映像提示による単音節の音声知覚

6N－3

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。このための基礎データ収集の一環として、単音節を対象に視覚と聴覚による認識傾向の違いを実験により調べた。先に報告した実験の結果[1]を更に詳細に分析した結果、音素グループごとに視覚と聴覚とで明らかな認識傾向の違いが見られた。また、今回新たに唇周辺に貼ったマークを抽出した映像、および唇部分を抽出した映像に対する視覚認識実験を行い、新たな知見を得たので報告する。

2. 音素グループ毎の認識傾向について

映像のみを提示した認識実験（話者：男性アナウンサー1名、被験者：男女各5名）において、子音の正解率は20%程度と低いが、唇音ｸﾞﾙｰﾌﾟの正解率は94%と高いことを報告した[1]。

視覚と聴覚による子音の認識傾向の違いを調べるため、調音位置と調音様式をもとに音素をｸﾞﾙｰﾌﾟ化し（後続する拗音(/y/)の有無については区別をせずに同じｸﾞﾙｰﾌﾟとした）、各音素ｸﾞﾙｰﾌﾟ毎の正解率を比較した(図1)。

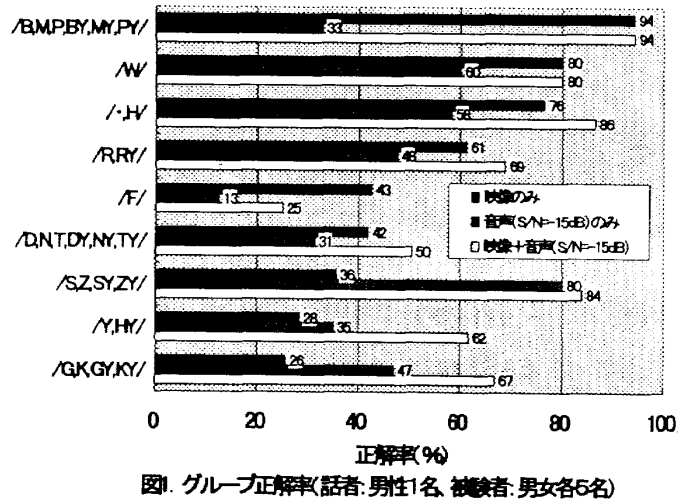

図1. グループ正解率(話者:男性1名、被験者:男女各5名)

図1より、映像のみ提示の場合、/b,m,p,by,my,py/（94%）、/w/（80%）、/',h/（76%、/'/は母音と撥音）、/r,ry/（61%）の順で正解率が高く、調音位置が唇(/b,m,p,by,my,py/、/w/)にあるものの正解率が高い。一方、音声（S/N=-15dB、hothﾉｲｽﾞ重畳）のみ提示の場合は、/s,z,sy,zy/(80%)、/w/（60%）、/',h/（58%）の順で正解率が高い。また、/b,m,p,by,my,py/と/s,z,sy,zy/については、映像のみ提示時と音声のみ提示時の間で正解率に大きく差があり、/y, hy/については、映像のみ、音声のみ提示時に比べ、映像と音声を提示した時に正解率が大きく向上した。

これらの各音素ｸﾞﾙｰﾌﾟに対する認識傾向の差を利用することにより、映像と音声を併用した音声の機械認識の認識率の向上が期待できる。

3. マーク抽出映像、唇抽出映像による視覚認識実験

3-1)実験内容

話者(女性社員1名)を正面から撮影しS-VHSﾋﾞﾃﾞｵに収録した映像（図1(a)）と唇部分を抽出した映像（図1(c)）、および話者の唇の周辺にﾏｰｸ（14点）を貼って撮影した映像（図1(b)）からﾏｰｸ部分を抽出した映像（図1(d)）を用いて、110単音節の視覚認識実験を行った。

図2. 提示映像

Auditory Perception of syllables by Visual Information.
H. Furuyama, H. Yashio, and I. Inoue.
Matsushita Electric Industrial Co., Ltd.
4-5-15 Higashi-Shinagawa, Shinagawa-ku, Tokyo 140, Japan.

3-2)実験結果

図3にｵﾘｼﾞﾅﾙ映像、ﾏｰｸ抽出映像、唇抽出映像を男女各2名の被験者に提示した時の単音節、母音、および子音の正解率を示す。ﾏｰｸ抽出映像、唇抽出映像提示時ともにｵﾘｼﾞﾅﾙ映像提示時よりも正解率が低下しており、単音節の正解率は約半分程度であった。

