THE ROLE OF PHONOLOGY IN SPEECH PROCESSING

Richard Wiese
Seminar für Allgemeine Sprachwissenschaft
Universität Düsseldorf
D-4000 Düsseldorf, FRG

1. Introduction

In this paper, I discuss the role of phonology in the modelling of speech processing. It will be argued that recent models of nonlinear representations in phonology should be put to use in speech processing systems (SPS). Models of phonology aim at the reconstruction of the phonological knowledge that speakers possess and utilize in speech processing. The most important function of phonology in SPS is, therefore, to put constraints on what can be expected in the speech stream. A second, more specific function relates to the particular emphasis of the phonological models mentioned above and outlined in § 4: It has been realized that many SPS do not make sufficient use of the suprasegmental aspects of the speech signal. But it is precisely in the domain of prosody where nonlinear phonology has made important progress in our insight into the phonological component of language.

From the phonetic point of view, phonological knowledge is higher level knowledge just as syntactic or semantic information. But since phonological knowledge is in an obvious way closer to the phonetic domain than syntax or semantics, it is even more surprising that phonological knowledge has been rarely applied systematically in SPS.

2. Prosodic factors in the variability of speech

One claim of this paper is that the proper use of phonology is one key to the successful handling of variability in speech. In (1), five versions of a common greeting in German are transcribed in a fairly close manner.

(1) Guten Morgen a. [ˌgu:tən 'mɔʁgən]
b. [ˌgu:tn̩ 'mɔʁ͓gn̩]
c. [ˌgun̩ 'mɔ(ɐ)gŋ̍]
d. [ˌgʊn 'mɔ(ɐ)ŋ̍]
e. [n̩mɔ̃ŋ̍]

The version (1a) is certainly overly careful even for speakers of the standard language in highly controlled situations. But it is precisely in front of the--ignorant--computer, that speakers might revert to a speech mode as the one in (1a). It has been noted that speakers talking to a SPS turn to careful, hyper-correct speech when repeating utterances that the system did not understand (Vaissière 1985: 204). If a system does not have the means for representing this very explicit form of speech, talking like in (1a) is no help for the system; in fact, it is even harder to understand for the system than the less careful speech. The SPS will almost necessarily fail to analyze the utterance although the speaker has made considerable effort to make himself understood.

On the other side of the scale of variability, there is reduction, assimilation and even deletion of sounds, which makes speech processing extremely difficult. (1b) might be the normative standard language version. Compared to (1a), the nasal consonants carry the syllabicity of the unstressed syllables. Also the r-sound will be more or less vocalized, and the final nasal will be assimilated to the plosive. (1c) and (1d) show further reductions in the segmental material. I assume that the various processes occur roughly in the steps given although nothing hinges on that. It is important, however, that the suprasegmental information is quite stable over all the reductions in the segmental material. (1a) to (1c) show the same number of syllables (as do d and e), and all versions share the same stress pattern. The unstressed syllables are the ones that disappear first, the syllable with secondary stress is reduced in (1e).

The conclusion is that reductions and omissions in speech are such that as much as possible is kept of the suprasegmental structure. Apart from this aspect, the example demonstrates a major problem for a SPS: The signal for what is regarded as one utterance can be, even in the abstract form given in (1), highly variable and context-dependent.

It is important to realize that phonology since its beginnings aims at the extraction of the relevant information from the speech stream. The concept of distinctive vs. predictable and accidental features is a cornerstone for all phonological theories. To see how this could be relevant for a SPS, we have to look at the structure of such a system.

3. The structure of a speech processing system

SPS analyze or synthesize speech in order to relate the speech signal to an utterance representation (text). The text could consist of the orthographic form of words or some other form closer to the representation of words in a mental lexicon. It is common for advanced SPS, however, to define an intermediate representation between the raw signal and the text. This representation, a symbolic code for phonetic categories, stands halfway between the unanalyzed signal and the textual or lexical representation. The broad structure of a SPS can therefore be depicted as (2).

(2)

Signal
↕
Symbolic Representation
↕
Textual or lexical Representation

As a first pass, the symbolic representation can be seen as a phonetic transcription, exemplified in (1). This reveals its intermediate nature: It codifies properties of the speech signal into discrete phonetic categories, but it also contains idiosyncratic features that are not part of the lexical representations or of the representation of the utterance.

