1065

AMERICAN
SPEECH-LANGUAGE-
HEARING
ASSOCIATION

JSLHR, Volume 40, 1111–1117, October 1997

On the Registration of Time and the Patterning of Speech Movements

Jorge C. Lucero
Kevin G. Munhall
Queen's University
Kingston, Canada

Vincent L. Gracco
Haskins Laboratories
New Haven, CT

James O. Ramsay
McGill University
Montreal, Canada

In order to study speech coordination we frequently average kinematic and other physiological signals. The averages are assumed to be more representative of the underlying patterns of production than individual records. In this note we outline different approaches to averaging and present a new nonlinear normalization technique that offers better information than ensemble averaging, linear normalization, or feature alignment methods. We suggest that this technique provides a clear estimation of pattern shape while preserving information on the variation over time.

KEY WORDS: time registration, averaging, speech movement, lip motion

In speech production research we commonly use waveform averaging for one of three general purposes: (1) to estimate the latency of a response to a defined stimulus, (2) to assess the central tendency and variability in movement trajectories in an experimental condition, and (3) to infer aspects of the control regime from trajectory shapes. The most common approach is to compute sample-by-sample averages from a single line-up event (e.g., ensemble averaging of responses to some phasic stimulation: Gracco & Abbs, 1985; Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984). We will refer to this approach as *un-normalized averaging*. A second approach involves defining two line-up points that span an event. Before averaging, the waveforms are linearly stretched or compressed to a common length (e.g., Smith, Goffman, Zelaznik, Ying, & McGillem, 1995). We will refer to this approach as *linearly normalized averaging*. A final approach involves defining two line-up points that span an event and stretching the data in a nonlinear fashion to a common length (e.g., Ramsay & Silverman, in press; Strik & Boves, 1991). In this approach, key events are registered by nonlinearly warping the trajectories. We will refer to this approach as *nonlinearly normalized averaging*.

Each of these approaches has advantages and disadvantages. The major disadvantage of un-normalized averaging is distortion due to variations in the time domain: the greater the variability in the component responses, the more likely that the average signal will not resemble the individual signals much beyond the line-up point. For example, averaging variable speech movement events that occur over more than about 50–100 ms will often result in a representation of the central tendency that in no way reflects any of the component waveforms. Un-normalized averaging has the major advantage that physical time is preserved, and

1092-4388/97/4005-1111

estimates of the timing relationship between a stimulus and a time-locked response are preserved. Other advantages include simplicity of processing and simplicity of interpretation of the average signal.

Linearly normalized averaging reduces the effects of differences in event duration but does not eliminate distortion due to nonlinear factors. For example, if the waveforms being averaged vary in duration in a nonuniform manner (e.g., through the insertion of pauses in an utterance or different combinations of vowels and consonants), the resulting normalized signals will retain variability in the timing of landmarks. The resultant variation represents a possible combination of factors, ranging from the phonetic composition of the speaking material to variations in the control process. Thus, unless the behavior being studied is linearly scaled in time, the interpretation of the resulting average signal is ambiguous. For short stretches of behavior the assumption of linear scaling may be acceptable. Under these conditions, a computationally simple and reasonably accurate average can be derived using this approach.

Nonlinearly normalized averaging overcomes the problem of variability in the timing of landmarks and produces an average that can be used to study the underlying spatiotemporal patterning or shape. The precise definition of *shape* is in itself a complex issue (e.g., Bookstein, 1991). Here, we simply mean the number and magnitude of key events (landmarks) in the data, such as peaks, cycles, and so forth. Distortion is avoided by aligning those events before averaging and thus minimizing phase differences. Variations in physical time are removed from the average but are available for examination in a separate measure derived from the alignment process.

Below we will illustrate the three methods of averaging using lip acceleration data, and we will describe in detail one technique for nonlinearly normalized averaging. We chose to process the lip acceleration instead of its displacement, because acceleration is directly related to the forces acting on the lips and thus is worth analyzing to study how lip motion is controlled. Moreover, because it has a more complex shape than displacement and velocity, nonlinearly normalized averaging will be more revealing of the differences between the various averaging techniques.

