

---

# Defining and Measuring Speech Movement Events

## RESEARCH NOTE

---

### Stephen M. Tasko

Army Audiology & Speech Center  
Walter Reed Army Medical Center  
Washington, DC

### John R. Westbury

Waisman Center and  
Department of  
Communicative Disorders  
University of Wisconsin  
Madison

A long-held view in speech research is that utterances are built up from a series of discrete units joined together. However, it is difficult to reconcile this view with the observation that speech movement waveforms are smooth and continuous. Developing methods for reliable identification of speech movement units is necessary for describing speech motor behavior and for addressing theoretically relevant questions about its organization. We describe a simple method of parsing movement signals into a series of individual movement “strokes,” where a stroke is defined as the period between two successive local minima in the speed history of an articulator point, and use that method to segment speech-related movement of marker points placed on the tongue blade, tongue dorsum, lower lip, and jaw in a group of healthy young speakers. Articulator fleshpoints could be distinguished on the basis of kinematic features (i.e., peak and boundary speed, duration and distance) of the strokes they produce. Further, tongue blade and jaw fleshpoint strokes identified to temporally overlap with acoustic events identified as alveolar fricatives could be distinguished from speech strokes in general on the basis of a number of kinematic measures. Finally, the acoustic timing of alveolar fricatives did not appear to be related to the kinematic features of strokes presumed to be related to their production in any direct way. The advantages and disadvantages of this simple approach to defining movement units are discussed.

**KEY WORDS:** speech, movement, segmentation, kinematics

---

**W**hen we consider the temporal organization of speech, we are often referring to the way particular events within the signal stream are sequenced or coordinated. The long-held view is that each utterance is built up from a series of discrete “sound units” joined together. Speech acoustic waveforms seem to prompt such a view because they include modulations in energy, every 100 ms or so, often corresponding to abrupt changes from periodic to aperiodic energy, or from the presence to the absence of energy. These changes serve as segmentation points for sound units—for example, the interval of frication marking the initial consonant /f/ in the word *fall*; a loud, phonated interval marking the following vowel /ɔ/; a lower-amplitude phonated interval marking the final consonant /l/. Speech acoustic waveforms segmented in this way imply that production involves a successive approximation of targets, each maintained for a while and linked to preceding and following targets by brief transitions. Moreover, this interpretation of speech obscures the fact that many separate physical events, such as muscle contractions and articulator movements, co-exist during each putative target or transition. Speech movement waveforms, representing position histories of fleshpoints or landmarks recorded from multiple articulators, highlight both the underlying complexity of speech production and also its continuity. Held positions among concurrent movements

are rare, and any single articulator usually moves smoothly and continuously from one local position extremum to another. Hence, there are few obvious intervals to designate as targets or transitions, and we are hard-pressed to say where any units, much less the classical and familiar sound units, are in such streams.

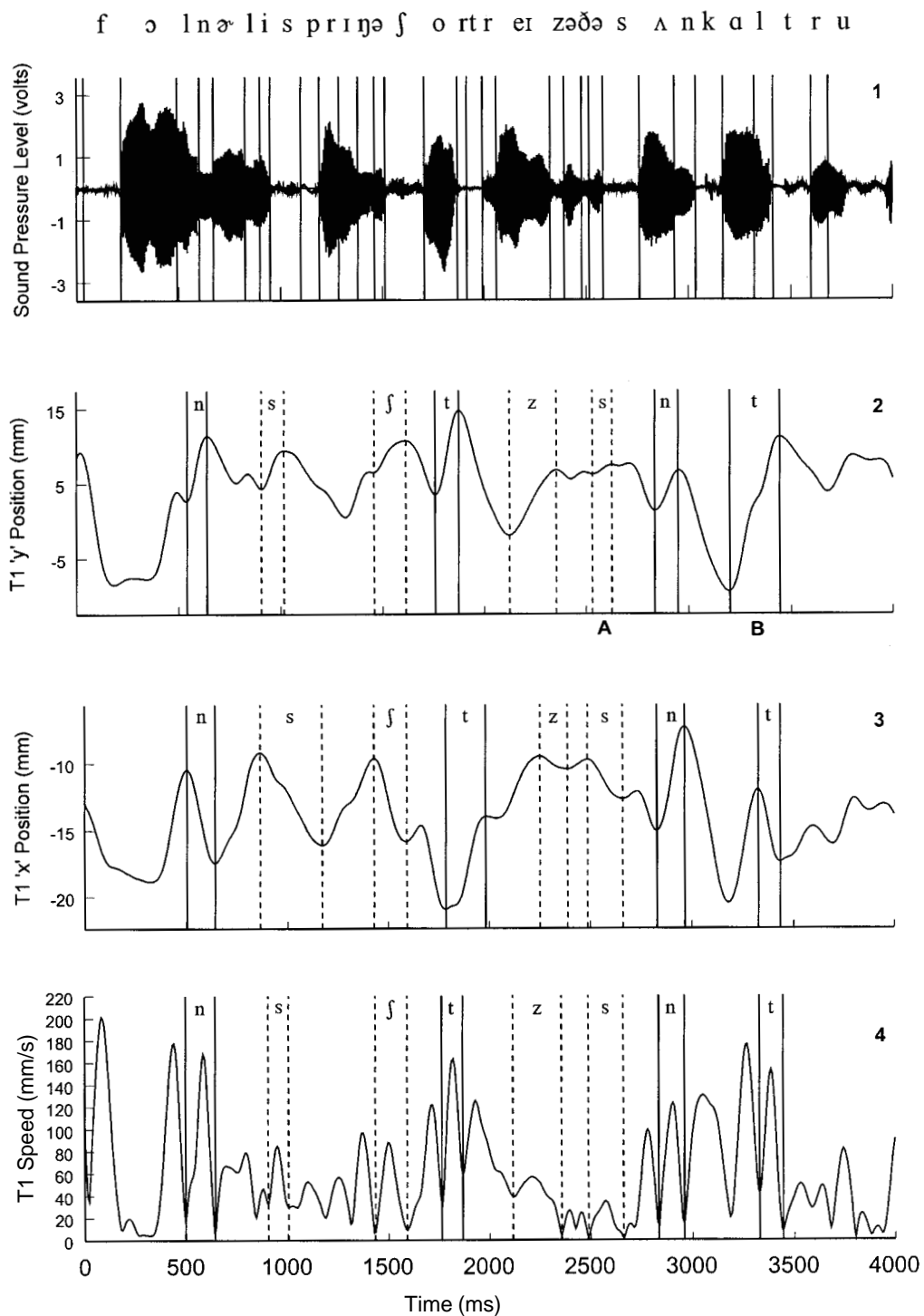
These problems are illustrated in Figure 1, where a set of signal streams associated with a speaker reading the phrase *...fall and early spring, the short rays of the sun call a true...* have been plotted. The top panel (1) shows the sound pressure wave, segmented and labeled phonetically. The remaining panels show the midsagittal vertical position (2), horizontal position (3), and speed of a small marker (labeled T1) attached to the speaker's tongue blade. From the kinematic channels, we see that the tongue blade marker does not exhibit extremely abrupt positional changes, nor does it exhibit periods of zero position change. This is what we mean when we say that speech movement is smooth and continuous. Two basic questions we might ask about movement waveforms are these: How can we parse movement waveforms into component "movement units" for descriptions and analyses that are overtly kinematic in nature? And assuming the parsing can be done, how will the resulting units be related to more familiar units derived from other types of speech analysis?

Returning to Figure 1, let us focus attention on eight acoustically defined intervals associated with alveolar stops (i.e., /t/ and /n/) and fricatives (i.e., /s/, /z/, and /ʃ/). These sound segments were chosen simply because they likely have associated tongue blade motion and, therefore, may serve as focal points for further discussion. If we look only at the vertical tongue blade position in panel 2, there is consistent elevation near the leading edge of each acoustically defined segment. Each elevation event has an onset and offset defined by local position minimum and maximum, respectively. Therefore, as a first pass, we might define each pair of elevation events as the occlusive movement unit that corresponds to each stop or fricative production. Solid lines mark *stop* units, and broken lines mark *fricative* units. Defining movement units in terms of excursion along a single movement dimension is a common practice, whether the relevant dimension is defined in anatomical terms (e.g., Löfqvist & Gracco, 1994, 1997), in more functional terms such as a principal component axis (e.g., Adams, Weismer, & Kent, 1993; Hertrich & Ackermann, 2000), or with respect to the orientation of the measurement instrument (e.g., Kelso, Vatikiotis-Bateson, Saltzman, & Kay, 1985; Ostry & Munhall, 1985). Having identified these movement units, it is then possible to evaluate characteristics of the units themselves—such as their amplitude, velocity, duration, and so forth—or their temporal relation to events or units within the acoustic or other movement signal streams.

