P-13

HMM を用いた数字音声認識
Spoken digit recognition using Hidden Markov Model

○池田　早希，島田　雅弘[1]
Saki Ikeda，Masahiro Shimada[1]

Abstract: We applied a de-fact standard approach of HMM (Hidden Markov Model) to speech recognition and performed a simulation study using HTK.　This report focuses up on the study of spoken digit recognition.

1.　概要

現在，音声認識は様々な分野で利用されている．一例として，近年の携帯電話器では文字を打たずに音を聞き取って自動的に打つ，また，雑音や聞き取りにくくても正確に聞き取る，ということなども可能となっている．

本論文では音声認識でよく用いられる HMM(Hidden Markov Model；隠れマルコフモデル）を用いて 6 人の話者から得られた数字音声サンプルを用い，その数字音声認識実験を行った結果について報告する．

2.　音声認識

音声認識を行う際にまず，発話者から得られた各音声サンプルについて特徴抽出を行う．その特徴抽出データとして通常 MFCC(Mel-Frequency Cepstrum Coefficients)が用いられる．

次に，その特徴抽出されたデータを用い HMM の学習を行う．音声認識は，学習済みの HMM と評価用の入力音声を用いて行う．

3.　HMM

Hidden Markov Model は時間と共に確率が変化するようなデータのパターン認識を行うことに有効な確率モデルであると言われている[1]．また，音声認識を行う HMM として状態が左から右へと遷移するモデルが一般的である．

なお，出力系列が与えられたときにどのような状態遷移が行われてきたかが隠れているので通常「隠れ」マルコフモデルというように呼ばれる．本研究ではこの HMM を用いて実験を行った．

4.　MFCC

MFCC は音声認識の音声の特徴量として広く利用されている．

MFCC を得る手順として，まず音声を低周波成分と高周波成分に分け，情報量が多い低周波成分を解析する[1]．その与えられた音声信号をディジタル化する．次にその高速フーリエ変換を行う．高速フーリエ変換は録音された音声の周波数変換を行う手法として広く用いられている．そしてメル帯域化を行った後に対数をとり最終的に特徴が抽出される[1].

5.　HTK

HTK(Hidden Markov Model Toolkit)[2]は HMM を使用する際のツールキットである．

これを用いることで，HMM の構築・学習・認識・評価といった一連の操作を行う[1].

Table 1 は本研究で使用したコマンドの一覧である．また，これとは別に，実験では録音された音声データとして wav データを用いたので，あらかじめ wave surfer[3]プログラムを用いラベル付けを行った．ラベル付けは，具体的には録音した音声の何秒から何秒までの間にどのデータが入っているのかの情報を付与することである．

次に HCopy コマンドを用い音声の特徴抽出(MFCC)を行った後に，実験に用いるデータを得る．また，HMM の初期構成を決定する際に HInit コマンドを用いる．

そして HRest コマンドを用い，HMM の各パラメータの繰り返し学習を行う．学習は，Baum-Welch アルゴリズムにより更新の値がある閾値より小さくなったときに終了する．このような HMM パラメータの学習後に単語認識を行う．

単語認識を行う前に，あらかじめ HParse コマンドを用い HMM をネットワーク表現に変換する．そのネットワーク表現を用い認識および評価を行う．

認識にはビタビアルゴリズムを実装した HVite コマ

1：日大理工・学部・数学

ンドを用いる.

ビタビアルゴリズムは各状態における確率の最大値を求めその最大値を与える経路の情報を保存するものである．また，認識結果の表示には HResults コマンドが用いられる.

コマンド名	機能
HCopy	特徴抽出
HInit	HMMの初期化
HRest	HMMの学習
HParse	ネットワーク表現に変換
HVite	ビタビアルゴリズムによる認識
HResults	認識結果の集計

Table 1．主要な HTK のコマンドと機能

6．認識実験

まず最初に予備実験として単数字 0~9 (/zero/，/ichi/，/ni/，/san/，/yon/，/go/，/roku/，/nana/，/hachi/，/kyuu/，/sil/)を一人の話者により 3 回ずつ録音した．ここで，sil は無音区間のことを指し，ラベル付けは「無音区間+数字+無音区間」のように行った．予備実験では合計 30 個の録音データを用意し，HMM の初期化をし学習を行った後にその学習に用いたデータと同じものを全て認識させた.

次に, 本実験として計 6 人(男性 3 人女性 3 人)の話者による単数字音声認識を行った．その際，1 人につき各数字に対して 10 個の録音データ(つまり 1 人につき合計 100 データを用意し 6 人で 600 の録音データ）を用意した. そしてこのデータのまず半分を用いて HMM の初期化をし学習を行い，残りの半分を認識データとして用い実験を行った.

7.　　実験結果および考察

Table2 に実験結果を示す.

予備実験では 90．00%の認識率が得られた．このような高い認識率が得られたのは実験に用いたデータが一人の話者のみによるものでもあり，学習させたデータをそのまま認識に用いたためであると考えられる.

本実験では予備実験より低い結果となってしまった. 本実験では多くのデータを学習させたほうが認識率が高い結果になると見込んでいたものの，予備実験とは異なり 6 人の話者から得た学習データとは別のデータを認識データとして用いたことが認識率の低下を招いた原因であると考えられる．さらに，実験に用いた録音データには男性および女性のデータが混在していることや，発話の仕方が各話者間で異なっていたことも一因であったと考えられる.

他に認識率低下の原因について，主に録音時の音量および雑音の混入などといったことも考えられる。

以上の実験結果から，認識率向上においてラベル付けが一つのポイントであると思われる．前述のように録音データは「無音区間+数字+無音区間」という形式に則ってラベル付けを行ったため，本来無音区間としてラベル付けされるべき部分に音声部が混入してしまうことやその逆も考えられ，それが認識率に大きく影響したとみられる.

なお，本実験においては 72.09%という結果が得られたが，その詳細については学術講演会当日に報告する予定である.

	予備実験	本実験
HMM	90.00%	72.09%

Table 2．実験結果

8．参考文献

[1]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，pp.151-171，2017 年 4 月.

[2]S. Young , G. Evermann， D. Kershaw， G. Moore, J. Odell， D. Ollason， D. Povey, V. Valtchev， P. Woodland :「The HTK Book」2002 年 12 月.

[3] http://www.speech.kth.se/wavesurfer/ : (wave surfer のダウンロード先)