A CASE FOR GLOBAL LISTENING STRATEGIES

J.C.T. RINGELING

Dept. of English
Utrecht University
Oudenoord 6
3513 ER Utrecht, The Netherlands

W. EEFTING

Dept. of Phonetics
Utrecht University
Trans 14
3512 JK Utrecht, The Netherlands

ABSTRACT

A case is made for global perceptual strategies. In poor listening conditions subjects appear to be able to perceive and comprehend elliptic speech, albeit with some difficulty. If sufficient semantic information is available, they seem capable of basing themselves on global characteristics in speech sounds, particularly on information related to place of articulation. The study pleads for the formulation of perceptual features to obtain a better insight into the processes operative in speech perception.

1. INTRODUCTION

When Zue (1) showed that a trained spectrogram reader can recognize a substantial number of words from spectral information, the discussion on invariant features in speech perception gained new ground. In the seventies many linguists did not take invariance very seriously, although some invariant features were generally accepted (see e.g. 2 and 3). Naturally, Zue´s success in reading spectrograms was partly caused by extensive use of linguistic expectancy to solve ambiguities, but it made clear that some sort of invariance must be present in speech, although perhaps of a different nature than had traditionally been accepted in terms of linguistic features.

Carlson, Elenius, Granström and Hunnicut (4) and more recently Veenhof and Bloothooft (5) have shown that it may theoretically be possible in many cases to come a long way in arriving at word identification by specifying the acoustic information on the basis of broad phonetic categories. They showed that a classification of phonemes into global categories such as plosives, nasals, fricatives, remaining sonorant or vowel, often provides sufficient information to limit the number of words in a cohort for recognition to take place. This insight that word recognition may be feasible on the basis of a broad phonetic classification has proved helpful in automatic speech recognition (6,7).

However, it is by no means certain if human perception can adequately use a broad phonetic classification in the process of listening to connected speech, and if so, it remains questionable whether listeners base themselves on the same phonetic categories as are frequently adopted in theoretical studies. If we wish to find out what phonetic underlying features can be used in human perception, it is imperative that listening tasks are constructed which vary the amount of acoustic information along global phonetic parameters. An attempt at such a task was an informal study by Ringeling (8), who demonstrated that Dutch listeners could fairly successfully identify sentences in which all consonants had been replaced for consonants that were similar with regard to place of articulation, in such a way that the phonotactic constraints of Dutch were not violated. The use of elliptic speech (see e.g. 9) thus served to manipulate the amount of acoustic information in the speech signal. An English example in ordinary orthography would be the conversion of the saying: ´no place like home´ to ´mow crafe wipone´. The resulting sentences sounded Dutch, but could not readily be understood. However, when redundancy of the acoustic signal was reduced by adding noise to the sentences, it turned out that listeners produced much better recognition scores on the same material. One of the most interesting findings of this study was that subjects were rarely aware of the manipulations that had been carried out. This suggests that a global phonetic analysis had taken place on the basis of similarity of place of articulation. In view of the task at hand, which drew heavily on an intensive use of linguistic expectancy, sentences with constraining context were understood much better than those with relatively neutral content.

Van der Woude (10) based a study on this idea. He investigated the theoretical possibility of arriving at unique identifiability of words by grouping consonants together, either on the basis of manner of articulation, or on the basis of place of articulation. On the basis of a random sample of 100 words from 68,000 word tokens (12,000 word types), he found that specification of Dutch words in terms of broad phonetic classes thus defined, did not yield a clear theoretical advantage to either classification. In his definition of patterns, leaving vowel-quality intact, he found

72 % unique identifiability for grouping consonants on the basis of changes in place of articulation and 78 % for grouping them together on the basis of changes in manner of articulation. Moreover, for those words that could theoretically not be identified uniquely, the remaining cohort of word candidates from a 12,000 wordtype lexicon never exceeded four and was only two in 80 % of the instances.

Yet, even if theoretically both types of classification would seem to qualify as potential approaches for listening strategies, it is evident that actual speech perception need not avail itself of these theoretical possibilities. In fact, it would seem highly unlikely that both strategies are equally effective, since it is well-known from the literature that perceptual confusions on the basis of changes in place of articulation are much more frequent than those on the basis of manner of articulation (see e.g. 11). It was therefore decided to undertake a preliminary study into the perceptual relevance of global phonetic listening strategies on the basis of place-changed and manner-changed consonants.

Miller and Isard as early as 1963 (12), showed that listeners can extract the linguistic content of a message if they have access to normal syntactic and semantic information when speech is presented under high levels of noise. If this linguistic information is also deteriorated, the listeners identification will suffer accordingly. We therefore expect that in the experiment reported on here, sentences with high semantic constraints will be identified correctly more often than neutrally constrained sentences. Moreover, on the basis of what was stated above, we will expect to find a discrepancy in recognition scores based on the amount of phonetic information. If the place-changed and manner-changed consonants lead to unique word patterns, better recognition scores are expected than if the resulting word patterns leave room for ambiguities.