Although there are many advantages to the approach illustrated in panel 2 of Figure 1, interpretation of events can be complicated by a number of factors. For example, some events are much easier to identify than others. The segment labeled "A" is a small amplitude excursion bounded by two extremely small negative-going excursions. If we look only at the movement waveforms, we might question the wisdom in defining these very small inflections as distinct movement unit boundaries. However, the acoustic signal reveals that this short period of speech contains a number of acoustic-phonetic boundaries. In this case, what looks like a questionable decision in identifying a movement is justified with knowledge of acoustic-phonetic events. Alternatively, the segment labeled "B" is characterized by a very large upward excursion. Although there is a small deflection in the positive-going signal, there is not a direction reversal and so "B" is identified as a single movement unit. However, comparison with the acoustic signal indicates that this movement unit extends over a number of acoustic-phonetic events. In this case, defining segment "B" simply as a stop closure may be questionable. These examples highlight familiar problems in movement unit definition and analysis (Harris, Tuller, & Kelso, 1986). Perhaps an even more important consideration when defining movement segments along any single movement dimension is that the time location of events (e.g., marked at direction reversals) is linked to the axis orientation. Changing the axis along which the movement is measured usually changes the time location of the event. This is clearly the case for some movement examples shown in Figure 1. Panel 3 shows horizontal movements of the tongue blade during the test phrase. Had we used the horizontal position history, instead of the vertical one, to define movement segments relative to the onsets and offsets of the acoustic segments, we would find that the horizontally defined movement units often do not align with the vertically defined units in any straightforward way. This fact raises the obvious question of what set(s) of measurements we should use to define the movement units.

One alternative would be to define a measurement unit in a way that is not influenced by the reference axes that are chosen to frame the data. Movement speed (i.e., the magnitude of the rate of change of position with respect to time, or simply  $\sqrt{(dx/dt)^2 + (dy/dt)^2}$ ), provides a suitable signal for such a definition because speed will be the same for a trajectory represented in an anatomically based coordinate system, as in this example, or (alternatively) in a "functionally" based coordinate system where the reference axes are given by the principal components of the trajectory itself (Westbury, Severson, & Lindstrom, 2000). Panel 4 of Figure 1 shows the speed of the tongue blade marker for the speech sample. We can see that the speed of the tongue blade

**Figure 1.** Panel 1 plots the sound pressure level wave of a male speaker orally reading "...fall and early spring, the short rays of the sun call a true..." The signal has been partitioned into presumed phonetic segments using conventionally accepted segmentation rules. Panel 2 plots the vertical position history of a marker placed on the tongue blade (T1). Panel 3 plots the horizontal position history of the T1 marker. Panel 4 plots the speed of the T1 marker. See text for details regarding solid and broken lines and labels.



marker never falls to zero. There is always movement, though there are moments when movement is relatively slow. In fact, the tongue blade speed history creates roughly a saw-toothed pattern in which troughs separate speed peaks. This fast-slow-fast-slow pattern is typical of speech and reflects alternating openings and closings of the vocal tract. As a first pass, a straightforward approach to defining movement units using the speed history would be to consider the local minimum values, corresponding to troughs, as unit boundaries. In this way a movement unit is defined simply as a period between two consecutive speed minima. Each unit contains a period of acceleration followed by a period of deceleration. Using this definition, a speech sample of any length may be parsed into a consecutive series of movement units, each with a set of describable features such as duration, extent, and maximum speed attained. Further, these movement units may be labeled according to such variables as phonetic identity, syllabic stress, and location within an utterance. In our example, we are able to define tongue blade movement units that temporally overlap the acoustically defined onsets of stops and fricatives. As was done before, the boundaries of these speed units are marked with solid lines for stops and broken lines for fricatives. There are at least two noteworthy observations. First, we can see that the edges of these units often (but not always) line up with a local extremum in one of the individual movement dimensions (panels 2 and 3). However, sometimes the extremum is in the "x" dimension and sometimes it is the "y" dimension. Second, if we group the stops and fricatives, the maximum speeds associated with fricative productions are typically smaller than those associated with stops. This is a simple example of the kinds of comparisons that can be made without concern that axis definition is influencing the results.

Establishing and evaluating approaches for movement unit definition has both theoretical and practical implications. The manner in which the speech units, whether they are movements, sounds, syllables, or words, unfold in time is at the heart of many theoretical discussions on speech production (e.g., Browman & Goldstein, 1992; Fowler, 1977; Kent & Minifie, 1977; Lashley, 1951; MacNeilage, 1970). From the perspective of gestural phonology, for example, it is not hard to imagine how the reliable identification of "surface" manifestations of gestures could be necessary for evaluating theoretical notions about their underlying specification. Reliable procedures for movement identification could also be useful in any attempts to understand how time is distributed in speech that are based upon well-known motor control principles such as Fitts' Law (Fitts, 1954) or the so-called Power Law (Viviani & Cenzato, 1985). Finally, on a practical front, some speech disorders, such as stuttering, are considered to be the result of disrupted timing within and

coordination across production units (Kent, 1984). Standards for unit identification and description can serve as important tools underlying comparisons of normal and disordered speech. The main purpose of this report is to provide a detailed description of a method of movement unit identification based on pellet speed and to summarize certain results that have been obtained from its application to streams of orofacial movements generated by a group of healthy speakers reading aloud.

---

## Method

### Speakers

Speech materials recorded from 18 speakers were selected for analysis from the University of Wisconsin X-ray Microbeam Speech Production Database (XRMB-SPD). The XRMB-SPD is a publicly available database that includes the acoustic signal and synchronous mid-sagittal-plane motions of eight articulator fleshpoints recorded from 57 neurologically and communicatively healthy, young, native-English speaking adults performing a range of speech and nonspeech oral activities. The age, sex, and dialect base (i.e., place of residence during linguistically formative years) of the 18 speakers selected for our analysis are shown in Table 1, along with some information about reading rate and speech stroke counts. Additional physical and demographic information can be found in the XRMB-SPD Handbook (Westbury, 1994).

### Speech Task

Speakers performed an oral reading of a slightly expanded version of the Hunter script (Crystal & House, 1982) read at a self-selected speaking rate. For technical reasons, the approximately 300-word reading passage was recorded in four separate records ranging from 21 s to 25 s in length. These four records served as the source for all analysis performed in this study. The Hunter script was selected because there is evidence that the phonetic distribution of the script approximates that of conversational speech (Crystal & House, 1982). To estimate the oral reading rate for each speaker, we derived the quotient of the total number of syllables read and total passage duration.

### Data Acquisition and Processing

Data acquisition was performed using the University of Wisconsin X-ray Microbeam system, according to the procedures described in Westbury (1994). Briefly, articulator motion was recorded by tracking the mid-sagittal positions of small gold pellets (2–3 mm in diameter) glued to locations within and around the oral cavity. The movements of five pellets were the focus of this

**Table 1.** Speaker characteristics, oral reading rates, and quantitative information on the data used for analysis. Values in parentheses indicate the number of strokes that fell below the arbitrary cut-off values discussed in the text.

Speaker	Gender	Age (yrs)	Ht (in)	Wt (lbs)	Dialect base	Reading rate (syll/s)	No. of words read (max=303)	T1 stroke count	T4 stroke count	LL stroke count	MI stroke count
S12	M	21	72	155	Marinette, WI	4.18	287	528 (436)	623 (508)	694 (516)	482 (367)
S15	M	22	66	130	Milwaukee, WI	3.94	290	682 (537)	441 (353)	629 (509)	497 (277)
S16	F	20	65	140	Kiel, WI	3.61	272	698 (574)	631 (478)	759 (478)	484 (419)
S20	F	25	67	125	Milford, MA	3.84	284	339 (283)	612 (515)	744 (478)	496 (402)
S24	M	20	72	150	Jefferson, WI	4.15	286	678 (592)	613 (487)	735 (477)	506 (374)
S26	F	24	65	140	Verona, WI	4.14	277	730 (584)	658 (507)	761 (493)	499 (429)
S31	F	20	62	130	New Holstein, WI	3.55	258	690 (523)	623 (462)	672 (510)	468 (386)
S34	F	21	68	145	Amery, WI	4.69	286	669 (558)	596 (431)	652 (537)	466 (371)
S39	F	24	66	140	Rochester, MN	4.51	290	719 (536)	614 (464)	736 (437)	470 (376)
S41	M	20	75	190	Milwaukee, WI	4.10	279	667 (534)	661 (524)	668 (493)	540 (342)
S45	M	21	69	173	Mishawaka, IN	4.34	288	730 (573)	642 (482)	767 (447)	506 (382)
S47	F	25	71	160	Indianola, IA	4.39	299	715 (542)	637 (499)	720 (463)	473 (382)
S48	F	21	61	150	Maywood, IL	3.80	266	761 (510)	628 (458)	701 (405)	469 (351)
S51	M	19	72	175	Madison, WI	4.40	289	742 (551)	678 (479)	747 (428)	537 (317)
S52	F	26	68	125	Kewanee, IL	4.40	303	696 (560)	635 (492)	665 (512)	445 (362)
S58	M	23	71	165	Fair Lawn, NJ	3.52	275	698 (544)	600 (489)	823 (418)	496 (397)
S59	M	29	69	160	Sauk City, WI	4.50	303	696 (592)	611 (512)	521 (401)	491 (424)
S63	M	21	71	145	Los Angeles, CA	4.33	286	686 (549)	617 (465)	673 (497)	516 (383)

study. Two pellets (T1 and T4) were respectively positioned on the blade and dorsum of the midline surface of the tongue. Two mandibular pellets were positioned at the gum line between the mandibular incisors (MI) and between the first and second molar on the left side (MM). The final pellet was positioned midline at the vermilion border of the lower lip (LL). These five pellets were selected because they represent fleshpoints distributed across three potentially independent articulators, allowing for comparisons between articulators for the same speaking task. The two tongue pellets representing the dorsum and the blade of the tongue were included to allow for comparison between tongue regions that have often been considered functionally distinct (Mermelstein, 1973).

