THE INTERFACE BETWEEN ACOUSTIC-PHONETIC AND LEXICAL PROCESSES

William D. Marslen-Wilson Uli H.Frauenfelder

Max-Planck Institute for Psycholinguistics,
Nijmegen, The Netherlands

Medical Research Council Applied Psychology Unit,
Cambridge, England

Abstract

Speech research and psycholingustic research into spoken language comprehension have a common interest in the processes of acoustic-phonetic analysis. This paper argues that this common goal should be reflected in a common research programme, integrating together the questions and the techniques of the two disciplines. Without such an integration, neither discipline can expect to achieve adequate answers to its characteristic questions.

Introduction

A fundamental goal of the phonetic sciences is to characterise the ways in which the acoustic signal is mapped onto an acoustic-phonetic level of mental representation. This objective is subsumed within one of the major goals of experimental psycholinguistics -- namely, to characterise the mapping from the speech signal onto a level of meaning representation. But despite this intimate inclusion relation, there has been surprisingly little direct contact between the two disciplines. Research in phonetics has paid only limited attention to the wider context within which the processes of acoustic-phonetic analysis presumably operate. By the same token, psycholinguistic research into spoken language comprehension has tended to neglect the complexities of the acoustic-phonetic input and its analysis.

This bidirectional indifference is doubly surprising when we consider just how strong the interdependence must be between the two disciplines at the point where their interests directly converge: that is, at the interface between acoustic-phonetic and lexical processes. From the phonetic perspective, the emphasis, naturally enough, is on the computation of acoustic-phonetic representations from the speech input. From the psycholinguistic perspective, the extraction of meaning depends upon access to the mental lexicon, and this in turn depends upon the ability of the system to map the speech input onto mental representations of lexical form via an acoustic- phonetic representation.

In other words, both disciplines are closely concerned with one and the same representation -- what we label here as the input representation -- mediating between the speech signal and the mental representations of lexical form. The goal of this paper is to demonstrate that it is both necessary and possible to investigate the properties of this representation from both perspectives. We want to show, on the one hand, how the input representation, and the processes mapping it onto the lexicon, are constrained by the signal and its acoustic-phonetic analysis, and, on the other, how the input representation and its construction are influenced and constrained by the target lexical representations, and by the properties of the language in general.

Research in our laboratories over the past three years has been guided by this dual perspective, aiming at the development of a unified picture of the early stages of the speech understanding process. The following sections give an overview of some of this research. In the first part, we will focus on some ways in which the properties of the speech signal constrain the input representation and lexical access, and on the consequences of this for the theoretical assumptions that have been used to justify the separation of acoustic-phonetic issues from the lexical level. In the second, we will discuss research into the processing structure of the

interface, focussing on the directionality of information flow within the system. In the concluding section of the paper we will turn to some research into the role of the listener's system of phonological knowledge in mediating the relationship between the signal and the lexicon, and the consequences of this for the input representation.

The speech signal and lexical access

The conventional division between psycholinguistic research into spoken word recognition and phonetic research into speech analysis is based on the assumption that lexical access is largely insulated from the detailed properties of the speech signal and the way it carries information over time. This assumption in turn depends on a number of further assumptions about the properties of the speech processing system. The most important of these -- as we argued here four years ago (9) -- 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, 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 role of these processes in providing the basis for a further mapping onto the mental lexicon. Thirdly -- and most crucial for the psycholinguistic neglect of the speech signal -- there is the assumption that the representation generated at the acoustic-phonetic level (the input representation) is highly abstracted from the detailed properties of the input to the acoustic-phonetic processor. In fact, psycholinguistic research into lexical access has standardly been conducted on the assumption that the input to the lexicon is a string of phonemic labels, and, indeed, that this is also an adequate characterisation of the properties of lexical form representations.

In this part of the paper we will argue that this cluster of assumptions is false. The detailed properties of the speech signal, and of the way it carries discriminating information over time, are tracked faithfully and continuously at the lexical level. The psycholinguistic problems of lexical access and selection cannot be isolated from the problems of acoustic-phonetic analysis.

The salient feature of the speech signal, considered as an information channel, is that it is based on a continuous sequence of articulatory gestures, which result in a continuous modulation of the signal. Cues to any individual phonetic segment are distributed across time, and, in particular, they overlap with cues to adjacent segments. This means that the speech signal is rich in what we can call **partial information** -- that is, anticipatory cues to the identity of an upcoming segment. As the listener hears one segment, he will also hear partial cues to the identity of the next.

