# TIMING MODELS FOR PROSODY AND CROSS-WORD COARTICULATION IN CONNECTED SPEECH

Mary E. Beckman
Department of Linguistics
Ohio State University
Columbus, OH 43210

## ABSTRACT

Gauging durations of acoustic intervals is useful for recognizing the phrasing and stress pattern of an utterance. It aids in the recognition of segments that are differentiated by duration, and it can improve segment recognition in general because knowing the stress and phrasing reduces the vocabulary search space. However, models of speech timing that compute acoustic segment lengths cannot capture spectral dynamics, and they rapidly become unwieldy in connected speech, where many effects interact to determine interval durations. I will review two results from recent work on articulatory dynamics that suggest a more workable alternative. Browman and Goldstein have developed a general model of the timing of articulatory gestures. Using this model they can describe many assimilations and apparent deletions of segments at word boundaries in terms of simple manipulations of intergestural timing, an account which should be useful for predicting the lenition pattern and for interpreting the resulting spectra in order to recover the underlying form. Beckman, Edwards, and Fletcher have applied Browman and Goldstein's model in examining articulatory correlates of global tempo decrease, phrase-final position, and sentence accent. Their data show that these three different lengthening effects are functionally distinct and suggest that the kinematics of formant transitions and amplitude curves can be used for distinguishing among the effects to parse the prosodic organization of an utterance.

## INTRODUCTION

Variation in timing is one of the most pervasive features of speech. It plays a role at all levels. A particular pattern of vowel lengthening, for example, can cue the segmental contrast between [ae] and [E] in 'bad' versus 'bed' and between the following [d] and [t] in 'bad' versus 'bat' (e.g., Nooteboom 1973; Klatt 1976; Raphael 1972). In speech synthesis, manipulating the timing pattern by changing the lengths of acoustic segments can also alter the perceived stress pattern or intonational phrasing of an utterance (e.g., Fry 1958; Klatt 1979; Scott 1982). It is hardly surprising, therefore, that knowledge of segment durations can improve speech recognition. For example, Deng, Lennig, and Mermelstein (1989) have shown that information about vowel interval durations dramatically increases recognition rates in a Hidden Markov Model isolated-word recognition system. Similarly, Lieberman (1960) showed

that vowel-interval durations augmented by rudimentary RMS amplitude measures can identify stressed syllables. Using interval durations to parse the stress pattern in this way can drastically reduce the search space in large-vocabulary isolated-word recognition systems (Waibel 1988). Knowing the stress pattern should prove even more crucial to recognition of connected utterances, because of the way that stress interacts phonologically with the phrasing to cue the prosodic organization of the utterance into words and larger phonological units (Nespor and Vogel 1986; Beckman, de Jong, and Edwards 1987). An accurate prediction of assimilations, deletions, and other lenition rules across word boundaries also depends on the phonologoical phrasing (Nespor and Vogel 1982; Zek and Inkelas 1987).

If knowledge just of acoustic interval durations can aid recognition in both isolated words and connected speech, what if we were to use finer measures of timing? There are many indications that knowledge of the temporal structure within acoustic segments could improve recognition even more. For example, in addition to being longer and having a lower first formant, [i] (as in 'beat') differs from [I] (as in 'bit') in having a faster, shorter second formant transition that starts later in the syllable (Neary and Assman 1986). Other tense-lax vowel pairs also show this difference in spectral kinematics. Similarly, in addition to being shorter in overall duration before a word-final voiceless obstruent, vowels tend to have shorter, faster first-formant transitions (Summers 1987). A better understanding of the control of such timing patterns in speech production could lead to more accurate accounts of the kinematic differences and to more wieldy predictions of interactions among the many factors that influence segment-interval duration.

In the last decade, we have made tremendous advances toward a better understanding of timing control by looking in detail at the kinematics of the articulatory gestures involved in producing speech. Following a proposal by Fowler et al. (1980), speech scientists have worked at applying a general model of motor control orginally developed to account for such things as the coordination of flexor and extensor muscles in maintaining gait across different terrains and speeds or the coordination of shoulder and elbow joints in different reaching tasks (e.g. Ostry, Keller, and Parush 1983; Kelso et al. 1985; Saltzman 1986).

