"Psychophysics Versus Specialized Processes in Speech Perception: An Alternative Perspective," in Psychophysics of Speech Perception, by M.E.H. Schouten (ed.), in press.

PSYCHOPHYSICS VERSUS SPECIALIZED PROCESSES IN SPEECH PERCEPTION: AN ALTERNATIVE PERSPECTIVE*

Dominic W. Massaro
Program in Experimental Psychology, University of California, Santa Cruz , Santa Cruz, CA 95064, USA

The title of this conference describes one of the two major contrasting frameworks for speech perception research during the last three decades. This point of view is that speech perception can be understood by the principles of auditory psychophysics. Speech involves complex auditory signals and the processing and perception of speech can be understood by the rules of processing complex auditory signals. Research representative of the paradigm has been contributed by Cutting and Rosner (1974), Kuhl and Miller (1975, 1978), Pastore, Ahroon, Baffuto, Friedman, Puleo, and Fink (1977), and Pisoni (1977). The other point of view, the antithesis of the first, is that speech perception represents the operation of a set of specialized processes unique to speech. This view began as the motor theory of speech perception (Liberman, Cooper, Shankweiler, & Studdert- Kennedy, 1967) and has evolved into an illustration of the modularity principle (Fodor, 1983; Liberman & Mattingly, 1985). Representative studies within this paradigm can be found in Best, Morrongiello, and Robson (1981), Eimas and Corbit (1973), Eimas and Miller (1980), and Repp (1982).

These two schools of thought have imposed a very narrow research environment, falling short of contributing to a basic understanding of speech perception. The empirical work and controversy has bounced back and forth resembling a tennis match in which the server demonstrates that speech perception is special and the opponent replies that it is not. Caught up in the controversy of psychophysics versus specialization, little has been accomplished with respect to the question of how speech is perceived. A resolution of the controversy is offered by a third perspective proven successful in other domains such as reading and categorization of natural objects (Massaro, 1984; Massaro, in press, d).

INFORMATION-INTEGRATION PERSPECTIVE

One salient aspect of pattern recognition involves the processing of multiple sources of information. Consider recognition of the word "performance" in the spoken sentence: "The actress was

*The writing of this paper and the research reported in the paper were supported, in part, by NINCDS Grant 20314 from the Public Health Service and Grant BNS-83-15192 from the National Science Foundation. Michael M. Cohen made important contributions to the research enterprises, Neil Appel helped with the references, and Ervin R. Hafter provided helpful feedback on an earlier version of this paper.

praised for her outstanding performance". Recognition of the critical word is achieved via a variety of bottom-up and top-down sources of information. Top-down sources include semantic and syntactic constraints and bottom-up sources include acoustic features and syllables making up the word. Phonological constraints also have been shown to contribute to perceptual recognition at the word level (Massaro & Cohen, 1983d). Integrating multiple sources of information appears to be central to pattern recognition, not just speech perception.

Historically, the present approach can be traced, in part, to Egon Brunswik's (1952, 1955) Probabilistic Functionalism. He proposed that there are many cues influencing perception but that these cues are equivocal and only probabilistically related to the objects of interest. Brunswik realized "the limited ecological validity or trustworthiness of cues ... To improve its (the organism's) bet, it must accumulate and combine cues" (Brunswik, 1955, pg. 207). Methodologically, Brunswik called for representative designs or experiments that are random samples from natural phenomena. We reject this method, however, in favor of experiments that manipulate the environment. Only by independently varying naturally correlated cues are we able to determine which cues are functional in perception. To this end, we employ factorial designs and functional- measurement techniques (Anderson, 1981, 1982) and test mathematical models of perceptual performance (Massaro & Cohen, 1983c).

According to the present framework, well-learned patterns are recognized in accordance with a general algorithm regardless of the modality or particular nature of the patterns (Massaro, 1979; Oden & Massaro, 1978). The model postulates three operations in perceptual (primary) recognition: feature evaluation, prototype matching, and pattern classification. Continuously-valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions. The model is called a fuzzy logical model of perception (abbreviated FLMP).

It is necessary to distinguish between environmental properties that are potentially informative about some object or event and the properties actually used in perception and recognition of the object or event (Massaro, in press c). The former might be called data and the latter information. With respect to bottom-up properties in speech, I have referred to them as acoustic characteristics and acoustic features, respectively (Massaro, 1975). One primary task of research is to determine which acoustic characteristics function as acoustic features in speech perception. Furthermore, it is necessary to specify the algorithms or computations that resolve the wide variety of acoustic features. The description does not end here, however, because the features, once resolved, must be combined or integrated to achieve a unitary perception. The processing of top-down sources of information must be described in the same manner as bottom-up sources. An adequate theory must also specify how the perceiver integrates bottom-up and top-down sources in real time during speech perception. Finally, decision processes must be revealed given that speech perception, like other forms of pattern recognition, represents a selection of one of several viable candidates or alternatives. I will

first discuss briefly the research areas of categorical perception, normalization, duplex perception, the McGurk effect, and trading relations, and contrast the two modal approaches to our approach to the study of speech perception. This latter approach not only offers a productive framework for describing the processes involved in speech perception, it provides major constraints on potential contenders for a theory of speech perception.

FIGURE 1, Schematic representation of Stages of Processing in Speech Perception.

Figure 1 gives a schematic representation of four stages of processing in categorization of speech. Sensory transduction transforms the physical stimulus into sensory data. The feature evaluation process assesses the sensory data with respect to the important dimensions of speech characterizing the speech segments in the language. The feature integration process integrates or combines the featural information from the different dimensions with respect to prototype representations in memory. The pattern classification process determines the relative goodness of match of the sensory information with the relevant prototypes in memory. The relative goodness of match is equal to the likelihood of identifying the speech event as an instance of the prototype category.

CATEGORICAL PERCEPTION

The speech-is-special school offered the phenomenon of categorical perception in its support (Liberman, et al., 1967). The contemporary field's repression of the concept of categorical perception makes transparent the sterility of this area of research. To this day I cannot understand why categorization behavior was (and continues to be) interpreted as evidence for categorical perception. At the risk of belaboring the obvious, I will illustrate very quickly how it is only natural that continuous perception should lead to sharp category boundaries along a stimulus continuum. Given a stimulus continuum from A to not-A that is perceived continuously, the goodness of A, abbreviated G(A), is an index of the degree to which the information represents the category A. The left panel of Figure 2 shows G(A) as a linear function of Variable A.

