Proceedings of The Institute of Acoustics

VISUAL PRESENTATION OF VOICING AND OTHER CUES AN AID TO LIPREADING

W.A.Ainsworth

Dept. of Communication and Neuroscience,
University of Keele, Keele, Staffordshire

INTRODUCTION

The task of a lipreader is not easy. Not only must he perceive the rapid and intricate movements of the lips, he must also continuously employ his knowledge of language and the current situation in order to resolve any ambiguities he sees. The words 'mob' and 'bomb', for example, look similar without sound.

One way of facilitating the task of the lipreader is to provide him with information which he cannot see. This information can be derived from the acoustic signal and then presented to a deaf lipreader by one or other of the remaining senses.

There are three categories of lipreading aids: visual, tactile, and cochlear implants. Each has been shown to be at least partially successful. Upton [1] has reported finding a visual aid useful to himself, Sherrick [2] has reviewed the prospects of tactile aids being used by deaf people, and Fourcin et al [3] have described tests carried out with external electrical stimulation of the cochlea.

Tactile stimulation has the advantage that it is independent of both the auditory and visual channels, so it does not interfere with lipreading. Conversely, however, it has no immediate connection with speech. In addition the resolution of the tactile channel is limited in both the temporal and spatial dimensions. Visual stimulation is attractive in terms of channel capacity, but may cause interference with lipreading. Cochlear stimulation ensures that the information enters the nervous system through the most appropriate channel, but there are obvious disadvantages of cost and surgical intervention. It appears, therefore that no category of aid is a priori superior, so investigations into the usefulness of each should be carried out.

In order to assess the appropriateness of visual aids a number of preliminary experiments have been carried out using vowel-consonant-vowel nonsense syllables and normal hearing subjects.

ANALYSIS

The speech signal was analysed as shown in Figure 1. An electret microphone converted the acoustic signal into a voltage whose intensity range was reduced by logarithmic compression. The signal was then low pass filtered at 400 Hz, detected, and applied to a threshold circuit. The threshold was adjusted so that this circuit acted as a voicing detector. The output of the threshold circuit was used to drive a red LED.

VISUAL PRESENTATION OF VOICING

In a second branch of the circuit the wide-band signal was detected and applied to a second threshold trigger. The output of this was combined with the output of the first branch such that a signal appeared if the input was sufficiently intense but was not voiced. The output was connected to a green LED. This circuit was sensitive to voiceless fricatives.

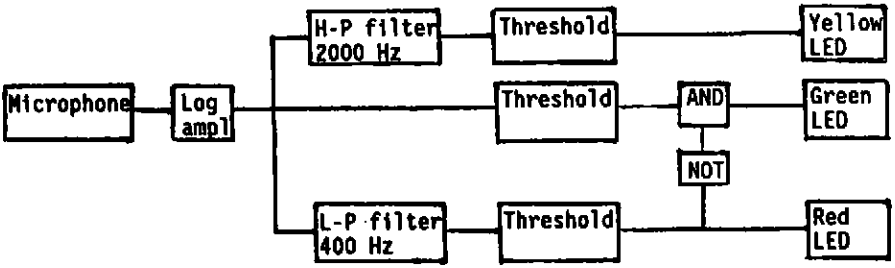

Figure 1. Schematic diagram of the speech analyser.

In a third branch of the circuit, the signal was passed through a high pass filter, cut-off frequency 2000 Hz, detected, and applied to a threshold trigger. This was connected to a yellow LED. This was activated by the presence of high frequency energy.

The LED's were arranged in a vertical column and attached to spectacle frames. The user viewed the speaker's face with the LED's to one side of the speaker's mouth. The LED's appeared as patches of coloured light when they were illuminated.

METHOD

The object of the experiments was to investigate whether lipreaders would be able to recognise the consonants in VCV syllables more accurately with the aid than by lipreading alone. Furthermore the recognition accuracy was to be measured with the use of one, two, and three LED's in order to study the effects of making more information available to aid identification.