Two recent results of this work seem particularly relevant to achieving better recognition models. One is Browman and Goldstein's (1987) application of their task-dynamic model to explain many common lenitions across word boundaries in casual or fast speech. The other is Beckman, Edwards, and Fletcher's (1989) application of the model in understanding the control of three different lengthening effects associated with slow tempo, phrase-final position, and nuclear sentence stress. In the next two sections, I will describe these two results and their implications for speech recognition in more detail.

## CROSS-WORD LENITIONS AND THE GESTURAL SCORE

One of the biggest problems in recognizing connected speech is coarticulation across word boundaries. This coarticulation can cause a drastic restructuring of the spectral characteristics of segments at the edges of words. Final segments can change by assimilation to the following word's initial segment, and they can even be seemingly deleted, as shown in the examples in (1), which are taken from Brown (1977) and Catford (1977).

(1)* a. assimilations

| | | | |
|---|---|---|---|
| /dhIs shap/ | -> | [dhIshshap] | 'this shop' |
| /huhndruhd paUndz/ | -> | [huhndruhbpaUndz] | 'hundred pounds' |

b. deletion

| | | | |
|---|---|---|---|
| /muhst bi/ | -> | [muhsbi] | 'must be' |

c. deletion and assimlation

| | | | |
|---|---|---|---|
| /graUnd prEshR/ | -> | [graUmprEshR] | 'ground pressure' |

Such lenitions are ubiquitous in casual or fast speech and are not uncommon even in fluent read speech. They can occur within the word as well as at word boundaries, as in the assimilative devoicing or deletion of the first vowel in [ptEIto] for 'potato' or the apparent deletion of the medial [t] in [twEni] for 'twenty'.

In these examples, we have described the lenitions as if they were discrete changes in the symbolic representation of the segment string. If the lenitions are approximated by an allophonic analysis in this way, the word-internal cases could be accounted for in isolated-word recognition systems by encoding all common patterns as variant pronunciations in the lexicon. This could be accomplished, for example, by providing separate spectral templates

---

* Here and elsewhere, I use the following ARPABET-like substitutions for the standard phonetic symbols:

[I] = high front lax vowel
[E] = mid front lax vowel
[ae] = low front vowel
[U] = high back lax vowel
[O] = low-mid back lax vowel
[uh] = mid-central or reduced vowel ("carrot" or schwa)
[R] = rhotacized mid-central vowel (i.e., syllabic [r])
[sh] = voiceless alveopalatal stop
[zh] = voiced alveopalatal stop
[dh] = voiced interdental fricative
[th] = voiceless interdental fricative
[D] = flap

for each variant pronunciation or by listing alternate paths in an allophonic-segment-based HHM model (Kopec and Bush 1985). Lenitions across word boundaries in connected speech can also be handled by pre-compiling alternate HHM paths for every possible transition (Bush and Kopec 1987), but this is feasible only when the vocabulary size is very small. Thus, cross-word segment lenitions cause a particular problem for large-vocabulary recognition systems even when explicit phonetic knowledge is incorporated in the form of allophonic variants for acoustic segments.

A possible solution is to base the lexical representation of the allophones not on alternate paths through discrete phonologically unanalyzed acoustic intervals, but rather on alternate specifications of acoustic features in a feature-based recognition system (Stevens 1986). The assimilation of [s] to [sh] in 'this shop' could then be handled by an explicit assimilation rule that changes the acoustic features associated with the [s] segment from [+anterior] to [-anterior] in the context of the following [-anterior] segment in the following word. The apparent deletion of the [t] in 'must be', similarly, could be handled by a rule deleting the features associated with [t] stop release in the context of a following obstruent segment. If this solution is adopted, the problem reduces to that of discovering the correct assimilation and deletion rules and the optimal acoustic feature system for stating these rules.

A disadvantage of this approach is that these coarticulatory assimilations and deletions look like a motley array of discrete rules when described in terms of feature changes and deletions. Among the ways that models of articulatory kinematics might contribute to speech recognition is in providing a more explanatory account of these cross-word lenitions, an account that better predicts the patterns of assimilation and apparent deletion that are likely to occur in any given context. Browman and Goldstein (1987) have suggested an account of common lenition patterns that unifies assimilations and deletions into a single process.