An optimal decision rule in a discrete judgment task would set the criterion value at .5 and classify the pattern as A for any value greater than this value. Otherwise, the pattern is classified as not-A. Given this decision rule, the probability of an A response, P(A) would take the step- function form shown in the right panel of Figure 2. That is, with a fixed criterion value and no variability, the decision operation changes the continuous linear function given by the perceptual operation into a step function. Although based on continuous perception, this function is identical to the idealized form

of categorical perception in a speech identification task (Studdert-Kennedy, Liberman, Harris, & Cooper 1970). It follows that a step function for identification is <u>not</u> evidence for categorical perception because it can occur given continuous information.

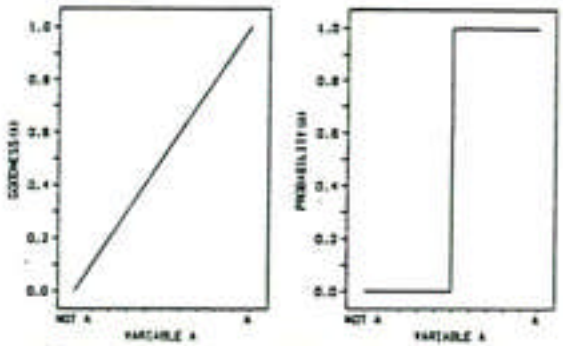

FIGURE 2, Left Panel. The degree to which a stimulus represents the category A, called Goodness(A) as a function of the level along a stimulus continuum between not-A and A. Right Panel. The probability of an A response, Probability(A), as a function of the stimulus continuum if the subject maintains a decision criterion at a particular value of Goodness(A) and responds A if and only if Goodness(A) exceeds the decision criterion.

If there is noise in the mapping from stimulus to identification, a given level of Variable A cannot be expected to produce the same identification judgment on each presentation. It is reasonable to assume that a given level of Variable A produces a normally distributed range of Goodness(A) values with a mean directly related to the level of Variable A and a variance equal across all levels of Variable A. If this is the case, noise will influence the identification judgment for the levels of Variable A near the criterion value more than the levels away from the criterion value. Figure 3 illustrates the expected outcome for identification if there is normally distributed noise with the same criterion value assumed in Figure 2.

If the noise is normal and has the same mean and variance across the continuum, a stimulus whose mean goodness, G(A), is at the criterion value will produce random classifications. The value of G(A) will be above the criterion on half of the trials and below the criterion on the other half. As the value of G(A) moves away from the criterion value, the noise will have a diminishing effect on the identification judgments. Noise has a larger influence on identification in the middle of the range of G(A) values than at the extremes since variability goes in both directions in the middle and only inward at the extremes.

This example shows that categorical decisions made on

the basis of continuous information produce identification functions with sharp boundaries, previously taken to represent categorical perception. Strictly speaking, of course, categorical perception was considered present only if discrimination behavior did not exceed that predicted from categorization (Studdert-Kennedy, et al., 1970). However, one should not have been impressed with the failure of discrimination to exceed that predicted by categorization if the discrimination task resembled something more akin to categorization than discrimination (Fujisaki & Kawashima, 1970, Pisoni, 1975). Even in this period of enlightenment, however, we have authors such as Eimas (1985) using the concept of categorical perception to describe typical categorization behavior.

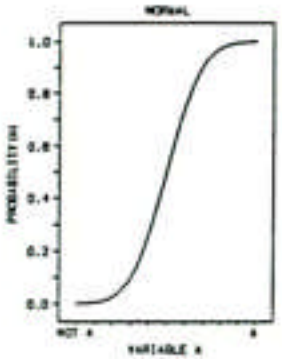

FIGURE 3. The probability(A) as a function of Variable A given the linear relationship between Goodness(A) and Variable A and the decision criterion represented in Figure 1, but with normally distributed noise added to the mapping of Variable A to Goodness(A).

This analysis of categorical perception also makes explicit at what level in the processing system categorical perception must be demonstrated. Categorization behavior alone cannot be taken as evidence for categorical perception, for it is the mapping of stimulus information to sensory information (feature evaluation in our model) that is relevant, not simply that mapping of stimulus information to perceptual judgment. The issue formalized in Figure 1 is whether the mapping of Variable A to Goodness(A) is continuous or categorical. I don't understand Repp's (this volume) opinion that the present analysis of the problem precludes any contribution of experience and attention. In fact, the McClelland and Elman Trace Model places categorical perception at exactly the level of mapping of Variable A to Goodness (A) (see later discussion). If categorization implies categorical perception, we have abandoned any interest in the processes leading to perception and have joined the behavioristic camp of psychological inquiry.

The psychoacoustician's answer to categorical perception was not to reject the concept but to attempt to show that it also occurs for nonspeech stimuli. Thus speech cannot be considered special because both speech and nonspeech are perceived categorically. An attractive but incorrect solution was the idea of natural auditory sensitivities accounting for perceptual categories in speech. Little processing is needed if the speech categories fall on opposite sides of some perceptual discontinuity in the auditory system. For example, the most popular distinction is voice onset time (VOT), the time interval between the onset of the release of a stop consonant and the

onset of vocal cord vibration (see Rosen & Howell, this volume). The psychoacoustic explanation for VOT has been in terms of temporal order judgments (TOJ). In 1959, Hirsh found that listeners needed about a 17 ms onset difference to determine which of two auditory events (e.g., tones) occurred first. If the 17 ms difference is viewed as a threshold, then stimuli with onset asynchronies longer than this value would be heard in one way and stimuli with onset asynchronies shorter than this value would be heard in another way. Voiced stop consonants would have VOT values less than the threshold, whereas voiceless stops would have VOT values greater than the threshold.

