Published in final edited form as:
Cognition. 1981 ; 10(1-3): 249–259.

Some current theoretical issues in speech perception

David B. Pisoni*
Indian University

The field of speech perception is a very rich interdisciplinary area involving researchers from a number of disciplines including psychology, linguistics, speech and hearing science, electrical engineering and artificial intelligence. AS a consequence, the particular issues under study are often approached in different ways by different research groups. Despite these superficial differences, however, there are a small number of basic questions that can be identified as the core of problems that people are working on today. In my view, the fundamental problem is speech perception is to describe how the listener converts the continuously varying acoustic stimulus produced by a speaker into a sequence of discrete linguistic units and how he recovers the intended message. This problem can be broken down into a number of more specific sub-questions. For example, what stages of perceptual analysis intervene between presentation of a speech signal and eventual understanding of the message? And, what types of operations occur at each of these stages? What specific types of perceptual mechanisms are involved in speech perception and how are they used in understanding spoken language?

In understanding spoken language we assume that various types of information are computed by the speech processing mechanisms. Some forms of information are transient, and last for only a short period of time; other forms of information are more durable, and interact with other sources of information from long-term memory (see Pisoni, 1978). The nature of these perceptual codes, and their interactions are among the major concerns in the field of speech perception today. In the remainder of this paper I will briefly review what I see as the major theoretical issues in speech perception. Some of these have been discussed in the past and continue to occupy a central role in speech perception research today; others relate to new problems to be pursued in the future.

Linearity, lack of invariance anal the segmentation problem

For more than thirty years, it has been extremely difficult to identify acoustic segments and features which uniquely match the perceived phonemes independently of the surrounding context. Often a single acoustic segment contains information about several neighboring linguistic segments (i.e., parallel transmission), and, conversely, the same linguistic segment is often represented acoustically in quite different ways depending on the surrounding phonetic context, rate of speaking and talker (i.e., context conditioned variation). In addition, the acoustic characteristics of individual speech sounds and words exhibit even greater variability in fluent speech because of the influence of the surrounding context.

The context-conditioned variability in the correspondence between acoustic signal and phoneme also presents enormous problems for segmentation of speech into phonemes or even words. Because of the failure to meet the linearity and invariance conditions, it has

*Reprint requests should be sent to: David B. Pisoni, Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405, U.S.A.

been difficult to segment speech into acoustically defined units that are independent of adjacent segments or are free from contextual effects of sentence contexts. It has been difficult to determine strictly by simple physical criteria where one word ends and another begins, especially in connected speech. Although segmentation is possible according to strictly acoustic criteria (see Fant, 1962), the number of acoustic segments is typically greater than the number of phonemes in the utterance and, moreover, no *simple* invariant mapping has been found between these acoustic attributes and perceived phonemes or individual words. Continued attention will no doubt be directed at describing the decoding strategies used by listeners in interpreting context-dependent acoustic patterns that correspond to phonemes in words. Other research will be concerned with how these cues are modified by the acoustic characteristics of sentences in fluent speech.

Internal representation of speech signals

Until very recently there was fairly good agreement among investigators working on human speech perception that at some stage of perceptual processing speech is represented internally as a sequence of discrete segments and features (see Studdert-Kennedy, 1974, 1976). There was less agreement, how ever, about the exact description of these features. Arguments were provided for feature systems based on distinctions in the acoustic domain, the articulatory domain, and systems which combine both types of distinctions.

Recently there has been a trend to view these sorts of traditional feature descriptions of speech sounds with some skepticism, particularly with regard to the role these features play in on-going perception (Ganong, 1979; Klatt, 1977, 1979; Parker, 1977). On reexamination, much of the original evidence cited in support of feature-based processing in perceptual experiments seems ambiguous and equally consistent with more parametric representations of speech. As a consequence, numerous investigators have begun to look more closely at how speech waveforms are processed in the peripheral auditory system, and what those more detailed representations may contribute to questions surrounding phonetic processing and particularly the problem of acoustic-phonetic invariance (Klatt, 1979; Searle *et al*, 1979; Zwicker *et al* 1979).

