して，尤度が最大となるパラメータを求めた．パラメータの組の大きさは $N \times N$ とする．N を $10, 30, 60, 100, 200, 400$ と変えて実験した．

N 通りの値をとるデータ $X^N = X_1^N, \ldots, X_{200}^N$ とする．図 5 は実験の結果得られたパラメータの組を使い，区間 [0.0, 1.0) の一様乱数で生成した観測値 Y の $t = 0, \ldots, 200$ の波形である．図 4 と図 5 のグラフを比較して評価した．図 5 は，N が大きくなるにつれて図 4 のグラフに近くなっている．つまり，特徴をつかんだ学習ができているといえる．この実験は，学習データの数に比べて，推定するパラメータの数が多い．どこかで過学習が起きていると予想できる．$N = 100$ 以上をみると，図 4 の特徴によく似たグラフが生成された．このことから，$N = 100$ で過学習が起きていると考えられる．

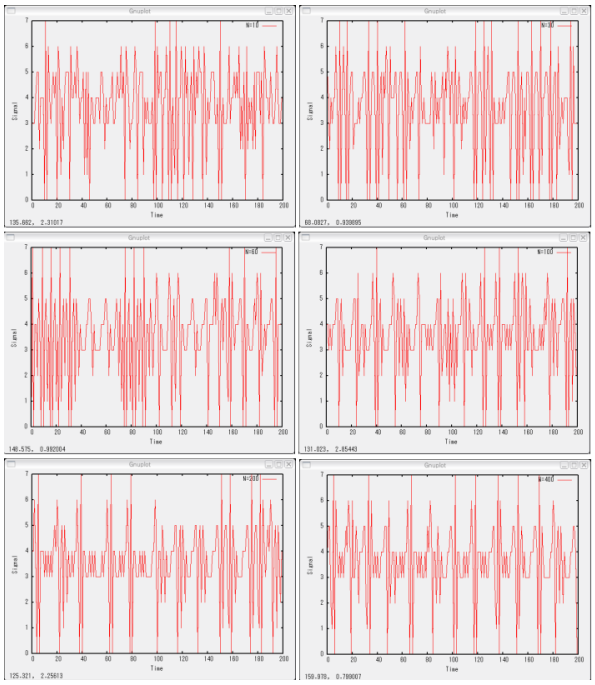

図 5: 左から右の順に，上：$N = 10, 30$，中：$N = 60, 100$ 下：$N = 200, 400$ である．横軸が時間， 縦軸は振幅を表している．

次からは学習に使われる音声信号の数を増やして同様の実験をした．

4.2 データ数の多い音声信号での学習

実験には「あ」「い」「う」「え」「お」の音声信号をそれぞれ量子化ビット数 3（8 値），10kHz でサンプリングして学習データとして使用した．データ数は表 1 の通りである．図 6 は学習データの「あ」と「い」の $t = 0, \ldots, 1000$ の波形を示している．N を 10, 60, 100, 200 として実験した． 結果は図 7，図 8 のようになった．図 6 と図 7，図 8 をそれぞれ比べる．図 7 をみると，学習データのグラフに似たグラフは生成されていない．一方，図 8 では，$N = 10$ 以外は学習データに似たグラフが生成された．これは,「あ」よりも「い」の方が単純な特徴をもつからだと考えられる．次に，生成された音を聞いてみる．どちらの母音もはっきりとは聞こえない．母音は似た波形が繰り返される周期性という特徴をもつ．しかし，生成された音声信号には周期性が見られない．これが，音で聞いたときにはっきりと聞き取れない原因だと考えられる．

表 1: 母音のデータの個数

音声	データ数 (個)
「あ」	2250
「い」	2250
「う」	2008
「え」	2370
「お」	1640

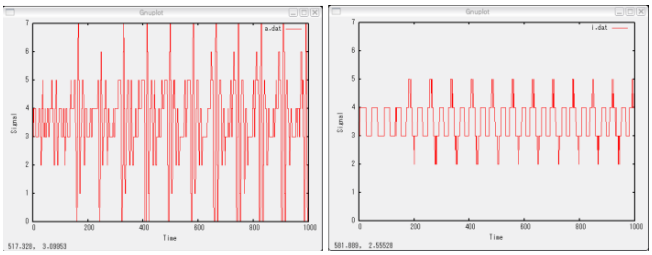

図 6: 左：「あ」，右：「い」それぞれの学習データ [0,1000] の波形．横軸は時間，縦軸は振幅を表している．

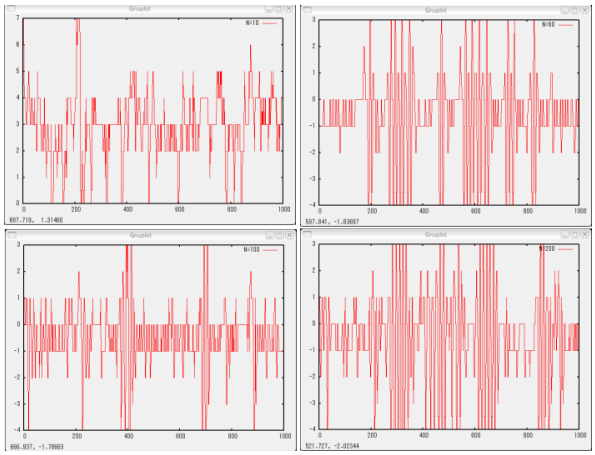

図 7: 「あ」の波形．(上) 左：$N = 10$，右：$N = 60$，(下) 左：$N = 100$，右：$N = 200$．横軸は時間，縦軸は振幅を表している．

