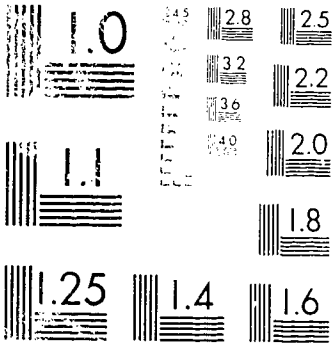

DOCUMENT RESUME

ED 168 304 FL 010 114

AUTHOR Cowart, Wayne
TITLE Fusions as a Source of Information on Higher-Order Influences Upon Speech Perception. CUNYForum, No. 4, 1978.
INSTITUTION City Univ. of New York, N.Y. Graduate School and Univ. Center. Program in Linguistics.
SPONS AGENCY National Institutes of Health (DHEW), Bethesda, Md.
PUB DATE 78
GRANT NIH-71-2420
NOTE 46p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Auditory Perception; *Linguistic Theory; *Perception; *Phonemics; *Phonology; *Psycholinguistics; Semantics; *Speech; Syntax; Visual Perception
IDENTIFIERS *Perceptual Fusions

ABSTRACT

This paper suggests that some features of the syntactic and semantic structure of sentences sometimes influence the phonemic analyses assigned to stretches of speech by the perceptual system. It is argued that the role of higher-order levels of linguistic analysis in speech perception can be productively studied. Theoretical issues appropriate for such study are outlined; the discussion centers on fusion phenomena occurring in speech perception, i.e. when the simultaneous presentation of two or more disparate stimuli causes an observer to perceive some third stimulus different from either of those actually presented. Pilot research confirming the existence of phonological process fusions is reported; these are suggested by existing fusion phenomena, but have not been previously observed. Among the general conclusions is the view that the speech perception system can exploit specifically phonological knowledge and that lexical knowledge appears to play some role in speech perception. Tabular material is appended. (EJS)

FUSIONS AS A SOURCE OF INFORMATION ON HIGHER-ORDER INFLUENCES UPON SPEECH PERCEPTION

Wayne Cowart
Queens College

1. Perceptual Fusions in Speech

Several theorists have suggested that various 'higher-order' levels of linguistic analysis play some role in speech perception.[1] This is meant to suggest, for example, that there are some features of the syntactic and semantic structure of sentences which sometimes influence the phonemic analyses assigned to stretches of speech by the perceptual system.[2]

Despite this theoretical recognition, there has been little direct investigation of higher-order influences. Most speech researchers seem to assume that the problem of learning something about the relation between the auditory and phonetic levels of description has sufficient logical and practical priority to make an investigation of higher-order influences inappropriate at this time.[3]

In this paper I will argue that the role of higher-order influences can be productively studied and I will try to outline some of the theoretical issues to which such studies should be relevant. The phenomena with which I will be mainly concerned are called "fusions."

This research was supported in part by a grant from the National Institutes of Health (NIH-71-2420) awarded to Haskins Laboratories, New Haven, Conn. Valuable assistance was provided at various stages by Drs. Helen S. Cairns and Michael Studdert-Kennedy of the City University of New York and Dr. T. Halwes of Haskins. Credit for errors and lapses of good sense belongs entirely to the author.

Perceptual fusions occur when the simultaneous presentation of two or more disparate stimuli causes an observer to perceive some third stimulus different from either of those actually presented. Fusions contrast with perceptual rivalries in which one stimulus appears to block the perception of the other. A considerable variety of fusion phenomena are known in speech perception, acoustic perception, and visual perception. Among these the most remarkable was one of the first to be reported. In 1877, A.L. Austin noticed that when two different photographs of the same person were viewed through a stereoscope, the observer typically perceives a third view of the subject which differs from either of the two views presented.[4] If one photograph shows a smiling expression and the second a frown, the perceived face bears an expression intermediate to these two. Even more remarkably, the photographs may be of two different people, a man and a woman, without destroying the integrated percept of a single face with features and an expression different than those in either photograph. Evidently, the visual system possesses some notion of a "possible human face" and within limits, construes incoherent face-like stimuli as a single possible face. It is difficult to imagine how such phenomena could arise without the involvement of some level of analysis more abstract than that which relates visual stimuli to the perceived colors, shapes and textures which go to make up the image of a face. Similarly, many of the fusions which occur in speech seem to require the involvement of a higher level of analysis than that which relates acoustic stimuli to segmental percepts.

In the next section I will try to state some of the theoretical questions to which I believe speech fusions may be relevant. Following this is a discussion of some published work on what have been termed "phonological fusions" and some suggestions as to what additional phenomena might be sought. Part 2 of the paper reports some pilot research which tentatively confirms the existence of fusion phenomena which are suggested by those that are already known but which have not previously been observed. Section 3 of the paper draws together some tentative conclusions and tries to point up some implications of fusions in speech perception.

1.1 Some Theoretical Questions About Speech Perception

Let us suppose that some of a speaker's knowledge of his language can be separated into knowledge about two or more distinct levels in a hierarchy such as (1).

(1)
$$\begin{array}{ll} \text{Level}_n & (L_n) \\ \vdots & \\ \text{Level}_1 & (L_1) \end{array}$$

Leaving aside problems of definition, we may take it for granted that there is sound motivation for positing various levels and that there is some reasonable sense in which some levels are said to be higher than others.[5] Given some framework such as this, there are four kinds of questions that naturally arise about the perceptual components of a theory of language performance. I will state these questions in general terms since the special cases of them that pertain to speech perception should be obvious enough.

