COMPARISON BETWEEN AUDITIVE AND AUDIO-VISUAL PERCEPTION OF PB-WORDS MASKED WITH WHITE NOISE

Carl Ludvigsen

1. Introduction

The basis of this paper is furnished by some experiments carried out two years ago at The State Hearing Center, Bispebjerg Hospital, Copenhagen. The purpose of these experiments was in broad outline to examine the influence on the discrimination score when PB-words were presented to listeners with a normal hearing belonging to three different age-groups, and the words were presented with and without the possibility of seeing the face of the speaker. A report on this experiment is given by Ewertsen et al. (1970).

The aim of the present paper is to study the answers obtained from the subjects under the two modes of presentation and four different signal to noise ratios (S/N).

1.1. The word material

The stimulus material consisted of four phonetically balanced lists (A, B, C, D) each containing 25 words. Due to the phonetic balancing the lists contained only mono- and disyllabic words, all commonly used Danish words. All disyllabic words had a trochaic stress pattern with the first syllable stressed and the second unstressed.

1.2. The presentation

A recording of the lists read by a male speaker whose dialect was close to standard Copenhagen was made on a video tape recorder for use on internal television. The speech signal from the sound track on the video tape was fed through an attenuator and then mixed with noise. The noise level was kept constant throughout the experiment and different signal to noise ratios were obtained by attenuating the speech signal. The noise was approximately "white" within the audible frequency range. The sound signal was presented monaurally through a pair of headphones. The visual signal i.e. a frontal picture of the speaker's face could be presented synchronously with the acoustic signal on a TV-screen in front of the subject.

Each of the four lists was presented four times to every subject. The first presentation was given through the headphones, with no picture on the TV-screen, for the second presentation the TV film was added, and finally approximately one month later these two presentations were repeated.[1] The S/N's were different for the four lists: list A: S/N = -20 dB, list B: S/N = -10 dB, list C: S/N = 0 dB, and list D: S/N = +10 dB.

1.3. The subjects

28 subjects participated in the experiment. For the sake of studying the influence of age on the discrimination score these subjects were selected from different age groups. Thus, 9 subjects were from 20-25 years old, 10 from 45-55 years old, and 9 from 65-75 years old. Audiograms were taken of all subjects, and only subjects with audiograms normal for their group of age participated in the experiment. For the present study, however, it was decided to disregard answers from the oldest group mainly because of the pronounced hearing losses at high

---

1) This was done in order to study the effect of retesting.

frequencies which are typical for their age-group. Furthermore two persons from the group 45-55 years were discarded because of hearing losses of respectively 30 and 35 dB for high frequencies. The remaining 17 subjects had audiograms differing less than 25 dB from the normal threshold in the frequency range from 125 Hz to 8000 Hz.

1.4. Experimental procedure

The subject was placed in a quiet room normally used for audiometry. The experimenter was placed in an adjacent room from where he could watch the subject through a window. Near the subject was placed a microphone which allowed the experimenter to hear the replies from the subject. The subject was instructed in the testing routine and was asked to repeat the words as he heard them. The experimenter registered whether the subject was able to repeat the words and if the subject answered incorrectly, i.e. with a word which was not identical to the stimulus, he noted this word on the list.

2. Analysis of answers

As the subjects were not forced to answer, three alternative types of reply are possible: 1) the subject may not have answered or he may have answered 2) with the correct word or 3) with an incorrect word. In the first case no information is obtained about the perception of the stimulus. In the second case some uncertainty exists concerning the cues used for identification of the word. Thus it is possible that a certain feature of a stimulus may not be detected by the subject although a correct answer is given, since the identification may have been based solely upon other features. The most useful source of information about the mechanism of perception seems to be the incorrect answers. Thus a comparison between an incorrect answer and the stimulus word provides information about which cues are detected and which are not.

2.1. Detection of number of syllables

If we assume that the subjects are unable to detect the number of syllables we shall expect the number of syllables found in the incorrect answers to be distributed approximately in the same way as in a normal word material and independently of the number of syllables in the stimulus word. This hypothesis can clearly be rejected from the material: Even at the most unfavourable S/N in the auditive tests the detection of number of syllables is very accurate. This is shown in TABLE 1 below.

