THE ENGLISH VOICING CONTRAST AS VELOCITY PERTURBATION

Robert F. Port and Fred Cummins

Department of Linguistics
Department of Computer Science
Indiana University
Bloomington, Indiana 47405
port@cs.indiana.edu

Abstract

How do the discrete phonological units of the lexicon map onto continuous-time articulatory gestures and continuous-time auditory signals? The distinctive feature of [*voice*] in syllable-coda position in English raises these questions with a vengeance. For minimal pairs like *buzz/bus, clamber/clamper, tens/tense*, etc, most measurable time intervals associated with the first syllable of these words are affected by the value of [*voice*]. Several of the rules in the traditional standard phonologies of English and many so-called 'phonetic implementation rules' serve to account for the various large and small temporal effects associated with the feature. We show that a very simple model for the English voicing contrast can be proposed that may account for these effects only if this phonological feature is phonetically defined as a *velocity perturbation of a periodic dynamical system* for English syllables. We summarize some evidence for the generalization that localized speaking-rate changes characterize a change in voicing. Then we suggest a general mathematical form for this dynamic effect that requires only a few parameters. This model implements the voicing feature as a perturbing forcing function for an underlying syllable oscillator.

1 The English Voicing Contrast

This paper explores some influences of English phonology on the temporal microstructure of speech. In English, when the [*voice*] value of a consonant or cluster in syllable-final position is changed, not only is there a difference in the presence of glottal pulsing, but the temporal pattern of the syllable is altered in complex ways. For example, the linguistically minimal change from *clamber* to *clamper* affects the durations of many of the segmental parts of the word. The phonological feature difference in this pair also differentiates other English pairs like *fuss-fuzz, rapid-rabid, lunch-lunge*, etc.[1] In particular, the goal of this paper is to propose a model to obtain the right kind of effects on speech timing in speech production. At the same time, however, we expect that the dynamic model we propose will also be an essential component of a model for auditory perception of speech by English-speaking listeners.

1.1 Temporal Microstructure of Voicing

To illustrate the problem, we show in Figure 1 some of the segmental durational effects of a change in voicing in two syllable types studied by Fourakis and Port [5]. The two voicing pairs are *dense-dens* and *dents-dends* (where the last word was described to subjects as a form of the hypothetical verb **dend* rhyming with *tend*) and were embedded in a carrier sentence read by 4 speakers. The display shows the mean durations (n=24) for each measurable segment after the initial stop, plotted cumulatively: the vowel, nasal, stop closure and final fricative. It can be seen that the voicing change from [−*voice*] to [+*voice*] has the effect of lengthening both the vowel and nasal portions of each syllable and shortening the stop and fricative parts of the syllable (significantly, of course).[2] The basic results here are well-known in the

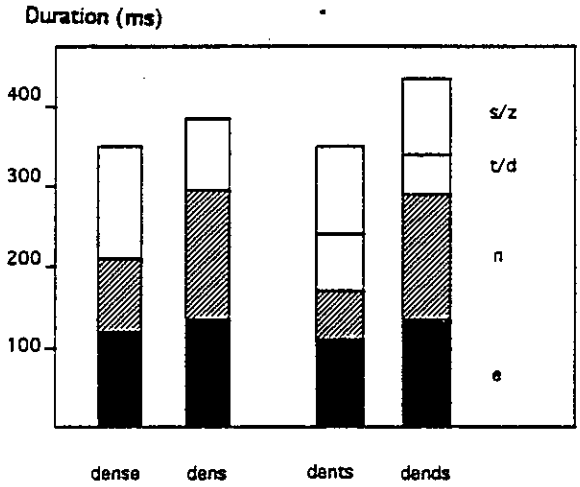

Figure 1: Cumulative segment durations for the minimal pairs, *dense-dens* and *dents-dends*. Data are means across 24 productions of each word by four (Anglophone) speakers of South African English [5].

literature [3, 13, 19]. Not only are these effects observed in production, but they also play a role in speech perception [21, 14, 17], indicating that listeners are able to exploit such temporal detail for making perceptual judgments – at least, they can if they are native speakers of English (cf. [16, 4, 18]).