1) Suppose that we want to understand how the language perception/comprehension system constructs representations of utterances at some level L_i. In general, when psycholinguists confront such questions their attention fastens on lower levels of representation which are presumed to mediate between L_i and the input form of the utterance. There are excellent reasons for this; it is not easy to see, for instance, how syntactic analyses can get far enough ahead of phonological analyses to exert much influence on the lower level analysis. Nonetheless, for any level L_i and some higher level L_j, it is necessary to ask what role, if any, analyses at L_j play in the determination of analyses at level L_i.[6]

2) Suppose we show that some higher level L_j does influence analyses at L_i; we must also ask which facts about L_j play a role. For instance, the phrase to please in the sentence John is easy to please seems to have the syntactic character of a verb phrase. But there is little reason to suspect that this kind of property has much significance for perceptual phonological analysis in such sentences. By contrast, suppose that the sentence A foul odor drifted out of the crypt, were produced in such a way that all parts were distinct except for crypt. Further suppose that the production of crypt could serve equally well as clipped in The gardener clipped the hedge. In this circumstance it seems reasonable enough to suggest that the syntactic role of /k(?)Ipt/ might determine whether the ambiguous segment is heard as /r/ or /l/. Thus for any higher level L_j we must determine which particular ones of the properties and relations defined at L_j have a role at L_i.

3) To show that facts about representations at some higher level can play a role at a lower level is not to show that they always do. Therefore a third question for perceptual theory concerns the circumstances under which available higher order analyses will be exploited. Suppose for instance that syntactic facts such as those involved in the previous examples can help to determine what listeners hear. A natural question is: Do they always have such influence? If we inserted an unambiguous production of clipped into the "foul odor" sentence, would it be heard as crypt anyway, or would this happen only where there was some ambiguity in the acoustic input?

4) It seems likely that higher order influences upon any level will be operative only within structurally defined limits. Thus a fourth question is: What are the domains within which any demonstrated influence reaching from level L_j downward to level L_i may operate? For example, consider the sentences in (1).

(2) (a) The airplane crashed

(b) The airplane from the United States which crashed had just taken off

(c) John was right to avoid that airplane. It crashed.

Suppose we construct productions of each of these sentences in which the underlined word is ambiguous between crashed and clashed. Further suppose that we have already demonstrated that the syntactic fact that the airplane in (2a) is a singular noun strongly biases listeners in favor of hearing the ambiguous item as crashed. Given

all this, the question remains open whether such an influence can reach between clauses as in (2b) or between sentences as in (2c).

This list of questions is not meant to be complete. It is simply an outline of some interesting issues that arise whenever there are higher order analyses which have some potential influence on lower order analyses. As will be seen in the next section, the published literature on speech fusions does not address most of these questions. This work has been done with quite different questions in mind. Nonetheless, it should become clear that speech fusions are a potentially rich source of information about higher order influences on speech perception.[7]

1.2 "Phonological" Fusions in Speech Perception

Over the last decade several researchers have observed a variety of perceptual fusions arising from linguistic stimuli. Most notable among these are R.S. Day (1968, 1970, 1974) and J.E. Cutting (1973, 1975, 1976). Cutting (1976) provides a useful categorization of these phenomena into six types. One of the categories he proposes seems especially relevant to the questions outlined above. These he calls "phonological fusions." This category includes instances in which a listener is dichotically presented with two different words but hears a third word which incorporates elements of both inputs. For instance, simultaneous presentation of PAY and LAY frequently results in the perception of PLAY. Following Day (1968), Cutting considers a fusion response to have occurred whenever "a stimulus pair, each item of which has n phonemes,

yields a percept of n + 1 phonemes" (1975: 106). The most frequently studied class of these fusions involve pairs of monosyllabic words which have the same vowel and final consonant but different initial consonants, a stop and a liquid. One of the conceivable reorderings of the inputs is invariably blocked, apparently because constraints on consonant clusters in English will allow only stop+liquid clusters.

Cutting has used the term "phonological fusions" because of this apparent role of phonological morpheme structure constraints.[8] For our purposes it will be useful to distinguish between various kinds of phonological influences on fusions, thus a more restrictive term will be used; I will call these events, "constraint fusions" (CF's).

Cutting (1975) argues that CF's are central nervous system phenomena on the grounds that when the same stimuli are presented binaurally (ie., with the two channels mixed and fed to both ears) the frequency of fusions falls to about one third the rate for dichotic presentation. If fusions are peripheral events the presentation of both words to both ears ought to increase their frequency. CF's are remarkably insensitive to other variations in the physical structure of the input. Intensity differences across channels of up to 15 db and pitch differences as great as 20 Hz have little effect on fusion frequency. Many fusions occur even when a liquid-inital stimulus begins 200 msec before the stop-initial stimulus on the opposing channel. Similarly, when the apparent sizes of the

vocal tracts producing the stimuli are in the ratio 5:6 there is little effect on fusions. There is also no evidence that the relative frequency of occurence of the particular words heard as fusions has anything to do with what observers hear in these experiments.

One physical factor which does affect fusion rate is the type of device used to generate the stimuli. Fusions occur far more frequently with synthetic speech than with natural speech. Since the patterns of responses are essentially the same across types, this discrepancy apprears to reflect mostly on the inadequacies of speech synthesis. Apparently synthetic speech determines phonemic percepts much less effectively than natural speech does.

The possibility that "semantic" factors influence fusions was explored in Cutting (1975) by inserting fusible stimuli in sentence contexts such as THE MINISTER ----S FOR US. The stimuli on opposite channels were identical except for the blank position which was filled by the fusible pair PAY/LAY or PAY/RAY. One result of the manipulation was roughly a 50% increase in the frequency of fusions. The second result depends upon a curious fact about fusions arising from stop + /r/ presentations. About 70 to 80% of all fusions generated by these pairs are heard as /l/-fusions such as PLAY. This proportion is reliably large under several manipulations explored by Cutting. There were no cases where a majority of stop + /r/ presentations were heard as /r/-fusions and there is no corresponding tendency to hear /l/ as /r/. The point of Cutting's sentence contexts was to try to create a biasing context in favor of

/r/-fusions. It didn't work. /l/ for /r/ substitutions were about equally frequent regardless of the plausibility of the /l/-fusions in the context. Futhermore this frequency was about the same as it had been for isolated presentations. Cutting concludes from this that sentence level semantic factors cannot counteract the /l/ for /r/ substitution effect. There are reasons, however, why we should reject this conclusion.[9]

First there are many complex semantic interrelations among the various elements of a sentence (cf. Katz (1972) and Jackendoff (1972)). As noted above, even if we show that one such relation does not affect a certain aspect of speech perception, clearly this is not equivalent to showing that none have any influence. Second, there appears to be no significant linguistic contrast between the "pray" and "play" cases. Though their internal structures and interpretations are obviously somewhat different, both are perfectly acceptable sentences. Neither the rules which govern the syntactic structure of English sentences nor those that govern the assignment of interpretations to sentences give us any reason to think that one or another of them is ill-formed. Differences between these sentences arise only when we ask which is more likely to be true, or to be believed, or (for social reasons) to be said. But these questions obviously must be judged against our knowledge of, or beliefs about, the objective world. A speaker's knowledge of English grammar contributes nothing to their answers.