One of these approaches has been to represent speech in the frequency domain as a sequence of magnitude spectra sampled about every 10 ms. or so. If these spectral samples are adjusted to take account of certain psychophysical facts about hearing (such as critical bands, spread of masking and the growth of loudness), a continuous representation can be obtained that is similar to a ‘neural spectrogram’. Efforts along these lines have directed attention to the problem of the neural representation of speech, and to questions having to do with psychological filtering. Very little is currently known about the way speech is represented at these early stages of processing.

There are, of course, numerous different ways of representing the *perceptual* dimensions of speech signals and these can be examined with regard to questions dealing with how a listener identifies the acoustic cues to various segmental contrasts. Moreover, there is the additional and quite separate problem of defining the proper psychological dimensions that can be used to characterize appropriate displays of these dimensions. The peripheral auditory system can be modelled as a frequency analyzer with continously varying neural signals as output. Thus, the perceptual dimensions and the ‘neural’ representations of speech should be based on what is currently known about the function of the peripheral auditory system and how it processes various types of acoustic signals.

Normalization problem

In addition to the problems arising from the lack of acoustic-phonetic invariance discussed earlier, there are also two distinct problems having to do with normalization of the speech signal. One is the talker-normalization problem. Talkers differ in the length and shape of their vocal tracts, in the articulatory gestures used for producing various types of phonetic segments, and in the types of coarticulatory strategies present in their speech. As a consequence, there are very substantial differences among talkers in the absolute values of the acoustic correlates of many phonetic features. Differences in stress and speaking rate as well as in dialect and affect also contribute to differences in the acoustic manifestation of speech. Clearly, the invariant properties cannot be absolute physical values encoded in the stimulus but instead must be relational in nature. Unfortunately, there is relatively little known at present about this form of perceptual normalization or about the types of mechanisms involved (however, see Remez *et al.*, 1981).

The second problem concerns time and rate normalization. It is now well known that the durations of individual segments are influenced quite substantially by an individual talker's speaking rate. However, the acoustic *durations* of segments are also affected by the locations of various syntactic boundaries in connected speech, by syllabic stress, and by the component features of adjacent segments in words ((see for example Gaitenby, 1965; Klatt, 1975, 1976, 1979; Lehiste, 1970). In addition, there are substantial differences in the durations of segments of words when produced in sentence contexts compared to the same words spoken in isolation. The rate at which speakers talk also influences the duration and acoustic correlates of various phonetic features and segments. Numerous low-level phonetic and phonological effects such as vowel reduction, deletion and various types of assimilation phenomena have been well documented in the literature. These effects are influenced a great deal by speaking tempo, dialect and surrounding phonetic context.

It has also been known for many years that duration can also be used to distinguish various segmental contrasts. Many phonetic and phonological segmental contrasts are also distinguished by redundant differences in duration as well as by their primary spectral correlates. Thus, the listener is faced with the problem of trying to ignore certain kinds of irrelevant durational information while trying to incorporate other kinds of distinctive durational information about segments, stress, prosody and syntactic structure (see Miller, 1981; Port, 1981 for reviews).

Units in speech perception

Another important and long-standing issue in speech perception is the choice of a minimal unit of perceptual analysis. Because of limitations of channel capacity, especially in the auditory system, raw sensory information must be recoded into some more permanent form that can be used for subsequent analysis. Is there a basic or 'natural' coding unit for speech perception? Many investigators have argued for the primacy of the feature, phoneme, syllable or word as their candidate for the basic perceptual unit. Other investigators have even proposed larger units for perceptual analysis of speech, such as clauses of sentences (Bever, Lackner and Kirk, 1969; Miller, 1962). The debate over the choice of a perceptual unit can be resolved if a strict distinction were made concerning the level of linguistic analysis under consideration. The size of the processing unit in speech perception varies from feature to segment to clause as the level of linguistic processing changes, and arguments over the question of whether there is one basic or primary unit are inappropriate since there are, in fact, many units that are used by the speech processing mechanisms.