The speech material employed consisted of video recordings of the twelve consonants P,B,M,T,D,N,K,G,S,Z,F,V spoken in a /aCa/ context. The recording consisted of six lists of all twelve stimuli, with the stimuli arranged in different orders in each list.

Experiments

In the first test, the subjects were asked to to try to identify each consonant by lipreading alone. They were provided with a response sheet showing the alternatives and were required to tick the consonant which appeared to have been spoken. They observed all the lists in order, beginning with list one.

In the second test, the subjects wore the aid but only the red light, indicating voicing, was illuminated. They were asked to read a short paragraph explaining that the light was illuminated for the vowels and the voiced consonants such as the nasals and voiced fricatives. The light went off briefly for the voiced plosives, B,D,G, and was extinguished for the voiceless plosives, P,T,K, and the voiceless fricatives. The subjects observed all six lists beginning with list two.

In the third test, the green light, indicating voiceless energy, was illuminated in addition to the red, voicing light. It was explained that this light should be illuminated briefly for the voiced plosives, more substantially for the voiceless plosives and F, and would fill the interval between the vowels for the fricative S. Again all six lists were used, but this time beginning with list three.

In the fourth test, all of the three lights were used. It was explained that the third, yellow light responded to intense, high frequency sound. This meant that the red and yellow lights were illuminated during Z, and the green and yellow lights during S. The subjects watched all six lists beginning with list four.

In the final test, the subjects repeated test one, lipreading alone, in order to check if there had been any learning effects. In this last test they began with list five.

Subjects

Nine subjects took part in the tests. They were volunteers studying psychology at university. They were paid for participating in the tests. They all reported that they had normal hearing and vision. Five females and four males took part.

They were tested individually, with each test being administered on approximately successive days. They sat about 1.5 m in front of a television monitor whose screen size was 45 cm. The face of the speaker occupied most of the screen. The tests took place in a room with the lights extinguished but with sufficient light coming through the blind for the subjects to read the response sheets.

The interstimulus interval was about 10 s, giving plenty of time for a response to be made. Each session lasted about 15 min.

RESULTS

The consonant recognition scores for the individual subjects are shown in Table 1. The mean score for lipreading alone was 32.7%. The addition of the voicing cue increased this to 44.2%. This increase is significant at the 0.01 level ($t=2.76$). Adding a second cue to indicate the presence of voiceless sound increased the recognition score to 52.8%. Again this increase is significant. This time at the 0.05 level ($t=1.78$). However, with the addition of a third cue, indicating the presence of high frequency energy, the score remained at 52.8%. (The actual increase was 0.05%, $t=0.011$, n.s.). There was was some learning during the course of the experiment. The second test with lipreading

alone produced a mean score of 37.2%, an increase of 4.5% compared with the first test (significant only at the 0.1 level, t=1.483).

Table 1. Consonant recognition scores (%) for (A) lipreading alone, (B) with the voicing indicator, (C) with the voiceless indicator, (D) with the high frequency energy indicator, and (E) lipreading alone.

Subject	A	B	C	D	E
SD (M)	31.9	47.2	48.6	45.8	31.9
LB (F)	34.7	58.3	54.2	62.5	41.7
JD (F)	36.1	62.5	65.2	68.1	45.8
JW (M)	33.3	43.1	63.9	65.3	37.5
LN (F)	34.7	48.6	55.6	51.9	44.4
DC (M)	23.6	29.2	40.3	40.3	36.1
EM (F)	38.9	31.9	54.2	36.1	31.9
PW (F)	37.5	36.1	38.9	41.7	41.7
AM (M)	23.6	41.2	54.2	63.9	23.6
Mean	32.7	44.2	52.8	52.8	37.2
S.D.	5.24	10.62	8.55	11.62	6.74

There was a difference in the performance of male and female subjects as shown in Table 2. The females were better at lipreading alone by an average of some 8.3% (t=3.35, significant at the 0.01 level) and they remained superior with the addition of the voicing cue (7.3% difference, t=0.963, n.s.). The addition of the second cue narrowed the difference to 1.8% (t=0.295, n.s.), whilst the third cue raised the male recognition score to 1.7% above that of the female subjects (t=0.20, n.s.). The second lipreading alone test showed approximately equal learning effects for both male and female subjects. The difference became 8.8%, compared with a difference of 8.3% in the first test (significant at the 0.01 level, t=3.366).