Description of the Data

Data was collected from a male native-speaker of English with no reported speech disorders, producing the sentence "Buy Bobby a puppy" in 20 trials at a slow rate. One trial was lost because of a recording error, leaving 19 trials for analysis. An OPTOTRAK (Model 3010, Northern Digital Inc.) system was used to transduce the three dimensional motion of an infrared emitting diode placed on the lower lip in the midline, digitized (12-bit resolution) at a sampling frequency of 200 Hz. The data were corrected for motion of the head and transformed to a coordinate system in which the origin is the incisor cusp and the horizontal and protrusion axes lie along the bite surface (Ramsay, Munhall, Gracco, & Ostry, 1996). For the present analysis, only the vertical component of lower lip motion was examined.

All further signal processing was done using Matlab. The raw data were filtered in the forward and backward directions using a low-pass fifth-order Butterworth filter with a 15-Hz cutoff frequency. Each record (trial) was then differentiated by evaluating the first differences to compute velocity, and the segment between the peak velocity of the first and last opening movement was extracted (see Smith et al., 1995). For each extracted record, the acceleration (second difference) was computed for further processing. The data were already reasonably smooth so that a good estimate of the acceleration was obtained. In situations where further smoothing is required, an algorithm such as spline smoothing with a roughness penalty in the fourth derivative (Ramsay et al., 1996) may be used. In the present study, application of this algorithm produced no appreciable difference in the resultant signals. The trials differed in length; the final number of data points for each acceleration record ranged from 348 (1.74 s) to 409 (2.04 s), with the mean at 368.47 (1.84 s).

Figure 1 shows a typical record (displacement, velocity, and acceleration) as a reference for the following analysis.

Computation of Averages

Un-Normalized Averaging

Figure 2 (a) shows the acceleration waveforms aligned at the start of each signal (i.e., the peak velocity of the first opening movement). The average shown in Figure 2 (b) was computed simply by taking the average point by point. Differences in the timing of landmarks (e.g., peak accelerations) may be analyzed with this technique. However, the average does not resemble the individual waveforms, and it becomes increasingly poorer as time increases because of a cumulative distortion.

Linearly Normalized Averaging

We examined two methods of linearly normalized averaging: resampling and Fourier series resynthesis (Smith et al., 1995). We will present the results only

Figure 1. Typical data record: (a) displacement (mm), (b) velocity (mm/s), and (c) acceleration (mm/s²) (time in s).

