REPORT: SPEECH PERCEPTION
(see vol. I, p. 59-99)

Reporter: Michael Studdert-Kennedy
Co-reporter: Hiroya Fujisaki
Co-reporter: Ludmilla Chistovich

Chairpersons: Antony Cohen and Louis C.W. Pols

REPORTERS' ADDITIONAL REMARKS

Michael Studdert-Kennedy gave a summary of his report. He mentioned that he might have misunderstood the aim of the work of the Leningrad group to some extent. He had thought that they were looking for phonetic segments in the acoustic signal, i.e. for acoustic segments that would be isomorphic with phonetic segments, but it appears from Ludmilla Chistovich's report that they are in fact looking primarily for acoustic segmentation, which will, e.g. be essential for the estimation of durational events.

Discussing the problem of feature detectors he mentioned that animals that have feature detectors and templates (e.g. the bullfrog and birds) have them because they need them, having to get along very soon after birth without parental help, but that is not the case with the human infant, who has a long period of parental care.

Concerning the problem of perception of sounds by means of an integration of a variety of cues, he emphasized that the idea that these cues may be held together by the underlying gesture should not be understood as a claim for a motor theory of perception, which implies that perception requires reference to the production system. The idea is that you perceive the production gesture directly like you perceive the movement of a hand by means of the light reflected from it. If the hand was moved inside a resonating chamber which had a source exciting it, you might hear the gesture instead of seeing it.

Studdert-Kennedy added a section on cerebral specialization not found in the original report. A written version of this addition is given below:

Cerebral specialization

Nonetheless, opposition between the two modes of lexical

access -- holistic, from "auditory contour", analytic, from phonetic segments -- should not be too sharply drawn. The work of Zaidel (1978a,b) with "split-brain" patients has demonstrated that holistic access is certainly possible. The cerebral hemispheres of such patients have been surgically separated by section of the connecting pathways (corpus callosum) for relief of epileptic seizure. The separation permits an investigator to assess the linguistic capacities of each hemisphere independently. Zaidel (1978 a,b) has shown that the isolated right hemisphere of such a patient, though totally mute, can recognize a sizeable auditory lexicon and has a rudimentary syntax sufficient for understanding phrases of up to three or four words in length. However, it is incapable of identifying nonsense syllables or of performing tasks that call for phonetic analysis, such as recognizing rhyme (cf. Levy, 1974). This phonetic deficit evidently precludes short-term verbal store, thus limiting the right hemisphere's capacity for syntactic analysis of lengthy utterances, and forces organization of language around meaning. Whether we assume a similar, subsidiary organization in the left hemisphere or some process of inter-hemispheric collaboration, it is clear that normal language comprehension could, at least in principle, draw on both holistic and analytic mechanisms.

At the same time, Zaidel's work provides striking support for the hypothesis, originally derived from dichotic studies, that the distinctive linguistic capacity of the left hemisphere is for phonological analysis of auditory pattern (Studdert-Kennedy and Shankweiler, 1970). Further support has come from electroencephalography (Wood, 1975) and, quite recently, from studies of the effects of electrical stimulation during craniotomy (Ojemann and Mateer, 1979). The latter work isolated, in four patients, left frontal, temporal and parietal sites, surrounding the final cortical motor pathway for speech, in which stimulation blocked both sequencing of oro-facial movements and phoneme identification.

This fascinating discovery meshes neatly with a growing body of data and theory that has sought, in recent years, to explain the well-known link between lateralizations for hand control and speech. Semmes (1968) offered a first account of the association by arguing, from a lengthy series of gunshot lesions, that the left hemisphere is focally organized for fine motor control, the right hemisphere diffusely organized for broader control. Subsequently,

Kimura and her associates reported that skilled manual movements (Kimura and Archibald, 1974) and non-verbal oral movements (Mateer and Kimura, 1977) tend to be impaired in cases of non-fluent aphasia. These impairments are specifically for the sequencing of fine motor movements and are consistent with other behavioral evidence that motor control of the hands and of the speech apparatus is vested in related neural centers (Kinsbourne and Hicks, 1979). In fact, Kimura (1976) has proposed that "...the left hemisphere is particularly well adapted, not for symbolic function per se, but for the execution of some categories of motor activity which happened to lend themselves readily to communication" (p. 154). Among these categories we must, incidentally, include those that support the complex "phonological" and morphological processes of manual sign languages, now being discovered by the research of Klima, Bellugi and their colleagues (Klima and Bellugi, 1979).