The basis for Browman and Goldstein's account is the gestural score. Browman and Goldstein, in conjuction with Saltzman and other colleagues at Haskins Laboratories, have developed a task-dynamic model in which utterances are represented as a principled orchestration of invariant articulatory gestures. The gestures are modeled as target-specific movements in a second-order linear spring-mass system. The orchestration specifies a given phasing for a gesture relative to the relevant surrounding gestures. The [t] of 'must be', for example, is represented as an overdamped gesture of a given stiffness and underlying amplitude specified for the task of making a complete closure with the tongue tip near the alveolar ridge. This alveolar closing gesture is specified as concurrent with either a ballistic abductive glottal gesture or a totally adductive glottal stop gesture, and as occuring at some time relative to the opening gesture from the word-initial [m] into the [uh] vowel. The [b], similarly is composed of a labial closing gesture coupled to a glottal approximation gesture, with the two gestures specified to occur at some time relative to the oral and glottal gestures of the preceding [t].

Under this account, the apparent deletion of the [t] can be modeled as the endpoint of a continuum of lesser to greater overlap between the tongue-tip gesture in the [t] and the labial gesture in the [b]. If the two gestures overlap to any extent, the release of [t] tongue-tip closure will be masked by the [b] labial closure. That is, the usual aerodynamic consequences of the [t] release -- namely, the burst, will be prevented by the closure upstream. In extreme cases, not just the release of the [t] but the entire tongue-tip gesture can be hidden by the labial gesture, as Browman and Goldstein have shown in their examination of the movements of the tongue tip and lower lip and other movement traces recorded at the Tokyo X-ray microbeam system (Kiritani et al. 1975). Nolan (1989) shows similar cases of overlap between dental and velar gestures as evident in patterns of contact measured by an electro-palatograph. In sequences such as 'late calls', the tongue-tip contact for the word-final [t] can overlap to a greater or lesser extent with the tongue-body contact for the following word-initial [k].

In Browman and Goldstein's task-dynamic model, assimilations such as the apparent substitution of [sh] for [s] in 'this shop', can also be specified as overlap. The two tongue-tip constriction gestures for the fricatives overlap in time in the same way as the [t] and [b] of 'must be'. In this case, however, the overlap involves the same vocal tract subsystem. Therefore, the kinematic consequence of the overlap is not a "hiding" of one gesture by the other, but a spatio-temporal "blending" of the two gestures, resulting in an uninterrupted [sh]-like spectral pattern.

Thus, examination of the articulatory patterns provides a single explanatory account of the motley array of cross-word lenition patterns. Both the apparent segment deletions and the feature assimilations can be described by a common articulatory mechanism. It seems likely that the same mechanism also will account for various sorts of manner lenitions, such as the flapping of [t] and [d] and stop consonants being produced as fricatives. In the gestural score, these will probably be represented as undershoot of the temporal or spatial target for the consonant when the consonant's closing gesture is blended with the opening gesture for the following vowel. That is, flapping and frication are probably simply two more examples of gestural overlap.

One advantage of this account is that the continuous phase settings of the gestural score correctly predict that there will be varying degrees of overlap, resulting in varying degrees of spectral masking by the following segment, unlike in the all-or-none segment deletion and assimilative feature-changing accounts. Since human listeners apparently can use the residual spectral information of the preceding vowel-formant transition to perceive the different between a deleted [t] in 'late calls' and no [t] in 'lake calls' (Nolan 1989), this is a desirable outcome. In a recognition system based on all-or-none feature changes, by contrast, near minimal pairs such as these can only be distinguished if there is disambiguating syntactic or semantic information in the context.

Finally, the gestural score account makes all types of segmental lenition fall out from manipulations of the timing pattern, and when combined with a model

of the articulatory correlates of tempo change and prosodic structure, should provide a better prediction of when lenitions will occur. That is, lenitions should occur more frequently at tempi and in prosodic contexts where articulatory gestures are phased more closely together.

## THE KINEMATICS OF TEMPO, PHRASING, AND ACCENT

While Browman and Goldstein have not yet provided an account of articulatory correlates of prosodic structure within their task-dynamic model, there is other recent work that suggests how several effects can be described using the gestural score. Such a description is obviously important, for many reasons. A first obvious reason is that the cross-word assimilations and deletions discussed in the preceding section are blocked by certain sorts of prosodic phrase boundaries. For example, the word-final [s] in 'this' would not assimilate to the following [sh] in any typical intonational phrasing for 'So the question is this: should we do it or not?'

