Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Perceptive evaluation of segmental cues

Carlson, R. and Granström, B. and Pauli, S.

journal: STL-QPSR
volume: 13
number: 1
year: 1972
pages: 018-024

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

II. SPEECH PERCEPTION

A. PERCEPTIVE EVALUATION OF SEGMENTAL CUES*

R. Carlson, B. Granström, and S. Pauli

Abstract

In our search for perceptually relevant cues in speech we have taken special interest in parts of the acoustic signal displaying rapid changes and short durations. In "tape splicing" experiments with natural voiceless stops we have found a greater amount of absolute identity of the explosion than might be expected from earlier studies. This finding should be taken into account in speech synthesis.

Using competing cues and selectively manipulated explosions we have a sensitive method of testing cues in identification tests.

For nasals and laterals, the importance of the consonant vowel boundary relative to the consonantal segment has been studied. Both place and manner cues are investigated by means of manipulated natural speech as well as speech synthesis experiments. Some 10 msec around the boundary proved to radically change the identity in natural speech. On the other hand, the detailed pattern of the consonantal segment seemed rather unimportant.

Introduction

Synthetic speech is a powerful tool in investigating the perceptual cues of speech. If we shall learn more about human speech there are however good reasons to check the results from such work with tests on natural speech, especially in cases where the synthesis equipment might have some insufficiences. We will deal with two such "cases". One is the brief explosions in stops which have quite a complex spectrum but short duration[1]. The other is the vowel consonant boundary in nasals and laterals that can be described as switching of complex resonators, which not readily is synthesized by standard formant synthesizers[2,3,4]. From the perceptual point of view there are reasons to suppose this study worth-while since auditory attention tends to focus on points of rapid change.

Stop cues in the stop explosion

Background

Most of the work on cues for stop consonants reviewed in the literature seems to be concerned with formant transitions and relative timing of voice onset. The brief explosion, often shorter than 20 msec, tends to attract less interest perhaps due to a belief that its relative distinctive importance

---

* This paper was presented at the International Conference on Speech Communication and Processing, Boston, Cambridge, USA, April 1972.

is rather low. A now classical experiment in Haskins Laboratories[5] showed that the burst was very context sensitive and displayed almost total overlap between [p] and [k] areas. These results are supported by a tape splicing experiment with natural speech[6]. Both experiments used vowels without transitions and this implicates that no manner cue except explosion was present in these stimuli.

However, there is good reason to suppose that the explosion as an event associated with total constriction and relatively distant from phoneme boundaries should be less coarticulated than for example the aspiration.

The interchangeability of explosion

The interchangeability of explosion was tested through manipulation of digitized natural speech. The explosion from [pɑ], [pi], [pu] was replaced by explosions (first 20 msec) from [kɑ], [ki], [ku] (nine possibilities). In one test the aspirated part was replaced by silence. The results from identification tests can be seen in Fig. II-A-1 together with the restructured results from Schatz' (ref.[6]) experiment. Rather great discrepancies can be seen partially depending on language differences but probably most due to the fact that Schatz used [k] explosions plus hVC, i.e. relatively transition free formants. In our experiment the absolute identity for [kɑ] [ku] explosions seems to overrule the often misleading transitions, even in the [i] context if the aspiration is replaced by silence. However, it should be noted that the unaspirated stops, though more distinct, sounded less natural. Similar experiments with [t] and [p] explosion also showed total dominance of the [t] explosion, which might be less surprising regarding the result from synthetic speech where the [t] explosion area has little overlap with the [p] and [k] areas. Furthermore it can be noted that [tɑ] without explosion is clearly heard as [pɑ]. If successively longer parts of the burst is removed the stops are first deprived of their place cues and later for extreme cutting their manner cues.

