SYLLABLE-BASED PHONOLOGICAL RULES AND THEIR IMPLICATIONS FOR SPEECH RECOGNITION

Daniel Kahn

Bell Communications Research ("Bellcore")
435 South Street
Morristown, NJ 07960 USA

Abstract

Rules can be written which describe with fair accuracy the perceived syllabic structure of English. Once syllabic structure is established, many important phonological rules find natural expression in terms of this structure. In particular, phonemes tend to be modified under the influence of conditions that exist within the syllable in which they reside or when they play a particular role within their syllable. These observations provide support for the syllable-based approach to speech recognition, but the explicit rules that arise from syllabic phonology are applicable to phoneme-based recognition as well.

1. Introduction

Phoneticians as well as workers in the field of automatic speech recognition (ASR) are well aware of the lack of anything close to a one-to-one correspondence between the phonemes of a language and acoustic events. While the complexity of the mapping from phoneme to sound does not preclude the creation of an effective ASR device whose basic unit of recognition is the phoneme, it is clear that the success of such an undertaking is dependent upon discovering the large set of relevant context-sensitive rules, making them explicit, and encoding them in the recognizer. Even proponents of such an approach recognize the enormity of the task (cf. Zue, 1985).

Whole-word template-matching (cf. Itakura, 1975; Rabiner & Levinson, 1981) is an approach to ASR which appears to obviate the need for the long and difficult program of discovery of the details of the phoneme-to-sound mapping. In this technique, no explicit decision is made regarding where in time each phoneme lies and what its identity might be. Rather, for each word in a vocabulary, a reference template is created consisting of a set of spectral representations computed at regular intervals in time, on the order of every 10 msec. The sequence of spectral representations of a word to be recognized is then compared to each of the templates (after time-normalization) and the unknown word is taken to have the same identity as the template to which it has the least total spectral "distance," appropriately computed.

Whole-word matching works very well for recognizing small vocabularies of words spoken in isolation. As vocabulary size increases, a disadvantage of this approach become apparent: a new template must be created, stored, and included in the distance calculation for each additional word in the vocabulary. In addition, much of the advantage of whole-word matching is lost in continuous speech, since word boundaries are not easily determinable and, in any case, cross-word-boundary phonology can greatly alter the isolated form of words.

It has occurred to several ASR researchers that most of the advantages of the phoneme-based approach (finite vocabulary size, straightforward extension to continuous speech in many cases) and of the whole-word template-matching method (no need for explicit representation of many complex contextual effects) can be combined in an approach to ASR in which the basic unit is the syllable or demisyllable. Inherent in the advocacy of syllable-based recognition is the assumption that most contextual variation on the part of phonemes is due to the influence of other phonemes within the same syllable, and that the effects of the environment outside the syllable in which a given phoneme lies can for the most part be considered second-order (cf. Fujimura, 1975; Mermelstein, 1975; Kahn et al, 1984).

In the last ten years several groups have taken important first steps toward the implementation of high-performance (demi)syllable-based recognition systems (e.g., De Mori et al, 1976; Ruske & Schotola, 1978; Zwicker et al, 1979; Hunt et al, 1980; Ruske, 1982; Rosenberg et al, 1983), and it is to be hoped that this work will continue.

I too have performed some (very preliminary) work in syllable-based (Kahn, 1982, 1983) and demisyllable-based (Rosenberg et al, 1983; Kahn et al, 1984) recognition, but the present paper is concerned with the linguistic motivation for the use of (demi)syllabic units in ASR. I believe, however, that not only does the phonological analysis discussed below argue for the wisdom of the (demi)syllable approach, but also that the explicit rule formulations that are an output of the syllable-based analysis can profitably be used in phoneme-based recognition.

2. The syllable in English phonology

In many languages it is obvious to native speakers how words of their language are to be syllabified, but English has both clear (*reply* = *re*–*ply*, not *rep*–*ly* or *repl*–*y*) and unclear (*pony* = *po*–*ny* or *pon*–*y*?) cases. This apparent indeterminateness has led the authors of many formal accounts of English phonology to deny the syllable a role in linguistic descriptions. This is unfortunate, because the concept of "syllable" is intuitively meaningful even to speakers of languages like English, and also because many phonological rules call out for descriptions in terms of the syllable, if only the concept could be formalized.

In Kahn (1980) I suggested an analysis of English syllable structure that I feel accounts well for both the clear and unclear cases of word syllabification, as well as for the syllabification of phrases in the case of continuous speech (where a syllable may extend across a word boundary). Most important, once syllabic structure is established in accordance with this analysis, many important phonological rules (sound modifications) can be expressed in a natural and compact way in terms of the syllable. In the limited space available here I will try to outline the analysis of English syllabification and discuss some examples of syllable-based rules. In all cases, I will have to omit details which may be significant but which do not, I believe, affect the correctness of the basic analysis.