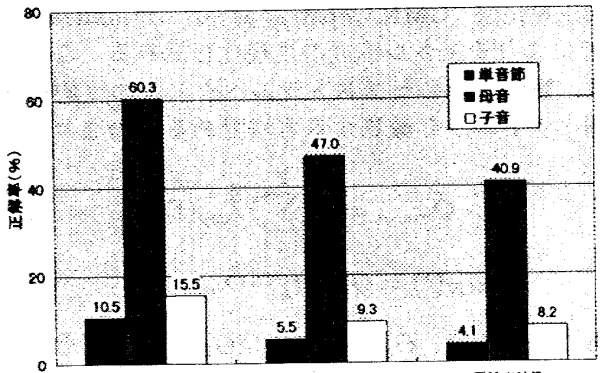

図3. 抽出映像提示時の正解率
（話者：女性1名、被験者：男女各2名）

図4に各母音の正解率を示す。母音の正解率は/a/と/e/では、3つの提示映像間での正解率に差はほとんどなく、/i/と/u/ではｵﾘｼﾞﾅﾙ映像、ﾏｰｸ抽出映像、唇抽出映像提示の順で正解率が低下した。また、/o/では、ｵﾘｼﾞﾅﾙ映像提示時に対してﾏｰｸ抽出映像、唇抽出映像提示時の正解率は低下したが、ﾏｰｸ抽出映像と唇抽出映像間での正解率の差はなかった。

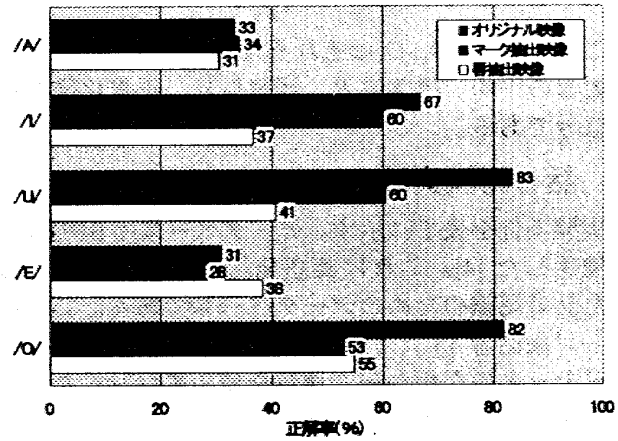

図4. 抽出映像提示時の母音の正解率
（話者：女性1名、被験者：男女各2名）

図5に各音素ｸﾞﾙｰﾌﾟ毎の正解率を示す。/d,n,t,dy,ny,ty/、/g,k,gy,ky/では、提示した3つの映像間での正解率の差はほとんどなかった。/f/、/s,z,sy,zy/、/r,ry/、/g,k,gy,ky/（、/w/）、については、ｵﾘｼﾞﾅﾙ映像提示時に対してﾏｰｸ抽出映像、唇抽出映像提示時の正解率は低下しているが、ﾏｰｸ抽出映像と唇抽出映像提示時の正解率の差はほとんどなかった。一方、/b,m,p,by,my,py/と/ﾝ,h/については、唇抽出映像提示時は、ｵﾘｼﾞﾅﾙ映像提示時に比べ正解率が大きく低下しているが、ﾏｰｸ抽出映像提示時は、ｵﾘｼﾞﾅﾙ映像提示時との正解率の差は唇抽出映像提示時と比較して小さい。これらのｸﾞﾙｰﾌﾟは、ｵﾘｼﾞﾅﾙ映像提示時で60%以上の高い正解率があり、映像と音声を統合した音声認識において性能向上への寄与が期待できるｸﾞﾙｰﾌﾟである。そして、ﾏｰｸ抽出映像でも同様に高い正解率であったことから、唇周辺のﾏｰｸﾎﾟｲﾝﾄとその動き情報は、映像を併用した音声認識において有効であると考えられる。

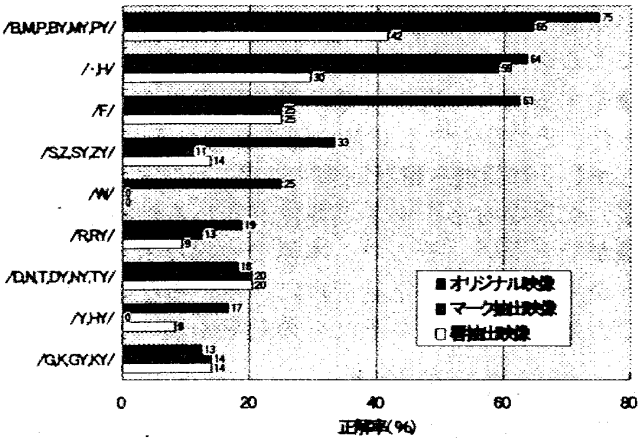

図5. 抽出映像提示時の音素グループ正解率
（話者：女性1名、被験者：男女各2名）

5. まとめ

単音節の視覚と聴覚による認識実験の結果について、子音を複数の音素ｸﾞﾙｰﾌﾟに分類した。各音素ｸﾞﾙｰﾌﾟに対する認識傾向の差を利用することにより、映像と音声を併用した音声の機械認識において認識率の向上が期待される。また、唇抽出映像とﾏｰｸ抽出映像を提示した認識実験結果より、唇周辺のマークポイントとその動き情報は映像を併用した音声認識において有効であると考えられる。

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。

6. 参考文献

1）古山他、「単音節の音声知覚における視覚情報と聴覚情報の関係」、55回情報処理全国大会、2-25(1997).