CONTEXTUAL VARIABILITY VERSUS PERCEPTUAL CONSTANCY IN SPEECH

GERARD F.J. VAN MEURS

1. INTRODUCTION

One of the most striking aspects of the perception of speech is the relatively great amount of perceptual constancy in the light of the enormous amount of acoustical variability in the speech signal. Listeners are able to divide speech in discrete segments, roughly corresponding to phonemes (Daniloff and Hammarberg, 1973). This perceptual reality (1) seems to suggest an acoustical reality as well. The history of speech perception started with the search for stable acoustic correlates of perceived phonetic contrasts. As far as this goal is concerned this search was rather unsuccessful. It is hard to find invariant cues for each phoneme, i.e. cues that accompany a phoneme in all accoustic-phonetic environments. Acoustically different signals can be perceived as one and the same phoneme, as demonstrated by Liberman et al. (1967) with their well known /di/ - /du/ example, in which the same /d/-percept is cued by very different transitions of the second formant. At the same time it is possible for one acoustical signal to be perceived as different phonemes: Liberman, Delattre and Cooper (1952) showed that the same synthetically produced burst of noise could be perceived as a /p/ or a /k/, depending on the following vowel. These facts have led to a theory of speech perception as an active process in which analysis of the continuous acoustical signal yields discrete phonetic percepts; in this view the cues to phonetic contrasts are always evaluated in the context in which they occurred. Some investigators however opt for a theory of acoustical invariance (see for instance Stevens and Blumstein, 1981). Without denying the possible role of context dependent cues they claim that the acoustical invariance corresponding to a given phonetic category is provided by the integration of several acoustic properties. Stevens and Blumstein extensively illustrate their point of view with the observation that the gross spectral properties of the initial 10-20 msec spectrum of stops provide more or less invariant information concerning the place of articulation.

The majority of studies on all sorts of aspects of speech perception have been concerned with isolated, natural or synthetic, utterances. The underlying assumption was that the results of these studies could be generalised straightforward to the perception of natural speech. Only since the last decade an increasing number of workers in the field of speech perception seem to acknowledge that this assumption is at least questionable (Koopmans-van Beinum, 1980; Pisoni, 1983; Pols, 1983; 1984). Although those studies with isolated stimuli have greatly contributed (and continue to do so) to our knowledge of the different cues to different phonemes, it should be noted that identifying isolated speech stimuli is a quite unnatural task that has very little in common with understanding natural speech. The most important deficiency in isolated speech is the lack of context: natural speech is always embedded in a context. The linguistic aspects of the context put heavy constraints on the possible identifications of phonemes at the syntactic, lexical/semantic, and pragmatic level. But even without these constraints the cues that accompany a given phoneme vary with the acoustic-phonetic context. The acoustic-phonetic aspects of the context (for instance phonemic environment, speaking-rate, duration, amplitude, spectral transitions, and characteristics of the speaker) are important in creating a referential framework for interpreting cues. Pols states this point as follows: "There is always interaction between the variation in speech sounds and the context which sets the referential framework. If there is no context then the listener's reference is ill-defined." (Pols, 1983, p.2).

In this paper an overview will be given of some studies concerning aspects of this acoustic-phonetic context. This overview is the start of a ZWO project concerning the effect of local context on the correct identification of acoustic-phonetic information in speech understanding. In this project the hypothesis will be tested whether isolated presentation of a segment of a few phonemes will always lead to a relatively low identification compared to the performance with the same segments in a broader local context, which only contributes acoustic-phonetic information.

2. ASPECTS OF THE ACOUSTIC-PHONETIC CONTEXT

Although it is well known that the syntactic, lexical/semantic and pragmatic aspects of the context in which natural speech is uttered can have great influences on the identification of phonemes and words (2), the attention in this paper is focused to some acoustic-phonetic aspects of the context: effects of the phonemic context (2.1), effects of speaker variation (2.2), effects of speaking rate (2.3), and more global aspects of the acoustic-phonetic context (2.4).

2.1 THE PHONEMIC CONTEXT

The acoustic realization of a given phoneme is to a very large extent dependent on the phonemic environment in which it is uttered. Rather then being produced as isolated sounds, phonemes are produced as overlapping and interacting movements of the articulators (coarticulation). In order to be able to decode the message from the acoustical signal, it is necessary for the listener to use the local phonemic context in which the message was produced. Liberman et al. (1967) argue that the speech code in which most phonemes are contextually coded is a necessary condition for achieving normal speech rate: "It is this kind of parallel processing that makes it possible to get high speed performance with low speed machinery." (p. 446/47).

Liberman et al. (1967) make a distinction between 'encoded' and 'unencoded' phonemes. The cues to the first set of phonemes are fully interwoven in the surrounding context; phonemes of the latter sort are more or less independent of the context (fricatives and the longer steady state vowels). Cole and Scott (1974) make a similar distinction; in their view the listener uses at least three classes of cues simultaneously to identify the speech sounds: invariant cues, context-dependent cues, and waveform envelop cues. Reviewing the research on consonant identification they conclude that it is possible to identify some phonemes, and to discriminate some groups of phonemes, on the basis of an invariant configuration of acoustical cues; particular members of these latter groups can only be identified with context-dependent cues. Liberman et al. (1967) point at some interesting differences between the two sets of phonemes:

1. The superior discrimination of physical differences near phonetic bounderies ('categorical perception') seems to be restricted to the set of 'encoded' phonemes.
2. The superior identification of speech stimuli presented to the right ear ('right ear advantage') seems to hold only for 'encoded' phonemes.

