Phonetica
Editor: K. Kohler, Kiel

Separatum
Publisher: S. Karger AG, Basel
Printed in Switzerland

Phonetica 1995; 52: 131–143

Walter F. Sendlmeier

FB 1:
Kommunikationswissenschaft,
Technische Universität Berlin,
Deutschland

Feature, Phoneme, Syllable or Word: How Is Speech Mentally Represented?

Abstract

Four experimental approaches frequently used in speech perception research are discussed with respect to their impact on word recognition models and their implicit assumptions on the mental representation of speech. These approaches are (1) reaction time experiments; (2) the procedure of click localisation; (3) the method of selective adaptation, and (4) the assessment of word similarities. The results of the studies vary as a function of the experimental procedure chosen. Phonetic features, single sounds, syllables and words as entities are alternatively favoured as primary perceptual units. A critical evaluation and an attempt at integrating the data lead to the assumption that the adult speaker/listener has different kinds of mental representation of speech at his/her disposal. Depending on the focus of perception, units of different sizes are primarily focused in the recognition process. This implies that the listener is able to modify his/her temporal analysis window to a certain extent. Nonetheless, as a default case, the syllable serves as the primary perceptual unit.

1. Introduction

It is only during the past two decades that any appreciable attention has been paid to the mental representation of speech. In this article, the question will be examined of how speech is mentally represented with respect to lexical processing. In the complex process of lexical access by humans, the decisive factor is the extraction of meaning from the speech signal. Psycholinguists and phoneticians who are interested in perception try to elucidate the

Prof. Dr. W. Sendlmeier
Technische Universität Berlin
FB 1: Kommunikationswissenschaft,
Sekr. EN8, Einsteinufer 17
D-10587 Berlin (Germany)

0031–8388/95/
0523–0131 $ 8.00/0

question of how this can be so easily managed by humans. Engineers try to construct machines which are able to recognize speech input nearly as well as humans. The process of lexical access is of central importance to the work of both groups of scientists. Since the utterances a person is required to process are infinitely variable, the meanings of all potential utterances cannot be stored in a person's memory in order to relate these to an acoustic input. Instead, what has to be stored are the meanings of discrete units utterances consist of. For the sake of simplicity, these units can be called 'words'. The part of the memory where the sound of words is connected with the meaning is called the 'lexicon' [see also Cutler and Norris, 1988].

Scarcely any of the leading models of word recognition contains explicit information about the phonetic and mental representation of words in the lexicon. Almost all models, however, contain more or less precise information about primary perceptual units to which – at least implicitly – the status of mental representation is attributed. According to some models, the code of access consists of distinctive features or phonetic segments like phonemes or allophones; according to others, it consists of syllables; and according to yet others, words are held to be represented as holistic entities without regarding any segmentation within word boundaries as necessary for the word recognition process. Accordingly, the problem of phonetic mental representation is closely linked to the question about the basic units of the auditory perception of speech.

Before investigating in detail the various approaches and their experimental examination, the term 'mental representation' has to be defined. When cognitive psychologists use the term 'mental representation' they start from the assumption that the human information processing system receives information from its surroundings, stores these data, if necessary transforms them, and eventually shows observable behaviour on the basis of the thus stored information. Mental representation refers to conditions which are internal to the system and which are believed to reflect external conditions [Engelkamp and Pechmann, 1988].

How do we get to know further details about the mental representations of phonetic phenomena? In other words: how is the phonetic aspect of lexical units represented in a speaker/listener? This question, which touches exactly the interface between phonetics and psycholinguistics, will not be discussed relating to physiology or introspection, but in terms of experimental observation of behaviour as a reaction to speech input. The objective is to identify systematic correlations between speech stimuli and specific listener reactions which can be explained by the assumption of certain mental representations in the listener. For this purpose, different approaches to experimentally examining the mental representation of linguistically defined units in speech will be discussed in the context of models of spoken word recognition. It will be argued that listeners have mental representations of various sizes available to them during perception, and that the size of representational units used by listeners in any given experiment depends upon the demands of the task. It will also be argued that the syllable is the default perceptual unit.

2. Sublexical and Lexical Representations

Many phoneticians and psycholinguists agree that sublexical representations play an important role in the process of word recognition. It is generally assumed that the processing of sublexical units is able to simplify the process of segmentation. There are two pos-

sibilities: either segment boundaries can be seen as a starting point for lexical access, or lexical hypotheses can be restricted to those word candidates that begin with certain sublexical units. In both cases, a superfluous classification of inappropriate lexical entries is avoided.

Nonetheless, a considerable range of viewpoints has been put forth about the precise function of sublexical units. For the following discussion, it seems useful to distinguish between two fundamental concepts. One view assumes a standardized sublexical representation, whereas the other view is based on the supposition that multiple representations work as intermediate constituents during the process of word recognition.