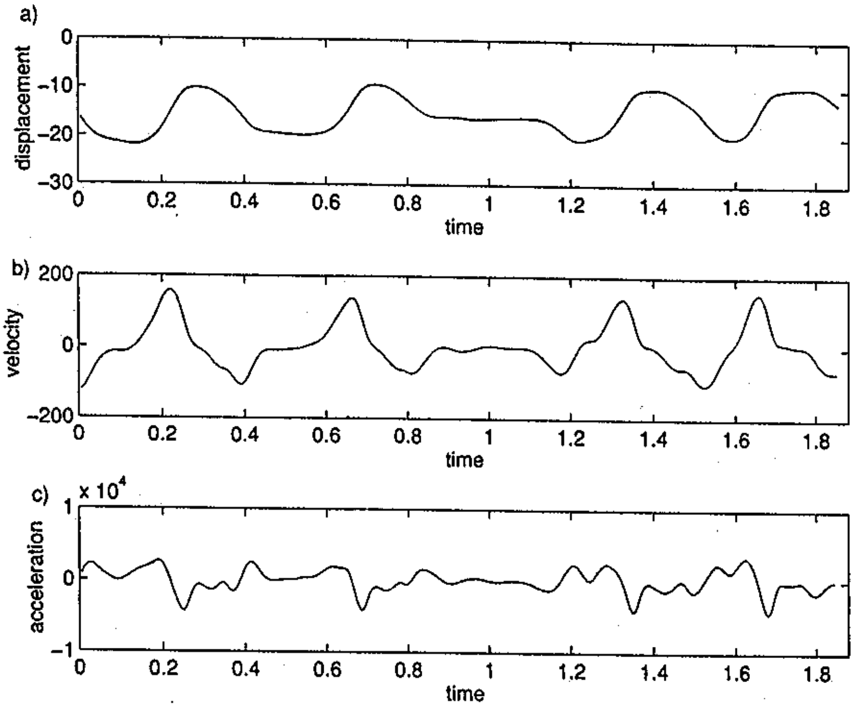
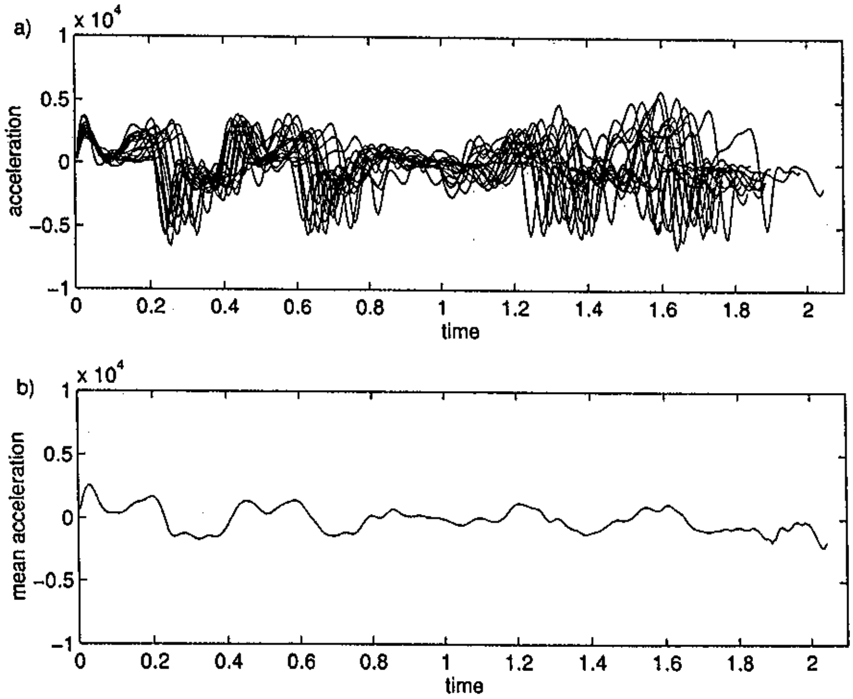


Figure 2. Un-normalized averaging: (a) un-normalized records, (b) average (acceleration in mm/s² and time in s).



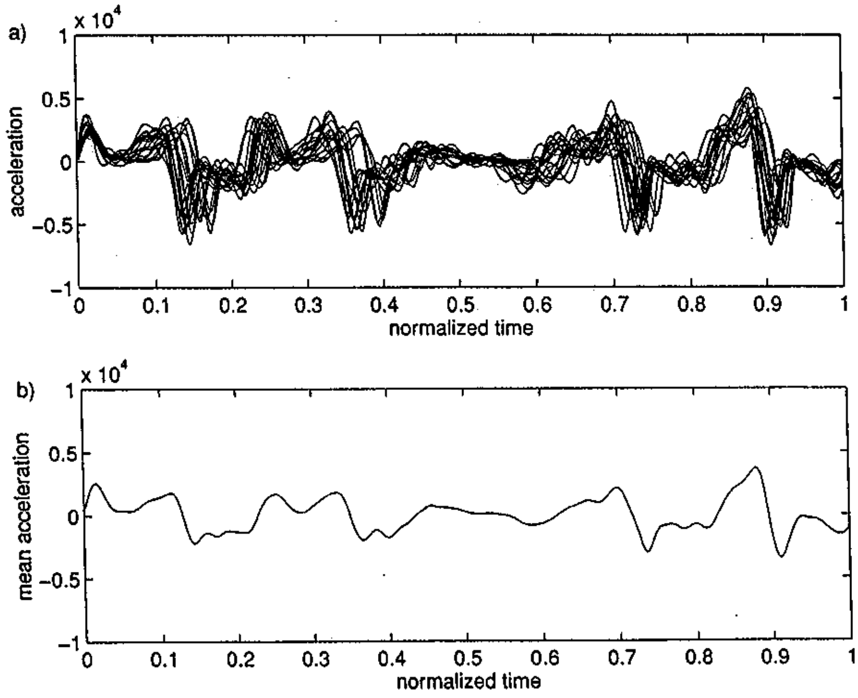
for the resampling technique, but the findings apply equally to both methods.[1]