TABLE 1

STIMULUS	NUMBER OF WORDS	NUMBER OF PRESENTATIONS	NUMBER OF CORR. ANSW.	NUMBER OF NO ANSW.	NUMBER OF WRONG ANSW.	NUMBER OF WRONG ANSW. WITH 1 SYLL.	NUMBER OF WRONG ANSW. WITH 2 SYLL.
2 SYLL. WORDS	12	408	52	298	58	0	58
1 SYLL. WORDS	13	442	33	360	49	47	2

TABLE 1. Answers from all 17 subjects pooled (list A, auditive test, S/N = -20 dB)

Table 1 shows that although more than 50 % of the answers are incorrect almost all of these contain the correct number of syllables. This finding also indicates that addition of the visual signal will not improve the detection of syllables appreciably.

2.2. Detection of the unstressed vowels in the disyllabic words

A cursory glance at the incorrect answers tells that the subjects rarely fail to detect the vowel in the second, unstressed syllable of disyllabic words. This impression holds true even for the tests with the smallest S/N and with auditive presentation only. In list A the second syllable of 11 of the 12 disyllabic words were of the type (C)ə(C) (zero or one consonant + schwa + zero or one consonant) and only one ended in a different vowel, viz. ʌ (the word was "tænder" (teeth)). To this word were given 10 answers; six of these were incorrect but all ended in unstressed ʌ. To the remaining 11 words 104 answers were obtained 52 of which were incorrect. 51 of these ended in Cə(C) and only one in a different vowel (unstressed, short i).

This very accurate detection of the unstressed vowels was to be expected: since approximately 90 % of Danish disyllabic words with stressed first syllable have schwa as unstressed vowel, the a priori uncertainty about the identity of the unstressed vowel is relatively small. On the other hand a reduction in intensity or duration of unstressed vowels might be expected to make the identification difficult.[2]

2.3. Confusions between stressed vowels

A larger percentual number of errors occur among the stressed vowels. And, of course, the nature of the errors depends heavily on the mode of presentation, auditive or audio-visual, and the S/N. In order to study the confusions among stressed vowels the words were grouped with respect to the acoustic quality of the stressed vowel. Obviously, this grouping can be done

---

2) Intensity curves of the stimulus material showed no reduced intensity for the unstressed vowels.

in more or less detail. For the present purpose the grouping was based on the frequencies of the two lowest formants, F_1 and F_2, as found from a frequency analysis of a tape recording of the stimulus material. The formant frequencies of vowels with marked transitions of F_1 and F_2 were measured at the most steady part of the vowel (generally in the middle) or, if no such part could be found (as in diphtongs), in the first part of the vowel. The letters used for transcription are given below with a few examples.

i	(fine, tit, bi)	[fi:nə, tʰit, bi·ʔ]
e	(dele, fedt, sne)	[de:lə, fet, sne·ʔ]
ɛ	(sæbe, mælk, æg)	[sɛ:bə, mɛlʔk, ɛ·ʔk]
æ	(stave, værst, sal)	[sdæ̈:və, væɒst, sæ̀·ʔl]
a	(nat, trække)	[nat, tʀage]
α	(aften, mig, leg)	[αfdən, mαi, lαiʔ]
ɑ	(varme, brand, barn)	[vɑ:mə, brɑnʔ, bɑ·ʔn]
y	(lyve, ny)	[ly:və, ny·ʔ]
ø	(løbe, dyppe, sø)	[lø:bə, døbə, sø·ʔ]
œ	(køn)	[kœnʔ]
ɶ	(gør, tørstig)	[gɶʀ tʰɶʀsdi]
u	(bule, skulder, jul)	[bu:lə, sgulɒ, ju·ʔl]
o	(hoste, sko)	[ho:sdə, sgo·ʔ]
ɔ	(kåbe, stå)	[kʰɔ:bə, sdɔ·ʔ]
ʌ	(slot, øjne)	[slʌt, ʌinə]
ɒ	(går, får)	[gɒ·ʔʁ], [fɒ·ʔʁ]

After the grouping confusion matrices were formed: one for each age-group and mode of presentation. As the confusions were distributed in the same manner for the two age-groups the answers from these were pooled for the subsequent examinations. From these matrices it followed that with a few exceptions the vowels in the incorrect answers were either identical to the stimulus vowel or (especially for a low S/N) with approximately the same first formant frequency. This observation is illustrated in Figs. 1-6.

