PHONEME IDENTIFICATION IN ISOLATED STIMULI AND IN CONTEXT

Louis C.W. Pols

1. INTRODUCTION

Because of the present emphasis on invariant cues in speech perception (Perkell et al., 1984) it is worthwhile to draw attention to the influence of physical aspects of (word or sentence) context on phoneme identification. With physical aspects I mean properties like tempo, duration, amplitude, spectral transition, and speaker characteristics, all of which are more or less independent of the syntactic and semantic attributes of a sentence, but which do set a referential framework within which specific properties of a speech segment can be evaluated. Some recent papers on phoneme identification study this context dependence (e.g. Repp et al., 1978; Marcus, 1978; Fitch, 1981; Johnson and Strange, 1982; van Heuven, 1983). In this paper some of these studies will be reviewed and some (preliminary) data of our own will be presented.

2. CONSONANT IDENTIFICATION AND DISCRIMINATION

In a recent paper by Repp (1983) about the perceptual equivalence of specific signal properties in phoneme identification (trading relations), it struck me that all five experiments were done with isolated, manipulated stimuli. The stimuli were of the type 'say-stay', 'say shop-say chop', 'goat-coat', 'slit-split', or 'ga-ka' and the properties traded against each other were respectively, duration of silence vs. F1 onset frequency; duration of silence vs. duration of fricative noise; voice onset time vs. amplitude of aspiration noise; silent closure duration vs. presence or absence of labial release burst; and voice onset time vs. F1 onset frequency. The parameter

values on the contina were chosen such that within-, as well as between-phoneme category discrimination experiments could be done. Repp's conclusion is that his data lend support to the classic dual-process view of speech perception, which represents bottom-up and top-down components. His final suggestion is to stop quibbling about the phonetic or auditory origin of trading relations and to start theoretical and empirical studies about the acquisition of phonetic categories and their internal representation.
However, I wonder if we ever will be able to disentangle the internal representation of phonetic category prototypes if we limit ourselves to phoneme identification and discrimination experiments with isolated stimuli. The so-called prototypes acquired by past experience of speaking and listening are certainly no fixed standards but will depend on talker, speaking style, and context, to mention just a few. In everyday speech perception all of these aspects are taken into account in a natural way, whereas in many laboratory experiments:

- the talker is ambiguous, for instance because of the manipulations with the speech material, or because of the use of synthetic speech;
- the speaking style is unspecified, because the stimuli are too short;
- the context is absent, because of the isolated presentation;
- or the context is artificial, if for instance determined by the range of stimuli in the experiment or by a so-called standard stimulus.

I also consider a set of synthetic CV stimuli along a continuum from /ba/ to /wa/ to be isolated stimuli, whereas for instance Maxwell and Landahl (1983) call this a 'phonetic context'. A remark in the discussion section of that same paper is also relevant to the present discussion. It reads: "... the results may be due in part to the nature of the synthesized stimuli, and therefore may not be representative of actual speech. Not only are the stimuli impoverished with respect to real speech, but they may not reflect the actual changes which occur in speech sounds at different speaking rates" (p. 126).

I certainly agree with this warning about results being influenced by the quality of the stimuli, and I should like to add that these restrictions are not just valid for the stimuli but also for the identification experiment itself. For most laboratory experiments the 'referential framework' towards which a specific utterance should be evaluated is absent or ill-defined. The set of stimuli within one experiment can become the temporary reference, or the listener will perhaps refer to some average, nonspecific, and probably unstable, standard with respect to phoneme duration, formant transition, amplitude relation and the like. This implies that it can be dangerous to extrapolate results of such experiments to natural speech, although valuable by themselves.

3. PHONEME IDENTIFICATION IN CONTEXT

So, in order to be more certain about the referential framework actually used by the listener in a phoneme identification experiment, I suggest to specify the framework as well as possible. One way of doing so is to embed the stimulus word in a neutral carrier phrase of the same talker or the same synthesizer. Whenever temporal (tempo, VOT, closure duration), level (level of burst, amplitude of aspiration), spectral (formant values), or dynamic (spectral onsets) attributes are involved, the context of the carrier phrase can provide a stable framework against which these attributes can be evaluated.