What is at issue here is a question similar to those discussed

in section 1.1. Just as we hypothesize that a speaker's knowledge of his language can be segmented into components concerned with levels of phonetics, syntax, or semantics, or mappings between these levels, we also hypothesize that some distinction is to be made between a speaker's knowledge of his language and his knowledge of the world. By hypothesis, if a listener knows some fact F which suggests that a speaker is unlikely to produce an utterance of a certain form, and the category of knowledge to which F belongs is involved in speech perception, then the listener's percept may be affected. In these terms, Cutting's experiment is relevant to the possible claim that a listener's general knowledge of the world (as opposed to his knowledge of his language) has no role in the determination of segmental percepts in speech. A change in the rate of /l/ for /r/ substitutions would have tended to falsify this claim.

Improving upon Cutting's experiment as a test of semantic influences involves some difficulty. The boundary between the syntactic and semantic domains is not pretheoretically given; both its existence and, if it exists, its 'location' are matters of dispute among theorists. For any given relation that might influence speech perception, there is no certainty how this relation will eventually be categorized for the purposes of linguistic theory. This need not, however, be an inhibition to experimentation in psycholinguistics. On anyone's view of the boundary between syntax and semantics, it is clear that within each of these categories there are numbers of particular kinds of relations obtaining between various constituents

within a sentence. Again, because one of these relations is shown to have a role in speech perception does not guarantee that any other will also have influence, even other relations that fall on the same side of some putative boundary between syntactic and semantic phenomena. Thus even if we had some definitive linguistic distinction between syntax and semantics it would still be necessary for psycholinguists to examine the roles of particular types of semantic and syntactic relations in speech perception. There is no reason why this program may not proceed while the boundary question remains open. From some viewpoints at least, psycholinguistic results might even contribute to the delineation of that boundary.

Returning to the question whether Cutting's experiment might be improved upon, an approximation to a 'clean' test of semantic influences might be achieved with examples such as COASTING YOUR CAR DOWN A STEEP HILL IS A (CLIMB/CRIME). Presumably the definition of CLIMB blocks its occurrence in this definition-like context.[10] The difficulty of constructing any variety of such examples is perhaps suggested by AN ENTHUSIASTIC ---- GATHERED where the possible fusions are CLOUD and CROWD. Here there is surely a semantic difficulty in attributing enthsiasm to an inanimate object, but this may be reflected in a syntactic marking on the adjective requiring that it be associated only with animate nouns.[11] Thus any affect on /l/ for /r/ substitutions that might be observed could not be confidently attributed to semantic influences.

One other higher-order linguistic effect on fusions has been reported. Day (1968) found that when one possible fusion was a real

English word and the other was not, there was a tendency for observers to hear the real word. This produced a significant effect on /l/ for /r/ subtitutions. Nonetheless observers do report substantial numbers of nonceword fusions with and without an acoustic representation of /r/ in second position.

In an attempt to understand the mechanism underlying the very high levels of /l/ for /r/ substitutions, Cutting performed several experiments which are relevant to the problems central to this paper. To determine whether observers could detect any differences between those /l/-fusions which were prompted by stimuli containing [r] and those which contained [l], Cutting performed a discrimination test which presented the contrasting types of stimuli to the observers in rapid sequence so that close comparisons could be made. Observers performed just above chance levels when asked to compare, for instance, a PLAY fusion based on an [l]-stimulus with one based on an [r]-stimulus.

On the chance that some aspect of the vowel following the liquid contributed to the substitution effect, a test was performed in which all formant transitions were removed from some of the liquid stimuli. Fusions occurred only 6% of the time. Since this suggests that the formant transitions in the liquids are critical to the fusions, another test used single transitions without vowels in place of the full liquid stimuli. Twelve different transitions were tried. Those showing a falling pattern or steady state yielded stop + liquid fusions on 8% or less of the trials. Only rising transitions produced

fusion rates above 8%. The most effective transition corresponded to the second formant transition occurring in both [l] and [r] stimuli.[12] This second forman[illegible]ced fusions on 24% of the trials where it was

It is instructive to add these numbers. If we assume that the effects of the stimuli without transitions and those having only transitions are additive, we would expect that combining these into a single stimulus would yield fusions about 30% of the time. In fact a stimulus such as this produced fusions on 60% of its trials. This is particularly striking considering that just the same information about initial consonant identity was available with the transition alone as with the transition in context with the following vowel and consonant. The significance of this observation will be considered below.

1.3 Further Questions

On the face of it the various constraint fusions reported by Cutting and Day show that there are at least some higher-order influences on speech perception. Phonetic sequences [lp] and [rp] are allowed across morpheme and syllable boundaries, though of course the liquids in such environments are not identical to those that might occur in PLAY and PRAY.[13] Nonetheless, it is only at the phonological level that there appear to be prohibitions agains these sequences. Thus the most natural account of the observations seems to involve reference to some of the morpheme structure principles of English.

MSC's also may contribute importantly to an explanation of the dramatic increase in fusions when the isolated second formant was followed by a VC sequence instead of appearing entirely alone on its channel. Presumably, the increase reflects an increase in the frequency with which the second formant is taken to be a linguistic stimulus. If the speech perception system continuously exploits MSC's, this is just what we'd expect since the MSC's would indicate that the most likely (in the sense of least marked) content of the interval preceding the VC is some C.[14] If the second formant used is specific to liquids, the MSC's could then supply all of the feature values except those that distinguish /l/ from /r/. Even these could be supplied if the system, in the absence of information to the contrary, simply assumed that the segment actually occurring is the least marked of these alternatives.

