G-11

HTK を用いた楽器音の連続音声認識

Sound recognition of musical instruments using Hidden Markov Model Toolkit

○佐藤　淳[1]
Atsushi Satou[1]

Abstract: Hidden Markov model is a statistical method which is most widely used for speech recognition. This paper reports an experimental study for the sound recognition of musical instruments.

１．概要

現在，スマートフォンやタブレットに音声コントロール等の音声認識を利用した技術が広く用いられている．本研究では，音響モデルを作成するうえで主流となっている隠れマルコフモデル(Hidden Markov Model, HMM)という確率モデルを用いて音響モデルを作成し，楽器音の連続音声認識を行った．その結果について報告する．

２．HMM とは

隠れマルコフモデル(Hidden Markov Model ; HMM)とは，音響モデルを作成するうえで，現在主流である音の時系列が確定していない場合に対して有効的な確率モデルである．時系列の長さを定めるために，各オートマトンの状態が自身に戻る(自己ループ)できるように，Figure1 で示すように状態遷移することで，任意の特徴ベクトル系列に対して確率を求めることができる．HMM の学習は, Baum-Welch アルゴリズムにより HMM のパラメータの変化量を閾値以下になるまで実行して行う．一方，HMM を用いた音声認識では，認識において Vitervi アルゴリズムを用いて，状態遷移系列の確率計算を行う．

また，HMM の特徴として，入出力系列が観測されても, 出力系列を生成する状態系列は複数通り考えられ, 状態遷移の様子を一意に定めることができない(隠れている)ために，「隠れマルコフモデル」という名称がつけられている[1].

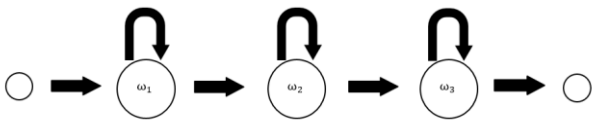

Figure1．HMM の構成

３．HTK と MFCC

Hidden Markov Model Toolkit(HTK)とは，HMM の構築，学習，認識，評価などを行うためのツールキットである[2]．本研究で用いた HTK コマンドは，Table1 で示したものである．

各コマンドでは，HSlab で音声を録音，ラベル付けを行い，HCopy で音声の特徴抽出をする．学習の音響モデルを作成するには, HInit で HMM の初期値を求め, HRest で Baum-Welch アルゴリズムで HMM の学習を行う．次に，HParse により認識のための文法ネットワークを作成し，HVite により Vitervi アルゴリズムによる近似計算を行い認識を行い, HResult で認識結果を求める．

また，本研究では，音声データから抽出する特徴パラメータの特徴量として，メル周波数ケプストラム(Mel Frequency Cepstral Coefficient ; MFCC)を用いた．MFCC とは，現在音声認識で主流となっている特徴量である．MFCC を求める手順として、フーリエ変換により求めたスペクトル情報を，人間の聴覚特性に合わせたフィルタを通して対数変換し，離散コサイン変換したものである[2].

コマンド	機能
HSlab	音声の録音、ラベル付け
HCopy	特徴抽出
HInit	HMM の初期化
HRest	HMM の学習
HParse	文法記述をネットワーク表現に変換
HVite	Viterbi アルゴリズムによる認識
HResult	認識結果の集計

Table1．HTK の基本コマンドとその役割

１：日大理工・学部・数学

４．認識実験 1

本研究の実験として、ド(do)，レ(re)，ミ(mi)の 3 つの音声を用いて 3 音の連続音声認識を行った．(音階はド(C5)，レ(D5)，ミ(E5)とした．)実験では，電子キーボードを用いて各音声ファイルのサンプリング周波数を 1/16000Hz として録音を行い，実験データとして Table3 の音声ファイルを準備した. (認識対象は、ド(do), レ(re)，ミ(mi)，無音空間(sil)の 4 つ．)Table2 の各音声ファイルを学習用データ，実験用データ(各データの数はド(do)27 個，レ(re)27 個，ミ(mi)27 個．)それぞれ用意し認識実験を行い，出力結果の認識率を四捨五入により小数点第 3 位までの値を求めた.

do-do-do	do-do-re	do-do-mi	do-re-do
do-re-re	do-re-mi	do-mi-do	do-mi-re
do-mi-mi	re-do-do	re-do-re	re-do-mi
re-re-do	re-re-re	re-re-mi	re-mi-do
re-mi-re	re-mi-mi	mi-do-do	mi-do-re
mi-do-mi	mi-re-do	mi-re-re	mi-re-mi
mi-mi-do	mi-mi-re	mi-mi-mi	

Table2．学習データ一覧

５．実験結果 1

Table3 に実験結果を示す. 実験では，ミ(mi)の認識率が高かったのに対し，ド(do)は 3.704%, レ(re)は 22.222%と低い認識率となった．また，ド(do)やレ(re)の出力結果がミ(mi)に誤認識されることが多かった．誤認識した原因として，学習データの個数が不十分であったと考えられ，また，学習データの不足により，ド(do)とレ(re)のフレーズの区別がつきにくくなり，出力結果がミ(mi)に偏ってしまったと考えられる.

入力＼出力	do	re	mi
do	3.704	11.111	85.185
re	0	22.222	77.778
mi	18.519	3.704	77.778

Table3．認識実験 1 の各フレーズの識別結果(%)

６．認識実験 2

第 5 章の実験手順と同様に，Table2 の学習用データをもう 1 組用意し，ド(do)54 個，レ(re)54 個，ミ(mi)54 個用意し HMM の学習を行い，再度認識実験を行った．次章では，認識実験 2 の結果について考察し，認識実験 1 のデータと比較する.

７．実験結果 2

認識実験の結果を Table4 に示す．この場合，ミ(mi)は 80%以上，レ(re)は 70%以上の認識結果となった．よって，認識実験 1 のデータと比較して全体的に認識率が上昇したことがわかる．また，認識実験 1 より誤認識が少なくなり，認識実験 1 で生じたド(do)とレ(re)のミ(mi)への誤認識も少なくなった．認識率が上昇した理由として，学習データの数を増やしたことで，認識率が向上したと考えられる．以上の実験から，学習データを任意の数増やすことで，認識率が上昇することがわかった.

入力＼出力	do	re	mi
do	51.852	33.333	14.815
re	0	70.370	29.630
mi	14.815	3.704	81.481

Table4．認識実験 2 の各フレーズの認識率(%)

８．終わりに

今回の実験から，学習用のデータを増やすことで，認識率を上げることができると考えられる.

前章の実験結果を踏まえ，学術講演会当日では，さらに学習用のデータを充実させ，その場合における認識率について発表する予定である．また，ファ(fa)，ソ(so)，ラ(ra)，シ(si)と認識対象の数を増やすことによって認識率を向上させた音響モデルを作成する予定である．(音階はファ(F5)，ソ(G5)，ラ(A5)，シ(B5)とする．)また，簡単な演奏を認識させ，認識結果についても学術講演会で発表する予定である.

９．参考文献

[1] 中川聖一：「音声言語処理と自然言語処理」，コロナ社，pp.37~41，2013 年 3 月.

[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，pp.130~147，2007 年 10 月.