IDENTIFIABILITY OF VOWELS IN SHORT SEGMENTS FROM FREE CONVERSATIONAL SPEECH

by H.J.A.G. Buiting

1. INTRODUCTION

The experiment which will be reported here forms part of a larger study which aims to find out to what extent Dutch vowels can be identified from very short segments of free conversational speech. An earlier study (Koopmans-van Beinum, 1980) revealed that the percentage of correct identifications (the identifiability) of unstressed Dutch vowels from free conversation, averaged over all vowels, is some 30%, whereas it was predicted that the identifiability of stressed vowels would be some 50%. The present experiment focusses on the question whether or not, and if so to what extent, additional information is comprised in the transitions from or to the surrounding consonants i.e. the identifiability of vowels as a function of the addition of transition is considered. The influence of different consonantal contexts on the identifiability of vowels is regarded as well. This experiment forms only the first part of the total investigation we have in mind with several more speakers. Therefore we restrict ourselves mainly to a description of the experiment and a presentation of the (preliminary) results.

2. METHODS

2.1 Stimuli

The stimuli were segments from free conversation of one trained male speaker. The conversation was recorded on a partable recorder (Tandberg Model H). Attention was payed to the identification of 8

vowels viz: /u/, /o/, /ɑ/, /œ/, /i/, /I/, /ɛ/ and /y/. To assess possible consonant dependency, vowels in systematically varying consonantal contexts were utilized. To this purpose 17 consonants were distributed over 10 groups according to articulation characteristics (see Table 1). For each of the 8 vowels mentioned before, a number of CVC combinations were chosen such that each group of consonants was represented. It appeared to be impossible to cover all groups for every vowel (see cells with ∅ in Table 1). In order to prevent that vowels of a too deformed nature would form part of the stimulus ensemble it was stipulated that the intended vowel was clearly recognizable within a short context (~1 sec.). In spite of this fairly selective procedure several CVC combinations had to be replaced as they appeared to be totally unrecognizable (creaky, very low pitch, strongly variating pitch etc.) after being stripped of their embedding contexts.

In this way 45 CVC combinations were selected (see Table 1). The CVC combinations were digitized (12 bit, sample frequency: 12.5kHz; prefilter: -3dB point at 4kHz, 24dB/oct).

The stimuli used in the listening tests were taken from the 45 CVC combinations by means of a speech editing system which enables determination of segmentation points by both visual and auditory inspection (Buiting, 1981).

Five categories were determined: the initial consonant (IC), the initial transition (IT), the stationary part (SP), the final transition (FT) and the final consonant (FC). The choice of these categories was based on the results of a pilot experiment. The influence on the identification scores of adding stepwise more and more of the transition to the stationary part appeared to be unclear. Therefore it was decided to treat the transitions as a whole.

Briefly the segmentation procedure was as follows:

First the beginning of the IC and the end of the FC were determined. In doing so great care was taken that no vowel could be heard (or seen) other than the one enclosed by the IC and FC. After that the stationary vowel part was established. It was considered to be that part of the vowellike portion which shows maximum amplitude and of

		p	t	k	b	d	g	f	s	v	z	m	n	l	r	w	j	h
BOER					/i										/f			
VOEL										/i				/f				
WOEN													/f			/i		
HOEF	/u/							/f										/i
GOEJ							/i										/f	
MOET			/f									/i						
WOR															/f	/i		
JOS									/f								/i	
TOP		/f	/i															
HON	/o/												/f					/i
DOG					/f		/i											
VOL										/i				/f				
WAS									/f							/i		
ZAL										/i				/f				
DAG					/i		/f											
HAL	/ɑ/													/f				/i
KAR				/i											/f			
JAN													/f				/i	
JUR															/f		/i	
WUS									/f							/i		
BUS					/i				/f									ø
ZUL	/œ/									/i				/f				
KUN				/i									/f					
SUUL									/i					/f				
JUUN													/f				/i	
ZUUR										/i					/f			
DUUR	/y/					/i									/f			
LUUT			/f											/i				
HUUW																/f		/i
ZIET			/f							/i								
BIEJ				/i													/f	
HIER	/i/														/f			/i
LIEF								/f						/i				
NIEW													/i			/f		
DIT			/f			/i												
MIS									/f			/i						
WIL	/ɪ/													/f		/i	ø	ø
RIN													/f		/i			
VIN										/i			/f					
HEP		/f																/i
VER										/i						/f		
BEN					/i								/f				ø	
WEL	/ɛ/													/f		/i		
FEK				/f				/i										
TEM			/i									/f						
INIT.		0	2	2	4	4	1	1	1	4	4	2	1	2	1	6	4	6
FIN.		2	4	1	0	0	2	2	5	0	0	1	8	8	8	2	2	0