Whether any such exploitation of MSC's ordinarily occurs cannot be determined from Cutting's or Day's results. The segments and environments used are too limited.

Only MSC's and the lexicon are implicated in the observations. A variety of phonological, syntactic and semantic principles which conceivably could be involved in fusions have yet to be investigated.

The only domain within which these phenomena have been shown to operate is the single morheme, and even here nearly all the research is concerned with initial consonant clusters.

The research described below is intended as a pilot project. It provides a modest extension of the work of Cutting and Day. It

would be more surprising to find that the phenomena observed do not exist than it is to find that they do.

The stimuli used were of two main types. The main set consisted of various nonsense words to which plural endings could be attached.

A secondary [illegible] real words together with past tense endings. The infl[illegible]ndings always appeared on one track only and always had the opposite voicing value as the final consonant on the opposite track. The question was whether the inflectional endings would ever become perceptually associated with the material on the other track and, in so doing, change voicing value. This potentially extends the work of Cutting and Day in several respects. It involves segments that have not figured in previous work, and these appear in quite different environments. Day's n+1 definition of fusions does not apply to these cases, though we can quite reasonably replace it with the notion that a fusion has occurred whenever the percept contains either particular segments or a sequence of segments that did not occur in the stimuli. Any fusions that arise with such materials presumably involve phonological rules instead of MSC's. I will call such fusions, "process fusions" (PF's). The domain of such fusions is necessarily two morphemes, and thus broader than that of CF's.

2. An Investigation of Process Fusions

The research outlined below was intended as a preliminary search for evidence of phonological process fusions.

2.1 Method

Tapes 1 and 2 were designed to provide a number of simultaneous presentations of two nonsense words, one with a plural marker attached and the other without. The question was whether this plural marker would ever be perceived in a different context and form than the [illegible] presented, eg S/REEG would ever be perceived as REEGS. Tape 3 is similar to 1 and except that four pairs of real words were used in conjunction with past tense markers instead of plural markers. Again the question was whether the inflectional element would ever be moved around and/or transformed by the perceptual system. Effects such as these would presumably reflect the involvement of phonological processes in fusion phenomena.

2.1.1 Materials: Stimulus tapes for these tests were prepared on the DDP-224 PCM system at Haskins Laboratories, New Haven, Conn. Recordings of natural speech productions of the stimulus words were recorded into digital form and stored individually on a disk file. The loudness of the various words was adjusted to approximately the same level by ear. Objective intensity varied over a range of approximately $\pm$ 4dB. Dichotic tapes were prepared by recalling items from the disk file and recording them on a digital version of the stimulus tape. Each trial consisted of the simultaneous presentation of two words, one on each channel. The information on the digital tape was then converted to analog form and recorded. These recordings were the stimulus tapes discussed below.

Tape 1 was built around a set of six pairs of nonsense words which had been designed so that both members of each pair had the same initial consonant and medial vowel. The final segments were stop consonants throughout. Within each pair the final consonants were always different in both voicing and place of articulation. The six pairs were reet [rit]/reeb, reet/reeg, rauk [rok]/raub, raup/raud, rauk/raud, and [illegible]/raug. A single vowel for all forms would have been preffered but none chould be found that would render all the forms meaningless. Each member of each pair was used with and without a plural marker. Since there were two items in each pair to which the plural could be attached and two channels on which each such combination could occur, each pair of words was represented in various forms four times in each block of trials. Thus there were 24 trials in each randomized block of trials (6 pairs x 2 plural locations x 2 channels).

With 20 sec. rest periods intervening between blocks, five blocks were recorded to make up Tape 1. Pre-testing with Tape 3 (described below) and similar materials showed that observers require about six seconds of response time between trials on this type of task. Six second inter-trial intervals were used on Tapes 1 and 2.

Tape 2 was modeled on Tape 1 and used two of the word pairs from that tape, reet/reeb and rauk/raub. The primary difference between the two tapes was in the introduction of a 30 msec. delay on one channel on every trial. Thus there were 2 word pairs x 2 locations for the plural marker x 2 delay relations(left delayed, right delayed) x 2 channels resulting in 16 trails per block of approximately the same form as in Tape 1. In another 16 trails per block the same word

appeared on both channels (4 words x 2 pluralization values x 2 delay values). Four blocks were recorded to make Tape 2.

Tape 3 used real words as stimuli, in particular TACK/TAG, RACE/RAISE, LOP/LOB, and RIP/RIB. Note that, unlike the pairs used in Tape 1, the final consonants on opposite channels contrast only in voicing. Each word was used in present and past tense forms. Otherwise, Tape 3 was similar in design to Tape 1. There were 16 trials per block resulting from 4 word pairs x 2 locations for PAST x 2 channels. Three independently randomized blocks were recorded to make Tape 3. Inter-trial intervals were 3 sec. throughout.

2.1.2 Procedure: Each observer was told at the beginning of his or her session that the task involved listening to some recordings of simultaneous word pairs. Observers were asked to report whether the two "voices" heard on each trial were saying the same word or two different ones, and what was said. Observers wrote their reports on forms. They were required to write a numeral "1" or "2" to indicate whether they thought the two words were the same or different before writing out the words themselves. Most observers were tested in the laboratory of the Gertz Clinic at Queens College. Observers were seated in a soundproof booth and heard the stimuli over a pair of Grayson-Stadler headphones driven by an Ampex tape recorder. Level adjustments were made by ear. Post hoc objective measurements showed average intensities of 82dB and 85dB in the right and left channels respectively. Intensities varied $\pm$ 4dB within each channel and pairwise differences on single trials were as great as 8 dB though these differences averaged a little less than 4 dB.

Two observers listened to stimulus tapes at the investigator's apartment. The tapes were played on a Sony model 366 tape recorder and Koss KO 727 headphones. Signal intensity was set at a comfortable level but was not measured.

