ABSTRACT AND LITERAL ASPECTS OF LEXICAL MEMORY FOR SPEECH

P. Standen
Department of Psychology
University of Western Australia

ABSTRACT - Many models of human speech recognition propose a lexicon where each word is represented by a single fixed string of abstract phonetic or phonemic elements. In contrast, recent models of long term memory require the storage of literal features of individual word tokens, including speaker's voice. In an experiment using the repetition priming phenomena as an indicator of lexical access, it is found that the lexicon is largely insensitive to variation in speaker's voice, even when an unfamiliar accent is present.

INTRODUCTION

Theories of speech recognition usually refer to an internal lexicon containing the information which distinguishes one word (or morpheme) from all others. In many accounts the burden of recognition is given primarily to perceptual units which detect phonetic or similar segments in the speech signal, and then match these in some trivial way with the abstract prototypical forms found in the lexicon. From this point of view, problems such as the need to normalize across speakers are to be solved by examining more sophisticated hypotheses about how segments are coded, for example by using bark-phone scales or by examining global properties of the spectrum (see Pisoni, 1985). This view of the lexicon as abstract in regard to particular instances also appears, for different reasons, in traditional linguistics and studies of perception and short term memory for linguistic stimuli.

A different picture arises when the lexicon is considered as part of long term memory. A number of studies show the storage of literal details such as speaker's voice that are incidental to the linguistic content of a message. Other work has shown that token based models of memory can account for effects previously thought to require storage of abstract information. This paper briefly outlines these abstract and literal views of memory, and reports an experiment examining the possibility that at least some speaker-specific information is retained at the lexical level rather than being normalized during segment identification.

PHONOLOGICAL VARIATION AND THE LEXICON

The lexicon is not often seen as a means for accounting for the variable instantiation of speech, perhaps due to the influence of linguistic theory which, in general, emphasizes the rules by which lexical strings are converted to sounds and places in the lexicon only those details which cannot be explained by rule. These rules are normally applied to produce only standard pronunciation citation forms. It is important, however, to distinguish between the formal description of the competence of an ideal speaker of a language, and the more pragmatic concerns of real listeners who have to cope with widespread variation in word forms.

Recent studies of variations from citation forms due to dialect, speaker differences and connected speech usage have provided a broader role for the

lexicon in accounting for the variable realizations of words. For example, in Nolan's (1982) categorization of segmentally based between-speaker differences, three of the four categories relate to the lexicon. Another recent development has been to examine the "psychological reality" of phonological theories, which has lead to a questioning of the importance of rule-based accounts compared to alternatives such as listing multiple forms in a single lexical entry, or otherwise indicating the range of variation in lexical entries (eg Linell, 1979). One item of evidence here is that both diachronic change and the development of speech in children proceeds by 'lexical diffusion' from one item to another rather than by simultaneous application of an abstract rule throughout the lexicon (Hooper, 1981).

Studies of speech perception do not uniformly rely on normalization at the phonetic level to account for variation. Nooteboom (1981), in particular, has shown how the perception of speech timing can be modeled by a lexicon which includes a wide range of durations for each auditory unit, removing the need for any active normalization process at the cost of increased memory storage. Other models in which words or morphemes are the primary perceptual targets and phoneme identification occurs after lexical access must provide similar mechanisms for normalization at the lexical level. One of the few attempts at this is Klatt's (1979) model, which shows the computational advantages of representing variability due to coarticulation and connected speech usage in the lexicon.

Although there are reasons for increasing the role of the lexicon in accounting for phonological variation, the present hypothesis that speaker characteristics may be stored at the lexical level does not imply that this is the only way of representing such information. Clearly listeners do have knowledge about the phonetic consequences of some forms of variation, such as accent differences (Flege, 1984), and it is also clear that perceptual normalization algorithms exist which can solve at least some of the problems of speaker variance at a low level.