2.1 Analysis of words and phrases into syllables

There is little controversy as to how many syllables a normally-spoken word contains. At the core of each syllable is exactly one vowel or other "syllabic" phoneme (like [n̩] of *button*). Each syllable will also contain zero or more non-syllabic phonemes (which I will imprecisely refer to as "consonants") before and after the vowel. Clearly any word-initial (-final) consonants must reside in the first (last) syllable of the word. Thus the question of interest is whether, in words of more than one syllable, to associate consonants that stand between two vowels with the preceding or following syllable.

In this regard, it is surely significant that any polysyllabic word of English can be broken down into syllables each one of which could stand alone as an English word without breaking the constraints on permissible *word*-initial and -final clusters. Thus English has words like *hamster*, corresponding to the permissibility of word-final /m/ and -initial /st/, but none like *hamkter* since there is no analysis of /mkt/ into permissible clusters. A natural conclusion from this observation is that English simply has a set of permissible *syllable*-initial and -final clusters, from which the facts about *word*-initial and -final clusters fall out as an immediate consequence.

The question remains how to correctly predict syllabifications in cases where more than one analysis is consistent with the cluster constraints (why *re*–*ply*, not *rep*–*ly*?). The answer appears to reside in the "maximal initial cluster" (MIC) principle: a syllable boundary is placed in a sequence of between-vowel consonants as far left as possible, consistent with the initial/final cluster constraints.

The MIC principle alone will, in general, predict correct syllabifications for what were referred to above as the "clear" cases. Even in the unclear cases, MIC appears to be correct, provided we look at overly precise, very-slow-speech pronunciations. In such speech we observe *po*–*ny*, not *pon*–*y*; *ci*–*ty*, not *cit*–*y*; *Pa*–*trick*, not *Pat*–*rick*.

Before returning to normal-rate syllabifications, it will be helpful to introduce a graphical representation of syllabification. Fig. 1 indicates that the word *reply* consists of two syllables, *re* and *ply*. Note that if we impose the natural constraint that the lines connecting syllables and phonemes may not cross, a whole class of syllabifications, like that in Fig. 2 in which the /r/ of *reply* is a member of the *second* syllable, become, quite appropriately, impossible to represent.

Now suppose that there are no further constraints on linking syllables and phonemes (aside from the one-syllable-one-vowel principle mentioned earlier). Then in addition to the syllabification of *pony* shown in Fig. 3, which, as noted above, is appropriate for the slow-speech pronunciation of this word, we might try to interpret the syllabification of Fig. 4. In Fig. 4, the /n/ of *pony* is shown as belonging simultaneously to both syllables, i.e., as being "ambisyllabic." I would suggest that this is the normal-rate syllabification of the word. The native speaker's inability to assign the /n/ of *pony* unambiguously to one or the other syllable in the normal-rate pronunciation of the word would then be attributed to the /n/ being ambisyllabic at normal rates (and in fact some phoneticians, in informal descriptions of English syllabification, have suggested that such consonants might be shared by two syllables). We can formalize the structural change in going from slow to normal speech as the addition of the line of association between /n/ and the first syllable.

The consequences of such an analysis go well beyond formalizing the intuition that certain consonants in English do, and others do not, reside fully in one syllable; there are phonological implications as well. For example, the simple rule "vowels become nasalized in English when followed by a nasal consonant in the same syllable" accounts for the /õ/ of *tone* and normal-rate *pony* alongside the /o/ of *poke* and slow-speech *pony*. French nasalized vowels are the result of a similar rule (*an* vs. *année*). Sect. 2.2 is concerned with examples of this type of rule.

We have not yet discussed under what conditions we observe ambisyllabicity; for ex., as opposed to *pony*, the syllabification of *reply* has the simple form given in Fig. 1 for both slow and normal speech. As discussed in more detail in Kahn (1980), it appears that the initial consonant of an *unstressed* syllable becomes ambisyllabic with a preceding vowel-final syllable. Thus it is the stress on the second syllable of *reply* that blocks ambisyllabification of the /p/.

