単音節の音声知覚における視覚情報と聴覚情報の関係

3Ｊ－1

松下電器産業（株）東京通信システム研究所
古山　浩志、八塩　仁、井上　郁夫

1．はじめに

視覚と聴覚情報の統合による映像検索への応用のための基礎検討として、複数の被験者に単音節を発声する話者の正面、横、斜め方向から撮影した顔画像と雑音を重畳した音声を提示して、認識実験を行った。提示した顔の向きと認識率の比較および誤認識の傾向をまとめ、視覚情報が音声知覚に及ぼす影響を調べた。

2．実験方法

映像は、正面、横（左側面）、斜めから撮影したS-VHS ビデオを単音節ごとにランダムな順序に編集し、14 インチモニターを被験者から約1m離れた位置に配置して、1単音節あたり約8秒の間隔で110 単音節を提示した。また、音声は hoth ノイズを重畳し、ヘッドホンを使用して被験者に提示した。

3．雑音レベルと認識傾向

話者(女性 1 名)の正面から撮影した顔画像と雑音(S/N=0、-5、-10、-15、-20dB)を重畳した音声を男女各2名の被験者に提示して正答率を調べた(図1)。

音声のみの場合、映像と音声を提示した場合のいずれもS/Nが低下するに従い正答率も低下するが、映像と音声を提示したときの正答率と音声のみを提示したときの正答率の差は、S/N が-15dB のときに最大（24%）であった。

以後の実験では、音声のみの場合と映像と音声を提示した場合の認識の傾向をみるために、認識率の差が最も顕著となるS/N=-15dB に設定した。

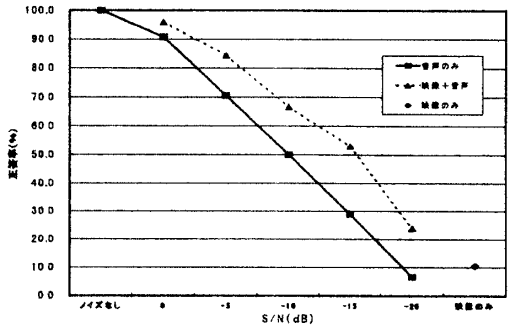

図1．S/Nと(単音節)正答率

4．顔の向きと認識傾向

話者(男性1名)の正面、横、斜めから撮影した映像と雑音(S/N=-15dB)を重畳した音声を男女各5名の被験者に提示し、音節、母音および子音の正答率を調べた(表 1)。映像のみの場合、音節の正答率は20%程度であったが、母音の正答率は 80%以上、子音の正答率は 20%前後であり、音節の正答率と子音の正答率がほぼ一致していた。また、斜め、正面、横の順で正答率が低下した。

表1.映像中の顔の向きと正答率(単位:%)

	映像のみ			音声のみ	映像と音声		
	正面	斜め	横		正面	斜め	横
音節	19.6	22.1	17.1	33.4	48.8	45.7	44.7
母音	85.2	86.1	82.5	91.4	95.5	96.0	96.0
子音	21.8	23.8	18.4	31.6	49.6	45.8	44.9

表 2(a)～(e)に母音のコンフュージョンマトリックス（行：提示単音節の母音、列：回答単音節の母音）を示す。映像のみを提示した場合、/a/と/e/の正答率が低い。また、正面の映像では/a/から/e/の誤答が多いが、斜め、横の順で/a/と/e/間の誤答傾向の逆転がみられた。また、音声のみを提示した場合には、/i/から/u/への誤答が多くみられたが、映像と音声を提示することにより正答率が向上した。

表3（a）～（c）に子音のコンフュージョンマトリックス(表中、/-/は母音と撥音）を示す。音声の

The Relationship between Audio and Visual Information on auditory Perception of Syllables.

H. Furuyama, H. Yashio, and I. Inoue.
Matsushita Electric Industrial Co., Ltd.
4-5-15 Higashi Shinagawa, Shinagawa-ku, Tokyo 140 Japan.

