唇周辺のマークを用いた音素認識のための基礎検討

6C－11　古山　浩志、八塩　仁、井上　郁夫
松下電器産業(株)東京通信システム研究所

1. はじめに

我々は、音声認識技術の映像検索への応用を目的とし、視覚情報と聴覚情報とを併用した音声の機械認識の認識精度を向上するための方式開発を行っている。

今回は、視覚情報を併用した音声認識方式開発の基礎検討として、唇周辺にマーキングして単音節を発声する話者を撮影、マークのみを抽出、マーク数を変えた映像を作成、被験者に提示して視覚認識実験を行った。また、抽出したマークを特徴パラメータに用いて、母音と一部の子音グループを対象とした音素の機械認識実験を行い、視覚認識と機械認識の結果を比較した。

2. マーク映像提示による視覚認識実験

話者(女性社員1名)の唇周辺にマーク(14点)を貼って撮影した映像(図1－(1))から色相を用いた色空間領域抽出によりマーク部分を抽出した映像(図1－(2))を作成した。このマーク抽出映像の先頭フレームのマーク位置から、次フレームのマーク位置を算出する処理を行い、各映像フレームのマーク位置座標を抽出し、マーク数が14、8、4の映像(図1－(3)～(5))を作成し、被験者に110単音節を提示して視覚認識実験を行った。

(1)オリジナル映像　(2)マーク抽出映像

(3)14マークのみ　(4)8マークのみ　(5)4マークのみ

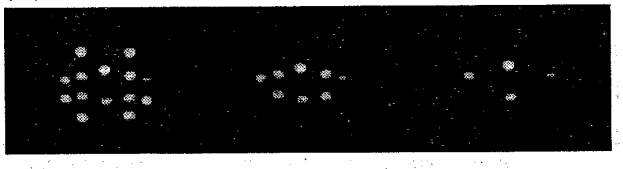

図1. 提示映像例

A Study for Phoneme Recognition using Marks Placed around Lips.
H. Furuyama, H. Yashio, and I. Inoue.
Matsushita Electric Industrial Co., LTD.
4-5-15, Higashi-Shinagawa, Shinagawa-ku, Tokyo 140-8632, Japan.

図2に男女各2名の被験者を対象に視覚認識実験を行ったときの単音節、母音、子音の正解率を示す。オリジナル映像、マーク抽出映像、マークのみ映像提示時の順で正解率は低下しているが、提示するマーク数による正解率の差はほとんどなかった。

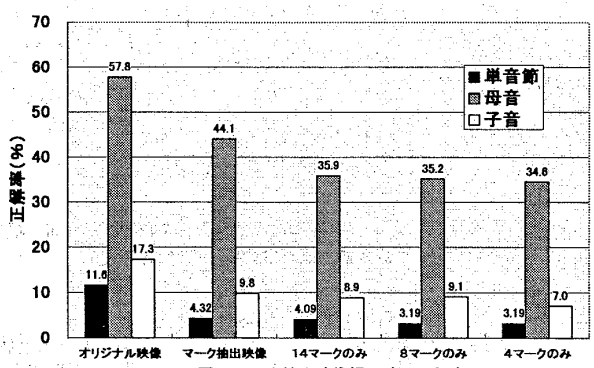

図2. マーク抽出映像提示時の正解率
(話者:女性社員1名、被験者:男女各2名)

図3に各母音の正解率を示す。/I/と/U/でマークのみ映像提示時に正解率が大きく低下した。

4マークのみ提示時に多くみられた誤答のパターンは、/A/→/E/(53%)、/I/→/E/(63%)で、/E/は/A/(24%)と/I/(13%)への誤答が多かった。/O/はオリジナル映像提示時には、発声時の唇の形状が同じ円唇となる/U/への誤答が多かったが、マークのみ提示時には、/E/への誤答が増えた。一方、/U/はオリジナル映像提示時に/O/への誤答が多かったが、マークのみ映像提示時には、/O/以外に/A/、/E/、/I/への誤答が増えた。

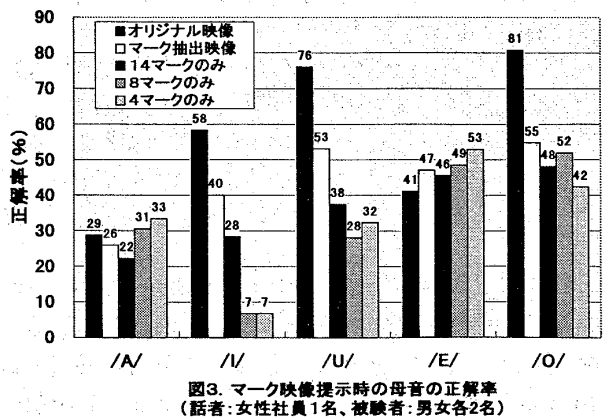

図3. マーク映像提示時の母音の正解率
(話者:女性社員1名、被験者:男女各2名)