Pellet positions were originally sampled at rates of 40 Hz (MI and MM), 80 Hz (T4 and LL), and 160 Hz (T1), and the simultaneous sound pressure level signal was recorded at 22 kHz. Following acquisition, the coordinate histories for each pellet underwent a series of processing steps. The coordinate histories of all pellets, including reference fiducials at the maxillary incisors and along the bridge of the nose, were low pass filtered at 10 Hz and resampled at 145 Hz. The MI pellet positions were then expressed in an anatomically based Cartesian coordinate system in which the abscissa was

located along the maxillary occlusal plane, and the ordinate was normal to the abscissa where the central maxillary incisors met the maxillary occlusal plane. The positions of the T1, T4, and LL pellets were then decoupled from and expressed relative to the position of the mandible, with a local origin at the mandibular incisor (MI) pellet and an x-axis defined by MI and MM, parallel to the mandibular occlusal plane. This data expression allowed for the assessment of tongue and lower lip positions independently of changes in mandibular position. The coordinate histories for all pellets were imported into a PC-based commercial data-analysis software package (S-Plus, StatSci, 1993) for all subsequent signal processing and analysis specific to this study.

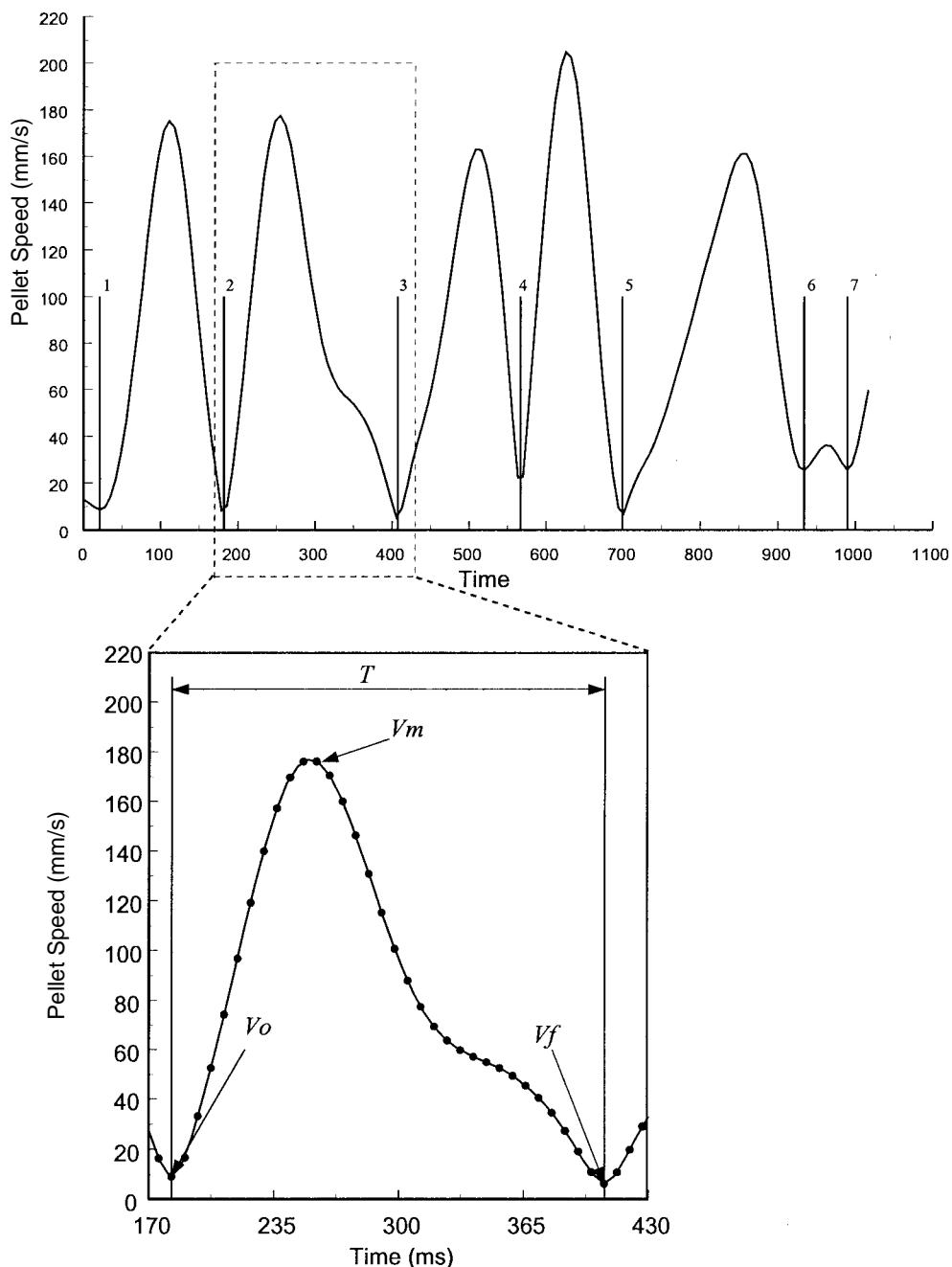
For example, the speed history  $v(t)$  for each pellet was derived using the following formula

$$v(t) = \sqrt{[\dot{x}(t)]^2 + [\dot{y}(t)]^2} \quad (1)$$

where  $\dot{x}(t)$  and  $\dot{y}(t)$  represent the velocity components of the first-order time derivatives of the position history  $[x(t), y(t)]$ . The derivatives were approximated using a three-point central difference method.

The upper panel of Figure 2 shows the speed history of the T1 pellet for approximately one second of connected speech. During this interval, speed varies in

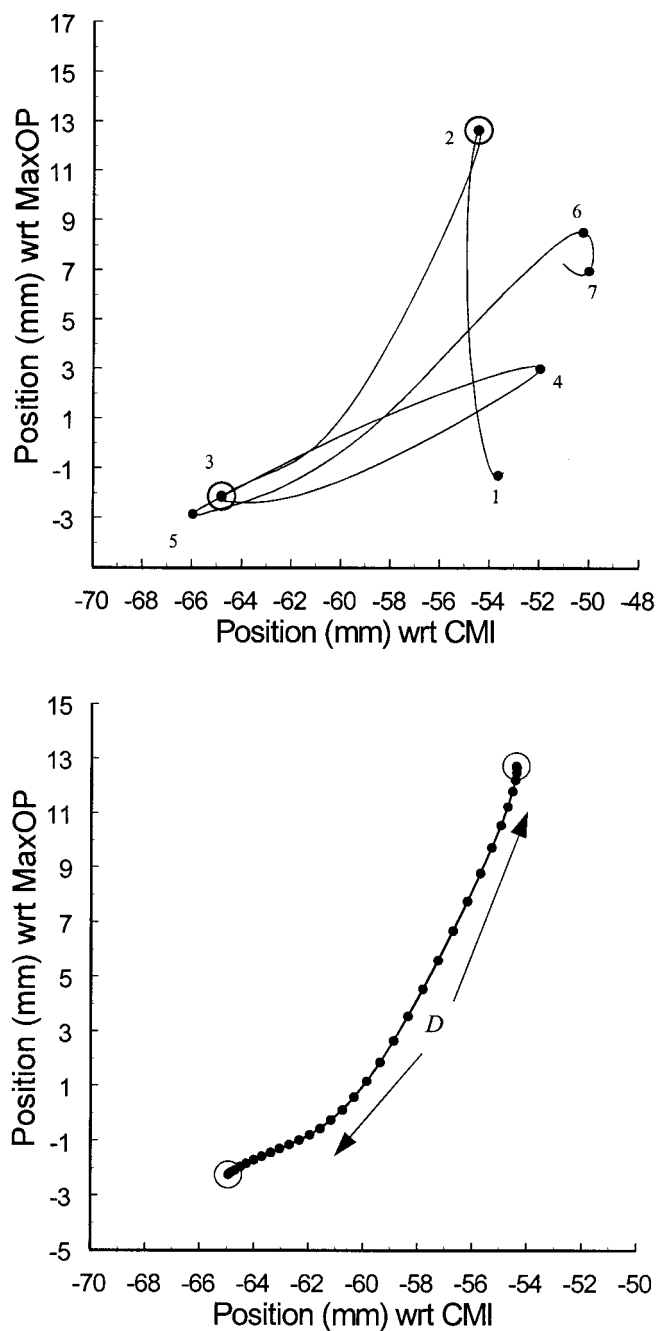
**Figure 2.** The upper panel plots the speed history of T1 during approximately 1 s of connected speech. The saw-tooth pattern in this example is typical across pellets and speakers. The signal has been parsed at local magnitude minima in the speed history in order to define individual movement strokes. The lower panel is an expansion of the region bounded by the box in the upper panel and plots an individual movement stroke. The following kinematic measures have been labeled:  $V_m$ , peak stroke speed;  $T$ , stroke duration;  $V_o$ , speed at stroke onset;  $V_f$ , speed at stroke offset.