For the resampling method, all the records were interpolated using cubic splines to a common length of 500 points, and an artificial time span of 0 to 1 s. was adopted for the interpolated records. Figure 3 shows the resampled waveforms and the average based on the resampled approach. We can see that the average is closer in shape to the original waveforms, compared with the previous un-normalized approach. However, we can also note some significant differences. Consider, for instance, the negative peaks in the individual waveforms, which have a much lower amplitude in the resultant average. In both the resampled average, Figure 3 (b), and the average derived from the Fourier series resynthesis (Smith et al., 1995), the major source of distortion is phase variability. The waveforms in Figure 3 (a) are slightly out of phase because of nonuniform timing changes in some trials. Thus, when the records are combined in an average, new "shapes" are created that do not exist in any of the individual records.

Nonlinearly Normalized Averaging

In Figure 3, the differences between the average and the individual waveforms are caused by variations in the timing of landmarks. A better way to obtain an average that reveals the patterning of events is to distort the time scale nonlinearly, so that landmarks become aligned. One simple technique could be to first identify important landmarks in each of the records—typically peaks or zero crossings—and then distort the time scale of each record to align the landmarks. This is a piecewise linear technique, because the time scale is stretched or compressed linearly between landmarks. However, this technique has some drawbacks. We must first decide which landmarks are important to align, and some of the selected landmarks may be missing in some of the curves. These problems can be overcome by defining a suitable fitting criterion for the alignment and by distorting the time scale nonlinearly so as to optimize

Figure 3. Linearly normalized averaging using spline interpolation: (a) interpolated records, (b) average (acceleration in mm/s²).



[1] For the Fourier series resynthesis we followed the procedure of Smith et al. (1995). As in that work, each record was expanded into a Fourier series with 10 components. The coefficients for the Fourier series were obtained using the discrete Fourier transform. For the acceleration data being analyzed 10 components were too few, and the resulting waveforms were far too smooth. This problem is trivial and could be corrected by using more components in the resynthesis. An additional problem of the Smith et al. technique is more serious. Taking a constant number of components of the Fourier series filters the records at different cutoff frequencies, depending on their duration. For example, using 10 components, a record from a fast speaking condition that is 0.7 s long will be filtered with a cutoff frequency of 14.0 Hz, whereas a record from a slow speaking condition that is 1.85 s long would be filtered with a cutoff frequency of 5.4 Hz. This will be a problem particularly when sentences produced at different rates are analyzed (as in Smith et al., 1995), because the degree of smoothing introduced for the various rates can be different.

that criterion. We describe here a new algorithm for nonlinearly normalized averaging that follows this approach (Ramsay & Silverman, 1997).

First we will develop the algorithm in the general case, and then we will apply it to the lip movement data. Let us call the set of data records for which we want an average (the acceleration records in our case) $x_i(t)$, with $i = 1,\ldots, N$, and N is the number of records, and their average $\bar{x}(t)$. All the records are assumed to have the same number of data points, from $t = 0$ to $t = 1$. For simplicity, we consider each data record as a continuous function.

