SEGMENTAL PHONOLOGY AND NON-SEGMENTAL PHONETICS

Patricia A. Keating
Phonetics Lab, Linguistics Department,
University of California, Los Angeles, USA

ABSTRACT

In this contribution I present the view that there is no fundamental problem in relating segmental, non-dynamic phonological representations to non-segmental, dynamic phonetic representations of speech, and that other kinds of theories of phonological representation are less suited to dealing with prosodically-conditioned variation.

INTRODUCTION

The question posed in this session is, "What benefits/problems flow from taking a dynamic/non-segmental approach to phonetics?". This question arises because most, though not all, phonologists have traditionally assumed that the segment is one basic level of phonological representation, and because most, though not all, phoneticians assume that there are no phonetic segments. Thus there is a mis-match between the two levels, which would seem to be undesirable. An apparent solution to this apparent problem has become increasingly popular: that the phonology should match the phonetics in being dynamic/non-segmental. In what follows I will present my own view on this issue: first, that there is no real problem crying out for a solution, and second, that a dynamic/non-segmental phonological representation creates new problems for phonetics, because it contains too much specific information.

In addition, I should note that I believe that there is good reason to assume that people have a tendency to construct a psychological representation in terms of segments. In short, I share the view that the widespread success of phonemic alphabets reflects (because it depends on) the ability of humans to readily construct segmental representations of words. (This is not to say that a person must have a complete and final segmental representation before learning to read, or that the orthography has no influence on the segmental representations.) I will not have time to defend that component of my position here, but it is certainly a motivating principle in what follows, because it means that some "benefit flows" from having at least a quasi-segmental phonological representation. Following Goldsmith [1], then, I see the language learner as being faced with the task of making structural sense of the speech signal by abstracting segments (and other higher-level units) from it.

"DYNAMIC" AND "NON-SEGMENTAL"

First, the terms under discussion, "dynamic" and "non-segmental", require some clarification, especially since they are not the same, and therefore might bring different benefits and problems into play. The question of whether representations are chunked into units is separate from the question of whether those units, or their components, are dynamic.

"Dynamic" seems the easier term of the two: if some component is dynamic, it is directly and inherently specified as time-varying. The phonological alternative is "static": no claim is made about how long a given property persists, how fast it comes on or turns off, etc. Static phonological representations (though non-segmental) are defended by e.g. Local [2].

"Segmental" is much the murkier term. (See Abercrombie [3] for an interesting historical review; see Pike [4] in particular for a full program of phonetic segmentation.) Phonetic segmentation, for example of spectrograms, usually refers to a strict division of the speech signal into discrete, non-overlapping, temporal slices which exhaustively parse the entire signal. Certainly this can be done up to a point: acoustic records of speech show a quasi-segmental character based on changes in the major manner features of the primary articulatory constriction. Thus conventions for acoustic measurement of "segment" durations are typically based on source characteristics

of the constriction (stop, fricative, approximant), rather than on voicing, nasalization, place of articulation (formant frequencies), or secondary constrictions. (This manner-based division of the signal is somewhat along the lines of McCarthy [5], where the segmental root node consists of the features Consonantal and Sonorant.) One non-segmental aspect of speech, then, is that other features of a segment do not have to line up with this basic manner division either grossly (e.g. nasalization may begin well before a nasal consonant) or in detail (e.g. voicing may continue for one or two pitch periods after the closure for a voiceless stop).

Another non-segmental aspect of speech is that when these manner features do not change, most notably in a sequence of resonants, no obvious segmentation emerges and our acoustic criteria are quite arbitrary. Put another way, acoustic signals may suggest some segmentation and perhaps support further segmentation, but not always corresponding to a phonological segmentation.

Hertz [6] takes an intermediate position on acoustic segmentation: phones are quasi-steady-state portions of the signal, while transitions are specific time intervals that come between phones. F2 is used as the primary basis for segmentation. Hertz and colleagues show that interesting phonetic generalizations can be made on the basis of this segmentation, e.g. that phones and transitions pattern differently in terms of durational changes.

Phonological segmentation is by no means the same thing as dividing up a spectrogram. The job of phonological segments is at least twofold: to indicate phonological precedence (which features come roughly at the same time vs. which are clearly in sequence), and to give a gross indication of notional time (a segment's worth of time). The same jobs are done by higher-level units too, of course; the segment is simply one level of such organization.

