Perception and mental representation of speech*

WALTER F. SENDLMEIER

Abstract

Starting from theoretical considerations concerning fundamental linguistic and psychological aspects of research in perception and the mental representation of speech, three experiments are presented. Employing a variety of experimental designs and methods and testing groups of subjects with different relevant characteristics, the experiments give evidence for a rejection of a feature- and phoneme-oriented approach in the investigation of speech perception.

Theoretical considerations

This paper deals with aspects of psychophonetic problems of speech perception. Emphasis is laid on speech as a phenomenon which has psychological reality rather than on speech as a sociological factor in the sense of Saussure's 'langue'. This implies concern with the analysis of the behavior of language users in speech production and speech perception as can be investigated by applying physical and psychological methods.

Alternative approaches dealing with speech from a sociological or formal perspective and governed by overall principles such as economy and formal simplicity lead to theoretical constructions, which — by aiming at something like an autonomous field of linguistics — deliberately neglect the actual behavior of language users. Such theoretical constructions appear as nothing but operational fictions with which the data they are supposedly based on can be handled and universal laws can be set up. Thus, the categories, structures, and systems which theoretically oriented linguists set up in their analyses of speech are not inherent in the language itself, but are products of the work of linguists dealing with their material.

This contrast between an empirically founded versus a theoretical approach reminds one of the relation between phonetics and phonology

Linguistics 27 (1989), 381–404

0024–3949/89/0027–0381 $2.00

in the 1930s. Although the representatives of structural phonology — especially those of the Prague school — were concerned with leaving aspects of perceptional psychology out of their theoretical concepts, they nevertheless were bound to refer to auditory judgments of acoustical phenomena in analyzing speech. These judgments were usually the product of their own subjective introspection. When within the field of phonology the distinctive-feature theory was introduced, the importance of auditory judgments for the description of acoustical features was stressed; however, the researchers did not verify the relation between stimuli and perceptual judgments by means of adequate phonetic experiments. In the field of generative phonology the relation between the acoustic stimulus and its human perceiver was neglected altogether, thus leading to the impossibility of empirically validating the theoretical constructs set up.

When, in the early 1950s, phoneticians and psychologists working experimentally started to investigate the relation between the linguistic unit and its processing by the human listener, they were guided by the concept of minimal pairs and the ensuing distinctive-feature theory developed by phonologists. Thus they focused on the smallest isolated and reduced units — presented in the form of synthesized signals to listeners in the laboratory to be identified and discriminated. Notwithstanding the valuable results obtained by such studies, one should be aware of the fact that the experiments were based on artificial acoustic phenomena which were far removed from their natural manifestations. When preparing the stimuli in such a way that perception depends on nothing but the stimulus itself, one runs the risk of producing artefacts of the laboratory or else of simply registering in terms of psychophysics the aptitude of the peripheral acoustic analyzer and of neglecting the governing cognitive processes in speech perception. If one aims at discovering the nature of the perception process by the manipulation of external stimuli and of their presentation, one can never rule out for certain that by modifying the stimulus presentation, the character of the stimulus processing may also be affected. This uncertainty may be regarded as the psychological equivalent of Heisenberg's *Unschärferelation*, and I believe that this is one of the fundamental problems in perception research. In order to minimize this dilemma, it appears to be necessary to investigate the process of speech perception by means of widely varying methods. Among these there should be methods which aim at proving the sensitivity of the human listener to certain parameters within the acoustic signal as well as methods which, by starting from more complex natural units, help to gain insight into the human processing of speech and its presuppositions by means of a variety of tasks.

For the prevailing theories within phonology, words were and still are nothing but linked sounds. Although the phonological functions are manifested not in the sounds but in the words, and although sounds hardly ever occur in isolation but only as parts of words, the single sound became the objective of phonological analysis due to a viewpoint which might be characterized as 'atomistic' in spite of the fact that the researchers considered structures on the paradigmatic axis. Such an approach, however, can describe only part of the acoustic properties of words relevant for perception.

Looking at the process of word recognition in everyday life situations, it becomes obvious that the listener — due to the situational context and strictly linguistic restrictions — does not consider all the words in his vocabulary alike as ones to be recognized. Instead, by interpretation of the existing restrictions, the listener arrives at a subset of his complete vocabulary which constitutes the frame for recognition. The process of identification and discrimination in word perception is strongly dependent on the size and type of such a subvocabulary. If this is so, generally valid statements about the relevance of certain phonetic features for the listener cannot be made, since the relevance of phonetic features varies with the relations between all the alternatives considered.

Whether phonological oppositions, as they were deduced from minimal pairs, really do exist in the consciousness of the language user must be questioned. Given, first, that most words in a language have no direct partner with which to form a minimal pair, second, that in everyday life one is rarely confronted with the task of having to distinguish between two words constituting a minimal pair, and third, that words are formed as units and that differences between them are only later deduced and made consciously explicit (Oksaar 1977), it seems that the phenomenon of phonological opposition has been widely overestimated (see also Barry 1980). Ladefoged (1984) pinpoints the problem of the ontological status of phonemes even more closely when he writes,

> Perhaps the most startling conspiracy — one that seems to have deceived by far the majority of linguists — is the appearance of phonemes. Accounts of human behavior in terms of phonemes are nearly always examples of what has been called the psychologist's fallacy — the notion that because an act can be described in a given way that [sic] it is necessarily structured in that way. As far as I can see, phoneme size units play only a minor role in human behavioral acts such as normal speaking and listening (1984: 92; for further details see also Ladefoged 1980).

The fact that the single sound and thus the concept of the phoneme stood and still stands so much in the center of phonetic-phonological approaches may certainly also be reduced to the fact that scientific

perspectives are always culturally and biographically influenced. Most leading linguists came and come from an area where speech is represented by an alphabetic system. In their analyses they were concerned either with languages for which the same may be said or else with languages that had not been fixed in written form so that they could transfer their well-known principles without having to discuss competing writing systems. In contrast to such an approach, Firth (1957), for example — influenced by his work with languages transcribed on a syllabic basis — suggested a linguistic system for analysis and description in which suprasegmental and phonotactic phenomena play the central role. Unfortunately Firth never really developed a system that people not in contact with him could use (Kelly 1988); metrical and CV phonology are a welcome current development in a similar direction.