Prosody, rhythm and speech timing

Most of the research in speech perception over the last thirty years, as well as the major theoretical emphasis, has been concerned with segmental analysis of phonemes. One seriously neglected topic has been the prosodic or supra-segmental attributes of speech, which involve differences in pitch, intensity, duration, and the timing of segments and words in sentences. At present there remains a wide gap between the research on isolated segments and features and prosodic factors (see Cohen and Nooteboom, 1975). It is clear, however, that this source of linguistic information serves to link phonetic segments, features and words to grammatical processes at higher levels of analysis (see Darwin, 1975; Huggins, 1972; Nooteboom *et al.*, 1978 for reviews). Moreover, speech prosody may also carry useful information about lexical, syntactic and semantic properties of the speaker's message. It would be of interest to know, for example, the extent to which syntactic and semantic variables influence the durations of phonetic segments and words, and whether listeners can and do use this sort of information in understanding spoken language (see Huggins, 1972, 1978; Klatt and Cooper, 1975).

Lexical access and word recognition

The problems of word recognition and the nature of lexical representations have been long-standing concerns of cognitive psychologists, although these problems have not been studied extensively by investigators working in the mainstream of speech perception. This is true because the bulk of work on word recognition was concerned with investigating visual processes with less attention directed to questions of spoken word recognition. Moreover, most of the interest in speech perception has been directed toward feature and phoneme perception which typically used isolated nonsense syllables. While such an approach is appropriate for studying 'low level' acoustical analysis of speech, it is not very helpful in dealing with questions surrounding how meaningful words are recognized in isolation or in connected speech.

There are several interesting and important problems in speech perception that touch intimately upon the process of lexical access and bear more directly on the nature of the various types of representations in the mental lexicon. For example, it is of considerable interest to determine precisely what kinds of representations exist in the mental lexicon. Do words, morphemes, phonemes, or sequences of spectral templates represent lexical entries? Is a word accessed on the basis of an acoustic, phonetic or phonological code? Are high frequency words recognized more-or-less automatically by very rapid search through a special precompiled network? Are less frequent words analyzed by general rules for morphological analysis?

One of the central problems in word recognition and lexical access deals with the interaction of sensory input and higher-level contextual information. Some investigators, such as Forster (1976) and Massaro (1977), maintain that early sensory information is processed independently of higher-order context, and that the facilitation effects observed in word recognition are due to post-perceptual processes involving decision criteria. Other investigators such as Morton (1969, 1979), Marslen-Wilson and Welsh (1978), Marslen-Wilson and and Tyler (1980), Cole and Jakimik (1978) and Foss and Blanck (1980) argue that context can, in fact, influence the extent of early sensory analysis of the input signal.

Klatt (1979, 1981) has recently proposed a model of lexical access that explicitly avoids any need to compute a distinct level of representation corresponding to a sequence of discrete phonemes. Instead, he has precompiled an abstract phonetic lexicon of all possible words into a network of sequences of spectral templates. These templates are context-sensitive much like the earlier 'Wickelphones' (Wickelgren, 1976) since they are supposed to

characterize the acoustic correlates of phones in different phonetic environments by encoding the spectral characteristics and transitions from the middle of one phone to the middle of the next. Klatt (1979) argues that this form of diphone concatenation is sufficient to capture much of the context-dependent variability observed for phonetic segments in spoken words. Much remains to be done to access these claims as valid psychological descriptions of the representation of words in the mental lexicon.

Phonetic and phonological recoding of words in sentences

One of the major difficulties encountered in speech perception is that each utterance of a language can be realized phonetically in many different ways. Obviously, it is unrealistic to store every possible utterance of the language in long-term memory since the number of different sentences and phonetic realizations is potentially infinite. While it might be possible to adopt this strategy in the case of machine recognition of speech in very limited context, such a strategy seems inappropriate in the case of human speech perception. In addition to general phonological processes which characterize certain uniform dialect differences in pronunciation among talkers, there are also sets of low-level phonetic implementation rules which characterize more specific acoustic-phonetic variations among individual talkers. Because the number of different phonological phenomena in language is quite large, and because of the idiosyncratic variability of individual talkers, sets of decoding rules have been formulated from careful study of the acoustic and phonetic properties of speech in various contexts. Rules such as these must also be assumed to be part of the perceptual strategies used by human listeners in understanding spoken language.

