



PHONOLOGICAL RULES AND SPEECH RECOGNITION

Charles Hoequist, Jr.*

ABSTRACT

There has been little interaction between speech recognition and linguistic phonology, due to the different aims of the two fields. It is proposed here that information about phonological processes can be of use in recognition. This paper compares two phonological rule components in a system having limited dialect normalization, one illustrating context-free rules and one making use of context sensitivity. It is argued that a context-free set of phonological rules is inadequate to deal with phonological processes in natural language.

INTRODUCTION

Techniques used for isolated-word recognition suffer badly degraded performance when confronted with continuous speech, resulting both from the absence of obvious acoustic word boundaries and from the increased variability of words owing to coarticulation at their boundaries. Fortunately, this increased variability is by no means random, but demonstrates regularities that can be (and in linguistic theory, are) formalized as phonological rules.

Given the desire to capture linguistic regularities in speech production, nothing seems to stand in the way of taking a reasonably comprehensive set of rules from some linguistic description and using them in a recognition system. The rewrite-rule formalism popular in linguistics is explicit enough to make this straightforward. Nonetheless, such a short step is impractical.

DIALECTS AND PHONOLOGICAL RULES IN LINGUISTICS

Most examples of phonological rules published by linguists have not been intended to describe English phonology, but to illustrate a particular phonological process or (more often) to argue for or against some aspect of rule formalism, and are thus extremely limited in scope. This is not meant to be an indictment of linguists. It is not the purpose of work in phonology to write a comprehensive descriptive grammar of a particular sound system. The point here is that the work done in that area does not have the same goals, and so is not easily transferable, to ASR.

A more serious problem is the formal power of most linguistic rules. Phonological descriptions of the last quarter-century have been overwhelmingly cast as a single set of context-sensitive transformational rules, that is, rules capable of rewriting the phrase markers making up a string, here a string composed of phonetic segment labels. Difficulties with this type of rule will be discussed later.

DIALECTS AND PHONOLOGICAL RULES IN ASR

The difference in goals is also evident in the lack of attention given in general linguistics to variation across speakers and dialects. In ASR, speaker variation is a major hurdle, and to date, dialect variation has been subsumed under it.

*Linguistics Department, University of Cambridge

Speech-recognition work to date often implicitly collapses dialectal and idiolectal variation together as part of the problem of speaker normalization. This makes the recognizer's task more difficult, since it collapses variation which is rule-governed for some group of speakers together with speaker-specific variation. To the degree that systematic variation can instead be predicted for all or part of the speaker population, the normalization the system needs to do is reduced. ASR work at Cambridge is therefore proceeding on the assumption of prior dialect identification as part of the initial calibration of the system to a speaker (ref 1). It is hoped that this lessen the amount of normalization necessary, as well as restricting rule application, and thereby reducing the number of hypotheses resulting from an input string.

But even if the rules in a phonological component are marked for which dialect they apply to, the context-sensitive transformational rule formalism is itself problematic. This is because it is in some cases impossible to uniquely reconstruct the pre-transformation string, because of the ability of transformational rules to modify phrase markers. In practice, this can lead to a given phonetic surface string being traceable to numerous possible phonological strings, analogous to syntactic ambiguity. As an input utterance grows longer, the possible parses multiply.

As in much recent work in syntax, one solution is to restrict the formal power of rules. To see whether this is satisfactory, two phonological parsers will be discussed here, one using exclusively context-free rules, the other allowing some context specification.

CONTEXT-FREE PHONOLOGICAL RULES

Using only context-free rules in the phonological component of a recognizer might seem hopeless at first; phonological processes are heavily context-dependent. As will be seen, the 'context-free' system discussed here does not ignore context, but encodes it in such a way that a context-sensitive formalism is claimed to be unnecessary.

The most developed example of a context-free phonological parser is to be found in Church (ref 2). There, a segment lattice with no word- or syllable-boundary specifications serves as the input to a chart parser operating with a set of context-free phrase structure rules. The parser outputs a string, enhanced by the insertion of syllable-boundary markers. This serves in its turn as input to a syllable-based lexicon.

The encoding of phonological processes, many of them dependent on segmental context, into context-free rules is done by encoding them into a hierarchical structure with phonetic segments as its lowest level. Contextual dependencies are encoded into higher-level units (syllables and feet), thus enabling the parser to handle many context-conditioned processes without requiring the generative capacity of a context-sensitive system. Nonetheless, the system (at least as presented by Church) suffers from two serious problems.

The first difficulty can be characterized as the perfect-input requirement. This refers to the system's need for error-free, fine-grained phonetic labeling performance. Even labeling in too broad a fashion can be disastrous. For example, one of Church's rules involves the realization of voiceless stops as aspirated if and only if they are dominated by a syllable-onset node (in other words, syllable-initial voiceless stops in English are aspirated; note how information on context-dependent realizations is encoded). If the recognizer's front end ever

fails to recognize a syllable-initial voiceless stop as aspirated, the parse fails. It is overoptimistic to expect flawless performance from any front end. This difficulty weighs heavily, because a proper parse in this system depends on the accurate labeling of such phonetic detail. The problem can be evaded by making the rules more general, but at the cost of dramatically increasing the number of valid parses formed from any given string.