The standard way to account for these effects in production is to propose *temporal implementation rules* [12, 19, 9]. These are rules that would specify how much to lengthen or shorten each phonetic segment that appears in each possible syllable type in English for each linguistic and pragmatic feature that has an influence on segmental duration. Thus, a hypothetical durational implementation rule for the [*voice*] contrast would require a different form for each possible syllable type (eg, CVC, CVNC, CVCC, etc) and then appropriate coefficients for each vowel and consonant that could appear in these segment slots – since so many segment durations are affected by other segments within the same syllable. And the rules would still presumably have to be stated in terms that are abstract enough to be applicable across a range

[1] Of course, [*voice*] also distinguishes syllable-initial stops and fricatives by means different phonetic cues. Our concern here is only with syllable-final position.

[2] These data are for speakers of South African English. American English is the same except that in *dense*, Americans tend to epenthesize a short stop before the [s] in *dense*, making it more similar to (but not identical with) *dents*. See [5].

of speaking rates. Thus, the first difficulty with this concept of temporal implementation rules is that there are a vast number of syllables. So the number of temporal implementation rules threatens to approach the number of possible syllables - even ignoring rate and stress effects. Secondly, such rules are merely specifications or injunctions -- to be 'obeyed' by some other system. They do not suggest how the actual durations could be achieved by the speaker [10, 6]. The dynamic model proposed below is a model for achievement of appropriately timed speech gestures. Finally, there are data that show that the voicing effect is actually continuous and affects all portions of the preceding vowel. This provides additional evidence against discrete segment-based temporal implementation rules [23].

1.2 General Form of the Model

The idea we explore in this essay is a means of accounting for the durational effects differentiating $-$[*voice*]-final from [$+$*voice*]-final syllables in a new way. Working within a dynamical-systems model for the speech production process [11, 2], we suggest that the temporal effects of the voicing contrast are due to *perturbation of the rate of motion* of an underlying syllabic oscillator. To implement this, one of the syllable types, the [$-$*voice*] syllable, is chosen as basic[3] and modeled as a simple harmonic oscillator. Then a perturbation function having two terms is proposed that causes a phase-dependent continuous distortion of the rate of motion of this oscillator. One term decelerates the rate of articulation of all speech gestures near mid-phase in the syllable, and the second accelerates the rate of articulation of any obstruent (i.e., stop or fricative) segments that occur late in the syllable. The deceleration would result in a lengthening of vowel or resonant segments near the middle of the syllable and the acceleration would cause shortening of any final obstruents. This perturbation function is presumably called for by the feature [$+$*voice*] in the syllable coda of lexical entries.

If a mathematical description of such dynamics could be obtained, it might provide an economical description for all syllable types and segmental components that is independent of speaking rate. Conceivably, the variable effects of the voicing change on specific segments in the syllable could be accounted for by their dynamic dependence on phase angle and by changes in the parameters of the perturbation function.

1.3 Toward a Dynamic Interpretation

The first step toward such a model is to clarify what the data suggest about the proposed perturbation. Continuous mappings from [$-$*voice*] to [$+$*voice*], such as might be produced by dynamic time warping of one onto the other, are hard to find in the literature (but see [15] for application of dynamic time-warping for a similar purpose), although mappings based on the durations of acoustically salient segments exist in many forms. In Figure 2, we have replotted old data [19, 5, 7, 24] for several syllable types under different conditions of word length, stress and dialect in a form that shows the amount of lengthening or shortening required on each segment to generate a [$+$*voice*] syllable from the corresponding [$-$*voice*] syllable.