We want to determine a strictly increasing and reasonably smooth transformation of time $h_i(t)$ (warping function) for each record $x_i(t)$, such that each record as function of the transformed time (registered record)

$$x_i^*(t) = x_i[h_i(t)] \tag{1}$$

is close in some measure to the average $\bar{x}(t)$. Such warping functions can be described by the homogeneous differential equation

$$\frac{d^2h_i}{dt^2}(t) = w_i(t)\frac{dh_i}{dt}(t) \tag{2}$$

for some suitable function $w_i(t)$. Note that a strictly monotone function has a nonzero first derivative, and we may consider the existence of the second derivative as a requisite for "reasonable smoothness." Thus, $w_i(t)$ is trivially the relation of the second derivative [curvature of $h_i(t)$] to the first derivative [slope of $h_i(t)$], that is, the relative curvature of $h_i(t)$. It defines $h_i(t)$, because the solution of Eq. (2) is

$$h_i(t) = C_0 + C_1\int_0^t\left[\exp\int_0^u w_i(v)dv\right]du \tag{3}$$

The coefficients C_0 and C_1 are selected so that the warping function satisfies $h_i(0) = 0$ and $h_i(1) = 1$.

The closeness of the registered records to the average may be evaluated through the measure

$$F(x_i;w_i) = \int_0^1\left[\bar{x}(t) - x_i^*(t)\right]^2dt \tag{4}$$

Thus, the warping functions are evaluated to minimize Eq. (4). After a set of warping functions has been computed, the average of the registered records $x_i^*(t)$ may be used in the place of $\bar{x}(t)$ to compute a new set of warping functions, and this process may be iterated until there is no significant change between two consecutively calculated sets of warping functions (or averages).

The optimizing criterion (4) can be extended by including a penalty for the roughness of the warping functions, as follows

$$F_\lambda(x_i;w_i) = F(x_i;w_i) + \lambda\int_0^1 w_i^2(t)dt \tag{5}$$

Note that a large value of the smoothing parameter λ will result in small values of the curvature of $h_i(t)$, which will approach the straight line $f(t) = t$.

In our calculations, first we resampled all the records to a common length of 500 data points, as in linearly normalized averaging. Then, we estimated the warping functions by expanding the function $\int_0^u w_i(v)dv$ in Eq. (3) in a linear basis, using the following technique (Ramsay & Silverman, 1997). We divide the time interval [0, 1] with a set of break points τ_k, $k = 0,\ldots, K$ satisfying $0 = \tau_0 < \tau_1 <\ldots< \tau_K = 1$. Next, we define the hat function

$$\Phi_k(t) = \begin{cases}(t-\tau_{k-1})/(\tau_k-\tau_{k-1}) \text{ if } t\in[\tau_{k-1},\tau_k]\\(\tau_{k+1}-t)/(\tau_{k+1}-\tau_k) \text{ if } t\in[\tau_k,\tau_{k+1}]\\0 \text{ otherwise}\end{cases} \tag{6}$$

for $k = 1,\ldots, K$ (for $k = K$, $\Phi_k(t)$ is not defined in the interval $[t_k, t_k + 1]$).

$$\int_0^t w(v)dv = \sum_{k=1}^{K}c_k\Phi_k(t) \tag{7}$$

where c_k are coefficients to determine so as to minimize the optimizing criteria. In our calculations, we used 11 equally spaced breakpoints ($K = 10$) and integrated Eq. (3) by a simple trapezoidal rule. The initial values of coefficients c_k were set to zero. We used Eq. (5) with $\lambda = 0.001$ to evaluate the optimal warping functions and iterated the calculations until the maximum difference between two consecutive averages was less than 2% (9 iterations). A sequential quadratic programming method implemented in Matlab was used for the optimization. At this point, there was no visual difference between consecutive averages.

Figure 4 (a and b) show the waveforms for the registered records and the average, respectively. Major events in the original waveforms have been aligned, and the average clearly preserves their shape.

Figure 5 (a) shows the warping functions $h_i(t)$, and (b) shows the difference between the warping functions and the straight line $f(t) = t$, that is, the curves show how much the transformed time departs from the normalized (resampled) time. They describe the differences between the timings of the unregistered waveforms and the computed average. Note that they allow us to analyze separately variability in the timing of events and variability in their magnitude. For instance, following the same approach as Smith et al. (1995), we may compute the standard deviation across the registered records at each data point and use their sum as an index of magnitude variability. This calculation may be applied to the warping functions, obtaining an index of phasing variability. Timing variability may be assessed separately by examining changes in the duration of the unnormalized records.