Results obtained at different S/N cannot immediately be compared since they come from different stimulus words. In order to make such a comparison meaningful, more information about the stimulus material is given in TABLE 2. This table shows for the three lists A, B, and C separately the total number of times a word containing a specific vowel was presented to a subject (e.g. in list A two words have /i/ as stressed vowel; these words are presented twice to seventeen subjects; consequently, the total number of presentations is 68). TABLE 2 also shows the total number of answers containing the correct vowel. The results are given for auditive (upper figures for each vowel) as well as audio-visual presentation (lower figures).

In Figs. 1-6 the number of mistakes between vowels is indicated by different types of lines according to the signature. An arrow on the line points towards the incorrect vowel.

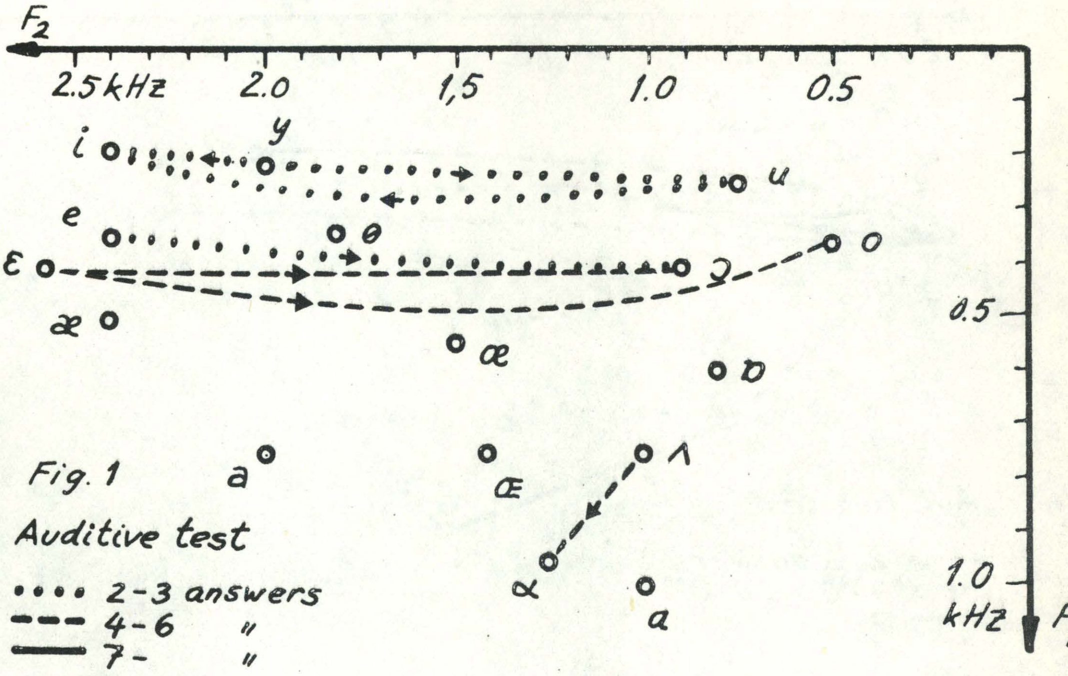
Fig. 1 shows confusions observed at the auditive presentation of list A (S/N = -20 dB). It follows that mistakes occur mainly between vowels with approximately identical first formant frequencies.

Fig. 2 shows that for the same words and the same S/N but with audio-visual presentation no confusions occur between rounded and unrounded vowels. This is in part surprising since no marked difference in lip configurations is present when pronouncing ʌ and ɑ. The explanation seems to be that all words containing ʌ in list A also contain a bilabial stop consonant. This is an easily detected visual cue and therefore the wrong answers observed in the auditive test will not appear in the audio-visual test as they contain no such consonants.

