P-3

県名データベースを用いた場合の CNC による連続単語音声認識

Connected Spoken Word Recognition Using Cascaded Neuro-Computational Model and Japanese Prefectures Database

○佐藤 淳[1]，　保谷 哲也[2]

Atsushi Satoh[1]，　Tetsuya Hoya[2]

Abstract: CNC(Cascaded Neuro-Computational model) is a neural network model that adopts a psychological approach. In this paper, we report simulation results using a Dataset for continuously spoken Japanese prefectures and compare with those obtained using HMM.

１．概要

音声認識の主流なモデルの１つとして, HMM(Hidden Markov Model)によるベイズ決定規則に基づいた確率モデルが挙げられる. 本研究では, CNC(Cascaded Neuro Computational Model)と呼ばれる心理学的アプローチを取り入れたニューラルネットワークモデルを用いて実験を行う．CNC は連続数字音声認識のようにクラス数が少ない場合，HMM と同等の認識結果が得られることが報告されている[1]．本研究では，47 都道府県が発話されたデータセットを用いた場合における連続単語音声認識実験により CNC と HMM の比較を行った．

２．音声認識

音声認識は基本的に以下の手順で行われる．まず，入力された音声情報から，認識に必要な情報を抽出する特徴抽出を行う．その特徴抽出により得られた情報を用いて，学習モデルを形成し認識器を形成する．認識の際には，入力データと学習されたデータベースとのマッチングを行い，類似度の最も高い結果を認識結果として出力する．

本研究では，音声特徴量として MFCC(Mel Flequency Cepstral Cofficiency)を用いて実験を行った. MFCC とは, 音声信号に高速フーリエ変換(First Fourier Transform, FFT)により得られた振幅スペクトルに対し，メルフィルタバンクをかけ圧縮した後, 離散コサイン変換(DCT)により得られたケプストラム成分である．

３．CNC(Cascaded Neuro-Computational Model)

CNC(Cascaded Neuro-Computational Model)とは，脳内の神経細胞の発火に基づいたニューラルネットワークモデルの一つである[2]．CNC は Figure 1 で示されるように 3 層から構成されており，Layer1(L1)では，入力された MFCC データをフレーム毎に(1)式の RBF(Radial Basis Function)によりユニットを構築する．

$$h_i(x) = exp\left(-\frac{\|x - c_j\|_2^2}{\sigma^2}\right) \qquad (1)$$

学習の際は第１層のユニットから発火されたデータはユニットに追加せず，発火されなかったデータを新たに第１層ユニットとして追加する．これを全ての学習パターンについて行う．また，連続単語音声認識において L1 では音声区間と非音声区間の音声情報を区別する必要がある．非音声区間を検出する際には，音声データの先頭と末尾にある非音声区間を検出し，L1 ユニットと発火したユニットを L1S ユニットとして再定義する[1]．Layer2(L2)学習の際は，L1 ユニット内の最大発火されたユニットのラベル情報のみを L2 に出力する．L2 では L1 の出力結果から，各単語毎の時系列順のラベル情報をユニットに格納する．次に非音声区間の L1 ユニットのラベル情報を 1 フレームまで削減し，音声区間内で単語の境界決定を行う．その際，各単語の境界決定は非音声区間ではなく，同等のテンポにより話されたものとして単語の境界決定を行う[1]．一方，認識では，入力されたフレームデータから順に孤立単語音声認識の手順で行う．その際，L1 で発火されたラベル情報を随時 L2 に送り，L2 ユニットで発火されたものから順に逐一出力する．これらを繰り返すことで，連続して音声パターンの分類を行うことができる. Layer3(L3)では, L2 の出力を各単語毎に集計し，最終的な認識結果として出力する．

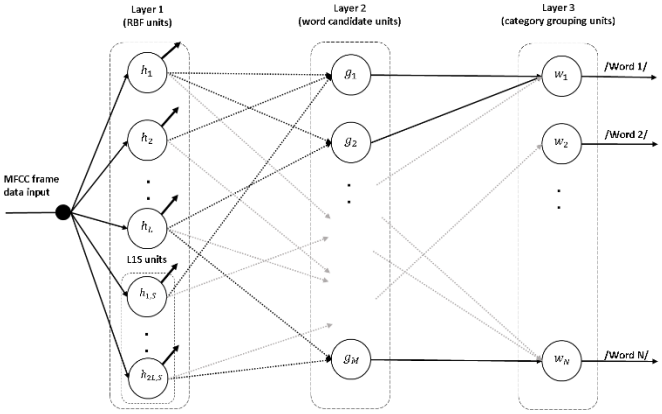

Figure 1. Cascaded Neuro-Computational Model

１：日大理工・院(前)・数学　２：日大理工・教員・数学

４．HMM(Hidden Markov Model)

