Variance and Invariance at the Word Level

Jacqueline Vaissière

J. VAISSIERE

in INVARIANCE AND VARIABILITY IN SPEECH PROCESSES
edited by J.S. Perkell and D.H. Klatt,
Lawrence Erlbaum Associates, Publishers, 1986

Jacqueline Vaissière:

Variance and Invariance at the Word Level

As is well known, the word represents the main point of interaction between all sources of knowledge shared by both the speaker and the listener (Klatt, 1977). Such sources of knowledge include the lexicon which contains the list of available possible words in a language, the syntax, the semantics, and the pragmatics. Those factors constrains sequences of words to form (in the ideal case) grammatically, semantically, and pragmatically correct sentences. The ultimate purpose of speech communication is to convey meaning from speaker to listener. As the basic information-carrying unit, Cohen points out, the word rather than the phoneme might be considered the building block in speech research. As for what concerns invariance at the word level, Cohen considers the listener's point of view and concludes that, "the only invariance we might claim is the one based on word types, to be seen as moulds in the shape of gestalts, stored in our mental lexicon, determined by the phonotactic constrains of the language." Cohen emphasizes the role of other types of information than acoustic information available to the listener to decode the successive words in the message, and cites the LAFS model (Klatt, 1979a) and the Logogen model (Morton, 1964) as possible models for speech perception. From Cohen's mentalist point of view (see Section 23a), the problems of segmentation into phoneme-size units and invariance of phoneme and feature become small relative to the problem of segmentation into word-size units and their invariance. I agree with the views expressed in Cohen's paper, most of which serve to point out the importance of the word as an unit in speech. However, I would like to try to fill in some of the details of these views.

WORD DEFINITION

The paper does not define in enough detail what a "word" is. The definition of the word, either as a unit of meaning or as an acoustic unit, is not clear cut. First, it may refer to the graphic form of the word, and blank spaces are considered boundary markers between successive words. Second, the word may also be defined as the basic unit of meaning, "the morpheme." A "graphic" word, however, may be composed by several morphemes, and morphemes (and not the graphic word) are considered by linguists as the basic unit of meaning. Moreover some words such as the grammatical words (articles or auxiliaries) do not have a meaning by themselves. Third, the separation of the "graphic words" into two word types and the notion of the prosodic word have been introduced by reference to the acoustic level for languages like French and English.

The words of the first type (type A), corresponding typically to function (grammatical) words, are generally very short words in terms of number of syllables, have a reduced duration, an average lower fundamental frequency and are not precisely articulated; the words of the second type (type B), typically lexical words, correspond to a local peak of prominence on one syllable (marked by durational and

fundamental frequency contrasts), with a lengthening of the word final syllable, and a strengthening of the word initial phoneme. Depending on the context, a lexical word may however behave as a word of type A and function word as a word of type B. Words of type A and of type B may be regrouped into what is called a single prosodic word (related to the notion of "rhythmic unit," "syntagm," "stress group," or "hat-pattern") primarily characterized by two F_0 movements in opposite direction. Such regrouping obscures the acoustic identity of the component words. Where the prosodic word starts and where it ends is difficult to determine, since there is no solid theoretical basis for such a demarcation.

For example, the two lexical words "petit garçon" are generally regrouped into a single four-syllable prosodic word. The two words may tend to separate however into two different acoustic units in certain circumstances, which depend on the speaker, the style, the rate of speech, the length of the words (shorter words may be regrouped more often), and the frequency of occurrence of the words in the lexicon and in the discourse. Also a long graphic, lexical word has a tendency to "spread" into two prosodic units depending on its morphemic composition (as marked in French by a lengthening of the final syllable of an internal morpheme). Integration or complete separation of two "graphic" words into a single acoustic unit, or separation of a graphic word into two "morphemic" units may not be a binary decision, but may vary along a continuum. In other words, the relationship among graphic words, units of meaning, and prosodic words is not necessarily a one-to-one correspondence.

REPRESENTATION OF THE WORD

Cohen does not express clearly what a gestalt at the level of the word may look like, since by definition, a gestalt must be invariant. A gestalt point-of-view is not incompatible with the view of the word as composed of smaller constituents (phonemes, for example) at least at some abstract level: Such a theory may be equally applied to the syllable, the demisyllable, the diphone, the phone, and the feature. The detailed acoustic shape of a word is known to be influenced by a large number of factors. By pointing out the importance of other sources of knowledge (such as syntax and semantics and phenomena such as phoneme restoration) Cohen suggests a gestalt of high complexity which questions the adequacy of the representation of the word as a string of phonemes.

One principle of phonetic transcription has been to use a set of symbols in the most economical and the most efficient way to represent the various utterances of a language. Linguists postulate two levels, or a continuum of levels, between an abstract representation—a phonemic level—and surface realizations—the phonetic level. A single phonemic symbol, such as the phoneme [R] or the phoneme [l] could be associated with spectrographically very different looking sounds, depending on the position of the phoneme in the syllable (e.g. initial, final, and in cluster) and in the word. It is not clear, for example, if all [R] and [l] allophones share even perceptual equivalence or invariant properties at the acoustic level. There may be some invariant cue across such positions, but the lack of invariance across different

SUPRASEGMENTAL FRAMEWORK AND SEGMENTAL FEATURES

There is a tendency in contemporary studies to separate the segmental and suprasegmental information contained in the signal and to interpret them separately. The relative duration of the events and the fundamental frequency contours delivered in parallel with the spectral information provide very useful information for both decoding the segmental information, and for detecting word boundaries, word stress, and word regrouping. However, the word has intrinsic segmental and suprasegmental characteristics which appear concomitantly. Striking similarities related to the word as an acoustic unit have been observed for a number of related and unrelated languages such as English, Swedish, French, and Japanese: word-final syllable lengthening, word-initial consonant lengthening, or word-initial allophones. Similarities also extend to the way with which the words are regrouped into larger units (F_0 rises associated with initiation, F_0 fall and lengthening with termination resetting of the base-line as boundary marker; see Vaissière, 1983a, for references). Such regularities may come from similar ways of representing words in the mental lexicon, and from general processes in composing sentences that are independent of the language spoken. However, there is a lack of knowledge of a large number of languages for definitive conclusions.