The role of the symbolic representation in SPS can be illustrated as follows. In speech recognition, it serves as a meeting-point for the two kinds of procedures called upon in systems of this kind. For bottom-up analysis of the signal, results are outputted as pieces of the symbolic representation. For top-down procedures, i.e., hypotheses about what might occur in the signal, the output is again some piece of the representation. The requirements and possibilities for bottom-up and top-down analysis define to a large extent which criteria the symbolic representation has to meet: Whereas the signal is highly speaker-dependent, the symbolic representation is not. On the other hand, while a lexical representation of a word would not include predictable phonetic information, the phonetic transcription as a symbolic representation would contain information of this kind. In speech synthesis, lexical representations can first be translated into a phonetic representation which is then transformed into signal components. This two-step procedure for the adjustment of the phonetic forms to context influences such as assimilation between adjacent words can possibly very efficient. If lexical representations are mapped directly onto speech signals, it is hard to see how adjustments of this sort can be performed systematically.

I have been deliberately vague about the nature of the symbolic representation, because there are various proposals to this question. A number of units have been used and discussed as the elementary recognition or synthesis units, e.g., the phone, the diphone, the phoneme, the demi-syllable, and the syllable. The basic requirement for a symbolic representation in a general-purpose SPS would be that it is able to denote as much information as can be extracted from the signal or be deduced from the lexical representation. Thus, if the system can compute the occurrence of an allophonic variant of some sound, then this allophone should be representable in the symbolic representation. Similarly, if it is detected that two syllables are present in the signal, this fact should be encoded in the representation.

These considerations lead to the conclusion that the symbolic representation might be richer as is often assumed in existing systems. We will now show that phonological theory can help to define an adequate symbolic representation which is both a code for expressing phonetic categories and a model of the phonological knowledge of the language user.

4. Some recent developments in phonology

There is a long tradition in phonology to distinguish between segmental and suprasegmental features. Segmental features are those of the individual segment; suprasegmental ones belong to a domain larger than one segment.

But it is by no means clear in advance where a feature stands in this classification. To give an example, segments are often specified by the feature syllabic . A segment is syllabic if it stands in the peak of the syllable. Thus, in (3a) all the segments marked with a vertical line are syllabic, all others are not.

(3) a. [ɪ̩ntɔ̩wne̩rʃə̩n] b. [n̩to̩ne̩ʃn̩] <u>intonation</u>

But here, there are other pronunciations of the same word with different syllabic elements, such as (3b). What remains constant is that for each syllable there is exactly one syllabic peak. This suggests that syllabicity is not a segmental feature but suprasegmental.

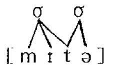
In this chapter, three examples are used to introduce some aspects of recent models in phonology. The examples are ambisyllabicity, vowel length and stress patterns; the constructs to deal with these are the syllable-node, the CV-tier and the metrical tree.

4.1. Ambisyllabicity and syllable structure

There is a common notation to mark syllable-boundaries by some symbol inserted into the segment string. But recent work on the syllable (such as Kiparsky 1979, Clements & Keyser 1983) has assigned to the syllable a more important role than just a boundary notion. That syllables are not just boundaries can be shown by the phenomenon of ambisyllabicity, which occurs in a number of languages.

It is well-known that in German words as <u>Mitte</u> or <u>lassen</u> the intervocalic consonants are a part of both syllables of each word. In view of this fact, it becomes a rather arbitrary and unmotivated decision to insert a syllable-boundary. But the syllable division and the ambisyllabic nature of some consonants can be naturally denoted if the syllable is given a hierarchial character. The notation for <u>Mitte</u> would then be as in (4), with 'σ' denoting the syllable node.

(4)

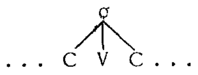


The segments and the syllable nodes appear on different rows or 'tiers' of the representation. This does away with the concept of the phonetic representation as a unilinear string. Elements on the different tiers are connected by 'association lines'. In the unmarked case, association is one-to-one, but in the case of an ambisyllabic segment association, association is one-to-many, as demonstrated by the /t/ in (4).

