P-13

県名データセットを用いた修正版 CNC による連続単語音声認識

Connected Spoken Word Recognition Using a Modified Cascaded Neuro-Computational Model and Japanese Prefectures Dataset

○清水秀樹[1]，保谷哲也[2]
*Hideki Shimizu[1], Tetsuya Hoya[2]

Abstract: Cascaded neuro-computational model(CNC) is a neural network model that seeks some psychological plausibility of human speech recognition, while processing real speech input. In this paper, we propose a modified CNC and report the simulation results using a speaker-independent dataset of connected Japanese prefectures names.

１．概要

音声認識の主流モデルとして，統計的手法である HMM(Hidden Markov Model)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に基づいたモデルの修正版を新たに提案し，それを用いて連続音声認識実験を行うことが目的である．また，実験では 47 都道府県名を複数の不特定話者により連続発話された 47 都道府県名のデータセットを用い，新たに提案された修正版 CNC と HMM との比較を行う．

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，それを文字に変換する技術である．音声認識は以下の手順で行われる．まず，入力された音声データから，認識に用いる特徴抽出量の抽出を行う．次に，その特徴量を用いて，学習モデルと認識器の生成を行う．認識の際には，入力データと学習済みのデータを用いてマッチングを行い，最も高い類似度の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 Frequency Cepstral Coefficients)を用いる．MFCC は与えられた音声信号に対し FFT(高速フーリエ変換)を行うことにより得られた振幅スペクトル対してメル帯域化，および，その対数をとった後に DCT(離散コサイン変換)を行い，ケプストラム係数を求め，さらに，ケプストラムに高次の情報を削除するリフタリングを行い，求められた低次の情報のことである [1].

３．CNC(Cascaded Neuro-Computational Model)

CNC(Cascaded Neuro-Computational Model)は，脳内の神経構造を基にした人工ニューラルネットワークである[2]．CNC は Figure 1 に示されるように Layer1(L1)は RBF ユニット，Layer2(L2)は単語候補ユニット，Layer3(L3)は単語ユニットから構成される 3 層構造を成す．

L1 では，MFCC 学習データがフレーム毎に入力として提示された際，(1)式で示される RBF(Radial Basis Function)におけるセントロイドベクトル c_i として適宜与えられる．つまり，L1 の学習はユニットの追加により行われる．

$$h_i(x) = exp\left(-\frac{\|x-c_i\|^2}{\sigma^2}\right) \quad (1)$$

L2 では，各フレームを L1 ユニットに入力した際に得られる最大発火されたユニットの時系列順のユニット番号等のラベル情報が各ユニットに入力される．また，各 L2 ユニットにてラベル情報の類似度を測定し，単語候補の出力を得る．L3 では L2 からの出力を各単語毎に集計し、最終的に孤立単語音声認識結果として出力される．

CNC を用いた連続単語音声認識では，L1 にて非音声区間を検出し，音声区間と非音声区間の区別を行う．L1 ユニットのうち，非音声区間において発火したユニットを L1S ユニットとし再定義する．L1S ユニットは，音声区間内で単語の境界を定めるのに用いられる[2]．L2 では，入力データに対し L2 ユニット内に保持されるテンプレートデータの長さに応じた各単語における類似度の測定が行われる．

上記の CNC を用い，計 47 クラスからなる県名データセットを用いた特定話者による連続単語音声認識において，HMM と同等の認識結果が報告されている[3].