In fact, most phonological representations are what we might call "semi-segmental". They are basically segmental in that there are segment slots of some kind (root nodes, Xs, whatever) which do these jobs of segments, but they are non-segmental to the extent that features are autosegmentalized, that is, features can belong to no segment slots (as in floating morphological features), or to more than one (as in geminates), and more than one value of a feature can belong to a single segment (as in affricates).

Finally, note that for the purposes of thinking about the relation between phonological and phonetic representations, it doesn't matter whether segmental phonological representations are underlying (as most phonologists assume) or derived (e.g. Archangeli and Pulleyblank [7]).

SEGMENTAL AND NON-DYNAMIC PHONOLOGY: IT'S NOT A PROBLEM

The traditional class of models of the relation between phonology and phonetics is known as "target and interpolation" models. That is, these are models that provide targets and interpolations between targets. Individual phonological feature values associated with segments (or, in some speech synthesizers, unanalyzed whole segments) specify "targets" in articulatory and/or auditory-acoustic domains. The targets are aimed at by the speech producing mechanism, which moves, or "interpolates", from target to target. Examples of work in this framework include [6], [8], [9], [10], [11], [12], [13], [14], and [15].

In this approach, then, there are three steps in getting from a discrete phonological representation to a continuous phonetic representation.

The first step is a general one, and the other two are done for each utterance.

Step 1: relate each feature to one or more parameters (articulatory or auditory-acoustic, as relevant). To some extent this correspondence will be the same across languages: some one parameter is most basic for a given feature; but to some extent this correspondence will vary across languages, because other parameters may also be used, or not. These expressions of features can be quite complex, but that is the nature of speech, not of the theory itself (contra Zsiga [16]).

Every theory must grapple somewhere with the dual facts that articulatory-acoustic correspondences are complex and that different articulations can be used together to produce or enhance a given acoustic end. For example, the feature Strident needs to control parameters of tongue shape, jaw (tooth) position, glottal opening size, and velic opening size. Task Dynamics theory (e.g. [17]) does this in terms of Coordinative Structures (where the example of Strident is more complicated than ones usually discussed in that framework), and Enhancement theory (e.g. [18]) does this in terms of redundant feature specifications (for which again Strident is a complex example).

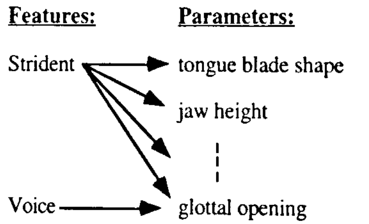

Figure 1. Features affecting multiple phonetic parameters.

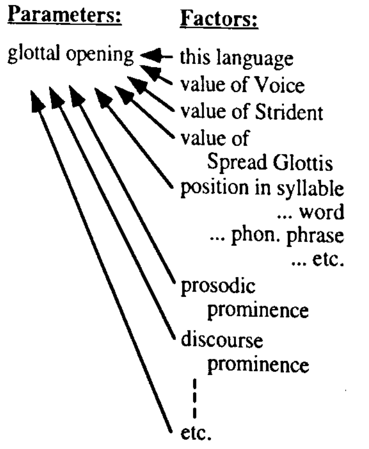

Figure 2. Multiple factors affecting the setting of a phonetic parameter.

Step 2: interpret any given (discrete) value of that feature as a (continuous) value in two dimensions: along the relevant parameters, for some interval of (notional) time. This continuous value will depend on many factors besides the phonological feature value, including values of other features in the same segment, and prosodic variables. For some parameters, more than one feature will determine the ultimate value. For example, Strident, which wants an open glottis for high airflow, competes with Voice, which wants approximated vocal cords for vibration, for control of glottal opening in a voiced strident, which makes voicing breathy, and/or difficult to sustain. Assignment of target values is called target evaluation by Pierrehumbert [8].

Step 3: connect successive values according to some mathematical function; this function may differ for different target values, parameters, languages, but is most commonly treated as linear.