An example of this is the presence of cues to the place of articulation of a word-final plosive in the formant structure of the preceding vowel. Thus, in the word **scoop**, the lips may move towards closure for /p/ during the vowel, while in **scoot** the tip and body of the tongue are brought forward to form closure for the /t/. Both movements, conditioned by the place feature of the consonants, produce differences in the formant frequency patterns towards the end of the vowel.

The question we have asked in recent research is whether this type of partial information is made available at the lexical level. How far is the on-line process of lexical access and selection sensitive to the continuous nature of information transmission in the speech signal, and to the availability of partial information as it accumulates over time? To the extent that such sensitivity can be demonstrated, then the separatist assumptions we listed above seem to fail. If word candidates can be accessed and identified on the basis of partial information about the identity of a sound segment, then this causes fundamental problems for the claim that the speech input is mapped onto representations of word-forms in the mental lexicon in terms of complete phonemes (or units of a similar or larger size)

We have investigated this question in a number of studies,carried out in English, Bengali, and Dutch (4, 7, 11, 12), which have used the speech gating task to trace the temporal microstructure of acoustic-phonetic uptake during spoken word-recognition. We focus here on the English studies, looking at the uptake at the lexical level of partial cues to word-final place and voice in CVC monosyllables (11,12).

These were experiments in which listeners heard gated fragments of CVC's, drawn from pairs contrasting in place (like **scoop/scoot**) or in voice (like **log/lock**). The words were presented in increments of 25 msec, focussing on the 125 milliseconds leading up to the closure of the vowel. Gate 0 in Figure 1 represents the gate at which the vowel terminated. The subjects were required at each increment to say what they thought the word was, or was going to become.

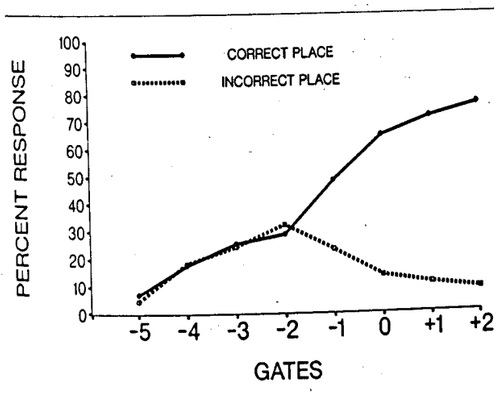

*Figure 1: Lexical responses to pairs of CVC's contrasting in place. The correct responses are cases where the subjects respond with the member of the pair with the correct place (e.g., **scoop**); incorrect responses are cases where they respond with other member of the pair (e.g., **scoot**).*

For the place contrasts, which here involved CVC's ending in voiceless plosives and matched for frequency, partial information as to place of articulation is conveyed by the changing spectral properties of the vowel as it approaches closure. The question at issue was whether this would affect lexical access and selection, as reflected in the subjects' responses at each gate. If so, then their responses should start to diverge before vowel closure (i.e., before Gate 0), and certainly before they hear the plosive release, falling some 80-100 msec after closure. The results, summarised in Figure 1, clearly show this early divergence, with a strong preference at Gate 0 for the word with the correct place of articulation.

For the voicing contrasts (involving pairs like **rip/rib** and **dog/dock**) we were asking similar questions, but looking now at a durational cue -- vowel length is a powerful cue to voicing in English. In the gating task, listeners hear the vowel slowly increasing in length over successive gates. Our question was whether they could exploit this information as it became available.

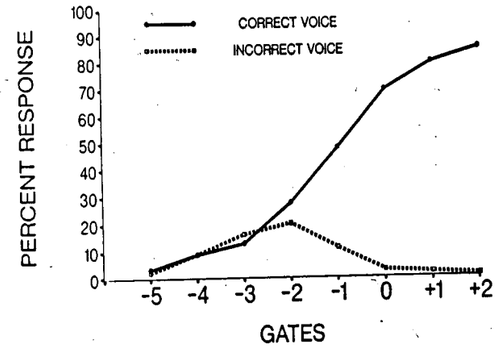

*Figure 2: Lexical responses to pairs of CVC's contrasting in voice. The correct responses are cases where the subjects respond with the member of the pair with the correct voice (e.g., **dock**); incorrect responses are cases where they respond with other member of the pair (e.g., **dog**).*

Turning to Figure 2, where Gate 0 again represents vowel closure, we see that after an initial period in which voiceless responses predominate, listeners start to successfully discriminate voiced from voiceless words as soon as the length of the vowel starts to exceed the durational criterion (around

135 msec from vowel onset for this particular stimulus set).