To this point we have been discussing the syllabification of words in isolation. Turning to continuous speech, let us note first that it is always at least possible to pause between words, so a reasonable approach to continuous speech would be to postulate an initial level at which syllabification is in accordance with the "word-is-an-island" rules of the preceding paragraphs, with additional lines of syllabic association across word boundaries added by "continuous-speech rules." The most important of these rules appears to add a line of association (e.g., the dotted line in Fig. 5b) between the final consonant of a word and the initial syllable of a following *vowel-initial* word. This rule of "trans-word-boundary ambisyllabification" (TWA) can be understood when it is recalled that the clearly preferred syllable structure among the world's languages is ...CV-CV..., not ...VC-VC... Within words, this fact is reflected in the MIC principle. MIC is powerless, however, in the case of a word that happens to start with a vowel. In continuous speech, the unnatural situation of a vowel-initial syllable is remedied, where possible, by TWA. Thus *rocket* and *rock it*, syllabically distinct in slow speech (solid lines of association in Fig. 5), become homophonous at normal rates (addition of dashed lines).

2.2 Rules sensitive to syllabic structure

Many important phonological rules of English (and other languages) are best described in terms of syllabic structure. The outline of English syllabic structure given above is sufficient to illustrate several of these rules.

It is well known that the voiceless stops, and in particular /t/, take very different form as a function of environment. For example, /t/ is an aspirated stop in *tack*, an unaspirated stop in *stack*, a "flap" in *city* (Am. and Can. pronunciation) and is glottalized in *sit*. I would suggest that the rules responsible for these forms state that /t/, underlyingly an unaspirated stop, is aspirated when only syllable-initial, flapped when ambisyllabic, and glottalized when following a vowel and not syllable-initial. It is straightforward to confirm that these rules operate properly in simple cases like the words just cited, but the rules make other testable predictions. Thus in the phrase *Let Ann do it* we expect - and observe - glottalized /t/ in *let* if there happens to be a pause after the word but flapped /t/ in continuous speech, where TWA has applied. Similarly, in overprecise speech, where the (within-word) ambisyllabification rule fails to apply, the /t/ of *city*, normally ambisyllabic and flapped, has syllable-initial association only, and is aspirated. Of course, rules such as the ones that account for the various allophones of /t/ *could* be stated without reference to syllabic structure, but they would be grossly complicated, and would in fact be restating the independently-needed rules of English syllabification within the specific allophonic rules (cf. Kahn, 1981).

In standard British English and in parts of the Eastern U.S., /r/ is deleted in certain environments where spelling and the more "conservative" dialects would have it pronounced. The rule accounting for these facts, as it entered the language, is clearly syllable-conditioned and takes very much the form of the /t/-glottalization rule. Thus /r/ is lost when not syllable-initial, as in *form*, *for me*, *for(pause)Ann*, but is retained in *forest*, where /r/ is syllable-initial by MIC (and, irrelevantly, also syllable-final at normal rate by ambisyllabification), and *for(no-pause)Ann*, where /r/ is syllable-initial by TWA. French "liaison" is a more complex, though clearly related, phenomenon. If we regard a word like *vous* as consisting of the phonemes /vuz/ at an abstract level, and delete /z/ when not syllable-initial, then the TWA-like rule of French will account for *vous avez* [vuzave] vs. *vous l'avez* [vulave].

There is another very large class of rules which are clearly syllable-conditioned but differ in having been "frozen" at the lexical level. In most dialects of English, the vowel of *car*, through the influence of the following back phoneme /r/ (which until quite recently was pronounced in *all* dialects), has a distinctly more back quality than the vowel /ae/ of *cat*, *cap*, etc. (As suggested by the spelling, the vowels of *car*, *cat*, etc. were at one time identical.) The /ae/ of words like *carry*, however, was unaffected by the rule that modified *car*. We can account for these facts by imposing the natural condition that /r/ be fully in the syllable of the vowel it follows for it to have the backing effect. In accordance with this rule, words like *card* also have the backed vowel. The rule is "frozen" in the sense that words whose base form became subject to the rule now show the backed vowel even in non-base forms which should not be subject to the rule. Thus *starry* has the vowel of *star*, not of *carry*. Similar rules have affected other vowels: *her*, *herd* (vowel modified by /r/) vs. *hem*, *herring* (not).

A similar rule, but in the domain of consonants, accounts for the loss of /g/ in *long* [loŋ] vs. its retention in *longer* [ŋg]. Basically, /ŋg/ is simplified to /ŋ/ except when /g/ is syllable-initial. In the case of words of the form VngC...V, this rule correctly predicts [ŋ] without [g] (e.g., *angstrom* and *Yngve*) *except* when C is such that /gC/ forms a permissible initial cluster: *angry* (cf. grow), *linguist* ["ling-gwist"] (cf. Gwendolyn). Previous, non-syllabic analyses of *ng* did not properly account for these facts and could be made to only through explicit reference to the differential behavior of *gs* etc. vs. *gr* etc.; but clearly the correct course is to state the latter distinction once and for all in the (independently-required) permissible-cluster rules.

