HIERARCHY OF LEVELS IN SPEECH PERCEPTION

V.B. KASEVICH, Y.M. SHABELNIKOVA

Dept. of Oriental Studies
University of Leningraa
Leningrad, USSR 199034

ABSTRACT

Speech perception is argued to be essentially a top-down process coming down stepwise from higher levels to lower ones, the higher levels being characterizable, from the phonological point of view, in terms of their prosodic features.

In Donald Norman's words, "it is usually thought that the analysis of speech requires levels of abstraction. For example, speech sounds are transformed into phonemes, and phonemes into words"/9, p.388/. The analysis of this kind, nicknamed'bottom-up', until very recently was accepted almost universally. The only alternative was presented by the one-time influential analysis-by-synthesis model which typically did not make use of the notion 'level of abstraction'. In other words, the predominant views link the very notion of levels to the more traditional bottom-up model, while its top-down counterpart,the analysis-by-synthesis model, is not thought to require the notion. The reason seems to be rooted in a rather narrow understanding of what levels of abstraction may be like: it is believed by many that at the outset of the process of speech perception man deals with the phonetically 'richest', i.e.the least abstract,characteristics of the incoming signal, the latter being step by step 'refined' so as to dispose of communicatively irrelevant details.

Yet the psychology of vision /II/ as well as the still earlier ideas of N.A. Bernstein /I/ suggest a valuable hint to the contrary. It is argued that at the first stages of visual perception man deals with highly generalized, and therefore abstract, features of the object to be perceived. Such features are sufficient to get a 'rough idea' of what is being seen, i.e. to assign the object to a very broad class of entities. If the actual setting is informative enough to provide ground for (subliminal) choice within the class, the object is recognized with all its relevant details without further analyzing its actual physical characteristics. If not, its lower-level features, less abstract and more specific, are to be brought into consideration until the information is made sufficient to identify the object.

It may be seen from the above very sketchy exposition that visual perception exhibits a clearly top-down character. On the other hand, it is hardly in line with the analysis-by-synthesis model -- at least not beyond the anticipation routines common to all perception strategies. It seems to be of primary importance that the perceptual process evolves as a stepwise progressing from a more abstract representation of an object to a more specific (concrete) one. That means, at the same time, that the process is hierarchical in nature. Levels of abstraction are also levels of control where the output of a higher level largely constrains and, for that matter, controls the functioning of the lower one.

If auditory perception is presumed to be essentially parallel to visual perception, then we may accordingly seek similar stages in processing the sound information. One of the crucial problems is singling out particular sound features to be assigned to higher levels of speech perception. Since in processing the speech signal the listener aims from the very outset at grasping its meaning, the features sought should be applicable to as big speech chunks as possible. Clearly such features are most likely to be prosodic (suprasegmental), i. e. pertaining to intonation, stress (accent) or tone.

One possible method of investigating the relative role of suprasegmentals and segmentals (syllables, vowels, consonants) in speech perception is artificial distortion of certain acoustic parameters responsible for realization of particular segmentals or suprasegmentals, which gives an opportunity to see their contribution to the process. For instance, Price and Levitt /IO/ report that insufficient prosodic information makes the /š/ - /č/ distinction unstable. They speak of prosody as of "an aid in initial parsing of a sentence" (p. 302). In our view, such data point to an hierarchically higher role of prosody as oompared to that of segmentals.

Bosshardt /3/ has observed that if parts of test sentences are interchanged (cf. Der Student schreibt seine Arbeit in den kalten Dachstube → In den kalten Dachstube der Student schreibt seine Arbeit → Seine Arbeit der Student schreibt in den kalten Dachstube) the changes affect comprehension because of the "perturbations in the suprasegmental information" (p.I93).

Krulee et al. /8/ argue that the prosodic system is "responsible for the segmentation of continuons speech into sentences, phrases, and words. It attempts to establish a context within which a second system, responsible for the processing of words and syllables, can operate"(p.53I). This, again, is tantamount to recognizing an hierarchy of levels in speech perception, albeit somewhat narrowly understood, for the interaction of prosody and "a second system" is reduced to segmentation and identification respectively.

In what follows we sought to bring to light the relative role of suprasegmentals and segmentals by experimentally changing the acoustic parameters responsible for their natural relationship.