Despite the long-standing interest in phonological processes by linguists and the importance they play in the acoustic-phonetic realization of spoken language, relatively little perceptual research has been directed toward these problems. With the use of synthesis-by-rule systems, sets of phonological and phonetic implementation rules can be formulated and the effects of variations and modifications in these rules can be studied with isolated words and words in sentence contexts (see Huggins, 1978).

Focused search and ‘islands of reliability’

There can be little doubt after some thirty years of research on speech that the acoustic signal contains a great deal of redundant information. A basic engineering goal has been to try to locate the most important information and code it in the most efficient way for transmission. In the same way, investigators concerned with human speech perception have tried to identify the ‘minimal cues’ for phonemes in the hope that once these could be identified the basic problem of recognition of speech could be solved. Unfortunately, there is a great deal more to speech perception and spoken language understanding than simply discovering the minimal cues for phonemes. The speech signal appears to be rich with salient and reliable information that listeners use in understanding the message. As a consequence, the basic problem becomes one of finding in the stimulus input these ‘islands of reliability’ that can be used to access various different sources of knowledge.

The term *focused search* has been used to characterize the strategies that listeners or intelligent machines use for inspecting the signal for information that can be useful at *any given point* in the perceptual process; focused search also specifically avoids information that does not provide useful support. Examples of such reliable information include: the presence of stressed syllables, the beginnings and ends of words, and the locations of various spectral changes indicating shifts in the source function.

Focused search emphasizes an important problem, namely, to identify those acoustic correlates of the signal that the listener relies on. The scope of a listener’s focused search

strategies varies substantially with the requirements of experimental tasks; what may be salient and reliable acoustic-phonetic information in one listening context may not be used at all in another. Research on this problem has shifted recently from experiments using isolated nonsense syllables which are manipulated in very precise ways to investigations directed at how listeners use these cues to perceive words in isolation and in sentence contexts where several diverse sources of knowledge can be used.

The principle of delayed binding

Human speech perception and spoken language understanding take place very rapidly in real time although relatively little is currently known about the processes and mechanisms that support such on-line activities. A good deal of the speech perception process occurs automatically and is therefore unavailable for direct conscious introspection. Do all decisions at all levels of the speech perception process take place immediately in real-time or are there selected processing delays at particular analytic levels pending additional information? What is the size of the scanning window over which low-level phonetic decisions are made? What depth of processing is required before a final and binding decision can be made about the segmental composition of the input signal? These are questions that are being pursued at this time by a number of researchers.

The ‘principle of delayed binding’ evolved from the ARPA speech understanding project (see Klatt, 1977 for a review). According to this principle, decisions at low levels of processing are not forced if the information is unreliable or insufficient to make a final decision (see also Miller, 1962). Of course, such a principle might be appropriate in computational situations where the front-end or basic acoustic-phonetic recognition device fails to perform as well as humans do. But we know from much of the earlier research on word intelligibility that human listeners can and do make binding low-level segmental and lexical decisions with extremely high accuracy even under very poor listening conditions. After all, if the quality of the acoustic-phonetic information is very good, as in high-quality natural speech, phonetic, lexical and even syntactic decisions can occur on-line quite rapidly. However, in situations where the speech signal is physically degraded or impoverished, the speed of perceptual processing may be substantially slower, and certain low-level decisions may well have to be delayed pending higher-order constraints (Miller, Heise and Lichten, 1951; Miller and Isard, 1963). In future research, it will be important to find out more about the perceptual and interpretative processes in human speech perception that are responsible for the very rapid processing and the seemingly immediate on-line interpretation of spoken language as it is heard.