The result suggests that in synthetic speech one could do with just a single explosion pattern for each phoneme without risking confusions such as in the Haskins' synthesis work, provided that transitions are appropriate or even of the [p] type and little emphasis is put on the aspiration. It also means that if we put a stop explosion before any aspiration plus vowel containing competing transitional cues, the explosion may carry so heavy

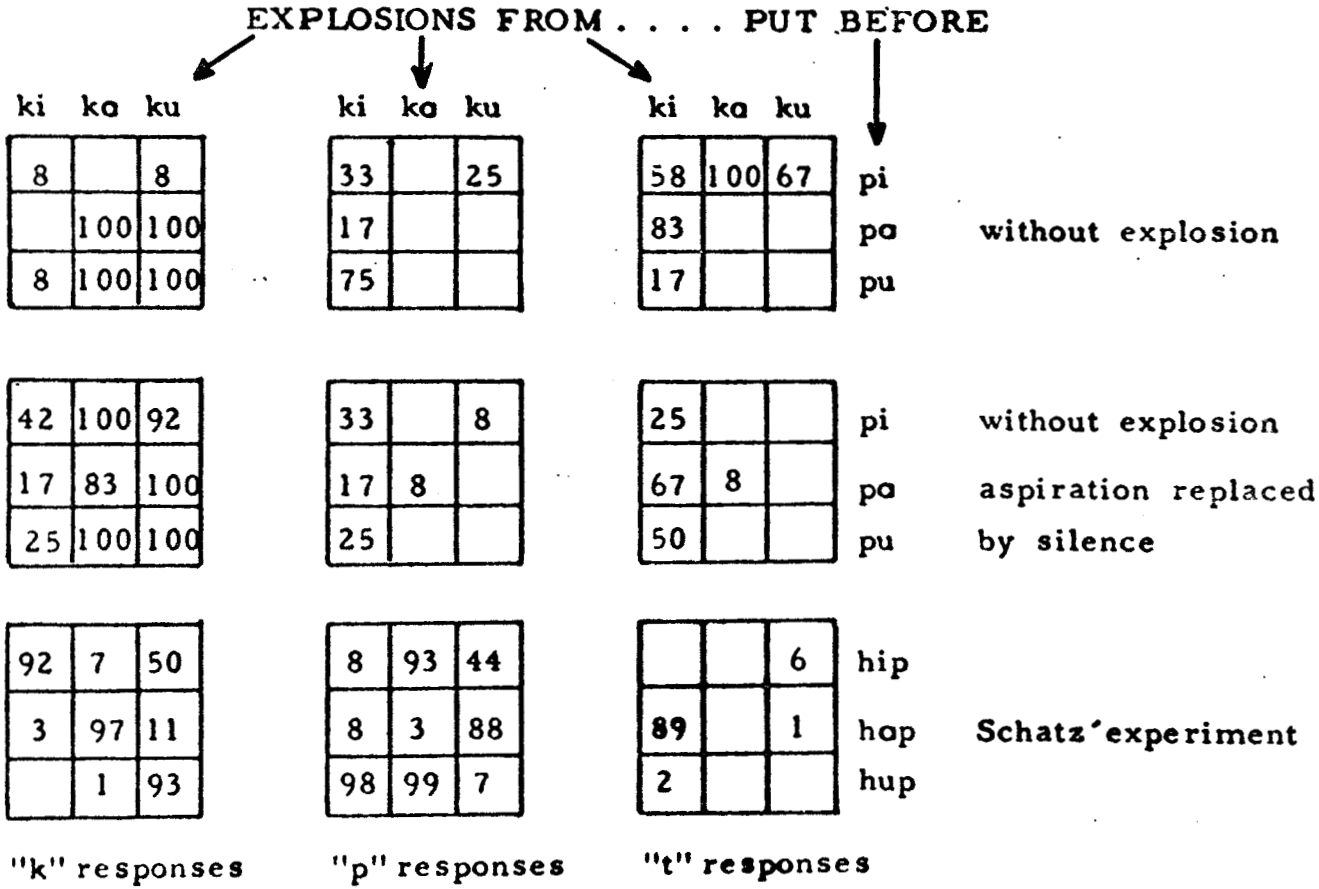
EXPLOSIONS FROM . . . . PUT BEFORE
ki ka ku
ki ka ku
ki ka ku
8 8
100 100
8 100 100
33 25
17
75
58 100 67 pi
83 pa without explosion
17 pu
42 100 92
17 83 100
25 100 100
33 8
17 8
25
25 pi without explosion
67 8 pa aspiration replaced
50 pu by silence
92 7 50
3 97 11
1 93
8 93 44
8 3 88
98 99 7
6 hip
89 1 hop Schatz'experiment
2 hup
"k" responses
"p" responses
"t" responses

Fig. II-A-1.

information about its origin that the transitions will be neglected in an identification situation. If we distort the explosion we will get information about which aspects of it are used in perception.