What we find, then, in Figures 1 and 2, is clear evidence for the immediate uptake of accumulating acoustic information. There do not appear to be any discontinuities in the projection of the speech input onto the lexical level. The speech signal is continuously modulated as the utterance is produced, and this continuous modulation is faithfully tracked by the processes responsible for lexical access and selection. As the spectrum of a vowel starts to shifts towards the place of articulation of a subsequent consonant, this is reflected in a shift in listeners' lexical choices, which becomes apparent about 25-50 msec before closure. As the duration of a vowel increases, the listener produces lexical choices that reflect these changes in duration, shifting from voiceless to voiced as the durational criterion is reached and surpassed, at about 50-75 msec before closure. There is immediate use of partial durational cues, just as there is immediate use of partial spectral cues.

Lexical processing is clearly not insulated from the detailed properties of the speech signal. Psycholinguists interested in the temporal structure of the speech understanding process cannot ignore the variations in information flow that stem from the continuous modulation of the incoming speech signal. By the same token, research into the properties of acoustic-phonetic processing will have to acknowledge the direct relevance of its subject-matter for processes at the lexical level.

The processing structure of the interface

We turn now a different perspective on the properties of the interface between acoustic-phonetic and lexical processes. This involves the structure of the interface viewed as an information-processing system. If we are talking about different levels of analysis during speech processing, and discussing the flow of information between these levels, then we need to specify the directionality of this flow, and to determine the contraints on how information at one level can affect processes at another level. Is information-flow strictly "bottom-up", in the sense that information flows in one direction only, from the speech signal, via an acoustic-phonetic processor, up to the lexical level. Or does the system allow for "top-down" effects as well, in the sense that information at a higher level can feed back to lower levels and directly affect the outcome of these lower level processes.

We have seen in the preceding section how the time-course of lexical access and selection is determined by the properties of the signal and its on-line acoustic-phonetic analysis. Information originating in the sensory input flows continuously in a bottom-up fashion to drive lexical processing. The further results of these gating studies (11, 12) looking at the effects of at least one lexical variable (the frequency of occurrence of the word being heard), suggest that the bottom-up (sensory) input has the priority in determining the outcome of lexical access and selection. Although word frequency did have an effect, with subjects initially tending to respond with more frequent words (for pairs of words where we explicitly contrasted frequency), the scope of these effects was severely limited. In particular, frequency only affected lexical processing under conditions where the available sensory information was sufficiently ambiguous or indeterminate to allow a choice between one or more alternatives. But these effects dissipate immediately as soon as more determinate bottom-up information became available.

This processing asymmetry between the importance of bottom-up and top-down processes is in apparent conflict with a strong current trend to assign an important role to top-down information-flow during speech processing. On this type of account, decisions about the content of the sensory input (i.e. the identity of segments) are affected by expectations coming from higher levels of processing. In effect, the perceptual output of the mechanisms of speech perception are assumed to vary as a function of the lexical context in which the speech input occurs.

Evidence for top-down information flow comes from a variety of sources, including studies of phonetic categorization and phoneme restoration.

Phonetic categorization (for example, the identification of stimuli falling at different points along some acoustic-phonetic continuum) has been claimed to be influenced by the lexical status of the item bearing the phonetic segment. In one such study (6) the voice onset time (VOT) of stimulus initial stop consonants was manipulated to construct a voicing continuum. These stimuli were chosen such that the lexical status of each stimulus changed from a word to a non-word, as in **dash/tash**, as a function of the voiced or unvoiced character of this initial phoneme. Subjects were asked to make a forced phonetic choice (i.e. between **/d/** and **/t/**). It was found that the lexical status of the item led to a shift in the location of the phoneme boundary along the VOT continuum, in the direction of word rather than non-word responses. This result was interpreted as showing that the perception of the identical speech sound can differ depending on its status at the lexical level -- whether it forms a word or a non-word.

The phenomenon of phoneme restoration has also been taken as evidence for a contribution of the lexical level to phonetic processing. Listeners typically report that an utterance sounds intact even when a part of it has been replaced by an extraneous noise. According to Warren (13), this ability to restore the missing speech sound shows that the perception of speech is mediated by higher levels, via a top-down processing link.