It is widely known that a specific language background influences a person's ability to segment utterances, supporting the view that the single sound cannot be taken universally as the central perceptional unit. Language teachers report that Chinese speakers familiar only with the traditional logographic system have considerable difficulties with segmentation below word level; in contrast, Vietnamese, whose language is phonetically-phonologically somewhat similar to Chinese but has an alphabetic system, do not have the same kind of problems. Obviously previous experience concerning acoustical categorization influences the language user in his ability to arrive at phonetic-acoustical analyses of certain kinds. Teachers who teach adult illiterates confirm that their students often have great difficulties in segmenting the sound stream within the boundaries of words. Another possibility of gaining insight into the primary perceptual unit lies in looking at the early stages of a child's language-acquisition process. In first-language acquisition research, it has become a fact that the child learns a word as bearing meaning corresponding to a certain object or class of objects. It seems plausible to assume that in this learning process phonetic characteristics are globally perceived; in other words, the child learns, for example, the word 'ball' as one phonetic unit and not as a combination of the single sounds /b/ + /ɔ:/ + /l/ or even as a matrix of 3 × 9 distinctive features. And a child never produces sequences such as /ba da ga/. In this context it seems interesting that Aramaic scholars who developed a syllabary did not notice, over hundreds of years, that syllables such as /ba da ga/ had something in common (Ladefoged 1984).

The attempt to embed the manifold findings from perception experiments into linear sequential models of information processing was influenced by the concept of linear computer programs, which were able to achieve a good portion of what was until then considered to be a

specifically human achievement. Within the information-processing paradigm the perception of simple stimuli presented against a homogeneous background can be described more or less adequately as a detection of discrete criterial properties and their conjunction. If, however, complex and structured stimuli such as words are considered, which are characterized by configurative properties, that is, by a variation of parameters across time, and if in addition these stimuli are partly masked by other heterogeneous stimuli, then a great number of identifications can no longer be adequately described by binary context-independent property detection. Auditory word perception in everyday life can be described more adequately if one takes into account that principles related to those addressed by Gestalt theory come into effect and that larger units than distinctive features or phonemes play a major role in information processing.

The question of how the auditory images of words are represented in the mind of the language user is closely connected with the foregoing considerations. Cognitive representation may be regarded not only as a result of the perception process, but also as a way of monitoring perception. More precisely the question arises of whether structural or configurative properties of acoustic sequences are already primarily represented as units which as a whole can be taken from precategorical storage. It seems plausible to assume that — concerning structured stimuli — the relevant information is represented and stands at the disposition of the user both as a unit and as a number of isolated elements. The question then arises as to whether the different levels of representation have to be structured hierarchically in order to exclude even theoretically the possibility of interference between the two kinds of representation during the process of information extraction. This, of course, would lead to the consequence that during the perception process, global or local perception would have to be regarded as dominant or at least as chronologically first (see also Neisser's considerations concerning preattentive monitoring [1967] and Broadbent's distinction between two kinds of selective attention [1977]). It is probable that the speed of speech production, the type of task, the context of perception, etc., determine the level of representation from which the relevant information for solving a perception task is primarily taken. This implies, however, that perceivers have to anticipate that level of representation which contains the information relevant for their actions by means of implicit knowledge about their perceptual surroundings and/or relevant contextual information. This implicit knowledge comprises the recipient's experience with the perceptual surroundings resulting in the fact that the number of levels of representation in which a stimulus is embedded may differ from individual to

individual. Whereas in psychophysics one can neglect interindividual differences as minimal (Sixtl 1967), one may expect from psycholinguistic and psychophonetic experiments that individual variation among subjects will be of considerable importance. One might then wonder whether the acquisition of a phonetically oriented writing system (as when learning to read and write alphabetically) will lead to a more differentiated organization of the acoustic representation of words and whether this affects the perception process.

In addition, it seems to make sense to assume that the perceptual activities of listeners vary with varying tasks, so that, for example, they can focus on different levels of representation interchangeably; thus listeners could switch to the level of syllables or even of single sounds when discriminating difficult words of a foreign language, and then switch back to word level or even a higher level later (Stoffer 1981).

Such a possibility of variation in perceiving the background of acoustically constant stimulus presentation can be seen in analogy with *Kippfiguren* in visual perception where one finds a change of figure and background. As in the laws of Gestalt psychology we find — besides the immediate phenomenal correlate of the stimulus parameters — an additional specification of the phenomenal unit. Similar scientific models which imply the representation of stimuli within different systems of similarity and conjunction at the same time have been successfully used, at least implicitly, in other psychological fields such as for example in the psychology of thinking. Flexibility in problem solving is based on the ability to change perspective and the contextual system (Prinz 1983).

Closely connected to the problem of the size in which acoustic perceptual units are represented is the question concerning the form of these representations. Here Wertheimer's concept of 'ideal types' (1923) or Rosch's related concept of 'prototypes' (1973, 1975, 1977) seem to be adequate alternatives to abstract feature matrices. Wertheimer supposes that among perceptual stimuli there are certain ideal types which function as anchor points in perception. Starting from the discovery of focal points in the perception of colors, that is from the proof that there are areas in the color spectrum which are universally perceived as more prominent than others, Rosch comes to the supposition that our physical world is divided in a typical manner and that categories in the human perceiver's memory are stored in the form of especially typical ('prototypical') representatives of the categories in question in analogous, nonanalyzed form. However, a number of categorizations are not universally but culturally determined and thus subject to learning; among these are the categories of linguistic signs. But how are prototypical representatives of acoustic units acquired by the listener if these representatives cannot be

deduced from the stimuli as such? It seems more plausible to assume that a listener generates a prototype from all the representatives of a category ever heard in the sense of a 'statistical mean' during the course of language acquisition. Attneave (1957), working with meaningless polygons, and Posner et al. (1967), using patterns of dots, were able to verify empirically that the human perceiver can form mental schemas (prototypes) for artificially presented categories. Especially interesting within the present context, however, are such experiments (for example by Bransford and Franks 1971; Posner and Keele 1968; Posner 1973; Reed 1973) in which it could be shown that the subjects generated mental representations in the form of a schema, although the correlate of this schema was never presented; thus the schema must have been deduced from other stimuli. If one supposes that acoustic units of different size (up to word level or at times even up to the level of phrases) are represented in the form of especially typical prototypes in an unanalyzed way, but not in the sense of a first-degree isomorphy, this implies an enormous capacity of long-term memory. The objections of scientists who, by referring to principles of economy, argue against the assumption of such memory-consuming representations, can be rejected in view of proofs of the enormous capacity of the human brain (Penfield 1969).