a “saw-toothed” pattern in which local minima (i.e., troughs) separate alternating local maxima (i.e., peaks). Each speed history was parsed into successive epochs at its troughs, located at those moments (interpolated to the nearest ms) corresponding to sign changes in the

first time derivative of the speed history itself. The pellet trajectory for the same speech interval shown in Figure 2 is illustrated in the upper panel of Figure 3. The numbered points in Figures 2 and 3 represent the speeds and positions respectively, at corresponding parsing

**Figure 3.** The upper panel is a plot of the 2-dimensional pellet trajectory for the sample used in Figure 2. The numbers labeling the solid circles correspond to the indices numbered in the speed history. The lower panel represents the portion of the trajectory between the larger open circles and corresponds to the speed history in the lower panel of Figure 2. The stroke distance  $D$ , which is defined as the sum of the segment lengths along the movement trajectory, is labeled.



moments. Note that the minimum speeds are usually associated with abrupt changes in direction of the pellet trajectory, though this does not have to be so. It is

easy to imagine a situation in which a pellet might move along a straight path, slowing down and then speeding up again, yielding a trough. Returning to Figures 2 and 3, it can be seen that even though the parsing was based solely on the speed history, the boundaries of successive epochs appear to occur at reasonable places along the trajectory (i.e., changes in movement direction). Successive speed minima have the effect of defining movement “strokes” in each pellet trajectory, strictly in kinematic terms, and not in acoustic or phonetic ones.

For each individual stroke, many descriptive kinematic measures can be made. We chose to make four such measures, defined below, as part of our initial analysis of the data. For purposes of illustration, these measures are marked in the bottom panels of Figures 2 and 3. The measures were—

*Stroke distance ( $D$ ):* the sum (mm) of the segment lengths between adjacent samples, along the length of a movement trajectory sampled  $n$  times.

*Stroke duration ( $T$ ):* the duration (ms) between successive speed minima.

*Peak stroke speed ( $V_m$ ):* the maximum speed (mm/s) generated within the duration of the stroke of interest.

*Boundary speed ( $V_o$  and  $V_f$ ):* the pellet speed (mm/s) at the onset and offset of the stroke. In the speech data set, the onset of one stroke is typically the offset of the previous stroke.

For each pellet-by-speaker condition, a data set was derived that included all movement strokes occurring within the entire speech task, beginning with the stroke immediately preceding the acoustic onset of the first word in the speech record and ending with the stroke immediately following the acoustic offset of the last complete word of the record. It was not uncommon for the fixed recording period to end before speakers could finish the reading passage. Therefore, the last word spoken varied across speakers, and thus the total word count varied across speakers. Within each speaker’s full data set, no initial attempt was made to “categorize” movements, phonetically or otherwise. This is a somewhat unorthodox approach to studying speech movement and is not meant to imply that categories such as phonetic context or syllabic stress are unimportant. Instead, our initial focus was to document features of speech movement strokes in a general way.

Subsequent to our first-pass examination of the data, we then extracted the subset of T1 and MI strokes that temporally overlapped the onset of all voiced and voiceless alveolar fricatives in the speech task. These sounds occur frequently enough in the passage ( $N = 88$ ) to give adequate samples for statistical treatment. The purpose of this second-level analysis was to learn about possible speech kinematic “signatures” associated with

this class of speech sounds. We use “kinematic signature,” analogously to Weismer and colleagues’ (Weismer, Kent, Hodge, & Martin, 1988) use of “acoustic signature,” to refer to a set of speaker-general features common to some defined class of speech events (e.g., sounds or words). Using the sound pressure wave and wide band spectrograms derived from it, all examples of /s/ and /z/ produced by each speaker were segmented acoustically, at judged onsets and offsets of aperiodic high-frequency energy. We excluded from the analysis all tokens that were perceptually distorted. Distortions were rare, constituting less than 2% of the total data set. Those movement strokes that temporally overlapped the onsets of acoustically defined fricatives were identified as the strokes most closely associated with forming the fricative constrictions. Analysis of the fricative data set was restricted to the T1 and MI marker pellets because these points are closest to the primary constriction.

## Results

### *Speech Acoustic and Kinematic Units and Their Relationship*

There is no simple way to map classical sound, syllable, or word units onto movement strokes and, hence, no simple way to talk about relationships between their respective “time boundaries” or locations. One way to illustrate this point is to compare counts of unit types in one of the Hunter passage sentences. The first sentence, beginning *In late fall and early spring...*, nominally includes 28 words, 31 syllables, and roughly 80 consonants and vowels (e.g., depending upon whether the typically elided /d/s are counted in *and* and the first syllable of *childhood*). Of course, the “true” number of such units produced by speakers would also depend upon the clarity or formality of their performances. If we were to imagine that each adjacent pair of sounds in this sentence should yield a tongue stroke spanning the “transition” between the adjacent “sound centers,” we would expect to find one fewer strokes than sounds. In fact, and on average across 18 speakers, the strokes in the speed histories associated with T1, T4, LL, and MI numbered only 64, 52, 57, and 42, respectively. The number of T1 strokes most closely approximated the nominal number of sounds in the sentence, whereas the number of MI strokes most closely approximated the nominal number of syllables. However, in general, there were typically fewer strokes than adjacent sound pairs, more strokes than syllables or words, and different numbers of strokes per articulator. These facts are perhaps no great surprise. We report them for two reasons. The first is to give a general sense of the parallel, anisomorphic distribution of classical segmentation units and movement strokes. The second is to suggest that the

functional significance of movement strokes associated with any single articulator, whatever that might be, is not the same as that of the classical units spanned by the strokes. In some way, then, strokes from different articulators must be juxtaposed and woven in time, perhaps like a *score* in gestural phonology (Browman & Goldstein, 1992), to yield the more familiar sounds, syllables, and words that are typically defined using either acoustic and/or linguistic criteria.

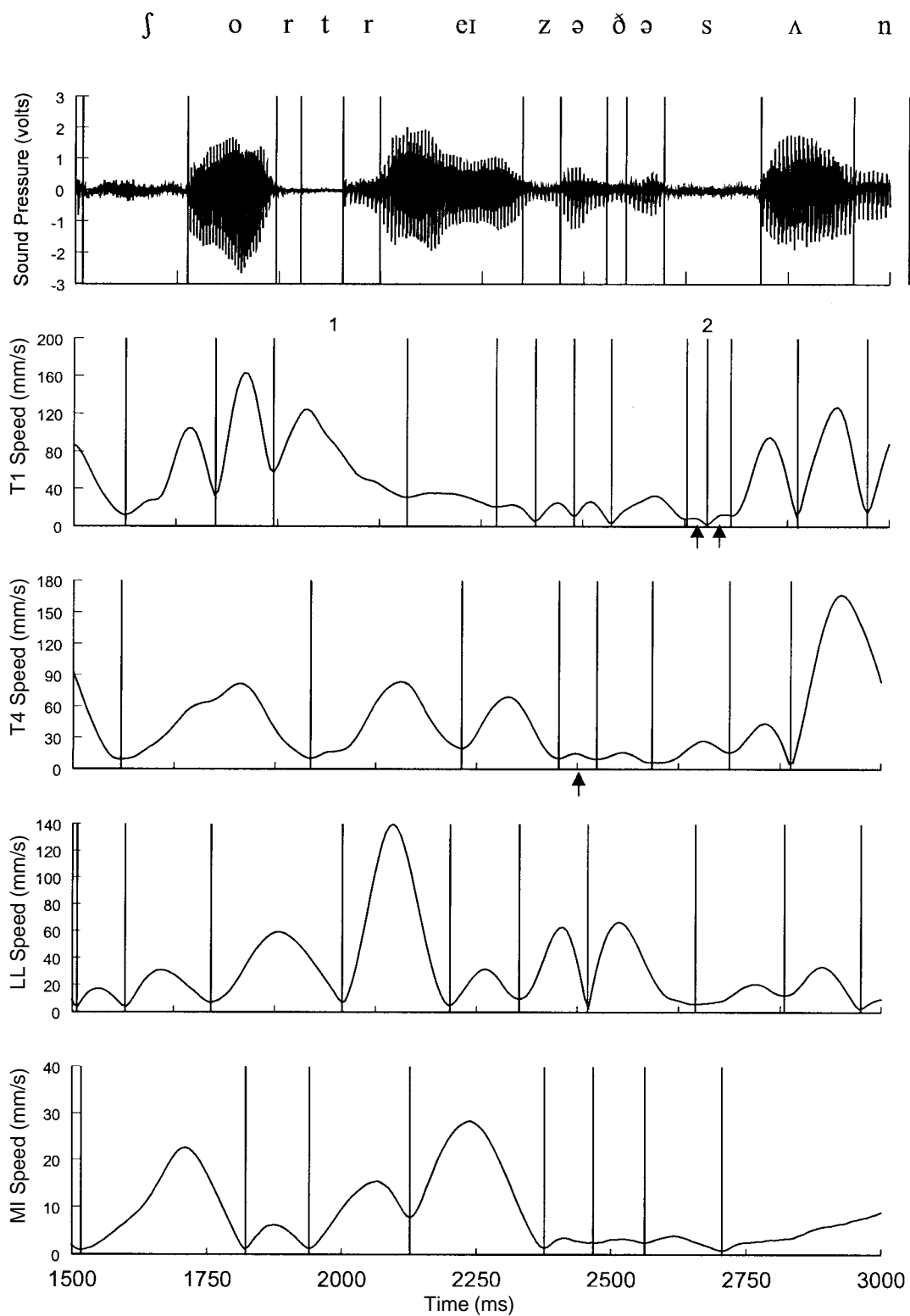
We illustrate differences in unit locations in a concrete way in Figure 4, which shows the sound pressure wave above and the four pellets’ speed histories below, corresponding to the nominal phrase *...short rays of the sun...* recorded from S15. The boundaries of sound segments in the speech wave have been marked using typical criteria (e.g., “discontinuities” in the wave represented by quick changes in its associated spectrogram: Olive, Greenwood, & Coleman, 1993), whereas the stroke boundaries are defined by successive speed minima. This phrase contains 5 syllables and about 13 broadly transcribed speech sounds. Respectively, there are 12, 7, 9, and 7 T1, T4, LL, and MI strokes associated with the phrase. For this example, the T1 stroke count closely corresponds to the number of sounds. However, the boundaries of the acoustic segments and the T1 strokes do not align or overlap in any systematic way. For example, the T1 stroke marked “1” overlaps more than three acoustic segments. On the other hand, the time point marked “2” has several small T1 strokes located within a single acoustic segment. Similar examples can be found for T4, LL, and MI movements.