Within the concept of a unitary type of sublexical representation, two linguistic units in particular are proposed as constituents. Sublexical representation is attributed either to the phoneme or to the syllable. The cohort model developed by Marslen-Wilson and Welsh [1978] is an example for a phoneme-based model. It is based on the assumption that the speech signal is sequentially processed in discrete sounds and that, in a separate step, this sequence is related to lexical representations. The lexical representations themselves are supposed to be composed of linearly concatenated segments. In principle, the approach of Mehler [1981] is very similar. He, however, regards the syllable as the constituent of the sublexical representation, for native speakers of French in particular.

Contrary to this view of unitary sublexical constituents, in a series of models recently developed, the interface between acoustic/phonetic and lexical processing is regarded as more complex. Different types of perceptual units are distinguished: segmentation and classification units for the analysis of the speech signal, and access units which are necessary to make contact with the lexicon.

Cutler and Norris [1988], for example, presented a model in which the process of segmentation takes place independently with respect to linguistic/phonetic classification and lexical access. According to their theory of word recognition, stressed syllables are the primary unit for lexical segmentation. Metrically strong syllables – roughly described as syllables that do not contain a reduced vowel – serve to determine the word boundaries. If a syllable is identified as strong by the listener, lexical hypotheses are activated that begin with this syllable. These syllables, however, do not function as classification units; the classification is performed with the help of subsyllabic units – possibly phonemes – which have the function of referring to the lexicon.

In contrast to the already mentioned models, which presume that most of the problems of speech recognition are solved by assigning certain segments of the speech signal to sublexical units, other models concentrate on the lexical level. This becomes very obvious in the 'Lexical Access from Spectra-Model' developed by Klatt [1979], for example. The basic assumption of Klatt's model is a direct mapping of the acoustic/phonetic input to a once established spectral sequence decoding network structure. The input waveform is analysed by computing a spectrum every 10 ms; a sequence of such spectra is compared with the spectral templates of the network. A word is identified by finding the path through the lexical decoding network that best represents the observed input spectra. Klatt pointed out that the strategy of not making any segmental decisions below the word level can help to avoid mistakes in classification and segmentation which could lead lexical processing into the wrong direction. There is no feature detector stage in his model either.

3. Experimental procedures

How have scientists tried to prove the adequacy of the assumed primary perceptual units? This problem has been dealt with in different ways. Apart from mere linguistic descriptions of diachronic and synchronic character, observations from first- and second-language acquisition as well as phenomena of sound confusion during the process of speaking and listening, and finally experimental studies have been used.

In attempting to deduce falsifiable hypotheses from these distinct models and to investigate them by empirical methods, one encounters a number of problems. The first problem is to design suitable experimental procedures which make it possible to analyse the process of word recognition listeners experience as immediate and integral in its parts with regard to time and function.

Another problem concerns the interpretation of results. The aim is to generalize from the behaviour of listeners in these experiments to the behaviour of listeners in everyday situations. The reason why this is so difficult is that most experimental procedures do not investigate the process of word recognition directly but through tasks that contain additional and possibly alternative strategies of processing [Sendlmeier, 1989]. The range of results which have been obtained by different experimental procedures makes the derivation of general conclusions about lexical access difficult. The conclusions drawn from the individual studies can, for the time being, only refer to the information processing of subjects confronted with the very specific tasks in those experiments.

In the following, four different experimental approaches and some of the most important results of these experiments will be dealt with. First, reaction time experiments will be described which are very popular in psycholinguistics. Second, the procedure of click localisation will be discussed, thirdly the method of selective adaptation, and finally a study that deals with the assessment of word similarities.

3.1 Reaction Time Experiments

In reaction time experiments, also known as 'monitoring tasks', listeners are asked to detect previously specified target units very quickly, such as, for example, single sounds or syllables in specific carrier items. These carrier items generally consist of words or meaningless syllables which are presented acoustically as a list or sentence. The basic idea of this experimental procedure rests on several assumptions. It is assumed, for example, that subjects construct or rather activate an internal representation of the target sound in order to manage such a task. This representation is then kept active during the analysis of the carrier items. Furthermore, the listener is purported to initiate a response when recognizing sufficient correspondence between an incoming signal and the representation of a target sound in order to indicate the detection by his reaction, that is by pressing a button. The reaction latencies are supposed to indicate the mode of the representation of speech as well as the temporal course of its processing. Two kinds of monitoring tasks can be distinguished: in the first approach, characteristics of the target elements, particularly their size, are manipulated while the carrier units are kept constant. In the second approach, the characteristics of the carrier items are modified while the target unit is kept constant.