2. METHOD

2.1 Stimuli

21 Sentences were synthesized using the diphone synthesis system, developed by Elsendoorn (13). By using diphones it was possible to preserve a natural flow of speech while changing the consonants at will. Each sentence was synthesized in three conditions:

place-changes: all consonants were systematically replaced for other consonants differing in place of articulation, in conformity with the phonotactic rules of Dutch. The feature voiced/unvoiced in these elliptic sentences remained unaffected.

manner-changes: idem, but differing in manner of articulation.

control: these were stimuli syntesized without manipulation of consonant features.

Three types of sentences were constructed:

sentences consisting of short words (non-unique pattern in terms of global perceptual categories) and neutrally constrained,

sentences consisting of long words (unique pattern in terms of global perceptual categories) and neutrally constrained,

proverbs/sayings, semantically highly constrained.

In corresponding sentences in the three conditions, overall intensity and intonation were kept identical. All sentences were masked with noise at an S/N-ratio of -6 dB, which had resulted in a 90 % correct recognition score of the 'control' sentences in a pilot experiment. Noise was turned on 1 second before the signal started and turned off .5 s after the speech signal had ended.

2.2 Subjects

21 native speakers of Dutch, aged 20 to 30, served as unpaid participants. No subject reported hearing defects. They were members of staff or students at Utrecht University. Some were phonetically trained, but none were familiar with the stimuli or the aims of the experiment.

2.3 Procedure

In a sound treated room, subjects listened to 3 trial sentences and 18 target sentences. The stimuli were presented binaurally over headphones at a comfortable listening level, using a Revox tape recorder. Each sentence was repeated after a 1 second interval. Items were preceded by a short 200 Hz tone. After each pair of sentences there was an interval of circa 13 seconds to give subjects time to write down their responses. Subjects were encouraged to write down partial responses as well, even if those consisted of separate sounds, fragments of sentences that seemed anomalous etc. Each subject heard each sentence only in one condition to prevent learning effects.

3. RESULTS

Reactions from the subjects and the amount of missing data (63 % of the sentences, 35 % of the content words) indicated that the task was considered quite difficult. In some instances subjects were aware that the material had been manipulated.

In table I the number of correctly identified content words is presented. The condition MANNER-CHANGED was by far the most unintelligible. On average only 3 % of the words were reported correctly. For neutrally constrained sentences in the PLACE-CHANGED condition circa 10 % of the words were identified correctly. It is in this condition that the powerful influence of linguistic constraints can most clearly be observed. In highly constrained sentences over 50 % of the words were recognized. No differences are found within conditions with respect to the type of words presented. Apparently greater word-length, and consequently a higher degree of uniqueness of the word pattern, did not facilitate recognition.

Table I: Number of word responses, subdivided into words reported correctly, incorrectly and failure to respond, for neutrally and highly constrained sentences, in the experimental conditions CONTROL, PLACE-CHANGED and MANNER-CHANGED

HIGHLY CONSTRAINED

	CONTROL	PLACE CHANGED	MANNER CHANGED
N	141	152	141
Correct	127 (90%)	79 (52%)	5 (3%)
Incorrect	3 (2%)	28 (18%)	64 (45%)
Missing	11 (8%)	45 (30%)	72 (52%)

NEUTRALLY CONSTRAINED

UNIQUE WORD-PATTERN

	CONTROL	PLACE CHANGED	MANNER CHANGED
N	160	146	148
Correct	119 (74%)	12 (8%)	4 (2%)
Incorrect	14 (9%)	35 (24%)	27 (19%)
Missing	27 (17%)	99 (68%)	117 (79%)

NON-UNIQUE WORD-PATTERN

	CONTROL	PLACE CHANGED	MANNER CHANGED
N	121	120	119
Correct	90 (74%)	12 (10%)	5 (4%)
Incorrect	17 (14%)	43 (35%)	43 (35%)
Missing	14 (12%)	65 (55%)	71 (61%)

From the data it appeared that correct sentence recognition in the PLACE-CHANGED and MANNER-CHANGED conditions was rare. 90 % of the control sentences were reported correctly when the context was highly constrained. In neutrally constrained sentences this percentage was circa 50 %. For the manipulated versions correct sentence recognition was always below 5 %, except for highly constrained sentences in the PLACE-CHANGED condition, which obtained a 34 % correct recognition score.