An important class of context-dependent cues is formed by the transitions from (VC) and to (CV) a vowel nucleus. Transitions are produced by changing the realized shape of the vocal cavity for one phoneme to the realization for the next. The transition from one phoneme to another is determined by the 'starting' phoneme as well as the 'destination' phoneme. In this way it is possible for transitions to convey parallel information about adjacent phonemes (3).

Steady state vowels produced in isolation can be regarded as laboratory artefacts that are never encountered in natural speech; it might therefore be hypothesized that vowels embedded in a consonantal context may be better identified than vowels produced in isolation. This hypothesis was confirmed in an experiment by Strange et al. (1976) on vowel identification. The difference in the identification scores between isolated and embedded vowels could not be attributed to deviant formant structures of the vowels, since these were found not to differ systematically for the two sorts of vowels. (See however Koopmans-van Beinum (1980); in her data the formant frequencies of the vowels in CVC words were slightly 'reduced' compared to those of vowels produced in isolation). According to Strange et al. "... no single, temporal cross section of a syllable conveys as much vowel information to a perceiver as is given in the dynamic contour of the formants." (p. 221). Rakerd, Verbrugge and Shankweiler (1984) replicated this effect of the consonantal context with a vowel monitoring task, although the magnitude of the effect was much smaller than in the original experiment of Strange et al.. This illustrates an important point in perception research: more often than not the magnitude - sometimes even the absence or presence - of an effect is dependent on the particular experimental design. Weenink (1984) questions the data of Strange et al. concerning the facilitating effects of a consonantal context; he gives more credit to the data of Macchi (1980), who failed to find a difference in identifiability between isolated vowels and vowels in a consonantal context when speaker, listeners, and response categories where carefully chosen.

The influence of the phonemic context is of course not restricted to the effect of consonants on the perception of vowels. Sherman (1952) investigated the influence of the vowel on the identification of the consonant in CV and VC segments from three different talkers. More specifically, she hypothesized that the presence of cues from the transitional period from the consonant to the vowel and vice versa might help the identification of the consonant. She found that the identification of consonants was indeed differentially affected by the presence of different vowels, and also, that this influence of the vowel was dependent on the position of the consonant in the syllable. The same phenomenon was recently investigated more systematically for the Dutch language by Klaassen-Don (1983); her conclusions point in the same direction:

1. Vowel transitions contain information about adjacent consonants; the amount of information depends on the specific consonant.
2. There is no clear dominance in perceptual information about adjacent consonants in VC over CV transitions (4): some consonants are better identified in VC syllable whereas other are better identified in CV syllables.
3. Vowel transitions in excerpts of running speech contain less information about adjacent consonants than vowel transitions in utterances produced in isolation.

The perception of consonant phonemes is not only influenced in vocalic environments, but also in consonant clusters; see for instance Van den Berg and Slis (1984) concerning the assimilation of voice in Dutch two-consonant clusters.

Apart from the influences on the spectral characteristics of speech sounds, the phonemic context also has effects in the temporal domain. The durations of both vowels and consonants depend to a substantial degree on the specific phonemic environment in which they are uttered. The relative duration of vowels seems to be determined by the nature of the following consonant: vowels before nasals tend to be longer than vowels before fricatives, whereas the latter tend to be longer than vowels before plosives. The durations of consonants is also influenced by the surrounding vowels and consonants. Most of the durational differences have an articulatory basis (Nooteboom and Cohen, 1984).

So far, only the mutual influences of adjacent speech sounds have been considered. But the interacting influences of phonemes (5) go far beyond the strict neighbouring sounds. Effects of coarticulation for instance were shown to operate over several consonant phonemes

(Bell-Berti and Harris, 1982). Some of the more remote effects of the acoustic-phonetic context will be treated in 2.2, 2.3, and 2.4.

2.2 SPEAKER VARIATION

The acoustic structure of speech shows marked differences among men, women, and children, and very often even the differences within a group of speakers of the same sex are considerable. These differences between speakers are most easily demonstrated as differences in the vowel spaces (6) of the different speakers. Despite the sometimes rather large differences in the formant frequencies of one vowel among different speakers, people are able to identify these acoustically different signals as instances of the same vowel. This remarkable performance has led investigators to hypothesize some normalization process: according to Joos (1948, cited in Ladefoged and Broadbent, 1957), listeners calibrate a speaker's vowel space on the basis of a short stretch of speech. This short stretch of preceding speech is necessary for creating a referential framework against which other vowels of the same speaker can be evaluated.

Since the acoustic structure of speech varies from speaker to speaker it might be hypothesized that a set of vowels from one speaker might be better identified than a random mixture of vowels from different speakers. This hypothesis was tested in the experiment on vowel identification of Strange et al. (1976). They found a statistically significant support for this hypothesis, although the effect of the consonantal context (see 2.1) far outweighed the effect of speaker variation. This effect of the single speaker versus the multiple speakers condition was replicated in one of the experiments of Verbrugge et al. (1976). In another experiment they performed a more direct test of the normalization hypothesis. In one condition of their experiment the /hVd/ test syllables of different speakers were preceded by three point vowels (/i/, /a/, /u/; in a /kVp/ syllable) that were thought to define a speaker's idiosyncratic vowel space to which later vowel tokens might be referred. Apart from the influence on the listener's response bias these precursors hardly had any effect. In a third experiment the /pVp/ test syllable was produced in a destressed position in a neutral carrier sentence. Presenting the excised /pVp/ syllables preceded by a precursor of point vowels produced in isolation actually increased the number of errors compared to the presentation without precursor. Presenting the destressed /pVp/ syllables in the sentences in which they were originally produced decreased the error rate. Verbrugge et al. (1976), and also Strange et al. (1976) do not consider their data as very convincing support for a proces of speaker normalization. Their observation that the talker- dependent acoustic variation does not seem to pose a major perceptual problem for the listener does not, however explain how the listener copes with this variation.