3.1.1 Variation of the Target Unit. Savin and Bever [1970] were the first to make use of monitoring tasks in order to study the psychological reality of phonological categories in the process of lexical access. In a pioneering study they varied the size of the target units and observed a quicker detection of syllables than of phonemes. From this they drew the

conclusion that phonemes are perceived only on the basis of already analysed syllables. In an analogous procedure, Foss and Swinney [1973] additionally found that words are recognized more quickly than syllables. The perception-orientated interpretation of these reaction time experiments was criticized by Norris and Cutler [1988], who regarded the quicker reaction time for larger target units in contrast to smaller ones as an artefact. They argued, for example, that in experiments with larger target units the subjects had more information at their disposal with regard to the specification of the target units and could therefore respond more quickly. This interpretation is open to further discussion. What seems less debatable, however, is the point that the reaction times for the different size units are comparable, independent of size, and that it is the discrepancy in size between the recognition unit and the carrier unit which increases the reaction time [see also Barry, 1980].

3.1.2 Variation of the Carrier Items. The second type of monitoring tasks is such that the target unit remains constant and the characteristics of the carrier items, for example the syllabic or prosodic structure, is varied. Treiman et al. [1982] compared the reaction times for the detection of single consonants in meaningless syllables with a varying degree of complexity of the consonant structure at the beginning of a syllable. They found shorter latent periods when the target unit corresponded to the initial phoneme in the carrier item with a simple consonant-vowel beginning than when the target unit appeared in complex initial consonant clusters. The authors attributed the longer reaction times for target units occurring in consonant clusters to the additional processing which is necessary to segment the more complex syllable initial cluster into its phoneme constituents. This was regarded as evidence for the onset of a syllable to function as a unit of speech perception, regardless of whether it consists of a single consonant or of a consonant cluster. Cutler et al. [1987a] showed, however, that the delay in detecting a consonant in an onset cluster is not due to the processing required to divide the cluster.

3.1.3 Language-Specific Differences in Lexical Processing. In addition, the question was raised whether language-specific differences can be observed in monitoring tasks. Cutler et al. [1986] presented English and French stimuli to each of their English- and French-speaking subjects. The native speakers of French showed quicker reaction times for both languages when the target unit corresponded to the first syllable of a carrier word. For example, the French-speaking subjects detected the target unit /pal/ more quickly in the word 'palmier' than in the word 'palace'. The target unit /pa/, however, was detected more quickly in the word 'palace'. The reactions of the English-speaking listeners did not reveal such a syllable effect.

Cutler et al. [1986] attributed this result to the differences in the syllable structures of both languages. The syllable boundaries in French are regarded as definite and clear whereas in English this is not the case. In lexical processing, French-speaking listeners make use of their knowledge of phonotactic restrictions in order to subdivide the speech signal into syllabic units. The authors held that the fact that in English a segment may belong to two phonetic syllables makes syllabic segmentation difficult and unreliable; as a consequence, English listeners do not make use of a strategy of syllabic segmentation.

3.2 Click Experiments

The so-called click experiments are another experimental approach to determine sublexical constituents and their significance in the word recognition process. In this procedure, subjects are confronted with the task of localizing short disruptive impulses – clicks –

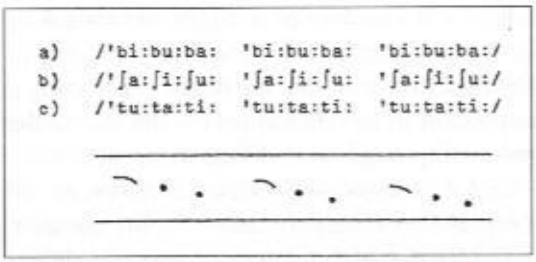

Fig. 1. Three examples of the utterances and their corresponding intonation contour used in the click localisation experiment by Barry. From Barry [1980].

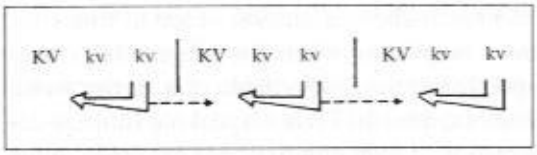

Fig. 2. Influence of stress and group coherence on the shift in click localisation. From Barry [1980].

which were artificially added to the speech signal. This procedure was used by Ladefoged and Broadbent [1960] for the first time and was also employed in a study by Barry [1980, 1984] in which certain phonetic categories are investigated as to their status as primary perceptual units. Barry worked with meaningless utterances consisting of nine syllables which corresponded to the phonotactic system of the German language. Examples of such utterances are given in figure 1.

The utterances were prosodically subdivided by rhythm and intonation into three trisyllabic groups, each with stress on the first syllable and falling intonation. On the second track of the tape recordings, a click – consisting of a very short and abruptly starting sinusoid of 1,000 Hz and 100 dB – was placed to coincide with the middle of the vowels or consonants. The listeners were asked to exactly localize the click in each utterance.