Experiment 1

Introduction

After these theoretical considerations I shall now present the results of three experiments conducted to find clues about the process of word perception. The first experiment is based on the question of whether awareness of speech as a sequence of single sounds is related to the degree of mastery of an alphabetic writing system. In addition to its relevance for basic research, this question is of some importance for practical work in the field of teaching illiterates.

Summarizing the results of experiments concerning the acquisition of the ability to segment utterances into phones (Bruce 1964; Savin 1972; Smith and Tager-Flusberg 1980; Magnusson et al. 1983), one can say that significant progress is made at the age of six or seven years. Since, however, this is the age at which children are confronted with reading/writing instructions, the present question cannot be answered by choosing young school children as subjects.

A different approach to verify the thesis of the relation between the

ability to segment phonetically and the knowledge of an alphabetic writing system can be seen in confronting adult illiterates with tasks the solution of which makes a conscious phonetic analysis necessary. This approach, which Morais et al. (1979) have already used in a pilot study with Portuguese adults on a more limited scale, was chosen for the present investigation.

If the ability to make a precise phonetic analysis is influenced by the degree of mastery of a writing system, then adult illiterates — depending on their previous knowledge — must fail to a varying degree in word-manipulation tasks which are based on such analytical processes. If, however, the ability to process speech in a phonetically analytic way is dependent on the process of general cognitive maturation, then illiterates and literates must be able to solve the given tasks equally successfully.

Method

The tasks used in the present experiment implied that in certain words single sounds (consonants) were to be manipulated in initial position. Table 1 gives the words used in this experiment. Each task consisted of 16 words, 8 of which were meaningful and 8 of which were meaningless. One-half of the sounds to be manipulated were single consonants in initial position, the other half were to be manipulated in consonant clusters. The test material contained only words for which, in a pretest, students could easily make the required manipulations. Before the test six practice items (two deletion, two addition, two substitution tasks) were offered, in the solution of which the instructor gave help and corrected wrong answers. In the actual test wrong answers were not corrected and no help was offered. The respective sounds in question were pronounced by the instructor according to their phonetic values, that is, leaving out the vowel quality in producing the consonants as much as possible. Subjects were told to delete, add, or substitute the respective sound. An example of instructions for a deletion task is, 'Leave out the "m" in the word *Mund* 'mouth'; what word do we get?'

Since the previous reading/writing knowledge of so-called 'functional illiterates' can vary considerably, the previous knowledge of the 30 subjects (aged 16 to 50) was estimated by their teachers on a graded scale from 1 to 7.

Results

The data were treated by methods of correlation and analysis of variance. The relation between previous knowledge in reading and writing and the

Table 1. *List of words used in Experiment I*

	Meaningful	Meaningless
1. Deletion task		
Instructions for the first item: 'Leave out the "t" in the word "Teiche"; what word do we get?'		
a. in isolation	*T*eiche — Eiche	*p*ache — ache
	*W*ende — Ende	*s*oncke — oncke
	*L*ehre — Ehre	*r*iede — iede
	*M*und — und	*n*usche — usche
b. in cluster	K*l*asse — Kasse	p*l*esche — pesche
	*T*rasse — Rasse	*k*rüsse — rüsse
	*/*liegen — liegen	*/*lieden — lieden
	Sch*m*erz — Scherz	sch*m*ilch — schilch
2. Addition task		
Instructions for the first item: 'Put an "m" in front of the word "acht"; what word do we get?'		
a. in insolation	acht — *M*acht	ift — *m*ift
	Ecke — *D*ecke	oppe — *g*oppe
	Amen — *R*ahmen	uhmen — *l*uhmen
	immer — *Sch*immer	annen — *sch*annen
b. in cluster	Trumpf — *S*trumpf	prant — *sch*prant
	Bau — b*l*au	geu — g*l*eu
	nie — *K*nie	noh — *k*noh
	Schutz — Sch*m*utz	kust — k*n*ust
3. Substitution task		
Instructions for the first item: 'Replace the "m" at the beginning of the word "Märchen" with a "p"; what word do we get?'		
a. in isolation	*M*ärchen — *P*ärchen	*n*ade — *p*ade
	*H*und — *F*und	*h*oss — *f*oss
	*B*uch — *T*uch	*b*iff — *t*iff
	*K*ind — *R*ind	*k*emp — *r*emp
b. in cluster	S*p*aten — S*t*aaten	sch*p*ullen — sch*t*ullen
	Sch*w*amm — Sch*l*amm	sch*w*end — sch*l*end
	sch*m*ecken — Sch*n*ecken	b*l*itt — britt
	B*l*ei — B*r*ei	sch*m*atten — sch*n*atten

ability to solve the presented tasks is shown in the correlation coefficients in Table 2. The general results show that (with r = .94) an almost linear relation is given between reading/writing proficiency and the ability to manipulate single sounds. The correlation coefficients for the separately estimated previous knowledge in reading and writing and the number of adequate solutions for all three types of tasks show a similarly clear relation. Looking at the different types of tasks, it becomes obvious that the correlation between reading/writing ability and the deletion tasks was highest.

Table 2. *Correlation coefficients according to the Pearson (Spearman) formula*

	Deletion	Addition	Substitution	Total
Reading	.923	.840	.822	.897
	(.937)	(.847)	(.863)	(.912)
Writing	.942	.772	.766	.922
	(.941)	(.791)	(.809)	(.934)
Sum	.947	.819	.807	.936
	(.948)	(.832)	(.848)	(.941)

The evaluation of the data by means of variance analysis proved significant effects for the main variables: type of task, meaningfulness of stimulus, and phonotactic characteristics, that is, isolated consonants versus consonants in clusters (for more detailed information see Sendlmeier 1987a). Within the present context, the results concerning the phonotactic variable are of primary interest. The mean of correct solutions in tasks involving isolated consonants proved to be significantly higher than the mean of correct solutions in those tasks in which consonants had to be linked together or in which a consonant cluster had to be separated ($F(1,29) = 130.13$; $p < .001$). This makes it obvious that consonant clusters are, to a high degree, regarded as units which cannot be further dissolved or composed of smaller units. Further experimental support is reported by Barry (1984), who demonstrated that reaction times in initial-consonant identification are reduced after training for /bV/ syllables but not for /br/ clusters.