It can be seen that there is lengthening of the vocalic and resonant portions of the syllable (implying deceleration) and a shortening of the obstruent portions (implying acceleration). A critical

[3]We somewhat arbitrarily treat the [$-$*voice*] syllable as basic, and we model its temporal pattern with a simple harmonic oscillator. Thus, the longer vowels before voiced consonants or in syllable-final position are achieved by our dynamic version of a 'lengthening rule'. Of course, the model would be essentially the same if the longer vowel were taken as basic.

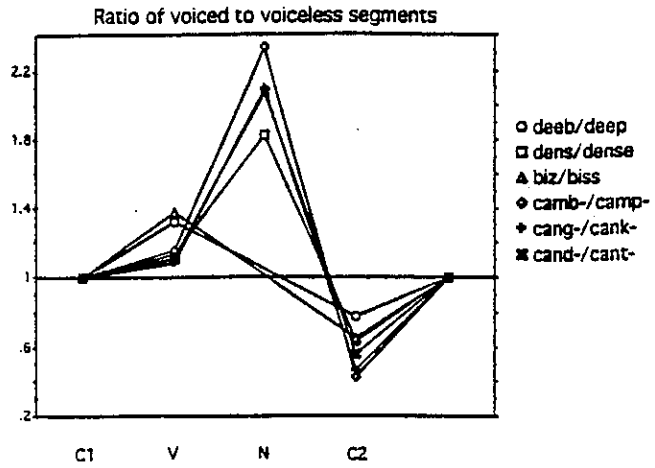

Figure 2: Proportion of lengthening and shortening required to transform each segment of selected syllables from voiceless to voiced. Thus, on the y-axis, 1 means no change in duration. The x-axis shows only the ordinal position of each segment in the syllable. Data are from [5] *dens/dense* by South African English speakers; [19] *deeb/deep* (Expt. 1); [7] *biz/biss* (Expt. 1, Phrase Internal); [24] *camber/camper, candor/canter.*

property is the position at which lengthening is replaced by shortening. It appears that the deceleration function has its greatest value close to this point. Thus, for example, the nasal in *camber-camper* tends to lengthen more than the vowel that precedes it does. Other data suggest that even the preceding voice-onset time (VOT) in syllables with initial aspirated stops is lengthened a small amount by the effect of [$+$*voice*] on the final consonant cluster [20]. Syllable-initial consonant closures appear to be totally unaffected by syllable-final voicing [19]. Altogether these results suggest that a schematized continuous rate-perturbation function for voicing in CVC syllables involves gradually decreasing articulatory rate until the phase angle for a final obstruent onset approaches, at which point articulation is accelerated.

In the next section, we present a sketch of a dynamic model for speech production which will provide a suitable framework for presentation of our own model for syllable-final voicing.

2 Dynamical Model of Speech Production

In a series of papers, Kelso, Saltzman, Browman and others [11, 22, 1, 2] have put forward a framework that attempts to reconcile the linguistic hypothesis that speech consists of a sequence of discrete, context-independent (phoneme-like) units with empirical observation of continuous, context-dependent overlapped articulatory movements. They propose a model having several hierarchically ordered, dynamic systems. At the highest level, there are partially ordered *speech gestures* that correspond very roughly to phonological autosegmental features and segments in the lexicon. These can be organized into a 'gestural score' [2] for a word or any stretch of speech with rows representing independently specifiable articulatory subsystems or 'tiers', and the horizontal axis representing the time dimension in some form. In variants of the gestural score, time can be (1) the sequential order of segments, (2) phase angle within a syllable cycle, or (3) absolute milliseconds, depending on one's purpose and the data to be represented.

Each gesture in this score exerts partial control over a set of *tract variables* (e.g., lip aperture, lip protrusion and velar low-

ering) which specify articulatory goals. Each goal is modeled as a point attractor in a low-dimensional space. The trajectories toward these goals are then given by damped second-order linear differential equations, incorporating inertial and stiffness coefficients. These coordinates are then transformed into the higher dimensional *articulator* space, where they specify the position, velocity and acceleration of the various articulators. Thus the general strategy is that higher level coordinative systems specify goals which are initially expressed in a low dimensional task space. These goal specifications are then transformed, via a one-to-many mapping, into the much higher dimensional space of the effectors.