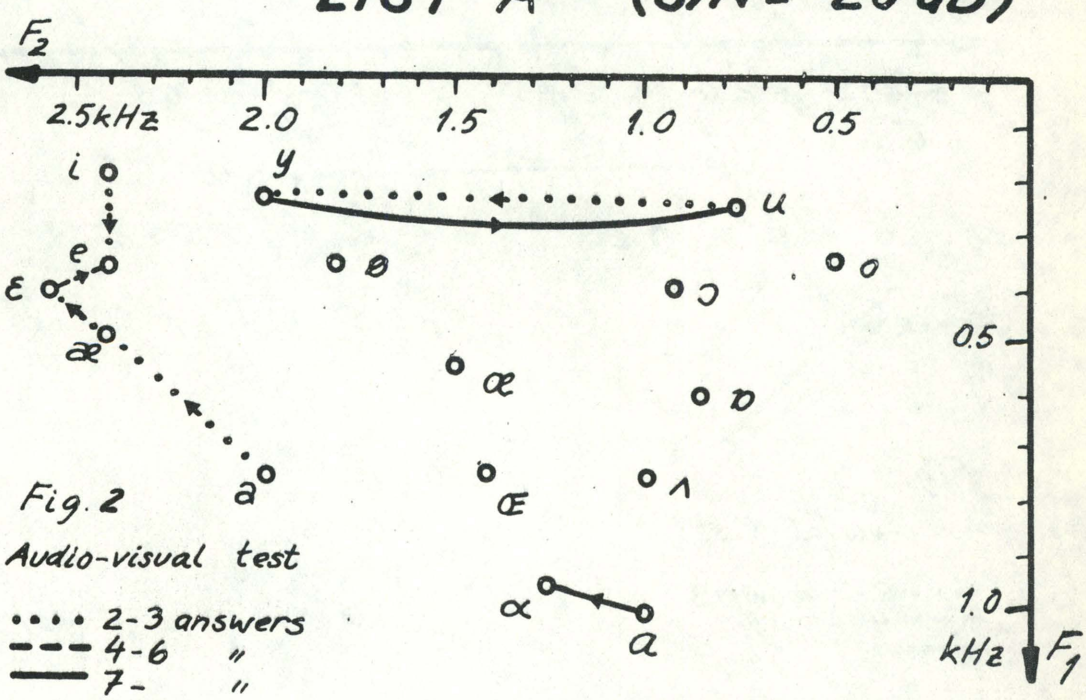
TABLE 2

	LIST A (S/N = -20 dB)				LIST B (S/N = -10 dB)				LIST C (S/N = 0 dB)			
	NUMBER OF: PRES.	ANSW.	NUMBER OF: CORR. ANSW.	CORR. VOWEL	NUMBER OF: PRES.	ANSW.	NUMBER OF: CORR. ANSW.	CORR. VOWEL	NUMBER OF: PRES.	ANSW.	NUMBER OF: CORR. ANSW.	CORR. VOWEL
i	68 68	8 50	1 22	6 48	34 34	16 24	1 6	7 24	68 68	66 68	58 68	65 68
e	68 68	19 55	8 37	15 53	136 136	99 131	55 114	77 127	136 136	114 131	67 114	98 122
ɛ	170 170	34 144	14 124	23 139	136 136	84 127	42 108	62 124	102 102	97 102	87 101	91 102
æ	68 68	17 58	5 40	14 55	34 34	21 30	1 11	20 20	68 68	63 67	50 58	63 67
a	34 34	5 24	1 4	4 16	68 68	39 62	7 54	20 58	34 34	30 34	29 34	29 34
α	102 102	35 96	29 92	32 94	34 34	29 28	6 15	17 28	68 68	65 68	62 66	65 68
ɑ	34 34	3 16	2 5	2 6	34 34	22 31	17 30	22 31	68 68	67 68	62 68	66 68
y	34 34	8 25	2 10	3 14	- -	- -	- -	- -	34 34	34 34	34 34	34 34
ø	- -	- -	- -	- -	34 34	20 30	6 23	15 30	- -	- -	- -	- -
œ	- -	- -	- -	- -	- -	- -	- -	- -	34 34	31 33	28 32	28 32
ɶ	- -	- -	- -	- -	34 34	27 33	16 30	24 33	- -	- -	- -	- -
u	68 68	5 38	0 27	0 31	68 68	53 66	46 60	48 61	34 34	23 32	17 31	19 32
o	34 34	8 28	2 21	5 25	68 68	59 67	47 62	57 67	34 34	33 34	31 34	33 34
ɔ	34 34	16 29	5 19	12 28	34 34	30 34	23 34	30 34	34 34	32 32	21 27	30 31
ʌ	102 102	26 94	13 76	18 93	102 102	74 101	58 96	67 101	102 102	99 99	81 91	90 99
ɒ	34 34	13 28	8 22	12 27	34 34	29 32	21 31	26 32	34 34	33 34	33 34	33 34