A major weakness in this type of evidence for top-down perceptual processes is that it ignores the temporal properties of the proposed top-down information flow. In fact, it is critical to determine when top-down information first becomes available to influence processing, and whether this influence operates quickly enough to affect the continuous and immediate bottom-up analysis. The importance of temporal variables in controlling information flow in lexical processing can be seen in a more recent study (2) of phonetic categorization, where time constraints were introduced. When subjects were required to make speeded phonetic decisions, a lexical effect was found only for slow responses, but disappeared for fast ones. This suggests that a certain amount of time, or more precisely - a substantial amount of acoustic information - needs to accumulate before the lexicon can exert its influence on the bottom-up analysis. The critical question arises whether this top-down influence comes into play before the bottom-up analysis has been completed.

We have conducted a number of experiments investigating the temporal properties of these proposed top-down lexical effects, in order to determine whether lexical constraints do in fact influence on-line processes at lower levels of analysis. We will present here a sample of this research (for more details, see Frauenfelder, Segui, and Dijkstra, this volume).

To trace the time-course of lexical effects, we selected words containing phoneme targets to be detected in different positions in the words. These targets occupied four different positions with respect to the words' uniqueness points (word onset, before uniqueness point, after uniqueness point, word offset) -- for an example set, see Table 1.

Table 1

Examples of Monitoring Stimuli

	WORD	NONWORD
ITEM ONSET	**Pagina**	**Pafima**
BEFORE UP	**jaPanner**	**joPammel**
AFTER UP	**olymPiade**	**arimPiako**
ITEM OFFSET	**bioscooP**	**deoftooP**

The Uniqueness Point (UP) is defined as the point at which a spoken word becomes uniquely identifiable, going from word-onset. Nonwords were created by changing one or more segments in the original word, while keeping the target's local phonetic environment as constant as possible. The dependent variable that we measured was the subject's latency to detect a previously specified phoneme target. The difference in these

detection latencies to phoneme targets in the same position in matched words and nonwords was taken to provide a measure of the lexical contribution to the phoneme detection process. Figure 3 shows these differences between words and nonwords as a function of target position.

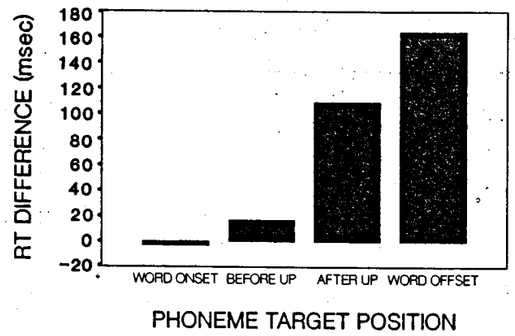

Figure 3: Mean differences between targets in words and non-words as a function of target position.

There are significant differences between words and nonwords, but only after the UP. This suggests that the lexicon exerts its effect on phoneme detection responses only at a point in processing where just a single candidate is still compatible with the sensory input. Listeners have already accessed full lexical information and have available the entire phonological description of the word (10). This severely limits the potential role of top-down lexical effects in speech processing. If there is a top-down influence on lower level processes -- and there has been some dispute as to whether phoneme detection tasks tap a sub-phonemic lexical level at all (e.g., 1)-- then these effects seem to come into play only after the bulk of the bottom-up processing has been completed. This defeats the potential processing function of top-down effects in theories like TRACE (8), where top-down information-flow to the phoneme level has the effect of tuning the responses of phoneme nodes as a function of their lexical environment.

Phonological aspects of the interface

We have spoken so far about the relationship between the acoustic-phonetic analysis of the speech signal, and the processes of form-based lexical access and selection. We now turn briefly to the potential role of phonological factors in determining the character of this interface.

Research in acoustic-phonetics and in lexical access typically assumes a fully transparent relationship between the signal and the lexicon - that the speech signal makes information available and that this information is straightforwardly mapped onto representations of lexical form. In fact, the listener's system of phonological knowledge may mediate this relationship (3,5), with consequences for the interpretation of the signal which are not deducible from the properties of the signal alone.

We see this, for example, in recent research that reveals phonologically based assymetries in the lexical interpretation of cues in the speech signal. These are studies looking at the perceptual consequences of vowel nasalisation (4,7,11), contrasting languages like Bengali, where nasal is distinctive for vowels, with languages like English, where it is not.