The drift of all this work is toward a view of the left cerebral hemisphere as the locus of interrelated sensorimotor centers, essential to the development of language, whether spoken or signed. To understanding of the speech sensorimotor system perceptual studies of dichotic listening will doubtless contribute. Indeed, important dichotic studies have recently found evidence for the double dissociation of left and right hemisphere, speech and music, in infants as young as two or three months (Entus, 1977; Glanville, Best and Levenson, 1977). However, dichotic work has not fulfilled its early promise, largely because it has proved extraordinarily difficult to partial out the complex of factors, behavioral and neurological, that determine the degree of observed ear advantage (cf. Studdert-Kennedy, 1975). For the future, we may increasingly rely on instrumental techniques for monitoring brain activity, such as the blood-flow studies of Lassen and his colleagues (Lassen, Ingvar and Skinhøj, 1978), induced reversible lesions by focal cooling (Zaidel, 1978b), improved methods of electroencephalographic analysis, auditory evoked potentials (Molfese, Freeman and Palermo, 1975) and, perhaps infrequently, direct brain stimulation.

References

Abramson, A.S. (1977): "Laryngeal timing in consonant distinctions", Phonetica 34, 295-303.

Campbell, R. and B. Dodd (in press): "Hearing by eye", Quarterly Journal of Experimental Psychology.

Entus, A.K. (1977): "Hemispheric asymmetry in processing dichotically presented speech and nonspeech stimuli by infants", in S.J. Segalowitz and F.A. Greber (eds.) Language development and neurological theory, 64-73, New York: Academic Press.

Glanville, B.B., C.T. Best and R. Levenson (1977): "A cardiac measure of asymmetries in infant auditory perception", Developmental Psychology 13, 54-59.

Kimura, D. (1976): "The neural basis of language qua gesture", in H. Whitaker and H.A. Whitaker (eds.) Studies in Neurolinguistics (vol. 3), New York: Academic Press.

Kimura, D. and Y. Archibald (1974): "Motor functions of the left hemisphere", Brain 97, 337-350.

Kinsbourne, M. and R.E. Hicks (1979): "Mapping cerebral functional space: competition and collaboration in human performance", in M. Kinsbourne (ed.) Asymmetrical function of the brain, 267-273, New York: Cambridge University Press.

Klima, E.S. and U. Bellugi (1979): The Signs of Language, Cambridge, Mass.: Harvard University Press.

Lassen, N.A., D.H. Ingvar and E. Skinhøj (1978): "Brain function and blood flow", Scientific American 239, 62-71.

Levy, J. (1974): "Psychobiological implications of bilateral asymmetry", in S.J. Dimond and J.G. Beaumont (eds.) Hemisphere function in the human brain, London: Elek.

Martin, J.G. (1972): "Rhythmic (hierarchical) versus serial structure in speech and other behavior", Psychological Review 79, 487-509.

Mateer, C. and D. Kimura (1977): "Impairment of non-verbal oral movements in aphasia", Brain and Language 4, 262-276.

Molfese, D.L., R.B. Freeman and D.S. Palermo (1975): "The ontogeny of brain lateralization for speech and nonspeech stimuli", Brain and Language 2, 356-368.

Nakatani, L.H. and K.D. Dukes (1977): "Locus of segmental cues for word juncture", JASA 62, 714-719.

Ojemann, G. and C. Mateer (1979): "Human language cortex: localization of memory, syntax and sequential motor-phoneme identification systems", Science 205, 1401-1403.

Semmes, J. (1968): "Hemispheric specialization: A possible clue to mechanism", Neuropsychologia 6, 11-26.

Stevens, K.N. and S. Blumstein (1978): "Invariant cues to place of articulation", JASA 64, 1358-1368.

Studdert-Kennedy, M. (1975): "Two questions", Brain and Language 2, 123-130.

Studdert-Kennedy, M. and D.P. Shankweiler (1970): "Hemispheric specialization for speech perception", JASA 48, 579-594.