The segmental features have to be interpreted depending on their position in the sentence. The word as an unit imposes strong constraints on the acoustic structuring of the utterance. Analysis of the velum behavior in sentences shows the velum to be typically higher in word initial position and in prestressed position than in other positions for both the nasal consonants (which require an open velopharyngeal port) and the oral consonants (requiring a close velopharyngeal port) (Vaissière, 1983b). A higher position of the velum corresponds to a greater tensing of the levator palatini. It may be hypothesized that the difference in velum height between positional allophones may be due to the superposition of a common suprasegmental feature, let say [+strong], corresponding to a greater tensing at word onsets, rather than to fluctuations associated with each phoneme. With this hypothesis, the aspiration of the so-called tense stops consonants in English [p, t, k] should be considered as being due to the same factor as to the partial devoicing of word initial lax stops [b, d, g], the glottalization of word-initial vowels, the higher position of the velum in word-initial position, or the fact that vowels in sentence context (at least for French) seem to be, ceteris paribus, more precisely uttered if the structuring of the sentence requires them to be uttered in the upper F_0 register of the speaker.

Aspiration, devoicing, glottalization, higher position of the velum or more precise articulation have a common characteristic: a greater tensing of at least one of the articulators, the vocal folds, the velum (levator palitini), or the tongue. Such extra tensing may be contradictory or not to the articulatory requirements of the underlying segmental feature(s). Less intra- and interspeaker variability in velum height was observed, when the velum was supposed to be low (nasal) and not suprasegmentally tense (-strong), or high (oral) and suprasegmentally tense (+strong), than in the combination (+nasal), (+strong) or (oral), (-strong) (Vaissière, 1983b). The

tensing is not only a function of position in the word, but also a function of the relative importance of the word as an information carrying unit in the sentence context. If a word is sufficiently stressed, it tends to be uttered with an higher fundamental frequency and [p, t, k] may be aspirated even in final or medial position. As a consequence, segmental and suprasegmental characteristics of speech should be considered as intimately connected. Therefore, their separate interpretation should be avoided.

Thus EMG data, articulatory positions, and the acoustic signal may be interpreted as the result of a combination of both segmental features and the suprasegmental framework. The relative importance of suprasegmental variables as compared with the realizations of the segmental features is speaker-dependent, at least for the velum (Vaissière, 1983b). The lack of electromyographic, articulatory, and acoustic invariance for the distinctive features as discussed in this volume may be partly explained first by context-dependent differences in suprasegmental variables, second by a speaker-dependent way of combining suprasegmental influences and the segmental features, and third by unpredictable gestures unrelated to the content of speech (see my comments on speech ready gestures, in Section 10b). Such a lack of observed invariance does not argue against the possible existence of invariance at a more central level, before the integration of the segmental and suprasegmental features into the entire speech event.

CONCLUSION

The acoustic signal is intrinsically highly structured. An important determinant of this structure is the temporal variation of the relative tenseness of the articulatory processes in the realization of a given segmental feature. The word plays a great role in determining this structure and consequently it is a very important unit for interpreting the acoustic correlates of the distinctive features. However, the syllable, the phrase, and the sentence also play important roles; there are no unambiguous criteria to decide which unit is most important. The word is an obvious building block in speech perception and production, but such a view is not incompatible with the use of the phonemes as building blocks for constructing words. So instead of following Cohen's suggestion by "replacing" one building block (the phoneme) by another (the word), it may be more effective to search for an integration of the different acoustic units into a single framework.

Caution is advisable in extending the conclusions of studies of nonsense words to real words. The distinctiveness of a feature at the lexical level may play an important role in its mental representation since there seems to be a natural tendency to reduce redundancy in the communication process.

The decoding of the speech signal by the listener is known to be a complex process involving various types of normalization. This process is far from being fully understood. But our present inability to deal with this problem should not lead to ignoring it in modeling speech understanding, to an underestimation of what is actually contained in the speech signal, and an overemphasis of the role of syntactic, semantic, and pragmatic constraints. No doubt, when the words are not in dicta-

tion form but embedded in a context, the pronunciation of the word may become "sloppier" probably because other sources of knowledge constrain the number of possible words in each position so that the full realization of the complete set of features may become redundant. Nevertheless, literate listeners are perfectly able to discriminate meaningful or nonsense words pronounced in isolation and to transcribe them phonetically. One of the problems with invariance is that researchers are still confused about what unit should be invariant or "more" invariant than the others (features, phonemes, allophones, or even words?) and at which level of the communication process (functional invariance, perceptual equivalence, acoustic invariance or articulatory control?). The question of invariance remains the central problem in speech research the importance of which is not reduced by the acknowledgement of the word as an important unit. Even if no invariance will be found in the future, the search for invariance may provide a useful hypothesis and an adequate framework to studying speech.