Conclusion

The bulk of research on speech perception over the last thirty years has been concerned principally, if not almost exclusively, with feature and phoneme perception in isolated contexts using nonsense syllable materials. This research strategy has undoubtedly been pursued because very substantial problems arise when one deals with issues such as spoken language understanding and the relationship between early sensory input or the problem of word recognition and its interface with higher levels of linguistic analysis. Researchers in any field of scientific investigation typically work on tractable problems that can be studied with existing methodology and paradigms. Relative to the bulk of speech perception research on isolated phoneme perception, very little is actually known today about how the early sensory-based acoustic-phonetic information is used by the speech processing system in tasks involving word recognition and sentence perception or how changes in the segmental and/or suprasegmental structure of the speech signal influence intelligibility and

comprehension of spoken language. These are problems of current interest that will no doubt be pursued over the next few years.

I believe that continued experimental and theoretical work in speech perception will be directed at new models and theories that capture significant aspects of the process of spoken language understanding. What is important at the present time, is to direct research efforts toward somewhat broader issues involving the use of meaningful stimuli in more naturalistic experimental tasks that require the listener's active deployment of phonological, lexical, syntactic and semantic knowledge to assign an interpretation to the sensory input. Past theoretical work in speech perception has not been very well developed nor has the link between theory and empirical data been very sophisticated. Moreover, work in the field of speech perception as in other areas has tended to be defined by specific experimental paradigms or particular phenomena (i.e., dichotic listening, categorical perception or selective adaptation). The major theoretical issues in speech perception often seem to be ignored. They receive little serious attention by investigators who are involved in working on the details of experimental problems that unfortunately bear only marginally on the primary perceptual and cognitive processes that are used in spoken language understanding. Although a very formidable task, research in the future will be focused more directly on the general problem of spoken language understanding. In my view, it is here that the greatest insights into language processing will be found in the next ten years.

Acknowledgments

This research was supported by NIH research grant NS-12179 to Indiana University in Bloomington. I an grateful to Jon Allen of MIT for many fruitful conversations regarding the issues discussed in this paper.

References

Bever TG, Lackner J, Kirk R. The underlying structure sentence is the primary unit of immediate speech processing. Percep. Psychophys. 1969; 5:225–234.

Cohen, A.; Nooteboom, S., editors. Structure and Process in Speech Perception. Heidlelberg: Springer-Verlag; 1975.

Cole, RA.; Jakimik, J. Understanding speech: How words are heard. In: Underwood, G., editor. Strategies of Information processing. New York: Academic Press; 1978. p. 68-116.

Darwin, CJ. On the dynamic use of prosody in speech perception. In: Cohen, A.; Nooteboom, SG., editors. Structure and Process in Speech Perception. Berlin: Springer-Verlang; 1975. p. 178-194.

Fant CGM. Description analysis of the acoustic aspects of speech. Logos. 1962; 5:3–17. [PubMed: 13891546]

Forster, KI. Accessing the mental lexicon. In: Wales, RJ.; Walker, E., editors. New Approaches to Language Mechanisms. Amsterdam: North-Holland; 1976. p. 257-287.

Foss DJ, Blank MA. Identifying the speed codes. Cog. Psychol. 1980; 22:l–31.

Gaitenby JH. The elastic word. Haskins Laboratories Status Report on Speech Research, SR-. 1965; 2:3.1–3.12.

Ganong WF. The internal structure of consonants in speech perception: Acoustic cues, not distinctive features. Unpublished manuscript. 1979

Huggins AWF. On the perception of temporal phenomena in speech. J. acoust. Soc. Am. 1972; 51:1279–1290. [PubMed: 5032944]

Huggins, AWF. Speech timing and intelligibility. In: Requin, J., editor. Attention and Performance VII. Hillsdale, NJ: Erlbaum; 1978.

Klatt DH. Vowel lengthening is syntactically determined in a connected discourse. J. Phon. 1975; 3:129–140.

Klatt DH. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. J. acoust. Soc. Am. 1976; 59:1208–1221. [PubMed: 956516]

Klatt DH. Review of the ARPA speech understanding project. J. acoust. Soc. Am. 1977; 62:1345–1366.