Studdert-Kennedy concluded by quoting Ludmilla Chistovich who as a conclusion of her report writes "We (our group) believe that the only way to describe human perception is to describe not the perception itself but the artificial speech understanding system which is most compatible with the experimental data obtained in speech perception research". He found that this was a very good statement of a heuristic programme, but emphasized that what is required is a constant interplay between the psycho-biological facts of the human behaviour and whatever robotic facsimile the engineers have managed to construct.

Hiroya Fujisaki summarized his report, giving a more detailed account of the first section on categorical perception based on slides illustrating his well-known dual coding model of discrimination. The fact that categorical perception appears in an apparent enhancement of discriminability on the phoneme boundary, and not in a suppression of discriminability within the category, was illustrated by reference to experiments with an r-l continuum presented to American and Japanese listeners. Categorization immediately after the auditory mapping and dominance of categorical perception on comparative judgement seems to be characteristic of the speech mode, but is also found in some cases of non-speech stimuli. Due regard should be paid to disturbances by noise (uncertainty) both in the categorical judgement process and in the retrieval process from the short term memory of timbre. The ability of categorical judgement is based partly on basic physical discreteness, partly on language specific criteria acquired through training in a specific language.

As for the perception of speech in context, Fujisaki emphasized that the importance of context can not be evaluated until we have studied the variability of phonemes in isolation.

Ludmilla Chistovich had been prevented from participating in the congress.

DISCUSSION

The discussion was opened by Kenneth Stevens, Sieb Nooteboom and Christopher Darwin.

Kenneth N. Stevens confined his remarks principally to the question of invariance versus non-invariance. It is obvious that when one produces phonetic segments in context, the articulators

have to move from one target to the next, and so the signal is clearly context-dependent. But if you examine the sound in the right way and look at the right places in the sound, you will see much less variability and more invariance for a given distinctive feature both in the context of other features in the same segment and in the context of adjacent segments. Stevens showed slides of the acoustic waveforms of the syllables ba,da ga,pa ta,ka. The samples were taken at the onset of the consonants and the spectra had been calculated in a specific way with a specific time window. He pointed out that in labials the gross shape of the spectrum was flat or falling and spread out in frequency. For the alveolars the spectrum was also spread out in frequency, but rising, or acute, and in velars it had a prominent peak in the mid frequency range. One may say there is compactness to the spectrum.

It is possible to devise algorithms or templates that will recognize each of these gross spectrum shapes - and the point is that if one looks at the gross spectrum shapes rather than at the details of where individual peaks are in the spectrum, one does see a considerable amount of invariance. Now, this is a physically measured spectrum with a linear time scale and with fixed bandwidths. What one should really do is to look at a spectrum as it is processed by the auditory system with the appropriate bandwidths and time constants of that system. At some level in the auditory representation that spectrum may well be influenced by what immediately precedes the spectrum. There are already neurophysiological data that would indicate that. The spectra would have to be brought more in line with what we know about psychophysics and the electro-physiology of the auditory system. But even at this acoustical level we see a measure of invariance for stop consonants, as far as place of articulation is concerned.

In this connection Stevens added some remarks on categorical perception. As one moves along the continuum from ka to ta, the auditory system does not treat the physical continuum as though you were moving continuously. As long as the sound has some sort of compact spectral peak it would sound pretty much the same, and it is only when this peak disappears that you will get a sudden change over to a different kind of sound. Stevens would argue that at some level of the auditory system there is some kind of unique response to each of the spectrum types characterized by the gross properties mentioned above.

Where should one look in the signal to find this invariance? Ludmilla Chistovich and Stevens agree that the places in the sound where there are rapid changes are the places which seem to contain a lot of information. If one looks at these places, one sees invariance not only for place of articulation but also for other distinctive features. The formant transitions are acoustic material that links these rapidly changing events with the relatively slowly changing events during the vowel. There is a tendency for a given phonetic feature to have invariant properties. Stevens would argue that the infant comes into the world endowed with mechanisms that are sensitive to these properties. It has a mechanism for classifying sounds, in particular features, as being similar. These relatively invariant primary acoustic properties help to define distinctive features and provide the signal with the kind of properties that enable the infant to learn language. The context-dependent effects which can go along with these primary properties can be used when necessary, perhaps in noisy situations or in rapid speech to supplement the primary cues.