An even more general reason for wanting a better description of the articulatory correlates of prosodic structure is that stress and phrasing interact with segmental duration patterns in ways that are very difficult to capture in computational models of acoustic interval durations (see, e.g., van Santen and Olive 1989; Riley 1989). Yet human perceivers clearly use the timing patterns of an utterance to parse the segments, stress pattern, prosodic structure, and overall tempo. It seems unlikely that in doing so, they perform the complicated computations that interval-based models use to predict the segment interval durations. A better model of speech timing could provide evidence as to what is actually being perceived when the timing patterns of an utterance are parsed to provide the perceptual cues to segmental and suprasegmental structures.

Work by Beckman, Edwards, and Fletcher (1989) suggests that articulatory kinematics can differentiate global tempo change from phrase-final lengthening, and both of these from the lengthening effect of accent or stress. We looked at the durations, displacements, and peak velocities for opening-gestures and closing gestures in the sentence-intial [pap] sequences in the sentences in (2):

(2) a. Pop, opposing the question strongly, refused to answer it.
b. Poppa, posing the question loudly, refused to answer it.
c. Poppa posed the question loudly, and then refused to answer it.

The underlining in (2) indicates the test sequences. In (2a), the sequence is final to an intonation phrase, whereas in (2b) it is not final. The sequence in (2b), in turn contrasts to the sequence in (2c) in bearing the nuclear accent in its phrase.

We had several speakers repeat these utterances at three self-selected speaking rates, and measured the kinematics of the jaw-opening and closing gestures into and out of the low vowel [a]. We found that slowing down tempo

overall works essentially by changing the stiffness of the articulatory system. Both the opening gestures and the closing gestures have smaller peak velocities at slower tempi, with essentially no change in displacement. Phrase-final lengthening looks like slowing down tempo, but localized to the closing gesture. The lengthening associated with accent, by contrast, did not significantly change the speed of either gesture. Instead it seemed that the accented vowel was longer because the closing gesture was later relative to the opening gesture. In terms of Browman and Goldstein's gestural score, accentual lengthening is a phase shift that lessens the overlap between the vowel gesture and the following [p] gesture.

This last result confirms the findings of Summers (1987), who compared the articulatory kinematics of accentual lengthening with the effects of voicing in a following final stop. The duration and velocity patterns he found for accent are similar to those in our experiment, whereas the effect of voicing was more similar to those of our final lengthening; the closing gesture out of the vowel was slower before a voiced stop. Voicing differed from final lengthening in affecting displacement slightly as well as velocity; the jaw did not open as far before the voiced stop.

This work has implications for the ways in which acoustic timing patterns can be used to recognize stress and prosodic phrasing. Other things being equal, jaw opening is correlated with first formant frequency and overall amplitude. Low vowels, with more open jaw positions, have higher first formants and greater amplitudes than high vowels, with less open jaw position. In keeping with these correlation, Summers (1987) found that the first formant was lower in [a] and [ae] before [b], as expected from the lesser jaw opening there. In a later perception experiment involving syllables synthesized to mimic the first formant patterns in his production experiment, he found that first formant frequency and transition speed could cue the difference between a following voiced versus voiceless stop.

Given our results concerning accent and final lengthening, then, we would expect that final lengthening should effect longer, slower first-formant transitions, whereas accent should not. Accent, on the other hand, should be associated with a greater average volume over the syllable nucleus, whereas final lengthening should result in gradually decreasing amplitude after an early loudness peak. We are testing these predictions in experiments presently underway. If they are borne out, then tracking formant kinematics and amplitude contours over a syllable should help interpret its overall duration pattern. A recognition system that incorporated these results would have much better recognition of the stress and phrasing pattern, with all the improvements in segmental recognition which that entails.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. and IRI-8902142. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and

do not necessarily reflect the views of the National Science Foundation or of the other co-PI's on Grant No. IRI-8902142. The discussion of task-dynamic models represented in this paper benefited by conversations with Catherine Browman, Louis Goldstein, and Elliot Saltzman. (However, the author alone is responsible for any errors in the understanding of their work.) The work on kinematic correlates of tempo, final lengthening, and accent was done in collaboration with Jan Edwards and Janet Fletcher and was supported by the NSF under Grants No. IRI-861752 and IRI-8858109 to Mary Beckman and Grant No. IRI-8617873 to Jan Edwards.

References

Beckman, M., De Jong, K., and Edwards, J. (1987). The surface phonology of stress clash in English. Paper presented at the 62nd Annual Meeting of the Linguistic Society of America, San Francisco, 27-30 December.