Klatt DH. Speech percetion: A model of acoustic-phonetic analysis and lexical access. J. Phon. 1979; 7:279–312.

Klatt, DH. Synthesis by rule of segmental durations in English sentences. In: Lindblom, B.; Ohman, S., editors. Frontiers of Speech Communication Research. New York: Academic Press; 1979.

Klatt DH. Lexical representations and processing strategies during speech production and perception. Psychol. Rev. (In press).

Klatt, DH.; Cooper, WE. Perception of segment duration in sentence contexts. In: Cohen, A.; Nooteboom, SG., editors. Structure and Process in Speech Perception. New York: Springer-Verlag; 1975.

Lehiste, I. Suprasegmentals. Cambridge, MA: MIT Press; 1970.

Marslen-Wilson WD, Tyler LK. The temporal structure of spoken language understanding. Cog. 1980; 8:1–71.

Marlsen-Wilson WD, Welsh A. Processing interactions and lexical access during word recognition in continuous speech. Cog. Psychol. 1978; 10:29–63.

Massaro, DW. Technical Report No. 423. Wisconsin Research and Development Center for Cognitive Learning, University of Wisconsin-Madison; 1977. Reading and listening.

Miller GA. Decision units in the perception of speech. IRE Transactions on Information Theory, IT-8. 1962:81–83.

Miller GA, Heise GA, Lichten W. The intelligibility of speech as a function of the context of the test materials. J. exper. Psychol. 1951; 41:329–335. [PubMed: 14861384]

Miller GA, Isard S. Some perceptual consequences of linguistic rules. J. verb. Learn. verb. Behav. 1963; 2:217–228.

Miller, JL. Eimas, PD.; Miller, JL. Perspectives on the Study of Speech. Hillsdale, NJ: Erlbaum Associates; 1981. The effect of speaking rate on segmental distinctions: Acoustic variation and perceptual compensation.

Morton J. Lateraction of information in word recognition. Psychol. Rev. 1969; 76:165–178.

Morton, J. Word recognition. In: Morton, J.; Marshall, JD., editors. Psycholinguistics 2: Structure and Processes. Cambridge: MIT Press; 1979. p. 107-156.

Nooteboom, SG.; Brokx, JPL.; deRooij, JJ. Contributions of prosody to speech perception. In: Levelt, WJM.; Floresd'Arcais, GB., editors. Studies in the Perception of Language. New York: John Wiley; 1978. p. 75-107.

Parker F. Distinctive features and acoustic cues. J. acoust. Soc. Am. 1977; 62:1051–1054.

Pisoni, DB. Speech perception. In: Estes, WK., editor. Handbook of Learning and Cognitive Processes. Vol. vol. 6. Hillsdale, NJ: Erlbaum Associates; 1978. p. 167-233.

Port RF. Combinations of timing factors in speech production. J. acoust. Soc. Am. 1981; 69:262–274. [PubMed: 7217524]

Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. Science. 1981; 212:947–950. [PubMed: 7233191]

Searle CL, Jacobson JZ, Raymont SG. Stop consonant discrimination based on human audition. J. Acoust. Soc. Am. 1979; 65:799–809. [PubMed: 447910]

Studdert-Kennedy, M. The perception of speech. In: Sebeok, TA., editor. Current Trends in Linguistics. Vol. vol. XI. The Hague, Mouton: 1974.

Studdert-Kennedy, M. Speech perception. In: Lass, NJ., editor. Contemporary Issues in Experimental Phonetics. New York: Academic Press; 1976. p. 243-293.

Wickelgren, WA. Phonetic coding and serial order. In: Carterette, EC.; Friedman, MP., editors. Handbook of Perception. Vol. vol. VIII. New York: Academic Press; 1976. p. 227-264.

Zwicker E, Terhardt E, Paulus E. Automatic speech recognition using psychoacoustic models. J. acoust. Soc. Am. 1979; 65:487–498. [PubMed: 489818]