A third group of observers listened to tapes in the Communication Research Laboratory of the Communication Arts and [illegible] at Queens College.

2.1.3 Observers: All observers were adults with normal hearing whose native language was American English. Thirteen observers were paid $2 each for their participation and two were unpaid volunteers. Data supplied by two observers was discarded because it was possibly distorted by an equipment malfunction.

2.2 Results

The written responses provided by observers were compared with the contents of the stimulus tapes. Whenever an observer reported hearing an inflectual ending conjoined to a final consonant with a different voicing value than the one to which the ending was attached in the input, this was taken as evidence that a fusion had occurred. Three varieties of these events were possible with Tapes 1 and 2 (see Table 1): Type A) the plural marker could 'move' to become attached to the final consonant presented on the opposite channel from the plural itself, Type B) the final consonant to which the plural was appended in the input could 'change' its voicing value with the plural remaining attached, or Type C) the plural could become attached to a consonant with a different voicing value than the input form of the plural itself and a different place of articulation value either

of the two presented final consonants. Responses of the third type were considered anamalous. On Tape 3 there were no place of articulation contrasts between the final consonants so the only question was whether the PAST marker was attached to a [illegible] as it had been in the stimulus.

4.2.1 Tape 1: Four observers listened to Tape 1 at the Gertz Clinic. Out of a total of 480 trials, fusions occurred on 20.4% or 98 trials. Two fusions occurred on 8 trials for a total of 106 fusions. There were significant variations in the frequency of fusions across observers (see Table 2).

Fusions were considerably more likely to occur when the input form of the plural marker was voiceless than when it was voiced; $z=7.25$, $p<.001$. This tendency was independently significant for two observers and all observers showed the same pattern of results (see Table 3). The fusions which occurred were unequally distributed among the three possible types of response when the input was voiceless; $x^2=21.6$, $df=2$, $p<.001$. Type A responses were most frequent. Type B responses somewhat less frequent and Type C responses much less frequent than both A and B (see Table 4).

A number of factors which conceivably could have affected the frequency and form of fusion events apparently did not. Frequency of fusions was unaffected by the ear to which the plural form was presented (see Table 5). Nor did it seem to matter which of two inputs was objectively louder. A comparison of the three voiced final consonant clusters which the observers reported hearing when the input was voiceless, [bz] [dz] [gz], uncovered no significant differences

in frequency between them. Each of the voicelss clusters, [ps] [ts] [ks], was about equally likely to stimulate fusi[illegible] the observers' identifications of the stimuli agreed with the experimenter's on 99.5% of 196 trials presented monaurally.

Four additonal observers listened to Tape 1 in the Communication Research Laboratory. Faulty equipment made it necessary to present the materials in binaural form, i.e., with the two channels mixed and with this signal supplied to both ears. Under these conditions fusions occurred with virtually the same frequency as in the dichotic presentations discussed above. Again there were significant variations in the frequency of fusions across observers; the tendency for the voiceless plural form to give rise to far more fusion events than the voiced form was also found in these observations (see Table 3). Because the effects of the equipment problems are unknown no further analyses were attempted.

2.2.2 Tape 2: Two observers listened to Tape 2 at the Gertz Clinic Laboratory and were found to have perceived far fewer fusions than other observers who had listened to Tape 1. To distinguish possible differences between observers from the effects of the 30 msec delays introduced on Tape 2 another two observers listened to both Tape 1 and Tape 2 at the experimenter's home. By comparing the frequency of fusions on Tape 2 with the frequency obtained with Tape 1 with the same observers it was possible to determine that the manipulations on Tape 2 did have a significant tendency to suppress fusions; $z=4.82$, $p<.001$ (see Table 7). Since numbers of phonetic feature

[15] appeared in the observers' responses to both of these tapes, the effects of Tape 2 on this kind of fusion was compared with that for process (plural) fusions. This revealed that there was also a significant suppression of feature fusions on Tape 2. Feature fusions were also significantly more common on both tapes than were process fusions (see Table 8) and process fusions were more severely suppressed by Tape 2 than were feature fusions (see Table 8).

2.2.3 Tape 3: One observer who had previously listened to Tape 1 also listened to Tape 3 at the experimenter's home. Fusions involving changes in the PAST marker occurred on 16.7% of the trials as compared with 12.5% CF's for the same observer on Tape 1. This difference was not significant. Fusions involving the perception of voiceless forms as voiced predominated in these data as they had in the data for Tape 1 for the same observer and the ratio of voiced to voiceless fusion percepts was essentially the same.

2.3 Discussion

Do these results demonstrate the existence of process fusions in speech perception? I believe the appropriate answer, though tentative, is yes.

Observers have reported hearing sequences which appear to involve the phonological rules of English. A fairly large number of these events has been observed and no observer failed to report at least a few. The materials used include two kinds of inflected forms, several different stems (only some of which are real words), and incorporate all of the stop consonants in English. The observers have

not, in general, found anything strange about what they heard. The exception to this is an observer who felt that one or two of the plural-like forms may have been a little odd. She was quite unsure, however, and estimated the number of odd productions well below the number of fusions she experienced. Three observers who had previous training in phonetic transcription listened to all or part of Tape 1. Though all reported some fusion events, none found the production of any of these odd.

The individual differences which appeared among observers in this study is also characteristic of the results obtained by Cutting (1975) and Day (1969; 1973).

The strongest evidence against the claim that the observed events are of the same general kind as those reported by Cutting and Day is in their apparent high sensitivity to small differences in onset time. Cutting (1975) obtained fusion rates above 10% with onset differences as great as 200 msec while here differences of only 30 msec produced a drastic reduction in fusions. Day (1970) found that offsets of up to 150 msec had little effect using natural speech stimuli. The significance of these differences remains unclear because of two factors. First, almost all previous work has used pairs of stimuli which contrast in their initial segments instead of their final segments. Second, the stimuli used in previous studies have been carefully constructed while those used here are comparatively quite crude. The temporal structure of paired stimuli varied considerably.