VISUAL PRESENTATION OF VOICING

Table 2. Comparison of male and female recognition scores (%) for the various experimental conditions (see Table 1).

	A	B	C	D	E
Male	28.1	40.2	51.8	53.8	32.3
S.D.	4.53	6.70	8.58	10.96	5.42
Female	36.4	47.5	53.7	52.1	42.5
S.D.	1.63	11.97	8.44	12.07	2.34
Difference (F-M)	+8.3	+7.3	+1.8	-1.7	+8.8

Confusion matrices were constructed for each of the conditions. From these the confusions of the phonetic features were derived. Table 3 shows the voicing confusions. With lipreading alone there is no indication of voicing, so it is expected that this feature would be recognised at chance level. It was actually perceived correctly 53.4% of the time. With the voicing light the number of correct responses increased to 68.4%. With the voiceless energy indicated in addition, the proportion of correct responses increased to 77.3%. The addition of a high frequency energy indicator, however, reduced this to 71.2%

Table 3. Percentage voicing confusions with (a) lipreading alone, (b) with voicing indicator, (c) with voiceless energy indicator, and (d) with high frequency energy indicator.

	Stimulus	Response	
		Voiced	Voiceless
(a)	Voiced	54.4	45.5
	Voiceless	48.0	52.0
(b)	Voiced	66.2	33.8
	Voiceless	28.7	71.3
(c)	Voiced	76.3	23.7
	Voiceless	21.4	78.6
(d)	Voiced	69.4	30.6
	Voiceless	26.4	73.6

It might be expected that place of articulation would be easy to recognise by lipreading alone. However, it is difficult to distinguish alveolar and velar consonants, as shown in Table 4. The effects of the aid were negligible. With lipreading alone place was correctly recognised in 85.6% of the stimuli, with the voicing indicator 83.7%, with the voiceless energy indicator 83.6%, and with the high grequency energy 84.5%.

Table 4. Percentage place confusions with (a) lipreading alone, (b) with voicing indicator, (c) with voiceless energy indicator, and (d) with high frequency energy indicator.

	Stimulus	Response			
		Bilabial	Alveolar	Velar	Labiodental
(a)	Bilabial	96.9	0.4	0	2.7
	Alveolar	0.7	94.4	4.2	0.7
	Velar	5.7	54.6	34.0	5.7
	Labiodental	0	1.9	0	98.1
(b)	Bilabial	98.1	1.3	0	0.6
	Alveolar	3.4	91.2	4.2	1.2
	Velar	4.7	61.7	30.8	2.8
	Labiodental	0	2.8	0.9	96.3
(c)	Bilabial	92.9	6.5	0	0.6
	Alveolar	1.5	89.8	6.4	2.3
	Velar	4.0	56.4	38.6	1.0
	Labiodental	1.9	1.9	0	96.2
(d)	Bilabial	95.5	1.9	0	2.6
	Alveolar	0.9	93.1	2.0	4.0
	Velar	5.6	59.3	34.3	0.8
	Labiodental	0	0	0	100.0

Table 5 shows the manner confusions for each condition. For lipreading alone, manner was correctly recognised in 59.1% of the stimuli. This increased to 70.1% with the voicing indicator. The addition of the voiceless energy indicator increased this to 76.7%. With the presence of high frequency energy indicated as well, manner was correctly recognised 77.2% of the time.

DISCUSSION

The twelve consonants employed in the test fall into three visual categories, the bilabials (P,B,M), the labiodentals (F,V), and the rest (T,D,N,K,G,S,Z). It might be expected, therefore, that a recognition score of 25% would be obtained. The actual score of 32.7% was somewhat better than this, although two subjects scored only 23.6%.