There are several limitations with this proposal that make it an unlikely possibility for perception of voicing. First, listeners do not experience two auditory events one before the other, but instead qualitatively different types of percepts. (Massaro, 1972; Warren, 1974a, 1974b). Second, the onset difference at threshold depends on other factors such as the overall duration of the auditory events (Patterson & Green, 1970). Third, there is little evidence that the critical cue to voicing is temporal order as opposed to the acoustic events during the onset of the speech segment. In a neglected study, Winitz, La Riviere, and Herriman (1975) opposed VOT and the quality of the burst and aspiration in stop consonants in initial position. As an example, the burst of /du/ was isolated and separated from the periodic portion by an interval appropriate for the VOT of /tu/. In this situation, listeners tended to hear /du/ showing that VOT itself was not the critical cue to voicing. Fourth, Rosen and Howell (this volume) illustrate that Hirsh's (1959) results are inconsistent with a threshold or discontinuity in the perception of relative onset time. Other limitations to a TOJ threshold in speech exist and are mentioned in other sections of this commentary.

At best, there may be only a couple of speech contrasts even approaching categorical speech perception (Howell & Rosen, 1983). These few cases are better described as slight irregularities or discontinuities rather than by sharp thresholds in the auditory processing system (Hary & Massaro, 1982; Massaro & Oden, 1980b). By irregularity is meant that the mapping between some perceptual variable such as the discrimination of which stimulus came first and some physical variable such as relative onset time is irregular. Speech, like nonspeech, is unlikely to be perceived categorically, even in those few rare cases of irregularity across some continuum.

Some psychoacousticians had the inappropriate reaction to the notion of categorical perception. Macmillan, Kaplan, and Creelman (1977) redefined categorical perception within the theory of signal detection which, of course, assumes continuous information. Although categorical perception was defined as a match between categorization and discrimination performance, the underlying assumptions were those of continuous perception. Thus, a match between discrimination performance and categorization performance certainly could not mean categorical perception. More generally, evidence consistent with the traditional definition of categorical perception is not necessarily inconsistent with alternative models assuming continuous perception (Massaro, in press b). Using the traditional paradigm of the categorization/discrimination task, Hary and Massaro (1982) demonstrated that sounds that appear to be perceived categorically in one context appear to be perceived continuously in another. Massaro

and Cohen (1983a) showed that the distribution of rating responses to speech syllables was significantly better described by a continuous than by a categorical model of perception. None of us like to be reminded of the sterility of a particular research enterprise. If we can face up to mistakes in the past, however, we may be less susceptible to repeating them in the future.

The idea of categorical perception is not feasible for the perception of continuous speech. It is well known that the acoustic characteristics of the speech signal vary considerably with speaker, rate of speaking, and the surrounding segmental contextual. Categorical perception would prove to be too inflexible to deal with these context variations. As an example, Summerfield (1982) showed that the influence of voice-onset time (VOT) in the discrimination of voicing could not be categorically perceived. The contextual modifications resulting from the influence of this variable in speech are significantly different from what could be predicted from nonspeech. For example, the position of boundaries between phonetic categories on VOT continua depends on other spectral properties, such as the onset frequency of the first formant. Analogous spectral manipulations in the nonspeech analogs of the syllables do not result in a similar dependency.

It is also of considerable interest that the demise of categorical perception poses a serious problem for the Trace Model of speech perception (McClelland & Elman, 1986). Their model produces categorical-like behavior at the sensory (featural) level rather than at simply the decision stage. This occurs because of the nonindependence between the feature and phoneme levels in the model (which contrasts with the independence assumption of our fuzzy logical model). In the Trace Model, a stimulus pattern is presented and activation of the corresponding features sends more excitation to some phoneme units than others. Given the assumption of feedback from the phoneme to the feature level, the activation of a particular phoneme feeds down and activates the features corresponding to that phoneme (McClelland & Elman, 1986, p. 47). This effect of feedback produces enhanced sensitivity around a category boundary, exactly as predicted by categorical perception. Categorical perception is infrequent in speech if it occurs at all, and we have exposed an important weakness in a powerful and comprehensive model of speech perception.

NORMALIZATION

A second area of research has to do with normalization processes in speech perception. It is well known, for example, that a given speech segment in a sentence will be interpreted differently depending on the rate of speaking. Thus, for example, Miller and Liberman (1979) showed that the identification of /ba/ versus /wa/ as a function of transition duration also depended on the rate of speaking the sentence. Pisoni, Carrell, and Gans (1983) showed that similar context effects occur with nonspeech. However, the size of the normalization effect appeared to differ in the speech and nonspeech tasks. In one case, the context effects appeared to be significantly larger for the speech stimuli, and in another the context effects appeared to be significantly larger for the nonspeech stimuli. It is incumbent upon the psychoacoustician to show direct correspondences between the nonspeech and speech signals, not simply a rough

approximation. On the other hand, I am not optimistic about this research strategy given the limitations in comparing speech and nonspeech (see discussion in the section Trading Relations). Sine wave analogs of speech heard as speech or nonspeech might provide a better assessment of a psychoacoustic explanation of normalization (see Best et al., 1981).

DUPLEX PERCEPTION

Another arena of research controversy involves duplex perception. In this situation a single stimulus gives rise to two different perceptions: one is speech, and the other is nonspeech. If an isolated formant transition (the chirp) is presented to one ear, while the rest of the speech sound (the base) is presented to the other ear, subjects report hearing both speech and nonspeech sounds lateralized at different locations (Nusbaum, Schwab, & Sawusch, 1983; Rand, 1974). The speech percept must result from the fusion of the two inputs, whereas the nonspeech percept must result from the isolated formant transitions. In addition, different rules seem to describe the perceptual processes involved with the two percepts. The formant transition is perceived differently in the context of being fused with the rest of the speech sound relative to its perception as an isolated, nonspeech sound. In the speech-is-special camp, this result is interpreted as evidence for a specialized process in speech perception (Liberman, 1982; Repp, Milburn, & Ashkenas, 1983). In turn, the psychoacousticians take pains to illustrate that similar processes can occur with nonspeech. As an example, Pastore, Schmuckler, Rosenblum, and Szczesiul (1983) showed that duplex perception also occurs for musical stimuli, which then weakens the argument for specialized processes. The debate continues in this area; for recent papers see Repp (1984) and Nusbaum (1984). Nusbaum (1984) provides a reasonable interpretation of duplex perception. Following the idea of integrating multiple sources of information in speech perception, it is reasonable that both the base and the chirp contribute to perceptual recognition of the speech segment. This interpretation, if formalized within the fuzzy logical model of perception, also accounts for the finding that the contribution of the base should increase as the relevant cue given by the chirp becomes ambiguous.