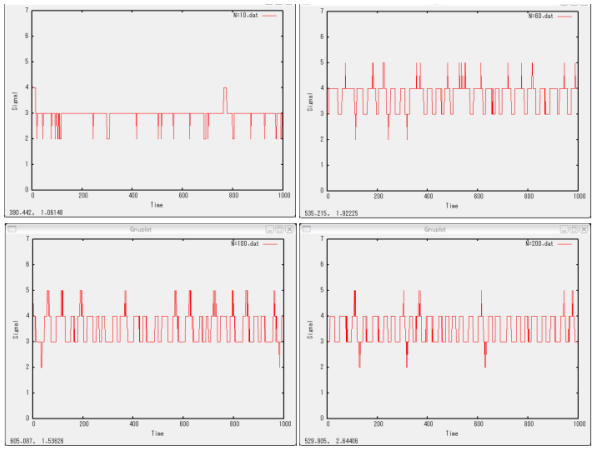

図 8: 「い」の波形．(上) 左：$N = 10$，右：$N = 60$，(下) 左：$N = 100$，右：$N = 200$．横軸は時間，縦軸は振幅を表している．

5　子音との合成

ここでは，学習の結果得られたパラメータを用いて「ま」「み」の音を生成した．母音の学習データは 4.2 節と同じデータを使用した．子音の学習データは,「ま行」を量子化ビット数 3（8 値），10kHz でサンプリングし，子音部分だけを取り出して使用した．$N = 100$ で学習した子音と母音のパラメータをの順に読み込んで音声信号を合成した．図 9 は録音した「ま」と「み」のグラフである．

合成の結果，図 10 の波形ができた．図 9 と図 10 のグラフをみると，どちらも子音部分と母音部分の違いがわかる．しかし，生成された音声信号の子音部分は録音した音声信号の子音部分とは似ていない．これは，学習データとして使用した子音のデータが「ま行」全ての子音部分を集めて学習させたからである．音として聞くと,「ま」とは聞こえないので,「あ」と聞き比べてみた．その結果，聞き分けることができた．合成した子音は音声信号に影響を与えていると考えられる．

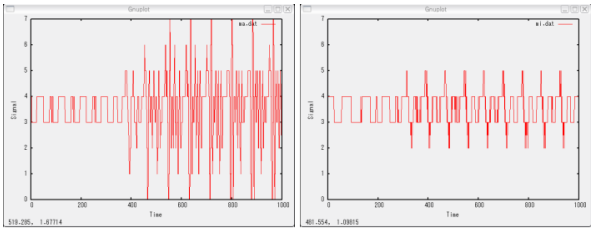

図 9: 録音した (10kHz,3bit) の波形，左：「ま」，右：「み」

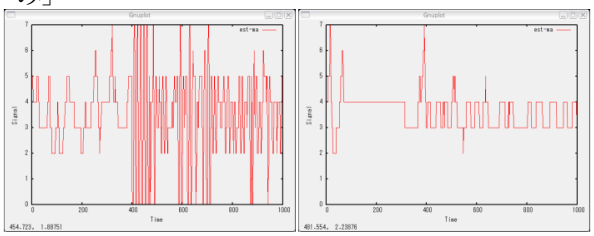

図 10: 合成した波形．左：「ま」，右：「み」．横軸は時間，縦軸は振幅を表している．

6　まとめ

本研究では，音声データ生成機能を持つ隠れマルコフモデルに現実の音声データを与え，モデルを学習した．学習したモデルから生成された音声信号を波形を見るだけでなく実際に聞いたところ，音質は学習データほど良くはなかったが，母音の違いや，子音の有無の違いは判別できる音が生成できた．今回，X から Y は決定論的に定まる関数を用いた．今後，$P(Y|X)$ のモデル化を工夫することで，よりよい音声合成が可能であるかも知れない．

参考文献

[1] H. Künsch, S. Geman, and A. Kehagias: “Hidden markov random fields,” The Annals of Applied Probability, Vol.5, No.3, pp.577-602, 1995.

[2] S. Geman and K. Kochanek. “Dynamic programming and the graphical representation of error-correction codes,” IEEE Trans. Information Theory, vol.47, no.2, pp.549-567, Feb. 2001.