Sieb Nooteboom had no disagreement with the description given by the reporters of the state of the art in speech perception research, but some comments with respect to the state of the art itself.

The underlying or most basic common goal of speech perception research is undoubtedly to understand the structures and processes by which a listener can recover from the acoustic signal what a speaker is saying to him. It is only when we have reached a basic understanding of speech perception in this sense that we can apply the insights gained to phonological explanation, improvement of synthesis by rule, etc. The most important of the processes involved may be labeled recognition. But experimental paradigms in our discipline draw heavily on forced-choice identification, discrimination, similarity judgements, and scaling, none of which studies recognition as a process in itself. In a typical recognition task each stimulus is presented once only and is potentially compared by the subjects with, for example, all possible words or morphemes in the language, whereas in identification stimuli are typically presented more than once and the response set is restricted by the task. With a very few notable exceptions (cf. Goldstein 1977, Marslen-Wilson and Welsh 1978, Cole and Jakimik

1978) recognition is not studied at all. In this respect research on reading, where considerable attention is paid to visual word recognition, is ahead of research on speech perception (Bouwhuis 1979).

Too much attention is focussed on phonemes and phonemic features at the expense of more comprehensive structures, words, morphemes, and prosodic structures, and their communicative function. For a listener to understand what a speaker is saying to him, he must generally recognize meaningful units. Words and morphemes are certainly the most important structures in speech perception. Most investigators seem to believe that once we understand how phonemes are extracted from the signal we can easily explain further linguistic processing. This is hardly true. We do not know whether word recognition is mediated by phoneme extraction, or rather, as recently suggested by Dennis Klatt (1979), by spectral templates, and we will never know until we turn to the study of word recognition. And even if word recognition turns out to be mediated by phoneme extraction, that is certainly not all there is to it (cf. the word completion effect in visual word recognition, Reicher 1969, Bouwhuis 1969).

The even more comprehensive suprasegmental or prosodic structures also contribute in several ways to a listener's recovery of what the speaker wants to say to him. It is a good thing that in recent years researchers have been paying more attention to prosodic structures. Attention has mainly centered around the connection between prosody and syntax, but Nooteboom thinks that two other functions are at least as important in daily speech communication. One is that differences in global pitch level, as well as the presence of normal intonational patterning, appear to increase the intelligibility of speech masked by speech (Brokx 1979). The other, and perhaps most important communicative function of prosody is to signal semantic focus (O'Shaughnessy 1978).

We should acknowledge that phonetics, and especially perceptual phonetics, has reached a stage in which it should not be limited to the study of consonants and vowels. Much is to be gained from widening the scope of the mainstream of our discipline.

References

Bouwhuis, D.G. (1976): Visual Recognition of Words. Unpublished Doctor's Thesis, Catholic University of Nijmegen

Brokx, J.P.L. (1979): Waargenomen Continuïteit in Spraak: het Belang van Toonhoogte. Unpublished Doctor's Thesis, Eindhoven University of Technology

Cole, R.A. and J. Jakimik (1978): "Understanding speech: how words are heard", in G. Underwood (ed.) Strategies of Information Processing, Academic Press

Goldstein, L. (1979): "Perceptual salience of stressed syllables", Chapter II of an Unpublished Doctor's Thesis, University of California Los Angeles, Department of Linguistics

Klatt, D.H. (1979): "Speech perception: a model of acoustic-phonetic analysis and lexical access", Journal of Phonetics 7, 279-312

Marslen-Wilson, W.D. and A. Welsh (1978): "Processing interactions and lexical access during word recognition in continuous speech", Cognitive Psychology 10, 29-63

O'Shaughnessy, D. (1976): Modeling Fundamental Frequency, and its Relationship to Syntax, Semantics, and Phonetics, Unpublished Doctor's Thesis, M.I.T., Cambridge, Massachusetts

Reicher, G.M. (1969): "Perceptual recognition as a function of meaningfulness of stimulus material", Journal of Experimental Psychology 81, 275-280.

Christopher Darwin started by quoting Ludmilla Chistovich (the same passage that is quoted by Michael Studdert-Kennedy at the end of his report). He concentrated his contribution on a discussion of the relation between computer speech recognition work and the human speech perception in the area of auditory feature extraction and phonetic segment identification.