The present writers' previous findings obtained for Chinese and Vietnamese /6; 7/ have shown that syllable tones and segmentals behave differently in the process of speech perception. When 'monotonized' by means of vocoder technique, i.e. deprived of the fundamental frequency differences responsible for the identity of lexical tones, the syllables are recognized considerably worse. The confusion of our 'toneless' syllables taken at its face value would be a trivial fact, for tones are known to be unalienable features of each syllable in Chinese or Vietnamese. Yet, less trivial is the quite consistent confusion of the test syllables' segmental components whose acoustic characteristics were kept intact in our experiments. On the other hand, white-noise masking(signal/noise ratio 0 dB), while resulting in a drop in recognition scores for segmentals, practically does not impair those for tones. The latter results, althogh quite natural, seem to be of crucial importence: in their absense one might be tempted to argue that any interference with the speech signal, whatever its nature, may lead to a poor recognition of the signal in its totality. However, our data seem to indicate that the recognition of tones is a prerequisite to that of segmentals -- but not the other way round. In other words, there must be an hierarchy in the processing of segmentals and tones which consists in the following: syllables are classified into tonal categories and the identification of segmentals is carried out by the listener within thus previously determined tonal classes of the syllables. When this natural process is made impossible because of inavailability of the tonal information, the listener finds himself at a loss facing the necessity of discriminating between all the possible syllables instead of using the overlearned strategy of operating within a tonally delineated class. Hence the multiple confusions reflected in our data.

The general aim of the experiments reported in this paper is to find out whether the process of speech perception in stress languages such as Russian can be visualized as structurally similar to what has been observed for tone languages. If the hierarchical relationships of perceptual strategies for stress languages do parallel those observed for tone ones, then the place of syllable tones of the latter should be taken up by word stress patterns of the former, for functions of the two are alike: in tone languages a tone marks prosodically a syllable-morpheme, the basic operative unit of the languages /5/, whereas in Russian-type languages it is the word prosodically patterned by means of stress that is the basic unit of a functionally similar type.

For obvious reasons, it is more difficult to artificially deprive syllables of a stress-language word of their prosodic features than to make Chinese or Vietnamese syllables 'toneless'. That is why we resorted to a different experimental technique, viz.: the final and the initial unstressed.CV syllables of two conjoined two-syllable words spoken without a pause were cut out so as to make one pseudo-word with both its syllables unstressed, e.g. X'IMU out of the sentence /duX'I MUžej n'e pugal'i/ 'The ghosts did not frighten our husbands'. As a result, we obtained pseudo-words composed of two stressless syllables. Similar pseudo-words, this time trisyllables, were obtained by means of mutual (crosswise) transplanting of the syllables, e.g. out of the two sentences: /a vot japonskaja p'evica cús'ita/ and /a vot japonskaja p'evica cus'íta/ 'Here is Tsushita, the Japanese singer', a third one was constructed, where all the constituent syllables of the word /cus'ita/ were made unstressed. All in all, about 20 isolated disyllables and the same number of trisyllables in carrier sentences were presented to 20 phonetically untrained subjects. The subjects were asked to write down what they heard without leaving blanks. In both experiments (with di- and trisyllables) recognition scores for vowels and consonants in 'stressless'words have been found to drop noticeably down to 60-70 and 76-83 per cent in di- and trisyllables respectively.

A word of caution would be in order here. The expression 'stressless', 'unstressed' when applied to words, should not be taken without some reservation. The reason is that the syllables of our pseudo-words are really stressless only with respect to their 'original' words. Within the artificial test words, the syllables, from a perceptual point of view, are to form a new syntagmatic hierarchy of their own, because for a Russian speaker/hearer independent 'stressless'words simply do not exist. In short, stress patterns of the pseudo-words are not absent but rather distorted. The same can be said of the 'monotonized' Chinese and Vietnamese syllables referred to above.

Turning back to the recognizability data, we must admit that the low intelligibility of the test words might be partly attributable to less pronounced syllable contrasts which is characteristic of any unstressed syllable /2/. Yet, there seems to be every reason to believe that what has most affected the performance of our subjects is the distortion of the stress pattern, the identification of the latter thus being an important precondition for the recognition of segmentals -- vowels and consonants. In other words, the situation appears to be similar to that observed for tone languages. It was more than once suggested in literature, that prosodic information is processed independently of that concerned with vowels and consonants. What is no less important, prosodic information seems to be processed prior to segmental information, which means, again, that perception procedures are hierarchical in nature: words appear to be classified first according to their stress patterns, and only then the identification of the words, including the segmental composition, is carried out within the previously determined stress-pattern classes. Both in tone and stress languages the interrelationship of such hierarchical processes exhibits the top-down direction, if we take suprasegmentals as belonging to higher levels of the phonological component and segmentals to lower ones.