Adding the voicing light should enable P and B to be distinguished from M because the light should be extinguished prior to the release of the plosive burst. However, this is unreliable as B is often prevoiced in intervocalic position. Similarly N and Z, which are voiced throughout, should be distinguishable from T,D,K and G, and V from F. This gives six categories, and predicts a recognition score of 50%. The actual score was 44.2%, but two subjects scored over 50%.

Table 5. Percentage manner confusions with (a) lipreading alone, (b) with voicing indicator, (c) with voiceless energy indicator, and (d) with high frequency energy indicator.

	Stimulus	Response		
		Plosive	Nasal	Fricative
(a)	Plosive	47.0	22.2	30.8
	Nasal	48.1	41.7	10.2
	Fricative	9.8	2.0	88.2
(b)	Plosive	57.9	17.4	24.7
	Nasal	24.5	67.3	8.2
	Fricative	9.1	1.8	89.1
(c)	Plosive	65.1	14.2	20.5
	Nasal	19.8	75.5	4.7
	Fricative	4.2	1.4	94.4
(d)	Plosive	62.5	17.8	19.7
	Nasal	18.0	82.0	0
	Fricative	2.4	0.5	97.1

The light indicating voiceless energy should enable more distinctions to be made. P has a longer burst of voiceless energy than B, although this cue is unreliable. Similarly D and G might be distinguished from T and K, and V from F. S is distinguishable from the other members of its group because of the continuous presence of voiceless energy between the vowels. This suggests that about nine categories might be distinguished, leading to a maximum recognition score of 75%. In practice the analyser did not produce the expected visual cue all of the time because of stimulus variability. Tests showed that the expected cue was generated about for about 90% of the stimuli, so the expected recognition score was 67.5%. The measured score was only 52.8%, although one subject achieved 65.2%. It would appear that both cues can be used, but not with 100% efficiently.

The light showing the presence of high frequency sound independently of the presence of voicing enabled Z to be distinguished from N, making ten categories and a predicted recognition score of 83.3%. With 90% reliability this reduces to about 75%. The actual score was 52.8%, with the best subject scoring 68.1%. Although some subjects improved with the additional cue, the performance of others deteriorated. This suggests that two or three visual cues representing the presence or otherwise of acoustic/phonetic features perhaps represents the maximum useful number in a lipreading aid.

CONCLUSIONS

It has been shown that a lipreading aid which attempts to present visually cues to the identification of consonants raised the recognition score of VCV syllables from 32.7% with lipreading alone to 52.8% with the aid. It is not yet known whether this increased performance will be retained with meaningful

sentences.

It was found that the lipreading performance of the female subjects who took part in the experiments was about 8% higher than that of the male subjects. However, the lipreading aid increased the performance of the male subjects more than that of the female subjects, so that with the aid their performances were very similar. This suggests that aids may be of more assistance to poor lipreaders.

Acknowledgements
The assistance of Dr G.V.Prosser with many pilot experiments is gratefully acknowleded. The video recordings were provided by Dr.Q.Summerfield and J.Foster of IHR, Nottingham. The work was supported by the Medical Research Council.

REFERENCES

[1] H.Upton, 'Wearable eyeglass speech reading aid', Proc. Conf. on Speech Analyzing Aids for the Deaf, Am.Ann.Deaf, Vol. 113, 222-229, (1968).

[2] C.E.Sherrick, 'Basic and applied research on tactile aids for deaf people: Progress and prospects', J.A.S.A., Vol. 75, 1325-1342, (1984).

[3] A.J.Fourcin, S.M.Rosen, B.C.J.Moore, E.E.Douek, G.P.Clarke, H.C.Dodson, and L.H.Bannister, 'External electrical stimulastion of the cochlea: clinical, psychological, speech-perceptual and histological findings', Brit.J.Audiol., Vol. 13, 85-107, (1979).