SOME REMARKS ON RECENT ISSUES IN SPEECH PERCEPTION RESEARCH

Hiroya Fujisaki, University of Tokyo, Tokyo, Japan

I understand that the role of my contribution is to supplement Michael Studdert-Kennedy's comprehensive and impressive report. Therefore I will not try to give here an extensive review, but will state my personal remarks on some of the issues in recent studies of speech perception.

Categorical Perception of Speech and Non-speech Stimuli

A number of recent studies (Cutting and Rosner, 1974; Miller et al., 1976; Pisoni, 1977; Pastore et al., 1977) have confirmed the earlier assertion (Lane, 1965) that the categorical effect in discrimination measurements (I prefer the above expression to the conventional "categorical perception") is not specific to speech perception. As it has already been shown by a rigorous psychophysical account of the measurement procedure (Fujisaki, 1971), the apparent enhancement of discriminability across a category boundary (*not* the suppression of discriminability within categories) is an artifact that accrues from the subject's ability to categorize the test stimuli and to retain the results in the short-term memory, regardless of whether the stimuli are speech or non-speech. In other words, the categorical effect is a consequence of the single fact that the subject possesses or is provided with a stable threshold for categorical judgment of individual stimuli, but the process of discrimination is clearly sequential since the comparative judgment for discrimination is mediated by the results of categorical judgment. This simply indicates the inherent inability of our test procedures to dissociate the two types of judgment. But why are people so eager to look into, and to produce still new examples of, this phenomenon, when, after all, discriminability plays only a minor role in the actual speech communication? One of the interesting outcomes of these efforts may be the indication of perceptual similarity between the VOT continuum of stop consonants and some non-speech continua, suggesting that the perception of speech categories might be based on some simple psychoacoustic properties rather than on complex speech-specific properties. Generalization of this finding to other speech sound categories, however, requires careful investigations since there exist a number of acoustic continua on which phoneme categorization is not universal, but is more or less specific to individual languages.

Levels of Processing and Selective Adaptation

There is little doubt that the identification of a particular segment of speech is a categorical judgment based on a number of acoustic properties (cues) detected from the continuous speech signal. For the sake of simplicity, we shall drop the issue of segmentation and defer the discussion of contextual effects to a later section. Conceptually, therefore, phoneme identification can be regarded as a two-stage process: property detection and decision. Neurophysiological evidences of signal processing in the visual cortex (e.g. Hubel and Wiesel, 1965), however, suggest that the detection of these properties is performed by a large number of neurons or neuron groups, arranged in a multi-level structure rather than a single-level structure; a set of primary properties being utilized for extracting a secondary property at the next level, and a set of secondary properties being further utilized for extracting a property of a still higher order at the next level, etc. Thus the conventional division of two levels (auditory vs. phonetic, or peripheral vs. central) may not be appropriate and the transition from the peripheral to the central processing may be more gradual than it is suggested by the terminology. It should also be noted that the extraction of individual properties need not be competitive, and the final decision is made after combination and temporal integration of the higher-order properties (Repp et al., 1978). The selective adaptation paradigm (Eimas and Corbit, 1973) is certainly a powerful tool to look into these mechanisms. Through systematic manipulation of the properties to be shared by the adaptors and the test stimuli as well as of the modes of stimulus presentation and response (e.g. Sawush, 1977a, 1977b), both structural and functional informations on these mechanisms have been accumulated. It is to be noted, however, that the adaptation is generally not restricted to one particular property detector nor to one particular level, and the resulting changes in a subject's response should be ascribed not only to changes in the sensitivity of the related property detectors but also to changes in the thresholds of categorical judgments both for phoneme identification and for stimulus rating. Further research on the elaboration of the paradigm, as well as its application to various speech sound categories other than the intensively studied voiced stops, would clearly lead to a deeper understanding of the processes of speech perception at least at the level of the phoneme.

Speech Perception in Context

Although the selective adaptation paradigm is successful in studying the mechanisms of speech perception by creating a very special context, the results are not directly applicable to the process of speech perception in an ordinary context, where individual phonemes generally follow one after another and overlap in their articulatory realization to form a continuous acoustic string. Both articulatory and acoustic studies of monosyllables, as the smallest units of a phoneme sequence, reveal the mutual character of coarticulatory influences between the vowel as the syllable nucleus and the adjoining consonant(s). These coarticulatory changes are, however, compensated for by perception. For example, it is a well-known fact that the perception of voiced consonants is severely impaired if we take away the formant transitions and leave only the bursts (e.g. Dorman et al., 1977). Likewise, the perception of the syllable nucleus is incomplete when we take away the formant transitions to the adjoining consonant(s) and leave only the stationary portion (e.g. Fujimura and Ochiai, 1963; Strange et al., 1978). Thus the vowel and the consonant(s) within a syllable complement each other in perception. In more generalized connected speech, however, the coarticulatory influences extend over the syllable boundaries, and the perception of a vowel within a syllable is found to be incomplete if the syllable is taken out of its context and presented in isolation, but is restored when the syllable is presented with its immediately adjacent syllable(s) (Kuwahara and Sakai, 1972). The perceptual mechanism of compensation for the coarticulatory variations of vowels has been investigated using synthetic disyllables of Japanese consisting of two vowels and non-speech stimuli with similar dynamic characteristics (Fujisaki and Sekimoto, 1975), indicating that the perception of a vowel in a dynamic context involves at least two distinct processes: extrapolation of incomplete formant transitions occurring both for speech and for non-speech, and short-term change of category boundaries occurring only for speech. Further investigation of the process of speech perception in the dynamic context is clearly necessary in order to elucidate the basis upon which the listener's knowledge of the language at the phonological, morphological, lexical, and syntactic levels, as well as the semantic and pragmatic information, is fully utilized in the understanding of spoken messages.

