ニューラルネットワークを用いた連続音声のわたり部の認識

5V−3

保住透　吉村宏紀　清水忠昭　菅田一博　井須尚紀

鳥取大学工学部

1. はじめに

連続音声の認識において, 音声信号のパターンマッチングを音節単位で行うことは, 非常に効果的である. そのためには連続音声を音節に分割(セグメンテーション)することが必要となる. しかし, 音声には変動要因が多い上, 音節間の調音結合の規則が不明確なため, セグメンテーションすることは非常に困難である.

本研究では, 連続音声中の“わたり部”を, LSP分析して得られるLSPパラメータの時間変動としてとらえ, ニューラルネットワークに“わたり部”を認識させることによって連続音声をセグメンテーションする方法を提案する. 認識に用いるニューラルネットワークをLSPパラメータの特性を考慮した構造にすることによって認識率の向上を計った. このニューラルネットワークが有効であることを実験的に確かめた.

2. ニューラルネットワークの構造

分析次数 p 次のLSPパラメータ $f_i(i=1,\cdots,p)$ の各 f_i は独立に時間変動している.(図1参照) 本研究ではこのLSPパラメータの特性をニューラルネットワークの構造に反映させた. 各 f_i の時間変動を認識するニューラルネットワークを p 個(以下, 第1段ネットワークと呼ぶ)を用意する. p 個の第1段ネットワークの認識結果を入力として着目部が“わたり部”であるかどうかを総合的に判断するニューラルネットワーク(以下, 第2段ネットワークと呼ぶ)を用意する. 従って, 第1段ネットワークと第2段ネットワークからなる2段構造のニューラルネットワークを考える.(図2参照)

第1段ネットワークはLSPパラメータの時間変動を1. 定常である, 2. 変動部分である, 3. 変動部分の開始点, または終了点を含んでいる, の3つのカテゴリに分類するネットワークとした. 第1段ネットワークに対する入力には,LSPパラメータの時間変動を認識させるため, LSPパラメータの時系列を与える.

第2段ネットワークに対する入力は p 個の第1段ネットワークの出力を与える. 第2段ネットワークの出力は現在の音声信号を1. 有声音部である, 2. わたり部である, 3. わたり部の開始点, または終了点を含んでいる, の3つのカテゴリに判定する.

3. 実験

実験で用いた音声資料のLSP分析の仕様は次のとおりである. フレーム長:25.6ms, インターバル長:19.2ms, 分析次数:12次.

本研究では,“わたり部”をLSPパラメータの時間変動としてとらえているため, 子音も“わたり部”に含まれる. 日本語の発音単位は,1. 母音のみ,2. 1つ又は2つの子音+母音 である. 従って,LSPパラメータの定常部分である母音の位置を特定することによって連続音声をセグメンテーションすることができる. そのため, 音声資料として日本語5母音 /a,i,u,e,o/ を連続発声したものを用いた.

音声資料を分析して得られるLSPパラメータは0.0〜5000.0(Hz)である. それに対し, ニューラルネットワークの入力は0.0〜1.0である. 本研究では一般の正規分布を,$N(0.5, 0.215^2)$ に変換する方法を用いてLSPパラメータの各 $f_i(i=1,\cdots,12)$ の値を $[0,1]$ とした値を第1段ネットワークの入力とした. ノイズを除去するために, 各 f_i

Segmentation of Speech Signal by Neural Network

Tohru Hozumi,Hiroki Yoshimura,Tadaaki Shimizu,Kazuhiro Sugata,Naoki Isu
Faculty of Engineering,Tottori Univ.

の変換後の値の 3 点移動平均値を入力値とした. 第 1 段ネットワークに入力する LSP パラメータの時系列は 8 時点 (8×19.2ms) とした. 従って, 各々の第 1 段ネットワークの入力ユニット数は 8 個, 出力ユニット数は 3 個である. 第 2 段ネットワークの入力は, 第 1 段ネットワークの出力であるから, 第 2 段ネットワークの入力ユニット数は 3(第 1 段ネットワークの出力ユニット数) × 12(第 1 段ネットワークの個数) = 36 個, 出力ユニット数は 3 個である.

各ニューラルネットワークの中間層の層数, ユニット数は複数の候補の中から実験によって決定した.

4. 結果・考察

本研究で考案したニューラルネットワークによる “わたり部” の認識率は 95%であった. 誤認識した部分はほとんどが “わたり部” の始点, 終点の近辺であった. これは “わたり部” の始点, 終点が明確でないため, ニューラルネットワークを学習させる際に用いた教師信号の “わたり部” の始点, 終点の誤差を完全には除去することができないためであると考えられる.

本研究で考案した方法で連続音声のセグメンテーションを行う場合, 他の方法と併用することによってさらに高精度のセグメンテーションが行えると思われる.

5. おわりに

本研究では,LSP パラメータの特性を考慮した構造のニューラルネットワークは連続音声のセグメンテーションを行うのに有効であることを実験的に示した.

ニューラルネットワークの構造に問題の特性を反映させたので, ネットワークの規模を小さくすることができ, 汎化能力を高めることができた. 本来は解決すべき問題の特性をニューラルネットワークが自分自身で構造に反映できることが望ましい. しかし, あらかじめ問題の特性が解っている場合にはネットワークの構造として与えておくことが汎化能力, 認識率などを向上させると考えられる.

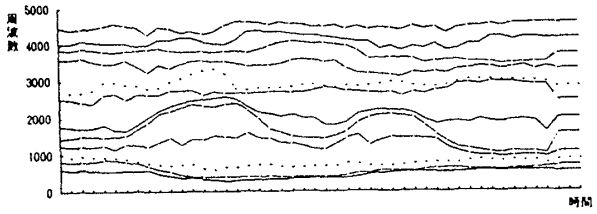

図1　ＬＳＰパラメータの時間変動

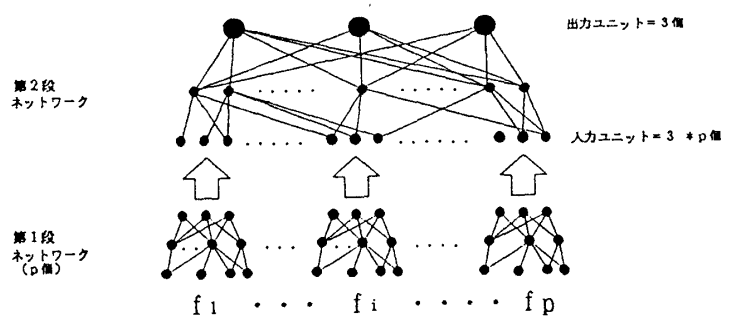

図2　ＬＳＰパラメータの特性を考慮したニューラルネットワークの構造の模式図