１：日大理工・院（前）・数学　２：日大理工・教員・数学

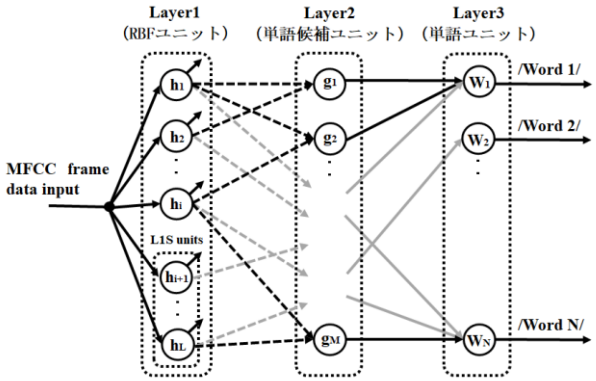

Figure 1.　Cascaded Neuro Computational Model

４．不特定話者県名データセットを用いた連続単語音声認識実験

本研究では，不特定話者県名データセットを用いた場合における CNC と HMM の連続単語音声認識実験を行った．その際，47 都道府県名(/Hokkaido/ ~ /Okinawa/)を 2~7 個連続で発話し，各県名について発話回数の偏りが生じないように 47 通りに組み合わせた単語列を用いる．発話者 10 名(男性 5 名，女性 5 名)により，各単語列 10 回ずつ録音したものの内，学習データ，テストデータとしてそれぞれ 4 個ずつ用い，学習話者を 8 名(男性 4 名，女性 4 名)とし，残りの 2 名(男性 1 名，女性 1 名)をテスト話者とした．

CNC[3]では$r1$(Radius 1)および$r2$(Radius 2)のパラメータ調整が必要である．それぞれ，$2 \leq r1 \leq 7$，$0 \leq r2 \leq 1$の範囲内で組み合わせ，CNC と HMM の比較実験を行った．また，ラベル総数における削除数，置換数，挿入数の割合を削除率(%)，置換率(%)，挿入率(%)とし誤認識について検討した．

５．実験結果

CNC と HMM の比較実験結果を Table 1 に示す．実験結果から CNC の認識精度は HMM の半分以下となり，置換数がラベル総数の約半分となった．これは，置換率から，L2 の認識にてテンプレートデータ，入力データそれぞれのラベル情報とユニットの番号が不一致のものが多かったことを示している．次章では以上を踏まえて，修正版 CNC を提案する．

	Acc(%)	削除率(%)	置換率(%)	挿入率(%)
CNC $r1 = 6.0\ r2 = 0.3$	45.52	4.66	44.04	5.78
HMM	97.64	0.35	1.89	0.12

Table 1.　CNC と HMM の認識精度

６．修正版 CNC

前章における実験結果の考察により，本研究では新たに修正版 CNC として次のように提案する：文献[3]で提案されたフレーム毎に最大発火した L1 ユニットのラベルから成る L2 への時系列入力データと各 L2 ユニット内のテンプレートデータとの編集距離を計算し L2 ユニットの出力を得るのではなく、(2)式で示されるような, テンプレートデータの各要素で示される L1 ユニットの発火値を基に出力を得る：

$$g_j = \frac{1}{N_{g_j}} \sum_{k=1}^{N_{g_j}} h_{t_j(k)}(x(k)) \qquad (2)$$

ここで，L2 において，テンプレートデータおよび入力データのフレームの長さが異なる場合がある．例えば，入力データがテンプレートデータのフレームの長さよりも大きい場合，入力データに対しテンプレートデータの長さに応じた各単語における類似度の測定が行われるため，境界を用い定めた単語のフレーム区間のうち，後部のフレーム区間が計算されなくなってしまう．そのため，(3)式のようにテンプレートデータtおよび入力データxの長さが異なる場合において，短い方に正規化させる必要がある．

$$t = [t(1), t(2), t(3)]$$
$$x = [x(1), x(2).x(3), x(4), x(5)]$$
$$x = [\{x(1), x(2)\}, \{x(3), x(4)\}, x(5)]$$
$$x' = [x'(1), x'(2), x'(3)] \qquad (3)$$

その際，データの長さを比較し，データの短い方に長さが合うよう長い方を前からまとめ，平均値または最大値を出力し，入力データx'として再定義する．これを L1 ユニット数回分行う．次にx'の各フレーム毎に平均または最大値等求めた後，入力データ$x'(k)$を(2)式の$x(k)$として用いる．

上記に述べた修正版 CNC およびフレーム長の正規化の詳細について学術講演会にて報告する予定である．

７．参考文献

[1] 荒木雅弘：「イラストで学ぶ音声認識」，講談社，pp.2-12,60-71，2016 年 7 月．

[2] Tetsuya Hoya and Cees van Leeuwen：”Connected Word Recognition Using a Cascaded Neuro-Computational Model”, Connection Science, Vol.28, No.4, pp.332-345, Aug.2016.

[3] Tetsuya Hoya :“ A modified cascaded neuro-computational model applied to recognition of connected spoken Japanese prefecture words”, J.Artificial Life and Robotics, pp.1-6, Aug.2019.