Figure 4 illustrates the implication of the articulator-related differences in the stroke counts of the first sentence. Across articulators, strokes seldom begin and end in the same place. There are exceptions. For example, T1, T4, and LL have strokes that begin at roughly the same point within the midpoint of the /j/ segment, and the T1 and LL strokes both end at about the same point. However, most strokes do not appear to align in simple ways. The timing of these events ought to be of interest for understanding how articulators are coordinated to meet communicative demands.

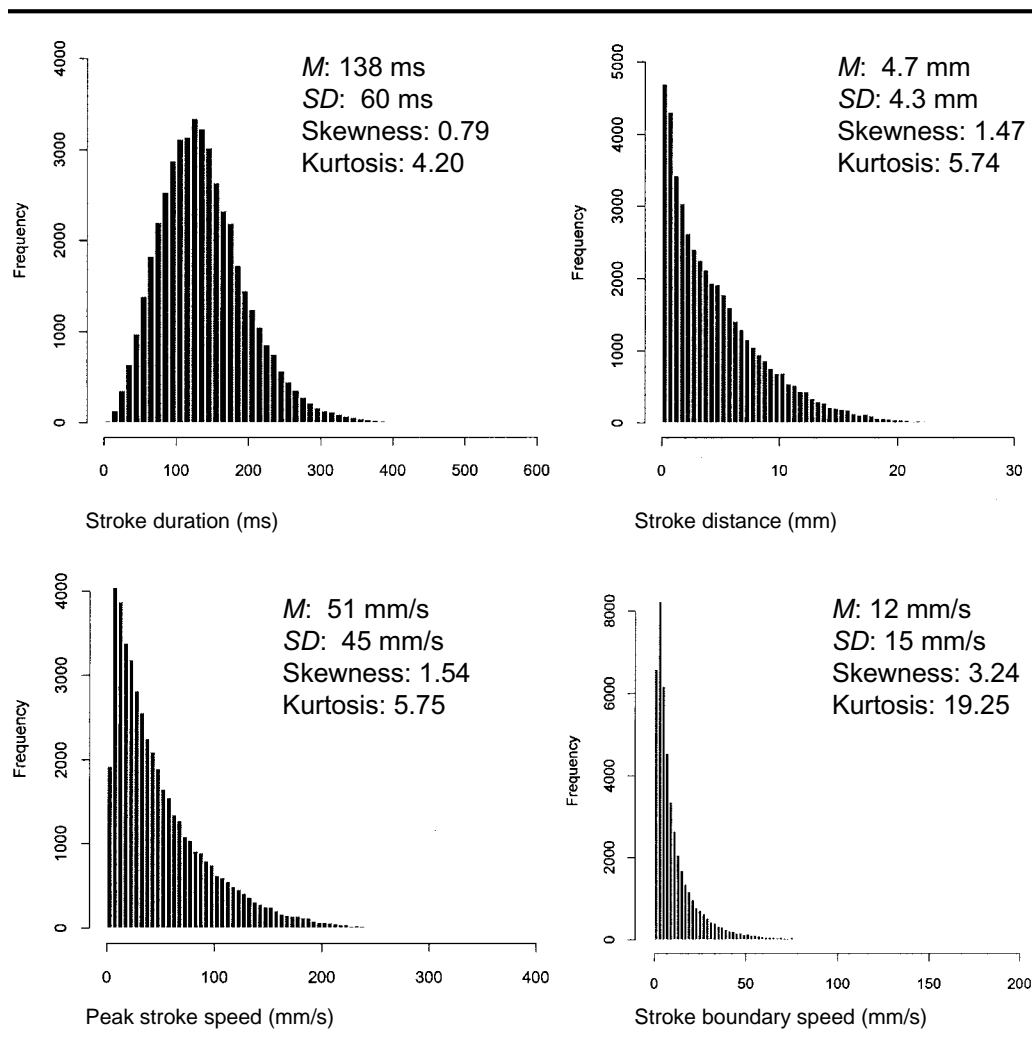
### *Kinematic Features of Movement Strokes*

The kinematic features of movement strokes take on a wide variety of values. Figure 5 plots frequency histograms. All observed strokes for all four articulator points across all 18 speakers ( $N = 45736$ ) are represented in the histograms. Along with each histogram are the distribution’s mean, standard deviation, skewness, and kurtosis. The normal distribution has a skewness equal to zero and a kurtosis equal to 3 (Kleinbaum, Kupper, & Muller, 1988). Skewness values less than and greater than zero respectively reflect distributions with

**Figure 4.** The top panel plots the acoustic wave, segmented using standard criteria. The bottom four panels respectively represent tongue blade (T1), tongue dorsum (T4), lower lip (LL), and mandibular incisor (MI) speed histories, segmented according to the methods described in the text. The arrows mark examples of strokes that have very short movement paths.



**Figure 5.** These frequency histograms show the distributional characteristics of each kinematic measure. Distributions are based on all strokes observed across the pool of speakers and pellets ( $N = 45736$ ). In the upper-right corner of each panel, a number of distribution descriptors are provided.



negative and positive skews. Kurtosis values that are less than 3 are leptokurtic (lighter tails than the normal distribution), and kurtosis values that are greater than 3 are platykurtic (heavier tails than the normal distribution). The top-left panel plots the distribution for stroke duration. On average, movement strokes are about 138 ms in duration. However, durations can range from about 15 ms to as much as a half a second. About half of the observations fall between 95 and 172 ms. The distribution has a slight positive skew and is heavier tailed than the normal distribution. The top-right panel shows the distribution for stroke distance. The average stroke distance is about 4.7 mm, with half the observed strokes having distances between 1.5 and 6.8 mm. Stroke distance ranges from close to zero to greater than 30 mm. The distribution has a pronounced positive skew and is heavy-tailed. The bottom-left panel shows the distribution for peak speed. The average peak

speed is about 51 mm/s, with half the values falling between 17 and 72 mm/s. Peak speeds can range from close to zero to as much as 400 mm/s. The skewness and kurtosis of the peak speed distribution have values very similar to those derived for the stroke distance distribution. The bottom-right panel plots the boundary speed distribution. The average speed at which a stroke boundary is defined is about 12 mm/s. Boundary speeds can be as small as zero and, perhaps surprisingly, can exceed 100 mm/s. About half the stroke boundary speeds fall between 3 and 15 mm/s. The distribution has an extreme positive skew and is platykurtic.

An important observation that can be made from Figure 5 is that, at the low end of the distribution, there is a plethora of strokes that are extremely short in duration and/or cover very small distances. In fact, the smallest strokes round down to zero mm in length and last about 15 ms. Clearly these very small strokes stretch

our notions about what constitutes a speech movement. The source of these small strokes may be traced to the method employed to identify them. Recall that the segmentation method identifies boundaries based only on direction reversals in the speed history and does not consider the magnitude of such reversals. Therefore, these small strokes indicate the presence of small, short-lived directional changes in the slope of the speed history. Examples of these small strokes may be illustrated in the second and third panels of Figure 4. Arrows identify strokes whose distances fall near the bottom of the group distribution. Although these events meet the operational definition of movement stroke, their importance is unclear.