The superior performance of listeners in identifying vowels of one speaker compared to their performance when listening to vowels of different speakers was demonstrated for the Dutch language by Van Balen (1977), and more recently by Weenink (1985) (7). Both Van Balen (1977) and Weenink (1984) disagree with the conclusions of Strange et al. concerning their minimization of normalization problems. Van Balen found that the difference between the single speaker and the variable speaker condition was greatest in the second half of the listening test; he interpretes this as fatiguing effects of listening to different voices.

In the classical study of Ladefoged and Broadbent (1957) subjects had to identify the vowel of synthetically produced /bVt/ segments that were preceded by a neutral precursor sentence. This sentence was synthetically generated with six different formant structures. The results clearly indicated that the identification of the vowel in the test word (the /bVt/ segment) was dependent on the auditory context (i.e. the formant structure of the vowels in the precursor sentence) in which it occured. These results seem to be at variance with the results of Verbrugge et al.. This variance cannot be explained by the fact that Ladefoged and Broadbent used synthetically manipulated stimuli since Dechovitz (1977) obtained similar results in a modified replication of their experiment with natural speech stimuli. A possible account for this discrepancy is the lack of perceived continuity between the three precursor vowels produced in isolation and the test vowels in the experiment of Verbrugge et al.. Experiments of Dorman, Raphael and Liberman (1979) and Dechovitz, Rakerd and Verbrugge (1980) suggest "... that discontinuities in the production of speech can constrain the stretch of signal over which the features are integrated in phonetic perception." (Dechovitz et al., 1980).

2.3 SPEAKING RATE

Changing the rate of speaking has consequences for the number and the durations of pauses and it also affects the articulation time of the phonetic segments in an utterance. These changes in articulation time imply a modification of the temporal and possibly also the spectral parameters of speech. In general, a higher speaking rate means that nearly all phonetic segments are shortened, although the various segments are not shortened proportionally. Whether a higher speaking rate also has spectrally reducing effects seems to be an unresolved issue (Miller, 1981). Some authors claim that a higher speech tempo results in a certain degree of spectral reduction that is mediated primarily by the temporal reduction. More recent evidence however fails to replicate those instances of spectral reduction. The data of Koopmans-van Beinum (1980) are inconclusive in this respect: the observed relation between vowel duration and formant shifts cannot be interpreted properly because speaking rate was not an explicit variable.

The observation that listeners seem to have no apparent trouble in identifying speech of different rates, together with the acoustical evidence for the temporal reduction due to variations in this rate suggests compensatory mechanisms in the process of speech perception. The perception of a variety of phonetic distinctions is indeed sensitive to the perceived rate of articulation. Obviously, the referential framework not only specifies spectral characteristics of a speaker's utterance, but also temporal ones.

Ainsworth (1974) investigated the effect of the rhythm of a precursor sentence on the identification of a test vowel. In an earlier experiment (Ainsworth, 1971) he had already demonstrated the importance of vowel duration as a cue to the identity of synthetic vowels. In this experiment subjects had to identify a synthetic vowel in a consonantal context (/hVd/), that was preceded by three repetitions of the vowel /ə/. The time between the precursor vowels was equal to their length; this length varied from 120 to 600 msec, in steps of 120 msec. The duration of the test vowel also varied. It was found that the rhythm of the precursor sequence influenced the perceived identity of the synthetic vowel, especially if the formant frequencies and the duration of the test vowel were such as to render its identity ambiguous. In a replication of this experiment it was demonstrated that the nature of the vowel in the precursor sequence did not interfere with the effect of the rhythm. These results point to the perception of the relative durations of speech sounds. The results should however be considered with caution for two reasons. In the first place it should be noted that the durations of both the precursor and the test vowels in the experiments of Ainsworth seem quite long in the light of more recent measurements on vowel durations (see for instance Klatt (1976) for the English language; Koopmans-van Beinum (1980) for the Dutch language). Secondly, from the experiments of Ainsworth it cannot be ruled out that the effects of the rhythm of the precursor sequence are effects of the duration of the interval between the precursor vowels, since this was confounded with the length of the precursor itself.

In one of the experiments of Verbrugge et al. (1976) subjects had to identify vowels in /pVp/ syllables that were rapidly spoken in a destressed position of a carrier sentence (see 2.2). Isolated presentation of these /pVp/ syllables resulted in more errors than isolated presentation of CVC syllables that were produced in isolation. A closer analysis of the errors revealed that the vowels in the /pVp/ syllables were evaluated as being produced in isolation; as a result of this perceptual bias most errors were responses of more 'centralized' vowels. "The tendency of listeners to select more 'central' vowel responses suggests that they underestimated the tempo at which the excised syllables were spoken." (Verbrugge et al., 1976, p.207). This tendency was even more pronounced in the error data of the subjects in the point vowel precursor condition. In this condition the /pVp/ syllables were preceded by a precursor containing the point vowels (/hi/, /ha/, and /hu/). This precursor was produced by the same speaker of the test syllable, but in isolation. Obviously, the precursor and the test syllable were produced at very different tempi. Analysis of the errors suggests that instead of calibrating listeners to the formant structure of a speaker's vowel space, the precursor calibrated the listeners to an inappropriate tempo, leading to a large increase of incorrect 'centralized' reponses.