For example, the tract variable goal for the bilabial gesture in [b] is given in context-independent terms of lip aperture in the gestural score. The associated context-dependent articulatory task variables are the yoked horizontal movements of the upper and lower lips, jaw angle, and independent vertical motions of the upper and lower lips. As the actual articulators have many more degrees of freedom than the more abstract tract variables, no one-to-one mapping from tract-variable coordinates to articulator coordinates exists. Rather, the dynamics are specified at tract-variable level and are transformed into articulator space with the aid of a suitable one-to-many transformation (since a tract variable goal can be reached in many ways in articulator space). (Need for this flexibility is demonstrated by experiments in which an articulator, like the jaw, is physically perturbed during speech, yet comes to adopt a slightly altered position nonetheless satisfying the speaker's requirements [11].)

Within the context of the model in [2], we might propose that the gestural score be supplemented with a new tier, the *Temporal Perturbation Tier*. The [*voice*] feature of syllable-final consonants inserts a symbol-like gesture (similar to those represented with Greek letters in [2]) onto this tier. The gesture may be just a deceleration (where there is no final consonant), or deceleration followed by acceleration (in the case where final voiced obstruents occur). These 'gestures' are different from the other gestures proposed since they do not produce distinct articulatory 'events'. They only change the local, intrasyllabic rate-of-articulation of all the other gestures. Still, this seems the simplest way to achieve the effect we need and yet retain the link to the phonological specification.

2.1 The Syllable as a Harmonic Oscillator

We propose to model the syllable as an abstract, coordinative timing unit – specifically, as a harmonic oscillator: $p'' + \omega^2 p = 0$. This oscillator should be viewed, for the moment, as a clock with unspecified angular rate whose phase angles can be used to specify critical time-points for the gestures associated with specific segments in the syllable. As this homogeneous system is autonomous, or time independent, the dynamics of the speech control system as a whole then depend on a single parameter of speaking rate, without recourse to any absolute-time clocking mechanism. This corresponds to our intuition that we may continuously vary our rate of speech, but, over a certain range of speaking rates, all further coordination takes care of itself.

By regarding coarticulatory phenomena as arising from intergestural competition for control of the articulators, the present model predicts that most coarticulation will be found within the domain of a single syllable. It also offers an answer to the question of why the syllable, which is required as an explanatory construct by so many phonological processes, does not have a clearly identifiable realization in the acoustic signal. If, as proposed, the syllable is a higher-level coordinative timing structure that is removed from actual articulator motion by a number of one-to-many coordinate transformations, the original timing information (phase angle) may only be present in very indirect form, as temporal patterning across the more directly observable segments.

2.2 Perturbation Model for [*Voice*]

We have so far only considered the underlying homogeneous system $p'' + \omega^2 p = 0$ which describes the abstract velocity profile and serves as a time scale for all the gestural tiers of the syllable. If we now perturb this system by the addition of a forcing function, $f(\theta)$, dependent upon phase angle, we can continuously alter the velocity of the gestures (cf. [8]). Taking the [$-$*voice*] alternate as basic and [$+$*voice*] to be perturbed, we select a perturbation function that provides the observed deceleration during those segments of the syllable nucleus which are lengthened, and acceleration in the shortened final obstruents. We therefore propose this preliminary and rather general version of a forcing function for the oscillator: $p'' + \omega^2 p = -ae^{-b(\theta-\delta)^2} + ce^{-d(\theta-\gamma)^2}$ where δ and γ are the peak locations (as phase angles) near the nucleus and coda respectively, a and c are amplitude terms specifying the perturbation minimum and maximum acceleration, and b and d control the rate of onset of the perturbation.

