P-7

Cascaded neuro-computational model を用いた連続音声認識について

A cascaded neuro-computational model for continuous speech recognition

○中田圭介[1], 会沢純行[2], 保谷哲也[3]
*Keisuke Nakata[1], Yoshiyuki Aizawa[2], Tetsuya Hoya[3]

Abstract: It is generally considered that human speech recognition is done by analyzing the levels of sub-word and word. A cascaded neuro-computational model proposed in the previous work models the process of human speech perception. In this paper, we investigate its potential applicability to continuous speech recognition.

１．はじめに

一般に, 人間の音声知覚は音素および単語レベルの２段階により音声を分析することによって行われると考えられている. ‘A cascaded neuro-computational model’ (CNC)はその音声知覚の過程を人工ニューラルネットワークによりモデル化したものである. 先行研究[2]では, このモデルを実際に数字音声パターン認識に適用し, HMM[1]と同様な認識率を得られたことが報告されている. 本論文では CNC を連続音声認識の識別に適用することについて述べる.

２．A cascaded neuro-computational model

CNC は Figure 1 に示されるように３層構造を成している. CNC の学習は各層のユニットおよび層間の結合を動的かつ追加学習的に行われる. より具体的には, 音声サンプルを mel-frequency cepstrum coefficients(MFCC)の形式で特徴抽出を行い, それをフレーム単位の入力としてネットワークに順次提示する. 各 MFCC フレームデータは３つの階層を通して順次処理される. 第一層では音素レベルの情報を扱っており, 単一フレームの MFCC データより得られたセントロイドベクトルを用いて, radial basis function(RBF)の計算をする. なお, 学習時には適宜ユニットの追加が行われる. 次に第二層および, 第三層において単語レベルの情報を扱う. まず第二層では, 第一層により最大限に活性化されたユニット情報を全フレーム分集積した結果を入力とすることにより各単語候補ユニットの出力が得られる. そして第三層では, 第二層で構成された各ユニットに対してカテゴリーの分類を行い. 最終的なパターン認識結果を出力する.

３．連続音声認識への適用

単語認識とは単一単語のみ(ex. /ichi/)を認識させることであるが, 連続音声認識では連続する単語

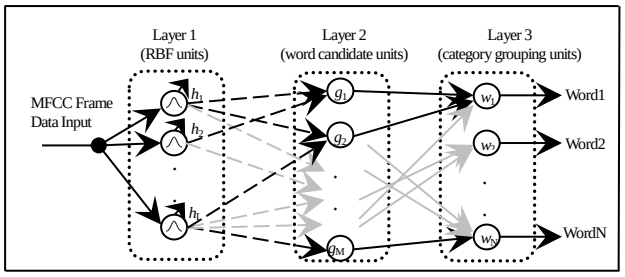

Figure 1. A cascaded neuro-computational model

(ex. /ichi-ni-san/)が認識対象である. 本研究では, その単語認識の方法を応用させることにより連続音声認識を行う. 学習時のネットワーク構築法については単語認識の場合と同様に行うが, テストの方法については以下の手順で行う.

(1) 特徴抽出後のテストデータの全フレームを, CNC に順次入力してゆき, 各フレームについて第一層で最大に発火したユニットナンバーを順に並べる.

(2) 学習時に第二層で作られたテンプレートベクトルとのパターンマッチングを行う. まず始めに, テストデータを第一フレームより数えたフレーム数だけを取り出す. その時のフレーム数を

Ave{g_i(i=1, 2 , …, M) ∈ [∀単語の word cadidate unit] (1)

とする. そして第三層より結果が出力される.

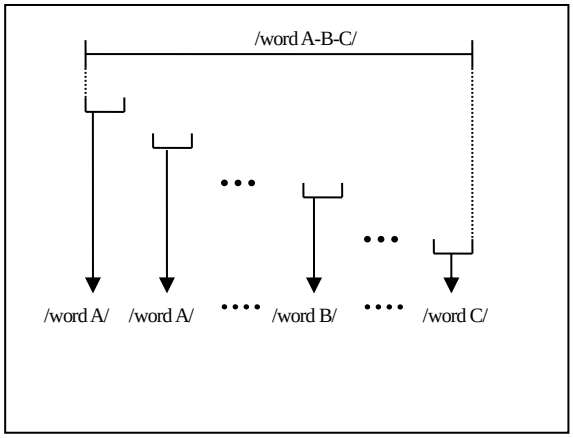

Figure 2. Continuous speech recognition pattern matching

1:日大理工・院・数学　2:日大理工・学部・数学　3:日大理工・教員・数学

(3) 次に，フレームと呼ばれる 10~30[msec]程度の短い区間に対する分析を時間方向に繰り返して行う．この１フレーム分の分析から一つの特徴パターンが得られ，特徴パターンの時系列が音声波形全体の特徴を表すことになる. (Figure 2).