Verbrugge and Shankweiler (1977) performed an experiment in which /pVp/ syllables in a neutral carrier sentence were produced at a slow and at a fast rate. The /pVp/ syllables were presented in isolation, in their original sentence frames, and in sentence frames of the opposite rate. Again it was shown that most errors in the isolated presentation of the /pVp/ syllables were reduced vowel responses. Interchanging the rate of the sentence frame and that of the /pVp/ syllable had an asymmetrical effect: the identification of the slow syllables in the fast sentence frames was not impaired, whereas the identification of fast syllables in slow sentence frames showed a substantial decrease in performance. In identifying the fast syllables in the slow sentences, subjects were obviously misled by the inappropriate rate of the sentence frame as most errors were reduced vowel responses. To account for the asymmetry the authors claim that it is possible that the identification of slow syllables was unaffected by the context because they are fully specified by the information they contain, whereas the fast syllables need information from the surrounding context.

In a replication of the work of Verbrugge and Shankweiler (1977) with a different speaker, a different consonantal environment and a different carrier sentence, Johnson and Strange (1982) had subjects to identify vowels in /tVt/ syllables that were produced at a normal or a fast rate in a neutral carrier sentence. Vowels in the normal rate syllables were identified very well, whether they were presented in isolation, in an inappropriate fast rate sentence frame, or in their original sentence frame. For the fast syllables however the context in which they were presented had a large effect on the accuracy of the performance. Most errors, 19%, were made if these fast rate syllables were presented in isolation; presentation of the fast rate syllables in the (inappropriate) normal rate sentence frame led to 11% identification errors, whereas presentation of these fast rate syllables in their original fast rate sentence frame reduced this figure to only 6%.

Vowels can be divided in intrinsically long and intrinsically short ones (Klatt, 1976; Johnson and Strange, 1982; Koopmans-van Beinum, 1980). Analysis of the errors that were made in identifying the fast rate syllables in the experiments of Verbrugge and Shankweiler (1977) and those of Johnson and Strange (1982) reveals that most of these errors were misidentifications of intrinsically long vowels, that were perceived as their spectrally similar short counterparts. The fact that the identification of intrinsically nonlong vowels in the fast rate syllables was hardly influenced by the context in which they were presented, even though these vowels showed about the same amount of spectral reduction as the intrinsically long vowels is taken as evidence that "... context exerted its primary influence on listener's sensitivity to temporal rather than spectral (i.e., target) information for vowel identity." (Johnson and Strange, 1982, p.1765). In a partial replication of one of the experiments of Johnson and Strange, Van Bergem and Drullman (1985) failed to find evidence for similar misidentifications of intrinsically long, rapidly spoken, Dutch vowels presented in an inappropriate context (that is, either in a normal tempo context, or in isolation). A possible cause of this failure to replicate the findings of Johnson and Strange is the rather large measured difference between the durations of the intrinsically long and intrinsically short vowels compared to the rate induced differences in vowel duration. It is also possible that the effect of the context is totally obscured by the very high performance of the subjects (ceiling effects).

The speech tempo not only modifies the temporal and spectral parameters of vowels, but also affects temporal cues to consonants. In her review Miller (1981) mentions a number of those characteristics that are known to be affected by the speaking rate. Again it is shown that the perceptual boundaries for some consonantal distinctions are not invariant, but dependent on the perceived rate of speech. The distinction between single and geminate consonants (i.e.'topic' versus 'top pick') is cued by the closure duration before the intervocalic stop. The crossover point is dependent on the rate of speech. In a Dutch variant of the 'slit-split' experiment Pols (1984) demonstrated that the duration of the silence between the /s/ and the /l/ determines whether /sl../ or /spl../ will be heard, and that the crossover point is dependent on a variety of context effects, including speaking rate. Voicing is determined by a variety of cues; the particular set of cues varies with the syllabic position of the consonant. Some of these cues are temporal, and known to be affected by the rate of speech. In an experiment Port (1976, cited in Miller, 1981) showed that both the duration of the closure and the duration of the vowel (cues to the voiced/voiceless distinction) were affected by the speaking rate, and that the contrast for both cues diminished as the speaking rate increased. More interestingly, it was suggested that at faster rates, when the temporal cues were less distinctive, the voicing distinction was more reliably cued by the absence or presence of glottal pulsing. In some cases the manner of articulation is (at least partly) cued by temporal aspects. Miller mentions the fricative/affricative, and the stop/semivowel distinctions: temporal variables that were shown to cue these distinctions were perceived in a relative way.

Several researchers have been concerned with the determination of the most informative part of the context for the perception of the speaking rate at a given target syllable. In general, it seems that the temporal characteristics (i.e. durations of vowels, durations of transitions) of the surrounding context all contribute to the perception of tempo, but that the effect increases with the proximity to the target syllable. Whether a particular part of the surrounding context plays a part in determining the speaking rate is also dependent on the perceived continuity of the utterance (see for instance Dechovitz, Rakerd and Verbrugge, 1980). But also the temporal characteristics of the target syllable itself are important determiners of the perceived tempo. Port (1978) showed that the /b/-/p/ crossover point depended more on the intrinsic tempo of the target word than on the tempo of the surrounding carrier sentence. Even later occurring parts of the context can have influences on the perceived rate of speaking. In one of the experiments of Johnson and Strange (1982) it was found that the stressed syllable after the test syllable was most informative concerning the identity of the vowel in the test syllable. In a speeded classification task, Miller (1981) showed that this use of the later occurring context is not confined to unnatural tempo decisions, but also takes place in real-time speech recognition.