The engineer does not have to make his system in a psychologically plausible fashion to make it work, but there does seem to be general agreement that speech recognition systems should take account of such relatively peripheral auditory phenomena as critical bands, middle-ear transfer function, growth of loudness and non-simultaneous masking although often the application to speech sounds has to be made on trust rather than on adequate psycho-acoustic data.

Chistovich, rightly, identifies as important the problem of how to represent the input parameters to an acoustic phonetic stage. She points out that theories of phonetic perception are going to be heavily influenced by the materials they have to work with. Thus, if speech understanding programmes are to be serious models of human perception we have to find ways of representing the input signal which are more psychologically plausible and more phonetically germane than a series of categorical labels representing the best, fitting one of a small (100-300) number of static spectral templates.

We have rather little idea what the parameters of an auditory representation should be. Probably it should represent all discriminable differences in the speech signal (taking the most liberal view of "discriminable"), rejecting none of the information to which the listener may need to be sensitive (cf. the work on early visual processing by Marr 1976), but on the other hand the representation must be explicit, organised along those dimensions that are most useful to subsequent processing. It is very different to state explicitly that, for example, there is a formant transition passing between two points in the frequency/time space than simply to represent the signal in a "neural spectrographic" form. The former requires extensive additional processing and important choices about what auditory dimensions to represent. These dimensions must allow not only phonetic classification but also the multitude of para- and non-linguistic decisions that we can make on a speech input, together with all those adjustments for speaker and rate of speech which bedevil recognition algorithms.

One property that a psychologically plausible auditory representation must have is to represent amplitude and spectral change explicitly rather than as a sequence of static events. Two experimental reasons can be given why this should be so:

First, the perceived loudness of a sound depends not only on its intensity but on the changes in intensity that precede and follow it. Jesteadt, Green and Wier (1978) have recently documented this effect which they call the Rawdon-Smith illusion after its co-discoverer (Rawdon-Smith and Grindley, 1935); they find that a rapid rise or fall in intensity is perceptually more salient than a slow change, so that subjects will, under suitable conditions match as equally loud two tones of the same duration and frequency that differ by 13 dB in intensity. Perceptually then, steady-states are (at least partly) defined by their edges, not vice-versa.

Second, the apparent perceptual spectrum of a sound is determined by the changes in spectrum that precede (and perhaps follow) it. Haggard and his colleagues (abstracted in Haggard et al., 1977/8) have shown that a flat spectrum can sound like, for example, [ɪ] if it alternates with a sound whose spectrum is the complement of [ɪ] (having zeroes where [ɪ] has poles).

As well as representing change explicitly, the auditory representation must allow auditory properties to be defined relative to a particular sound source. Silence, for example, is not absolute but rather a property of an assumed source. If a continuous formant pattern is perceptually divided into two assumed speakers by rapid alternations in pitch (Nooteboom, Brokx and de Rooij, 1976) then each speaker appears silent while the other is speaking and, with suitable choice of formant patterns, this perceptually induced silence can cue stop consonants (Darwin and Bethell-Fox, 1977).

Finally, Darwin wanted to make it clear that he finds the interaction between psychological theory and computer algorithm extremely stimulating. It is too easy for someone working with synthetic speech as a tool for investigating human perception to equate the auditory or phonetic dimensions used by the brain with the control parameters of his synthesizer. Trying to write an algorithm to detect, say, voice-onset time is a sobering experience for anyone used to generating beautiful synthetic continua. Algorithms applied to large quantities of natural speech are an invaluable complement to the necessarily restricted psychological experiment.

But if such joint perceptual and computer endeavours are to produce a theory of speech perception rather than a pot-pourri of micro-theories, each concerned with particular phonetic distinctions, we need to be more concerned with the general constraints on speech sounds. What is it that lets us hear as an additional extraneous noise the badly synthesized part of an utterance? Or what allows us to hear speech through a masking pattern that, on a spectrogram, deceives the eye (Liberman and Studdert-Kennedy, 1978)? The answer for some is in "directly perceiving" the articulation, but we are a long way from being able to write an algorithm that can directly perceive.