Another distinction between the present results and those of

Cutting lies in the failure to find a contrast between the binaural and dichotic modes of presentation. In Cutting's (1975) data fusions were three times more likely in the dichotic mode. What we should make of this difference is especially unclear since the binaural presentations reported here were made on equipment of uncertain quality. If the effect persists in further work with PF's, it may indicate that these phonemena are of a quite different and possibly simpler sort than CF's.

Future work with PF's should provide information on four factors: relative onset time, differential intensity effects, differences in fundamental frequency, and mode of presentation. All of these can be better studied with synthetic speech. This will also make the results more readily comparable to published work on CF's.

Assuming that PF's belong to the same general type as CF's, it should be emphasized that each bears a quite different relation to the phonological structure of English. CF's reflect rearrangements of phonetic sequences in which all of the segments may retain their original identities.[16] PF's don't necessarily involve any linear reordering but do necessarily involve a change in the phonemic identity of one segment. Furthermore, English MSC's have an entirely negative role in CF's; they rule out one possible ordering of the phonetic sequence. With PF's both of the possible fusion products often have the same relation to the MSC's'; either both are acceptable when considered as single morphemes, or both are not. Thus the MSC's do not choose between them. Additionally, there is a specific phonological rule which determines the voicing value of

the final segment if we assume that a morpheme boundary does precede it. Thus the phonological principles of English provide more positive direction in PF phenomena. Finally, it should be noted that if the voiced fusions which predominated in the present study are regarded as single morphemes, they would all be in violation of an MSC. But with the much less frequently occurring voiceless fusions, if we make the same one morpheme assumption, both forms are tolerated. In other words, MSC's not only cannot account for which fusions occurred, so long as we make the one morpheme assumption, they rule out just those fusion forms which occurred most frequently.

3. General Conclusions

For the sake of discussion in the following I will assume that some kind of process fusion has been shown to exist. The issue here is, how might this relate to the theoretical questions set out in 1.1.

3.1 Some Prospective Answers

With respect to Questions 1 and 2 of Section 1.1, the research reported here, together with that of Cutting and Day, strongly endorses the view that the speech perception system can exploit specifically phonological knowledge. The fusions reported by Cutting and Day apparently require reference to the MSC's of English while those reported here apparently involve the phonological processes of English. It is conceivable, of course, that something corresponding to these kinds of linguistic knowledge can be expressed in terms

that mention only acoustic events and distinctive features, but the awkwardness and redundancy of any such proposal seems likely to rob it of any plausibility.

In addition, Day (1969) provides evidence that the appearance of a form in the English lexicon contributes to its fusibility. Thus lexical knowledge also appears to play some role in speech perception. At present fusion research contributes no evidence for the involvement of syntactic, semantic or pragmatic factors in speech perception, though the experiments along the lines of those suggested in 1.2 may develop such evidence.

It is difficult to get general answers to the third question raised in Section 1.1 from fusion experiments since they necessarily involve a signal which is in some sense degraded. Dichotic stimuli which contrast radically across channels clearly stress the speech perception system to some degree and thus may be regarded by the system as special cases to which extraordinary resources must be applied. One argument against this view is simply that it doesn't feel that way. Observers report that the difficulty they have in doing the tasks reported in Section 2 lies primarily in the difficulty of remembering what they've heard, not in hearing it in the first place. Thus the hallmarks of stress which we see in the data, the variations in the responses to constant stimuli, are apparently not reflected in an experience of degraded stimuli. Cutting's evidence relative to MSC's is much more compelling. Observers simply could not distinguish fused from natural stimuli much above chance. If fusions were significantly 'harder' to ar-

rive at, this difference should have shown up on the discrimination task.

In any case we need not let this question rest at this point; there may be other experimental paradigms that will help us determine the circumstances under which phonological knowledge becomes involved in speech perception. In particular, we can get a more objective measure of degree of stress in fusions by combining fusion stimuli with reaction time task. If we get reaction time measures for both fused and natural occurrences of, say, PLAY, and if the fused percept involves some extraordinary resource, then we would expect this to be reflected in longer reaction times for the fusion trials.

With respect to Question 4, Cutting's and Day's research involves lignuistic contexts of no more than a single morpheme and only a single syllable within that. The new research reported above involves processes reaching across one morpheme boundary. Obviously, if higher order linguistic knowledge is shown to be involved, it is likely to come from somewhat larger domains.

In sum, fusion research has contributed to answers to the theoretical questions sketched in 1.1 and seems able to contribute more still. In general it seems to provide a most promising site for research on the involvement of higher order linguistic knowledge in speech perception.

3.2 A Surprise

Before concluding this paper I would like to suggest that we not take the news that MSC's can become involved in speech percep-

tion too calmly.[17] There appears to be little reason from the standpoint of either linguistic or psychological theory to anticipate this development. On the face of it MSC's seem to be relevant to psychological problems such as the acquisition or invention of new lexical items and (if we assume a psychological parallel to the role assigned to them in Chomsky and Halle's (1968) markedness theory) to the problem of recovering full phonetic matrices from the lexicon. None of this seems to entail any involvement in perception. Nor does it seem especially convincing to suggest that the speech perception system should edit its output to allow out only those things which are possible words in the listener's language. Those items which are not words could be detected solely by reference to the lexicon and those things which are neither words nor possible words could be prevented from becoming part of the lexicon by the MSC's at the point of entry to that system.

But suppose that in ordinary speech the acoustic signal and the principles that mediate between the acoustic level and the phonemic level do not, in themselves determine what the complete phonemic analysis will be. Suppose that the analysis to this point yields only an incompletely specified matrix of features. In this situation the MSC's might 'save' the system by making it possible to fill out complete phonemic matrices on the basis of fragmentary phonetic inputs. Or, to put it more positively, early perceptual deployment of MSC's may be able to make a major contribution to the speed and reliability of speech by reducing the amount of information

that must be encoded in the acoustic signal, i.e., by reducing the effective bit rate. Furthermore, MSC's can contribute importantly to the discovery of various linguistically significant boundaries by ruling out certain phonetic sequences within morphemes. The occurence of unacceptable sequences can effectively mark boundaries simply because certain sequences can't occur unless they include a boundary.