Selective operations on the [t] explosion

A simple description of stop explosions could be that [t] is a noise above a certain frequency limit, [k] a bandpassed noise, and [p] a noise with a slope in the spectrum. In that case, a filtered [t] explosion but before e.g. a [p] aspiration and vowel, the identity should be [t] if the low pass limit is high enough. Fig. II-A-2 shows this situation when we used a 10 msec [t] explosion from [tɑ] followed by [pɑ] with removed explosion. Variation of the [t] explosion level is used as a parameter. A high passed version is included in the figure.

The result shows the importance of energy above 4 kHz in a [t] explosion. The level has to be of some magnitude to avoid masking effects. It should be noted that a [t] explosion contains energy well below 4 kHz which is of little importance[7].

We simplified the explosion to bandpassed white noise and made a similar experiment which showed that the low pass limit must be above 4 kHz to get [t] responses. The high pass limit was of little importance.

Selective operations on the [k] explosions

There were no [k] responses in the test described above. This indicates that bandpassed noise is a too simple approximation of a [k] explosion.

A repetition of the experiment described above with the [p] explosion exchanged for a [k] explosion gave the following result:

1. Total [k] explosion - 100 % identification of [k].
2. The explosion high passed 1.4 kHz - 100 % identification of [k].
3. The explosion high passed 2.0 kHz - 0 % identification of [k].
4. The explosion low passed 2.0 kHz - 50 % identification of [k].

These findings are not astonishing except for the low passed explosion which shows the importance of the frequency area above 2 kHz. The main formant of the [k] explosion was centered around 1700 Hz.

In a further study the [k] explosion was synthesized and put before a natural [pɑ]. The result indicates that the F1-residue at about 500 Hz in explosion

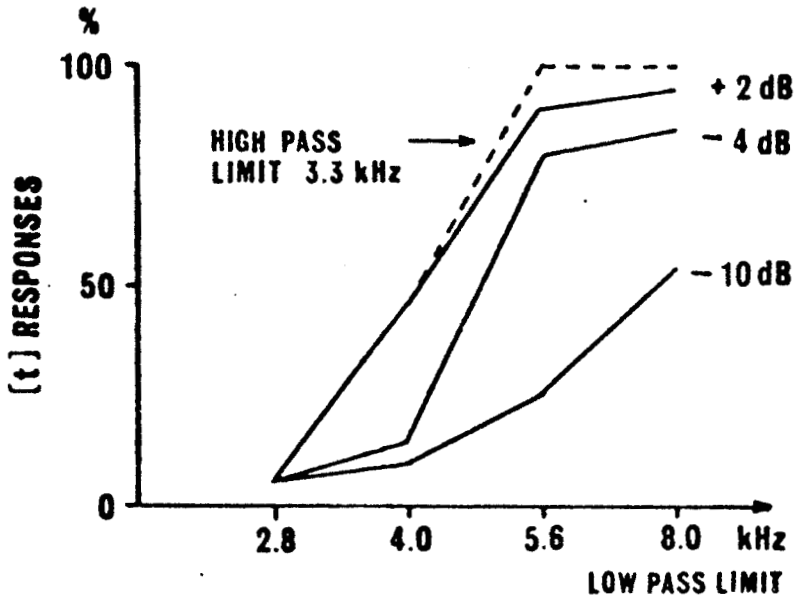

Fig. II-A-2. Result of identification test with filtered [t] explosion before [pɑ].

spectra could be of some importance and hence should not be neglected in speech synthesis.

Manner and place cues in nasals and laterals

Background

Laterals and nasals may be regarded as a homogeneous class with respect to complex resonator articulation and midsagittal oral closure as basic features. The similarity between [n] and [l] is obvious when looking at spectrograms. One difference, however, lies in the energy concentration around 3 kHz of the [l] occlusion, but could it serve as the main cue of an [l+V] sequence? In the case of [m] and [n] it is known from experiments with natural speech that the occlusion may be interchangeable without change of the identity[8]. This leads us to the question how much influence the beginning of the vowel has on the identity.