Discussion

These data provide evidence for the fact that functional illiterates, depending on their proficiency in reading and writing, have a more or less distinct concept of single acoustic segments. Thus the argument cannot be upheld that the single sound and the underlying concept of the phoneme could be a natural category which is either genetically fixed or which appears universally in the course of cognitive development. Even though the concept of speech as a sequence of discrete single sounds has proved useful on a descriptive metalevel to set up general rules in modern linguistics — from the *Junggrammatiker* to generative phonologists — the result of the present experiment leads to the conclusion that phonemes and the ensuing units are not necessarily psychologically real. This result seems to have even more impact since the types of tasks were deliberately constructed so as to resemble the procedures of phoneme analysis by

minimal pairs. It is interesting to note that even a number of students serving as subjects in the pretest with literates had considerable difficulties with the manipulation of single sounds in words which do not show a one-to-one correspondence between sound and graphic symbol. In the word *Axt* 'axe', for example, the deletion of /k/ or /s/ proved to be quite difficult.

Experiment II

Introduction

The second experiment is based on the concept of similarity. The concept of similarity as a fundamental psychological reality, which was stressed by William James as early as 1890, has increasingly found recognition in psychological research over the past 30 years. Phonetic relations of similarity between the words of a language or vocabulary can be perceived by the human listener. These relations of similarity which may become relevant as the source of potential confusions in communication form the objective of the following experiment.

The goal of the investigation is to find out what dimensions are fundamental for perceived similarities between words. Of special interest is the question of how far word similarities can be explained by comparing single sounds, that is, on the segmental level. Since no comparable experiments have been conducted so far, this investigation is to be regarded as a pilot study. The general underlying assumption is that judgments of similarity can satisfactorily be described by only a few dimensions.

Method

For the experiment three vocabularies were constructed, each consisting of 12 meaningless words. Meaningless material was chosen to avoid semantic relations that interfered with phonetic similarity. The words were constructed so as to avoid minimal pairs within a vocabulary. All sound combinations within the phonotactic constraints of German were allowed. The three vocabularies are presented in Table 3. The reason for subdividing the material into three vocabularies was to facilitate the experimental procedure, in which every word of a vocabulary had to be compared twice in a different order of presentation with every other word.

The words in each word pair were acoustically presented with a time interval of 500 ms; between the word pairs there was a pause of four seconds. Before the presentation of the word pairs of a vocabulary, the 12 words composing it were presented twice in a series. The tests were administered in three separate sessions; subjects were 23, 19, and 21 students respectively who had never received any phonetic-phonological training. The subjects were instructed to expect only meaningless words, since only the phonetic similarities were to be rated. In order to facilitate their judgments they were to indicate their ratings on a seven-point scale from very dissimilar to very similar.

Table 3. *The three vocabularies used for similarity judgments*

I	II	III
fœp	ʃtrɛmpf	pfirtsɛcfán
tsvɪp	klɔstən	ómɛrə
ʃprɛk	ʃvɪnt	ʃlíctafo:
nɔs	knɔlp	akríftəs
kɪrps	ʃlɪrpəl	fatʃékstax
flœx	ɪzə	tsáflɪçstɛp
təlʃt	ɔltəs	glœmbzvɛnda
pɛnst	ʃprœxst	malɛrú:
ʃtrɔt	rɛl	ɛsománt
tɪç	ʃœn	ʃtɛlasíc
pœrst	fɛŋkə	takótə
mɛk	œnə	malómə

The similarity judgments were evaluated by multidimensional scaling (MDS). A satisfactory degree of the goodness of fit (Kruskal's stress formula 1) for the MDS solutions for all three vocabularies was attained for the three-dimensional solutions. Nevertheless, the following figures illustrating the quality of the first two dimensions are based on the four-dimensional solution.

Results

For vocabulary I, the first dimension accounting for the greatest part of the variance distinguishes stimuli according to the qualities of vowels in the monosyllabic words, as can be seen in Figure 1. Words containing a back vowel in the right half of the figure are distinguished from words containing front vowels in the left half of the figure. The degree of openness of the vowels is the criterion for the second dimension. This also

applies for the word /pɑrst/ in which the /r/ following the vowel was pronounced as a centralized /a/, so that the diphthong also contained an open element. These distributions indicate that in their similarity judgments the listeners were primarily guided by the quality of the vowel. It seems noteworthy that the geometric configuration shows a strong resemblance to the vowel triangle. For the third and fourth dimensions, the quality of the initial consonant was the distinguishing feature; initial stops were found in opposition to initial fricatives and nasals on the third dimension, whereas on the fourth dimension words beginning with labials were opposed to words beginning with an alveolar or velar consonant.

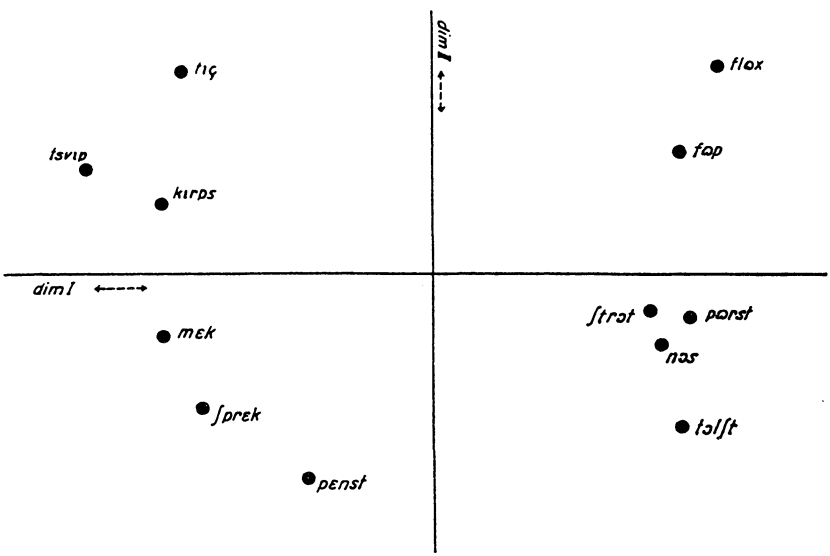

Figure 1. *MDS solution according to the first two dimensions for the relations of similarity of 12 words of vocabulary I*

Figure 2 shows the geometric configuration of the words of vocabulary II along the first two dimensions. The number of syllables, independent of the quality of the single sounds, may serve as an explanation of the first dimension, whereas vowel quality seems to be crucial for the second dimension. With regard to the position of the word /ɑnə/, an explanation based on the dichotomy 'stressed back vowels' vs. 'stressed front vowels' is not entirely satisfactory. It seems more adequate to recur to whether the words contain the vowel /ɔ/ in interpreting the second dimension. The third dimension may be described by the differing number of sounds in a word; words with five sounds are found between those with three and

those with seven sounds. For the fourth dimension, again vowel quality is the distinguishing criterion, the presence or absence of a back high vowel defining the poles on this dimension.