In most experimental studies that are reviewed by Miller (1981) the rate-specific cues to certain distinctions (i.e. vowel duration; transition duration) were also providing information about the rate of speaking. This problem, that a given acoustic property may both specify a phonetic contrast and also convey information about the speech tempo is examined for one specific case by Fitch (1981). In her study she showed that the perceived voicing of intervocalic plosives, as well as the perception of the speaking rate were both cued (at least partly) by the duration of the pre-plosive vowel and the duration of the closure. She demonstrated that it is possible to disentangle this confounding of cues by proposing cues of a higher order: the voiced/voiceless distinction seemed to be more reliably cued by the ratio of the durations of vowel and closure. More recent evidence however shows that this closure -to-vowel ratio is neither a very reliable nor an invariant cue to the perception of voicing (Luce and Charles-Luce, 1983). Obviously, the identification of phonemes and the perception of speech rate are dependent on very fine, local aspects of the surrounding context. Nooteboom (1981) avoids the contamination of cues by proposing that each word (or word-like unit) is perceptually represented by several words (or word-like units) corresponding to different speaking rates. In his version of Morton's logogen model word recognition is a passive process: it takes place as soon as a certain 'response strength' is reached.

2.4 MORE GLOBAL EFFECTS OF THE ACOUSTIC-PHONETIC CONTEXT

In his plea for the existence of a special mode for speech perception Repp (1981) distinguishes trading relations and context effects. Trading relations occur among different cues for the same phonetic contrasts. Repp speaks of context effects if a phonetic distinction is affected by the following or preceding context that is not part of the set of direct cues. This context may be adjacent or more remote. A typical example of a trading relation is the perceived voicing of intervocalic plosives that is cued by the duration of the preceding vowel and the duration of the closure. A typical example of a close context effect is the finding of Mann and Repp (1980; mentioned in Repp (1981)) that an ambiguous sound between /ʃ/ and /s/ is more often perceived as an instance of /s/ in a /u/ context than in a /a/ context. Examples of the effects of the more remote context are the effects of speaker variation (2.2) and effects of speaking rate (2.3).

Acoustic-phonetic information at the syllable-, word-, or sentence-level provides indirect cues to the interpretation of certain phonetic distinctions. A demonstration of the effects of indirect cues from the acoustic-phonetic context is given in an experiment by Pols and Schouten (1985). Subjects were given a two-way choice in ambiguous Dutch sentences which allowed two different intervocalic plosive consonants in the same position, i.e. 'Hij belde/telde twee keer' ('He rang/ counted twice'). Various parts of the VCV segment (plosive burst, vocalic transitions) had been deleted. The same VCV segments were also presented in isolation. In one condition of the experiment the burst and the VC transition had been deleted from the VCV segment, leaving only the CV transition intact. Presentation of the segments in a semantically non-informative context led to better recognition of the corresponding plosive consonants, compared with isolated presentation. Since the number of available direct cues was the same in both presentations the difference in recognition scores ought to be explained by indirect cues: the acoustic-phonetic information at the sentence level that facilitates the interpretation of the cues to the plosive consonant identity in the vocalic transitions.

In one of the experiments of Nooteboom and Doodeman (1980) the duration of a silent gap between a CVC word containing a test vowel and the rest of a sentence was manipulated, and this was found to affect the cross-over point of the vowel boundary (that is, the point were the perception of /ɑ/ and /a/ were equally likely). Nooteboom and Doodeman conclude that the results of their experiment "... confirms the hypothesis that the internal criterion for vowel length identification is primarily affected by the specific-phonetic structure of the surrounding speech material, ..." (Nooteboom and Doodeman, 1980, p.285). A related phenomenon is touched upon by Dechovitz, Rakerd and Verbrugge (1980) in their replication of an experiment of Dorman, Raphael and Liberman (1979). In this replication speakers produced the sentences 'Let's go shop' and 'Let's go' (phrase final intonation). Insertion of a silent period between 'go' and 'shop' induced the expected perception of 'chop' instead of 'shop' (see Dorman, Raphael and Liberman, 1979). In another condition 'shop' was appended to the second sentence ('Let's go'). Manipulation of the silence between 'go' and 'shop' did not have any effect at all: 'shop' remained 'shop' under all circumstances. Most probably, the perceptual continuity between 'Let's go' and 'shop' was disturbed by the phrase final intonation of the sentence.

Still another demonstration of more global aspects of the acoustic-phonetic context at the sentence level is given by Luce and Charles-Luce (1983) in their experiments on the invariance of the closure-to-vowel ratio. They measured this ratio at syllable final positions in CVC test words that were produced with vowels of different intrinsic durations, in different positions within the sentence, and with different subsequent local phonetic contexts (the first phoneme following the test word was either a reduced vowel /ə/ or a /t/). It was shown that both vowel duration and the closure-to-vowel ratio varied with all these factors.

