Speech rate and segmental perception or the role of words in phoneme identification

THE COGNITIVE REPRESENTATION OF SPEECH
T. Myers, J. Laver, J. Anderson (editors)

Speech Rate and Segmental Perception or the Role of Words in Phoneme Identification

S.G. NOOTEBOOM

Institute for Perception Research, Eindhoven

INTRODUCTION

Suprasegmental properties of speech, for example those related to intonation, rhythmical grouping and speech rate, form one class of factors causing acoustic variability of linguistically invariant units. Studies of the effects of suprasegmental properties of speech on the perception of discrete linguistic units may therefore in principle be used to gain insight into the general problem of how such discrete units are extracted from the variable speech waveform. By way of example, one particular class of such studies, those concerned with the effect of speech rate on phoneme perception, will be examined in some detail. This class of studies is typical of a much wider range of speech perception studies because of the unquestioned assumption that phonemes are immediate and natural response categories in speech perception tasks, and that studying the perception of phonemes is essential to our understanding of speech perception in general. In fact, it can be said that most perception researchers are eagerly trying to find out how phonemes are extracted from the speech waveform. My present view of this is that they are most probably on the wrong track. I will argue in this paper that linguistic processing of speech, and particularly word recognition, is not mediated by phonemes, but that rather phoneme perception as studied in phoneme identification tasks is mediated by word recognition. If this is correct, it leads to a reinterpretation of current data obtained in phoneme identification experiments. More importantly, it suggests that it may be high time to replace phonemes by words as the main focus of attention in speech perception research.

In what follows I shall first indicate the class of studies of speech rate and segmental perception used by way of example in this paper. I shall then attempt to account for these data within a phoneme-based view of speech perception, concluding that any such attempt is unsatisfactory. Next I will argue that a satisfactory account of these and other data from phoneme identification experiments can be based on the view that phoneme identification is mediated by word recognition or word identification. Finally I make a plea for being careful in interpreting data from identification experiments and for increasing our efforts in the area of word recognition studies.

SPEECH RATE AND SEGMENTAL PERCEPTION

The experiments to be considered here are conceptually very simple. Generally, there are two experimental variables. One is the duration of a particular acoustic segment, for example vowel segment duration, chosen so that at one extreme of the scale of durations a particular categorical response, say a phonemically short vowel,

is favoured, whereas at the other extreme the opposing categorical response, a phonemically long vowel, attracts most perceptual judgements. The segment duration at which 50% of responses in either response class are obtained in a forced choice identification task is defined as the perceptual boundary or phoneme boundary. The other experimental variable is speech rate of the utterance or part of the utterance the test segment belongs to. The dependent variable of interest is the shift in perceptual boundary corresponding to a shift in speech rate. Figure 1 presents an example of some data obtained in our Institute. The test vowel segment was embedded in the last syllable of a meaningful carrier phrase. Both the duration of the test segment and the rate of the vocoderized carrier phrase were experimentally varied. The mean phoneme boundary for 10 subjects is plotted as a function of speech rate. It may be seen that the perceptual boundary, defined as the estimated duration giving 50% of either response type, depends on the speech rate. A faster speech rate gives a shorter duration for the perceptual boundary.

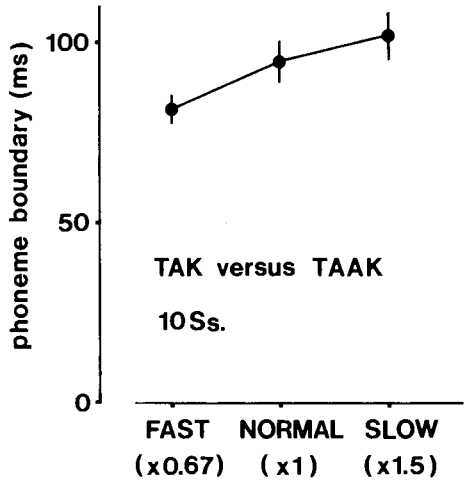

FIGURE 1 *Phoneme boundaries between dutch /a/ and /a:/ on the dimension of acoustic vowel duration. Test segments were embedded in a monosyllable word /tak/ or /taak/ placed at the end of a meaningful carrier phrase. The duration of the carrier phrase up to and including the consonant preceding the test segment was made 0.67, 1, or 1.5 times normal by changing the readout speed of the synthesis part of a computer controlled channel vocoder. Standard deviations over 10 subjects are indicated.*