Additional examples of syllable-conditioned rules could easily be cited. At this point, however, let us note that a common feature of the rules that have been discussed is that they involve major changes, as viewed by the phonetician. That is, these rules delete segments or replace one well-defined phonetic element with another. Another class of rules, not generally considered to be in the realm of traditional phonology, deals with phenomena at a lower level. Thus, for ex., the phonetician (and the native speaker) hears the /i/'s of *bee* and *Dee* to be identical, even though the initial parts of the two vowels are spectrally quite distinct, due to the formant-transition phenomenon. Although the separation between a phoneme causing an acoustic modification and the modified phoneme is sometimes surprisingly large, it is probably fair to say that the strongest effects are found within the syllable and thus might be regarded as simply very-low-level syllable-based phonetic rules (cf. Malmberg, 1955; Fujimura, 1975, 1976).

3. Conclusion

This paper has been concerned with syllable-based phonetics and phonology and their relevance to ASR. Whether one is attempting to predict what phonemes are allowable in a particular environment or the precise acoustic shape of a given phoneme, local syllabic structure is most often found to be significant. In ASR systems based on syllabic units, such dependencies come "built-in." Even to the worker committed to phoneme-based ASR, however, syllable-based phonology is relevant because it offers compact and explicit formulations of many phoneme realization rules.

References

De Mori, R., Laface, P., & Piccolo, E., "Automatic detection and description of syllabic features in continuous speech," *IEEE Trans. on ASSP 24:5*, 365-79 (1976).

Hunt, M. J., Lennig, M., & Mermelstein, P., "Experiments in syllable-based recognition of continuous speech," *Int. Conf. on ASSP*, 880-3 (1980).

Itakura, F., "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. on ASSP 23:1*, 67-72 (1975).

Fujimura, O., "Syllable as a unit of speech recognition," *IEEE Trans. on ASSP 23:1*, 82-87 (1975).

Fujimura, O., "Syllables as concatenated demisyllables and affixes," *J. Acoust. Soc. of America, 59:* S55(A) (1976).

Kahn, D., *Syllable-based Generalizations in English Phonology*, Garland Publishing, New York (1980).

Kahn, D., "Syllable-structure specifications in phonological rules," in *Juncture*, Aronoff, M. & Kean, M.-L., eds., Anima Libri (1981).

Kahn, D., "A syllable-parsing algorithm for telephone-quality speech," *J. Acoust. Soc. of America, 72:* S30 (1982).

Kahn, D., "A syllable-based connected-digit recognizer for continuous speech," *J. Acoust. Soc. of America, 74:* S31 (1983).

Kahn, D., Rabiner, L. R., & Rosenberg, A. E., "On duration and smoothing rules in a demisyllable-based isolated-word recognition system," *J. Acoust. Soc. of Am., 75:2*, 590-8 (1984).

Malmberg, B. (1955), "The phonetic basis for syllable division," in *Readings in Acoustic Phonetics*, Lehiste, I., ed., MIT Press.

Mermelstein, P., "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. of Am., 58:4*, 880-3 (1975).

Rabiner, L. & Levinson, S., "Isolated and connected word recognition - theory and selected applications," *IEEE Trans. on Comm., 29:5*, 621-659 (1981).

Rosenberg, A. E., Rabiner, L. R., Wilpon, J. G., & Kahn, D., "Demisyllable-based isolated word recognition system," *IEEE Trans. on ASSP 31:3*, 713-26 (1983).

Ruske, G., "Automatic recognition of syllabic speech segments using spectral and temporal features," *Int. Conf. on ASSP*, 550-3 (1982).

Ruske, G. & Schotola, T., "An approach to speech recognition using syllabic decision units," *Int. Conf. on ASSP*, 722-5 (1978).

Zue, V. recently estimated that another "30 to 40 years" of work in acoustic phonetics for ASR remained to be done and that what is currently known is just the "tip of the iceberg" (IEEE-ASSP Workshop on Speech Recognition, Dec., 1985).

Zwicker, E., Terhardt, E., & Paulus, E., "Automatic speech recognition using psychoacoustic models," *J. Acoust. Soc. of Am., 65:2*, 487-98 (1979). 2/18 09

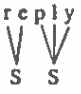

Figure 1

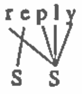

Figure 2

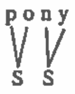

Figure 3

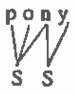

Figure 4

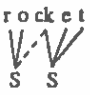

Figure 5a

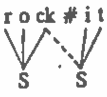

Figure 5b