The [l] [n] distinction

The relatively close similarity between [n] and [l] is apparent in an experiment with natural speech, where the oral occlusion plus the first 0 - 3 glottal pulses (GP) of the following vowel were interchanged in [na] and [la] (Figs. II-A-3 and II-A-4). (The number of responses not "n" or "l" were less than 5 %.) Two cases of filtering were tried: low passing of the complete exchanged part or of just the occlusion. It is clear that the nonfiltered stimuli changed from random to definite identity after one GP. The curves in Fig. II-A-3 move more rapidly towards 100 % than the corresponding curves in Fig. II-A-4, because the vowel adjacent to nasals have to be nasalized.

An interesting difference between the two modes of filtering is seen. If the vowel is unfiltered the "l" responses increase. The possible explanation is that a more abrupt change of intensity in the phoneme boundary is a cue distinguishing [l] from [n]. From an articulatory point of view it might well be so since the lateral passage, especially in context with high vowels, tends to close shortly before or simultaneously with the opening of the midsagittal passage.

The [m] and [n] distinction

In a similar experiment with [m] and [n] there is an almost total shift in identity between 0 and 3 GP in the exchanged part (see Table II-A-1).

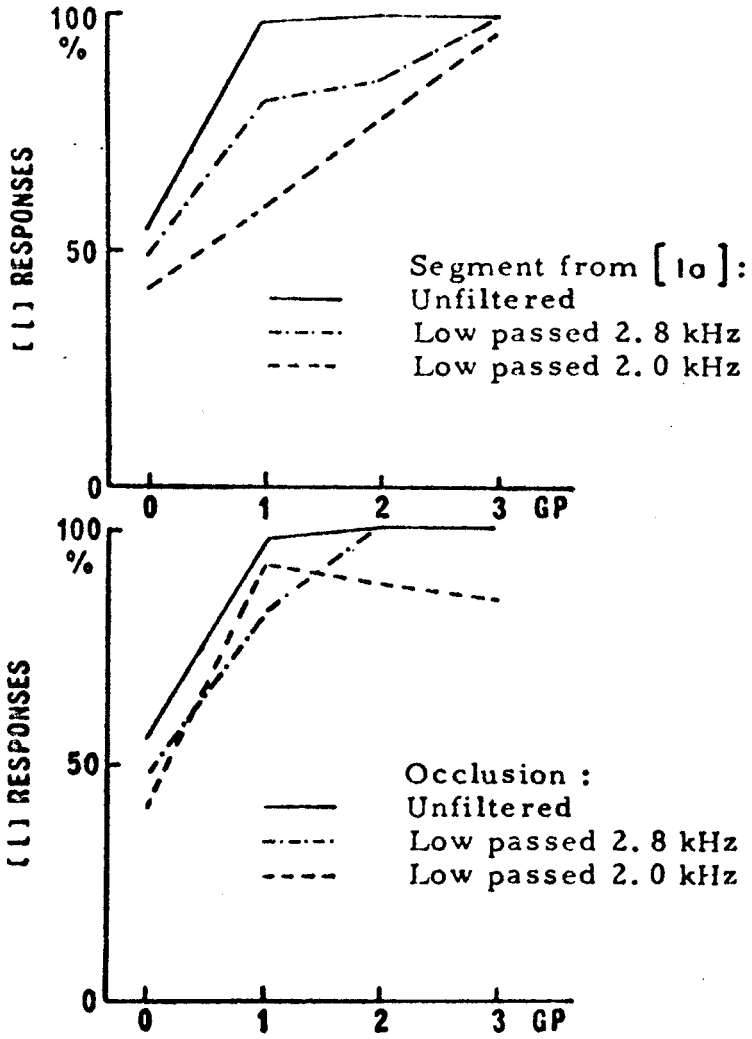

Fig. II-A-3. Result of identification test with occlusion and first 0-3 vowel pulses in [na] exchanged for the same part in [la].