Figure 4. Nonlinearly normalized averaging: (a) registered records, (b) average (acceleration in mm/s²).

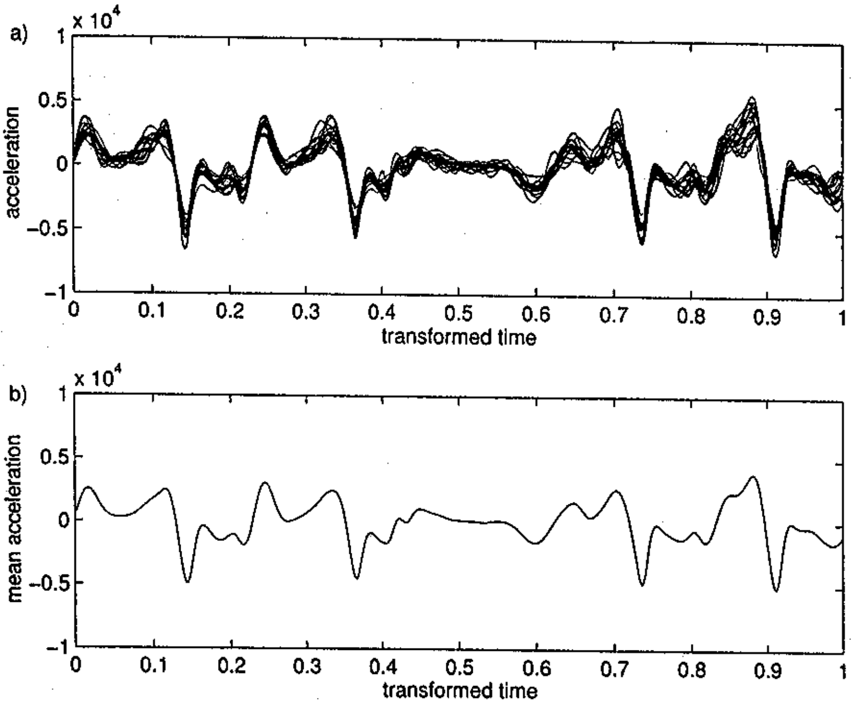
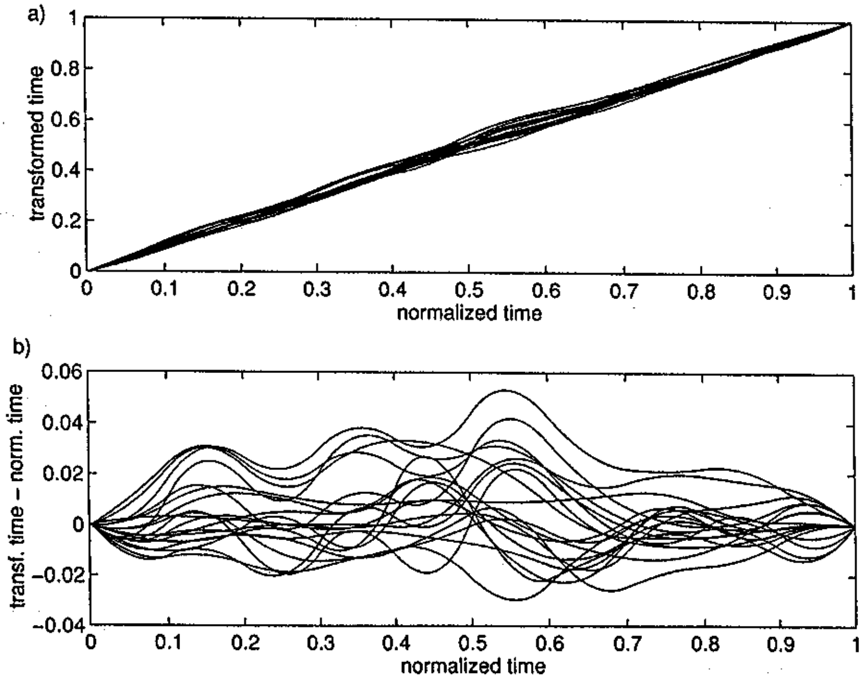


Figure 5. (a) Warping functions $h_i(t)$ for the registered records in Figure 4 (a); (b) difference between the warping functions and the normalized (resampled) time $f(t)$.