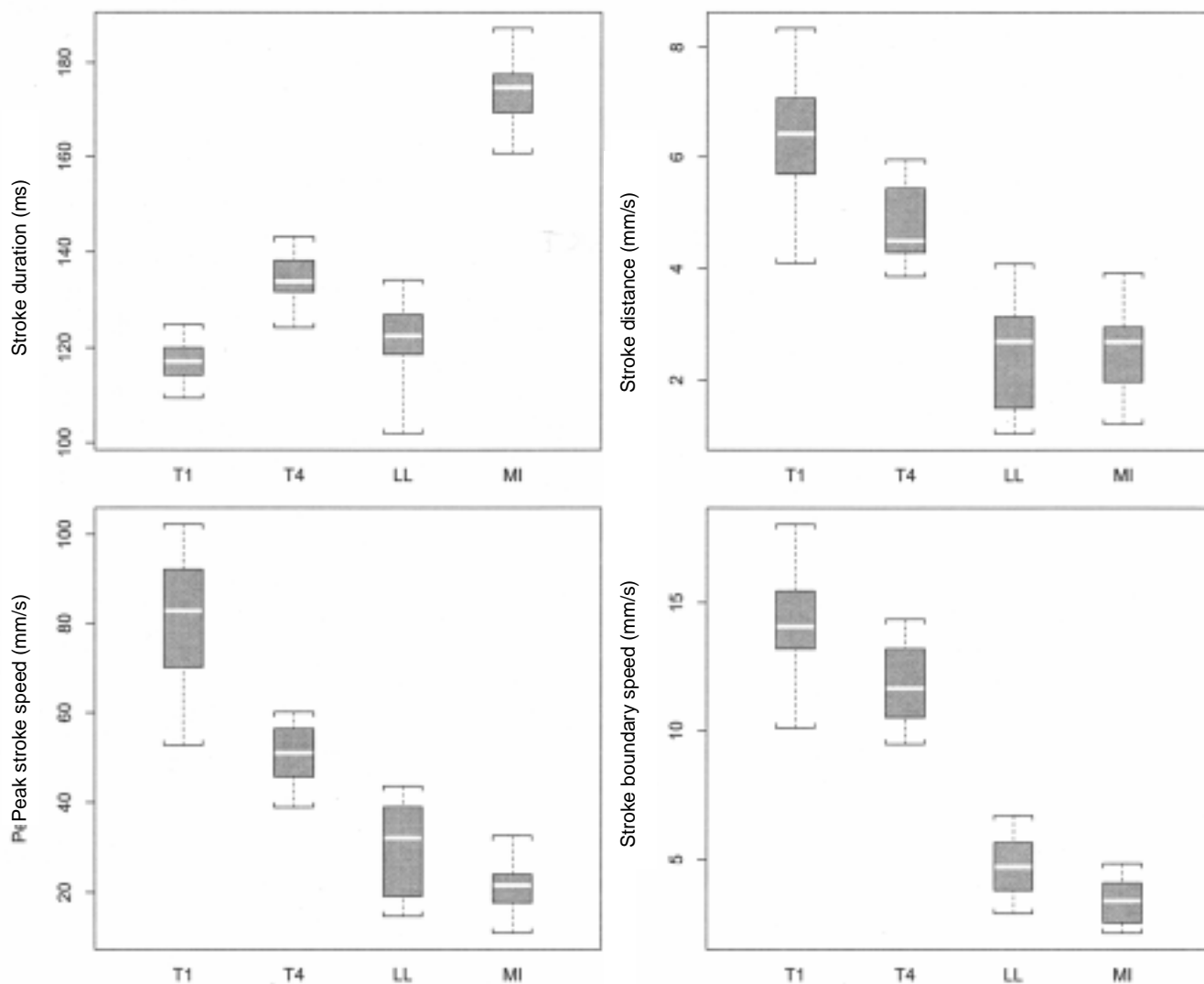
To gain some rudimentary insight into the significance of these events, strokes of an arbitrarily defined small size were identified and located in each pellet's signal stream for a single speaker (S15). The cutoff for evaluation was approximately based on the lower quartile of the distribution of stroke distances for each pellet. Therefore, all strokes less than 2 mm in length were evaluated for T1 and T4, and all strokes smaller than 1 mm in length were evaluated for LL and MI. The temporal boundaries of each stroke were compared to their location relative to the acoustic signal. For T1, about 40 % of the strokes shorter than 2 mm occurred in acoustically silent regions, typically at sentence boundaries. The remaining 60% of these strokes were coincident with speech-related acoustic energy. Of these strokes, almost two thirds occurred in the temporal vicinity of fricative production. The vast majority of these strokes (85%) occurred near fricatives that would be broadly labeled lingual (i.e., /s, z, ʃ, θ, ð/), and about three quarters of these lingual fricatives were alveolar (i.e., /s, z/). The remaining one third of the strokes that occurred during speech were found in a variety of contexts, including, but not limited to, lingual stops, glides, vowels, and diphthongs. For T4, only about 20% of the strokes shorter than 2 mm occurred in acoustically silent regions. Of the remaining strokes, about 60% occurred near fricatives. About three quarters of these strokes were associated with lingual fricatives, and approximately 85% of the lingual fricatives were alveolar. As with T1, the remaining strokes occurred across a number of phonetic contexts. In summary, T1's and T4's small strokes were distributed across very similar acoustic-phonetic environments. This similarity suggests that T1 and T4 might be behaving in kinematically similar ways. This would make sense, given their anatomical linkage. However, evaluation of their temporal locations indicated that only about a third of T1 and T4 strokes overlap temporally, and when they do, there is little similarity in their overall duration. Therefore, it appears that certain phonetic environments (e.g., alveolar fricatives) increase the probab-

ity of very short T1 and T4 strokes, but not necessarily in identical ways for each fleshpoint. For LL, only about 15% of strokes shorter than 1 mm occurred in acoustically silent regions. Of the remaining strokes, fricative contexts again dominated almost half the strokes. A vast majority of the strokes were near lingual fricatives (95%), and about three quarters of these lingual fricatives were alveolar. The remaining small LL strokes were observed in a variety of acoustic-phonetic contexts. MI resembles LL in that only about 15% of strokes measuring less than 1 mm occurred during acoustically silent regions. However, the MI strokes that occurred coincident with speech-related acoustic energy did not appear to cluster disproportionately within one or more acoustic-phonetic contexts. These small MI strokes were typically longer in duration than small T1, T4, and LL strokes and often crossed many acoustic-phonetic boundaries.

To summarize, although our test speaker's small strokes appear in a variety of acoustic-phonetic environments, the T1, T4, and LL strokes appear in substantial proportions within or around fricative segments. Of course, it is uncertain whether this finding is a speaker-specific idiosyncrasy or generalizes to the group. This can be resolved only with a careful microanalysis of these events on a much larger speaker sample and is beyond the scope of this study. We did, however, employ our arbitrary cutoff distances to identify the total number of small strokes in all 18 speakers' data. The number of strokes before and after this truncation for each pellet can be found in parentheses on the right side of Table 1. In general, this population of strokes, although great in number, accounts only for about 5% of the total distance the pellets moved and for about 15% of the total movement time.

Figure 6 shows box and whisker plots of the speaker distribution based on median values for each pellet and kinematic measure. The lower and upper limit of each box, respectively, represents the first and third quartiles of the data. The whiskers represent the range of the data. There are some clear trends in the data based on pellet identity. For the measure *stroke duration*, across the speaker pool T1 and MI, respectively, have the shortest and longest durations. Typical values for T1 range between 110 ms and 125 ms, whereas MI durations range between 160 ms and 190 ms. The T4 distribution lies between T1 and MI, whereas LL durations largely overlap T1 and T4. For the measure *stroke distance*, T1 strokes typically cover distances ranging from 4 mm to 8 mm. T4 stroke distances range from 4 mm to 6 mm. MI and LL stroke distances have largely overlapping distributions, ranging from about 1 mm to 4 mm. For *peak speed*, T1 exhibits the largest peak speeds, with typical values ranging from 50 mm/s to 100 mm/s. T4 peak speeds fall between 40 mm/s and 60 mm/s, whereas LL peak speeds range from 15 mm/s to 45 mm/s. MI peak

**Figure 6.** These box and whisker plots show the distributional characteristics of the speakers' medians for each kinematic measure for the T1 (tongue blade), T4 (tongue dorsum), MI (mandibular incisor), and LL (lower lip) markers. The box defines the interquartile range. The unfilled line in the box marks the group median. The whiskers mark the minimum and maximum values.