McGURK EFFECT

An area of research that has captured much of my effort is speech perception by ear and by eye (McGurk & MacDonald, 1976). In this situation, watching a speaker articulate speech influences what the perceiver hears. This result obviously has no psychoacoustic explanation, and this provided much hope for the framework of speech as special. As summarized by one esteemed researcher, "Both motor (speech is special) theorists and Gibsonians went dancing through the streets of every major city in the Eastern U.S. the day in 1976 that McGurk and MacDonald's paper in Nature hit the newsstands." Once again, the problem seems to be that the speech-is-special camp did not consider that alternative explanations are also consistent with the McGurk effect (Massaro & Cohen, 1983b; Massaro, in press e). We might expect that bimodal perception resulting from sight and sound can occur in other situations such as watching and listening to a musician pluck or bow a string on a violin. The visual capture effect in which the sight of an object can attract localization of a sound

source might be used as the counterexample against a speech-is-special interpretation of the McGurk effect.

If not a psychoacoustic explanation of the McGurk effect, a more general physical explanation is still a remote possibility. Perhaps there is some inherent relationship between the mouth configurations and the sound configurations, independent of their representation of vocalization. Kuhl and Meltzoff (1982, 1984) found that five-month-old infants recognized cross-modal correspondences of the vowels /i/ and /a/. The infants viewed a film showing two side-by-side images of a talker articulating /i/ and the same talker articulating /a/, in synchrony, with one of the two vowel sounds. The infants looked longer at the face matching the sound than at the nonmatching face. To test the physical explanation, Kuhl and Meltzoff (1984) used pure tone analogs of /a/ and /i/. There was a complete reversal of the finding with speech; infants now looked longer at the articulation of /i/ when the pure tone analog of /a/ was played, and analogously for the pure tone analog of /i/ (Kuhl & Meltzoff, 1984). Subject to the limitation of a speech-nonspeech comparison, the experiment offers little hope for the adequacy of physics (analogous to the inadequacy of pure psychoacoustics) as an explanation of bimodal speech perception.

TRADING RELATIONS

The final area of research has to do with what are called trading relations in the perceptual categorization of speech. Multiple cues influence a perceptual discrimination, and these cues can be traded off for one another. As an example, voice-onset time can be traded off against the first formant frequency and transition at the onset of voicing in the identification of the voicing of initial stops and fricatives (Massaro & Cohen, 1976, 1977; Summerfield & Haggard, 1974). Trading relations have been interpreted by the speech-is-special camp as meaning that articulatory processes must intervene in the integration of these diverse cues. Once again, trading relations can be found in other domains in addition to speech. Oden (1981), for example, has shown that visual properties about cups and bowls are evaluated and integrated as predicted by the fuzzy logical model, the same model with a history of success in the domain of speech perception.

The term "trading relations" is incomplete and possibly misleading to describe the contribution of multiple sources of information in speech perception. It might seem reasonable to say that voice onset time trades off with the frequency of the first formant at the onset of voicing. But it seems unreasonable to use trading relations to describe the contributions of lexical information and voice onset time to the perception of voicing (Ganong, 1980; Massaro & Oden, 1980b). In the latter case, it is more obvious that there are multiple sources of information contributing to the perceptual interpretation of the message. The same is true in the former case, and it is necessary for a theory to describe how the sources of information are evaluated and integrated to give the tradeoff that is observed.

The historical study of this problem in speech perception may also be of interest. Early workers at Haskins Laboratories manipulated multiple cues to perceptual categorization but did not assess how the

cues were integrated (Hoffman, 1958). Stevens and Klatt (1974) varied both voice- onset time and the onset of the first formant but did not manipulate these in a complete factorial design (see also Sawusch & Pisoni, 1974). Massaro and Cohen (1976) manipulated two cues to voicing in a factorial design and tested mathematical models of their integration. Sometime afterwards, workers at Haskins Laboratories began using factorial designs and studying trading relations and arguing that these depict a special speech processer (Repp 1977, 1982).

Analogous to the nonspeech studies of categorical perception, we are now witnessing a flurry of experiments purportedly illustrating trading relations with nonspeech (Diehl, this volume). Several do not make direct comparisons between speech and nonspeech, precluding a direct comparison between the two. Unless a direct comparison is provided, we have no measure of whether the trading relations are the same in speech and nonspeech domains. For those studies involving a direct comparison, some kind of model is necessary to evaluate the similarities and differences between speech and nonspeech situations. Our fuzzy logical model of perception (Massaro, in press a) permits direct assessments of the information value of each property involved in the trading relation and the nature of the integration process generating the trading relation. Both of these questions are fundamental to assessing any psychoacoustical bases for trading relations.

Diehl (this volume) assessed the psychoacoustic basis for the tradeoff of vowel duration and closure duration of consonants in the perception of voicing of medial stop consonants. Square wave analogs were created by replacing the formants of the speech with square waves. Subjects judged the speech syllables as the voiced or voiceless alternatives and judged the nonspeech as having a short or long silent period (closure) in the middle of the sound. Figure 4 gives the average results for the speech and nonspeech continua. Although superficially the results appear to be comparable, a fine-grained analysis reveals large differences between the speech and nonspeech effects. This difference can be highlighted by fitting the results with two very different models: the FLMP and a weighted averaging model. These models make different predictions about the joint effect of two cues. The FLMP predicts that the contribution of one cue increases with increases in the ambiguity of the other cue, leading to a statistical interaction, as given in the top panel in Figure 4. The weighted averaging model, on the other hand, predicts additive effects and thus parallel curves similar to those shown in the bottom panel of Figure 4.

Quantitative descriptions of the results reinforce this graphical analysis. The FLMP gave a better description of the speech results than did the weighted averaging model, and the opposite outcome emerged for the nonspeech results. A fine-grained analysis appears to reveal important differences between speech and nonspeech analogs. Integrating multiple sources of information appears to be a psychological rather than a psychoacoustic phenomenon.

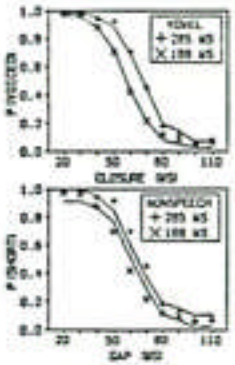

FIGURE 4. Top panel. Proportion of voicing judgements as a function of closure duration and vowel duration. The lines give the predictions of the fuzzy logical model of perception. Bottom panel. Proportion of short judgements as a function of silent gap and the duration of the nonspeech preceding the gap for square-wave analogs of the speech syllables. The lines give the predictions for a weighted averaging of the two properties of the nonspeech (results from Diehl et al., this volume).