Similarly, Picket and Decker (1960) showed that the perceptual boundary value between a single and a geminate stop cued by the stop closure duration falls at systematically decreasing durations with increasing speech rate. A qualitatively similar relation was found for the phoneme boundary between voiced and voiceless stops on the dimension of voice onset time (Summerfield and Haggard, 1972), the phoneme boundaries between long and short consonant and vowel phonemes (Ainsworth, 1972, 1974; Fujisaki, Nakamura and Imoto, 1975; Nooteboom, 1977, 1979), the boundary between fricative and affricative as cued by noise duration (Repp, Liberman, Eccardt, and Pesetsky, 1978), the boundary between intervocalic voiced and voiceless stop cues by silent interval duration (Port, 1977, 1978). An exception is reported by Marcus (1978) who found no effect of speech rate within

one-word utterances *slit* and *split* on the perception of a /p/ as cued by a silent gap introduced between /s/ and /l/, possibly because the perception of a stop consonant is not cued by the perceived duration of a silent gap but rather by the presence or absence of a silent gap. In all cases where the perceptual distinction depends on the perceived duration of a particular acoustic segment, the perceptual boundary varies with speech rate. How can we account for this rather general phenomenon?

SOME ATTEMPTS TO ACCOUNT FOR RATE EFFECTS ON PHONEME IDENTIFICATION

A simple model predicting the relation between perceptual boundary and speech rate is this: suppose that the transformation of physical time into perceptual time is controlled by a clock, and that the pace of this clock is set by the speech rate of the attended utterance (cf. Summerfield and Haggard, 1972). In its simplest form this model would predict that if we increase speech rate by a factor of two, the perceptual boundary on a durational boundary would be halved. This prediction is at variance with all relevant data, except those obtained by Fujisaki et al. (1975) who assessed phoneme boundaries between Japanese long and short phonemes, both vowels and consonants, at four different rates of speech. Their data show perfect adjustment of the phoneme boundaries to speech rate. However, in all other cases the shifts in perceptual boundary are smaller than predicted by a simple external clock whose pace is completely adjusted to speech rate.

The employment of a rate-controlled clock would, of course, be unrealistic from the standpoint of an efficient use of the cues available in the speech waveform. It is known that not all segment types in speech are equally affected by speech rate in the production of speech (Kozhevnikov and Chistovich, 1965; Lehiste, 1970). At least equally important is the observation that often the size of rate-induced shifts in perceptual boundaries cannot easily be predicted from the size of changes in speech rate in the surrounding speech material. Although there is a good correspondence between the production and perception data of Picket and Decker (1964), Summerfield (1975) found rate-induced shifts in the produced voice-onset time of the order of tens of milliseconds, whereas the corresponding shifts in perceptual boundaries were an order of magnitude smaller. Similar discrepancies were observed for systematic effects of speech rate on spoken vowel duration and the corresponding shifts in perceptual boundaries between short and long vowel phonemes in Dutch (Nooteboom and Doodeman, 1979). Apparently, the relation between speech rate and durational perceptual boundaries is not as straight-forward as suggested by the clock model.

Öhman (1975) proposed a view of speech perception according to which a listener decodes the speech signal by projecting it into his internal model of the vocal apparatus of the speaker. Thus the perceptual model would partly be a production model associated with a set of physical states definable in terms of physical concepts. Speech perception, just as the perception of moving cars, or walking people, would be the perception of physical system-state histories:in perceiving speech we hear another person's vocal organs move. A similar view of speech perception, expressed in somewhat less general terms, seems to be taken by Repp, Liberman, Eccardt, and Pesetsky (1978), and by Summerfield (1979), in relation to, among other things, data on speech rate and segmental perception. Repp et al. provide some data showing that the "listeners integrate a numerous, diverse, and temporally distributed set of acoustic cues into a unitary phonetic percept. These several cues have in common only that they are products of a unitary articulatory act. In effect, then, it is the articulatory act that is perceived". This view of speech perception is

basically at variance with the earlier attempts to account for effects of speech rate on segmental perception. This is so, because within this view one cannot distinguish between cues to segmental perception and cues to speech rate. As Summerfield (1979) puts it: "the acoustical substrate for the direct perception of rate and the acoustical elements whose interpretation rate must mediate not only co-occur, they are one and the same". It follows, then, that rate-induced shifts in perceptual boundaries are not caused by differences in perceived speech rate, but rather by rate-induced changes in other acoustic cues which follow from the same articulatory act as the one under investigation. Therefore the time-window over which speech rate may seem to be effective would reflect the time window over which "unitary articulatory acts" have acoustic results.[1]