Variation due to accent presents an interesting case because the possibility of listeners using phonological rules to map accented productions to standard forms would appear to be less than with other types of variation, particularly when the accent is strong or unfamiliar. Not only will listeners be unfamiliar with the phonetic and suprasegmental categories of the 'foreign' language, but the speaker may have an incomplete knowledge of the target pronunciation, including coarticulatory and word boundary phenomena, timing, lexical stress and suprasegmental aspects, making the utility of rules somewhat unclear. The outcome of these processes may not be a simple change in the realization of phonetic elements, but a word with different numbers and types of elements. Such changes also occur with non-accented speaker differences, dialect, and connected speech variations, although to a lesser extent.

Storing speaker dependent tokens in the lexicon could reduce the need for complex computation to map accented words to standard forms, in the same way that other forms of variation have been handled in the models described. A parallel solution appears in machine recognition models where insufficient knowledge of the phonological rules relating speakers is accommodated by the use of separate tokens for each word.

ABSTRACT AND LITERAL MODELS OF MEMORY

Early studies of memory suggested a distinction between sensory memory (precategorical acoustic store) and an abstract linguistic long term store

(eg Crowder and Morton, 1969), indicating that literal details were lost within a few seconds. Continuing studies of short term memory for sublexical stimuli such as vowels and syllables support the distinction between auditory and phonetic memory, although the relationship between the experimental methods used and word recognition is questionable (Nooteboom, 1981). Morton's (1969) logogen model is the primary example of an abstract theory of the lexicon. This model is based on data showing that recognition of words in noise can be facilitated by presentation in a sentential context, by increased frequency (in the language), and by repetition during the course of an experiment. The success of the model depends on these effects arising through changes to a single unit. Since both context and frequency do not involve literal details, and since Morton found that repetition of words was insensitive to literal detail, an abstract unit, the logogen, was the most parsimonious explanation.

These views are contradicted by more recent evidence that long term memory does retain such incidental information as modality of presentation, speaker's voice and temporal contextual detail, even when this information is irrelevant to the tasks involved in initial processing of the word (Craik and Kirsner, 1974). These studies use conscious recollection of details of previous encounters with words. Other studies of the categorization of both linguistic and nonlinguistic stimuli are consistent with the storage of instances rather than prototypes (see Jacoby and Brooks, 1984), and show how apparently abstractive behaviour can arise from the way memory is searched and the natural distribution of instances. Objections to instance based models due to their large memory requirements and inefficient organization and retrieval properties have been met by distributed storage models, in which literal memory traces are stored and abstractive information emerges as a result of the storage system. These models have been applied to a variety of memory phenomena and recently to word recognition and the role of the lexicon (see Rumelhart and McClelland, 1985).

REPETITION PRIMING

Much of the development of these abstract and literal views of memory has not directly concerned lexical memory as required in speech recognition. Doubts exist about the relationship between segment and word identification studies, about the strategic nature of conscious memory retrieval, and about the permanency of effects attributed to 'long term' memory when tested in a typical experimental session. A more direct tool for examination of the lexicon is the word repetition effect. A variety of tasks, including those requiring recognition of words in noise and lexical decision, show improved performance for words that subjects have also seen earlier in the experiment. There are a number of features of this effect which encourage interpretation in terms of a long-lasting modification to lexical memory (see Monsell, 1985). For example, it does not normally apply to nonwords, and morphological relatives of a word produce priming while visually or auditorally similar but semantically unrelated words do not, suggesting that the effect is not due to overlap at the segment level. The effect is long lasting and does not seem to depend on conscious retrieval of memory, as it is also found in amnesics who obviously lack 'explicit' memory.

It is known that priming occurs to some extent even when prime and test stimuli are not identical forms. For example, a change from upper to lower case in visual stimuli produces little or no decrease in priming, while presenting the prime visually and the test word auditorally results in a drop to about 50% of the normal priming effect. The general interpretation of these results is that full transfer indicates access to the same lexical

unit, while partial transfer arises from a two level system with incomplete inter-level transfer (Kirsner, Milech and Standen, 1983). If lexical units contain voice-specific detail, when prime and test words are in different voices some reduction of priming compared to identical repetition is expected. The experiment described below tests this prediction.