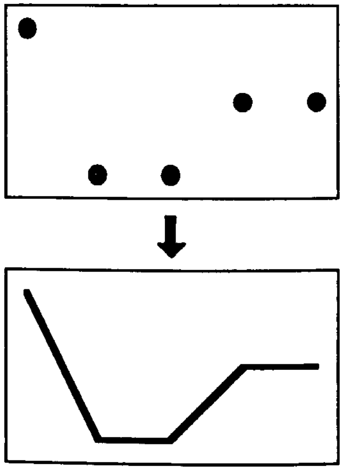

Figure 3. Interpolation between successive targets on some paramter

The evaluation for each feature is temporally independent in the sense that, for example, the different articulations for a single segment can begin and end at different times. That is, target and interpolation models require a theory of target alignment in the same way that Articulatory Phonology requires a theory of gestural phasing. The difference is that with a segmental phonology, these alignments are not considered to be lexically specified.

A property of both target and interpolation models and Articulatory Phonology is what I call phonetic underspecification. Phonetic underspecification means that not every segment has to have a specification, or target, for every feature. This has a strong effect on speech when interpolation functions do not care whether adjacent featural target specifications are from adjacent segments or not; targets are connected up through an empty time interval between them. This means that the effects of target specifications will extend further in time than the time interval occupied by the targets themselves. This is a way of getting dynamic effects without having the targets themselves be dynamic.

The diagnostic for phonetic underspecification, then, is variability across contexts. If there is no specification, then what you see will depend entirely on the surrounding specifications, which will trigger interpolation through the unspecified span in a temporally-gradient fashion.

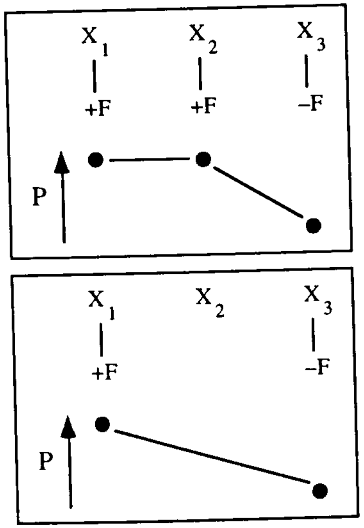

Figure 4. Interpolation between specified (top) and underspecified (bottom) values of some parameter.

On this account, then, much allophonic variation, especially variation that is coarticulatory or assimilatory in character, is generated by the quantitative operations (target evaluation and interpolation) just described. It can be seen that many allophones that have been previously described featurally (that is, by the change or acquisition of a feature value in a segment) are not described featurally here. As just one example, many vowel allophones that might be noted in a narrow phonetic transcription can be derived by interpolation, not by feature changes (e.g. Choi [15] on Marshallese). This does not mean that phonological feature spreading or changing cannot occur, but it certainly means that it does not occur as often as has been posited in the past. *The position that a segmental phonology means that every case of allophonic variation must involve segmental allophones is a straw man.*

Next, note that targets come from -- are projected by -- phonological feature values. Thus, when there is a phonological contrast, and therefore feature specification, there must be one or more phonetic targets (depending on how many parameters implement a feature). The targets will be the main influence on the parameter contour at that time. But if at some point in time there is no contrast that uses a given parameter, there will be no target at that time on that parameter, and the influence of context will obviously be strong. That is, contrast can restrict contextual variability while lack of contrast can give rise to contextual variability.

I developed this idea as the "window" model of surface phonetics. In this model, targets are not single values. Instead, they are ranges of permitted values. For articulation, you can think of these as constraints that say how much it matters how precise an articulation is. Some targets are very narrow ranges or windows; they permit little variation. Other targets are wide ranges or windows; they permit correspondingly more variation. In effect, windows turn phonetic underspecification from an all-or-none proposition to a gradient proposition.

With target ranges rather than points, interpolation becomes a more complicated function. For articulation, the general idea is to go as slow as possible while still making it into the required range. Guenther [19] has implemented a neural net model of articulation that incorporates a windows-like idea. Guenther's ideas differ from mine in certain respects but his model shows that interpolation between articulatory target ranges in accord with motor control principles is possible.