Beyond perceptual considerations, a speech perception theory which exploits the segmental and sequentional redundancy of language holds some promise for helping to solve some very perplexing problems in language acquisition. Linguistic markedness theory has been designed primarily to capture the notion of naturalness with respect to a variety of linguistic structures including segments, morphemes, rules, vowel systems and consonant systems.[18] Thus if it were to turn out that MSC's reflect universals of language and that they are available to children very early in life, they could make a major contribution to the acquisiton of all the phonetic, phonological and lexical knowledge that a child presumably must acquire in order to learn the syntactic and semantic properties of his language.

The study of linguistic redundancy as represented in markedness theory may be a rich source of hypotheses for psychologists interested in speech perception as well as language acquisition.

Notes

1. Theories of speech perception which emphasize the role of higher order levels of analysis are discussed in Stevens (1972, 1973) and Chomsky and Halle (1968:24).

2. Most research in speech perception has not distinguished between phonetic and phonemic levels. Rather, it has been primarily concerned with the relation between the auditory level and some imprecisely defined level at which percepts of single phones are available. See Studdert-Kennedy (1974) for a thorough review of the speech perception literature. For the purpose of this paper it will be assumed that the sequence of phones perceived in speech is the output of a phonemic analysis (in a sense consistent with Chomsky and Halle (1968)) and that, in general, strictly phonetic information is not available to consciousness.

3. See Studdert-Kennedy (1974:5) for some comments on the role of higher-order levels of analysis in speech perception.

4. See Ross (1976) and references cited there for a discussion of Austin's observation and for reports of a number of very interesting fusion effects arising in binocular vision.

5. Chomsky (1975: 105-117) for a discussion of the notion of a level in liguistic theory.

6. These remarks are not meant to imply that there is necessarily and relation between the levels of analysis appropriate to a linguistic theory and those employed by performance devices which realize the languages described by the theory. I do assume, however, that in general all of the properties which an ideal grammar attributes to a sentence of the language it describes are properties which a perceptual device must discover in the course of processing an utterance of that sentence.

7. I take it to be well-established that there are at least some ways in which higher levels can influence speech perception. Studdert-Kennedy (1974: 2-5) reviews relevant evidence.

8. The facts about segmental and sequential redundancy were first incorporated into generative theories as language specific redundancy rules and served to fill in the missing values in incompletely specified phonemic matrices in the lexicon (cf. Halle (1962)). There were several inadequacies with this approach and it was replaced by the theory of markedness which attempts to account for most redundancies in terms of universal principles. See Postal (1968), Chomsky and Halle (1968), Cairns (1969) and, for a review of both present and earlier notions of markedness, Hyman (1975, Chap. 5). Except for some considerations raised in Section 3, the fact that many aspects of the phonological structure of a particular language reflect universal constraints will not be important to the issues reviewed in this paper.

9. I will discuss only one of Cutting's four sentence contexts in detail but the same criticisms apply to the remaining three. The others used were THE TRUMPETER PLAYS/PRAYS FOR US, THE COALS ARE GLOWING/GROWING AGAIN, and THE TREES ARE GLOWING/GROWING AGAIN.

10. For a discussion of semantic well-formedness conditions and some examples which may suggest experimental possibilities see Katz (1972, chapters 1 and 2) and Jackendoff (1972: 17-21).

11. For a discussion of the syntactic role of seemingly semantic notions see Chomsky (1965: Chapter 2).

12. The acoustic distinction between the [l] and [r] stimuli consists solely in the slope of the third formant transition.

13. Of course the [l] and [r] would not be pronounceable forward of the stop unless they were preceded or followed by some V. The point however, is that it is apparently the MSC's which represent this fact.

14. See Chomsky and Halle (1968: 404-407) for the relevant rules.

15. See Cutting (1976) for discussion of phonetic feature fusions. These are events in which the listener hears a segment which appears to combine distinctive features of disparate inputs, eg., when a dichotic presentation of /ba/ and /ta/ is heard as either /da/ or /pa/.

16. The /l/ for /r/ substitutions are in no way necessary in order to satisfy the constraints of English and thus are not motivated by either MSC's or rules.

17. See Note 8.

TABLE 1

Types of responses counted as fusions for Tapes 1 and 2

	Input	Observer Reports Hearing...
Type A	reets/reeg	reegs...
Type B	reets/reeg	reeds...
Type C	reets/reeg	reebs...

TABLE 2

Tape 1: Frequency of trials on which process fusions occurred

Observer	Percent Fusion trials*	Number of Fusion trials
1	14	17
2	13	15
3	33	40
4	22	26

Total trials = 4 observers x 120 trials each = 480

Number of trials on which fusions occurred = 98

Mean frequency of fusion trials across observers = 20%

Total of fusion percepts reported (8 trials produced 2 fusions each) = 106

* A "fusion trial" is one on which the observer reported hearing at least one word which involves a fusion.

TABLE 3

Tape 1: Dichotic presentation; Frequency of fusion trials by observer

Observer	Number of fusion trials occurring: Total	With Voiceless Input	With Voiced Input	
1	17	15	2	$x^2 = 9.94$, $p<.01$
2	15	10	5	$x^2 = 1.67$, NS
3	40	39	1	$x^2 = 36.1$, $p<.001$
4	26	17	9	$x^2 = 2.46$, NS
Combined	98	81	17	

Tape 1: Binaural presentation; Frequency of fusion trials by observer

Observer	Number of fusion trials occurring: Total	With Voiceless Input	With Voiced Input	
5	39	39	0	$x^2 = 39$, $p<.001$
6	18	14	4	$x^2 = 5.56$, $p<.02$
7	28	26	2	$x^2 = 20.57$, $p<.001$
8	15	15	0	$x^2 = 15$, $p<.001$
Combined	100	94	6	

* "Voiceless input" refers to the voiceless form of the plural marker, ie., [s].