In the experiments of Pickett and Pollack (1963; Pollack and Pickett, 1963) the intelligibility of excerpts from fluent speech increased with the duration of the speech sample. It should be noted of course that these results - although related to the above - cannot be explained entirely by acoustic-phonetic factors since the unit of intelligibility was studied at the level of words in a meaningful context. The trade-off between acoustic-phonetic and linguistic (semantic/syntactic) information is nicely illustrated by the fact that the tempo of speech hardly had any effect: the extra acoustic-phonetic information in the slower tempo segments was compensated for by more linguistic information in the fast tempo speech. This effect could not be explained by the length of the segment as such: resynthesized segments that were stretched in time without shifting the speech frequencies did not result in better intelligibility. Another demonstration of this interaction between linguistic and acoustic-phonetic information is given by Ringeling (1984). He demonstrated that the final consonant is identified much better in meaningful CVC words than in their (minimally different) nonsense equivalents.

3. METHODS OF INVESTIGATION: SOME SUGGESTIONS

There is a rich variety of experimental methods in phonetics. In this section some methods will be suggested that seem suited to study the effects of the local context on the interpretation of acoustic-phonetic information.

The main objective of the ZWO project mentioned in the introduction is to investigate whether speech segments presented in isolation will be identified less accurately than the same segments presented in a broader acoustic-phonetic context. More precisely, the objective of the project is to evaluate the effect of physical aspects of the local context on the interpretation of acoustic-phonetic information. This has implications for the size of the segments to be used in relevant experiments. In spoken word recognition listeners use information from all available sources: pragmatic, semantic/lexical, syntactic, and acoustic-phonetic information. In order to confine the recognition process to acoustic-phonetic information nonsense words or nonsense syllables (8) should be used, and this should also be stressed in the instructions to the subjects. The size of the segments should not be too large, since larger segments might contain enough intrinsic information to render the acoustic-phonetic information unambiguous (see Verbrugge and Shankweiler, 1977; Johnson and Strange, 1982). These restrictions seem to suggest the use of meaningless segments of the VCV, CVC, CV, and VC form.

The same sort of general remarks can be made concerning the context in which the test segments are to be presented. This context must in some way present a referential framework to which the acoustic information of the test segment can be referred; this context should however not specify the identity of the test segment by means of pragmatic, semantic/lexical, or syntactic information.

The effect of context can be investigated at the word/syllable level, and at the sentence level. At the syllable/word level it is possible to present the test segments preceded by a precursor of only a few syllables. One variable in this setup might be the presence or absence of the target phoneme of the test segment in the precursor. Another possibility is the presentation of the test segment in a broader context of the original utterance in order to establish perceptual continuity. Although this might have more effect on the identification of the test segment than presenting it after a precursor of syllables or words (see Dechovitz, Rakerd and Verbrugge, 1980), this method has one severe drawback: extending the context of a test segment introduces quite often semantical/lexical cues to the identity of the segment. In studying the effects of the local acoustic-phonetic context at the sentence level it is feasable to present test segments preceded by a full sentence precursor. Again it can be argued that the perceived discontinuity between the precursor sentence and the test segment attenuates possible effects. The perceptual continuity can be maintained by presenting the test segment in its original sentence frame, that is the sentence frame in which it was produced. In this case it is necessary that the semantic content of the sentence is neutralized. This can be accomplished by embedding the test segment in a neutral carrier sentence (see for instance Johnson and Strange (1982); they presented tVt segments in the following carrier sentence: 'Was it the tVt sound that you heard ?'). Another approach, that avoids the possible habituation effects of the repeated presentation of the same carrier sentence is the presentation of the test segment in the partially masked original sentence frame. This masking should inhibit the semantic information, but retain at least some of the long-term and short-term physical characteristics of the context. An effective masking noise would probably be a speech-envelope following noise as described by Horii, House, and Hughes (1970).

An interesting paradigm from the word recognition domain is introduced by Grosjean (1980): " Our paradigm - the gating paradigm - entails presenting a spoken language stimulus repeatedly and increasing its presentation time (duration from onset) at each successive pass. Depending on the questions at hand and the level of analysis, the stimulus can range from a simple CV syllable to a complex sentence and the presentation time (or gate) can also be made to vary. The subjects's task is to guess the stimulus being presented after each pass and to give a confidence rating based on the guess." (Grosjean, 1980, p.267). Grosjean used his technique to compare the word recognition process with gated nouns that were presented in isolation or in a context (9). This use suggests that this task might also be a promising one for studying the effects of the local acoustic-phonetic context on phonetic identification. Apart from the gating time needed for correct identification, this task also gives a set of incorrect responses and a corresponding set of confidence ratings, that can be compared for the gated segments presented in isolation to the gated segments in a context.

Salasoo and Pisoni (1982) extended this gating paradigm. In their procedure all content words in a sentence were replaced by envelope-following noise; in successive presentations of these sentences segments of this noise were replaced by segments (gates) of the original word. They also varied gating direction and redundancy. Words could either be presented in isolation, embedded in a semantic anomalous, syntactic context (i.e. 'The end home held the press') or embedded in a meaningful (semantic) context. The authors were primarily interested in the differences between the uses of knowledge-sources in the syntactic and the semantic context. The difference between the performance in the isolated versus the syntactic context condition - reflecting the combined effects of syntactic and acoustic-phonetic information - is hardly mentioned. Still it should be noted that some 82% of the word-length is needed to identify words in isolation, whereas this figure is reduced to 72% in the syntactic context. What is needed is a task to separate the syntactic from the acoustic-phonetic effects. Again, this seems to suggest the use of nonsense words, perhaps in a context of nonsense words. Salasoo and Pisoni also investigated the sequential nature of the gating paradigm. They hypothesized that this nature could possibly influence the recognition process in a number of ways: the repeated presentation of parts of the stimulus might facilitate the recognition; on the other hand the previously given (erroneous) responses may have influences on later occurring responses. To this end they varied the gating duration also between subjects. Most within-subjects effects were replicated in the between-subjects design. In the syntactic context however, the identification threshold (the gate duration at which 50% of the subjects correctly identified the gated word) was lower in the single presentation (between-subjects) condition: " The accumulation of conflicting top down semantic and syntactic information with successive repetitions appear to constrain the use of word-initial acoustic-phonetic information." (Salasoo and Pisoni, 1982, p. 134).