The most important results were: when asked to localize the clicks on the level of single sounds, the subjects gave correct answers only in 55% of the cases, whereas on the syllabic level the correct answer rate was 85%, and on the prosodic level (within the stress group) it was as high as 93%. This result already indicates the difficulties in assigning exactly placed clicks to single sounds and at the same time demonstrates the higher degree of integrity of the units syllable and stress group. One might argue that these results do not necessarily indicate that syllables and stress groups are 'psychologically real units', because it seems hardly surprising that detection accuracy is higher for stress groups than for syllables than for segments, simply because stress groups are bigger than syllables, which are bigger than segments.

However, in analysing the localization errors on the single sound level it is remarkable that consonant click localizations were delayed to a much higher degree than anticipated; thus, in most cases they were shifted to the following vowel. In contrast, vowel clicks were almost always anticipated when incorrectly localized. Due to this symmetry the numerous segmentation errors only very rarely led to syllable mistakes. Thus one can deduce a strong integrity of the simple consonant-vowel-syllable which certainly is based on the acoustic structure of this syllable type with the known function of formant transitions from consonant to vowel. Similarly, when syllable errors occurred, the click was almost exclusively misplaced within a stress group with the force of attraction of the stressed syllable dominating the coherence of the stress group (see fig. 2).

There was a clear overall tendency to stay within the boundaries of the superordinate unit when there were shifts on the lower level. According to Barry, the number of errors that occurred in click localisation indicates that the single sound does not exist as a primary perceptual unit. Even though the lin-

ear input of the acoustic signal is reflected in an accumulation of errors which are characterized by the shifting of clicks to an immediately adjacent sound, the much higher degree of accuracy of localization as regards the syllable unit, however, points at the click localisation to have taken place only after the syllables were perceived. Unlike the single sound, the syllable as well as the prosodic unit of the stress group have to be regarded as psychologically real units in the first steps of speech processing.

3.3 The Search for Feature Detectors

When looking for primary perceptual units in the process of word recognition, subphonemic units such as distinctive features were also favoured. The search for complex auditory feature detectors started in the early seventies. The experimental approach chosen, however, was not based on neurophysiology but on the psychology of perception. In an experimental procedure known as 'selective adaptation', recognition tasks were carried out which had been used before in studies on categorical perception. To exemplify this procedure, the categorization task for the voiced/voiceless distinction will be explained. For an identification experiment, the two syllables /ba/ and /pa/ constituting the extremes of a voice onset time (VOT) continuum were chosen. Then a series of intermediate stimuli were generated which were located in equidistant steps between the poles of the continuum. Listeners were asked to categorize each of the stimuli as /ba/ or /pa/. They divided the stimuli into two groups in such a way that no gradual transition occurred from one group to another. Instead there was an abrupt change of category. The results of such a study are shown in figure 3. On the abscissa, the ba/pa stimuli with increasing VOT are given; on the ordinate, the percentages of voiced identification responses /b/ are indicated.

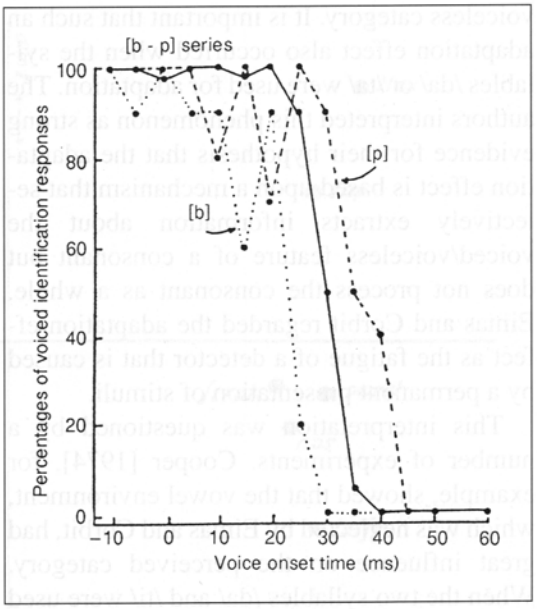

Fig. 3. Percentages of voiced identification responses (/b/) obtained with and without adaptation for a single subject. The solid lines indicate the unadapted identification functions, and the dotted and dashed lines the identification functions after adaptation. The phonetic symbols indicate the adapting stimulus. From Eimas and Corbit [1983, p. 104].