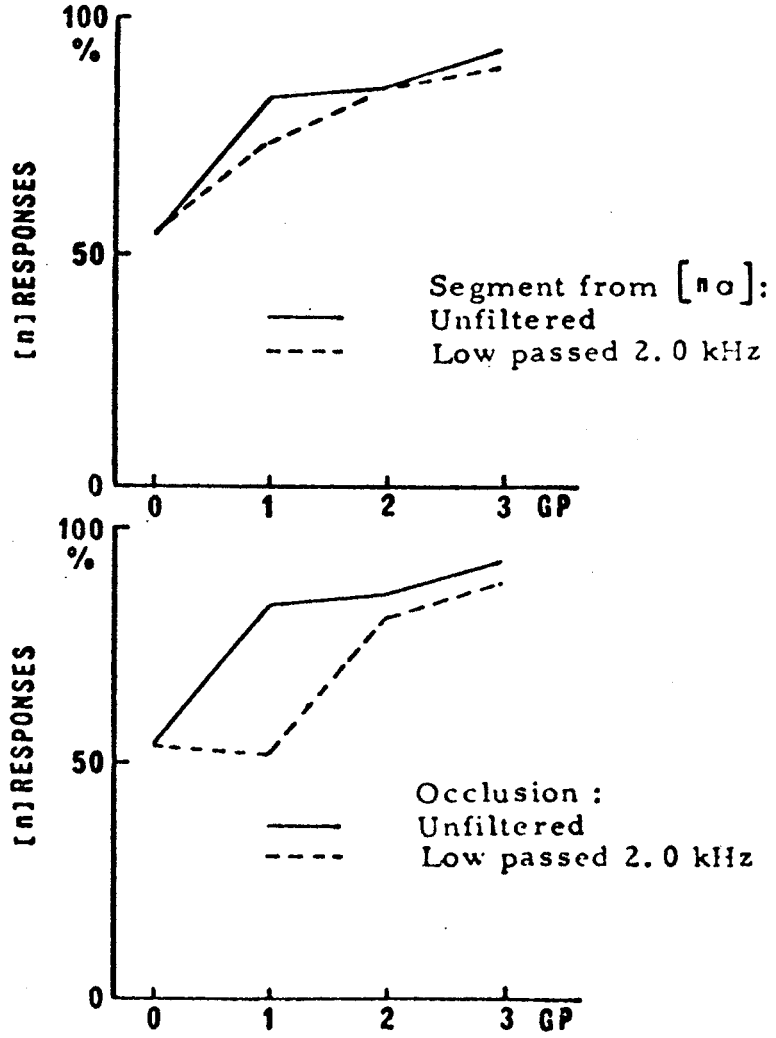

Fig. II-A-4. Result of identification test with occlusion and first 0-3 vowel pulses in [la] exchanged for the same part in [na].

Table II-A-1.

1.	UNDISTORTED	98 % correct
2.	WRONG OCCLUSION	95 % "
3.	WRONG OCCLUSION + 3GP	10 % "

This does not mean that the occlusive segment always lacks a place distinguishing importance as shown in the experiments with synthetic speech below.

Synthesis of nasals

The purpose of the following experiment was to find out if identification and acceptability of nasals were improved if the single formant (N1) nasal branch in the OVE III synthesizer was provided with an additional second formant (N2) in parallel with B1.

Description of the experiment. First the best possible /ma/, /na/, /mi/, and /ni/ stimuli were synthesized. These are referred to as prototypes. Fig. II-A-5 shows /mi/. The same A0- and AN-contours were used in all stimuli. In /mi/ and /ni/ F1 is raised during the nasalized vowel. N2 was chosen to 2.2 kHz in /m-/ and 3 kHz in /n-/. N1 and N2 were added with equal phase. In this way we got a zero between N1 and N2 which is often seen in spectra of nasals. The amplitude relation between N1 and N2 were chosen after a study on natural speech.

The four prototypes were varied with one acoustic cue at a time, all other variables kept constant. For each of these prototypes we made six vowels.

Two listening tests were performed: one identification test (IT) ("What can you hear?") and one preference test in which each prototype was compared with its variants ("Which of these two do you prefer?"), the results of which are shown in Fig. II-A-6.

Description of acoustic cue variations and results. The identification of the prototypes was on an average better than 95 %.

(1) Variation of N2. The following variations were introduced: N2 = 1, 1.4, 2.2, and 3 kHz. (The N2-bandwidth is approximately 300 Hz, and the N2-level is approximately 10 dB below N1.) The results from the IT are shown in Fig. II-A-7. The position of N2 is apparently critical in /i/-context, especially in /ni/, but not in /a/-context.