Additivity also appears for speech when the two properties function psychoacoustically as illustrated in van Heuven's (this volume) study. Vowel decay time and friction-rise time should have comparable acoustic effects in a vowel-consonant sequence. Both properties contribute to perceived silence between the vowel and consonant. The silent interval cues the difference between the affricate chop and the fricative shop, with longer intervals cueing the affricate. In sequences of say chop/shop, the effects of vowel decay time and friction rise time appeared to have additive effects on the categorization of the affricative-fricative. Increasing the decay time of the vowel and increasing the rise time of the consonant both increased the likelihood of an affricate categorization, supposedly by increasing the perceived silence between the vowel and consonant.

The psychoacoustical basis of trading relations has been assessed using nonhuman subjects, again repeating a strategy used in the categorical-perception controversy (Kuhl & Miller, 1975, 1978; Kuhl & Padden, 1982, 1983; Waters & Wilson, 1976).

Could spacing be improved

The surprising finding of nearly equivalent voice-onset-boundaries as a function of place of articulation for chinchillas and humans could have been a result of the synthetic speech stimuli that were used. The stimuli of Lisker and Abramson (1970) allowed other properties of the stimulus to vary with changes in the speech variable of interest. In these stimuli, the duration of the first format transition covaried with differences in the higher formants that were varied to cue place of articulation. This property could have been responsible for the differences in the VOT boundary as a function of place of articulation. Later versions of synthesized speech without this property do not replicate the large differences that were originally found (Miller, 1977, Massaro & Oden, 1980a, Oden & Massaro 1978). Hence, the original trading relation that was observed appears to have a psychoacoustic rather than a psychological explanation.

Rejecting the psychoacoustical basis for trading relations by showing differences between speech and nonspeech is perhaps a losing battle. The reason, analogous to studies of categorical perception, is the persistent possibility that an inappropriate nonspeech stimulus was used. Workers at Haskins Laboratories have solved this problem to some extent by using sine-wave analogs of speech and testing subjects in both speech and nonspeech modes of perceiving these signals (Best, et al., 1981). They have succeeded in demonstrating large differences between the two modes of perceiving, but without a formal analysis we don't know enough about how the two modes differ. The Best, et al. (1981) results appear similar to those of Diehl in that nonspeech integration appears to be more additive relative to the multiplicative integration of speech. When the signals are heard as nonspeech, subjects might use only one of the two varying properties to categorize the sounds. When they use two properties, they appear to integrate the two nonspeech cues according to a weighted- averaging rule. When the signals are heard as speech, however, the two cues are integrated in such a way that the least ambiguous cue has the greatest impact on the decision.

Diehl (this volume) claims that the integration of closure duration and closure pulsing has an auditory basis, perhaps the glottal pulsing biasing the observer to hear the interval as short. However, the actual relationship between the amplitude of voicing during the closure and the closure interval for voiced stops differs for different languages (Lindau & Ladefoged, 1986). It seems that these two variables are independently variable in articulation and different languages combine them differently to convey voicing information. This arbitrariness between the two dimensions precludes any psychoacoustic explanation of the integration of the two cues. In our framework, the integration of the two cues will reflect how these cues are used by the speaker to inform the listener.

Psychophysical explanations should be able to illuminate the contrasts between short and long vowels. Two vowels that have highly similar formant frequencies tend to differ from one another in duration. If psychoacoustics has any relevance, the relationship between formant structure and duration should be systematic. The normally longer vowel should be heard as longer relative to the normally shorter vowel, even though presented at the same duration. For example, a vowel with a high F2 might be heard as longer than a vowel with a low F2. In this case, duration might acquire cue status, as suggested by Stevens, Keyser, and Kawasaki (1986). Consider the three pairs /i/-/I/, /u/-/U/, and /ʌ/ - /a/ and their first two formants and average durations shown in Table 1.

For two of the pairs, F1 is lower in frequency for the longer member and higher for the longer vowel of the third pair. The frequency of F2 is higher for the longer member of two of the pairs and lower for the longer member of the third pair. Thus there is no systematic relationship between formant frequency and relative duration. Clearly it is unlikely that the integration of these two characteristics results from some low- level auditory interaction.

Table 1. Average values of first and second formants and the durations of six vowels (from Peterson & Barney, 1952, and Peterson & Lehiste, 1960)

	/i/	/I/	/u/	/U/	/ʌ/	/ɛ/
F_1(Hz)	270	390	300	440	660	530
F_2(Hz)	2290	1990	870	1020	1720	1840
Duration (ms)	240	180	260	200	330	200

Having observed limitations with nonspeech and nonhuman subjects, I offer in their place developmental and cross-linguistic comparisons with respect to trading relations. In an early observation, Simon and Fourcin (1978) found differences in the contribution of F1 to perceived voicing of initial stop consonants as a function of development and language. Onset frequency of F1 is more informative for voicing in English relative to French and interestingly French children acquire this cue sometime later than their English cohorts. If this result is reliable, it goes well beyond what could be predicted by a psychoacoustic interpretation. If this psychoacoustic interpretation is broadened to include an important contribution of perceptual learning, however, it can postdict the results. While achieving this worthwhile goal it would also become much less like its parsimonious predecessor and more like the speech-is-special (Liberman 1982) and information-integration (Massaro & Oden, 1980b) viewpoints.

Cross-linguistic comparisons offer a direct assessment of the psychoacoustic basis of trading relations. If a trading relation between two properties of the speech signal exists for subjects of one language but not for subjects of another, there is little merit to a psychoacoustic explanation. The results would be compatible with the integration of the two properties in one case and not in the other. This outcome would supposedly occur when a given property is ecologically valid in one language but not in the other.