Even if it were reasonable to expect good fine-grained labeling, a phonological-rule recognizer component would still suffer serious deficiencies owing to its inability to handle segmental context. Since segment labels cannot be altered, a context-free parser's job is really to judge whether its input is phonotactically possible. This fails to address the existence of phonological processes involving segment insertion (e.g. epenthetic [t]), deletion (e.g. disappearance of schwa vowels) or neutralization (e.g. reduction to schwa of vowels which are different in the lexicon). Such rule-governed alterations in a segment string are also troublesome because they can yield surface realizations violating the phonotactics of English. To avoid rejection of such strings, a recognition system would have to loosen its restrictions on allowable input, with (as before) the consequence of a sharp rise in the number of valid parses resulting from an input.

CONTEXT-SENSITIVE RULES

The example of context-sensitive parsing discussed here is the 'two-level' parser developed first by Koskenniemi (ref 3). Though his intention was to use it for morphological decomposition of an input string, it is usable in phonological rule implementation as well. The program package used at Cambridge is that developed by Ritchie et al. (ref 4).

The core of two-level parsing is its encoding of rules into nondeterministic, finite-state automata, which simulate context-sensitive rules. Since context-sensitive rules are formally more powerful than finite-state machines, it would be possible to write rules to generate strings not correctly analyzable by any finite-state device. However, phonological and phonetic processes in natural language do not ever seem to produce such outputs. The single rewriting of symbols which the automata can perform (rather than the unlimited rewriting of true context-sensitive rules) seems to be adequate. These automata can be envisioned as moving simultaneously along two 'tapes' (hence the name 'two-level'), one representing the input and one representing a graph through a lexicon having a tree structure.

The rules check whether the current input character can be matched to the current lexical character. If all the rules allow the current pairing, the next character is taken from the input and checked for any allowable matches in the lexicon, and the process is repeated. This continues until the end of the input is reached (a successful parse) or no pairing of lexical and input characters is possible at some point.

This shares with the context-free system the advantage of not needing word boundaries to be specified in advance, and so avoiding the problems caused by the reliance of a whole-word matcher on reliable word boundaries. It shows some structural differences, for example, in contrast to Church's parser, the rules are context-sensitive, and the parse is left-to-right. Additionally, in the two-level system, the lexical search takes place as part of the parse. Therefore, a parse may fail simply because an input word is not present in the lexicon (the output of the Church parser is independent of the lexicon). An advantage of the simultaneous parse and

lexical search is that many unusable analyses, consisting of phonotactically legitimate nonwords, are filtered out early. More important than any of these is of course the possibility of limited context specification in the rules that rewrite segment labels, thus allowing for the restoration of deletions, removal of insertions and recovery of neutralizations.

CONCLUSION

Many phonological processes often expressed in a context-sensitive formalism can be expressed in a more restricted system. However, the inability of context-free phonological rules to alter the segments constituting their input puts them at a severe disadvantage when faced with the output of those processes in speech production that alter segment identity (e.g. neutralization) or create surface violations of phonotactic rules through insertion or deletion of segments. Though it is possible to create a recognizer without context-sensitive rules between the input and the lexicon (recognizers can function without rules at all), part of the input variation due to context-sensitive processes must then be taken care of somewhere else.

While a rule component allowing context sensitivity is superior in this respect, it cannot be said to be optimal. The rules' formal power enables them to undo the effects of processes which defeat context-free rules and also overgenerates mappings between the input and the lexicon. The obvious next step is to find ways to prune these mappings. Two methods are already in use. One is that of automatically checking the lexicon to see whether it contains the segment sequences the rules produce, the other is the reliance on an early identification of the speaker's dialect to determine which rules will be applied to an input string. Further reduction in the number of hypotheses could be achieved using syntactic knowledge, and the linking in analysis of rules found to co-occur in speech.

REFERENCES

1. W. J. Barry, Automatic Identification of Regional Accent: Theory and Practice. Cambridge Papers in Phonetics and Experimental Linguistics 4 (1985).
2. K. W. Church, Phrase-Structure Parsing: A Method for Taking Advantage of Allophonic Constraints (Ph.D. dissertation, distributed by IULC, June 1983).
3. K. Koskeniemi, Two-Level Morphology. Texas Linguistic Forum 22, (1983).
4. G. D. Ritchie, A. W. Black, S. J. Pulman, and G. J. Russell, The Edinburgh/Cambridge Morphological Analyser and Dictionary System (Prototype: Version 2.2) User Manual. Cambridge University Computer Laboratory (1986).

ACKNOWLEDGMENTS

This paper is based on work carried out as part of the Linguistics Department's component (SERC grant GR/D/42405) of Alvey research project MM1/069 on Automatic Speech Recognition, which involves Cambridge University, the MRC Applied Psychology Unit, and STC Technology Ltd.