THE EXPERIMENT

It is not currently clear to what extent modality-specific lexical units are sensitive to incidental details such as speaker's voice. Jackson and Morton (1984) report that a change in speaker produces no loss of priming, but their experiment used a relatively conservative example of voice variation between male and female speakers. Further, there are reports of priming being reduced by changes in other incidental aspects of stimuli, such as size, typefont and orientation of printed words (Jacoby and Brooks, 1984).

To provide a stronger test of the effect of speaker differences, a voice with an unfamiliar and quite marked accent (Spanish) was used. This accent involved significant differences in the realization of phonetic elements, and frequently had a different phonetic composition and lexical stress to the standard Australian pronunciation. This condition was compared with others in which (a) prime and test stimuli were the same token from the test speaker, the standard exact-repetition condition; (b) the two stimuli involved a different token from the same speaker; and (c) two speakers having the same standard pronunciation were used, one male and one female. The latter condition was run as part of another experiment, which in all other respects was identical to that described here.

In the first phase of the experiment subjects were asked to decide whether they were familiar with the words. To check that subjects had a lexical representation for the low frequency words used, items with which they were unfamiliar were removed from analysis of priming data. The test task involved recognition of words presented in a babble distractor. Test words were spoken in a male voice with standard pronunciation. In order to remove individual variation in susceptibility to the noise, and to provide a baseline performance level that reduced the chance of ceiling or floor effects in the recognition task, the 50% recognition threshold was estimated for each subject prior to the test section of the experiment, using a modified maximum liklihood method. An additional benefit of this procedure is in providing a baseline measure that is comparable across experiments varying in subjects and word sets. The threshold estimation procedure and other experimental details are described elsewhere (Standen, Kirsner and Dunn, in preparation).

RESULTS AND DISCUSSION

Exact repetition resulted in an increase in accuracy to 65.4%, compared with the baseline figure of 50% for words not previously encountered. The scores for the other three conditions were very similar, and only slightly less than that for exact repetition. A change of token alone produced 64.2%, the gender change only condition showed 62.9%, and the accent plus gender change resulted in 63.1% accuracy. These figures show that lexical representations are substantially abstract with regard to speaker's voice, even when an unfamiliar accent is involved. There is no evidence that a change in the segmental and suprasegmental specification had any more effect than the change in the realization of segments involved in the male-female variation. It seems likely that similarly high priming would be found for other forms of phonological variation such as those due to connected speech and dialect.

The small loss of priming found for speaker change might be thought to be due to sampling error. However examination of data from the previous voice priming experiment (Jackson and Morton, 1984), experiments varying features of visual words and other voice priming experiments conducted in our laboratory, show that numerically, if not statistically, such an effect is found on the overwhelming number of occasions (Standen et al, in preparation). It therefore seems that there is some very small effect due to a change in surface form.

The present results do not support the claim by proponents of literal models that priming is proportional to the similarity of the two instances (Jacoby and Brooks, 1984; McClelland and Rumelhart, 1985). There is some evidence from studies of priming in visual words that extreme variations from standard forms produce much weaker priming. It may be that even the accented voice used here does not represent the greatest degree of variation as a result of the requirement that words be readily recognized in isolation. In continuous speech, semantic and syntactic information can augment bottom-up analysis and words are routinely identified from only partial phonetic information. However extreme variations may also result in the use of qualitatively different word recognition strategies (Standen et al, in preparation).

Storage of speaker's voice

If we accept that lexical units are largely insensitive to variations in speaker's voice, as indicated by the high degree of priming, token based models of the lexicon loose much of their appeal. The problem of accounting for memory of speaker's voice remains, though. Possibly this can be resolved by empirical examination of Jacoby and Brook's (1984) claim that such effects are "often small and sometimes short lived". Another possibility is that voice is stored in a system independent of the lexicon. It is known that voice information is retained independently of the words involved, particularly if associated with a face (Legge, Grossman and Pieper, 1984), and that retention of voice information is improved when the semantic contents of the message make it salient (Fisher and Cuervo, 1983).