A recent cross-linguistic study provides a simple rejection of the psychoacoustical-basis hypothesis. The contrast of interest was postvocalic voicing as cued by the duration of the preceding vowel and the duration of the consonant (Denes, 1955). Like the two pronunciations of use, the words peas and peace differ in the voicing of the final consonant. Since Denes (1955), we have known that both vowel duration and the aperiodic fricative duration contribute to this distinction in English (Derr & Massaro, 1980; Massaro & Cohen, 1976, 1977; Raphael, 1972). Flege and Hillenbrand (1986) observed that the /s/-/z/ distinction is not learned in Swedish and Finnish, since neither language possesses a /z/ phoneme. If a psychoacoustic basis existed for the trading relation between vowel and consonant durations in English, then learning this new distinction should be an easy charge for Swedish and Finnish speakers acquiring English. However, this distinction is difficult to learn and the question is whether the cues used by English listeners are learned by these individuals acquiring English as a second language. Both experienced and inexperienced

listeners were tested on the peas-peace contrast. Five durations of the periodic vowel were factorially combined with five durations of the aperiodic frication and subjects were instructed in English to categorize the words as peas or peace.

The results revealed that the Swedish and Finnish listeners did not use fricative duration as a cue to the /s/-/z/ contrast in English, but based their identification decision on only vowel duration. Although this result generates a variety of interesting questions, for our purposes, it weakens a simple psychoacoustic explanation of the trading relation in English. There is nothing inherent in the auditory resolution of the English syllables that leads to the tradeoff between vowel duration and consonant duration. If there were, the identification results for the Swedish and Finnish subjects should have been identical to those for the English speakers. The trading relation exists because the English listener integrates these two cues in perceptual recognition, not because the auditory system naturally categorizes short vowel-long consonant syllables as one class and long vowel-short consonant syllables as another class. The evaluation of vowel duration and consonant duration appear to occur relatively independently of one another, as assumed in our fuzzy logical model (Derr & Massaro, 1980; Massaro & Cohen, 1977, 1983b). They are integrated by native English speakers because both are informative about the identity of voicing of the final consonant.

The observation that Swedish and Finnish speakers are not influenced by consonant duration also weakens the appeal of consonant/vowel (C/V) ratio as the cue to voicing (Port & Dalby, 1982). If C/V ratio were used, consonant duration would necessarily have an influence on perceptual categorization. Conceptualizing vowel duration and consonant duration as independent sources of information about voicing, however, describes the results parsimoniously (Massaro & Cohen, 1983b). Consonant duration does not acquire cue value for the Swedish and Finnish speakers because their language does not have a /z/ phoneme. Learning English as a second language does not seem to change this situation, possibly because these individuals continue to speak their native language. In this case, the cue value of various sources of information is not easily normalized to take into account the language currently being perceived.

Remaining questions are why the Swedish and Finnish listeners used vowel duration as a cue and whether they use consonant duration as a cue for other contrasts such as stop consonants in medial position. Vowel duration would be functional for stops in final position (Raphael, 1972) and this could have generalized to the new /z/-/s/ contrast in English.

We have parallel results comparing Chinese and English subjects on the perception of a vowel contrast (i)-(y) that exists in Chinese but not English (Massaro, Tseng, & Cohen, 1983). The English subjects had no experience with Chinese. The vowels differ not only in their formant pattern but also in loudness (Fant, 1973) in that (i) tends to be louder than (y). Chinese listeners should know this (at a procedural not necessarily a declarative level), but English listeners should not. Massaro et al. (1983) utilized this logic in comparing Chinese and English speakers on the contribution of F0 pattern to identification of Chinese vowels. Given that loudness has no ecological validity in

distinguishing English vowels, the English subjects can be conceptualized as serving as a "chinchilla" control group. Thus, Chinese but not English should hear a louder vowel as more like (i) than like (y). To test this hypothesis, five levels of formant structure between (i) and (y) were factorially combined with 3 amplitude levels, producing a total of 15 syllables. Both the Chinese and English subjects were simply instructed to identify each syllable as (i) or (y). The results provide evidence for a psychological integration of formant structure and amplitude for Chinese listeners. It is not a psychoacoustic integration because the English subjects are not influenced by amplitude even though they use formant structure in the same manner as the Chinese subjects. Figure 5 shows that the Chinese reveal a significantly larger effect of amplitude when the formant structure is ambiguous, exactly as predicted by the fuzzy logical model.

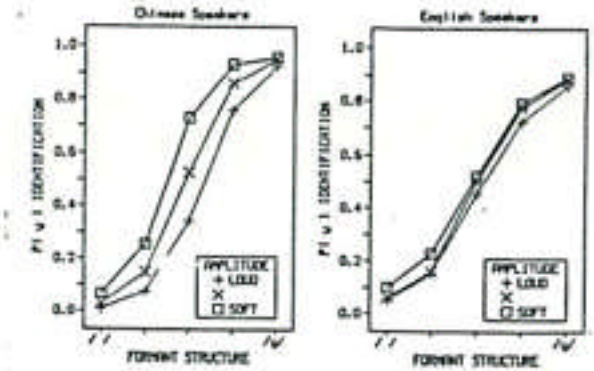

FIGURE 5. Proportion of /y/ identification for Chinese and English subjects as a function of the Formant Structure of the test syllable; the amplitude of the syllable is the curve parameter.

The English speakers show no hint of a similar effect, even though they discriminate the formant structure as well as the Chinese. The small effect of amplitude on identification for the English speakers was not statistically significant, nor, was the interaction of formant structure and amplitude.

In summary, the contribution of multiple sources of information appears to be a fundamental characteristic of speech perception. These sources of information differ for different languages and, therefore, the integration of the sources is not easily accounted for by psychoacoustic principles. The situation is better conceptualized as pattern recognition in which multiple sources of information are brought to bear on a decision. The FLMP provides a good quantitative description of the integration of the multiple sources of information across a variety of speech contrasts. The model not only describes the integration of acoustic sources but also their integration

with visible speech and with phonological, lexical, syntactic, and semantic sources of information (Glucksberg, Kreuz, & Rho, 1986; Massaro, in press c).

