Measures of articulatory variability in VCV sequences

Jorge C. Lucero
Department of Mathematics, University of Brasilia, Brasilia DF 70910-900, Brazil
lucero@mat.unb.br

Anders Lö̈fqvist
Haskins Laboratories, 270 Crown Street, New Haven, CT 06511,
and Lund University
lofquist@haskins.yale.edu

Abstract: Functional data analysis is used to examine articulatory variability across repetitions in normal speech, under different movement constraints. A temporal normalization technique is applied to align trajectories of lips, jaw, and tongue in vowel-consonant-vowel sequences. Next, an index of amplitude variability is computed, defined as the mean standard deviation between peak velocities of the consonantal closure by the active articulator, in each VCV sequence. The results show that articulatory variability varies as a function of both the phonetic requirements of the consonant and the biomechanical characteristics of the articulatory structures involved.

PACS numbers: 43.70.Aj, 43.70.Jt, 43.72.Ar [DO]
Date Received: November 15, 2004 **Date Accepted:** January 20, 2005

1. Introduction

The issue of constancy and variability in speech production has been the subject of much experimental and theoretical work (e.g., Adams *et al.*, 1993; Gracco and Abbs, 1985; Lucero, 2004; Lucero *et al.*, 1997; Sharkey and Folkins, 1985; Smith and Goffman, 1998; Smith *et al.*, 1995; Smith, 1995; Ward and Arnfield, 2001). Although it appears that variability is larger in children than in adults, and also in speakers with neurological disorders compared to normal subjects, the degree of variability in normal speech articulation is largely unknown. The same is true for the potential articulatory, acoustic, and perceptual mechanisms governing constancy and variability.

In this paper, we examine articulatory variability across repetitions in normal speech. Since articulator trajectories are time-varying continuous curves, computing measures of variability presents the difficulty that, theoretically, an infinite (continuous) amount of information is required to describe the trajectory shapes. This problem can be adequately treated by a functional data analysis [FDA; Ramsay and Silverman (1997)] approach. FDA constitutes a set of analytical tools to explore patterns and variability in sets of data obtained from observations of a repeated physical process. Although data is recorded discretely, FDA assumes that such data may be described by smooth functions of time which may be evaluated at any particular instant of time. The main advantage of this approach is that it takes account of the underlying continuity of the physiological system generating the data, and thus it may capture temporal relations in the data owing to this continuity (Ramsay and Silverman, 1997).

FDA techniques and applications to speech analysis were first introduced by Ramsay *et al.* (1996). There, data smoothing and functional principal component analysis were applied to lip trajectories during the production of utterances, to determine the main modes of variation of lip motion. Later, Lucero *et al.* (1997) applied a nonlinear temporal normalization technique to decompose a set of lip acceleration records into a common pattern and components of shape and timing variability. This technique was also used to evaluate irregularity of voice signals (Lucero and Koenig, 2000), and to assess variability of glottal gestures in children versus adults (Koenig and Lucero, 2002). FDA has also been applied to fit a differential equation to lip

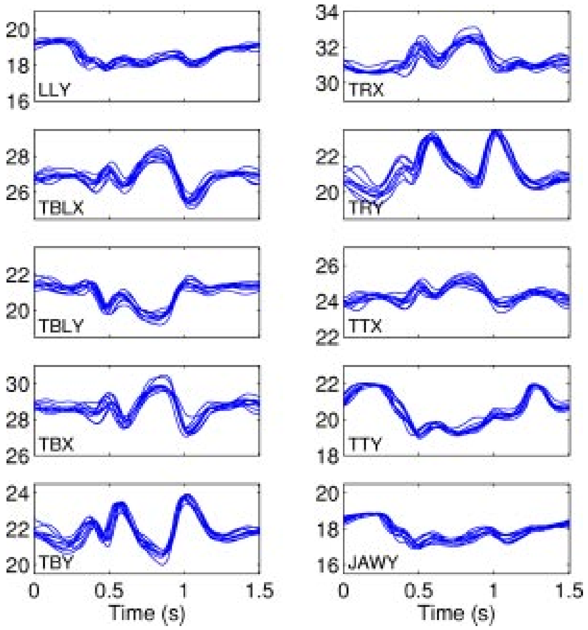

Fig. 1. Measured articulator trajectories for subject LK, utterance "say aka again." Vertical dimensions in cm.

trajectory data, and assess the variability of the forces acting on the lips (Lucero, 2002).