HMM(Hidden Markov Model)は音声認識を行うために用いられる，現在主流なモデルの 1 つである．HMM では音声の時系列の長さを定めるために，各オートマトンの自己遷移を考慮することで，任意の特徴ベクトル系列に対し確率を求めることができ，時間変動など特徴系列の長さが定まっていないデータに対し有効な手段の一つである[3]．HMM の学習は，Baum-Welch アルゴリズムにより HMM のパラメータ変化量を閾値以下になるまで繰り返すことによって行われる．一方，認識時には，Vitervi アルゴリズムを用い状態遷移系列の確率計算で最大尤度で示された経路を出力結果とする．

本研究では HTK(Hidden Markov model Toolkit)を用いて HMM の学習および認識を行った．

５．連続単語音声認識実験

本研究では, 特定話者における CNC と HMM の連続単語音声認識実験を行った．その際，10 名(男性 5 名，女性 5 名)の話者から，47 都道府県名(/Hokkaido/ ~ /Okinawa/)を 2~7 音連続で発話した 47 通りの組み合わせを考慮したデータベースを用いた．47 通りの組み合わせは，各クラスの発話回数に偏りが生じないように定めた．実験用データは，各話者で 10 回ずつ録音したものを使用し, その内, 8 個をトレーニングデータとして, 残り 2 個をテストデータとして使用した(Dataset 1).

CNC では Theta および Radius の 2 つのパラメータ調整が必要であるが, それぞれ, $0.7 \leq \theta \leq 0.8$，$4.5 \leq r \leq 5.5$の範囲内の組み合わせで実験を行った．

HMM では無音区間(sil)をクラスに含む monophone モデルを構築し, 連結学習により繰り返し学習した後，混合数を上昇させることを繰り返した．連結学習における繰り返し学習回数はそれぞれ 10 回，また，混合数を$1 \rightarrow 2 \rightarrow 4$と増加させた．実験結果では，混合数 4 で連結学習を 10 回行ったモデルを用いた場合のスコアを最終的な認識結果とした．また，ラベル総数における削除数，置換数の割合を削除率(%)，置換率(%)として誤りの遷移についても検証を行った．

６．実験結果①

Dataset 1 を用いた場合の CNC と HMM の実験結果を Table 1 に示す．実験結果では CNC の認識精度が HMM より約 7%低くなった．また，CNC の置換率は HMM と比較して 3 倍以上となった. 原因として, CNC では，入力されたフレームを順番に認識するため，Dataset 1 におけるフレーム数が長いパターン(/Hokkaido/)と短いパターン(/Chiba/，/Gifu/，/Mie/，/Shiga/，/Nara/，/Saga/)における認識率が他単語と比較し低くなったため，CNC では単語毎に正確に学習ができていないと考えられる．次章では，CNC の認識精度が低くなった原因について検証実験を行い,その結果について報告する．

	Acc(%)	削除率(%)	置換率(%)
CNC θ =0.7, r=4.7	68.87	4.48	24.65
HMM	75.42	5.66	7.19

Table 1．CNC と HMM の認識精度(Dataset 1)

７．検証実験

実験では，フレーム長の長さおよび類似したクラスが含まれたパターンが CNC の認識精度に影響するのかどうかを検証する．その際，Dataset 1 の中から男性 2 名を選択し，実験を行った(Dataset 2)．また，Dataset 2 の中からフレーム長が長いクラス(/Hokkaido/)と短いクラス(/Chiba/，/Gifu/，/Mie/，/Shiga/，/Nara/，/Saga/)，類似した単語[Yama]を含むクラス(/Yamagata/，/Yamanashi/，/Toyama/，/Wakayama/，/Okayama/，/Yamaguchi/)を除外したデータを用いて実験を行った(Dataset 3). Dataset 2 と Dataset 3 における CNC と HMM の比較実験を行った．

８．実験結果②

実験結果を Table 2 に示す．実験結果から，Dataset 2 の CNC モデルは Dataset 1 のモデルと比較して，認識精度が約 6%向上し, 削除率および挿入率についても減少した．以上のことを踏まえ，CNC では，①単語長の検出，②類似単語の分類の改善を行う必要があると考えられる. 今後は, 様々なデータセットを用いて, CNC の認識精度が下がる原因について探る予定である．

	Acc(%)	削除率(%)	置換率(%)
CNC(Dataset2) θ=0.7, r=4.9	75.00	6.25	16.98
CNC(Dataset3) θ=0.7, r=4.9	81.57	4.21	11.70
HMM (Dataset2)	95.27	0.66	0.08
HMM (Dataset3)	89.25	0.66	0.33

Table 2．Dataset 2 と Dataset 3 における実験結果

９．参考文献

[1] Tetsuya hoya and Cees van Leeuwen，“Connected Word Recognition Using a Cascaded Neuro-Computational Model,” Connection Science，pp.1-14，Aug. 2016.

[2] Tetsuya Hoya and Cees van Leeuwen, “A cascaded neuro-computational model for spoken word recognition,” Connection Science，vol.22，pp.87-101，Feb. 2010.

[3] 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．