References

Darwin, C.J. and C.E. Bethell-Fox (1977): "Pitch continuity and speech source attribution", Journal of Experimental Psychology: Human Perception and Performance 3, 665-672

Jesteadt, W., D.M. Green and C.C. Wier (1978): "The Rawdon-Smith Illusion", Perception and Psychophysics 23, 244-250

Haggard, M.P., G. Yates, M. Roberts and Q. Summerfield (1977-8): "Onset and offset spectra in the analysis of complex sounds", Annual Report 1-2, M.R.C. Institute of Hearing Research, Nottingham, U.K., 12-13

Klatt, D.H. (1977): "Review of the ARPA speech understanding project", JASA 62, 1345-1366

Klatt, D.H. (1979): "Speech perception: a model of acoustic-phonetic analysis and lexical access", J. Phonetics 7

Liberman, A.M. and M.G. Studdert-Kennedy (1978): "Phonetic perception", in R. Held, H. Leibowitz and H.L. Tenber (eds.) Handbook of Sensory Physiology VIII, "Perception", Heidelberg Also in Haskins SR-50, 1977, 21-60

Marr, D. (1976): "Early processing of visual information", Phil. Trans. Roy. Soc. B. 275, 483-524

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

Rawdon-Smith, A.F. and G.C. Grindley (1935): "An illusion in the perception of loudness", British Journal of Psychology 26, 191-195.

Dennis Fry expressed his admiration for Michael Studdert-Kennedy's report and for the amount of ingenious experimental work covered by the report. He only wanted, as a supplement, to put forward what he considered to be some brute facts about speech perception seen from the point of view of the acquisition of speech. All reporters mentioned this as an important aspect, but only in passing.

The first fact is that the child always proceeds from the referent to the sound distinction, never the other way about. He is paying attention to something in his environment and that gives him the motive to notice a sound distinction. Therefore this use of acoustic factors probably depends very much on an attentional factor, perhaps more than on the capacities for making these distinctions (cf. Carney and his co-authors).

The second fact is that the child evolves his own acoustic cues. It is essential to remember that every individual is free to evolve his own cues. The only constraint is that they must lead him to the right decision, that is to say to be able to recognize the word or whatever it is that has come in.

This means that the child attempts to learn to deal with the phonetic or perceptual space which is engaging his attention, not the whole phonetic perceptual space, and he starts with very simple cues, expanding the system of cues, that is, developing a larger and larger part of the possible phonetic perceptual space as the different references and the distinction between them make it necessary to do so. - And this whole development goes through reception first. You have to be able to receive, to distinguish, before you begin to produce; there is interaction between reception and production.

Dennis Fry thinks that all this is learnt. The fact that in different languages you get very different modes of dealing with the acoustic input is crucial, and the fact that once you have learnt one language you have difficulty in perceiving distinctions not made in your mother tongue, also shows that these things are learnt. Fry is not convinced of the existence of invariants or of any substratum of universal stuff, perhaps with the exception of the ability to distinguish between silence and sound.

As for the interaction between perception and production we do not keep it sufficiently in mind that every human individual being is hearing a completely unique version of his own sounds. Therefore no human being can make a perfect, and not even a very good match between the sounds he is producing and what he hears from somebody else. It is therefore important that the child develops a cue system which enables him to deal with what comes in. When he sends stuff out, he has only to ensure through his feedback that he is implementing the cues which he is using to listen to somebody else. You have only this amount of match. - Therefore Fry rejected the idea of a motor theory, also in the form that listeners should have to infer something about the vocal tract of the other person. This is not necessary if the whole thing is done on the basis of these cues.

Björn Lindblom showed slides of a distance metric box and of a block diagram of auditory analysis inspired by Manfred Schroeder, which, starting from a harmonic spectrum, converting the frequency scale into a Bark scale, and adding an auditory filter and a masking pattern, leads to two quasi-auditory excitation patterns, a quasi masking pattern and a loudness-density pattern. In accordance with Plomp he thinks that the perceived difference between two static stimuli depends on the area between two curves in the auditory excitation pattern. On this basis he and his co-workers try to explain: (1) the F2' data that have come out of the experiments by Carlson, Granström and Fant (in this respect the results are very positive), (2) Flanagan's difference limen data (which Lennart Nord has had some success in explaining), (3) dynamic events, e.g.: Is a vowel formant target identified better in a

dynamic than in a static context? (Karin Holmgren has found that it is not.) This latter result is not totally in agreement with the point that Darwin made, i.e. that the human speech perception mechanism is primarily sensitive to changes, although Lindblom, generally, agrees completely with this point of view.