A test of the hypothesis that speech rate affects segment identification only via its effect on durational cues immediately resulting from the articulation of that segment is provided by Summerfield (1979). Summerfield showed that the perceptual boundary on the dimension of voice-onset time between /b/ and /p/ in the utterance *why are you /biz/?* versus *why are you /piz/*, is affected by the duration of the syllable *you* and the duration of the vowel */i/*, but not by the durations of *why* and *are*: the effect of rate appeared to be entirely due to those segment durations which are regularly affected by the production of a voiced versus a voiceless stop. This may be taken to support the perception-of-articulation-model in the strict form that speech is perceived as sequences of articulatory acts each of which is cued by all acoustic attributes regularly associated with its occurrence in the production of speech. Speech rate can only affect segmental perception via its effect on these acoustic attributes, but does not in itself control the processing of these attributes.

So far the perception-of-articulation-model seems to be quite satisfactory. I will now mention two sets of data which, although not immediately concerned with speech rate, refute the model. The first set of data has been described by Nooteboom and Doodeman (1979), and concerns the perceptual boundary between a Dutch short vowel /a/ and a long vowel /a:/ on the dimension of vowel duration. The test segments were part of a monosyllabic real word, being either *tak* (Engl. branch) or *taak* (Engl. task), which was embedded in a meaningful carrier phrase. There were two experimental variables, the duration of the test segment and the duration of a silent gap introduced immediately after the monosyllable containing the test segment. The perceptual boundary value increased with increasing silent gap duration. This poses a problem to the perception-of-articulation-model in its strict form, because the silent gap cannot reasonably be intrepreted as an acoustic result of the articulation of a short or a long vowel. The association between vowel duration and silent gap duration is not that both result from the same unitary articulatory act required by the to-be-identified phoneme, but is of a different order. Vowels in prepausal syllables are regularly lengthened, and therefore the occurrence of a perceived speech pause increases the expected duration in the prepausal syllable, thus causing a shift in the perceptual boundary between short and long vowels.

The second set of data has been reported by Ganong (1978) who showed a clear and consistent bias towards categorization that made words as against categorization that made nonwords in a phoneme identification experiment. Thus in the pair *gift* versus *kift* there was a bias towards *gift* and in the pair *giss* versus *kiss* there was a bias towards *kiss*. Such lexical effects on phoneme identification, even in a binary forced choice task that in many respects resembles a simple discrimination task, cannot easily be explained by assuming that speech perception is the perception of articulatory acts. More likely, speech perception is the perception of words, or in the case of nonsense items, word-like units.

PHONEME IDENTIFICATION VIA WORD RECOGNITION

The tentative explanations given so far of current data on segmental perception and speech rate have one property in common. They are in line with the assumption that phonemes are natural and immediate response categories in a speech perception task. This is a corollary of a basic idea underlying much contemporary work in the domain of speech perception, namely that linguistic processing of speech, and particularly recognition of spoken words, is mediated by phonemes. I propose to reject that view and instead to start from the not altogether original, but perhaps to some speech researchers heretical, assumption that word recognition is not mediated by phonemes but, on the contrary, phoneme identification is mediated by word recognition. This proposal has some nice analogies with the relation between words and phonemes in phonological analysis: phonemes are secondary units of analysis, words are primary units of analysis. Phonemes are found by comparing minimally distinct words, words are not found by combining phonemes. The assumption that in a speech perception task, even in a phoneme identification task, words become available as responses before phonemes do is supported by reaction time experiments showing that listeners react faster to meaningful units than to phonemes or meaningless syllables (McNeill and Lindig, 1973; Foss and Swinney, 1973; Rubin, Turvey and Van Gelder, 1976) and by experiments showing that reaction times to phonemes in spoken sentences are sensitive to transitional probabilities of the words in which these phonemes are contained (Morton and Long, 1976). Further arguments against phonemes as discrete units necessarily mediating between the acoustic signal and word recognition may be derived from the course of language acquisition (Morton and Smith, 1974), from the less complex relation between the acoustic signal and words on the one hand than between the acoustic signal and phonemes on the other, and from the earlier mentioned experiment of Ganong (1978) showing lexical effects in a phoneme identification task. One way of interpreting the dependence of reaction times to phonemes on properties of the words these phonemes belong to is to assume that a phoneme comes available as response only after the word has been recognized. Thus phoneme identification would depend on word recognition.