In a preliminary pilot experiment phoneme identification was investigated for various segments excerpted from a read text. The smallest segments were of the type CVC and VC^nV, in which C^n stands for a cluster of n consonants. Segments were selected within words as well as over wordboundaries. Larger segments had the structure VC^nVC^mV, whereas also word groups were presented. Nine subjects were asked to write down exactly what they heard by using orthographic spelling. Although the vowels and consonants were indeed better identified in the longer segments than in the short ones, this experimental paradigm had several drawbacks: it was not always possible to define the stimulus uniquely, because in continuous speech not every phoneme is well articulated. Also the interpretation of the responses of the listeners caused many problems because writing habits interfered with writing down what was heard exactly. For instance a VCV segment /orə/ from 'het horen van' should be scored as 'ooru', but many subjects wrote 'oru' or 'ore'. Some of these problems can be avoided by using trained subjects and phonetic symbols. Several alternative approaches are suggested above.

Notes

(1) This is a controversial issue. Some authors doubt about the status of phonemes as the natural response category: Nooteboom (1981) for instance argues that spoken word-recognition is not mediated by phonemes, but that instead phoneme-recognition is usually mediated by the recognition of words.

(2) See for instance Lieberman (1963). He had speakers produce a set of words in a highly redundant context: the same set of words was also produced in a non-redundant context. Both sets of rapidly spoken words, excised from their sentence frames and embedded in a broadband noise, were presented to subjects for identification. The words that were produced in a non-redundant context were identified more accurately than their equivalents produced in a redundant environment.

(3) Apart from being cues to certain pairs of phonemes the transitions seem to play an important part in preventing the speech stream to segregate in a vowel-stream and a consonant noise-stream, and in retaining the perception of the correct temporal order of the phonemes (Cole and Scott, 1974; Bregman and Dannenbring, 1973).

(4) In a study of Sharf and Hemeyer (1972) it was found that VC segments contained more information concerning the place of articulation of the adjacent consonant than their corresponding CV segments.

(5) At this point it doesn't matter whether these effects have an articulatory basis or not; fact is that they are perceptually present.

(6) Most vowels can be identified on the basis of the first two formants. A convenient way of representing a speaker's vowel space is the representation of this speaker's vowels in the F1 - F2 plane.

(7) Weenink, D.J.M. (1985). Unpublished data.

(8) More formally, the segments to be used should have no lexical entries in the native language of the listener.

(9) In this experiment Grosjean (1980) validated his procedure by replicating some well known word recognition effects. He also falsified

the assumption of the Marslen-Wilson word recognition model (Marslen-Wilson, 1980; 1981) by demonstrating that the so called word-initial cohort is not entirely based on acoustic-phonetic information.

References

Ainsworth, W.A. (1971). Duration as a cue in the recognition of synthetic vowels. *Journal of the Acoustical Society of America*, *51*, 648 - 651.

Ainsworth, W.A. (1974). The influence of precursive sequences on the perception of synthesized vowels. *Language and Speech*, *17*, 103 -109.

van Balen, C.W. (1977). Different views on problems of normalization. *Proceedings of the Institute of the University of Utrecht*, *2*, 32 - 46.

Bell - Berti, F., & Harris, K.S. (1982). Temporal patterns of coarticulation: Lip rounding. *Journal of the Acoustical Society of America*, *71*, 449 - 454.

van den Berg, R.J.H., & Slis, I.M. (1984). The perception of assimilation of voice as a function of segmental duration and lingtic context. Report of the Institute of Phonetics Nijmegen, (IFN), University of Nijmegen.

van Bergem, D., & Drullman, R. (1985). Invloed van het spreektempo spreektempo van de context op de herkenning van Nederlandse klinkers in normaal en snel uitgesproken tVt - uitingen. *IFA Report*, nr. 77.

Bregman, A., & Dannenbring, G. (1973). The effect of continuity on auditory stream segregation. *Perception and Psychophysics*, *13*, 308 - 312.

Cole, R.A., & Scott,B. (1974). Toward a theory of speech perception. *Psychological Review*, *81*, 348 - 374.

Daniloff, R.G., & Hammarberg, R.E. (1973). On identifying coarticulation. *Journal of Phonetics*, *1*, 239 - 248.

Dechovitz, D.R. (1977). Information conveyed by vowels: a confirmation. *Haskins SR*, *51/52*, 213 - 219.

Dechovitz, D.R., Rakerd, B.D., & Verbrugge, R.R. (1980). Effects of utterance continuity on phonetic judgements. *Haskins SR*, *62*, 101 - 116.

Dorman, M.F., Raphael, L.J., & Liberman, A.M. (1979). Some experiments on the sound of silence in phonetic perception. *Journal of the Acoustical Society of America*, *65*, 1518 -1532.

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

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*, 267 - 283.

Horii, Y., House, A.S., & Hughes, G.W. (1971). A masking noise with speech-envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America*, *49*, 1849 - 1856.