For English listeners, the presence of nasalisation in a vowel is an unambiguous signal that they are hearing an oral consonant followed by a nasal vowel. But for Bengali listeners, where phonetically equivalent vowel nasalisation holds both for nasal vowels preceding oral consonants and for oral vowels preceding nasal consonants, the presence of nasalisation is ambiguous. What one sees, however, in a gating task carried out with Bengali listeners (7), is a very strong bias to interpret nasality as signalling the underlying marked value [+nasal] for the language. They interpret nasalisation as signalling the presence of a nasal vowel followed by an oral consonant -- the exact opposite of the interpretation of the same acoustic feature in a language like English. This choice of the Bengali listeners is only explicable if one takes into account the structure of their phonological systems. It is not explicable just in terms either of the signal, or of the representations of lexical form, taken on their own.

A different kind of asymmetry in the interpretation of nasalisation (7,11,12), is a difference in the signal value of the presence as opposed to the absence of nasalisation. When a vowel is nasalised in English, this has a strong effect on lexical choice, ruling out word-candidates where oral vowels are followed by oral consonants. The absence of nasalisation, in contrast, seems to have weaker effects, and does not prevent listeners from selecting CVC's ending with nasal consonants.

This asymmetry may reflect the status of the nasal feature for vowels in English. Because English has no nasal vowels, it is likely that the abstract specification of English vowels does not include the feature [nasal]. This means that when an unnasalised vowel is being heard, there is nothing in the abstract representation of lexical items ending in nasals that could exclude these as possible responses. If a vowel has no nasality feature, then the absence of nasalisation cannot be a discriminant property of the input. In contrast, when the vowel is nasalised, this is a positive cue to the status of the following consonant, and is treated as such by the listener.

These are only preliminary investigations of some asymmetries in the interpretation of acoustic cues at the lexical level. But if we are correct in suggesting that the formal properties of phonological representations can help determine the lexical interpretation of the speech input, then this has important implications for how we should investigate the properties of the acoustic-phonetic processing space within which the listener accesses the mental lexicon. We are arguing, then, not just for an integration of the questions and the techniques of speech research and of psycholinguistics, in studying the interface between acoustic-phonetic and lexical processing, but also for the full engagement in this enterprise of the associated linguistic disciplines.

REFERENCES

1) Cutler, A., Mehler, J., Norris, D. & Segui, J. Phoneme identification and the lexicon. *Cognitive Psychology*, 1987, (in press)

2) Fox, R.A. Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 1984, 10, 526-540.

3) Frauenfelder U. H. & Lahiri, A. Understanding words and word recognition: Can phonology help? Paper presented at the MPI Conference on lexical processing and representation, June, 1986.

4) Frauenfelder, U.H., Wessels, S., & Marcus, S.M. Manuscript in preparation, Max-Planck Institute for Psycholinguistics, Nijmegen.

5) Frazier, L. Structure in auditory word-recognition. *Cognition*, in press, 1987.

6) Ganong, W.F. III. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 1980, 6, 110-125.

7) Lahiri, A., & Marslen-Wilson, W.D. Manuscript in preparation, Max-Planck Institute for Psycholinguistics, Nijmegen.

8) McClelland, J.L. & Elman, J.L. The TRACE model of speech perception. *Cognitive Psychology*, 1986, 18, 1-86.

9) Marslen-Wilson, W.D. Perceiving speech and perceiving words. In M.P.R. v. d. Broecke & A. Cohen (Eds.), *Proceedings of the Tenth International Congress of Phonetic Sciences*. Dordrecht: Foris, 1984.

10) Marslen-Wilson, W.D. Function and process in spoken word recognition. In H. Bouma & D.G. Bouwhuis (Eds.) , *Attention and Performance X: Control of Language Processes*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1984.

11) Warren, P., & Marslen-Wilson, W.D. Continuous uptake of acoustic cues in spoken word-recognition. *Perception and Psychophysics*, in press, 1987 (a)

12) Warren, P., & Marslen-Wilson, W.D. Cues to lexical choice: Discriminating place and voice. Manuscript, Department of Experimental Psychology, University of Cambridge, 1987 (b)

13) Warren, R.M. Perceptual restoration of missing speech sounds. *Science*, 1970, 167, 392-393.