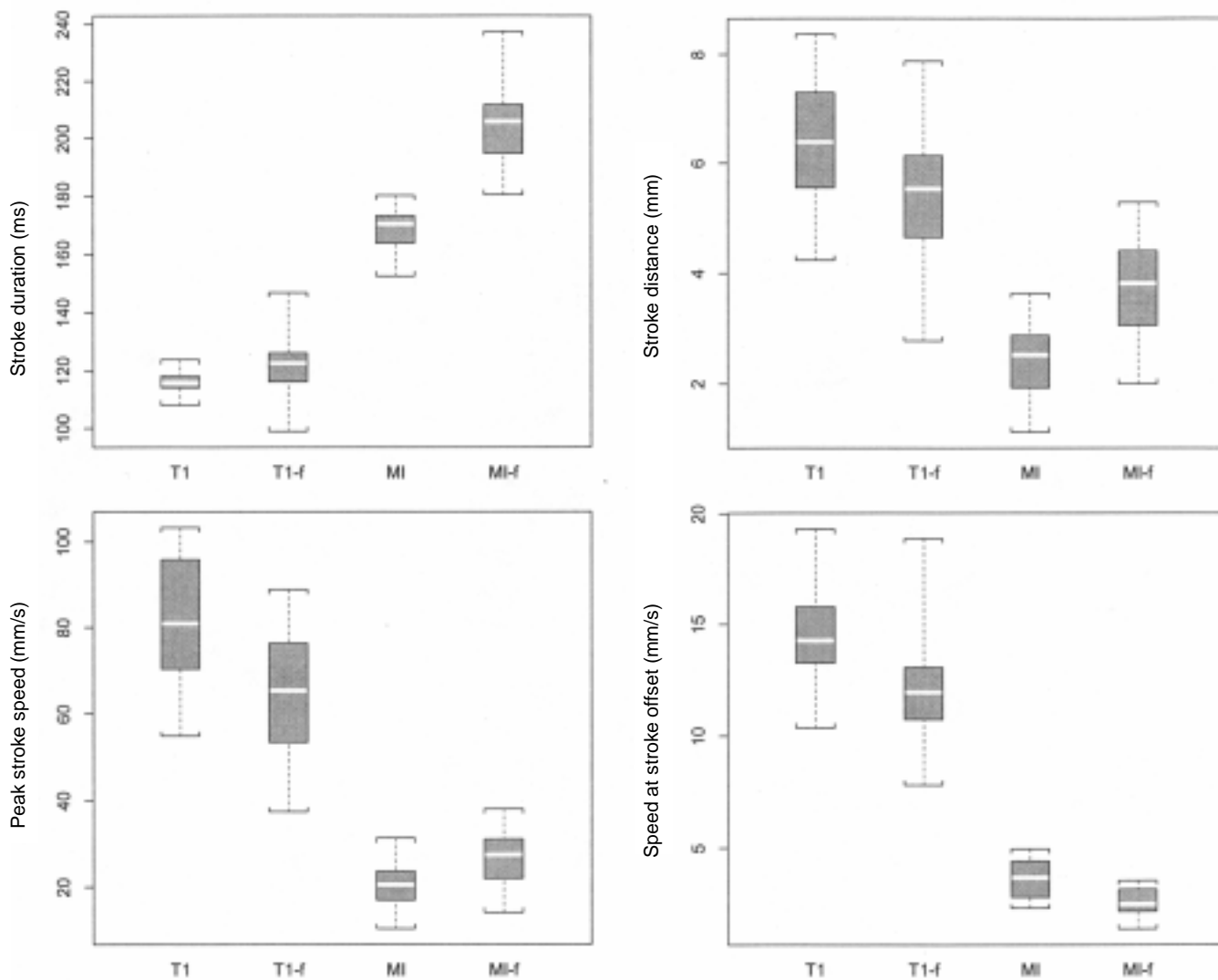
speeds are the smallest, with magnitudes ranging from 10 mm/s to 30 mm/s. *Boundary speeds* generally follow the same trend as peak speed, with T1 and T4 exhibiting larger median values than LL and MI. Differences were evaluated statistically using a series of Wilcoxon signed rank sum tests. Experiment-wise Type I error rate was held to less than 5% using a Bonferroni correction for multiple comparisons ( $p < 0.0001$ ). For stroke duration, all differences were statistically significant, with the exception of the T1-LL comparison. For stroke distance and peak stroke speed, all comparisons were significantly different, with the exception of the LL-MI comparison. Boundary speeds were significantly different for all pair-wise comparisons. To summarize, there are a number of observations worth highlighting. T1 typically traverses the longest distance at the highest

speeds over the shortest durations. Conversely, MI typically moves the shortest distances at the lowest speed over the longest durations. In general, T4 behaves like T1, but with less extreme values. LL is similar to MI for stroke distance and peak and boundary speeds, but has markedly shorter durations.

### ***Kinematic Features of Strokes Associated With the Alveolar Fricative Sound Class***

Figure 7 uses boxplots to compare the distribution of speaker medians of the full data set (less those associated with fricative production) to an alveolar fricative subset for stroke duration, stroke distance, peak stroke speed, and speed at stroke offset across T1 and MI pellets. All

**Figure 7.** These box and whisker plots show the distributional characteristics of the speakers' medians for each kinematic measure for the tongue blade full data set (T1), tongue blade fricative data set (T1-f), the mandibular incisor full data set (MI), and the mandibular incisor fricative data set (MI-f). The box defines the interquartile range. The unfilled line in the box marks the group median. The whiskers mark the minimum and maximum values.



pair-wise comparisons were evaluated using the Wilcoxon signed rank sum tests, and experiment-wise Type I error rate was held to less than 5% using a Bonferroni correction for multiple comparisons ( $p < 0.0001$ ).

We can see in the upper left panel of Figure 7 that, for T1, the average duration of fricative-related strokes does not markedly differ from the average duration of speech strokes in general. However, for MI, fricative-related strokes were significantly longer than speech strokes in general. The average stroke distance and peak stroke speed were significantly reduced for the T1 fricative data set and significantly increased for the MI fricative data set as compared to speech strokes in general. The speed at stroke offset was significantly reduced

for the fricative data set for both pellets (T1 and MI) as compared to speech strokes in general. To summarize, during fricative production the T1 strokes are smaller in amplitude and peak speed, whereas MI strokes are longer in duration and larger in peak speed amplitude as compared to speech strokes in general.

It has been the custom to assume that durations of acoustically defined speech sounds reflect production constraints. As an example, Lehiste (1970) suggested that the shorter durations of alveolar stops, relative to velars or labials, could probably be linked to a greater intrinsic capability for fast movement on the part of the tongue blade relative to the tongue dorsum or lips. This

general line of thought entails the notion that strong relationships should exist between sound durations and features of movements associated with their production. We used the fricative subset to explore this idea, applying multiple correlation to determine the strength and nature of within-speaker associations between fricative durations and the properties of movement strokes in their temporal vicinity. Linear combinations of stroke duration, peak stroke speed, and stroke distance for the T1 and MI pellets were correlated separately with acoustic segment duration, giving two correlation coefficients for each of the 18 speakers. All correlation coefficients were positive, but the proportions of variance accounted for were never large. For T1, for example, the median  $r$  across the speaker group was 0.40, with individual speaker values ranging between 0.24 and 0.59. For MI, correlations were consistently smaller. The median  $r$  across the group was 0.27, with individual speaker values ranging from 0.17 to 0.47. The values of correlation coefficients varied from speaker to speaker and pellet to pellet. Comparisons of  $r$  values for individuals indicated that speakers with weaker correlations for T1 did not necessarily demonstrate weaker correlations for MI. In general, these results would seem to show that it is difficult to accurately predict the durations of acoustically defined sound segments from closely aligned kinematic measures, though it is important to remember that it is no easy matter to determine which movement strokes, in which articulatory kinematic waveform(s), should be matched against sounds in the speech pressure waveform. Thus, the value of this result may be uncertain. However, what is not uncertain is the fact that the analysis itself would not be tractable without an explicit procedure for parsing kinematic waveforms into strokes.

---

## Discussion

This report describes an approach for parsing kinematic signal streams into movement units. The approach is probably most suitable for data derived from point-tracking techniques such as electromagnetic articulography, reflective/infrared video, Selspot, or x-ray microbeam. We believe there are some distinct advantages associated with this parsing method. First, movement strokes identified by the method match our intuitions about what a stroke ought to be—in short, a single period of acceleration followed by a single period of deceleration. In the general movement literature, intervals with one local maximum in speed are often considered to form the class of skilled movements (cf. Nelson, 1983). A second advantage of the method is that its strokes can be defined without reference to any terms external to the geometry of bodily movements. No appeals to other, notoriously difficult, “speech units”

such as targets, transitions, sounds, syllables, or the like are necessary. Thus, the approach can be applied to all types of motor tasks (e.g., typing, manual sign, developing movements of the limbs or masticatory apparatus, singing) and is essentially independent of any assumptions about goals associated with such tasks. This lack of dependence on task goals makes this method well suited for comparing motor behavior across different structures and tasks in health and disease. For practical reasons, we have focused our own study on articulatory movements generated during oral reading. A final advantage of the parsing method is that it is well suited to automated stroke identification. In the future, stroke inventories might be generated quickly and efficiently from large samples of movements and movers, allowing for large-scale comparisons across different disease conditions, dialects, languages, tasks, and performer characteristics. For example, we might expect certain dysarthric populations to exhibit patterns in their stroke inventory that are observed regardless of task, whereas other dysarthric populations may only be distinguished on more task-specific inventories.