Lloyd H. Nakatani agreed that phonetic perception is fundamental to speech perception and that, as Studdert-Kennedy said: "Perhaps all these years of studying C-V syllables have not been wasted after all", but now it is important to concentrate more work on prosody and bring more linguistic facts in. In prosody the cues are complex, and there are great idiolectal differences between talkers. We cannot continue generalizing from the Haskins speech synthesizer to the whole population. In some recent papers in JASA a new technique that attempts to cope with more complex perceptual phenomena has been described.

Dennis Klatt emphasized that you should not set up a dichotomy between phonetic segmentation and the possibility of going directly to larger units, like the word. Both phenomena are well motivated. Phonetic segmentation is supported by the fact that the speech production process manipulates units such as segments, and by the fact that one must have a method for understanding new words. But going directly to the word restricts the phonetic strings to look for and helps solving ambiguities. It also helps to interpret durational cues, because, e.g., stress plays a role. One possibly has to build into our model of the perceptual system kinds of constraints that will make for optimal decisions.

Klatt's second point was that there is no logically necessary connection between feature extraction and phonetic labelling. The features may lead directly to words. One should investigate the feature problem by building very simple models of perception, trying if simple psycho-acoustic distance metrics can be used to make predictions of the sort that are made by phonetic data or not. If not, it points to feature detectors. Probably some of the natural quantal categories will come out of very simple assumptions about the peripheral system and the distance metrics.

The context effects mentioned by Darwin will be troublesome for distance metrics, but this does not prevent a solution. The distance metric is going to be a change-over-time kind of metric.

Osamu Fujimura mentioned a recent study at Bell Laboratories by Marian Macchi treating the role of consonantal transition in perceptual identification of vowels which has been published in Speech Communication Papers edited by J.J. Wolf and D.H. Klatt 1979. In contrast to what Strange et al. reported (JASA 60, 1976, p. 213-24), Macchi's result demonstrates that vowels in isolation can give rise to a very high accuracy of identification when appropriate care is exercised concerning dialectal problems and the possible difficulty in orthography (Macchi used rhyming tasks instead). It is possible that dialects vary considerably in the phonetic characteristics of gliding, even for so-called monophthongal vowels in English, and these gliding effects are particularly important in the case of isolated vowels as opposed to syllables ending in a consonant, because the VC transition in the latter case reduces or perceptually obscures such gliding effects.

Dominic W. Massaro: It is recommendable to utilize an information processing approach in speech perception, because the goal of this approach is to delineate the stages of processes that occur between the acoustic stimulus and the meaning in the mind of the observer. It has been found that even at an early stage of processing where you are taking raw feature information and integrating it together it is necessary to incorporate what the listener knows in terms of speech he or she has heard before, in terms of constraints in the language, and in terms of possible words or nonwords and so on. So even at this early stage we have to develop models that allow the contribution of higher order processes. Rather than opposing bottom-up and top-down processes, what has to be developed are specific formal models that describe the integration of both sorts of information.

As for features Massaro has found that they are not binary. In fact, listeners have knowledge about the degree to which a feature is present in the speech chain.

Pierre L. Divenyi took up the problem of categorical perception as treated by H. Fujisaki. He found that the problem whether perception, and categorical perception in particular, is articulatorily or auditorily bound is an artificial one. In Fujisaki's second stage there may even enter non-speech auditory events. At the higher stage of perception there is no time for a detailed analysis. Categorical perception is a result of applying an

a priori decision process about what to pick from the signal, and this results simply in discrimination peaks and categories.

Steve Marcus argued that intermediate levels between the acoustical signal and the perceived word are only hypothetical constructs. It appears from split-brain studies that in the right hemisphere word recognition is obtained by an acoustic-lexical mapping system. It would be parsimonious to assume that the left hemisphere used the same system, and that the further possibility of the left hemisphere for segmental analysis would be used for special tasks only, such as CV-recognition, rhyme detection and learning of new words. An intermediate stage seems to be necessitated by current work on the combination of acoustic and visual-articulatory cues (lip reading) in speech perception. It would be interesting to examine whether split brain patients can use lip reading.