An immediate consequence of adopting the idea that phoneme identification depends on word recognition is that one must have some ideas about word recognition in order to interpret data of phoneme identification experiments. I assume, then, that word recognition is mediated by a whole array of independent, parallel word recognition elements which actively respond to features from different sources (cf. Morton, 1969; Marslen-Wilson and Welsh, 1978). For the present purpose it is most relevant to consider the nature of that part of the internal specification of these word recognition elements that makes them respond to the acoustic input. I suggest that this internal specification is not in terms of phonemes, but rather in terms of areas in a recognition space defined by the set of auditory features which each individual language user has acquired in order to distinguish between the lexical units of his language. As language users in the course of language acquisition are forced to learn that many different acoustic forms have to be recognized as a single lexical unit, each area in auditory recognition space may cover a whole range of auditory feature combinations. Each point in such an area corresponds to a particular combination of auditory feature values and may have a particular strength of its association with the corresponding word response: some auditory feature combinations lead more easily to a word response than others.

How does this relate to phoneme identification experiments? In such experiments phonemes are embedded either in real words or in nonsense items. In the case of real words, a subject has recourse to existing word recognition elements and can respond with one of the alternative phonemes each time he identifies the word containing that phoneme. Note that often no phoneme response is required, but a

phoneme perception is inferred by the experimenter from a word response. In the case of nonsense items, the subject has to set up ad-hoc recognition elements for these nonsense items on the basis of the instruction and the initial presentations. These ad-hoc recognition elements can then be used in the same way as the regular ones. The ability to create new recognition elements is essential to language acquisition and therefore must naturally belong to a language user's competence.

WORD RECOGNITION, SPEECH RATE, AND PERCEPTUAL BOUNDARIES

Let me now attempt to relate the present view of phoneme identification to the earlier discussed data on speech rate and segmental perception. I assume that in the internal specification of a word or a word-like unit for each auditory segment that can have a perceptual duration the whole range of potential durations is specified. Each sequence of segment durations (durational pattern) has a particular strength of its association with the word response (response strength). Those durational patterns that are most to be expected in normal speech for the word concerned have the greatest response strength, less likely durational patterns have a smaller response strength. For example, a durational pattern that would be normal within a given speech rate would have a great response strength, just as great as a durational pattern that would be normal in another speech rate. But a durational pattern that belongs partly to one and partly to another speech rate would have a relatively small response strength. In this way detailed tacit knowledge on systematic covariations of segment durations in speech is contained in the distribution of response strength over auditory recognition space for each word recognition element. If two auditory word forms corresponding to two different word responses differ only in the duration of a single auditory segment, there is, when the set of values of all other segment durations is fixed, one duration of that segment for which the association with both word responses is equally strong. This duration corresponds to the perceptual boundary measures in a phoneme identification experiment. Its value on the dimension of auditory segment duration will be the mean of the two values that would correspond to the highest response strengths for the two words in the given durational pattern. Of course, when this durational pattern changes, for example with a change in speech rate, the two optimal values will also change, probably both in the same direction, and therewith the point of equal strength will also shift. In this way shifts in perceptual boundaries can be explained as immediate and passive effects of the distribution of response strengths for different words over the entire set of possible combinations of auditory feature values. No active normalization processes, either operating on subjective durations or operating on internal criteria, are called for. This cheapness in mental calculations is bought at the expense of a very uneconomical storage of word recognition elements. But precisely because of the lack of necessary mental calculations this uneconomical storage has the great advantage that matching between acoustic input and word recognition elements does not take more time than strictly necessary. This means that the effect of memory noise on stimulus presentation is reduced to a bare minimum.

Because the distributions of response strengths must, in the acquisition of language, have arisen from systematic covariations of auditory feature values in the past experience of the individual language user, the present model accounts for all of the data on speech rate and segmental perception that are accounted for by the earlier discussed perception-of-articulation-model, without making the assumption that listeners construct an internalized model of the vocal organs of each speaker they are listening to. In addition the model can naturally handle effects of other sources of information, be they prosodic, lexical, syntactic, or semantic, on phoneme identification, because in speech perception different sources of information come together at the level of word recognition (cf. Cole and Jakimik, 1978; Marslen-Wilson and Welsh , 1978).

If both the model and the data are taken seriously, they together confirm that recognition elements contain detailed information on systematic covariations of auditory feature values in normal speech. Interestingly, this is not only true of regular words but also of nonsense items which have never been heard by the subjects before the experimental session. This suggests that the pattern of response strengths associated with potential configurations of auditory feature values for a particular word recognition element, does not necessarily arise from extensive auditory experience with the word concerned, but may be generated by the subject on the basis of his experience with other words. Thus, whereas the identification or recognition of a particular word or word-like unit may be completely passive, the creation of a new recognition element has to be an active process generalizing from the information patterns of existing recognition elements.