LIST A (S/N=-20 dB)



LIST A (S/N=-20 dB)



LIST B (S/N = -10 dB)

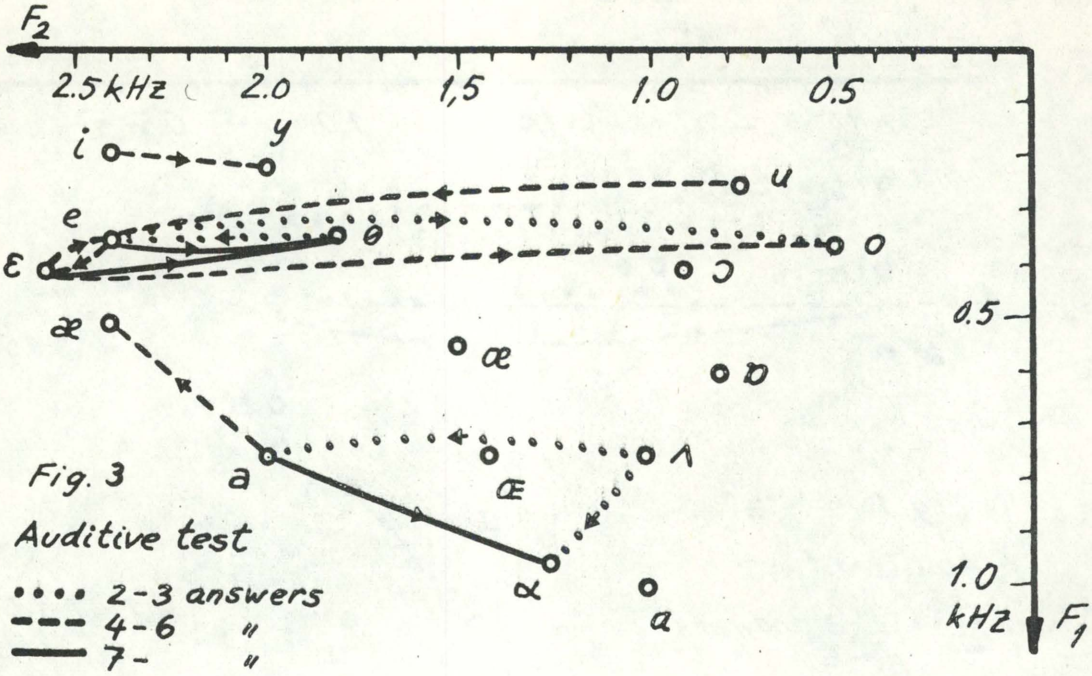


Fig. 3

Auditive test

LIST B (S/N = -10 dB)

