サポートベクターマシンによる動画像からの発話情報の認識

室蘭工業大学　小野 功一(P)、魚住 超(AP)、長嶋　友輝(MC)、大林 由英(PD)

1. はじめに

聴覚障害者の対話手段の１つとして口話法がある。口話法の中でも、相手の口の動きから何を話しているのかを理解する事は「読唇術」などと呼ばれており、その習得は困難とされている。本研究はパターン認識を行う学習モデルの一種、サポートベクターマシン(SVM)を用い、口唇領域の動画像からの発話認識を行い、その認識特性について考察していく。

SVM の主な特徴としては、1. 未学習のデータに対し優れた学習性能、2. 誤データに対する許容量を設定可能、3. 特徴ベクトルを完全に分離するための非線形拡張が可能で、かつそれに伴う高次元演算負荷

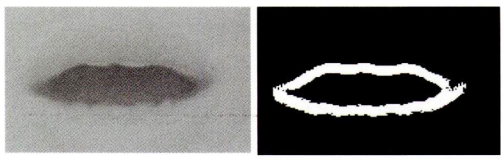

図１グレイスケール（左）輪郭抽出（右）

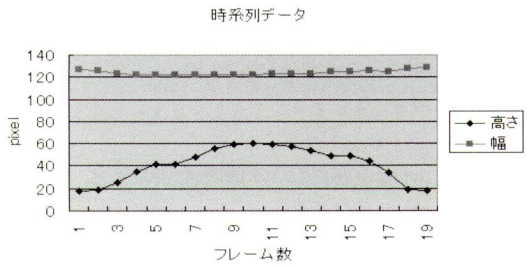

図２　パラメータの時系列データ

を減らすことができる、といった点が挙げられる。

2. パラメータの抽出

パターン認識に用いるパラメータとして、画像からの取得が容易な口唇領域の幅・高さの時系列データを用いる。パラメータは、図 1, 2 のように動画像処理によって口唇輪郭付近のデータから抽出する。また、発話時間の違いによる誤認識が考えられるためデータの時間軸正規化をおこなった。発話認識実験に用いたデータの特徴を表１に示す。

対象	:	特定話者・口唇領域の動画像
データ区間	:	発話開始時点から終了時点まで
形式	:	AVI ファイル
フレームレート	:	約　30 frame/sec
種類	:	孤立単音 46 音
データ数	:	2300 サンプル（46×50 データ）

表１　データセットの概要

3. 発話認識実験

本研究では孤立単音の認識に絞り実験を行った。データセットのうち、それぞれの孤立単音データ一種類毎の全 50 サンプル中で n 個の教師データと 25 個のテストデータを無作為に取り出し、25 個のデータの認識実験を行った。

予備実験として母音のみの認識を行った。教師およびテストのデータ双方に母音のみを用意した場合の平均認識率　83.2%(n=1)，　94.4%(n=10)，93.6%(n=25) と n の増加に応じて改善され、かつ 90% 以上の精度で認識が可能である。

さらに n=25 の場合、孤立単音 46 全ての認識実験結果について、図 3 に示す。

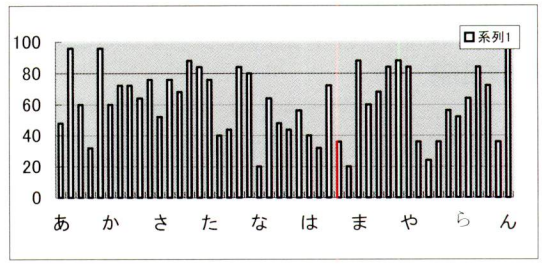

図 3 全孤立単音の認識率

口唇の輪郭抽出で得られた幅・高さという単純なパラメータによる認識手法をとったが、時系列変化を取り入れた結果として、特定話者であるとはいえ、母音については、ほぼ 100%，　46 音全音認識においても単純平均で 60～70%程度の認識が可能であることが示された。

4. まとめ及び今後の課題

特定話者の発話から特徴認識に絞り、SVM による発話認識の性能についての評価を行ったが、単純な特徴量の時間変化を判別基準として、子音を含む孤立単音の認識の可能性について実験を行い、発話情報の画像による認識可能性について考察した。

今後は発話の個人差の問題、そして単音と連続発話での違い等の調査が必要となる。またサンプル数をさらに増やした場合での認識率の安定性の考察も実際に読唇を実現するシステムへむけての重要な課題である。

また、今回は採用した時間推移に関するデータの正規化をおこなうことにより、時間軸に対する単位幅が動的に変わる量の認識を SVM で可能としたが、同じ発音の波形が異なる発話時間長の場合でも時間軸に対して相似形である場合、高認識率が保障されるが、実際に認識率の悪いデータに関してはこの仮定から外れた特徴を持っている場合であった。故にその正規化の方法についての吟味も必要であると考える。また、時間軸のみでなく幅、高さの絶対値についても何らかの正規化をかけた上で認識が必要であると考えられる。

認識率の評価については、各音の認識率の平均ではなく、実際の口話時での各音の出現比率に依存した形での評価が望ましいと考えられその点でのアプローチも今後検討してゆきたい。

参考文献

長嶋 友輝：室蘭工業大学 修士論文 2003, 3 月