Table 1

In the matrix the CVC combinations from which the stimuli are taken, are displayed along the vertical axis. The groups of consonants and their elements are displayed along the horizontal axis. A $/_i$ or $/_f$ denotes that the corresponding consonant and thereby the corresponding group is represented. A $/_i$ means that the consonant is initial, whereas a $/_f$ means that the consonant is final. A zero (ø) denotes that no representative could be found. The number of times every consonant shows up in the CVC combinations is given as well. Again initial and final consonants are treated separately.

which the periods change only slightly. Finally the end of the IC and the beginning of the FC were estimated. If possible it was chosen there where the characteristic periodic wave-pattern of a vowel passes into another kind of structure. If such a change in the wave-pattern could not be observed, which was common when w, j,h,r,l,m and n were involved, it was chosen arbitrarily midway between the beginning of the IC and the start of the SP or between the end of the SP and the end of the FC. The speech segments between the end of the IC and the beginning of the SP, and between the end of the SP and the beginning of the FC were considered to be the initial and the final transitions respectively.
With respect to the categories IC, IT, SP, FT and FC 9 speech segments were taken from every CVC combination viz:

SP	V	(V=vowel t=transition C=consonant)
IT+SP	tV	
SP+FT	Vt	
IT+SP+FT	tVt	
IC+IT+SP	CtV	
SP+FT+FC	VtC	
IC+IT+SP+FT	CtVt	
IT+SP+FT+FC	tVtC	
IC+IT+SP+FT+FC	CtVtC	

these speech segments are called items.
In gating out use was made of a cosine window with a duration of 2ms (25 samples) and with the zero level on the extremities of the segment involved. The items were all scaled with respect to the highest signal value occurring. In this way 9x45 = 405 items were prepared. They composed the stimulus ensemble in the perception experiment.

2.2 Subjects

The subjects were paid volunteers, naive with respect to the task and all native speakers of Dutch. All reported to have no hearing losses. 80 subjects took part in the experiment of which 73 were students and 7 were engaged in various professions.

2.3 Procedure

The 9 different speech segments as defined in section 2.1 represented 9 different stimulus conditions. The 45 items per condition were presented in random order. The order was different for each stimulus condition. The 9 stimulus conditions were presented separately in a fixed order (in principal from most difficult to most easy) equal to the one given in Table 3. The items were separated by a pause of 3 seconds. After presentation of each entity of 15 items a marking tone followed. Entities and marking tones were separated by a pause of 3 seconds as well. The whole was preceded by a set of 15 learning items, all belonging to the first (V) stimulus condition. The learning items too were separated by a pause of 3 seconds.
Before the experiment began the subjects were informed that in the course of the experiment consonants would become more often and more clearly audible also. The score documents contained 8 response alternatives that corresponded to the vowels used in the experiment (orthographically given). The subjects were asked to encircle the vowel they believed to hear. The stimulus condition consisting of the 'longest' items (CtVtC) was presented twice in different random order. The second time subjects were asked to write down the total CVC combination to see whether the consonants could be identified. The pause between the items was then 4 seconds. The listening tests were held over a period of four days at the 'Instituut voor Toegepaste Taalwetenschap.' (ITT) in Amsterdam (Tape recorder: Tandberg model 1021; headphones MBK 800). On each day a group of about 20 subjects was tested. The stimuli however were presented individually over headphones. The overall loudness was set at a subjectively determined comfortable level.

3. RESULTS

The identifiability of vowellike segments, (the percentage of correct responses), varied as a function of a number of factors. Four of them were 'controlled' i.e. the type of vowel, the initial conso-