Interestingly, Guenther is developing his target ranges as an implementation of Lindblom's H&H dimension [20]. A small window is a kind of hyperarticulation because it requires more careful speech to reach the small target and it limits coarticulation. So the target sizes encompass all styles of speech, but slower speech, and more careful speech, would be modeled as a shrinking of the targets, while faster/less careful speech would use the fully expanded targets. He also proposes to follow up a result of deJong et al. [21] and deJong [22], that phrasal prominence results in a decrease in contextual variation and thus involves hyperarticulation. This can be modeled straightforwardly as a decrease in target range of the head of a prosodic domain (though his model as it stands needs some modification to generate more extreme articulations under hyperarticulation). Not only are vowel articulations hyperarticulated under stress, but onset consonants show hyperarticulation effects of stress on their oral [22] and glottal [23] gestures.

I would take this approach even farther. It seems likely that hyperarticulation characterizes not only prosodic heads, but also at least some prosodic edges. Consider how edges of words are treated in English (and plausibly other languages). Wordfinally, lengthening is the demarcative property (e.g. [24], [25]), specifically lengthening of the closing (VC) gesture of the final rime (e.g. [26]) rather than the final consonant constriction interval, which is in fact shorter [27]. Word-initially, on the other hand, the initial consonant's constriction interval is lengthened [27] and both oral and velic gestures are more constricted ([28], [29]--see review by Browman and Goldstein [30]) and glottal opening is greater (again, reviewed by [30]). Most strikingly, and more generally, the size of the glottal opening for /h/ and for aspiration gets bigger the larger the prosodic domain in which the consonant is initial ([23], [31]). Thus it is not only heads of prosodic domains that are hyperarticulated; initial edges are also hyperarticulated, even above the word level.

This means that prosodic structure, not only at the syllable and word levels, but also at different phrasal levels, probably plays an enormous role in determining what we think of as purely "segmental" characteristics, like degree of stricture, as well as what we think of as "suprasegmental" characteristics, like duration.

THE ALTERNATIVE HAS A PROBLEM

We have seen how a distinction can made between phonological feature values, which characterize physical properties only very abstractly, and phonetic targets, which have specific spatial and temporal quantitative values as a function of many factors, including but not limited to the phonological feature values. Suppose instead we want phonological representations to include much of this detail: some indication of how fast and how long a movement will be, a more explicit indication of its exact spatial goal or target, more information about the relative alignments of different movements -- say, a theory like Articulatory Phonology. How will all the prosodic variation discussed above be dealt with?

Browman and Goldstein [30] broach this issue. However, they do so by focusing on different kinds of variation at different levels. Variation in gestural parameter specifications and in phasing are both considered at the word level, where lexical stress, position in syllable, and position in words are all relevant variables. As long as such variation occurs within the word, it can be incorporated into the lexical representation itself, where it is redundant information, but useful in specifying how the word is to be actually pronounced. Phrasally, however, Browman and Goldstein consider differences in phasing only as they occur between words. These different phasings leave the lexical representations intact and simply slide them around relative to one another. Even American English flapping of final alveolars before vowel-initial words is said to involve no

change in the word-final alveolar gesture itself, presumably because if it did, that would be harder to account for.

Yet within-word variation in gesture parameters and possibly in phasing does occur as a function of postlexical structure, and in fact may be completely pervasive. In that case, we cannot say that lexical specification tells us how to pronounce a word, only how to pronounce it in some particular context. Which prosodic position should be taken as the basis for the lexical representations? Should some position be taken as canonical, and other variants derived from it by some kind of readjustment? This would go against the spirit of the whole endeavor, because the lexical information would be misleading just because it is precise. Or should a list of all possible alternatives be precompiled, along with indices so you select the right one for any given occasion? Not only does this again go against the spirit of the theory, but it requires that there be some finite number of possibilities. Or should gestures in lexical representations indicate only ranges of spatial and temporal variation, with more precise values to be determined postlexically? Or should lexical representations be segmental and non-dynamic, as they are in Zsiga's [16] version of Articulatory Phonology? In these last two cases, the segmental-and-nondynamic and non-segmental-and-dynamic theories will turn out to be much more alike than they now seem.

REFERENCES

[1] Goldsmith, J. (1976), *Autosegmental phonology*, PhD dissertation, MIT.

[2] Local, J. (1992), "Modeling assimilation in nonsegmental, rule-free synthesis", *Papers in Laboratory Phonology II* (eds. Docherty & Ladd):190-223.