FROM SHIFTING PHONEME BOUNDARIES TO WORD RECOGNITION

There are three major points that emerge from the present reinterpretation of data on speech rate and segmental perception. The first point is this: speech rate as such does not control the processing of acoustic cues to segmental perception, i.e. listeners do not derive from the input speech a measure of the current speech rate which then in turn affects the processing of further acoustic material. The seeming effect of speech rate on the contribution of durational cues to speech perception apparently stems from the effect of speech rate on the production of other durational cues which together with the one under investigation determine what is being perceived. The implication of this finding is that no process of perceptual normalization on speech rate is called for to explain the shifts in perceptual boundaries brought about by changes in speech rate.

The second point is a criticism of the overemployment of forced choice identification as an experimental task in studies of speech perception. I see at least three reasons why one should be hesitant in generalizing from phoneme identification to the normal perception of speech:

1) Due to the lack of useful information in the neighbourhood of the perceptual boundary, the subject may be forced to employ sources of information which in the normal perception of speech never, or hardly ever, play a role.
2) For the same reason responses to stimuli in the neighbourhood of the perceptual boundary are delayed in time, often several hundreds of ms. This makes perceptual boundaries sensitive to information following the test segment in a way which is probably not representative for normal speech perception.
3) Due to the forced-choice character of the task, the limited number of alternatives and the employment of overlearned, highly stable internal criteria, phoneme identification is extremely accurate, often more accurate than one would expect from discrimination measurements. Phoneme identification seems to be an excellent way of measuring just noticeable differences (cf. Schouten, 1978) but these may have little bearing on normal speech perception.

In view of these caveats, it is fair to say that phoneme identification may be a valuable and precise analytical tool for probing a subject's tacit knowledge of, among other things, the sound structure of his language but it is a much less valid tool for studying how this tacit knowledge is applied in the normal perception of speech. This conclusion, of course, reaches much farther than the limited class of speech perception studies used as an example in this paper. Phoneme identification is the most popular experimental task among students of speech perception (not among their subjects).

Thirdly, I would like to make a plea for giving more attention to word recognition in studying suprasegmental effects on segmental representations. If indeed word recognition is as central to speech perception as is argued by Cole and Jakimik

(1978), Marslen-Wilson and Welsh (1978) and Marslen-Wilson (1979b) and as I have assumed in this paper, the effects suprasegmental structures have on recognition may be more interesting than the effects they have on phoneme identification. Although some work has been done relating to this topic (cf. Kozhevnikov and Chistovich, 1965; Blesser, 1969; Cutler, 1976; Brokx, 1979), it is fair to say that only little is known about the features extracted from the speech waveform and contributing to the recognition of words, prosody, syntax, and meaning, and how these features together determine what is perceived. I am convinced that further theorizing and experimental work in this area will benefit if the word instead of the phoneme is made the primary focus of attention.

Note

1. Data from Repp et al. (1978) can be understood in this way. Perceptual boundaries were measured between *shop* and *chop* in *why don't you say shop/chop again*? in a two-dimensional plane, defined by noise duration and silence duration, for both a slow and fast speech rate. Within the same speech rate boundary values for noise and silence duration were positively related, but, for equivalent values durations, more silence was needed in the fast utterance frame than in the slow frame to convert the fricative into an affricative. This may seem paradoxical when one tries to explain it from an effect of perceived speech rate on the cue value of silence duration. Assume, however, that the alleged effect of speech rate is entirely due to the difference in duration of the vowel in *say*, immediately preceding the silence, and that in production there is a regular negative correlation due to compensation between this vowel duration and the following silence duration of the type found by Kozhevnikov and Chistovich (1965) and De Rooij (1979). In that case the shortened vowel in *say* will create an expected longer silence duration for the listener and vice versa. This would explain the data in a natural way, and remove the paradox. Unfortunately, Repp et al. do not provide production data. The point I want to make, however, is that we have to consider the possibility that all so-called rate effects on segment identification can be explained from a subject's tacit knowledge about production regularities immediately related to the segments concerned, without perceived speech rate having any part in it.

Acknowledgement Some of the ideas expressed in this paper were inspired by discussions I have had with my colleage Steve Marcus on the human recognition of spoken words.