Although I just expressed my doubts about the use of isolated, manipulated stimuli, I must of course admit that many valuable results have come out of such experiments. Moreover, frequently these experiments could hardly have been done in another way. Strong attributes, like the importance of formant transitions or VOT, or strong trading relations will come out anyhow, but there are indications that more subtle relations like the importance of vocalic transitions vs. invariant burst onset spectra, or vowel identification in within- or between-speaker conditions, depend on the presence or absence of context. Before reviewing a few of these recent studies it is only

fair to say that the use of context most certainly will not solve all problems. When a constant neutral carrier phrase is used, a listener could adapt to that phrase to such an extent that it might just as well be omitted. On the other hand, if different phrases are used, the syntactic and semantic content could influence the phoneme identification. This syntactic and semantic content in itself is very basic for speech perception but not the first aim for many phonetic listening experiments. Because of interactions at a local level (at the phoneme and word boundary), (phonetic) context can also have an influence on phoneme identification.
It is not yet clear how much context is needed to specify the reference. Fitch (1981), for instance, suggests that a two-syllable word of the type /da$^{b}_{p}$i/ suffices to specify the speaking rate.
Johnson and Strange (1982) start with a full carrier phrase ("Was it the tVt sound that you heard") in order to study the identification of vowels in tVt syllables spoken in normal and rapid rate. Identification performance for isolated syllables was not as good as it was for syllables presented in the original utterance. Most long-short vowel confusions occurred when long vowels from the rapid sentences were presented in normal rate carrier phrases. If only parts of the carrier phrase were presented, the presence of the stressed word 'sound', immediately following the tVt syllable, appeared to be essential for the accurate identification of intrinsically long vowels in rapidly spoken syllables.

4. SOME OBSERVATIONS OF OUR OWN ON PHONEME IDENTIFICATION IN CONTEXT

In a pilot study we used comparable Dutch sentence material ("Nu krijgt de tVt 'n beurt"), also spoken at a normal and a rapid rate. We got more rate-dependent vowel confusions if not only the vowel part was misplaced but together with that also the closure interval following the test vowel. This is understandable if one realizes that with a higher rate not only the vowel part is shortened but also this silent interval. When this whole vowel-plus-silence segment was substituted in a normal rate carrier phrase we got some long-short

vowel confusions. More specific data will be presented at some later time.

Pols and Schouten (forthcoming) and Pols (1984) used naturally spoken sentences in a plosive identification experiment. Fifty pairs of meaningful Dutch sentences were composed for which the only difference per pair was the use of one 'intervocalic' plosive or another. An example in English would be "the car is open" vs. "the bar is open" with the opposition k/b. The listener knew in advance which sentence out of 50 would come, he only had to indicate which of the two given plosive consonants he heard in that particular condition. The experimental variable was the amount of speech signal deleted (VC vocalic transition and/or plosive burst and/or CV vocalic transition). In order to evaluate the effect of sentence context in this plosive identification experiment, the VCV segments were also presented in isolation, again with various parts deleted. The intriguing questions here are of course

- what is the importance of the vocalic transitions relative to the burst for plosive identification, and
- to what amount is this balance influenced by the presence or absence of a neutral sentence context.

It appears that, especially for Dutch voiced plosives, the transitional information is more effective in the sentence context than in isolation. Or, in other words, experiments with isolated stimuli can give the wrong impression about the relative importance of various components in the signal since the referential framework towards which dynamic information can be compared, is missing in isolated stimuli.