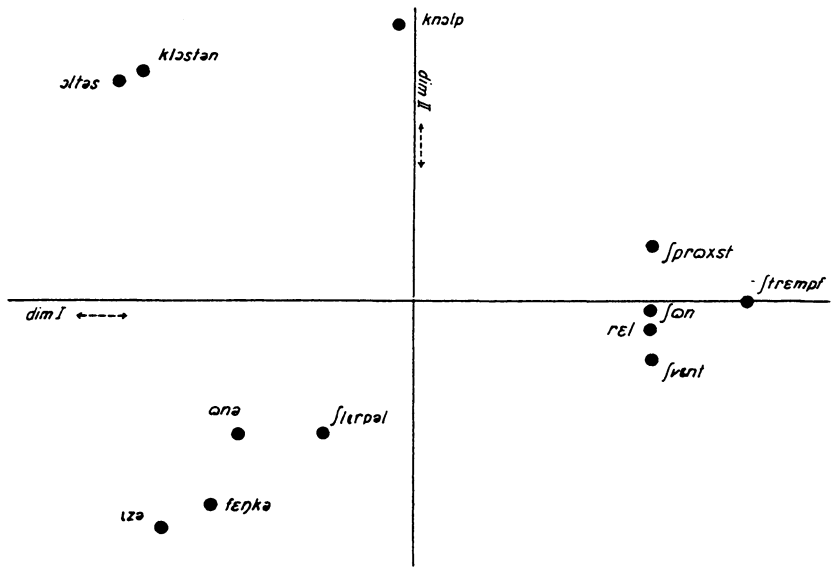

Figure 2. *MDS solution according to the first two dimensions for the relations of similarity of the 12 words of vocabulary II*

Figure 3 shows the MDS solution for the first two dimensions for vocabulary III. The first dimension ensues from the differentiation between words which have consonant clusters in which at least the first element is voiceless, and those which do not. The second dimension is characterized by differing positions of stress: on the one hand, words are stressed on the extreme syllables, and on the other hand, stress lies on the syllables in medial position. The third dimension can again be explained by the position of stress. Now words with stress on the first or medial syllable are opposed to those with stress in final position. The fourth dimension contributes very little to the reduction of stress in the analysis and may not be a real factor. It is characterized by the complex relation of the vowels within a word. Whereas one group of words shows an opening of the vowels toward the final /a/, this maximally open vowel /a/ stands in initial or medial position in the other group of words, so that here the vowel /a/ is always followed by vowels with a minor degree of opening. The two words which are neutral as to the mentioned criterion are placed between the two extremes.

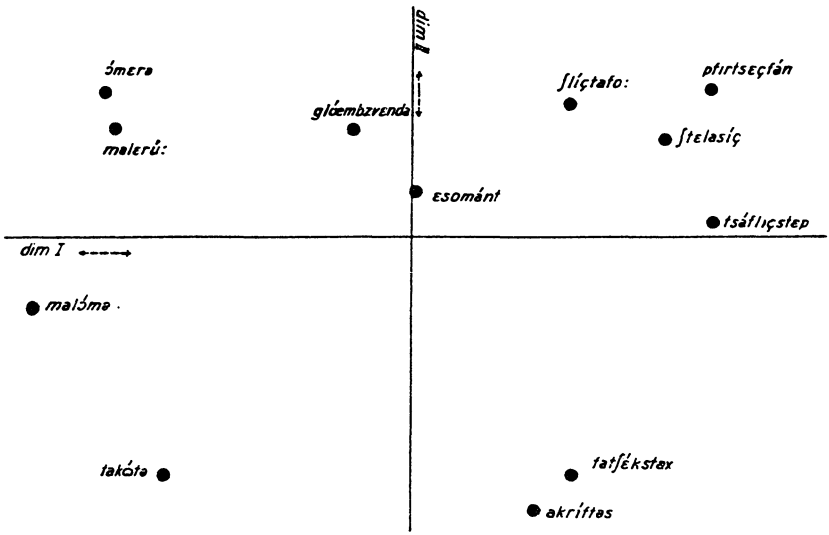

Figure 3. *MDS solution according to the first two dimensions for the relations of similarity of the 12 words of vocabulary III*

Looking at the judgments of similarity for all three vocabularies, one can note as a general result that the relations of similarity between the words of each vocabulary do not vary in an unsystematic way; rather, it becomes obvious that they are structured according to certain phonetic parameters.

It is remarkable that in the judgments the quality of single consonants plays a subordinate role for the perceived degree of similarity. Only for vocabulary I — and here only as regards the third and fourth dimensions — can single consonants help to interpret the dimensions; however, even there it is not the similarity of single consonants as such that is of importance, but only the similarity of such sounds as occur in the same exposed, that is initial, position. Looking at the vowels, however, one finds that their quality is of great importance, especially for the similarity judgments of short one-syllable words. Yet, in more complex stimuli, the relevance of vowel similarity is of less impact than properties which characterize the words as a whole. For example, the listeners primarily took the number of syllables as a basis for their judgments of word similarity in vocabulary II; for vocabulary III, the similarity judgments can be interpreted as a result of word structures, more precisely, the presence or absence of consonant clusters and the position of word stress.

The same investigation with phonetically trained listeners led to similar results (Sendlmeier 1987b). Besides the quality of vowels and the position of consonant clusters, which can explain the distribution of stimuli for vocabulary I, in vocabulary II the identity of consonants, but again only in initial position, and the quality of the final sound proved to be relevant. It is quite remarkable that the dimensions for the third vocabulary are almost identical to those of the untrained listeners. This result can be seen as proof of the hypothesis that, on the one hand, increasingly global properties are considered as criteria for similarity for increasingly complex stimuli, and that, on the other hand, this effect occurs independently of phonetic training, which only seems to be relevant in judging simpler stimuli insofar as single sounds receive more recognition. It should be noted that these results are in line with those in a number of other studies in which different methodological approaches were chosen. Boothroyd (1984), who carried out recognition experiments with hearing-impaired subjects, found that the number of syllables proved to be the most robust characteristic of words, followed by accent patterns and quality of stressed vowels. Cutler (1984) was also able to show that errors in whole-word recognition experiments with degraded speech quality invariably match the target words in terms of stress patterns. Furthermore Cutler and Norris (1988) present data according to which a strong syllable triggers segmentation of the speech signal, at least in a stress-timed language. They argue that segmentation at strong syllables in continuous speech recognition serves the purpose of detecting the most efficient locations at which to initiate lexical access.