In experiments on selective adaptation Eimas and Corbit [1973] used such series of synthesized stop-vowel stimuli with varying VOT. Beyond the described categorization task, in the adaptation experiment the /ba,pa/ syllables at the two extremes were used as adaptors. One of these syllables was each presented to the listeners for 1 min twice per second and followed by a test item (a stimulus from the VOT series) that had to be categorized. The identification scores including adaptation were compared with those without adaptation. The adaptation with the voiced plosive /ba/ resulted in a shift of the phoneme boundary towards the voiced category by about 5–15 ms. Similarly, the adaptation with the syllable /pa/ resulted in a shift towards the

voiceless category. It is important that such an adaptation effect also occurred when the syllables /da/ or /ta/ were used for adaptation. The authors interpreted this phenomenon as strong evidence for their hypothesis that the adaptation effect is based upon a mechanism that selectively extracts information about the voiced/voiceless feature of a consonant but does not process the consonant as a whole. Eimas and Corbit regarded the adaptation effect as the fatigue of a detector that is caused by a permanent presentation of stimuli.

This interpretation was questioned by a number of experiments. Cooper [1974], for example, showed that the vowel environment, which was neglected by Eimas and Corbit, had great influence on the perceived category. When the two syllables /da/ and /ti/ were used each as adaptors in the two series of stimuli /ba/-/pa/ and /bi/-/pi/, only stimuli with the same vowels showed an adaptation effect. Thus, it became clear that no extraction of features took place on the level of single sounds, but that the overall structure of the stimuli was categorized. Nevertheless, some authors still hold on to the concept of the extraction of context-independent features, as, for example, Stevens [1986] in his 'analysis-through-synthesis' word recognition model.

3.4 Word Similarities

Evoking judgements on word similarities is a further approach to gaining information about the relevance of perceptual units in speech processing. Sendlmeier [1987a] was concerned with the question whether units of different sizes are able to function as primary perceptual units dependent on certain characteristics of the situation of perception. The aim of the experiment was to find out which phonetic dimensions are crucial for the perception of word similarities. Three vocabularies were constructed each consisting of twelve meaningless words in order to avoid interferences between semantic relations and phonetic similarities.

These words were constructed in such a way that no minimal pair occurred within one vocabulary. Each word of each vocabulary was presented twice acoustically with every other word in varying sequence to the listener. The listeners were asked to estimate the pairs of words as to their phonetic similarity on a seven-point graded scale. The similarity judgements were analysed by means of multidimensional scaling (MDS) based on a four-dimensional solution.

The results of an MDS solution is illustrated in figure 4, which shows the first two dimensions for one of the vocabularies which exclusively contained monosyllables. Spatial distances reflect the extent of perceived similarity; items that lie closely together were judged as similar by the listeners, whereas those that are distant from each other were regarded as very dissimilar.

The distribution of words for these first two dimensions can be explained by the quality of the vowels. Along the first dimension – i.e. along the horizontal axis – the words are distributed according to the distinction of front vowels in the left half versus back vowels in the right half. Along the second dimension – i.e. the vertical axis – there is a distribution according to the degree of openness of the vowels. Thus one can conclude that for monosyllabic words the quality of vowels is of primary importance in similarity judgements.

In summary the other results of the study were: the quality of single consonants was of minor importance for the similarity judgements; only for monosyllabic words and here only for the explanation of the third and fourth dimensions, single consonants – when occurring in the same exposed, i.e. initial position – contributed to the explanation of the respective dimension. In more complex stimuli, vowel similarities were of minor influence on

