Allophonic and Phonotactic Constraints are Useful

Kenneth W. Church
NE43-307
Massachusetts Institute of Technology
Cambridge, MA. 02139
KWC@MIT-ML

This paper argues that allophonic and phonotactic cues are a source of constraint, not a source of noise as many speech researchers have assumed in the past. These constraints are formulated so that they can be exploited with well-known parsing techniques.

1. Parsing at the Phonetic Level

It is well known that phonemes have different acoustic/phonetic realizations depending on the context.[1] For example, the phoneme /t/ is typically realized with a different allophone (phonetic variant) in syllable initial position than in syllable final position. In syllable initial position (e.g., Lorn), /t/ is almost always released (with a strong burst of energy) and aspirated (with /h/ like noise), whereas in syllable final position (e.g., ca[), /t/ is often unreleased and unaspirated. It is common practice in speech research to distinguish acoustic/phonetic properties that vary a great deal with context (e.g., release and aspiration) from those that are relatively invariant to context (e.g., place, manner and voicing).[2] In the past, the emphasis has been on invariants; allophonic variation is traditionally seen as problematic for recognition.

(1) "In most systems for sentence recognition, such modifications must be viewed as a kind of 'noise' that makes it more difficult to hypothesize lexical candidates given an input phonetic transcription. To see that this must be the case, we note that each phonological rule [in an example to be presented below] results in irreversible ambiguity - the phonological rule does not have a unique inverse that could be used to recover the underlying phonemic representation for a lexical item. For example,... schwa vowels could be the first vowel in a word like 'about' or the surface realization of almost any English vowel appearing in a sufficiently destressed word. The tongue flap [£] could have come from a /t/ or a /d/." Klatt (MIT) [8, pp. 548-549]

This position is representative of much of the speech recognition literature, especially during the ARPA speech project. One can find similar statements by Cole and Jakimik (CMU) [2] and by Jelinek (IBM) [5]. I prefer to view both variant and invariant cues are helpful: variant cues reveal properties of the suprasegmental context and invariant cues reveal properties of the local segmental identity. This much has been observed elsewhere. For example, the following minimal pairs have been used by many authors to show that allophones of /t/ can be distinctive.

(2a)	a tease / at ease	aspirated / flapped
(2b)	night rate / nitrate	unreleased / retroflexed
(2c)	great wine / gray twine	unreleased / rounded

Unfortunately, these allophonic constraints on syllable structure and word stress have never been adequately integrated into a practical recognition system. I have attempted to remedy this situation in [1], where I proposed (and partly implemented) a recognizer that exploited contextually dependent cues (e.g., aspiration) by parsing the input utterance into syllables and other suprasegmental constituents using phrase-structure parsing techniques (e.g., Earley's Algorithm [3]). Invariant constraints were applied in the usual way to match portions of the utterance with entries from the lexicon.

1. This research was supported (in part) by the National Institutes of Health Grant No. 1 P01 LM 03374-01 and 03374-02 from the National Library of Medicine.

2. *Place* refers to the location of the constriction in the vocal tract. Examples include: labial (at the lips) /p, b, f, v, m/, velar /k, g, ŋ/, dental (at the teeth) /s, z, t, d, l, n/ and palatal /š, ž, č, j/. *Manner* distinguishes among vowels, liquids and glides (e.g., /l, r, y, w/), fricatives (e.g., /s, z, f, v/), nasals (e.g., /n, m, ŋ/) and stops (e.g., /p, t, k, b, d, g/). *Voicing* (periodic vibration of the vocal folds) distinguishes sounds like /b, d, g/ from sounds like /p, t, k/.

2. An Example of Lexical Retrieval

It might be helpful to work out an example in order to illustrate how parsing can play a role in lexical retrieval. Consider the phonetic transcription, mentioned above in the citation from Klatt [8, pp. 548-549]:

(3) [dʃjəhIʃ[Itɪtam]

It is desired to decode (3) into the string of words:

(4) Did you hit it to Tom?