A similar approach has been used previously by Strik and Boves (1991) to average various physiological waveforms related to speech, using a dynamic programming algorithm to estimate the optimal warping functions. Although the basic idea is the same, their algorithm results in nondifferentiable warping functions and requires users to select one of the records as a reference for the registration (e.g., the record with median length, if available) instead of working directly with the average.

Discussion

Addressing issues in the neural control of speech often requires substantial inference because the sensorimotor processes that generate the movements and associated muscular actions are unobservable. An important adjunct to experimental design and improved imaging technologies is the use of powerful analytical techniques to uncover important aspects of the speech motor control process. We have presented an approach to data visualization that provides a clear advantage over simple movement averaging and linear normalization. The rationale is that if sequential movements reflect a patterning process the only way to understand the process and identify the pattern is to faithfully reconstruct the pattern from the observations. Nonlinearly normalized averaging preserves the shape of movement patterns while retaining the option to investigate variability in their production. Variations in absolute time can be easily obtained using simple and standard measures, while nonlinearities in relative time are represented in the warping function. Note also that it is easy to modify the present technique using optimizing criteria other than Eq. (5), such as adding a term for roughness penalty if some degree of smoothing is desired or using derivatives (or weighted combinations of them) in the first integral if registration at different levels is desired (Ramsay & Silverman, in press).

Although the nonlinearly normalized averaging seems well suited to the study of the shape of trajectories, it has limitations. This technique tends to align the landmarks with larger magnitude over the smaller ones. Fortunately, the larger landmarks are usually the most important ones, but this might not necessarily always be so. Also, nonlinearly normalized averaging implicitly assumes that the individual waveforms have similar shape. If large differences in shape are present, for example, different number of major cycles, then the registration might not work properly. Finally, the choice of what method to use to average physiological data obviously depends on the purpose of the research. However, any form of normalization sacrifices the representation of physical time in order to sharpen the visualization of the patterns.

Acknowledgments

We are grateful to A. Smith and G. Ying for generously providing their time normalization software, and to L. Coady for her help in the collection of experimental data. This research was funded by NIH grant DC-00594 from the National Institute on Deafness and Other Communication Disorders and NSERC.

References

Bookstein, F. L. (1991). *Morphometric tools for landmark data: Geometry and biology.* Cambridge, UK: Cambridge University Press.

Gracco, V., & Abbs, J. (1985). Dynamic control of the perioral system during speech: Kinematic analyses of autogenic and nonautogenic sensorimotor processes. *Journal of Neurophysiology 54*, 418–432.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance 10*, 812–832.

Ramsay, J. O., Munhall, K. G., Gracco, V. L., & Ostry, D. J. (1996). Functional data analyses of lip motion. *Journal of the Acoustical Society of America 99*, 3718–3727.

Ramsay, J. O., & Silverman, B. W. (in press). *Functional data analysis*. New York: Springer-Verlag.

Smith, A., Goffman, L., Zelaznik, H. N., Ying, G., & McGillem, C. (1995). Spatiotemporal stability and patterning of speech movement sequences. *Experimental Brain Research, 104*, 493–501.

Strik, H., & Boves, L. (1991). A dynamic programming algorithm for time-aligning and averaging physiological signals related to speech. *Journal of Phonetics, 19*, 367–378.

Received January 22, 1997

Accepted April 28, 1997

Contact author: Jorge C. Lucero, Department of Psychology, Queen's University, Kingston, Ontario, Canada K7L 3N6. Email: lucero@pavlov.psyc.queensu.ca