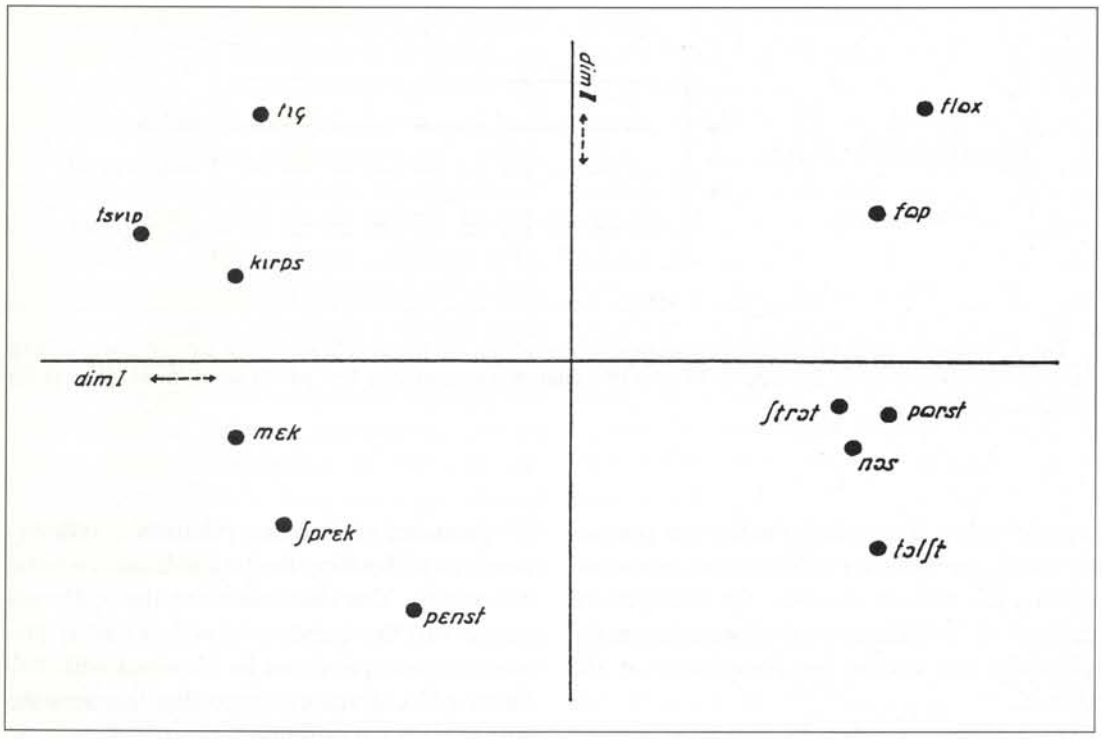

Fig. 4. MDS solution according to the first two dimensions for the relations of similarity of the 12 words of vocabulary I. From Sendlmeier [1989, p. 393].

the perceived word similarity. Features that characterize the words as a whole – such as the number of syllables, presence versus absence of consonant clusters or the patterns of word stress – were of much greater importance than features of single sounds. These results can be seen as an indication to the fact that global properties are increasingly regarded as criteria for word similarity of more complex stimuli.

4. A Model of Phonetic Mental Representation

On examination of all the results of the different experimental procedures, it becomes apparent that in a number of very different test conditions listeners used units of different sizes in the process of word recognition. These results lead to the model of phonetic mental representation illustrated in figure 5.

The underlying assumption is that the adult speaker/listener has several kinds of mental representation at the phonetic level at his/her disposal simultaneously. The most important of these are: the word, the syllable, the phoneme and the phonetic feature. It should be noted, however, that these different units are not different abstraction levels of representation, but different kinds of representation within one level, i.e. the phonetic level. These different kinds of representation are simultaneously at the disposal of the listener/speaker once he/she has established them. The kind of

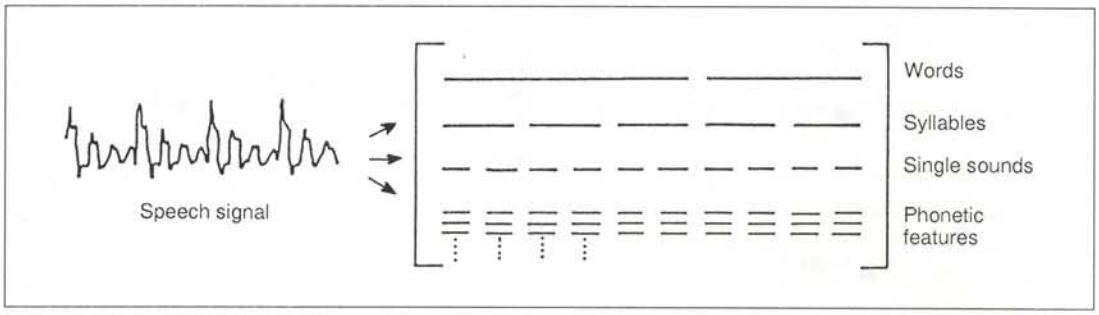

Fig. 5. Different kinds of mental representation of words on the phonetic level which are simultaneously at the disposal of the listener; the listener focuses that kind of representation first which seems most efficient for word recognition.

representation from which the listener primarily takes the relevant information when accessing the lexicon depends, for example, on the type of task, the context of perception, the speaking rate and/or the complexity of the stimuli.

In addition, it seems pertinent to assume that the perceptual activities of a listener vary not only with varying tasks, but that he/she may also interchangeably focus on different kinds of representation while solving one particular task. Thus, a listener can switch to single sounds or even distinctive features when discriminating, for example, minimal pairs or difficult words such as proper names, words of a foreign language or pseudowords, and he/she can then switch back again to the syllable or even word level when progressing in the recognition process [Sendlmeier, 1987b]. In other words, it is held that the listener is able to modify his/her temporal analysis window to a certain extent; he/she will do so to the degree necessary for speech recognition. The demonstration that a listener can detect or manipulate a unit of any size does not necessarily indicate, however, that that particular unit is constructed during normal speech processing. But it does show that the listener can make use of acoustic-phonetic information at that level of granularity (feature, phoneme, syllable, word) in performing the task at hand. To put it differently: The demonstration that different answers to the question of which unit is primary in perception can be obtained with different tasks, gives evidence that listeners are able to attend to different levels of information in speech perception. Those levels are not necessarily all computed during normal recognition, but the data impose that the representations are able to facilitate decision-making at different levels of analysis.