In practice, the lexical retrieval problem is complicated by errors in the front end. However, even with an ideal error-free front-end, it is difficult to decode (3) because, among other things, there are extensive rule-governed changes affecting the way that words are pronounced in different sentence contexts, as Klatt's example illustrates:

(5a) Palatalization of /d/ before /y/ in did you
(5b) Reduction of unstressed /u/ to schwa in you
(5c) Flapping of intervocalic /t/ in hit it
(5d) Reduction of schwa and devoicing of /u/ in to
(5e) Reduction of geminate /t/ in it to

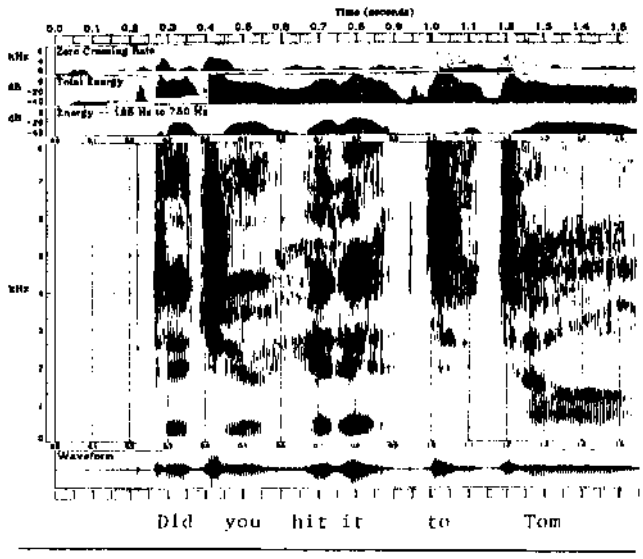
It is very difficult for the recognition device to "undo" these phonological transformations. Inverse transformational parsing is generally considered among computational syntacticians to be unlikely to succeed. Even Stan Petrick (personal communication), one of the few proponents of inverse transformational parsing at the syntactic level, agrees that inverse transformational parsing methods are unlikely to work well with phonological rules. In particular, allophonic processes often appear to neutralize phonemic distinctions. For example, the voicing contrast between /t/ and /d/, which is usually distinctive, is almost[3] completely lost where both /t/ and /d/ are realized in American English with a tongue flap [ɾ].

3. Parsing and Matching

As an alternative to inverse transformational parsing, I will factor the lexical retrieval problem into two (hopefully simpler) sub-problems: (a) parse the segment lattice into syllable structure, and (b) match the resulting constituents against the lexicon. As suggested above, variant phonological cues help constrain the parsing step; invariant cues restrict the matching process.

3. The contrast is not completely lost. In general, vowels are longer before voiced consonants than before unvoiced consonants. Thus, the underlined vowel in rider tends to be longer than the corresponding vowel in writer.

Fig. 1. Did you hit it to Tom?



Allow me to illustrate the approach with Klatt's example (enhanced with allophonic diacritics to show aspiration and glottalization):[4]

(6) [dʃjəhIʃ[ɪʔtʰɪtʰam]
↑↑ ↑

Using the phonotactic and allophonic constraints on syllable structure:[5]

(7a) /h/ is always syllable initial, *phonotactic*
(7b) [ʃ] is always syllable final, *allophonic*
(7c) [ʔ] is always syllable final, and *allophonic*
(7d) [tʰ] is always syllable initial. *allophonic*

the parser can insert the following syllable boundaries:[6]

(8) [dʃjə # hIʃ # ɪʔ # tʰɪ # tʰam]

It is now it is relatively easy to decode the utterance with lexical matching routines similar to those in Smith's thesis [9]. (See figure 2.) First, clusters (e.g., onsets, peaks, codas) are looked up in a syllabary, and then syllables are looked up in a lexicon of words.

4. Klatt analyzed the two /t/s in *Did you hit it to Tom* as a single geminated /t/. I prefer to analyze the region as two /t/s, one glottalized and one aspirated. My analysis is supported in part by presence of two peaks in total energy at 0.95 and 1.02 seconds (see spectrogram).
5. This formulation of the constraints is oversimplified for expository convenience; see [4, 6, 7] and references therein for discussion of the more subtle issues.
6. In [1], the constraints are formulated in terms of a phrase structure grammar so that the syllable boundaries can be inserted with a standard context-free parser (e.g., Earley algorithm [3]).