It is unlikely that movement strokes, as we've described them, map directly onto abstract speech production “units” or “targets,” however they might be defined. Interpreting kinematic movement patterns in terms of underlying control is difficult without a sophisticated understanding of the interaction between neuromotor activity and the peripheral biomechanical environment that results in movement. Biologically plausible vocal tract models and well-articulated speech production theories are critical to such interpretation. However, we contend that a methodological convention (kinematic, acoustic, or otherwise) that makes few assumptions about the underlying organization of speech is still a profitable endeavor because it focuses on developing a reliable, organized, descriptive account of the speech behavior. This is consistent with the view that understanding the cause of variation within any phenomenon should begin by first describing the variation itself (Fisher, 1925).

When we apply the parsing method to speech kinematic data, we gain new insights that are general to such behavior and specific to some sounds and sound classes. Insights of the first type relate to stroke features that are most typical for the lower jaw, tongue blade and dorsum, and lower lip. For example, jaw strokes, as a class, tend to be relatively longer in time than those of the lower lip or parts of the tongue. Tongue blade and dorsum strokes tend to be relatively longer in distance, and because they have relatively short durations are also relatively faster in terms of peak and boundary speeds. Lower-lip strokes tend to be more tongue-like in their durations, but jaw-like in their distances and peak and boundary speeds.

New insights specific to the class of alveolar fricatives include the fact that jaw strokes in their vicinity tend to be relatively longer in time and distance than the grand average computed across the entire speech task. At the same time, associated tongue blade strokes tend to be relatively shorter in time and distance. Finally, data from one of our speakers (the only one analyzed in this way) seem to show that relatively many unusually small articulatory strokes occur near fricatives. These last strokes may represent fine-tuning movements, important for narrow constrictions that are acoustically sensitive to small changes in their dimensions (Stevens, 1989). Each of these general observations about fricatives reflects a classical problem in the study of speech movement, and that is to understand the relationship between movement and sound. The usual way to attack this problem is to locate some sound or set of sounds in the speech pressure waveform and then ask what movements occur in their neighborhood. But clearly, we can mirror this discovery path as we did in our exploration of the class of very small strokes, first by locating movements of a type and then asking what sounds usually occur in their neighborhood. Just as we might ask of individuals, "How do you sign your name?" we can also ask of signatures, "Who wrote this?" The point is the same: to pair up writers and things written, movements and sounds made.

At least two limitations of the parsing method described in this report should be acknowledged. The first relates to the fact that the method is suitable only for position histories of points moving along lines, within planes, or in three-dimensional spaces. There is no obvious way to apply the method to the movement of, say, a rigid body (e.g., the lower jaw) translating *and* rotating in a plane or three-dimensional space. We cannot reduce such "complicated" motion to the coordinate histories of a single point and, hence, cannot derive the scalar magnitude of the first time-derivative of position for any discrete moment during the motion. This problem is clearly worse if our interest extends to "whole-body" movements of solid but plastic objects such as the tongue or soft palate. All movements analyzed for this study really must be understood to represent points on articulators, and not the articulators themselves, commonly conceived as whole, functional bodies.

The second limitation of the method has to do with potential errors in detection of beginning and endpoints of strokes. We locate these at successive local minima in the speed history, where markers are moving most slowly and the signal-to-noise ratio is lowest. Stroke boundary identification errors due to noise in the original estimates of marker positions will affect both stroke duration and extent, and the former effect would probably be the larger of the two. One possible "solution" to this

problem could involve restricting stroke boundaries to regions where certain thresholds on speed, or the rate at which it changes, are exceeded. This type of restriction would yield some "complex" strokes, with more than one local maximum in speed. Whether strokes of this type are appropriate or useful is an issue for future study.

The parsing method we describe for making boundary decisions for speech strokes is conceptually similar to the velocity-zero-crossing method that has been applied to one-dimensional data for at least the last 20 years (Adams et al., 1993; Hertrich & Ackermann, 2000; Kelso et al., 1985; Löfqvist & Gracco, 1994; Ostry & Munhall, 1985). Our method differs from the earlier standard, incorporating both dimensions of fleshpoint trajectories in the midsagittal plane, and in a way that is insensitive to the imposition of reference axes on such data. Thus, an important advantage of our method is that it does not entail an arbitrary choice of one dimension as more important than any other. The notion that our method is an improvement on some earlier standard should probably not signal the end of the developmental line for kinematic parsing methods. Boundary decisions might also be made using something other than, or in addition to, local minima in the speed history. A likely candidate for such a criterion is the (instantaneous) curvature of the trajectory. Curvature is defined as the scalar change in a curve's direction with respect to arc length (Hurley, 1980). We noted earlier that local speed minima are usually associated with abrupt changes in the direction of point trajectories (i.e., periods of high curvature). An empirical relationship between the scalar quantities of speed and curvature has been established for limb movement (Viviani & Cenzato, 1985), oculomotor movement (de'Sperati & Viviani, 1997), and recently for speech-related orofacial movement (Tasko & Westbury, 2001). This relation between speed and curvature has been used to segment drawing movements into hypothetical units of action. Laying speech stroke boundaries at locations where speed is low *and* curvature is high may yield strokes that are more intuitively and theoretically (Viviani & Cenzato, 1985) appealing than strokes defined using speed alone. This is a topic we plan to pursue in the future.

## Acknowledgments

This research was supported by NIH grants DC00820 and DC03723 and served as part of a doctoral dissertation completed at the University of Wisconsin. Portions of data analysis and manuscript preparation was supported by NIH grant DC03659. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

## References

- Adams, S. G., Weismer, G., & Kent, R. D.** (1993). Speaking rate and speech movement velocity profiles. *Journal of Speech and Hearing Research, 36*, 41–54.
- Browman, C. P., & Goldstein, L.** (1992). Articulatory phonology: An overview. *Phonetica, 49*, 155–180.
- Crystal, T. H., & House, A. S.** (1982). Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America, 72*, 705–716.
- De'Sperati, C., & Viviani, P.** (1997). The relationship between curvature and velocity in two-dimensional smooth pursuit eye movements. *Journal of Neuroscience, 17*, 3932–3945.
- Fisher, R. A.** (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
- Fitts, P. M.** (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology, 47*, 381–391.
- Fowler, C. A.** (1977). *Timing control in speech production*. Bloomington: Indiana University Linguistics Club.
- Harris, K. S., Tuller, B., & Kelso, J. A. S.** (1986). Temporal invariance in the production of speech. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 243–267). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hertrich, I., & Ackermann, H.** (2000). Lip-jaw and tongue-jaw coordination during rate-controlled syllable repetitions. *Journal of the Acoustical Society of America, 107*, 2236–2247.
- Hurley, J. F.** (1980). *Intermediate calculus*. Philadelphia: Saunders College.
- Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E. L., & 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.
- Kent, R. D.** (1984). Stuttering as a temporal programming disorder. In R. F. Curlee & W. H. Perkins (Eds.), *Nature and treatment of stuttering: New directions* (pp. 283–301). San Diego, CA: College-Hill Press.
- Kent, R. D., & Minifie, F. D.** (1977). Coarticulation in recent speech production models. *Journal of Phonetics, 5*, 115–133.
- Kleinbaum, D. G., Kupper, L. L., & Muller, K. E.** (1988). *Applied regression and other multivariate methods*. Belmont, CA: Duxbury Press.
- Lashley, K. S.** (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). New York: Wiley.
- Lehiste, I.** (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.
- Löfqvist, A., & Gracco, V. L.** (1994). Tongue body kinematics in velar stop production: Influences of consonant voicing and vowel context. *Phonetica, 51*, 52–67.
- Löfqvist, A., & Gracco, V. L.** (1997). Lip and jaw kinematics in bilabial stop consonant production. *Journal of Speech, Language, and Hearing Research, 40*, 877–893.
- MacNeilage, P. F.** (1970). Motor control of serial ordering in speech. *Psychological Review, 77*, 182–196.
- Mermelstein, P.** (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America, 53*, 1070–1082.
- Nelson, W. L.** (1983). Physical principles for economies of skilled movements. *Biological Cybernetics, 46*, 135–147.
- Olive, J. P., Greenwood, A., & Coleman, J.** (1993). *Acoustics of American English speech*. New York: Springer-Verlag.
- Ostry, D., & Munhall, K.** (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America, 77*, 640–648.
- Stevens, K. N.** (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3–45.
- Tasko, S. M., & Westbury, J. R.** (2001). *Kinematics and the geometry of speech movement*. Manuscript submitted for publication.
- Viviani, P., & Cenzato, M.** (1985). Segmentation and coupling in complex movements. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 828–845.
- Weismer, G., Kent, R. D., Hodge, M., & Martin, R.** (1988). The acoustic signature for intelligibility test words. *Journal of the Acoustical Society of America, 84*, 1281–1291.
- Westbury, J. R.** (1994). *X-ray microbeam speech production database user's handbook*. Madison: University of Wisconsin at Madison, Waisman Center.
- Westbury, J. R., Sevenson, E. J., & Lindstrom, M. J.** (2000). Kinematic event patterns in speech: Special problems. *Language and Speech, 43*, 403–428.

---

Received April 5, 2001

Accepted November 20, 2001

DOI: 10.1044/1092-4388(2002/010)

Contact author: Stephen M. Tasko, PhD, Department of Speech Pathology and Audiology, Western Michigan University, Kalamazoo, MI 49008. E-mail: stephen.tasko@umich.edu