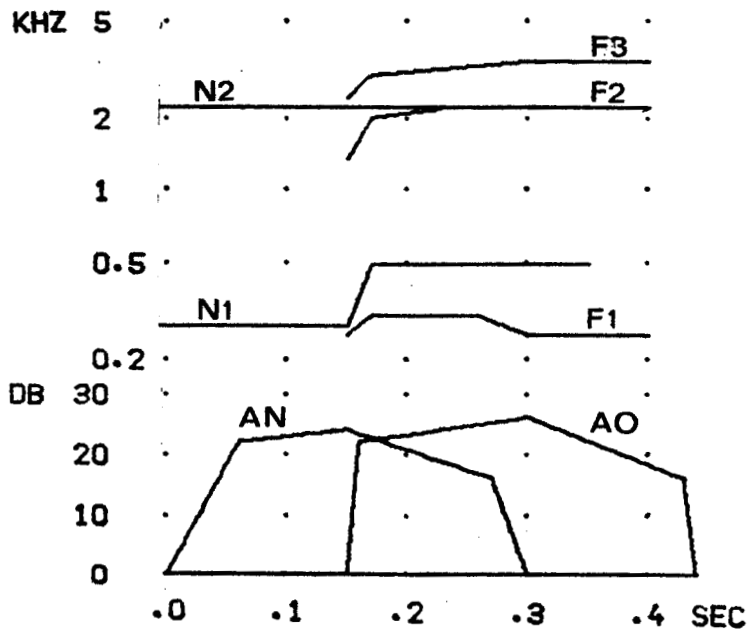

Fig. II-A-5. Prototype MI stimuli.

IDENTIFICATION TEST

	mi	ma	ni	na
N1+N2 (Prototypes)	95	100	90	100
N1-N2	95	95	95	100
N1 only	75	100	45	95
No F1-raise	100		75	
No init. trans. in F2 & F3		80		

Correct identification in %.

PREFERENCE TEST

	mi	ma	ni	na
N1-N2	30	48	35	48
N1 only	5	3	18	0
No F1-raise	53		35	
No init. trans. in F2 & F3		53		

Preference of the 'acoustic variants' to the prototypes in %.

Fig. II-A-6.