４． 実験

本研究における実験では，学習データ（データセット１）として，７人の特定話者により得られた各数字音声（データセット１）につき７０パターンずつ用意した．また，テストデータ（Figure 3）として，同一の話者により各数字につき１５パターンずつ用意した．実験方法として以下の３種類を考慮した．

（１）特定話者実験Ⅰ: 一人のテストデータ，学習データ全て用いた認識実験.

（２）複数特定話者実験: テストデータ，学習データを複数人分用いた認識実験.

なお，学習データには実験で用いる話者の全てのデータを用い，テストデータには任意の数のデータのみを用いて実験を行った.

（３）特定話者実験Ⅱ: 一つのテストデータを３等分に切り取り，それぞれを学習データとして用いた認識実験する．この実験結果については，学術講演会当日に発表する予定である.

データセット１: /ichi/, /ni/, /san/, yon/, /go/, /roku/, /nana/, /hachi/, /kyu/の９つの数字

データセット２: /ichi-ni-san/や/ni-ni-ni/など３つの数字を組み合わせたもの

５． 実験結果

連続音声認識実験（１）～（２）の結果を Figure 4 および, 5 に示す．（なお，図の横軸は話者，縦軸は認識率である．）

６．終わりに

本研究では, Figure 4 や 5 に示されるように，特定話者実験について数字の組み合わせや，無音部分の切り取り方によって結果に大きく影響することが観測された．今後の課題としては，複数特定話者認識実験における認識率向上が挙げられる．また，３等分に切り取ったデータを用いた場合の認識率がどのように推移していくかを確かめ，その結果を学術講演会当日に発表する予定である.

７．参考文献

[1] 荒木雅弘 “フリーソフトでつくる音声認識システム”，森北出版株式会社・2007, pp. 130-147.

[2] Hoya Tetsuya and van Leeuwen, Cees (2010), ‘A cascaded neuro-computational model for spoken word recognition, ’ Connection Science, 22: 1, 87-101.

118	123	147	222	239	355	369
416	451	456	573	642	645	745
789	867	871	898	932	967	983

Figure 3. テストデータで用いられた各数字列音声

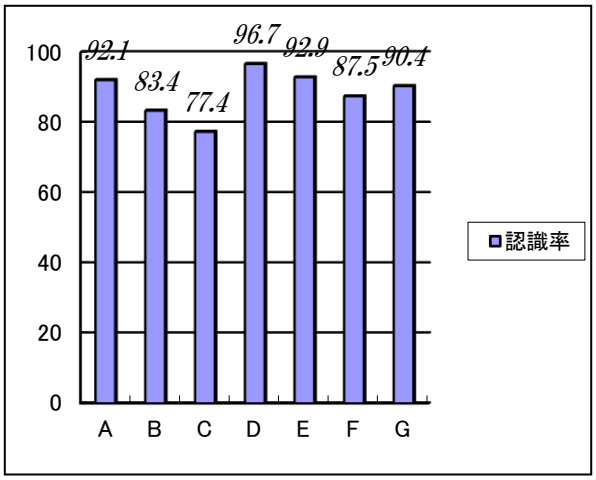

Figure 4. 特定話者実験による認識率

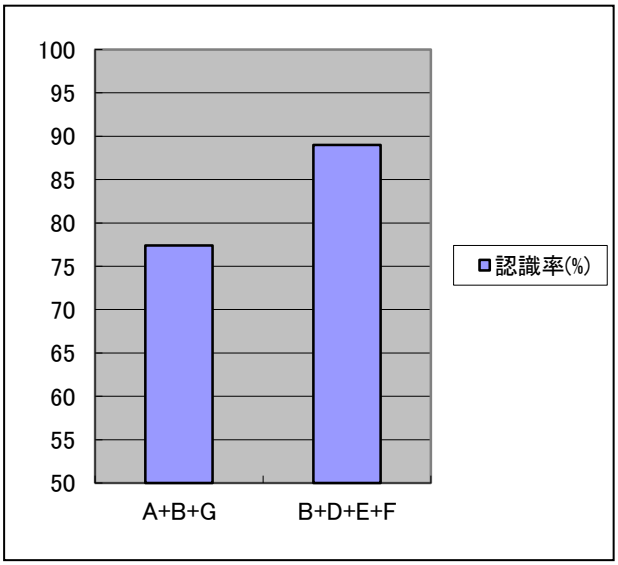

Figure 5. 複数特定話者実験による認識率