4. DISCUSSION AND CONCLUSIONS

In this experiment we hoped to learn something about the type of acoustic and non-sensory information listeners may employ when poor listening circumstances force them to use a global perceptual analysis. Because of the preliminary nature of the experiment, our conclusions can only establish promising areas for further research:

a. Changing manner of articulation does not appear to be a salient characteristic in the identification of spoken sentences.

b. Changing place of articulation appeared to yield satisfactory results in case sufficient linguistic constraints were available. Subjects' comments indicated that the message can be reconstructed properly and phonetic distortions mostly go unnoticed.

c. Word structure did not turn out to help the listeners in identifying the words correctly, although subjects did attempt to respond to uniquely patterned words more frequently than to non-uniquely patterned words, as can be seen from the percentages of incorrectly identified words. It may well be that uniqueness of word-pattern plays a more salient part if stimulus material is presented in which word boundaries are better available to the listener.

Although the outcome of the experiment clearly shows that PLACE-CHANGED manipulation plays a more important part than MANNER-CHANGED manipulation, the actual recognition scores remain disappointingly low if linguistic constraints are weak. It should be kept in mind, however, that our quest for global perceptual features was hampered by the choice of synthesized material. We did not synthesize plosive-like sounds or nasal-like sounds, but used substitutions of existing phonemes. This means that the listeners were purposely deluded. In view of this, the outcome of the experiment is quite promising . It may well be possible to arrive at core-features underlying perception in the future.

These features may be rather different from what we have traditionally used in articulatory or linguistic terminology. It is, for instance, noteworthy that in multidimensional scaling techniques, when applied to perceptual studies, the dimensions often do not correspond to traditional feature classifications. Similarly, in studies on broad phonetic classifications (such as 4 and 5) non-traditional as well as traditional 'features are used.

We find it important that research should be carried out into the perceptually salient features so as to arrive at a set of variables that are of primary importance to speech perception. The variables in use now, are sometimes haphazard and only used 'because they appear to work'. If we obtain a better understanding of the fundamentally important variables in speech perception, many issues may become more accessible. Notice in this respect that Van der Woude (6) found no theoretical reason to prefer a classification based on PLACE-CHANGED consonants to one that was based on MANNER-CHANGED consonants. But actual perceptual strategies evidently favour a PLACE-CHANGED approach. Nevertheless, we are by no means certain yet, if a PLACE-CHANGED categorization is the best possible approach in global listening strategies. It would be highly counterintuitive if this was not the case, but we will need to lay bare the fundamental features of speech perception first.

REFERENCES:

1. V.W. Zue (1983) Proposal for an Isolated Word Recognition System Based on Phonetic Knowledge and Structural Constraints, in A. Cohen and M. Van den Broecke (eds.): *Abstracts of the Tenth International Congress of Phonetic Sciences*, Foris, Dordrecht, The Netherlands: pp. 299 - 305.
2. K.N. Stevens (1975) Potential Role of Property Detectors in the Perception of Consonants, in G. Fant and M.A.A. Tatham (eds.): *Auditory Analysis and Perception of Speech*, Academic Press, New York, London.
3. M. Umeda (1977) Consonant Duration in American English, *JASA* 61: pp. 846 - 858.
4. R. Carlson, K. Elenius, B. Granström and S. Hunnicut (1985) Phonetic and Orthographic Properties of the Basic Vocabulary of Five European Languages, *STL-QPSR* 1: pp. 63 - 93.
5. T. Veenhof and G. Bloothooft (1987) Statistics of Sequences of Broad Phonetic Classes in Newspaper Dutch, *PRIPU* 12.1: pp. 39 - 56.
6. D.W. Shipman and V.W. Zue (1982) Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems, *Conference Record, IEEE 82, Int. Conf. on Acoustics, Speech and Signal Processing*: pp. 546 - 549.
7. D.P. Huttenlocher (1986) A Broad Phonetic Classifier, *Proc. of the ICASSP-Tokyo*: pp. 2259 - 2262.
8. J.C.T. Ringeling (1986) Luisteren is Gokken, *Toegepaste Taalwetenschap in Artikelen* 25.2: pp. 28 - 36.
9. Z.S. Bond (1981) Listening to Elliptic Speech: Pay Attention to Stressed Vowels, *Journal of Phonetics* 9: pp. 89 - 96.
10. C. Van der Woude (1987) A Theoretical Look at Global Perception, unpublished M.A. Thesis, English Dept., Utrecht University, The Netherlands.
11. G.A. Miller and P.E. Nicely (1954) An Analysis of Perceptual Confusions Among Some English Consonants, *JASA* 27.2: pp. 338 - 352.
12. G.A. Miller and S. Isard (1963) Some Perceptual Consequences of Linguistic Rules, *Journal of Verbal Learning and Verbal Behavior*, 8: pp. 217 - 228.
13. B.A.G. Elsendoorn (1984) Heading for a Diphone Speech Synthesis System for Dutch, *IPO Annual Progress Report*, 19: pp. 32 - 35.