4.2. Vowel length and the CV-tier

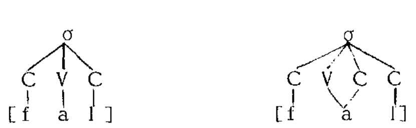
The syllable is probably more complex than is assumed in (4). This can be illustrated by the facts of vowel length. In German, which has contrastive vowel length, it appears that long vowels take up the space of a short vowel plus a consonant or of a diphthong (two short vowels). This is shown, e.g., by the fact that the maximal number of final consonants is 4 in a word with a short vowel (Herbst), but 3 in a word with a long vowel (Obst). To give formal recognition to the idea that a long vowel uses two positions in the syllable, although it is only one segment, yet another tier can be introduced into the syllable, called the CV-tier. It consists only of the elements C and V, where V denotes the syllabic nucleus of the syllable and C a consonantal position in the syllable. A syllable, then, is of the form (5); the maximal number of C-positions has to be determined for each language. The fact noted above that every syllable has exactly one syllabic nucleus can be expressed by letting V be an obligatory constituent of the syllable in the schema (5).

(5)



We have now a new formalism to express (phonological!) length not as a segmental feature such as long but as an association between the segmental tier and the CV-tier. The minimal pair Fall 'fall' vs. fahl 'pale' would be given the structural representation (6). With a given number of consonants following the V-position, the system also explains the fact that long vowels allow one consonant less in the syllable than short vowels.

(6)

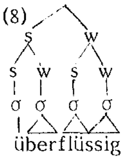


By treating phonological length as an association between tiers, I do not imply that all durational aspects of speech can be handled this way. There are other important timing relations between segments that determine intelligibility and naturalness of synthetic speech (see Huggins 1980). These have to be represented by other means, but are (partly) effects of the prosodic structure. Well-known examples include phrase-final lengthening and stress- timing vs. syllable timing.

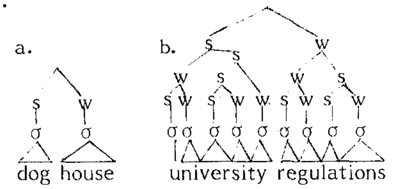
4.3. Stress patterns and the metrical tree

Moving up one or two levels in the prosodic hierarchy, there is the fact that strings of syllables occur in distinct accentuation patterns. It is part of the phonological competence of speakers of a language to be able to judge the accentual weight of a syllable with respect to its neighbouring syllables. In metrical models, this competence is formally expressed by metrical trees with syllables as terminal nodes. To give an example, the adjective überflüssig 'superfluous' has the highest degree of prominence on the first syllable, and the third syllable is relatively stronger than the last one. If a binary tree such as (8) is constructed over the syllables, and the nodes are labelled 's' (strong) and 'w' (weak), these accentual relations can be expressed easily and adequately. Syllabic and segmental detail is ignored in the examples here.

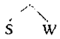
(8)



(9)


a.

b.

Interpreting accent as an abstract pattern over larger units has several advantages. It is, e.g., possible to give simple configurations as accent patterns for certain types of constructions. Compounds consisting of two words (in English as well as German) can be assigned the accent pattern  , independently of its internal accent pattern. (8) and (9) illustrate the situation. As (9b) shows, word-internal accent relations can become quite complex. This is not the point to discuss how trees of this kind are constructed, nor can we present alternatives that have been suggested.

A set of difficult questions arises if we ask how accentual patterns of this kind are realized phonetically. Notice that the metrical tree itself is quite uninformative in this respect. But this may turn out to be an advantage, since it is clear that there are a number of phonetic parameters correlating with accent. Intensity, length, F_0-movement, and vowel quality have all been identified as potential factors. But it may even be the case that listeners perceive an accent for which there is no cue in the signal. This is not so surprising, if accent is part of the phonological competence, and if at least some word-internal accents do not carry much information. Given that this is roughly a true picture of the situation, then it is a good strategy to have rather abstract accent representations which can be realized phonetically in a very flexible manner--and sometimes not at all.