Johnson, T.L., & Strange, W. (1982). Perceptual constancy of vowels in rapid speech. *Journal of the Acoustical Society of America*, *72*, 1761 - 1770.

Joos, M.A. (1948). Acoustic phonetics. *Language* Suppl. 24, 1 - 36.

Klaassen-Don, L.E.O. (1983). The influence of vowels on the perception of consonants. Doctoral dissertation, University of Leiden.

Klatt, D.H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59*, 1208 - 1221.

Koopmans-van Beinum, F.J. (1980). Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions. Doctoral dissertation, University of Amsterdam.

Ladefoged, P., & Broadbent, D.E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98 - 104.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431 - 461.

Liberman, A.M., Delattre, P.C., & Cooper, F.S. (1952). The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology, 65*, 497 - 516.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6*, 172 - 187.

Luce, P.A., & Charles-Luce, J. (1983). Contextual effects on the consonant/vowel ratio. *Research on Speech Perception Progress Report, 9*, 3 - 37.

Macchi, M.J. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *Journal of the Acoustical Society of America, 68*, 1636 - 1642.

Mann, V.A., & Repp, B.H. (1980). Influence of vocalic context on perception of the /ʃ/ - /s/ distinction. *Perception and Psychophysics, 28*, 213 - 228.

Marslen-Wilson, W.D. (1980). Speech understanding as a psychological process. In J.C. Simon (Ed.): *Spoken language generation and understanding*. Dordrecht: Reidel.

Marslen-Wilson, W.D., & Tyler, L.K. (1981). Control processes in speech understanding. *Philosophical Transactions of the Royal Society of London., B. Biological Sciences, 295*, 317 - 332.

Miller, J.L. (1981). Effects of speaking rate on segmental distinctions. In: P.D. Eimas and J.L. Miller (Eds.): *Perspectives on the study of speech*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*, 165 - 178.

Nooteboom, S.G. (1981). Speech rate and segmental perception or the role of words in phoneme identification. In G.E. Stelmach and P.A. Vroon (Eds.): *Advances in psychology, vol. 7: The cognitive representation of speech.* Amsterdam: North-Halland Publishing Company, 143 - 150.

Nooteboom, S.G., & Cohen, A. (1984). Spreken en verstaan: Een nieuwe inleiding tot de experimentele fonetiek. Assen: Van Gorcum.

Nooteboom, S.G., & Doodeman, G.J.N. (1980). Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America, 67*, 276 - 287.

Pickett, J.M., & Pollack, I. (1963). Intelligibility of excerpts from fluent speech: effects of rate of utterance and durations of excerpts. *Language and Speech*, *6*, 151 - 164.

Pisoni, D.B. (1983). Contextual variability and the problem of acoustic-phonetic invariance in speech. *Research on Speech Perception Progress Report*, *9*, 245 - 257.

Pollack, I., & Pickett, J.M. (1963). The intelligibility of excerpts from conversation. *Language and Speech*, *6*, 165 - 171.

Pols, L.C.W. (1983). Variation and interaction in speech. *IFA Report* nr. 74.

Pols, L.C.W. (1984). Phoneme identification in isolated stimuli and in context. *Proceedings of the Institute of Phonetic Sciences Amsterdam*, *8*, 33 - 40.

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

Port, R.F. (1976). The influence of speaking tempo on the duration of stressed vowel and medial stop in English Trochee words. Unpublished doctoral dissertation, University of Connecticut.

Port, R.F. (1978). Effects of word-internal versus word-external tempo on the voicing boundary for medial stop closure. *Haskins SR*, *55/56*, 189 - 198.

Rakerd, R.R., Verbrugge, R.R., & Shankweiler, D.P. (1984). Monitoring for vowels in isolation and in a consonantal context. *Journal of the Acoustical Society of America*, *76*, 27 - 31.

Repp, B.H. (1981). Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception. *Haskins SR*, *67/68*, 1 - 40.

Ringeling, J.C.T. (1984). Reducing redundancy in normal, soft and whispered speech: a study on native and near native perception. Doctoral dissertation, University of Utrecht.

Salasoo, A., & Pisoni, D.B. (1980). Sources of knowledge in spoken word recognition. *Research on Speech Perception Progress Report*, *8*, 105 - 145.

Sharf, D.J., & Hemeyer, T. (1972). Identification of place of consonant articulation from vowel formant transitions. *Journal of the Acoustical Society of America*, *51*, 652 - 658.

Sherman, D. (1952). The influence of vowels on the recognition of adjacent consonants. *The Journal of Speech and Hearing Disorders*, *17*, 198 - 212.

Stevens, K.N., & Blumstein, S.E. (1981). The search for invariant acoustic correlates of phonetic factors. In P.D. Eimas and J.L. Miller (Eds.): *Perspectives on the study of sppech*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Strange, W., Verbrugge, R.R., Shankweiler, D.P., & Edman, T.R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, *60*, 213 - 224.

Verbrugge, R.R., & Shankweiler, D. (1977). Prosodic information for vowel identity. *Journal of the Acoustical Society of America*, *61*, S39 (abstract).

Verbrugge, R.R., Strange, W., Shankweiler, D.P., & Edman, T.R. (1976). What information enables a listener to map a talker's vowel space? _Journal of the Acoustical Society of America_, _60_, 198 - 212.

Weenink, D.J.M. (1984). Literature overview on perceptual and physical normalization of speaker variation. _Proceedings of the Institute of Phonetic Sciences Amsterdam_, _8_, 5 - 17.