Contents of the lexicon

The present results require lexical units to respond to inputs which cannot be represented by a single phonetic string. This may be conveniently modeled by the use of multiple forms in a lexical entry, as suggested by the perceptual and linguistic models described earlier. However, although the accented words were very different from their standard pronunciation counterparts, the present evidence does not rule out the use of a transfer function which operates at a sublexical stage. Subjects may have used the available acoustic overlap of accented and standard pronunciation, some rapidly acquired phonological rules, and lexical identity constraints such as phonotactic rules to restrict lexical hypotheses. Clearly, since the accented words were correctly recognized, some of these sources of information must have been used. Further evidence is required before the hypothesis of multiple specifications in a lexical entry can be accepted.

In summary, models in which speaker-specific or instance-specific tokens are the basis of long term lexical memory are not supported. Rather the data suggests that lexical units are abstract over even larger variations from standard citation forms than previously suggested. Future research must examine ways in which the lexicon can supplement or replace computationally demanding perceptual normalization processes, and the extent to which phonological variation is handled by units specific to individual words.

REFERENCES

CRAIK, F. I. and KIRSNER, K. (1974). "The effect of speaker's voice on word recognition", Quarterly J. Exp. Psych. 26, 274-284.

CROWDER, R. G. and MORTON, J. (1969). "Precategorical acoustic storage (PAS)", Percept. Psychophys. 5, 365-373.

FLEGE, J. E. (1984). "The detection of French accent by American listeners", J. Acoust. Soc. Am. 76, 692-707.

HOOPER, J. B. (1981). "The empirical determination of phonological representation", in "The Cognitive Representation of Speech", edited by T. Myers, J. Laver and J. Anderson (North Holland: Amsterdam), pp 347-359.

JACKSON, A., and MORTON, J. (1984). "Facilitation of auditory word recognition", Mem. and Cognit. 12, 568-578.

JACOBY, L. L. & BROOKS, L. R. (1984). "Nonanalytic cognition: memory, perception and concept learning", in "The Psychology of Learning and Motivation, Volume 18", edited by G. Bower (Academic: London), pp 1-47.

KIRSNER, K., MILECH, D. and STANDEN, P. (1983). "Common and modality-specific processes in the mental lexicon", Mem. and Cognit. 11, 621-630.

KLATT, D. H. (1979). "Speech perception: a model of acoustic-phonetic analysis and lexical access", J. Phonet. 7, 279-312.

LEGGE, G. E., GROSMANN, C. and PIEPER, C. M. (1984). "Learning unfamiliar voices", J. Exp. Psychol.: Learning Memory Cognition 10, 298-303.

LINELL, P. (1979). "Psychological Reality in Phonology" (Cambridge University Press: Cambridge).

McCLELLAND, J. L. and RUMELHART, D. E. (1985). "Distributed memory and the representation of general and specific information", J. Exp. Psychol.: General 114, 159-188.

MONSELL, S. (1984). "Repetition and the lexicon", in "Progress in the Psychology of Language, Vol 1" edited by A. W. Ellis (Erlbaum: London).

MORTON, J. (1969). "Interaction of information in word recognition", Psychol. Rev. 76, 2, 165-178.

NOLAN, F. (1983). "The Phonetic Bases of Speaker Recognition", (Cambridge University Press: Cambridge).

NOOTEBOOM, S. G. (1981). "Speech rate and segmental perception or the role of words in phoneme identification", in "The Cognitive Representation of Speech", edited by T. Myers, J. Laver and J. Anderson (North Holland: Amsterdam), pp 143-150.

PISONI, D. B. (1985). "Speech perception: some new directions in research and theory", J. Acoust. Soc. Am. 78, 381-388.

STANDEN, P., KIRSNER, K. and DUNN, J. (manuscript in preparation). "Surface form in word recognition".