	V	tV	Vt	tVt	CtV	VtC	CtVt	tVtC	CtVtC
BOER	44	74	55	75	96	70	91	80	85
VOEL	79	76	66	65	71	71	64	66	76
WOEN	73	78	56	68	81	71	65	79	73
HOEF	56	38	88	76	68	94	94	90	90
GOEJ	70	75	81	84	85	55	91	55	93
MOET	11	29	36	23	41	41	44	53	46
WOR	41	51	36	69	50	49	48	59	74
JOS	23	11	1	3	6	11	14	4	60
TOP	60	60	65	63	74	40	84	53	68
HON	91	94	86	96	98	55	98	85	86
DOG	75	74	81	88	75	70	93	66	93
VOL	80	88	74	93	84	79	98	90	91
WAS	18	9	13	5	39	36	25	33	50
ZAL	65	63	55	71	75	69	71	69	70
DAG	50	66	73	71	84	73	95	73	89
HAL	48	56	41	53	69	54	64	68	69
KAR	5	14	6	8	61	6	40	10	66
JAN	44	44	39	36	75	33	43	38	44
JUR	68	81	74	75	61	89	69	86	69
WUS	53	68	65	86	83	75	83	90	90
BUS	30	55	26	61	63	45	56	70	68
ZUL	15	35	25	39	19	39	14	66	23
KUN	44	78	80	90	69	84	83	88	86
SUUL	78	88	61	70	75	59	69	61	63
JUUN	83	84	79	85	80	75	79	75	80
ZUUR	73	79	35	31	73	23	56	53	60
DUUR	31	71	31	41	50	33	50	30	34
LUUT	61	70	86	83	74	89	93	96	96
HUUW	65	75	64	76	70	60	61	78	70
ZIET	45	63	84	80	89	91	96	89	94
BIEJ	21	49	10	19	39	8	38	15	16
HIER	15	20	4	9	31	3	24	11	18
LIEF	4	6	3	8	18	4	13	5	21
NIEW	14	19	16	20	11	10	31	25	31
DIT	84	85	83	89	89	89	76	95	74
MIS	24	38	15	26	29	15	21	46	59
WIL	5	1	3	3	0	4	1	9	8
RIN	63	60	66	64	63	30	54	46	55
VIN	70	83	55	81	55	59	43	71	45
HEP	36	45	28	28	71	25	45	31	33
VER	35	20	26	19	44	34	43	31	53
BEN	34	23	25	29	59	20	51	28	29
WEL	4	4	1	4	0	3	3	3	1
FEK	38	23	56	36	45	25	70	23	31
TEM	48	45	41	31	40	46	51	33	44

Table 2

The percentages of correct responses per item. All items in one column belong to one stimulus condition. Above every column the stimulus condition is given. In this and the following tables percentages are rounded off to whole numbers.

nant, the final consonant and what we called the stimulus condition. The 'values' the variables took were such as defined in section 2.1. It is clear that the variables initial/final consonant and stimulus condition were coupled. For the sake of clarity however we treat them separately.

In Table 2 the identifiability is given for all items. At first sight it seems to vary in quite an unstructured way. To get some insight in the way the identifiability varies as a function of one or more variables one has to average over the other variables. Three such procedures were carried out. They resulted in:

1 the identifiability as a function of the stimulus condition
2 the identifiability as a function of the stimulus condition and the type of vowel
3 the identifiability as a function of part of the stimulus condition (V,tV,CtV,Vt,VtC) and initial/final consonant.

In the following the outcomes of the first averaging procedure will be discussed in some detail, whereas the result of the other two procedures will be discussed only briefly.

ad 1 The average identifiability for each of the 9 stimulus conditions is given in Table 3. The effect of the stimulus conditions on the identifiability appeared to be significant ($p<.001$).

STIM. COND.	%
V	46
tV	52
Vt	47
tVt	52
CtV	58
VtC	47
CtVt	58
tVtC	54
CtVtC	58

GRAND MEAN: 52%

Table 3

The average percentages of correct responses as a function of the stimulus condition.

An additional Tukey-HSD analyses was carried out with respect to the 9 stimulus conditions. The homogeneous subsets are given in Table 4. From Table 4 it can be concluded that an extension to the 'left' (the beginning) very often results in a significantly better overall identifiability, whereas extensions to the 'right' (the end) do not, or hardly, change the overall identifiability of the vowel.

46	47	47	52	52	54	58	58	58
V	Vt	VtC	tVt	tV	tVtC	CtVt	CtV	CtVtC

Table 4

The homogeneous subsets that resulted from a Tukey-HSD analysis applied to the items with respect to the 9 stimulus conditions on 5% level. The identifiability per stimulus condition is given as well.