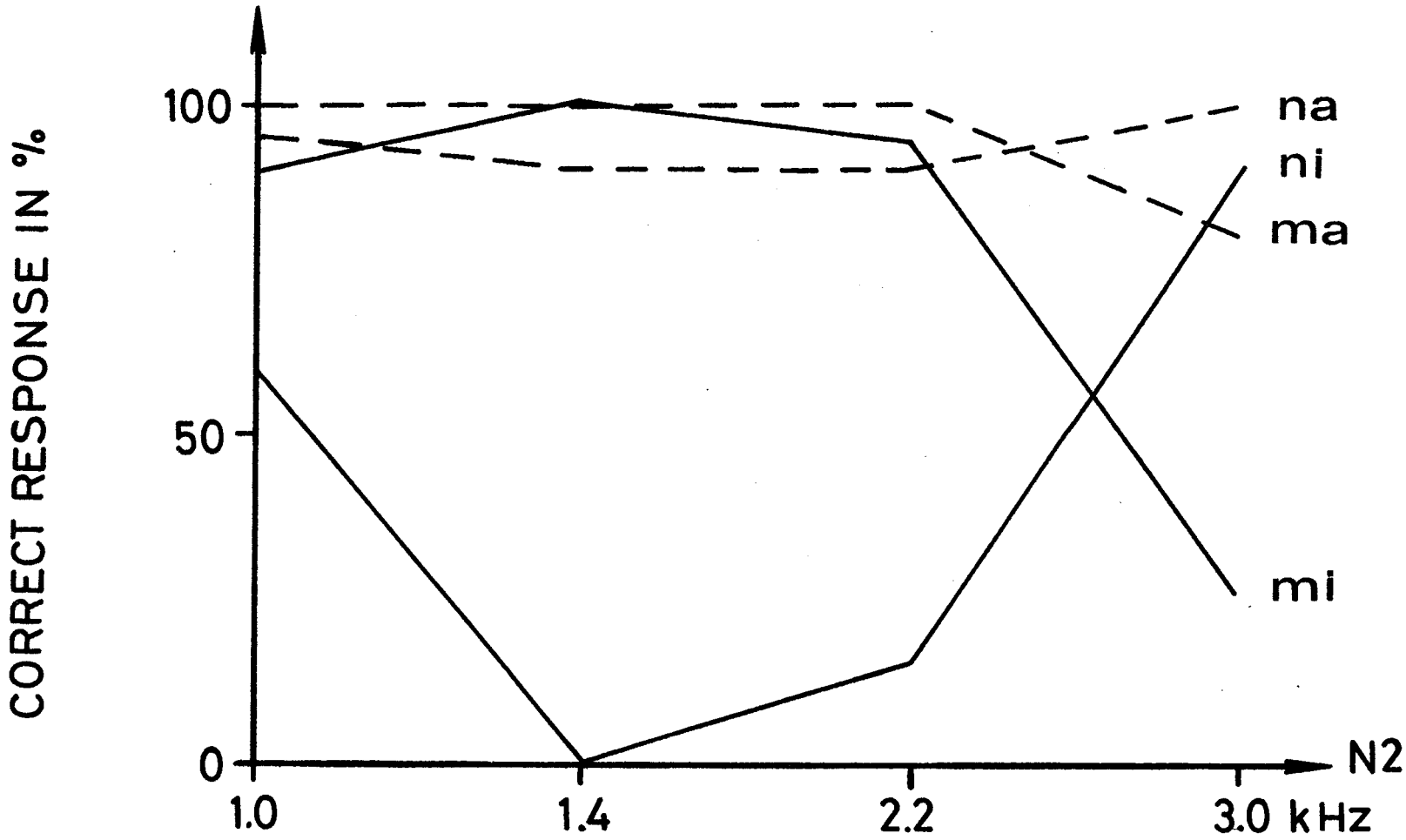

Fig. II-A-7. Correct response in the IT as a function of the N2 frequency.

It is not possible to use the same N2 in /m/ and /n/ in all vowel contexts. This finding does not conform with the results from the previous cutting experiments with natural speech. A pronounced effect is the shift in response from /n/ to /m/ when N2 of the /n/-prototype is shifted to 2.2 or 1.4 kHz. When N2 is shifted further down to 1.0 kHz the mixed /m/ and /n/ responses indicate the low frequency boundary of an /m/-cue frequency region.

(2) N2 added to N1 in opposite phase does not influence the identification, and the preference is influenced in /i/-context only where there is a significant preference for equal phase.

(3) Omission of N2. The identification in /a/-context is not affected but in /i/-context, especially for /ni/ there is a significant decrease in identification. A pronounced effect is the loss of naturalness as evidenced from the high (90%) preference for the two-formant nasals.

(4) Omission of F1-raise in the /i/ vowel. An effect was found in /ni/ only, both identification and preference being decreased.

(5) Omission of the initial transition in F2 and F3 in /na/. The identification is decreased, but the preference is not influenced.

Conclusions of the synthesis experiment

(1) The identity of nasals versus stops and laterals is largely bound to the continuity of nasal formants into the following vowel (general finding, not quantitatively tested).

(2) A two-formant nasal murmur provides better naturalness and identification than a single formant nasal murmur.

(3) The location of the second nasal formant N_2 is an effective place cue in /i/-context.

(4) Formant transitions in the vowel following release are effective place cues.

References

(1) Fant, G.: "Stops in CV-syllables", STL-QPSR 4/1969, pp. 1-25.

(2) Fant, G.: Acoustic Theory of Speech Production ('s-Gravenhage 1960, 2nd edition 1970).

(3) Fujimura, O.: "Analysis of nasal consonants", J.Acoust.Soc.Am. 34 (1962), pp. 1865-1875.

(4) Fujimura, O.: "Spectra of nasalized vowels", MIT-QPR No. 58, (1960), pp. 214-218.

(5) Liberman, A., Delattre, P., and Cooper, F.S.: "The role of selected stimulus-variables in the perception of unvoiced stop consonants", Am. J. of Psych. (1952), pp. 497-516.

(6) Schatz, C.: "The role of context in the perception of stops", Language 30 (1954), nr 1.

(7) Halle, M., Hughes, G., Radley, J-P.: "Acoustic properties of stop consonants", J.Acoust.Soc.Am. 29 (1957), pp. 107.

(8) Malécot, A.: "Acoustic cues for nasal consonants", Language 32 (1956), pp. 274-284.