【22】

氏名	呉 簡彤（こ かんと）
博士の専攻分野の名称	博士（医学）
学位記番号	第 9914 号
学位授与年月日	平成 3 年 10 月 14 日
学位授与の要件	学位規則第4条第2項該当
学位論文名	Neural Network Vowel-Recognition Jointly Using Voice Features and Mouth Shape Image （音声・口形特徴量を併用するニューラルネットを用いた母音認識）
論文審査委員	（主査）教授 田村 進一 （副査）教授 松永 亨　教授 井上 通敏

論文内容の要旨

〔目 的〕

本研究は，口形画像を一つの有効な情報として用いることにより，ニューラルネット処理による音声認識を行うものである。その目的の一つは聴覚障害者・咽頭摘出者相互間や健聴者とのコミュニケーションを補助するシステムを開発することである。口形は元々聴覚障害者のコミュニケーション手段として使われてきた。聴覚障害者が手話を用いて，コミュニケーションを行うときには，手話動作をみるだけでなく，口の動きや表情も読み取って，トータルとして話を理解している。実際，聴覚障害者は口形のみから話の内容の70—80%を理解するといわれている。本研究はさらに一般的に音声認識の精度向上，高騒音時の音声認識にも応用可能である。

〔方法ならびに成績〕

口形から音声を認識する研究は，日本語発話時の口の形や口の動きを調べた基礎的研究，音声認識の後処理に口形特徴を用いて，単語の認識率を高めようとした研究，ｘ—ｙトラッカーを用いたり，口紅を用いて安定した口形情報を得てそれをもとに音声認識を行った研究，画像処理により口紅など補助手段を用いずに口内部領域を抽出し，それにもとづいて音声認識を行った研究，口唇画像から得られるパラメータの時間変化特徴と音声信号のスペクトルより，子音の同定を行った研究などがある。また，当研究室ではニューラルネットを用いて，音声データのみから音声認識を行ってきている。

本研究では，音響的特徴と口形画像特徴を同時にニューラルネットに入力し，母音の識別を行う。その際の音声信号は，男11人および女3人の学生の／ａ／－／ｏ／の母音データ（各人各母音10回発

声），およびそのときの口形画像データを作成した。音響的特徴としてはFFTによるパワースペクトルを用いる。これらはFFTによる特徴抽出をうけた後，対数変換される。この処理のスペックはサンプリング周波数10 kHz，量子化ビット数8 bit，フレーム長20ms，FFTサンプル点数128，同特徴数64，FFTウィンドウはハミング窓である。口形画像としては，原画像（濃淡画像200×100），二値化画像（10×5，歯を含まない場合および歯を含む場合），二値化画像の幾何学的特徴（有効なx方向，y方向投影長および実面積からなるデータ・セットを用いる）の三種を用いる。ニューラルネットについては，3層のバックプロパゲーション（Back Propagation）ニューラルネットを用いる。音声特徴が入力されるユニット数は64で一定であるが，画像特徴の数が画像データのタイプにより異なるので，入力層のユニットの総数はそれにしたがって異なり，64－264となる。中間層のユニットの数は，8または12である。出力層のユニット数は/a/－/o/に対応して5である。学習パラメータの学習定数ηは0.1－0.4，安定化定数αは1と設定する。これら音響的特徴と画像的特徴を同時にニューラルネットに入力し，認識率の比較を行う。これにより，ニューラルネットによる認識に有効な画像特徴の種類やレベルをしらべる。

結果では，音声のみの場合の認識率は全体として82.7％で，学習話者を徐いた場合は80％である。音声＋画像の場合には，音声＋二値化口形画像三特徴を除いて，いずれも認識率が増加し，音声＋二値化画像（歯を含む）の場合92.4％の最大不特定話者認識率が得られた。歯を含まない場合および歯を含む場合の音声＋二値化画像の認識率は大体同じである。これらは認識率が高いだけでなく，学習時間の点でもすぐれている。音声＋濃淡画像の場合は，音声＋二値化画像の場合より，認識率が低い（危険率7.2％）。音声＋二値化画像の三つの特徴の場合の認識率は音声のみの場合と比べて，認識率は少し低下しているが，大きな差はない。画像データのみの場合の認識率は概ね5割強であり，大きな差はみられない。したがって，このサイズの画像のみに基づく不特定話者の音声認識は困難であるといえよう。なお，今の場合，濃淡画像と二値化画像の差はみられなかったが，音声データを加えた先の実験では，濃淡画像のほうが認識率はやや低かった。このことは，画像データと音声データが別々に処理されるのではなく，相互に関連して認識が行われていることを示唆している。この一つの解釈としてはつぎのようなことが考えられる。すなわち，濃淡画像データのみの場合のような不確かな状況下では，情報ロスのない原画像がよいが，音声データが与えられたようなかなり候補が絞られた状況下では領域をはっきりと示す二値化画像が最終判断には有効となる。

〔総　括〕

(1) 音声特徴と口形画像特徴を組み合わせてニューラルネットに入力し，母音認識を行うことは有効である。

(2) その際，二値化も特徴抽出の一種であると考えると一般的に，適切な特徴の選択・抽出により，認識率の向上に加えて学習時間の減少も期待できるが，不適切な特徴を用いればそれは期待できない。

(3) ニューラルネットは本研究のように，多様な情報をもとに学習および認識を行う場合に有効に働

く。

(4) 本方式は聴覚障害者や咽頭摘出者の口の動きおよび発声から音声認識を行う場合の基礎となろう。

論文審査の結果の要旨

従来ほとんどの音声認識の研究は音声の特徴のみを用いて行われてきた。本研究はそれに加えて口形画像特徴量を同一のニューラルネットに入力して音声認識を行っている。このような研究は，聴覚障害者，咽頭切除者などのコミュニケーション補助や，高騒音下での音声認識などに役立つことが期待される。

本研究では，口形画像の色々な特徴量と音声のパワースペクトルを共にニューラルネットに入力し，比較を行っている。その結果，(1) 口形認識に従来用いられてきた3種の幾何学的特徴量は不特定話者認識にはまったく無効であり，(2) 二値化画像そのものを用いた場合，母音認識率は12%向上する，(3) 濃淡画像そのものを用いた場合，それは6%向上する，ことを明らかにしている。

このような口形を利用した新しい音声認識方式の提案とニューラルネットに適した画像特徴を明らかにしたことは，音声認識，ニューラルネット，および画像認識の研究に寄与する点が大きく，学位に値する業績と認められる。