Experiment III

In the two preceding experiments, groups of listeners were asked to give metalinguistic judgments or to follow metalinguistic instructions. The third experiment — which may be regarded as a pilot study — was differently designed in that an identification task was of central importance. This experiment can be placed in the tradition of *Aktualgenese* — an approach which can be traced back to the Gestalt psychologist Friedrich Sander — but it also has an affinity with the gating paradigm (see for example Grosjean 1980). *Aktualgenese* in Sander's sense stands for the actual genesis of a Gestalt. In the following experiment the property of wholeness (*Ganzheit*) is attributed to the unit word. The word *Melone* 'melon' was chosen here as stimulus. In this context it is important to note that — as a rule — that which is a part and that which is the whole is not determined a priori but can vary with perspective. Thus even that which is

part in one case can become the whole with further differentiation, and vice versa, the whole can become a part when regarded from an overall perspective.

As parts of the stimulus, segments of five periods each were taken from the transitions of the speech signal. After defining the parts — that is, the five periods each from the transitions of the word *Melone* — the amplitudes between the segments were set to zero, thus creating silent intervals between the segments (see Figure 4). The duration of these silent intervals is defined primarily by the original quasi-stationary sections.

Starting from this destruction of the word, the silent intervals between the segments were extended or shortened successively so that each silent interval was changed by the same absolute amount. Altogether 14 stages of different length of silent intervals between the segments were produced

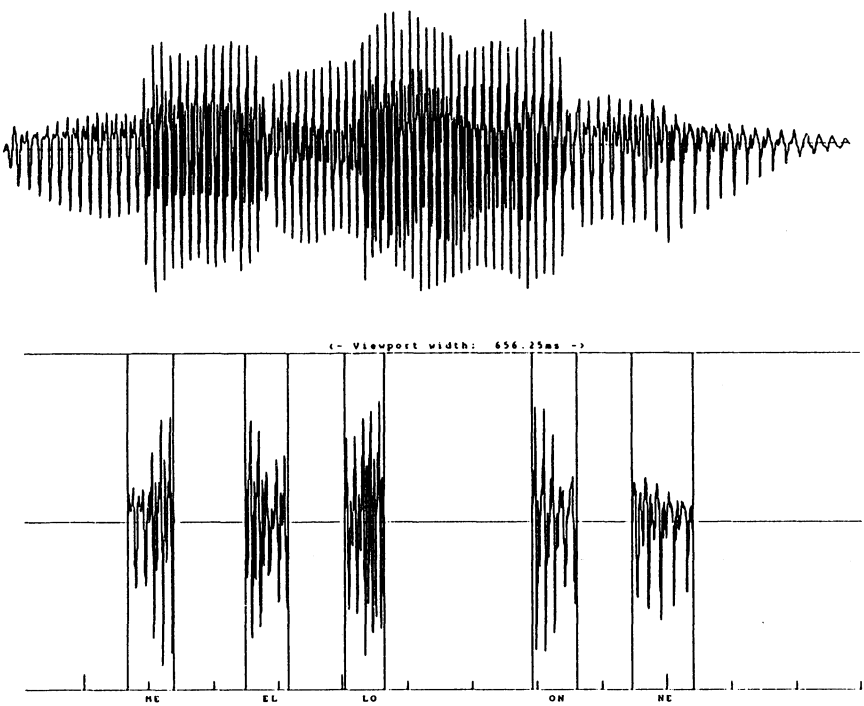

Figure 4. *Speech wave form of the word* Melone *'melon' (a) and a stimulus immediately derived from this speech signal consisting of segments of five periods each, taken from the transitions of the speech signal with the amplitudes between these segments set to zero (b); from this stimulus 13 further stimuli were produced by extending or shortening the silent intervals*

(see Table 4). These stimuli were taped and ordered in such a way that the stimulus with the greatest length of silent intervals (three seconds), which will be called the 'isolated presentation of the segments', stood at the beginning and the stimulus without silent intervals stood at the end of the sequence, as can be seen from Table 4. Stimulus 6 represents the original stimulus after the destruction of the word by clearing the amplitude of the quasi-stationary sections. In other words, the interval between the segments decreases from step to step of the sequence. Each stimulus — that is each of the different versions of the word *Melone* — was presented three times in immediate succession. Ten subjects (undergraduate students at the university of Bonn) were asked to describe what they heard in the form of orthographic symbols as best they could. In the case of an actual word being recognized, this word was to be written down. In addition, the subjects were asked to rate their respective certainty of judgment according to the three categories 'certain', 'more or less certain', and 'uncertain'.

The evaluation of the data showed that the isolated segments were hardly recognized at all, and also parts of the segments which as diphones contain the properties of two sounds were only very seldomly correctly identified. The number of correctly perceived single sounds and diphones increased with decreasing silent intervals between the segments. In the course of the first half of the experiment almost all subjects recognized a

Table 4. *Duration of silent intervals between reduced segments (five periods from the transitionary parts of the word* Melone*) in relation to the original stimulus (6); see Figure 4*

	(ms)	me	el	lo:	o:n	nə
0	+3000	3000	3000	3000	3000	
1	+ 150	224	202	293	193	
2	+ 50	124	102	193	93	
3	+ 30	104	82	173	73	
4	+ 20	94	72	163	63	
5	+ 10	84	62	153	53	
6	± 0	74	52	143	43	
7	− 10	64	42	133	33	
8	− 20	54	32	123	23	
9	− 25	49	27	118	18	
10	− 30	44	22	113	13	
11	− 35	39	17	108	8	
12	− 40	34	12	103	3	
13	no intervals	–	–	–	–	

Gestalt in the form of a word; however, this was not in every case the original word but one which is acoustically very similar, more precisely the word *Millionen* 'millions', which in German is more frequently used than *Melone*.