REFERENCES

1. Anderson, N.H. (1981). Foundations of information integration theory. New York: Academic.
2. Anderson, N.H. (1982). Methods of information integration theory. New York: Academic.
3. Best, C.T., Morrongiello, B. and Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception and Psychophysics, 29, 191-211.
4. Brunswik, E. (1952). The conceptual framework of psychology. Chicago: University of Chicago Press.
5. Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. Chicago: University of Chicago Press.
6. Brunswik, E. (1956). Perception and the Representative Design of Psychological Experiments. Berkeley, CA: University of California Press.
7 Cutting, J.E. and Rosner, B.S. (1974). Categories and boundaries in speech and music. Perception & Psychophysics, 16, 564-570.
8. Denes, P. (1955). Effects of duration on the perception of voicing. Journal of the Acoustical Society of America, 27, 761-764.
9. Derr, M.D. and Massaro, D.W. (1980). The contribution of vowel duration, F0 contour, and frication duration as cues to the /juz/jus/ distinction. Perception and Psychophysics, 27, 51-59.
10. Eimas, P.D. (1985). The perception of speech in early infancy. Scientific American, 252, no. 1, 46-52.
11. Eimas, P.D. and Corbit, J.D. (1973). Selective adaptation of linguistic feature detectors. Cognitive Psychology, 4, 99-109.
12. Eimas, P.D. and Miller, J.L. (1980). Contextual effects in infant speech perception. Science, 209, 1140-1141.
13. Fant, G. (1973). Speech sounds and features. Cambridge, MA: MIT Press.
14. Flege, J.E. and Hillenbrand, J. (1986). Differential use of temporal cues to the /s/-/z/ contrast by native and non-native speakers of English. Journal of the Acoustical Society of America, 79, 508-517.
15. Fodor, J.A. (1983). Modularity of Mind. Cambridge, Mass.: Bradford Books.
16. Fujisaki, H. and Kawashima, T. (1970). Some experiments on speech perception and a model for the perceptual mechanism. Annual Report of the Engineering Research Institute. University of Tokyo, 29, 206-214.
17. Ganong, W.F. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6, 110-125.
18. Glucksberg, S., Kreuz, R.J. and Rho, S.H. (1986). Context can constrain lexical access: Implications for models of language comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, ~~no.4~~, 323-335.
19. Hary, J.M. and Massaro, D.W. (1982). Categorical results do not imply categorical perception. Perception and Psychophysics, 32, 409-418.
20. Hirsh, I.J. (1959). Auditory perception of temporal order. Journal of the Acoustical Society of America, 31, 759-767.

21. Hoffman, H.S. (1958). Studies of some cues in the perception of the voiced stop consonants. Journal of the Acoustical Society of America, 33, 1035-1041.
22. Howell, P. and Rosen, S. (1983). Natural auditory sensitivities as universal determiners of phonemic contrasts. Linguistics, 21, 205- 235.
23. Kuhl, P.K. and Meltzoff, A.N. (1982). The bimodal perception of speech in infancy. Science, 218, 1138-1141.
24. Kuhl, P.K. and Meltzoff, A.N. (1984). Infants' recognition of cross- modal correspondences for speech: Is it based on physics or phonetics? Journal of the Acoustical Society of America, 76, Suppl. 1, S80(A).
25. Kuhl, P.K. and Miller, J.D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar-plosive consonants. Science 190, 69-72.
26. Kuhl, P.K. and Miller, J.D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. Journal of the Acoustical Society of America, 63 905-917.
27. Kuhl, P.K. and Padden, D.M. (1982). Enhanced discrimination at the phonetic boundaries for the voicing feature in macaques. Perception and Psychophysics, 32 542-550.
28. Kuhl, P.K. and Padden, D.M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. Journal of the Acoustical Society of America, 73, 1003-1010.
29. Liberman, A.M. (1982). On finding that speech is special. American Psychologist, 37, 148-167.
30. Liberman, A.M., Cooper, F.S., Shankweiler, D.P. and Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431-461.
31. Liberman, A.M. and Mattingly, I.G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.
32. Lindau, M. and Ladefoged, P. (1986). Variability of feature specifications. In J.S. Perkell and D.H. Klatt (Eds.), Invariance and Variability in Speech Processes. Hillsdale, NJ: Lawrence Erlbaum Associates, 464-478.
33. Lisker, L. and Abramson, A. (1970). The voicing dimension: Some experiments in comparitive phonetics. Proceedings of the Sixth International Congress of Phonetic Sciences, Prague, 1967, (Academia, Prague), 563-567.
34. Macmillan, N.A., Kaplan, H.L., and Creelman, C.D. (1977). The psychophysics of categorical perception. Psychological Review, 84, 452-471.
35. Massaro, D.W. (1972). Preperceptual images, processing time and perceptual units in auditory perception. Psychological Review, 79, 124-145.
36. Massaro, D.W. (1979). Reading and listening. In P.A. Kolers, M. Wrolstad, and H. Bouma (Eds.), Processing of Visible Language 1, New York: Plenum. 331-354.
37. Massaro, D.W. (1984). Building and testing models of reading processes. In P.D. Pearson (Ed.), Handbook of Reading Research. New York: Longman. 111-146.
38. Massaro, D.W. (in press a). A fuzzy logical model of speech perception. In W.A. Lea (Ed.), Towards Robustness in Speech Recognition. Apple Valley, Minnesota: Speech Science Publications.
39. Massaro, D.W. (in press b.). Categorical partition: A fuzzy logical model of categorization behavior. In S. Harnad (Ed.), Categorical Perception. New York.