The Roles of Prosody

Although prosody is not a well-defined concept, I consider it as a set of functions imposed upon a sequence of phonemes for the purpose of transmitting information concerning some linguistic units that are larger than the phoneme, such as word, sentence, and paragraph. Word prosody is almost synonymous with word accent (or intonation) and is used to transmit lexical information concerning homonyms. Sentence prosody consists of prominence, intonation, and rhythm, which are used to transmit or supplement both semantic and syntactic information of a sentence. Paragraph prosody (Lehiste, 1975; 1978), a relatively new concept, may be regarded as transmitting the structural information of a discourse. In addition to these major functions, prosody also contributes to facilitate segmental perception and to maintain the coherence of an utterance (Nooteboom et al., 1976), but I consider the latter functions to be rather subsidiary. These prosodic functions are realized mainly through the medium of suprasegmental features such as pitch, loudness and quantity (duration) of segments as well as of pauses, but may also be manifested by some segmental features such as phonemic quality of vowels (e.g. word accent in English). In spite of the importance of these functions in speech perception, comparatively little effort seems to have been spent in studying their perceptual effects. This may be firstly because of the lack or insufficiency of their formal descriptions, secondly because of the lack of analysis techniques to obtain quantitative acoustic formulations, and thirdly because of the increased difficulty in the preparation of synthetic speech stimuli of larger duration necessary for free and precise control of suprasegmental features. However, studies on perception of word accent and/or sentence intonation have recently been published on Japanese (Fujisaki and Sugito, 1976), on Dutch ('t Hart, 1976), on Danish (Thorsen, 1976), on Thai (Abramson, 1977), on Estonian (Eek, 1977), etc. The perceptual role of duration for expressing syntactic information has also been demonstrated for English (Lehiste et al., 1976). Perceptual reality of isochrony has been discussed and demonstrated using natural and synthetic speech (Lehiste, 1977; Higuchi and Fujisaki, 1978; Sato, 1978). On the other hand, perceptual roles of acoustic correlates of paragraph prosody, such as the peak in the fundamental frequency, pre-boundary lengthening, and pause duration, have been investigated using natural and spectrally-inverted utterances (Lehiste, 1975; 1978).

Development and Impairments of Speech Perception

While the main interest of phoneticians may reside in the understanding of speech perception by an adult with normal hearing and language abilities, much could be learned from the study of developmental processes in young children (e.g. Fourcin, 1978), as aptly pointed out by Studdert-Kennedy. Studies of speech perception in hearing-impaired children (Fourcin et al., 1978; Waldman et al., 1978) are indispensable for finding systematic methods of training and for designing useful aids. Specially designed rhyme tests using natural utterances (Risberg, 1976) or identification tests using synthetic stimuli (Yokkaichi and Fujisaki, 1978) are useful for efficient collection of data on segmental perception, while the ability of identifying intonation contours can be tested by using natural utterances (Risberg and Agelfors, 1978). Furthermore, the perceptual ability of children and adults with language comprehension impairments can be tested by synthetic stimuli with various temporal characteristics (Tallal et al., 1976; Tallal and Newcombe, 1978), allowing one to locate the processing of rapid transitions at the dominant hemisphere. Studies of speech perception in its developmental stages as well as in the pathological cases can thus shed light on the process of speech perception in normal adults and can also lead to a better use of our knowledge for the alleviation of the impairments.

References

Abramson, A. (1977): "The phonetic plausibility of the segmentation of tones in Thai phonology," Haskins Labs. Status Rept. on Speech Research SR-53(1), 73-77.

Cutting, J. and B.S. Rosner (1974): "Categories and boundaries in speech and music," Perc.Psych. 16, 564-570.

Dorman, M.F., M. Studdert-Kennedy, and L.J. Raphael (1977): "Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," Perc. Psych. 22, 109-112.

Eek, A. (1977): "Experiments on the perception of some word series in Estonian," Estonian Papers in Phonetics 1977, 7-32.

Eimas, P.D. and J.D. Corbit (1973): "Selective adaptation of linguistic feature detectors," Cogn.Psych. 4, 99-109.

Fourcin, A.J. (1978): "Acoustic patterns and speech acquisition," Speech and Hearing, University College London 3, 143-172.

Fourcin, A.J., S. Evershed, J. Fisher, A. King, A. Parker, and R. Wright (1978): "Perception and production of speech patterns by hearing-impaired children," Speech and Hearing, University College London 3, 173-204.