The correct original word was not recognized until stimulus 6 onward; most frequent recognition occurred for stimuli 8 and 9. Parallel to the increase of correct recognition, the grade of estimated certainty of judgment changed from an 'uncertain' phase to 'more or less certain', and then to a 'certain' phase during the progression through the 14 stimuli. The phase, consisting of two or three stimuli, where judgments were 'more or less certain' could thus be characterized as a phase of experiencing a preliminary Gestalt (*Vorgestalterlebnis*). In this phase the subjects became significantly more restless and tense. Upon word recognition this emotional tension was released suddenly, in the form of Bühler's '*Aha-Erlebnis*' and was accompanied by a shift of the estimated certainty of judgment from 'more or less certain' to 'certain'. This can be seen from the example of an answer sheet in Table 5. The correct recognition increased up to the reduction of the silent intervals to 25 ms. This result might be explained by the fact that the segments do not cover the transitions completely, in other words that they do not reach the beginning and end points of the transitions, but start and end in their frequency above or below the original extremes and are thus shortened in their temporal extension. Following the assumption that a listener relates the length of transitions to an estimated rate of speech, and thus to the

Table 5. *Sample answer sheet*

Stimulus	Orthographic transcription of the perceived	Certainty of judgment certain	more or less	uncertain
0	bi, i, d, ä, e			x
1	mi inuponä			x
2	miinuponä			x
3	miinuponä			x
4	miinluplonä			x
5	miiluponä			x
6	miiluponä		x	
7	millionen		x	
8	melone		x	
9	melone	x		
10	melone	x		
11	melone	x		
12	melone		x	
13	melone		x	

expected length of the stationary parts, shorter silent intervals could also be regarded as adequate.

It should be noted that the perceived sounds change from a chaotic rumbling to a recognizable word although the phonetic signal information contained in the segments remains constant. Since nothing but the length of silent intervals between the segments is changed this must be the immediate and only reason for the observed phenomenon, if one regards perception as being determined exclusively by external stimuli. An explanation of the findings as a result of a learning process can be excluded according to Huggins's (1975, 1981) findings. In his investigations of the phenomenon of 'auditory streaming' he also offered his subjects words with silent intervals of varying length, in random succession, however. He found that the rate of word recognition was completely independent of the sequence in which the stimuli were presented.

If one looks only at the stimulus parameters in order to explain the word identification — as is done in pure bottom-up approaches — one will find no answer to the question of why the change of the duration of the silent intervals has such an obvious effect on the perception. When investigating this question one necessarily changes perspective from looking at the external stimulus to focusing on the perception process. The latter can be explained as follows: by shortening the silent intervals, a perception process is initiated or facilitated in the course of which the individual segments are related to each other and are thus used for a mutual contrast formation. Such a process of relating elements is an active perceptual accomplishment on the part of the listener and not a physical property of the external stimulus itself (for a more detailed discussion of different word-recognition models see Sendlmeier 1987c).

What could be the result of such a process of relating the elements? First of all, the dissimilarity of the objects could be perceived; dissimilarity is a relation. In addition, however, the elements of the relation could themselves be changed by the relation in the perception process. In our experiment this could mean, for example, that an isolated, reduced segment /o:n/ would be perceived more precisely when presented together with other elements like /el/ or /nə/ than in isolation. Whether certain parts of the acoustic surroundings have different degrees of impact cannot be decided at present. The fact that such a comparison is made, and can be made useful, also seems to depend to a considerable degree on the temporal difference in the spacing between the offered segments. In our experiment the optimal temporal conditions seem to exist for stimuli 7 to 9. An investigation by Martin and Bunnel (1981) showed that listeners are indeed sensitive to mutual influence of segments, even if those segments do not stand in immediate juxtaposition to each other.

In perceiving language the percipient is striving for an optimal compromise between object adequacy and giving sense to the perceived object. The external stimuli offer a frame but at the same time allow for many possibilities of the final interpretation, since the listener is not searching for the acoustically most esthetic solution but for a basis for acting adequately. Models of the perception process which assume a successive and independent identification of distinctive cues or single sounds or diphones and which attribute to the listener, among other things, certain detectors which begin to function independently and are caused to function by the external stimulus, cannot explain the results of this experiment; in fact, they are contradictory to the present findings.

In a more linguistic terminology, our approach could be described as follows: the listener's internal structure is based on categorization on the paradigmatic axis — that is, on the setting-up of a relation between the external stimulus and the structure represented 'in absentia'. This process is influenced by setting up relations on the syntagmatic axis — that is, relations between the externally 'in praesentia'-represented stimuli. The relations on the syntagmatic axis lead to a constant changing and adjustment of the internal structure which is the basis for decisions on the paradigmatic axis. On the other hand, the actual internal structure determines the frame for the construction of syntagmatic relations to a considerable degree. Thus, we must again take a mutual influence into account. The phenomenon of 'speaker normalization', for example, could be interpreted in this sense also.

In further pursuing the Gestalt-psychological approach, one will have to pay special attention to those properties which can be attributed exclusively to the whole, that is to phenomena of word prosody, such as change of pitch, stress pattern, or the duration of a word, which as a whole are related to the parts and which also determine the perception of the parts. In this context an experiment by Miller (1981) deserves mention. She showed that perception of the initial consonant as /w/ or /b/ is also determined by speech rate and thus by the duration of the whole word in relation to the duration of the transition from the respective consonant to the following vowel (see also Ventsov 1983). This proves that the individual parts are related to all other parts of the whole and at the same time to the whole as such.

Conclusion

In presenting the results of the three experiments, I have tried to draw attention to aspects of the human speech-processing mechanism which

cannot adequately be described by feature detectors or phoneme-oriented concepts. My aim was to show with these experiments, in which different levels of perspective were chosen, that the processing of acoustic units like words varies depending on the nature of the linguistic stimuli and on the attentional properties of the listener.

It was shown in the first investigation that the ability to manipulate single sounds stands in close relation to the degree of proficiency in alphabetic reading and writing, providing that the single sound, and thus also the phoneme, do not have an a priori psychological reality. The result of the second experiment was that the listener — when judging word similarities of simple stimuli (one-syllable words) — uses features of single sounds to a greater or lesser degree depending on previous phonetic training, whereas in the case of growing complexity of the stimuli, increasingly more global features (number of syllables, intonation) are used as criteria for similarity independently of phonetic training. In the third experiment, where the listeners had to master an identification task, it was shown that, although the tasks suggested a detailed acoustic analysis, the listeners did not primarily identify single sounds independently of each other but set up hypotheses about parts of the word and then tested these hypotheses by relating the parts to each other.

In spite of these results I should like to stress that I do not mean to question the relevance of single sounds for speech perception generally; rather, my aim was to point out phonetic aspects which are not covered by the principles of phonology which, however, can be of great importance in the auditory processing of linguistic stimuli.

In closing I would like to point out that in Karl Bühler's work of 1934 some theoretical considerations already offer clues to the heterogeneity of the modes of speech perception that have just been empirically illustrated. Bühler postulated that in speech processing, modes which take the wholeness of the sound stream into account and analytic modes of perception become equally relevant.

