G-11

# ＨＭＭを用いた音声認識

## Speech recognition using Hidden Markov Model

○会沢　純将[1]
Yoshiyuki Aizawa[1]

Abstract: Hidden Markov Model is a useful model for speech recognition. In this paper, the simulation results of connected digit recognition using both Julius and HTK are reported.

### １．概要

現在,音声認識システムの研究では，隠れマルコフモデル(Hidden Markov Model : HMM)が広く用いられている．本研究では, HMM を利用した HTK および Julius２つのツールを使い，数字の連続音声認識を行った．本論文では，その実験結果について報告する．

### ２．音声認識

音声認識とは，音声信号のパターン認識を指す．また一般に，人間は外界から得た情報から，認識に必要なデータを抽出し，脳内のデータベースとパターンマッチングを行い，そのマッチングした情報を，音や文字など意味のある情報として認識していると考えられている．このような人間の脳内で日常的に行われてる情報処理をコンピュータで実現することが，今日に至るまでに一つの大きな目標となっている．

### ３．HMM

HMM は，時間的に変化する性質を持つ問題(音声認識や画像認識)に対して有効な確率モデルであることが広く知られている[1].HMM では，各状態に，次の状態遷移する確率と自己ループする確率，および状態内での特徴ベクトルの確率が学習により決定される．また，一般に音声認識においては，左から右へと遷移していくモデルが扱われる(Figure 1).

さらに ある出力系列が与えられた時に，どのような状態遷移が行われたかが隠されているため，隠れマルコフモデルと呼ばれる[2].

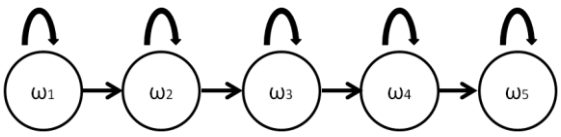


**Figure 1.** HMM における left-to-right 構造

### ４．HTK

HTK(Hidden Markov Model toolkit)とは, HMM の構築，学習，認識，評価などのパターン認識において必要な一連の操作を行うことのできるツール群である[2]．本研究において HTK で使用した主なコマンドは，Table1 に示した 6 個であり，その概略は以下のようである: まず，HCopy コマンドで，音声データから指定したパラメータの MFCC(Mel Frequency Cepstrum Coefficient)を抽出し，また, HInit コマンドにより, Viterbi アルゴリズムを用いて, HMM の初期値を設定する．次に, HRest コマンドにより，HInit コマンドを用いて初期化された HMM に対し，Baum-Welch アルゴリズムによる学習を閾値以下になるまで行う．さらに，文法規則を示したファイルを作り，認識のための HMM ネットワークを HParse コマンドで作成した後，パターン認識を HVite コマンドを用いて行い，最終的に，その結果を HResults コマンドで集計する．

| コマンド名 | 機能 |
|---|---|
| HCopy | MFCC の抽出 |
| HInit | HMM の初期化 |
| HRest | HMM の学習 |
| HParse | ネットワーク表現への変換 |
| HVite | Viterbi アルゴリズムによる認識 |
| HResults | 認識結果の評価，集計 |

**Table 1.** HTK の主なコマンド

### ５．MFCC

一般に，音声認識では,音声特徴量として MFCC が主に利用されている．通常, MFCC はマイクなどにより入力された音声信号をフーリエ変換(FFT)して得られるスペクトル情報に対してフィルタバンクによる分析を行った後，離散コサイン変換(DCT)を行うことで求められる．

１：日大理工・学部・数学

## ６．Julius

Julius とは，音声認識システムの開発，研究のためのオープンソフトウェアである[2]. Julius では，音声認識を行う際に使用する言語モデルとして，N-gram が使用されており，数万語の語彙を対象とした認識にも対応している．

また，音響モデルとしては HMM が使用されるが，前項で述べた HTK で作成した HMM を Julius により使用することも可能である．

## ７．N-gram

現在，音声認識における言語モデルの主流は統計的言語モデルである．その中でも，広く利用されている言語モデルが N-gram モデルである[3].

N-gram とは，文章内における単語の出現確率が，直前の N-1 個の単語にのみ依存すると考え，その確率を求める手法である: 一般に単語列 $\omega_1, \omega_2, \ldots, \omega_n$ に対して，その出現確率を $P(\omega_1, \omega_2, \ldots, \omega_n)$とすると N-gram の推定は次式により計算される．

$$P(\omega_1, \omega_2, \ldots, \omega_n) = \prod_{i=1}^{n} P(\omega_i | \omega_{i-N+1} \ldots \omega_{i-1}) \quad (1)$$

また, Julius では 2 パス探索が採用されており第 1 パスでは 2- gram, 第 2 パスでは 3-gram により探索を行うことで,より精度の高い音声認識が行うことができる．

## ８．予備実験

本研究における予備実験では，HTK による簡単な単語認識実験を行った．「はい」，「いいえ」の２語をそれぞれ５回ずつ録音し, HTK を利用して, HMM の学習を行い，認識を行った．

実験では，それぞれの音声データに,「はい」,「いいえ」,「無音部分」のラベル(音声データ内での区分)を付け，「はい」,「いいえ」,「無音部分」の HMM を初期化および学習し，それを用いて認識実験を行ったところ，認識率は８割となった．ここで，「無音部分」をラベル付けすることなくそれぞれの音声データから切り取り，「はい」,「いいえ」のみで HMM 初期化，学習，および認識を行ったところ，認識率が１０割に向上した．これは HMM の学習の際に認識に不必要な無音部分を切り取ったためだと考えられる．

次に述べる連続数字音声認識実験では HMM の有用性を調べるための実験であるので，本実験における結果を踏まえて，録音した音響データの無音部分を切り取り，学習および認識に両方に使用するものとした．

## ９．連続数字音声認識実験

実験では，１ ~ 9(/ichi/, /ni/, /san/, /yon/, /go/, /roku/, /nana/, /hachi/, /kyu/)の数字を組み合わせた 21 通りの日本語数字列(Table2)を考慮した．また，各数字列は３つ Table2 に示されるような３つの数字単語からなり，それぞれ５人の話者により発音・録音された 15 パターンを実験データとして用いた．その内，10 パターンを学習に使用し，残りの 5 パターンを本実験の評価データとした．

本研究において，前述の予備実験同様, HMM を利用した音声認識を行うために，HTK と Julius の両方を用い，HTK および Julius の認識結果を比較する予定である．また，言語モデルによる認識結果の比較と HMM による連続音声認識の有用性について考察する予定である．

以上述べた連続音声実験の結果について，学術講演会当日に報告する予定である．

| | | | | | | |
|---|---|---|---|---|---|---|
| 118 | 123 | 147 | 222 | 239 | 355 | 369 |
| 416 | 451 | 456 | 573 | 642 | 645 | 745 |
| 789 | 867 | 871 | 898 | 932 | 967 | 983 |

**Table 2.** 数字列一覧

## １０．　参考文献

[1] Richard O. Duda , Peter E. Hart , and David G. Stork : “Pattern Classification Second Edition”, John Wiley & Sons , April 2009, pp125.

[2] 荒木雅弘 :「フリーソフトでつくる音声認識システム」, 森北出版株式会社, October 2007, pp.22-23,130-147.

[3] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄 :「音声認識システム」, 株式会社オーム社, October 2006, pp.22-23.