I consider most trading relations, as described by Repp (1983) for isolated stimuli, of the context-dependent type and I therefore expect different results if these experiments are repeated with more natural stimuli embedded in an appropriate context. In fact Repp himself points out that this context or anchoring effect is a non-negligible factor in his experiment. I quote: "... the abnormally high performance level in the Within condition gives rise to suspicion. Indeed the author's observation as a pilot subject suggested

that the consistent presence of the 0-msec standard on every trial may have acted as an anchor .... If so, the trading relation evident in the Within condition may derive from the perception of phonetic contrasts, rather than from a psychoacoustic interaction" (p. 355). He then decides to use different standards in each test block. One can complain about this contrast effect, one can also consider it a natural thing in speech perception. Repp et al. (1978) actually showed strong interaction between speech rate and silent interval duration preceding the fricative noise for the affricate-fricative switch-over point in a carrier phrase context ("Why don't we say shop/chop again").

We evaluated this context effect with a Dutch variant of the 'slit-split' identification experiment by using the 'slijt-splijt' /slɛit-splɛit/ opposition in Dutch. We used both words spoken in isolation as well as in a carrier phrase ("Dit hout slijt/splijt niet"), and at a normal and a fast speech rate. The words, or sentences, were presented for 'slijt/splijt' identification to ten naive subjects while the silence durations were varied, see Fig. 1.

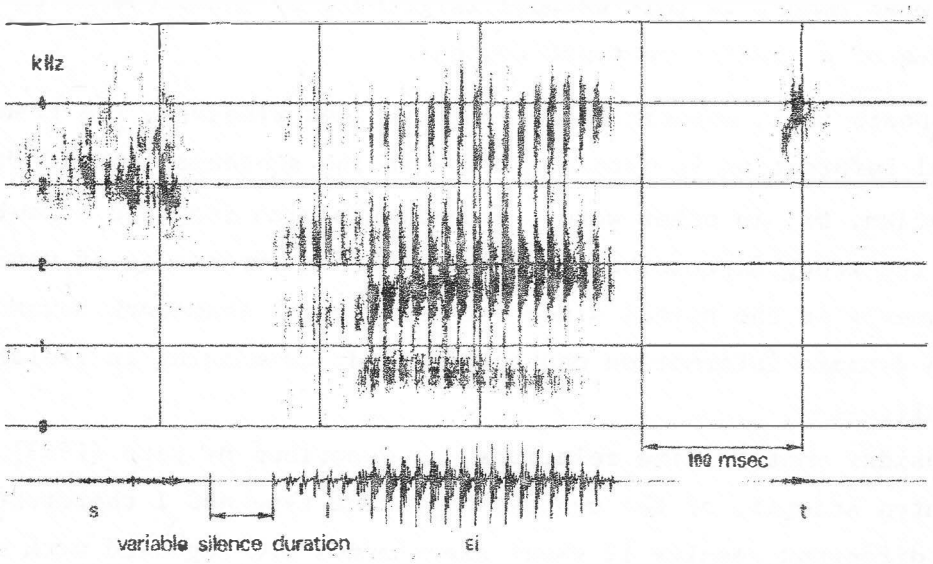

Fig. 1. Waveform and digital spectrogram of one of the stimuli in the slijt/splijt identification experiment. In order to achieve the various stimuli, the indicated silent interval was varied in duration.