Here, we use the FDA temporal normalization technique to detect segments of articulator trajectories with more or less variability, under different movement constraints. We also define an index which allows us to compare variability across articulators and utterances. Our general purpose is to evaluate the usefulness of this approach, and explore its application to find movements which are more constrained than others, differences between active and passive articulators and between different sounds, and similar issues.

2. Data

We collected horizontal and vertical displacement data of lips, jaw, and tongue in vowel-consonant-vowel sequences, from four normal English-speaking adults, producing ten repetitions of /api/, /ata/, /aka/ embedded in the carrier phrase "Say_again." The speakers were naive to the specific purposes of the experiment, although they had had experience with this recording technique before. Data were collected at a sampling rate of 625 Hz, from receivers on the jaw, upper and lower lips, and at four locations on the tongue: one on the tip (TT: tongue tip), one as far back as possible (TR: tongue root, typical position was 5.0–5.5 cm behind the tongue tip, measured with the tongue protruded), and the other two spaced evenly between the front and back extremes (TBL: tongue blade, and TB: tongue body). Two additional receivers, placed on the nose bridge and the upper incisors, were used for the correction of head movement; thus, they were not used for any measurements. For the jaw and lip receivers, only displacements in the vertical direction were considered as this is the principal direction of movement (Ramsay *et al.*, 1996). The upper lip data were recorded but not used, because the upper lip motion is small [it can be used for analyzing a derived measure of lip aperture (Löfqvist and Gracco, 1997; Löfqvist, 2004)]. The jaw was not decoupled from the lip. Since we are interested in the lower lip as an end effector, it should be measured with the jaw component included. For the tongue signals, movement trajectories were decomposed into *x* and *y* components. Figure 1 shows an example of measured trajectories.

3. Temporal normalization

All signal processing was performed using MATLAB. As preparation for the temporal normalization, the sets of records for each subject and utterance were first vertically aligned by subtracting the mean of the set. Next, all records were resampled to a common length of 201 points using spline interpolation, and this length was attributed an artificial time span of 0 to 1. Finally, the records were put into vector form

$$\mathbf{x}_i(t)=[\mathrm{TTY}_i(t),\mathrm{TTX}_i(t),\ldots,\mathrm{JAWY}_i(t)]^T, \tag{1}$$

where $i=1,\ldots,N$, N is the number of records in each set (for each subject and utterance), and $\mathrm{TTY}_i(t)$, $\mathrm{TTX}_i(t),\ldots,\mathrm{JAWY}_i$ are the recorded trajectories.

The records in each set were aligned in time by nonlinearly distorting their time scales through suitable warping functions $h_i(t)$. The warping functions were computed optimally, by minimizing the following measure of the distance of the aligned records $\mathbf{x}_i^*(t)=\mathbf{x}_i[h_i(t)]$ to their average $\bar{\mathbf{x}}^*(t)=N^{-1}\Sigma_{i=1}^{N}\mathbf{x}_i^*(t)$

$$C=\sum_{i=1}^{N}\left\{\int_0^1\|\bar{\mathbf{x}}^*(t)-\mathbf{x}_i^*(t)\|^2\,dt+\int_0^1\|D\bar{\mathbf{x}}^*(t)-D\mathbf{x}_i^*(t)\|^2\,dt+\lambda\int_0^1 w_i(t)dt\right\}, \tag{2}$$