5. Some consequences for speech processing

It is sometimes asked in speech processing work what should be the recognition or synthesis unit of SPS. The survey of phonological theory in § 4. reveals this to be a pseudo-question. There are hierarchies of units, and as far as they participate in phonological/phonetic processes, they are real and should be used in SPS. Therefore, the symbolic representation intermediate between the acoustic signal and the final representation of the utterance (see (1)) should be richer in structure than is generally assumed. It is not a string of units, but a multi-layered system of units. Some ingredients of this representation have been introduced above.

If prosodic information including the syllable is so important for speech processing, one might conclude that the use of a higher level unit such as the demi-syllable or the syllable is strongly recommended. But a consideration of some results of the morphology-phonology interaction shows this to be a precipitated conclusion.

Very often, wordinternal morpheme boundaries do not match syllable boundaries. If the phonetic information for the words *dog* and *bus* would be stored as the syllable templates [dog] and [bʌs], there would have to be additional templates for the plural forms [dogz] and [bʌs], [sɪz]. But plural formation in English is a very regular process, consisting of the affixation of a segment and a few rules depending on the nature of the final segment of the stem. Only if this segmental information is available to the system, a general algorithm for plural formation can work. Taking syllables as unanalyzable wholes would mean the spelling out of each plural form in the lexicon, thus nearly doubling the number of lexical representations. There are numerous similar examples in the morphology of languages like English and German.

In particular, there seem to be the following advantages in using a multi-linear representation of the kind sketched above. First, the representations derived from prosodic theories almost force the utilization of all kinds of information in the speech signal, especially suprasegmental information. This leads to a higher degree of predictability for segments. Take the example of word boundary detection, which is a crucial task for all SPS for connected speech. Different languages have different domains of syllabification. In some languages, e.g. English and German, the lexical word is the regular domain for syllabification. (Clitics, such as *it's* or *auf'm* (from *auf dem*) are the main exceptions.) But this is by no means a universal rule. In Mandarin Chinese, there is a good correlation between morphemes and syllables, which holds just as well as the one between words and syllables in English. In French, on the other hand, the domain for syllabification is a larger unit, say, the intonational phrase. It is the implementation of this kind of knowledge that makes it possible for a SPS to utilize information about syllable boundaries for the detection of word boundaries.

Secondly, the handling of both interspeaker and intraspeaker variation requires a framework in which the phonetic representation includes extensive prosodic structure. First, the rules governing variable speech (including fast-speech rules) are largely prosody dependent, as was illustrated in (1). An adequate formalization of the rules is thus only possible on the basis of prosodic representations. Second, extracting the relevant phonetic cues from the signal becomes easier if prosodic parameters are taken into account as fully as possible. Both vowel and consonant recognition is improved by taking into account F_o-values in the local context.

I have not addressed the computational side of the representational problem. It might be argued that a multilinear representation of the kind envisaged here is much harder to compute and represent in an actual SPS. But intelligent systems are quite able to deal with hierarchical or heterarchical objects of different kinds. Also, Woods (1985: 332) mentions the possibility of using cascaded ATNs for speech processing. Interlocking chains of ATNs could apply to recognize features, to bundle features into segments, to build syllables from segments, to combine syllables into words and to derive stress patterns for these words.

The general picture of a SPS assumed in this paper is that of a knowledge-based, intelligent system. I would like to stress that the phonological component is only component in such a system. But it is perhaps a component whose potential value has not been fully explored.

References

Clements, G.N. & S.J. Keyser (1983) CV-Phonology. A Generative Theory of the Syllable. Cambridge, Mass.: MIT-Press.

Huggins, A.W.F. (1978) 'Speech timing and intelligibility.' In: Requin, J. (ed.): Attention and Performance VII. Hillsdale, N.J.: Erlbaum.

Kiparsky, P. (1979) 'Metrical structure is cyclic.' Linguistic Inquiry 10, p. 421-441.

Vaissiere, J. (1985) 'Speech recognition: A tutorial.' In: Fallside, F. & W.A. Woods (eds.) Computer Speech Processing. Englewood Cliffs, N.J.: Prentice Hall, p. 191-242.

Woods, W.A. (1985) 'Language Processing for Speech Understanding.' In: Fallside, F. & W.A. Woods (eds.): Computer Speech Processing. Englewood Cliffs, N.J.: Prentice Hall, p. 305-334.