Therefore the conclusion is that on the average, at segmental level, the transitions from the preceding consonant to a vowel contributes to the identifiability of that vowel. It is noted that the amount of contribution to the identifiability is quite small compared to the overall variance.
Other variables apparently have a stronger impact on the identifiability of vowellike segments.

ad 2 The identifiability as a function of the type of vowel and the stimulus condition is given in Table 5. Though no variance analyses has been applied it seems fair to conclude that the identifiability is strongly influenced by the type of vowel. The results indicate that the particular vowel involved has a more pronounced effect on the identification score than the stimulus condition. The overall effect with respect to the stimulus condition is reflected here, deviations however occur.

Vowel	V	Vt	VtC	tVt	tV	tVtC	CtVt	CtV	CtVtC
/u/	55	64	67	65	61	71	75	74	77
/o/	62	57	51	69	63	60	73	65	79
/ɑ/	38	38	45	41	42	49	56	67	65
/œ/	42	54	66	70	63	80	61	59	67
/y/	65	59	56	64	78	66	68	70	67
/i/	20	23	23	27	31	29	40	38	36
/I/	49	44	39	53	53	53	39	47	48
/ɛ/	32	30	25	25	26	25	44	43	32

Table 5

The identifiability as a function of the type of vowel and the stimulus condition. The order of the stimulus conditions is the same as in Table 4.

ad 3 The identifiability as a function of the initial/final consonant and (part of) the stimulus condition is given in Table 6. The results are difficult to interpret because several groups are poorly represented c.f. /r/ in initial, and /w/ and /j/ in final position.

	p-,t-,k- (n=4)	g-,f-,s- (n=3)	m-,n- (n=3)	l- (n=2)	r- (n=1)	w- (n=6)	j- (n=4)	b-,d- (n=8)	v-,z- (n=8)	h- (n=8)
V	39	62	16	33	63	32	54	46	58	52
tV	49	62	28	38	60	35	55	62	63	55
CtV	61	68	27	46	63	42	57	69	64	68
	-p,-t,-k (n=7)	**-g,-f,-s (n=9)**	**-m,-n (n=9)**	**-l (n=8)**	**-r (n=8)**	**-w (n=2)**	**-j (n=2)**	**-b,-d (n=0)**	**-v,-z (n=0)**	**-h (n=0)**
V	48	37	61	47	39	39	46	-	-	-
Vt	63	40	61	41	34	40	46	-	-	-
VtC	57	47	54	47	38	35	31	-	-	-

Table 6

The identifiability as a function of the initial/final consonant and part of the stimulus condition. Initial consonants are shown first.

Nevertheless an interesting aspect was observed: it made quite a

different whether a group of consonants was initial or final with respect to its influence on the identifiability. This held for every appropriate stimulus condition c.f. /g,f,s/ and /m,n/. Again the effect seemed more pronounced than the influence of the stimulus condition, which nevertheless was reflected in the results.

For each of the 9 stimulus conditions confusion matrices were calculated. An attempt to interpret the results in terms of first and second choice did not lead to much so far. We hope to carry out such analysis fruitfully when the data of the forthcoming experiments are available.

The task with respect to the identification of the full CtVtC stimuli appeared to be too difficult. In most cases nonsense responses or no responses at all were given. This may have been caused by the fact that the CtVtC items were taken from the CVC combinations in such a way that only one vowel (the one enclosed by the consonants) could be heard, which possibly resulted in consonants too mutilated to be identifiable one way or another.

The results achieved so far indicate that all the investigated variables (vowel, consonantal context, stimulus condition) do significantly influence the identifiability of vowellike segments. This does not imply that the probability of a correct response can be predicted from knowledge of those variables. Two reasons may account for this:

- The influence of the variables on the identifiability may interact in a complicated manner, c.f. WEL, WUS (Table 2)
- The identifiability may very well be influenced by factors like intonation, prosody, tempo etc. that are not observed in this experiment and that depend upon the particular 'place' the segment occupies in current speech. These factors in their turn may of course have some interaction with each other and with the variables controlled by us.

The results and arguments presented so far are of a preliminary and tentative nature. More quantitive and firm data will be presented after prosecution of the forthcoming experiments with four speakers.

REFERENCES

Buiting, H.J.A.G. (1981). SESAM Speech Editing System Amsterdam. IFA-publication nr 70.

Koopmans-van Beinum, F.J. (1980). Vowel Contrast Reduction. An Acoustic and Perceptual Study of Dutch Vowels in Various Speech Conditions. Doctoral Thesis, University of Amsterdam.

ACKNOWLEDGEMENTS

This investigation is supported by the Netherlands Organization for Advancement of pure Research (Z.W.O.), project nr 17-21-19.
Wil Fagel was indispensable in the statistical analysis.