A function such as this will provide the observed effects of segment lengthening and shortening by *continuously* varying the rate of motion of the syllable. Figure 3 shows the acceleration function for both the simple oscillator (dark line) and the oscillator perturbed by the hypothetical decelerate-accelerate function

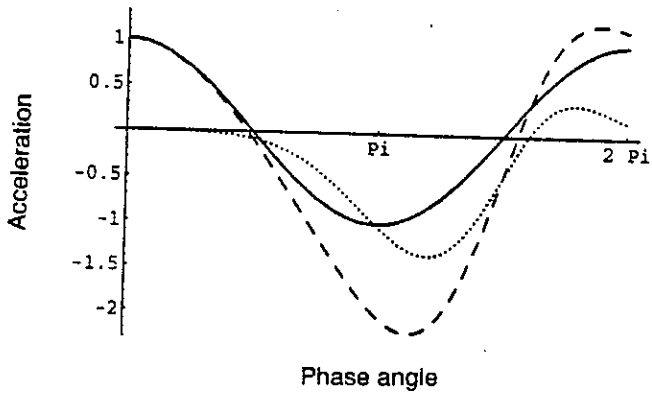

Figure 3: A plot of acceleration for a simple homogeneous oscillator (black curve) together with the perturbed oscillator (dashed) and the perturbation function itself (dotted). The equation in the text computed these curves using the parameter values: $a = 1.5, b = .5, c = .9, d = 1.0, \delta = 4, \gamma = 5.2$.

required to transform, say, *dense* into *dens* (dashed line). It was computed using the equation above with plausible values for the relevant parameters. The dotted line shows the combined perturbation function itself for these curves, illustrating gradual deceleration through the mid region and then acceleration through the syllable-final obstruent. This figure shows that there need be no underlying assumption of discrete segments. Instead, it is assumed that all the simultaneous gestures of speech production will be lengthened or shortened in a certain way depending on their relative position with respect to the perturbation function.

It will be noted that this function is quite general. It has separate parameters that specify both the phase angle and degree of rate change for each of the two perturbations independently. The deceleration and acceleration gestures are simply superimposed on each other and on the underlying syllable acceleration. Thus, for syllables with no coda obstruents, like *bull* or *Sam*, only the deceleration term applies, with a final [$-$*voice*] obstruent (e.g.,

/t) or obstruent cluster (e.g., *stamps*), both deceleration and acceleration terms exist. In some cases, it has been observed that the total syllable duration is unaffected by the change in voicing (e.g., in *dibber-dipper* and *deeber-deeper*) giving rise to the hypothesis that speakers are manipulating the *ratio* of the vowel to the final consonant in these cases [19], but in other cases the total syllable duration is considerably longer in the [+*voice*] case (as they do not in Figure 1 above). This is evidence that velocity perturbation provides a better model, since the amount of deceleration need not equal the amount of acceleration (as occurs in both Figures 1 and 3). Clearly, more data is required to allow better specification of the perturbation function.

3 Conclusions

Under a segmental analysis, the influence of the feature [*voice*] in syllable-final obstruents appears inordinately complex, and requires the postulation of a huge array of segment-based phonetic implementation rules, plus an additional, thusfar unspecified, mechanism for actual segmental duration control. Our model, however, accounts for all these effects by the perturbation of the velocity of an abstract syllable oscillator which in turn controls the dynamics of articulator motion. Even though the data which provided the initial evidence for such a rate-perturbation function were based on segmental measurements, the assumption of segmentation is incompatible with the model. Instead, we require only some way to specify temporal location in terms of syllabic phase angle and the ability to impose a very local, intrasyllabic perturbation on the rate of instantaneous articulation to account for the effects of voicing on syllables with final obstruents or without them. This seems to be the first model for a linguistic segmental feature that demands description in dynamic terms. Thus, this model provides a bridge between phonological segments and their phonetic implementation in time, and, at the same time, demonstrates that not even phonology can be described independently of the dynamic mechanisms of speech articulation.

Acknowledgments. We are grateful to Joseph Stampfli, James Townsend and Sven Anderson for comments and discussion. This research was supported in part by the Office of Naval Research, Grant Number ONR-N00014-91-J-1261.