TABLE 4

Tape 1: Distribution of fusions by type

Input Form	Total	Fusion Type*		
		A	B	C
Voiceless, i.e., [s]	89	45	34	10
Voiced, i.e., [z]	17	8	4	5
Column Totals	106	53	38	15

x^2 Test for uniformity of distributions across categories

Voiceless	$x^2 = 21.6$, $p<.001$
Voiced	$x^2 = 1.53$, NS
Combined	$x^2 = 20.74$, $p<.001$

* See page 26 for a description of the types.

TABLE 5

Tape 1: Frequency of fusion by observer and by ear to which the plural element was supplied

Observer	Observer Total	Ear to which plural supplied		
		L	R	
1	17	6	11	$x^2 = 1.47$, NS
2	15	11	4	$x^2 = 3.27$, $p<.1$
3	40	21	19	$x^2 = .1$, NS
4	26	15	11	$x^2 = .36$, NS
Column Totals	98	53	45	

Difference of proportion of fusions occurring when plural element is supplied to the left vs. right ear

Left 22.08% of 240 trials

Right 18.75% of 240 trials

$z = .906$, NS

TABLE 6

Tape 1: Distribution of fusions according to cluster perceived on trials with voiceless input, [s]

Cluster perceived	[bz]	[dz]	[gz]
Fusions	23	33	23

$x^2 = 2.53$, df = 2, NS

Tape 1: Distribution of fusions according to cluster presented on trials with voiceless input [s]

Cluster presented	[ps]	[ts]	[ks]
Fusions	21	30	28

$x^2 = 1.70$, df = 2, NS

TABLE 7

Tapes 1 and 2: Comparison of freqency of fusions across tapes

Observer	Tape 1			Tape 2			
	Total Trials	Fusions	Percent	Total Trials	Fusions	Percent	
9	120	35	29.2	64	2	3.1	$z = 4.21$, $p<.001$
10	120	15	12.5	64	1	1.6	$z = 2.5$, $p<.05$
Combined	240	50	20.8	128	3	2.3	$z = 4.82$, $p<.001$

45

TABLE 8A

Tapes 1 and 2: Frequency of phonetic feature fusions by observers

Observer	Tape 1 Total Trials	Tape 1 Fusion Trials	Tape 1 Percent	Tape 2 Total Trials	Tape 2 Fusion Trials	Tape 2 Percent	
9	120	67	55.8	64	3	4.7	$z = 6.8$, $p<.001$
10	120	70	58.3	64	17	26.6	$z = 4.11$, $p<.001$
Combined	240	137	57.1	128	20	15.6	$z = 7.66$, $p<.001$

46

Tapes 1 and 2: Proportion of all trials yielding fusions by type (process vs. phonetic feature) and by materials; Data from Observers 9 and 10 combined

	Tape 1	Tape 2
Process fusions	20.8%	2.3%
Phonetic feature fusions	57.1%	15.6%

42

TABLE 8B

Tapes 1 and 2: Test for similarity of the distributions between process fusions and phonetic feature fusions

		Fusion Trials	
		Tape 1	Tape 2
Process fusions	(a)	50	3
Phonetic feature fusions	(b)	137	20
Expected value of (b) given (a)		148.1	8.9
Expected value of (a) given (b)		46.3	6.8

x^2 (for (b) given (a)) = 14.72, $p<.001$

x^2 (for (a) given (b)) = 2.39, $p<.15$

References

Cairns, C. E. (1969) "Markedness, neutralization, and universal redundancy rules." Language 45. 863-885.

Chomsky, N. (1965) Aspects of the Theory of Syntax. Cambridge: MIT Press.

Chomsky, N. (1975) The Logical Structure of Linguistic Theory. New York: Plenum.

Chomsky, N., and M. Halle (1968) The Sound Pattern of English. New York: Harper and Row.

Cutting, J. E. (1973) "Levels of processing in phonological fusion." Doctoral dissertaton, Yale University.

_____ (1975) "Aspects of phonological fusion." Journal of Experimental Psychology: Human Perception and Performance 104:2. 105-120.

_____ (1976) "Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listenings." Psychological Review 83:2. 114-140.

Day, R. S. (1968) "Fusion in dichotic listening." Doctoral dissertation, Stanford University.

_____ (1969) "Temporal-order judgements in speech: Are individuals language-bound or stimulus-bound?" Haskins Laboratories Status Report on Speech Research SR-21/22. 71-87.

Day, R. S. (1970) "Temporal-order perception in a reversible phoneme cluster." Journal of the Acoustical Society of America 48. 95. (Abstract)

_____ (1973) "Digit span memory in language-bound and stimulus-bound subjects." Journal of the Acoustical Society of America 54. 287. (Abstract)

_____ (1974) "Differences in language-bound and stimulus-bound subjects in solving word-search puzzles." Journal of the Acoustical Society of America 55. 412. (Abstract)

Fodor, J. A., T. G. Bever, and M. F. Garrett (1974) The Psychology of Language. New York: McGraw-Hill.

Halle, M. (1962) "Phonology in generative grammar." Word 18. 54-72 Also in J. A. Fodor and J. J. Katz (eds), The Structure of Language. Englewood Cliffs, N.J.: Prentice-Hall.

Hyman, L. M. (1975) Phonology: Theory and Analysis. New York: Holt, Rinehart and Winston.

Jackendoff, R. S. (1972) Semantic Interpretation in Generative Grammar. Cambridge: MIT Press.

Katz, J. J. (1972) Semantic Theory. New York: Harper and Row.

Leiber, J. (1975) Noam Chomsky: A philosophic overview. New York: St. Martin's Press.

Postal, P. M. (1968) Aspects of Phonological Theory. New York: Harper and Row.

Ross, J. (1976) "The resources of binocular perception." Scientific American 234:3. 80-86.

Stevens, K. N. (1972) "Segments, features, and analysis by synthesis." In Language by Ear and Eye: The Relationships Between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. Cambridge, Mass.: MIT Press, pp. 47-52.

_____ (1973) Potential role of property detectors in the perception of consonants. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 110, pp. 155-168.

Studdert-Kennedy, M. (1974) "Speech perception." Haskins Laboratories Status Reports on Speech Research SR-39/40.