40. Massaro, D.W. (in press c.). Information-processing theory and strong inference: A paradigm for psychological inquiry. In H. Heuer and A.F. Sanders (Eds.), Perspectives on Perception and Action. Hillsdale, N.J.: Erlbaum.
41. Massaro, D.W. (in press d.). Integration multiple sources of information in listening and reading. In D.A. Allport, D.G. MacKay, W. Prinz, and E. Scheerer (Eds.), Language Perception and Production: Shared Mechanisms in Listening, Speaking, Reading and Writing. Academic Press.
42. Massaro,D.W. (in press e). Speech perception by ear and eye. In B. Dodd and R. Campbell (Eds.), Hearing by Eye: Experimental studies in the psychology of lipreading. Hillsdale, N.J.: Erlbaum.
43. Massaro, D.W. and Cohen, M.M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. Journal of the Acoustical Society of America, 60, 704-717.
44. Massaro, D.W. and Cohen, M.M. (1977). Voice onset time and fundamental frequency as cues to the /zi/-/si/ distinction. Perception and Psychophysics, 22 373-382.
45. Massaro, D.W. and Cohen, M.M. (1983a). Categorical or continuous speech perception: A new test. Speech Communication, 2, 15-35.
46. Massaro, D.W. and Cohen, M.M. (1983b). Consonant/vowel ratio: An improbable cue in speech. Perception and Psychophysics, 33, 501-505.
47. Massaro, D.W. and Cohen, M.M. (1983c). Evaluation and integration of visual and auditory information in speech perception. Journal of Experimental Psychology: Human Perception and Performance, 9, 753- 771.
48. Massaro, D.W. and Cohen, M.M. (1983d). Phonological context in speech perception. Perception and Psychophysics, 34, 338-348.
49. Massaro, D.W. and Oden, G.C. (1980a.). Evaluation and integration of acoustic features in speech perception. Journal of the Acoustical Society of America, 67, 996-1013.
50. Massaro, D.W. and Oden, G.C. (1980b.). Speech perception: A framework for research and theory. In N.J. Lass (Ed.), Speech and Language: Advances in Basic Research and Practice, 3, New York: Academic Press. 129-165.
51. Massaro, D.W., Tseng, C.Y., and Cohen, M.M. (1982). Vowel and lexical tone perception in Mandarin Chinese: Psycholinguistic and psychoacoustic contributions. Quantative Linguistics, 19, 76-102.
52. McClelland, J.L. & Elman, J.L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18, 1-86.
53. McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.
55. Miller, J.D. (1977). Nonidependence of feature processing in initial consonants. Journal of Speech and Hearing Research, 20, 510-518.
56. Miller, J.L. and Liberman, A.M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. Perception and Psychophysics, 25, 457-465.
57. Nusbaum, H.C. (1984). Possible mechanisms of duplex perception: "chirp" identification versus dichotic fusion. Perception and Psychophysics, 35, 94-101.
58. Nusbaum, H.C., Schwab, E.C., and Sawusch, J.R. (1983). The role of "chirp" identification in duplex perception. Perception and Psychophysics, 33, 323-332.
59. Oden, G.C. (1981). Fuzzy propositional model of concept structure and use: A case study in object identification. In G.W. Lasker (Ed.).

Applied Systems Research and Cybernetics. Elmsford, NY: Pergamon Press.
60. Oden, G.C. and Massaro, D.W. (1973). Integration of featural information in speech perception. Psychological Review 85, 172-191.
61. Paap, K.R. (1975). Theories of speech perception. In D.W. Massaro (Ed.), Understanding Language: An Information Processing Analysis of Speech Perception, Reading and Psycholinguistics. New York: Academic Press. 151-204.
62. Pastore, R.E., Ahroon, W.A., Baffuto, K.J., Friedman, C., Puleo, J.S., and Fink, E.A. (1977). Common-factor model of categorical perception. Journal of Experimental Psychology: Human Perception and erformance, 3, 686-696.
63. Pastore, R.E. Schmuckler, M.A., Rosenblum, L., and Szczesiul, R. (1983). Duplex perception with musical stimuli. Perception and Psychophysics, 33, 469-474.
64. Patterson, J.H. and Green, D.M. (1970). Discrimination of transient signals having identical energy. Journal of the Acoustical Society of America, 48, 894-905.
65. Peterson, J.H. and Barney, H.L. (1952). Control Methods used in a study of the vowels. Journal of the Acoustical Society of America, 24, 175-184.
66. Peterson, G.E. and Lehiste, I. (1960). Duration of syllable nuclei in English. Journal of the Acoustical Society of America, 32, 693-703.
67. Pisoni, D.B. (1977). Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 61, 1352-1361.
68. Pisoni, D.B., Carrell, T.D., and Gans, S.J. (1983). Perception of the duration of rapid spectrum changes: Evidence for context effects with speech and nonspeech signals. Perception and Psychophysics, 34, 314- 322.
69. Port, R.F. and Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. Perception and Psychophysics, 32, 141-152.
70. Rand, T.C. (1974). Dichotic release from masking for speech. Journal of the Acoustical Society of America, 55, 678-680.
71. Raphael, L.J. (1972). Preceding vowel duration as a cue to the voicing of the voicing characteristic of word-final consonants in American English. Journal of the Acoustical Society of America, 51, 1296-1303.
72. Repp, B.H. (1977). Interdependence of voicing and place decisions. Haskins Labs, New Haven CT, September 1977 (unpublished).
73. Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.
74. Repp, B.H. (1984). Against a role of "chirp" identification in duplex perception. Perception and Psychophysics, 35, 89-93.
75. Repp, B.H., Milburn, C., and Ashkenas, J. (1983). Duplex perception: Confirmation of fusion. Perception and Psychophysics, 33, 333-337.
76. Sawusch, J.R. and Pisoni, D.B. (1974). On the identification of place and voicing features in synthetic stop consonants. Journal of Phonetics, 2, 181-194.
77. Simon, C. and Fourcin, A.J. (1978). Cross-language study of speech- pattern learning. Journal of the Acoustical Society of America, 63, 925-935.
78. Stevens, K.N., Keyser, S.J., and Kawasaki, H. (1986). Toward a phonetic and phonological theory of redundant features. In J.S. Perkell

and D.H. Klatt (Eds.), Invariance and Variability in Speech Processes. Hillsdale, NJ: Lawrence Erlbaum Associates. 426-449.
79. Stevens, K.N. and Klatt, D.H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. Journal of the Acoustical Society of America, 55, 653-659.
80. Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., and Cooper, F.S. (1970). The motor theory of speech perception: A reply to Lane's critical review. Psychological review, 77, 234-249.
81. Summerfield, A.Q. (1982). Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops. Journal of the Acoustical Society of America, 72, 51-61.
82. Summerfield, A.Q. and Haggard, M.P. (1974). Perceptual processing of multiple cues and contexts: Effects of following vowel on stop consonant voicing. Journal of Phonetics, 2, 279-295.
83. Warren, R.M. (1974a). Auditory pattern discrimination by untrained listeners. Perception and Psychophysics, 15, 495-500.
84. Warren, R.M. (1974b). Auditory temporal discrimination by trained listeners. Cognitive Psychology, 6, 237-256.
85. Waters, R.S. and Wilson, W.A. Jr. (1976). Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants. Perception and Psychophysics, 19, 285-289.
86. Winitz, H., La Riviere, C., and Herriman, E. (1975). Variations in VOT for English initial stops. Journal of Phonetics, 3, 41-52.