Fig. 2. Lexical Matching

parsed transcription		decoding
dɪʃə	→	did you
hɪt	→	hit
ɪ?	→	it
t^hɪ	→	to
t^ham	→	Tom

A demonstration parser has been implemented and tested on a set of linguistic transcriptions. The program performs as well as can be expected with these methods. The program finds the intended decoding and no others, except when there are errors in the lexicon, grammar, input transcription or when higher level constraints (e.g., syntax, semantics, pragmatics) are required.[7] The emphasis here has been on methodology, rather than bottom-line performance. To focus on numbers at this early stage in speech research seems premature, especially in light of the fact that no machine to date is as good as a human listener (or a spectrogram reader) at producing the input transcriptions.

4. Concluding Remarks

In the past, lexical retrieval has been viewed as a single step process. Instead of parsing the input transcription into allophonically and phonotactically well-formed substrings and passing just those substrings onto the lookup routines, previous systems have tended to pass all n(n-1)/2 substrings onto the lookup routines. My proposal is more efficient because it will discover that most of the substrings are ill-formed and need not be looked up in the lexicon.[8] In addition, it may be easier to predict what will happen with very large lexicons, because my approach depends more on fundamental grammaticality constraints than accidental gaps in the lexicon.

In conclusion, the parsing and matching approach depends on the hypothesis that a syllable-like constituent structure is an appropriate intermediate level of representation. Just as syntacticians have argued for the introduction of constituent structure on the grounds that noun phrases, verb phrases and sentences seem to capture crucial syntactic generalizations (e.g., question formation, wh-movement), so too, I might argue (along with certain phonologists such as Kahn [6]) for the introduction of syllable structure because syllables, onsets and codas capture important allophonic and phonotactic generalizations such as aspiration, tensing and laxing. If this constituency hypothesis is appropriate for the analysis of speech, then it seems natural to propose a syllable parser for processing speech, by analogy with sentence parsers that have become standard practice in the natural language community for processing text.

7 For example, given a transcription of *This is the CBS* ..., the system will produce the word lattice *This is the* (OR C *see sea)* (OR *B be)* 5 See the Appendix IV of [1] for some more sample output.

8 Of the n(n-1)/2 possible substrings, it can be shown that (in many cases) only 4n of them can be allophonically and phonotactically well formed [1 §6.4.1]. Reducing the search space in this way results in substantial savings, assuming that lexical lookups are more expensive than testing for well formedness. The validity of this assumption depends on the size of the lexicon and the machine architecture. It might be a very reasonable assumption, if, for example, the lexicon is very large and a lexical lookup is very likely to induce a paye fault.

References

1. Church, K., *Phrase-Structure Parsing: A Method for Taking Advantage of Allophonic Constraints,* Ph.D., MIT, 1983 (also to appear, LCS and RLE publications, MIT).
2. Cole, R., and Jakimik, J., *A Model of Speech Perception,* in R.Cole (ed.), *Perception and Production of Fluent Speech,* Lawrence Erlbaum, Hillsdale, N.J., 1980.
3. Earley, J., *An Efficient Context-Free Parsing Algorithm,* CACM, 13:2, February, 1970.
4. Fujimura, O., and Lovins, J., *Syllables as Concatcnative Phonetic Units,* Indiana University Linguistics Club, 1982.
5. Jelinek, F., course notes, MIT, 1982.
6. Kahn, D., *Syllable Based Generalizations in English Phonology,* Indiana University Linguistics Club, 1976.
7. Kiparsky, P., *Metrical Structure Assignments in Cyclic,* Linguistic Inquiry, 10, pp. 421-441, 1979.
8. Klatt, D., *Scriber and Lafs: Two New Approaches to Speech Analysis,* chapter 25 in W. Lea (ed.), *Trends in Speech Recognition,* Prentice-Hall, 1980.
9. Smith, A., *Word Hypothesization in the Hearsay-II Speech System,* Proc. IEEE Int. Conf. ASSP, pp. 549-552, 1976.