G-5

県名データベースを用いた修正版 CNC による連続単語音声認識
第 2 層における認識手法改良の検討と考察

Study of the Modified Cascaded Neuro-Computational Model Applied to Connected Spoken Word Recognition of Japanese Prefectures

○小川春[1]，保谷哲也[2]
*Haru Ogawa[1], TetsuyaHoya[2]

Abstract: In the previous work, the modified cascaded neuro-computational model (m-CNC) was proposed as a psychologically plausible model of human speech recognition, while it can handle real speech input. In this paper, we study the modified CNC in more depth for a further improvement, focusing upon the second layer, using a speaker-independent dataset of connected Japanese prefecture names.

１．概要

本研究では，心理学的アプローチに基づいた連続音声認識を行う人工ニューラルネットワークモデルの修正版 m-CNC(Modified Cascaded Neuro-Computational Model) における挙動について理解を深め，それを用いて連続県名音声認識を行うことが目的である．より具体的には，実験では 20 都道府県名を複数の話者により連続発音されたデータを用い，特定話者，不特定話者それぞれの場合について修正版 m-CNC による連続県名音声認識実験を行い，その得られた実験結果より第 2 層における認識手法の改良について検討する．

２．音声認識

音声は呼気が声帯を振動させてできる基本周波数に対し，これが声道を通る際に一部の周波数帯が共振することで周波数スペクトルに包絡構造が加わる．基本周波数は音程(ピッチ)を決定し，声道共振によって音声が特徴付けられる．

本研究では発話の内容を推定する事が目的であるため，基本周波数を除き，スペクトル包絡構造の低次成分を抽出したものとして MFCC(Mel Frequency Cepstral Coefficient)を特徴量として用いる．

３．Modified Cascaded Neuro-Computational Model

m-CNC(modified Cascaded Neuro-Computational Model)は，その前身である CNC モデル[1]の修正版であり，CNC 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．CNC は全 3 層で構成されており，Layer1(L1)は音素を推定するための RBF ユニット層，Layer2(L2)は各単語候補とテンプレートとの類似度を出力する層，Layer3(L3)は L2 の類似度を単語グループ毎に収集する事で各単語グループの発火値を決定する層である．

L1 では，MFCC 学習データがフレーム毎に入力された際，その発火値が閾値を下回った場合，(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 としてユニットが新たに追加される．また，CNC を用いた連続単語音声認識では，音声データの非音声区間を利用し，非音声区間で発火する RBF を L1S ユニットとして再定義する．

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

L2 では，L1 においてフレーム毎に最大発火した RBF ラベルが時系列データとして入力され， L2 に保存された内部テンプレートとの類似度を測定し，最大発火した単語候補ラベルと発火値が出力として得られる．学習の際には最大発火した単語グループラベルと教師ラベルが異なった場合新たなテンプレートとして L1 で得られた音素ラベル列が保存される．

L3 では，単語グループ毎に L2 の発火値を収集し，最も類似度が高い単語グループラベルを出力として決定される．

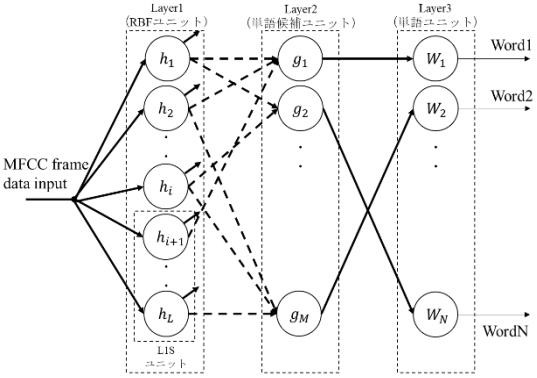

Figure1. Cascaded Neuro-Computational Model

１：日大理工・院（前）・情報２：日大理工・教員・情報

４．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，特定話者，不特定話者それぞれの場合において，県名データセットを用いた CNC と m-CNC による連続単語音声認識実験を行った．この際，検討のため実験における分類対象を一時的に発音が類似しない 20 都道府県に限定し，2~7 個連続で発話され，各都道府県による発話回数の偏りが無いように組み合わされた 22 通りの男性 5 名による連続単語音声によって実験を行った．

また，特定話者認識の際は学習話者を 5 名，テスト話者を 1 名とし，不特定話者認識の際は学習話者を 4 名，残りの 1 名をテスト話者とした．

m-CNC[2]では L1 の RBF で用いる *r1*(Radius 1)及び L2 の RBF で用いる *r2*(Radius 2)のパラメータ調整が必要である．本研究では，それぞれ，$3 \leq r1 \leq 6$，$0.1 \leq r2 \leq 1$ の範囲で組み合わせ，連続単語音声認識実験を行った．

５．実験結果

CNC と m-CNC による特定話者と不特定話者それぞれの実験結果を Table1 および Table2 に示す．実験結果から m-CNC の認識精度は特定話者認識では高い認識率を得られたが，不特定話者認識では特定話者の場合と比べて劣る結果となった．また，置換率が最も増加していることから，不特定話者認識において L2 の認識における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い事を示している．次章ではこのことを踏まえて，新たに修正版 m-CNC を提案する．

	Acc(%)	削除率(%)	置換率(%)	挿入率(%)
特定話者 r=5	96	2	2	0
不特定話者 r=5	70	2	27	1

Table1. CNC による認識精度

	Acc(%)	削除率(%)	置換率(%)	挿入率(%)
特定話者 r1=3.2 r2=0.4	99	0	0	1
不特定話者 r1=5.8 r2=0.7	86	1	9	4

Table2.m-CNC による認識精度

６．修正版 m-CNC

従来の CNC および m-CNC では，L2 の認識において L1 で最大発火した音素ラベルのみを考慮しており，L1 学習時の閾値が高いことから，2 番目に発火した RBF であっても十分な発火値を有している可能性がある．そのため，本研究では新たに修正版 m-CNC として次のように提案する： 2 番目，3 番目に最大発火…というような比較的RBFによる発火値が大きめのものを考慮しなかった事による認識率の低下を防ぐため，(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：

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

ここで，L2 において，テンプレートデータ及び入力データのフレーム長が異なる場合がある．この場合，テンプレートデータまたは入力データに計算されないフレーム区間が発生してしまう．そのため，フレーム長が短いデータを長いデータに対して 1 フレームずつずらしながら適用し，最大発火した値を出力とする．もしくは，正規化することにより，フレーム長が長いデータに合わせるように短いデータを前から伸ばし，一時的にデータを再定義する方法も考えられる．

しかしながら，L2 認識時に入力データを L2 内部テンプレート長の最大値でフレームを区切って認識する事から，内部テンプレート長に大きな差があった場合に認識に影響が出る事が推測されるため，本研究ではフレームをずらしながら適用する手法を修正版 m-CNC として新たに提案する．

７．実験結果

５．と同様のデータセットにおける修正版 m-CNC による実験結果を Table3 に示す．修正版 m-CNC が m-CNC よりも認識率が低下する理由の考察について学術講演会で発表する予定である．

	Acc(%)	削除率(%)	置換率(%)	挿入率(%)
特定話者 r=3.3	78	0	19	3
不特定話者 r1=4.6	69	1	25	5

Table3.修正版 m-CNC による認識精度

８．参考文献

[1]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.

[2]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.