図4に子音グループ[1](抜粋)の正解率を示す。オリジナル映像提示時、およびマーク抽出映像提示時に正解率の比較的高かった/B、P、M、BY、PY、MY/

(唇音)のグループは、マークのみ映像提示時も同様に他のグループと比較して正解率が高い。

/D、N、T、DY、NY、TY/のグループについては、オリジナル映像提示時よりも、マーク抽出映像提示時、マークのみ映像提示時に正解率が上昇した。このグループは、オリジナル映像提示時に子音の欠落という誤答パターンが多くあったが、マーク抽出映像提示時、マークのみ提示時は、子音の欠落が減少した。

/S、Z、SY、ZY/のグループは、マーク抽出映像提示時、マークのみ提示時ともに、オリジナル映像提示時と比較して大きく正解率が低下した。このグループは、オリジナル映像提示時は、/D、N、T、DY、NY、TY/への誤答が多かったが、マーク抽出映像、マークのみ映像提示時には、/D、N、T、DY、NY、TY/への誤答とともに子音の欠落が増加した。

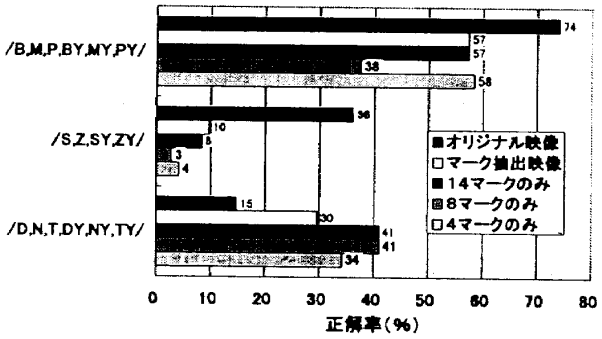

図4. マーク映像提示時の子音の正解率
(話者:女性社員、被験者:男女各2名)

3. マークを用いた機械認識実験

図1で示した唇周辺のマーク座標を特徴量として、39単音節(5母音と撥音、子音は/S/、/D、T、N/、/B、P、M/のみ)を対象に機械認識を行った。

データは男性アナウンサー1名と男女社員3名の話者の映像を使用した。また、音声データから音素の始終端を決定し、母音と子音の始終端の中間フレーム画像(静止画)を使用した。

抽出した特徴パラメータ(マーク位置座標、唇の上下の長さ(縦)、唇の左右の長さ(横)と縦横比)に対してK-means法によりクラスタリングを実行し、クラスタの中心ベクトルと入力データの距離を算出し、距離が最小となるクラスタに対応する音素を認識結果として出力した。

女性社員のデータに対して、特徴パラメータを14座標点、8座標点、4座標点、縦、横、縦・横、縦・横・縦横比としたときの母音の正解率は順に46.2%,46.2%,30.8%,35.9%,33.3%,61.5%,64.1%となり、今回の実験では、座標点(4〜14点)を特徴パラメータとして用いた時よりも、4座標点から算出した縦・横・縦横比を用いた時の正解率の方が高くなった。

また、話者4名のデータに対して、クローズ条件(条件1:認識対象となる話者のデータのみを用いてクラスタリングを実行、条件2:認識対象となる話者を含む4名のデータを用いてクラスタリングを実行)での母音の認識率(特徴パラメータは縦・横・縦横比とした)は、順に66.0%、59.0%となった。また、オープン条件(認識対象となる話者を含まない3名のデータを用いてクラスタリングを実行)での母音の認識率は、53.8%であった。

表1にオープン条件でのコンフュージョンマトリックス(行:入力音素、列:認識出力音素、単位%)を示す。母音での誤答は、/A/→/E/、/I/→/U/、/U/→/O/、/E/→/I/と/U/と/A/、/O/→/U/が多い。

子音グループでの誤答は/S/→/D、T、N/、/D、T、N/→/S/、/B、P、M/→/D、T、N/が多い。また、歯茎音である/S/と/D、T、N/の識別はうまくできていないが、唇音(/B、P、M/)と歯茎音(/D、T、N/と/S/)は比較的良好に識別できている。

表1. クラスタリングによる認識結果
条件:オープン(話者:男性アナウンサー1名、男女社員3名)

(1)母音

	A	I	U	E	O	NN
A	25	6	3	56	9	0
I	7	71	14	7	0	0
U	0	11	71	0	18	0
E	16	44	19	22	0	0
O	6	3	13	0	78	0
NN	0	0	0	0	0	100

(2)子音グループ

	/S/	/D,T,N/	/B,P,M/
/S/	20	75	5
/D,T,N/	27	73	0
/B,P,M/	2	17	82

4. まとめ

単音節を発声する話者の唇周辺に貼ったマークを対象として、視覚認識実験と機械認識実験を行った。いずれの場合も母音は、{/A/、/E/、/I/}間と{/U/、/O/}間の誤答が多いという傾向があった。子音についても、唇音グループと歯茎音グループ間の識別は比較的よくできているということがわかった。

今後は、誤認識の多い音素間を識別するための特徴量の検討、今回対象としなかった子音グループの認識実験、および機械認識における時系列データの取り扱いについて引き続き検討を行っていく予定である。

なお、本研究は通信・放送機構からの委託研究テーマ「インテリジェント映像技術の開発」の一環として行っているものである。

5. 参考文献

1)古山他、「映像提示による単音節の音声知覚」、56回情報処理全国大会、6N-03(1998).