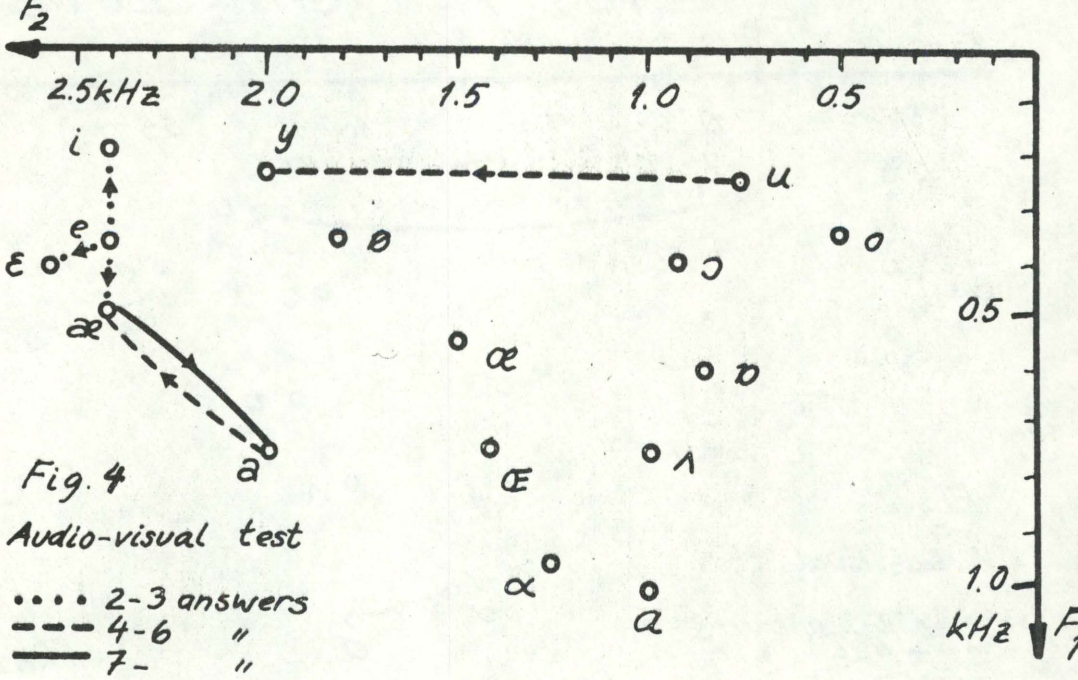
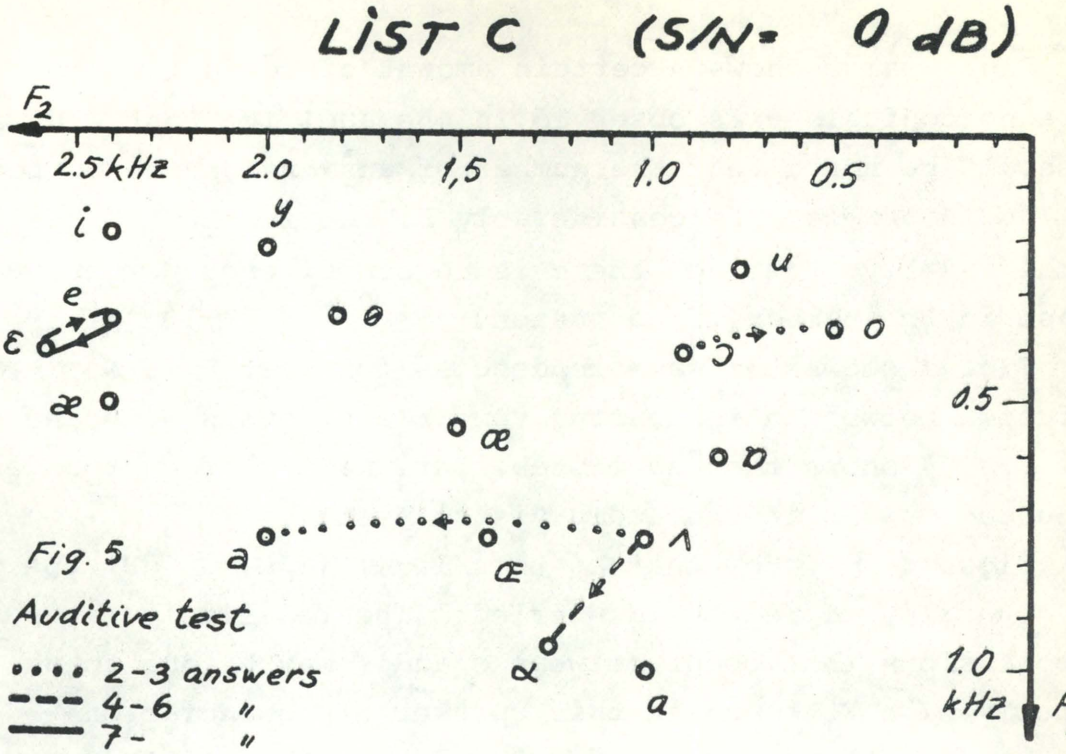


Fig. 4

Audio-visual test


LIST C (S/N= 0 dB)
F2
2.5kHz
2.0
1,5
1.0
0.5
0.5
1.0
kHz
F1
i
y
u
e
ɛ
ø
o
ɔ
æ
œ
ɒ
a
Œ
ʌ
α
ɑ
Fig. 5
Auditive test
•••• 2-3 answers
--- 4-6 "
—— 7- "