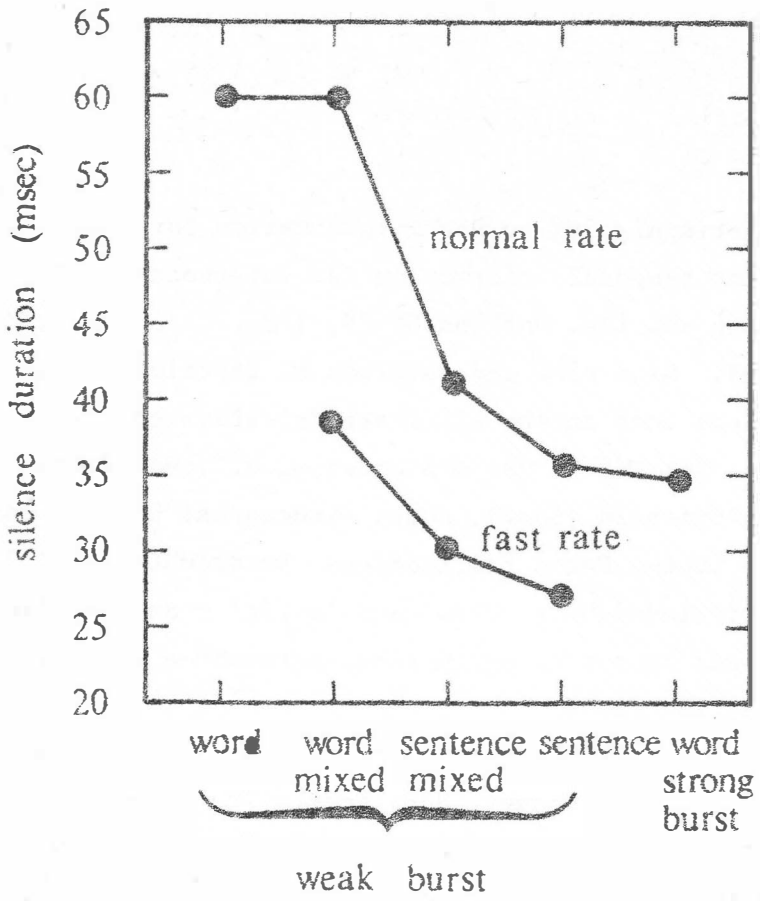

Fig. 2. Fifty-percent cross-over points, in terms of duration of silence interval between /s/ and /l/, from 'slijt' to 'splijt' identification for various conditions. See text for more details.

In the mixed conditions both normal and fast rate stimuli were mixed in one block. The 'weak burst' and 'strong burst' conditions are related to the trading relation as studied by Repp (1983) and have to do with the absence, or part-presence, of the plosive burst of /p/, respectively. Some preliminary results are presented in Fig. 2. For each condition the 50-percent cross-over from 'slijt' to 'splijt' is indicated. For silence durations longer than the value of the cross-over point in that condition the word is heard as 'splijt', and for smaller values as 'slijt'.

Without going into detail with respect to these preliminary results it will be clear that the value of the cross-over point varies a lot, from about 25 to 60 msec, and is most certainly not a fixed standard. This variation is caused by various context effects, like normal/fast rate, weak/strong burst, and isolated presentation vs. word in sentence context. These results once again show that the physical aspects of a context influence phoneme identification in a natural way.

REFERENCES

Fitch, H.L. (1981). Distinguishing temporal information for speaking rate from temporal information for intervocalic stop consonant voicing, Haskins SR-65, 1-32.

Heuven, V.J. van, (1983). Rise time and duration of friction noise as perceptual cues in the affricate-fricative contrast in English, In: M.P.R. van den Broecke, V.J. van Heuven, and W. Zonneveld (Eds.), Sound Structures: Studies for Antonie Cohen, Foris Publications, Dordrecht, 141-157.

Marcus, S.M. (1978). Distinguishing 'slit' and 'split' - an invariant timing cue in speech perception, Perception and Psychophysics, 23, 58-60.

Johnson, T.L. & Strange, W. (1982). Perceptual constancy of vowels in rapid speech, J. Acoust. Soc. Amer., 72, 1761-1770.

Maxwell, E.M. & Landahl, K.L. (1983). The stop-glide contrast and considerations of phonetic context, MIT Speech communication group Working Papers Vol. III, 119-127.

Perkell, J. et al. (Eds.) (1984). Proceedings of Symposium on Invariance and Variability of Speech Processes, Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale NJ.

Pols, L.C.W. (1984). Variation and interaction in speech, to be published in J. Perkell et al. (Eds.), also as report nr. 74 from Institute of Phonetic Sciences, University of Amsterdam, 1983, 23 pp.

Pols, L.C.W. & Schouten, M.E.H. (forthcoming). Plosive identification in ambiguous sentences, presented for publication.

Repp, B.H., Liberman, A.M., Eccardt, T. & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative and affricate manner, J. Exp. Psychol. (HPP), 4, 621-637.

Repp, B.H. (1983). Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization, Speech Communication 2, 341-361.