Fujimura, O. and K. Ochiai (1963): "Vowel identification and phonetic contexts," JASA 35, 1889 (A).

Fujisaki, H. and T. Kawashima (1971): "A model of the mechanisms for speech perception —— Quantitative analysis of categorical effects in discrimination ——," Ann.Rept.Engg.Res.Inst., Fac.Engg., Univ.Tokyo 30, 59-68. Also to appear as "On the modes and mechanisms of speech perception —— Analysis and interpretation of categorical effects in discrimination," in Frontiers of Speech Communication Research, B. Lindblom and S. Öhman (eds.), London: Academic Press.

Fujisaki, H. and S. Sekimoto (1975): "Perception of time-varying resonance frequencies in speech and non-speech stimuli," in Structure and Process in Speech Perception, A. Cohen and S.G. Nooteboom (eds.), 269-280, Berlin/Heidelberg/New York: Springer-Verlag.

Fujisaki, H. and M. Sugito (1976): "Acoustic and perceptual analysis of two-mora word accent types in the Osaka dialect," Ann. Bull. RILP, Univ. Tokyo. 10, 157-172.

't Hart, J. (1976): "Psychoacoustic backgrounds of pitch contour stylisation," IPO Annual Progress Report 11, 11-19.

Higuchi, N., H. Fujisaki, and S. Sekimoto (1978): "Production and perception of segmental durations in spoken Japanese," JASA 64, Supplement No.1, S113 (A).

Hubel D.H. and T.N. Wiesel (1965): "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," J.Neurophysiol. 28, 229-289.

Kuwahara, H. and H. Sakai (1972): "Perception of vowels and C-V syllables segmented from connected speech", Journal of the Acoustical Society of Japan 28, 225-234.

Lane, H. (1965): "The motor theory of speech perception: A critical review," Psych.Rev. 72, 275-309.

Lehiste, I. (1975): "The phonetic structure of paragraphs," in Structure and Process in Speech Perception, A. Cohen and S.G. Nooteboom (eds.), 195-203, Berlin/Heidelberg/New York: Springer-Verlag.

Lehiste, I., J.P. Olive, and L.A. Streeter (1976): "Role of duration in disambiguating syntactically ambiguous sentences," JASA 60, 1199-1202.

Lehiste, I. (1977): "Isochrony reconsidered," JPh 5, 253-263.

Lehiste, I. (1978): "Temporal organization and prosody —— perceptual aspects," JASA 64, Supplement No.1, S112 (A).

Miller, J.D., C.C. Weir, R.E. Pastore, W.J. Kelly, and R.J. Dooling (1976): "Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception," JASA 60, 410-417.

Nooteboom, S.G., J.P.L. Brokx, and J.J. de Rooij (1976): "Contributions of prosody to speech perception," IPO Annual Progress Report 11, 34-54.

Pastore, R.E., W.A. Ahroon, K.J. Baffuto, C. Friedman, J.S. Puelo, and E.A. Fink (1977): "Common-factor model of categorical perception," J.Exp.Psych. 3, 686-696.

Pisoni, D. (1977): "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops," JASA 61, 1352-1361.

Repp, B.H., A.M. Liberman, T. Eccardt, and D. Pesetsky (1978): "Perceptual integration of acoustic cues for stop, fricative, and affricate manner," Haskins Labs. Status Rept. on Speech Research SR-53(2), 61-83.

Risberg, A. (1976): "Diagnostic rhyme test for speech audiometry with severely hard of hearing and profoundly deaf children," STL-QPSR 2-3/1976, 40-58.

Risberg, A. and E. Agelfors (1978): "On the identification of intonation contours by hearing-impaired listeners," STL-QPSR 2-3/1978, 51-61.

Sato, H. (1978): "Temporal characteristics of spoken words in Japanese," JASA 64, Supplement No.1, S113 (A).

Sawush, J.R. (1977a): "Peripheral and central processes in selective adaptation of place of articulation in stop consonants," JASA 62, 738-750.

Sawush, J.R. (1977b): "Processing of place information in stop consonants," Perc.Psych. 22, 417-426.

Strange, W., J.J. Jenkins, and T.R. Edman (1978): "Dynamic information specifies vowel identity," JASA 63, Supplement No.1, S5 (A).

Tallal, P., R. Stark, and B. Curtiss (1976): "The relation between speech perception impairment and speech production impairment in children with developmental dysphasia," Brain and Language 3, 305-317.

Tallal, P. and F. Newcombe (1978): "Impairment of auditory perception and language comprehension in dysphasia," Brain and Language 5, 13-24.

Thorsen, N. (1976): "An acoustical investigation of Danish intonation: Preliminary results," Annual Report of the Institute of Phonetics, University of Copenhagen 10, 85-148.

Waldman, F.R., S. Singh, and M.E. Hayden (1978): "A comparison of speech-sound production and discrimination in children with functional articulation disorders," L&S 21, 205-220.

Yokkaichi, A. and H. Fujisaki (1978): "Identification of synthetic speech stimuli by hearing-impaired subjects," JASA 64, Supplement No.1, S19 (A).