In experiments with trisyllables it has been also observed that the effect of stress pattern distortion on recognition of vowels and consonants is the least if the 'de-stressed' word is to be found in final position where sentence stress is typically located. The data are reminiscent of the findings for measuring reaction times in phoneme monitoring experiments as reported by Foss et al. /4/. According to the authors, the targets on words bearing sentence stress are responded to more rapidly than targets on words outside the sentence stress. All such results seem to indicate that the perception of segmentals is dependent not only on prior identification of lexical stress pattern but also on that of sentence intonational pattern (cf. Bosshard's experiment referred to above).

In order to further probe into the nature of the hypothesised relationship, we designed one more experiment where distorted was the sentence intonational pattern. Out of two sentences of the type /p'et'a igrajet na g'itar'e/ and /p'et'a igrajet xarašo/ the third was constructed by means of interchanging and transplanting the words, namely, /na g'itar'e p'et'a igrajet xarašo/ 'The guitar, Pete plays well'. Similar test sentences were constructed with meaningless words, the words and the sentences being modeled after their meaningful prototypes, cp. /na g'itar'e p'et'a igrajet xaraso/ and (meaningless) /na d'ikal'e t'ep'i udlor'it šalaso/. 45 such sentences were presented to the seme team of subjects, first I5 meaningless test sentences, then their I5 meaningful counterparts, and, last, control original sentences.

The results show that the intonation pattern distortion equally impairs recognizability of vowels and consonants. The effect is especially significant for trials with the meaningless sentences which show 60-70 per cent recognition for segmentals.

Our data do not provide a sufficient basis for determining whether recognition of segmentals is directly dependent on identification of intonation contours or segmentals and intonational patterns are associated via lexical stress (the latter option seems preferable).

As can be seen, the 'top' -- i.e. the higher level -- is not identified here directly with semantics. At the same time, it is precisely because of the immediate association with communicatively relevant meaningful units such as sentences, phrases, words, that prosody takes up the role of starting point in speech recognition processes.In fact,prosodic description,i.e.a description in terms of suprasegmentals, serves to provide a first-approximation abstract representation of the meaningful unit to be recognized. The representation is open either to direct semantic interpretation (if the context is highly suggestive) or to further elaboration and transformation by means of bringing into play additional low-level information about segmentals.

We are not going to argue that speech perception is a unidirectional, strictly serial process. Highly plausible is the existence of modules operating in parallel. Besides, if the initial hypothesis about a word or, say, phrase is rejected as violating some regularities of mapping prosodic structures onto segmental ones the process is started anew -- thus acquiring a shuttle-like character in its functioning.

REFERENCES

I. N.A.Bernstein, The Coordination and Regulation of Movements. London: Pergamon Press, I967.

2. L.V.Bondarko, The syllable structure of speech and distinctive features of phonemes . - Phonetica, I969, vol.20, N I, p.I-40.

3. H.-G.Bosshardt, Suprasegmental structure and sentence perception. - In: H.W.Dechert and M.Raupach (eds.),Temporal Variables in Speech. Studies in Honour of Frieda Goldman-Eisler. The Hague-Paris-New York, I980,p.I9I-I98.

4. D.Foss, D.A.Harwood, and M.A.Blank, Deciphering decoding decisions: Data and devices. - In: R.A.Cole (ed.), Perception and Production of Fluent Speech. Hillsdale: Lawrence Erlbaum Associates, I980, p.I65-I99.

5. V.B.Kasevich, Phonological Problems in General and Oriental Linguistics. Moskow: Nauka Publishers, I983 (in Russian).

6. V.B.Kasevich and E.M.Shabelnikova, Segmentals and suprasegmentals in speech perception. - In: Proceedings of the 9th International Congress of Phonetic Sciences. Copenhagen, I979.

7. V.B.Kasevich and E.M.Shabelnikova, Tempo, rhythm, and the choice of strategy in speech perception. - In: Abstracts of the I0th International Congress of Phonetic Sciences. Dordrecht, I983.

8. G.K.Krulee, D.K.Tondo, and F.L.Wightman, Speech perception as a multi-level processing system. - Journal of Psycholinguistic Research, I983, vol. I2, N 6, p.53I-554.

9. D.Norman, Copy-cat science or does the mind really work by table look-up? - In: R.A.Cole (ed.), Perception and Production of Fluent Speech, p. 38I-395.

I0. P.J.Price and A.G.Levitt, The relative role of syntax and prosody in the perception of the /š/ - /č/ distinction. - Language and Speech, I983, vol.26, pt.3, p.29I-303.

II. I.Rock, An Introduction to Perception. New York - London: Macmillan Publishing Co., Inc. - Collier Macmillan Publisher, I975.