Beckman, M., Edwards, J., and Fletcher, J. (1989). Prosodic structure and tempo in a sonority model of articulatory dynamics. Paper presented at the Second Conference on Laboratory Phonology, University of Edinburgh, 30 June-4 July, 1989.

Browman, C., and Goldstein, L. (1987). Tiers in articulatory phonology, with some implications for casual speech. Paper presented at the First Conference in Laboratory Phonology. [Written version to appear in J. Kingston and M. Beckman, eds., (1990) Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech. Cambridge: Cambridge University Press.]

Brown, G. (1977). Listening to Spoken English. London: Longman.

Bush, M.A., and Kopec, G.E. (1987). Network-based connected digit recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 35.

Catford, J.C. (1977). Fundamental Problems in Phonetics. Bloomington, IN: Indiana University Press.

Fowler, C.A., Rubin, P., Remez, R.E., and Turvey, M.T. (1980). Implications for speech production of a skilled theory of action. In B. Butterworth, ed., Language Production I. London: Academic Press.

Fry, D.B. (1958). Experiments in the perception of stress. Language and Speech 1, 126-152.

Kelso, J.A.S., Vatikiotis-Bateson, E., Saltzman, E.L., and Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. Journal of the Acoustical Society of America 77, 266-280.

Klatt, D.H. (1979). Synthesis by rule of segmental durations in English sentences. In B. Lindblom and S. Öhman, eds., Frontiers of Speech Communication Research, 287-400. Academic Press.

Klatt, D.H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the Acoustical Society of America 59, 1208-1221.

Kiritani, S., Itoh, K., and Fujimura, O. (1975). Tongue-pellet tracking by a computer-controlled X-ray microbeam system. Journal of the Acoustical Society of America 57, 1516-1520.

Kopec, G.E., and Bush, M.A. (1985). Network-based isolated digit recognition using vector quantization. IEEE transactions on Acoustics Speech and Signal Processing 33, 850-867.

Neary, T., and Assman, P. (1986). Modeling the role of inherent spectral change in vowel identification. Journal of the Acoustical Society of America 80, 1297-1308.

Nespor, M., and Vogel, I. (1982). Prosodic domains of external sandhi rules, In H. van der Hulst and N. Smith, eds., The Structure of Phonological Representations, Part I. Dordrecht: Foris.

Nespor, M., and Vogel, I. (1986). Prosodic Phonology. Foris.

Nolan, F. (1989). The descriptive role of segments: Evidence from assimilation. Paper presented at the Second Conference on Laboratory Phonology, University of Edinburgh, 30 June-4 July, 1989.

Nooteboom, S.G. (1973). The perceptual reality of some prosodic durations. Journal of Phonetics 1, 25-45.

Ostry, D.J., Keller, E., and Parush, A. (1983). Similarities in the control of speech articulators and the limbs: Kinematics of tongue dorsum movement in speech. Journal of Experimental Psychology: Human Perception and Performance 9, 622-636

Raphael, L.J. (1972). Preceding vowel duration as a cue to the voicing characteristics of word-final consonants in English. Journal of the Acoustical Society of America 51, 1296-1303.

Riley, M.D. (1989). Statistical tree-based modeling of phonetic segment durations. Journal of the Acoustical Society of America, 85, Suppl. 1, S44.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: a preliminary model. In H. Heuer and C. Fromm, eds., Generation and Modulation of Action Patterns (Experimental Brain Research Series 15), 129-144. New York: Springer-Verlag.

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary. Journal of the Acoustical Society of America 71, 996-1007.

Stevens, K.N. (1986). Models of speech recognition II: a feature-based model of speech recognition. In P. Mermelstein, ed., Proceedings of the Montreal Satellite Symposium on Speech Recognition (Twelth Internation Congress of Acoustics), 67-68.

Summers, W.V. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. Journal of the Acoustical Society of America 82, 847-863.

Summers, W.V. (1988). F1 structure provides information for final-consonant voicing. Journal of the Acoustical Society of America 84, 485-492.

van Santen, J.P.H. (1989). Diagnostic tests of segmental duration models. Journal of the Acoustical Society of America, 85, Suppl. 1, S43.

Waibel, A. (1988). Prosody and Speech Recognition. Morgan Kaufmann.

Zek, D., and Inkelas, S. (1987). Phonological phrasing and the reduction of function words. Paper presented at the 62nd Annual Meeting of the Linguistic Society of America, San Francisco, 27-30 December.