[3] Abercrombie, D. (1991), *Fifty Years in Phonetics*, Edinburgh University Press.

[4] Pike, K. L. (1949), *Phonetics*, University of Michigan Press.

[5] McCarthy, J. (1988), "Feature geometry and dependency: a review", *Phonetica* 45:84-108.

[6] Hertz, S. R. (1991), "Streams, phones, and transitions: toward a new phonological and phonetic model of formant timing", J. Phon., 19(1):91-109.

[7] Archangeli, D. & D. Pulleyblank (1994), *Grounded Phonology*, MIT Press.

[8] Pierrehumbert, J. B. (1980), *The phonology and phonetics of English intonation*, Ph.D. dissertation, MIT.

[9] Pierrehumbert, J. & M. Beckman (1988), *Japanese tone structure*, Linguistic Inquiry Monograph, MIT Press.

[10] Clements, G. N. & S. R. Hertz (1991), "Nonlinear phonology and acoustic interpretation", *Proceedings of the XIIth ICPhS* , vol 1:364-373.

[11] Clements, G. N. & S. R. Hertz (1995), "An integrated representational system for phonology and acoustic phonetics, with a case study of English vocalic nuclei", MS.

[12] Huffman, M. (1990), *Implementation of Nasal: timing and articulatory landmarks, UCLA Working Papers in Phonetics* 75.

[13] Cohn, A. (1990), *Phonetic and Phonological Rules of Nasalization, UCLA Working Papers in Phonetics* 76.

[14] Cohn, A. (1993), "Nasalization in English: phonology or phonetics", *Phonology* 10:43-81.

[15] Choi, J. D. (1992), *Phonetic Underspecification and Target Interpolation: an acoustic study of Marshallese vowel allophony, UCLA Working Papers in Phonetics* 82.

[16] Zsiga, E. (1993), *Features, gestures, and the temporal aspects of phonological organization*, Ph.D. dissertation, Yale U.

[17] Saltzman, E. & K. G. Munhall (1989), "A dynamical approach to gestural patterning in speech production", *Ecological Psychology* 1:333-382.

[18] Stevens, K. & S. J. Keyser (1989), "Primary feature and their enhancement in consonants", *Language* 65:81-106.

[19] Guenther, F. H. (1994), Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psych. Rev.*, in press.

[20] Lindblom, B. (1990), Explaining phonetic variation: A sketch of the H & H theory, in W. J. Hardcastle and Alain Marchal (eds.) *Speech Production and*

Speech Modelling, Kluwer Academic Publishers, pp. 403-439.

[21] deJong, K., M. Beckman, and J. Edwards (1993), The interplay between prosodic structure and coarticulation, *Language and Speech* 36:197-212.

[22] deJong, K. (1995), "The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation", *JASA* 97(1):491-504.

[23] Pierrehumbert, J. & D. Talkin (1992), "Lenition of /h/ and glottal stop", *Papers in Laboratory Phonology II* (eds. Docherty & Ladd):90-116

[24] Oller, D. K. (1973), "The effect of position in utterance on speech segment duration in English", *JASA* 54:1235-1247.

[25] Wightman, C. W., S. Shattuck-Hufnagel, M. Ostendorf, & P. J. Price (1992), "Segmental durations in the viscinity of prosodic phrase boundaries", *JASA* 91:1707-1717.

[26] Beckman, M., J. Edwards, & J. Fletcher (1992), "Prosodic structure and tempo in a sonority model of articulatory dynamics", *Papers in Laboratory Phonology II*, (eds. Docherty & Ladd):69-86.

[27] Byrd, D. M. (1994), *Articulatorry timing in English consonant sequences, UCLA Working Papers in Phonetics* 86.

[28] Krakow, R. A. (1989), *The articulatory organization of syllables: a kinematic analysis of labial and velar guestures.* PhD dissertation, Yale U.

[29] Vaissiere, J. (1988), "Prediction of velum movement from phonological specifications", *Phonetica* 49:48-60.

[30] Browman, C. & L. Goldstein (1992), "Articulatory Phonology: an overview", *Phonetica* 49:155-180.

[31] Jun, S. (1993), *The phonetics and phonology of Korean Prosody*, PhD dissertation, Ohio State U.