The position that a listener carries on in his/her analysis of the speech signal only as far as necessary is further supported by research by Cutler et al. [1987b]. Using the phoneme-monitoring technique, they observed that both a prelexical and a lexical code can be accessed, one of the decisive factors being the manner in which the target word was processed. The lexical code seems to be accessed when the task requires that the target word is processed semantically. Such semantic recognition can be regarded as the default case in everyday language processing. A prelexical code appears to be used when semantic processing of the word is not required or not possible, e.g. in case of rare proper names, unusual words or pseudowords. These units do

not have a lexical representation the listener can rely on. Thus his/her efforts in processing these items are interpretable as reflecting the operation of attentional processes at a prelexical level. The discrimination between prelexical and lexical representations, though, should not be expressed in terms of the phonology/phonetics relation which – although it has often been the object of sophisticated discussions – is nothing else than a simple type/token relation.

The data presented support the position that information at different granularities can be made available depending on the task demands. Nonetheless, from the experimental results, a default case can be postulated. It seems most likely that the syllable functions as primary intermediate sublexical unit during the process of word recognition. This view is further supported by the fact that the listener can use the continuity of spectral and prosodic information in speech processing only from units of the size of syllables upwards. This auditory streaming is a prerequisite for the robustness of spoken word recognition in cases of a distorted speech signal or competing signals [Sendlmeier, 1985].

Cutler [1976] found that listeners detect a target phoneme faster when it occurs in a monosyllabic word receiving sentential stress than when the syllable is unstressed. This is still true when the local acoustic cues to stress are removed by cross-splicing, indicating that the listener can selectively enhance processing of syllables which are anticipated to be important. These findings illustrate two important facts: first, attention can enhance speech processing, and, second, attention can be allocated as precisely as a single syllable. There is also evidence from language-specific research that the syllable is prominent in speech processing (see 3.1.3). In cross-language studies, English speakers tended to segment speech at the onset of stressed syllables, a lexical segmentation strategy highly effective for English because the great majority of English lexical words in fact begin with strong syllables. For understanding French, such a procedure is less useful, because French has a prosodic structure quite different from the English stress rhythm. The segmentation procedure applied by French listeners is based on the syllable. Although the strategies vary, both document the central role of the syllable in segmenting speech. More generally, it should be noted that stress patterns cannot be determined without a concept of the syllable. Although in some languages the determination of the syllable boundaries is largely uncertain and so some cases of ambisyllabic segments exist, the native speakers of these languages also do have at least implicitly a concept of the syllable in the sense of Tillmann's phenomenological unit [Tillmann and Mansell, 1980]. Thus, it seems well justified to postulate that the syllable as carrier unit in prosody has a crucial function in segmenting and classifying the speech signal in the recognition process.

Closely related to the problem of which size the phonetic perceptual units are is the question of the form of their representation. Here the concept of ideal types, which the Gestalt psychologists established for visual perception [Wertheimer, 1923], or the related concept of prototypes seem to be adequate alternatives to abstract feature matrices. The representation in the form of prototypes is postulated here for all kinds of representations at the phonetic level. It seems likely that in the course of the language-acquisition process a listener generates a prototype in the sense of a statistical mean from all the representatives of a phonetic category ever heard. If one supposes that phonetic units of different sizes – up to words or even up to short phrases – are represented analogously in the form of prototypes, this implies an enormous capacity for

long-term memory. Objections raised by scientists who with reference to – though up to now uncertain – principles of economy argue against such a supposition of storage-consuming representation can be rejected in view of the almost unlimited capacity of the human brain [Penfield, 1969]. It seems plausible that language users develop a language-specific segmentation procedure to exploit language-specific rhythmic regularity. The analytic processes necessary for the development of prelexical and lexical representations will make use of those aspects of the input they find useful. Linguistic rhythm appears to be an extremely obvious and an easily exploitable property to the speech to which the infant is exposed [Cutler et al., 1992]. This view is supported by the fact that speech to infants indeed tends to exhibit a far more marked prosodic structure than adult-directed speech [Fernald and Simon, 1984; Friederici, in press].

All previously described approaches are committed to traditional views, unanimously assuming the memory to be a passive constituent of the human speech-processing system. A totally different approach, which has recently been advocated in the form of connectionist models [e.g. McClelland and Elman, 1986], postulates that the memory and the access to the memory cannot be separated. Memory performances are regarded as patterns that can be found implicitly in combination patterns among a large number of simple processing constituents. The memory is thus regarded as an active part of information processing. This approach, however, does not contradict the view put forth in the above model, which assumes that listeners are flexible in focussing phonetic units of different sizes as primary perceptual units.

Acknowledgements

I am indebted to Anne Cutler and William Barry for their helpful comments on an earlier version of the manuscript.