Secondly, Marcus argued that there is no empirical justification for assuming a phonemic level. It could also be a continuous real time integration, perhaps using some temporal reference points, which may be purely acoustically determined. The fact that initial phoneme detection times are dependent on factors affecting word recognition speaks against the role of phonemes in perception.

Herbert Pilch. Like Sieb Nooteboom H. Pilch regretted that the study of speech perception has been limited to controlled responses to synthetic stimuli. Our goal must be to understand speech perception in routine communication.

Prosodics signal neither syntax nor sentence meanings, but discourse structuring in the rhetorical sense. Monotonous reading fails to achieve communication, whereas intact prosodic performance can outweigh severe aphasic disturbances in phonemes and syntax.

Routine perception works on the basis not of specific linguistic elements (such as phonemes, syllables, words, sentences) but of total messages. Minimal distinctions may be hard to grasp.

The listener may, however, shift the focus of his perception from the total message to any particular element, i.e. perceive the speech signals as (a) a message, as (b) a linguistic structure, or as (c) noise. In case (a) he may miss the message, in case (b) the structure (cf. H. Pilch: Auditory Phonetics, Word (in print)).

James Pickett: Taking up Studdert-Kennedy's hypothesis that we perceive the speech movements directly, Pickett proposed that we should attempt to set up features of movement (What is moving? where is it going? how is it moving? how is it related to preceding and following movements?) and see where it leads.

Adrian Fourcin: Referring to Dennis Fry's contribution Fourcin confirmed that children do indeed go from the recognition of very simple physical features to levels which are more recondite and varied in the spectral form of the signal. So with the voiced-voiceless opposition you go initially, in the earliest years, from three to five, from a skill of discrimination based on whether voicing is there or not, to a skill based on the onset of the first formant as flat or rising.

Children who are totally deaf can learn to produce clear stress contrasts by means of a visual display of auditorily relevant information. Moreover, by using an auditory pattern approach and giving them an electrical stimulation of the cochlea you can teach totally deaf children to make discriminations based on their pattern knowledge and give them a categorical ability to discriminate which is not at all based on any motor references.

But in order to communicate at a fast rate you have to use a sort of parallel processing technique which is necessarily dependent on your knowledge of coarticulatory constraints.

T.M. Nearey reported that Assmann (cf. vol. I, p. 221) obtained the same results as Marian Macchi (see Fujimura's contribution to the discussion), i.e. a much higher recognition of isolated vowels than should be predicted according to Strange et al., when factors of dialect, orthography, etc. were controlled.

Hiroya Fujisaki emphasized that the role of prosody may be quite language specific. Further, he showed a number of slides illustrating his acoustical and perceptual investigation of Japanese accent.

Michael Studdert-Kennedy concentrated his final remarks on four points:

1. The problem of recognizing dynamic vowels against isolated ones is very complicated. O. Fujimura has showed that centers of vowels extracted from running speech are not readily identified and do need the surrounding formant transitions. Percent correct identifications is probably not the most sensitive measure for that question.

2. Studdert-Kennedy had not attempted to argue that we have no acoustic property detectors. Presumably there is some system within the brain that is able to pick up acoustic properties, but the question is whether there is any grounds for supposing that those property detectors are opponent detecting systems, and whether there is any ground for supposing that they have been adapted for linguistic purposes. In this regard he would rather go with Kenneth Stevens and suppose that language is simply exploiting properties of the auditory system rather than the other way around.

3. In answer to Steve Marcus: To what extent you use auditory contours in listening is an open question. But Studdert-Kennedy would give most of Marcus's data an exactly opposite interpretation. For instance, the fact that phoneme recognition comes after word recognition has nothing to do with perceptual processes, it is a question of experimental tasks and of bringing things into consciousness.

4. Studdert-Kennedy found the data on child language acquisition very important, for instance the work by Boysson-Bardies and by Lise Menn. Another field of research which is highly relevant for the problem of speech perception is that of sign language. Many of the processes of acquisition resemble quite closely the processes of acquisition of spoken language which suggests that what we are dealing with is a very general system that is highly flexible and adaptable to a variety of different circumstances.