Perceiving Speech and Perceiving Words

W.D. Marslen-Wilson
Nijmegen, the Netherlands

1.

Psycholinguistic research into spoken language comprehension, and phonetic research into the processes of acoustic-phonetic analysis, are both, in principle, part of the same general domain of inquiry. Both disciplines are concerned with aspects of the process whereby human listeners map from sound onto meaning. This implies, therefore, a close dependence between them.

In the past, however, there has been surprisingly little direct contact between the two disciplines. Research in phonetics – as, for example, Nooteboom (1979) has documented – tends to pay little attention to the wider functional context within which the processes of acoustic-phonetic analysis presumably operate. Conversely, psycholinguists – even those working on spoken word-recognition – tend to neglect, or simply ignore, the complexities of the acoustic-phonetic input to the processes they are studying.

We can take for granted that psycholinguists should pay more attention to acoustic-phonetic issues. What is less straightforward is the claim that phoneticians should pay more attention to psycholinguistic issues. Nonetheless, this is what I will try to establish here. I will do so with particular reference to the relationship between the acoustic-phonetic analysis of the speech signal and the perception and identification of spoken words.

Two questions need to be examined here. First, how far does the study of spoken word-recognition also raise important acoustic-phonetic questions? Second, how far has research in acoustic-phonetics in fact provided an adequate basis for an approach to these questions?

2.

The first point to be made concerns the extent to which further progress in understanding spoken word-recognition depends on developments in acoustic-phonetics. In the past, research on spoken word-recognition has been so general in the kinds of claims it made about the recognition process that it was not necessary to pay close attention to the acoustic-phonetic substrate for this process. It did not really matter what the imput to the word-recognition process was since the issue never really arose of how individual spoken

words were discriminated from each other (although this question certainly did arise very early on in research on machine recognition of fluent speech). Recent research, however, has led to the development of psycholinguistic theories of spoken word-recognition that do require a much more precise specification of the properties of speech analysis.

These developments arise from some observations of the rapidity and the immediacy with which the speech signal is mapped onto the mental lexicon (c.f. Cole and Jakimik, 1980; Grosjean, 1980; Marslen-Wilson 1975; 1980; 1983; Marslen-Wilson and Tyler, 1975; 1980). A wide variety of different experiments converge on a highly consistent estimate of the average "recognition-time" for words heard in a normal utterance and discourse context – where the term "recognition-time" refers to the amount of sensory input, measuring from word-onset, that needs to be heard before a listener can start behaving as if he or she has correctly identified the word in question. The estimate of this average recognition-time for words in context is of the order of 200 msec.

Not only is this remarkably fast, but also it is remarkably *early*, relative to the total duration of the words being identified. For the kinds of experiments involved, the words averaged 375-420 msec in length. This means that words in context can reliably be identified when little more than half of the acoustic input corresponding to that word in the signal could have been heard. This in turn implies that listeners are highly efficient in their use of the acoustic-phonetic information carried by the speech signal. More recent results (Marslen-Wilson, 1983) show that listeners are in fact *optimally* efficient in their use of this information.

The notion of optimal efficiency can, in principle, be defined as the extraction of the maximum information-value from the signal, in real-time as it is heard. The term "information-value" can itself be related to the definition of information in terms of the number of alternatives between which a given signal can allow a receiver to discriminate (Shannon and Weaver 1949). If we assume some set of possible messages that a given signal can transmit, to a given listener in a given context, then the speech signal can be viewed as providing a continuous flow of potential discriminative information with respect to this set of possibilities.

If the set of possibilities involved is the complete set of words in the language, known to a given listener, then the information-value of the signal is defined with respect to the information that the signal provides, over time for the discrimination of the correct word from among the initial total set of alternatives. Experiments using an auditory lexical decision task show that listeners are indeed able to identify the word being uttered at precisely that point in the word at which the theoretically sufficient acoustic-phonetic information becomes available (Marslen-Wilson, 1983; see also Tyler and Wessels, 1983).

These results, and other considerations, lead to a model of spoken wordrecognition in which there is a multiple accessing of possible word-candidates

early in the word. The subsequent selection of the correct candidate depends on the manner in which the accumulating sensory input not only matches the specifications (in the mental lexicon) of the correct word, but also fails to match the specifications of the incorrect words. The recognition of the correct word becomes possible, as experimentally demonstrated, as soon as the signal diverges sufficiently from the specifications of all other possible words.

An approach of this kind therefore stresses the implications for the identification of individual words of the discriminative information accumulating as the signal is heard. It is clear that the evaluation and development of such an approach depends on a satisfactory analysis of the nature of the input to these word-discrimination processes. Under what description are the products of acoustic-phonetic analysis delivered to the word-recognition system? What aspects of the original signal are preserved or discarded in the process of analysis? With respect to which set of discriminative categories should the information-value of the signal be evaluated?

3.

If, for an answer to these questions, we now turn to the main body of acoustic-phonetic research, we do not receive a coherent answer. One is faced with a remarkable diversity of different and incompletely specified proposals, where the products of speech analysis range from strings of phonemic labels, to bundles of probabilitically weighted features, to direct perceptions of speech events.