みの場合、誤答は子音の欠落が比較的多くみられたが、/y/のつく単音節とつかない単音節のグループ間での誤りは少ない。一方、映像のみを提示した場合には、/y/のつく単音節とつかない単音節のグループ間での誤りが比較的多く見られたが、唇音（/b、m、p、by、my、py/）をグループとした[1]ときの正答率は94%と高い。しかし、唇音を/y/の有り、無しでグループに分類すると、/b、m、p/の正答率は 89%と高いが、/by、my、py/は 55%と正答率が低くなる。

一方、映像と音声をともに提示した場合には、/b、p、m/と/by、my、py/のグループに分類した場合、それぞれ 91%、88%の高い正答率が得られた。これは映像による唇音と非唇音の識別と音声による/y/の有り無しの識別の相乗効果によるものと考えられる。

5．まとめ

hoth ノイズを重畳した単音節音声の認識において、母音では映像のみを提示した場合には/a/と/e/間の誤りが多く、音声のみの場合には/i/の誤答が多いという認識傾向の差がみられた。また、子音では映像のみの場合、唇音と唇音以外のグループ間での誤りが少ないという傾向が、音声のみの場合には/ y /のつく単音節とつかない単音節のグループ間での誤りが少ないという傾向が見られた。これらの認識傾向の差が映像と音声を提示したときの認識率の向上に寄与しているものと考えられる。

また、映像のみを提示した場合、正面顔画像では/a/から/e/への誤答が多いが、斜め、横の順で誤答傾向の逆転がみられた。

このような映像と音声、あるいは顔画像の向きによる認識傾向の差を利用することにより、映像を併用した音声認識において認識率の向上が期待できる。

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研究開発」の一環で行っているものである。

6．参考文献

1）積山他、「単音節の読唇における混同行列の分析」、信学技報、IE-87-127,29-35(1988).

表2.母音のコンフュージョンマトリックス（単位%）

(a)映像（正面）のみ

	A	I	U	E	O
A	66	0	0	32	0
I	3	85	6	3	1
U	0	2	97	0	1
E	13	1	1	82	0
O	1	0	1	0	97

(d)映像（正面）＋音声

	A	I	U	E	O
A	98	1	0	0	0
I	5	88	5	0	1
U	1	1	96	0	1
E	1	0	0	96	1
O	1	0	0	0	97

(b)映像（斜め）のみ

	A	I	U	E	O
A	77	0	0	19	1
I	3	85	6	1	2
U	2	3	94	1	0
E	23	1	1	71	0
O	0	0	0	0	98

(e)音声のみ

	A	I	U	E	O
A	97	0	1	0	1
I	7	71	14	1	5
U	1	3	96	0	0
E	0	4	2	92	1
O	0	0	2	1	95

(c)映像（横）のみ

	A	I	U	E	O
A	80	1	0	16	0
I	3	87	1	3	3
U	3	0	93	1	2
E	42	1	1	51	1
O	1	0	2	1	94

表3.子音のコンフュージョンマトリックス（単位 %）

(a)映像(正面)のみ

	・	H	B	M	P	BY	MY	PY	D	N	T	DY	NY	TY
・	47	30								3	2			
H	50	26									4			
B			2	50	34		4	6			2			
M		2	10	50	30		4				2			
P			14	34	42			4						
BY				13	3		60	20						
MY				23	17	7	30	13		3				
PY			7	30	23	7	13	13			3		3	
D									17	13	20		7	3
N	2								6	14	2		6	4
T		2							4	12	16		6	4
DY	20									40			10	
NY	3								3	7	3		13	13
TY	2	2								10	8		4	18

(b)映像(正面)＋音声

	・	H	B	M	P	BY	MY	PY	D	N	T	DY	NY	TY
・	88	3								2				
H	62	18	4							2				
B	4		60	14	12	4	6							
M			36	54	8		2							
P	2		44	16	28		4				2			
BY					3	60	20	3					3	
MY				3		23	70						3	
PY			3			20	50	17					3	
D									30	17	3			
N	16								6	36			4	
T	2								16	4	46			
DY	20									10				
NY													63	
TY										8		4		28

(c)音声のみ

	・	H	B	M	P	BY	MY	PY	D	N	T	DY	NY	TY
・	58	5	7	5				2		5				
H	44	8	8	4	2	2			2	2		2	2	
B	26	6	18	8	2	2		2		2				
M	18	8	10	20	4		6			4	2		2	
P	28	10	10	10	14					4			2	
BY						27	10						3	
MY				3		3	17						37	
PY						13	3	10					10	3
D	7		10						10	7				
N	6		8	2		2			10	26				
T	4		6		2				14	4	26			6
DY	10		10							20				
NY	3			3		3	13	3					30	
TY			2	2					2					18