G-27

HTK を用いた音声認識

Speech recognition using hidden Markov model toolkit

○横井優樹[1]
*Yuki Yokoi[1],

Abstract: Hidden Markov model toolkit is a toolkit for speech recognition. In this study, the effect of using correct class labels (as well as the transcriptions each describing both speech onset and duration) upon the recognition performance is investigated via simulations.

１．はじめに

本研究では，Hidden Markov Model Toolkit (HTK)を用いて，HMM モデルを生成し，孤立単語音声認識および連続数字音声認識の実験を行った．本論文では，時間詳細付ラベルの有無によって生じる認識率の変化を考察し，時間情報の有効性について報告する．

２．音声認識

音声認識とは，人間の発する音声をコンピューター上で認識させ，文字に変換して表示することや，発した声の特徴を抽出し，発話者を分析することである．最近では，スマートフォンに音声認識機能や音声認識を活用し音声を分析する事にも用いられており，様々な場面で実用化されている．

３．Hidden Markov Model (HMM)

HMM は，音声認識に対して有効な確率モデルであり，現在主流となっている音声認識モデルの１つである．HMM は，各状態に次の状態に遷移する確率（自己ループを含む）とそれぞれの特徴ベクトルの確率を学習することによって求められる[1].

また，ある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，隠れマルコフモデル（HMM）と呼ばれている．

HMM の確率計算においては，ビタビアルゴリズムが用いられる．ビタビアルゴリズムは，ある時点での推移で確率の高くなる方を採択し，それ以外の計算を打ち切るものである．HMM の学習においては，通常 Baum-Welch アルゴリズムが用いられ，学習過程は初期のパラメータを適当に設定した後，与えられた学習データの統計的な性質を反映し，繰り返し計算により調整を行うようなアルゴリズムである[1].

また．連結学習を用いることで，認識精度を向上させたい場合やラベルの時間詳細情報が無い場合にも，繰り返し学習により，統計的な性質を反映できる．HMM の各状態に出力確率分布を定義するには通常 GMM(Gaussian Mixture Model)が用いられる．これにより、複数の正規分布の出力の重みづけ和により定めることができ、HMM モデルの精度を向上を図る．

４．Mel-Frequency Cepstrum Coefficients (MFCC)

音声認識の音声特徴量として主に使用されているのが MFCC である．一般的に，音声信号を FFT（高速フーリエ変換）して求められる振幅スペクトル情報をメルフィルタバンクにより圧縮し，DCT（離散コサイン変換）により求められるケプストラム成分が MFCC である．

５．Hidden Markov Model Toolkit（HTK）

HTK は，HMM の構築，学習，認識，評価をする為に必要なツール群である[1]．本研究で使用したコマンドは，Table 1 の通りである．

コマンド名	機能
HCopy	特徴抽出
HInit	HMM の初期化
HRest	HMM の学習
HCompV	HMM の初期化（ラベル無）
HERest	連結学習
HHEd	混合数増加
HParse	文法記述を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　　HTK の基本コマンド

１：日大理工・学部・数学

６．実験手順（１）

まず，録音した音声を正解ラベル付けをした後，HCopy コマンドで特徴を抽出し，状態数，状態遷移，混合数を決めて HMM 構成情報ファイルを作成する．次に, HInit コマンドで MFCC から HMM の初期値を求め，HRest コマンドにより，初期化された HMM を Baum-Welch アルゴリズムで変化量が閾値以下になるまで学習を繰り返す．そして，HParse コマンドにより HMM のネットワーク形式に変換し，HVite コマンドを用いて認識を行う．最後に，HResults コマンドで評価を行う[1].

７．実験手順（２）

まず，連結学習を用いてモノフォンモデルを作成する．その際，HCompV コマンドにより，MFCC を用いて平均ベクトルおよび分散共分散行列を計算し初期値を求め，HERest コマンドを数回繰り返し連結学習を行う．次に，HHEd コマンドで連結学習したモデルの混合数を上げ, さらに HERest により連結学習をし, HHEd で混合数を上げる．その後は，実験手順（１）と同様に認識，評価を行う[2].

本研究では，連結学習を１０回行い，混合数を４まで増加させて実験を行った.

８．検証実験

本研究では，まず検証実験として不特定話者による孤立単音認識実験を行った．「nara」,「yamagata」の２単語を５人の話者から１回ずつ録音し，その内の４つの音声を HMM の学習データに使用し，１つの音声を評価データに使用した.

実験では録音した音声に，「無音(sil)」,「nara」,「yamagata」の時間情報を含む正解ラベルを付けた場合の実験(検証実験(1))と，時間詳細ラベルを含めない場合の実験(検証実験(2))を行い認識率の変化を比較する．実験手順（１），（２）の流れを Figure.1 で示す.

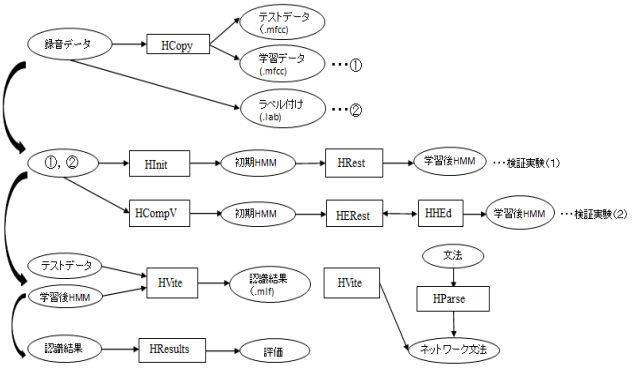

Figure 1　　実験手順の流れ

９．検証実験結果

検証実験（２）を行った結果，認識率は 90%であった．一方，検証実験（１）を行った時の認識率は 100%となった。これは、「無音(sil)」の部分を考えずに音声を認識しているからだと考えられる。

以上，検証実験（１）と（２）を比較をすると，微差ではあるが時間詳細付きの正解ラベルの有無によって認識率が変化しているため，時間詳細を含むラベル付けには有効性があるということがわかる．次に示す連続数字音声認識実験では，検証実験と同様に時間詳細付きラベルの有効性について考察するものである.

１０．連続数字音声認識実験

本実験では、Table2 で示した 21 通りの連続音声を５人の話者からそれぞれ 5 パターン用意し，実験データとして用いた．その内の４パターンを HMM の学習データとし，１パターンをテストデータとした．尚，この実験では不特定話者による連続数字音声認識で，時間詳細付き正解ラベルの有無における認識結果の変化について検証する.

以上述べた連続数字音声認識の実験結果については，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 連続数字音声認識実験に用いる数字列モデル

１１．参考文献

[1]荒木雅弘:「フリーソフトでつくる音声認識システム」,森川出版株式会社,October 2007,pp.22, pp.112-127, pp.130-147.

[2]土屋雅稔,山本一公:「音声言語処理と自然言語処理 演習」コロナ社,March 2013, pp.15-pp.28.

[3]会沢純将:「HMM を用いた音声認識」,日本大学理工学部学術講演会論文集(2013).