References

Barry, W. J.: Die Verarbeitung akustischer Information in der lautsprachlichen Wahrnehmung (Institut für Phonetik und digitale Sprachverarbeitung, Kiel 1980). Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK), vol. 13.

Barry, W. J.: Segment or syllable? A reaction-time investigation of phonetic processing. Lang. Speech *27:* 1–15 (1984).

Cooper, W. E.: Contingent feature analysis in speech perception. Percept. Psychophys. *16:* 201–204 (1974).

Cutler, A.: Phoneme-monitoring reaction time as a function of preceding intonation contour. Percept. Psychophys. *20:* 55–60 (1976).

Cutler, A.; Mehler, J.; Norris, D.; Segui, J.: The syllable's differing role in the segmentation of English and French. J. Mem. Lang. *25:* 385–400 (1986).

Cutler, A.; Butterfield, S.; Williams, J. N.: The perceptual integrity of syllable onsets. J. Mem. Lang. *26:* 406–418 (1987a).

Cutler, A.; Mehler, J.; Norris, D.; Segui, J.: Phoneme identification and the lexicon. Cogn. Psychol. *19:* 141–177 (1987b).

Cutler, A.; Norris, D.: The role of the strong syllables in segmentation for lexical access. J. exp. Psychol. hum. Percept. Perform. *14:* 113–121 (1988).

Cutler, A.; Mehler, J.; Norris, D.; Segui, J.: The monolingual nature of speech segmentation by billinguals. Cogn. Psychol. *24:* 381–410 (1992).

Eimas, P. D.; Corbit, J. D.: Selective adaptation of linguistic feature detectors. Cogn. Psychol. *4:* 99–109 (1973).

Engelkamp, J.; Pechmann, T.: Kritische Anmerkungen zum Begriff der mentalen Repräsentation. Sprache Kogn. *7:* 2–11 (1988).

Fernald, A.; Simon, T.: Expanded intonation contours in mothers' speech to newborns. Dev. Psychol. *20:* 104–113 (1984).

Foss, D. J.; Swinney, D. A.: On the psychological reality of the phoneme: Perception, identification and consciousness. J. verbal Lern. verbal Behav. *12:* 246–257 (1973).

Friederici, A.: Biologische Grundlagen von Spracherwerb und Sprachverarbeitung; in Sendlmeier, Mentale Repräsentation. Z. Semiot., in press.

Klatt, D.: Speech perception: a model of acoustic-phonetic analysis and lexical access. J. Phonet. *7:* 279–312 (1979).

Ladefoged, P.; Broadbent, D. E.: Perception of sequence in auditory events. Q. J. exp. Psychol. *12:* 162–170 (1960).

Marslen-Wilson, W.; Welsh, A.: Processing interactions and lexical access during word recognition in continuous speech. Cogn. Psychol. *10:* 29–63 (1978).

McClelland, J. L.; Ellman, J. L.: The TRACE model of speech perception. Cogn. Psychol. *18:* 1–86 (1986).

Mehler, J.: The role of the syllable in speech processing: infant and adult data. Phil. Trans. R. Soc. Lond. B *295:* 305–333 (1981).

Norris, D. J.; Cutler, A.: The relativ accessibility of phonemes and syllables. Percept. Psychophys. *43:* 541–550 (1988).

Penfield, W.: Consciousness, memory and man's conditioned reflexes; in Pribram, On the biology of learning (Harcourt, Brace & World, New York 1969).

Savin, H. B.; Bever, T. G.: The non-perceptual reality of the phoneme. J. verbal Leran. verbal Behav. *9:* 295–302 (1970).

Sendlmeier, W.: Psychophonetische Aspekte der Wortwahrnehmung (Buske, Hamburg 1985).

Sendlmeier, W.: Auditive judgements of word similarity. Z. Phonet. Sprachwiss. Kommunikationsforsch. *40:* 538–547 (1987a).

Sendlmeier, W.: A model for the phonetic mental representation of words. 11th Int. Congr. Phonet. Sci., vol. 1, pp. 68–71 (Academy of Sciences of Estonia, Tallinn 1987b).

Sendlmeier, W.: Perception and mental representation of speech. Linguistics *27:* 381–404 (1989).

Stevens, K. N.: Models of phonetic recognition. II. A feature-based model of speech recognition; in Mermelstein, Proc. Montreal Satellite Symp. on Speech Recognition, Twelfth Int. Congr. Acoustics, 1986.

Tillmann, H.-G.; Mansell, P.: Phonetik (Klett-Cotta, Stuttgart 1980).

Treiman, R.; Salasoo, A.; Slowiaczek, L. M.; Pisoni, D.: Effects of syllable structure on adults' monitoring performance. Progr. Rep. 8, Indiana University Speech Lab. Report (1982).

Wertheimer, M.: Untersuchungen zur Lehre von der Gestalt. II. Psychol. Forsch. *4:* 301–350 (1923).