At least one distinguished acoustic-phonetician, confronted with these difficulties, has concluded that the best approach to the question of how the signal is mapped onto lexical representations is, in effect, to renounce the whole framework of classical phonetics (Klatt, 1979; 1980). Instead one should opt for the kind of "brute force" computational solution, based on direct matching to spectral templates without any intervening phonetic analysis, that is exemplified in the **harpy** speech recognition system (Lowerre and Reddy, 1978). It is likely that this conclusion is too pessimistic. Nonetheless, it is clear that acoustic-phonetic research, for all its advances over the past thirty years, has failed to satisfactorily answer those questions that are most critical for researchers working on other aspects of language processing. In part this is no doubt due to the fact that acoustic-phonetics, just like any other branch of the study of human language, is extremely difficult; that it can't be expected to have found all the answers yet. But in part it may also be the consequence of the set of assumptions that permit, and even encourage, the current *de facto* separation between research on speech analysis and research on spoken word-recognition.

The most important of these assumptions seem to be the following. First, one must assume that there are two distinct levels of perceptual representation computed during speech analysis. These correspond, respectively, to an

acoustic-phonetic level of analysis and to a lexical level. Secondly, and crucially, one must assume that the properties of the acoustic-phonetic level, and of the processes that map from the speech signal onto this level, can be determined solely with reference to phenomena internal to this level, and without reference to the functional goal of these processes. Without reference, that is, to the role of these processes in providing the basis for a further mapping onto the mental lexicon (which in turn provides the basis for the extraction of communicative meaning).

Thirdly, one has to accept the direct translatability of results obtained in the phonetics laboratory, typically using either citation forms of synthetic speech, to the perceptual situation of the listener hearing fluent conversational speech. That is, one must assume that the kinds of relationships observed in the laboratory between a given speech signal and a given phonetic contrast, will also hold in the often different conditions of normal speech production and comprehension.

4.

It is not possible to state categorically that these assumptions are either false or misleading. But they are at least open to serious question. Consider, in particular, the second assumption, that speech analysis is most appropriately studied in functional isolation. In the case of spoken word-recognition, for example, one finds that it is by studying word-recognition in its functional context – as it contributes to the processes of language comprehension in utterances and discourses – that one can place the strongest constraints on possible models of lexical access (Marslen-Wilson, 1983; Marslen-Wilson and Welsh, 1978). In the same way, it may be that by examining the processes of speech analysis in their proper functional context – as part of the process of speech understanding – that one can place constraints on theories of speech analysis that could not be derived just by attempting to study these processes in isolation.

If, for example, as current analyses of spoken word-recognition suggest, one can predict precisely when a given word should become discriminable, then it should also be possible to determine just which aspects of the sensory signal are employed in making these discriminations. This, in turn, would surely have implications for one's assumptions about the speech analysis process that produces the basis for these effects.

Whether or not this particular strategy turns out to be fruitful remains to be seen (but see Streeter and Nigrom 1979). But the general point remains. Many psycholinguistic questions about the processes of spoken word-recognition are inescapably acoustic-phonetic questions as well. And it seems most unlikely that these questions can be resolved without a proper contact between the two disciplines – both in theoretical analysis and in experimental practice.

References

Cole, R.A., and Jakimik, J. (1980). A model of speech perception. In: R.A. Cole (Ed.), *Perception and Production of Fluent Speech.* Hillsdale, NJ: LEA.

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

Klatt, D.H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics,* **2**, 279-312.

Lowerre, B.T., and Reddy, D.R. (1978). The Harpy speech understanding system. In: W.E. Lea (Ed.) *Trends in Speech Recognition.* Englewood Cliffs, NJ: Prentice-Hall.

Marslen-Wilson, W.D. (1975). Sentence perception as an interactive parallel process. *Science,* **189**, 226-228.

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

Marslen-Wilson, W.D. (1983). Function and process in spoken word-recognition. In: H. Bouma and D.G. Bouwhuis (Eds.), *Attention and Performance X.* Hillsdale, NJ: LEA.

Marslen-Wilson, W.D., and Tyler, L.K. (1975). Processing structure of sentence perception. *Nature,* 1975, **257**, 784-786.

Marslen-Wilson, W.D., and Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition,* **8** 1-71.

Marslen-Wilson, W.D., and Welsh, A. (1978). Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology,* **10**, 29-63.

Nooteboom, S.G. (1979). More attention for words in speech communication research? In: B. Lindblom and S. Ohman (Eds.), *Frontiers of Speech Communication Research.* London: Academic Press.

Shannon, C.E., and Weaver, W. (1949). *The Mathematical Theory of Communication.* Urbana: University of Illinois Press.

Streeter, L.A., and Nigro, G.N. (1979). The role of medial consonant transitions in word perception. *Journal of the Acoustical Society of America,* **65**, 1533-1541.

Tyler, L.K., and Wessels, J. (1983). Quantifying contextual contributions to word recognition processes. Manuscript, MPI for Psycholinguistics, Nijmegen.