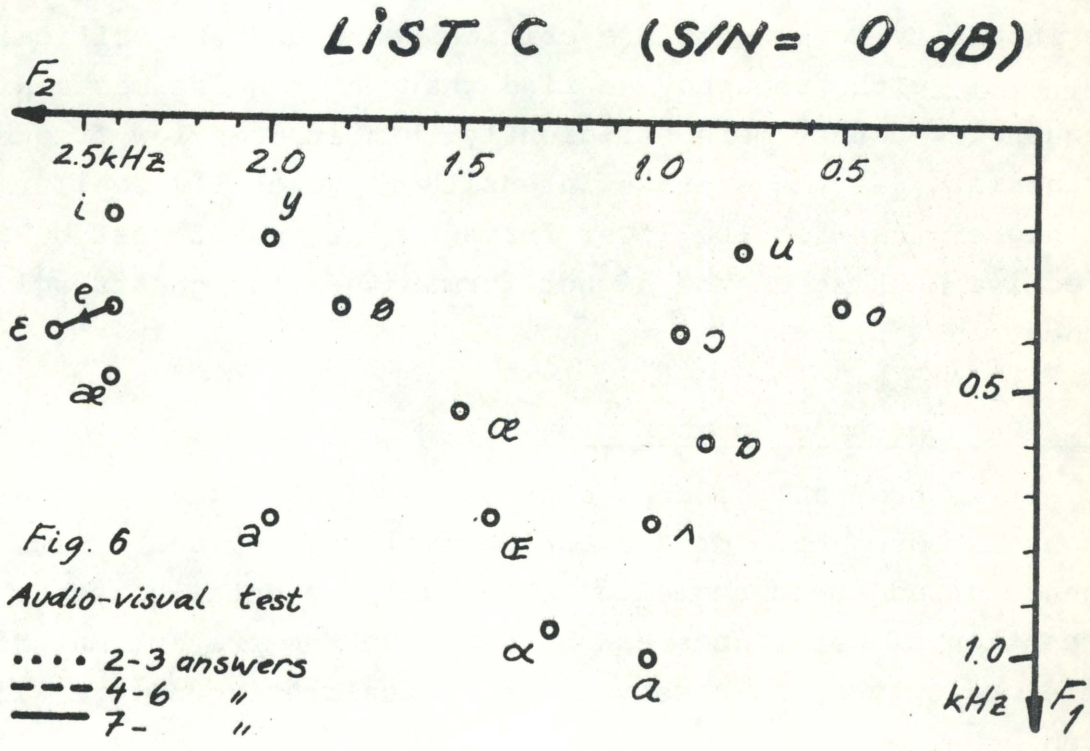

LIST C (S/N= 0 dB)
F2
2.5kHz
2.0
1.5
1.0
0.5
0.5
1.0
kHz
F1
i
y
u
e
ɛ
ø
o
ɔ
æ
œ
ɒ
a
Œ
ʌ
α
ɑ
Fig. 6
Audio-visual test
•••• 2-3 answers
--- 4-6 "
—— 7- "

Fig. 2 also shows a certain amount of confusion among vowels where no confusion was observed in the auditive test. However, it should be noted that the number of answers given (correct as well as incorrect) are considerably higher in the A-V test than in the auditive test and there is no significant percentual increase in the number of confusions.

Fig. 3 shows the same tendencies for list B (S/N = -10 dB): confusion between neighbouring vowels and vowels with the same F1.

Fig. 4 shows that no mistakes are made between rounded and unrounded vowels in the audio-visual test.

Figs. 5-6 show that at this level (S/N = 0 dB) the vowels are generally correctly identified. The relatively large number of confusions that occur between e and ε may be due to an exceptionally high F_1 found in this speaker's pronunciation of e.

No diagrams are shown for the test with list D (S/N = +10 dB) since no confusion between vowels was observed.

The results obtained here fit with the well-known masking proporties of white noise. As masking is roughly determined by the intensity level within a critical band and the critical bandwidth grows with frequency, we find that white noise masks high frequencies rather more efficiently than it does low frequencies. Furthermore, as the average intensity is generally smaller for the higher than for the lower formants the result must be an effective masking of the higher formants (although strongly dependent on S/N).