References

[1] C. Browman and L. Goldstein. Towards an articulatory phonology. *Phonology Yearbook*, 3:219–252, 1986.

[2] C. Browman and L. Goldstein. Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M. E. Beckman, editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, chapter 19, pages 341–376. Cambridge University Press, Cambridge, England, 1990.

[3] P. Denes. Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, 27:761–764, 1955.

[4] J. Emil Flege, M. L. Munro, and L. Skelton. Production of the word-final English /t/-/d/ contrast by native speakers of English, Mandarin and Spanish. *Journal of the Acoustical Society of America*, 92(1):128–143, 1992.

[5] M. Fourakis and R. Port. Stop epenthesis in English. *Journal of Phonetics*, 14:197–221, 1986.

[6] C. Fowler, P. Rubin, R. Remez, and M. Turvey. Implications for speech production of a general theory of action. In B. Butterworth, editor, *Language Production*, pages 373–420. Academic Press, 1981.

[7] P. C. Gordon. Context effects in recognizing syllable-final /z/ and /s/ in different phrasal positions. *Journal of the Acoustical Society of America*, 86:1696–1707, 1989.

[8] B. Kay, E. Saltzman, and J. A. S. Kelso. Steady-state and perturbed rhythmical movements: A dynamical analysis. Technical Report SR-103/104, Haskins Laboratories, New Haven, Connecticut, 1990.

[9] P. A. Keating. Universal phonetics and the organization of grammars. In V. Fromkin, editor, *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, pages 115–132. Academic Press, 1985.

[10] J. A. S. Kelso and B. A. Kay. Information and control: a macroscopic analysis of perception-action coupling. In H. Heuer A. F. Sanders, editors, *Perspectives on Perception and Action*, chapter 1, pages 3–32. L. Erlbaum, 1987.

[11] J. A. S. Kelso, E. Saltzman, and B. Tuller. The dynamical perspective in speech production: Data and theory. *Journal of Phonetics*, 1986.

[12] D. Klatt. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59:1208–21, 1976.

[13] D. H. Klatt. Vowel length is syntactically determined in connected discourse. *Journal of Phonetics*, 3:129–140, 1975.

[14] L. Lisker. *Rabid* vs. *rapid*: a catalogue of cues. *Haskins Laboratories Status Report on Speech Research*, 1985.

[15] M. Macchi, M. Spiegel, and K. Wallace. Modelling duration adjustment with dynamic time warping. In *ICASSP*, pages 333–336, New York, New York, 1990. IEEE. IEEE.

[16] M. J. Munro. *Perception and production of English vowels by native speakers of Arabic*. PhD thesis, University of Alberta, Edmonton, Alberta, 1991.

[17] R. Port and J. Dalby. C/V ratio as a cue for voicing in English. *Journal of the Acoustical Society of America*, 69:262–74, 1982.

[18] R. Port, J. P. Mora, and C. deJonge. Usefulness of temporal detail for word identification by native and non-native listeners. Indiana University. To be submitted, 1992.

[19] R. F. Port. Linguistic timing factors in combination. *Journal of the Acoustical Society of America*, 69:262–274, 1981.

[20] R. F. Port and R. Rotunno. Relation between voice-onset time and vowel duration. *Journal of the Acoustical Society of America*, 66(3):654–662, 1979.

[21] L. J. Raphael, M. F. Dorman, F. Freeman, and C. Tobin. Vowel and nasal duration as cues to voicing in word-final stop consonants: spectrographic and perceptual studies. *Journal of Speech and Hearing Research*, 18:389–400, 1972.

[22] E. Saltzman and K. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382, 1989.

[23] V. Summers. Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analysis. *Journal of the Acoustical Society of America*, 82:847–863, 1987.

[24] E. Vitikiotis-Bateson. The temporal effects of homorganic medial nasal clusters. In *Research in Phonetics*, 4, 199–233. Department of Linguistics, Indiana Univ., 1984.