Received 7 December 1987
Revised version received
29 August 1988

Max-Planck-Institut für
Psycholinguistik, Nijmegen
University of Bonn

Note

* Correspondence address: Institut für Kommunikationsforschung und Phonetik, Universität Bonn, Poppelsdorfer Allee 47, D-5300 Bonn 1, Federal Republic of Germany.

References

Attneave, F. (1957). Transfer of experience with a class-schema to identification learning of patterns and shapes. *Journal of Experimental Psychology* 54, 81–88.

Barry, W. J. (1980). Die Verarbeitung akustischer Information in der lautsprachlichen Wahrnehmung. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel* 13.

—(1984). Segment or syllable? A reaction-time investigation of phonetic processing. *Language and Speech* 27, 1–15.

Boothroyd, A. (1984). Auditory perception of speech contrasts by subjects with sensorineural hearing loss. *Journal of Speech and Hearing Research* 27, 134–144.

Borg, J. (1981). *Anwendungsorientierte multidimensionale Skalierung*. Berlin: Springer.

Bransford, J. D., and Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology* 2, 331–350.

Broadbent, D. E. (1977). The hidden preattentive processes. *American Psychologist* 32, 109–118.

Bruce, D. J. (1964). The analysis of word sounds by young children. *British Journal of Educational Psychology* 34, 158–159.

Bühler, K. (1934). *Sprachtheorie*. Stuttgart: Fischer.

Cutler, A. (1984). Stress and accent in language production and understanding. In *Intonation, Accent and Rhythm*, D. Gibbon and H. Richter (eds.), 77–90. Berlin: de Gruyter.

—, and Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology, Human Perception and Performance* 14, 113–121.

Firth, J. R. (1957). *Papers in Linguistics*. London: Oxford University Press.

Grosjean, F. (1980). Spoken word recognition and the gating paradigm. *Perception and Psychophysics* 28, 267–283.

Huggins, A. W. F. (1975). Temporally segmented speech. *Perception and Psychophysics* 18, 149–157.

—(1981). Speech perception and auditory processing. In *Auditory and Visual Pattern Recognition*, D. J. Getty and J. H. Howard (eds.). Hillsdale, N.J.: Erlbaum.

James, W. (1954 [1890]). *The Principles of Psychology*, 8th ed. New York: Dover.

Kelly, J. (1988). Firthian phonology in prospect and retrospect. Paper presented at the 6th International Phonology Meeting in Krems, Austria.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language* 56, 485–502.

—(1984). Out of chaos comes order; physical, biological, and structural patterns in phonetics. In *Proceedings of the Xth International Congress of Phonetic Sciences*, M. van der Broecke and V. van Heuven (eds.), 83–95. Dordrecht: Foris.

Magnusson, E., Naucler, K., and Söderpalm, E. (1983). Form or substance? The linguistic awareness of preschool children and school children investigated by means of a rhyming test. In *Abstracts of the Tenth International Congress of Phonetic Sciences*, A. Cohen and M. P. R. van der Broecke (eds.). Dordrecht: Foris.

Martin, J. G., and Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America* 69, 559–567.

Miller, J. L. (1981). Some effects of speaking rate on phonetic perception. *Phonetica* 38, 159–180.

Morais, J., Carry, L., Alegria, J., and Bertelson, P. (1979). Does awareness of speech as a sequence of phones arises spontaneously? *Cognition* 7, 323–331.

Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.

Oksaar, E. (1977). *Spracherwerb im Vorschulalter. Einführung in die Pädo-linguistik*. Stuttgart: Kohlhammer.

Penfield, W. (1969). Consciousness, memory and man's conditioned reflexes. In *On the Biology of Learning*, K. H. Pribram (ed.). New York: Harcourt Brace Jovanovich.

Posner, M. I. (1973). *Cognition: An Introduction.* Glencoe: Scott Foresman.

—, Goldsmith, R., and Welton, K. E. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology* 73, 28–38.

—, and Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology* 77, 353–363.

Prinz, W. (1983). *Wahrnehmung und Tätigkeitssteuerung.* Berlin: Springer.

Reed, S. K. (1973). *Psychological Processes in Pattern Recognition.* New York: Academic Press.

Rosch, E. (1973). Natural categories. *Cognitive Psychology* 4, 328–350.

—(1975). Cognitive reference points. *Cognitive Psychology* 7, 532–547.

—(1977). Human categorization. In *Studies in Crosscultural Psychology*, vol. 1, N. Warren (ed.). London: Academic Press.

Sander, F. (1926). Über räumliche Rhythmik. *Neue psychologische Studien* 1, 123–158.

—(1927). Über Gestaltqualitäten. In *Bericht über den 8. internationalen Kongreß für Psychologie, Groningen, 1926*, F. Sander (ed.), 183–189. Jena.

—(1928). Experimentelle Ergebnisse der Gestaltpsychologie. In *Bericht über den 10. Kongreß der deutschen Gesellschaft für Psychologie, Bonn, 1927*, F. Sander (ed.), 23–88. Jena.

Savin, H. B. (1972). What the child knows about speech when he starts to learn to read. In *Language by Ear and by Eye*, J. F. Kavanagh and J. G. Mattingly (eds.). Cambridge, Mass.: MIT Press.

Sendlmeier, W. F. (1987a). Die psychologische Realität von Einzellauten bei Analphabeten. *Sprache und Kognition* 6, 64–71.

—(1987b). Auditive judgements of word similarity. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 40, 538–547.

—(1987c). A model for the phonetic mental representation of words. In *Proceedings of the 11th International Congress of Phonetic Sciences,* Academy of Sciences of the Estonian S.S.R. (ed.), vol. 1, 68–71. Tallinn.

Sixtl, F. (1967). *Meßmethoden der Psychologie.* Weinheim: Beltz.

Smith, C., and Tager-Flusberg, H. (1980). The relationship between language comprehension and the development of metalinguistic awareness. Paper presented at the 5th Annual Boston University Conference on Language Development.

Stoffer, T. (1981). Wahrnehmung und Repräsentation musikalischer Strukturen. Funktionale und strukturelle Aspekte eines kognitiven Modells des Musikhörens. Unpublished dissertation, University of Bochum.

Ventsov, A. V. (1983). What is the reference that sound durations are compared with in speech perception? *Phonetica* 40, 135–144.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt, II. *Psychologische Forschung* 4, 301–350.