2.4. Perception of consonants

It is generally accepted that the identification of a consonant is based both on the consonantal segment and on its influence on adjacent segments. Obviously no detailed study of the perception of consonants can be based on the present material. Therefore, only a few observations of qualitative nature will be given here.

2.4.1. Labials

The most notable observation is the difference in discrimination of bilabials (p, b, m) as well as labio-dentals (f, v) found for the two types of presentation. While the discrimination of these consonants (especially that of f and notably for S/N $\leq$ -10 dB) is remarkably poor in the auditive tests, the discrimination in the audio-visual tests is very high, even for S/N = -20 dB. Generally the audio-visual detection of f, v, and m (word initially) is almost perfect, i.e. even in the incorrect answers these consonants are always found in the correct positions when they were present in the stimulus word. The consonants p and b are often mutually confused for S/N $\leq$ -10 dB but if a bilabial stop occurs in the stimulus word then a bilabial stop will be found even in incorrect answers.

This finding agrees well with earlier observations. The auditive discrimination between voiced and voiceless consonants in white noise is rather good (see e.g. Miller and Nicely 1954) and so is the visual detection of bilabials and labiodentals (see e.g. Woodward and Barber 1960). A generalization of these findings to audio-visual perception agrees with the above mentioned observations (remembering that Danish b is voiceless).

2.4.2. Voiceless fricatives

Another interesting observation is that voiceless fricatives are very hard to detect in white noise. This is not very surprising, but furthermore voiceless fricatives (especially s) as well as the affricated stop t occur in the incorrect answers where the stimulus word had no such consonants. This must be due to the pronounced similarity in acoustic quality between white noise and fricative sounds.

2.4.3. Voiced consonants

At a S/N = -20 dB wrong answers to stimulus words containing voiced consonants (m, n, ŋ, l, v, j, ð, r, ɣ) generally contain voiced consonants but with several confusions between them. At more favourable S/N the consonants are discriminated more accurately. The material gives no support to the theory that nasals are detected as a separate group as found by Miller and Nicely (1955). One reason for this could be that in American English vowels are generally strongly nasalized in nasal surroundings.

2.4.5. Stops

The stops in Danish are all voiceless and the difference between b, d, g and p, t, k is mainly one of aspiration (t is also somewhat affricated). The material from the auditive tests gives no indication that confusion within these groups[3] should be more likely than confusion between the groups or even with other consonants.

3. Conclusion

The results obtained in this paper show in agreement with earlier findings (see e.g. O'Neill 1954, Sumby and Pollack 1954) that the visual signal of a speaker's face considerably improves the detection of certain speech segments especially when the signal to noise ratio is unfavourable. The improvement is particularly conspicuous in the detection of bilabials and labio-dentals but also in separating rounded vowels from unrounded. The results are obtained from a discrimination test of isolated words and it may be expected that the influence of the visual signal is less pronounced in the perception of running speech, since e.g. the syntactical structure of preceding strings will make detection of certain segments redundant. And furthermore the articulation will generally be less distinct.

---

3) viz. the group b, d, g and the group p, t, k.

It is also shown that the masking of white noise is not uniformly distributed over the frequency range and that the masking is one of higher frequency components mainly. Thus if the background noise is to give approximately the same decrease in redundancy for all components of the speech signal, another type of masking sound with less intensity in the higher frequency region must be used.

References

Ewertsen, H.W., Nielsen, H.B., and Nielsen, S.S. 1970: "Audio-visual Speech Perception", Acta Otolaryngologica 263, p. 229-230.

O'Neill, John J. 1954: "Contributions of the Visual Components of Oral Symbols to Speech Comprehension", Journal of Speech and Hearing Disorders 19, p. 429-439.

Miller, G.A. and Nicely, P.E. 1955: "An Analysis of Perceptual Confusions Among Some English Consonants", Journal of the Acoustical Society of America 27, p. 338-352.

Sumby, W.H. and Pollack, I. 1954: "Visual Contribution to Speech Intelligibility in Noise", Journal of the Acoustical Society of America 26, p. 212-215.

Woodward, M.F. and Barber, C.G. 1960: "Phoneme Perception in Lipreading", Journal of Speech and Hearing Research 3, p. 212-222.