where D denotes the first derivative, λ is a roughness penalty coefficient, and $w_i(t)$ is the relative curvature of $h_i(t)$ [computed as $w_i(t)=D^2h_i(t)/Dh_i(t)$]. This process is done iteratively, by computing a new average after aligning the records, and realigning them again (Ramsay and Silverman, 1997). Three iterations were performed, after which there was no visual difference between consecutive averages. After the alignment, the average of the normalized records represents the underlying pattern common to all records, and the difference of each normalized record to the average is the component of amplitude variation introduced by it (Lucero *et al.*, 1997; Lucero, 2004).

Note in Eq. (2) that the distance measure was computed on both the displacement records (first integral) and their first derivative (second integral). Inclusion of the first derivative provides a finer alignment, since it is more variable than the displacement and thus contains more events to align (Ramsay and Silverman, 1997). It also includes a smoothness constraint (third integral), to control the allowed distortion of the time scales. The vector norm in Eq. (2) is computed as the weighted measure

$$\|\bar{\mathbf{x}}^*(t)-\mathbf{x}_i^*(t)\|^2=[\bar{\mathbf{x}}^*(t)-\mathbf{x}_i^*(t)]^T[\mathrm{diag}(\alpha_i)][\bar{\mathbf{x}}^*(t)-\mathbf{x}_i^*(t)], \tag{3}$$

where α_i are weights, given by

$$\alpha_i=\left[\int_0^1\bar{\mathbf{x}}_i^{*2}(t)dt\right]^{-1}. \tag{4}$$

The norm for the derivatives in the second integral of Eq. (2) is similarly computed. In this way, the distance measures in Eq. (2) for each component of $\mathbf{x}_i(t)$ are normalized according to the size of the component, so that all of them (including displacements and derivatives) have the same relative weight in the total measure.

The mean and standard deviations of the normalized records were computed in each set, to visualize the distribution of amplitude variability along the trajectory pattern. For quantitative assessments, we also computed an index of amplitude variability, defined as the mean standard deviation of normalized records across the region between peak velocities of the consonantal closure by the active articulator, in each VCV sequence. The target productions were segmented out of the carrier phrase interactively by viewing the audio signal and the kinematic signals. All the labeling of peaks, valleys, and zero crossings was made algorithmically by selecting a temporal window in the signals. In particular, the points of peak velocities were ob-

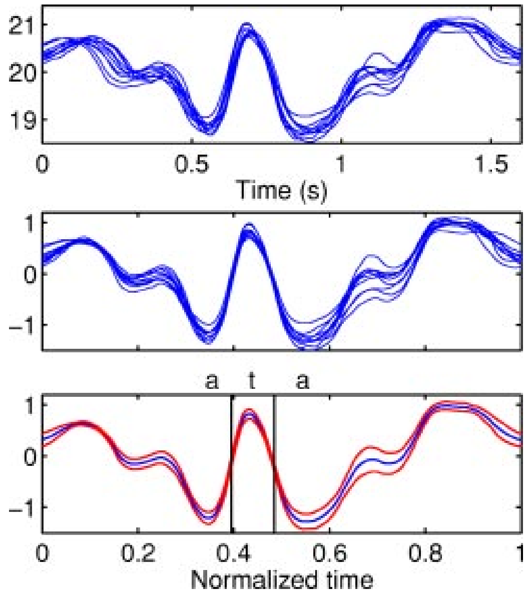

Fig. 2. Vertical displacement of tongue tip for subject DR, utterance "Say ata again." Top: recorded displacement. Middle: normalized records. Bottom: mean trajectory (blue curve) plus/minus standard deviation (red curves). The vertical lines mark the peak velocity of the tongue tip raising and lowering movements, and delimit the region over which the variability index was computed. Vertical dimensions in cm.

tained from the first derivative of the signals, computed using a three-point algorithm with spline smoothing (Ramsay and Silverman, 1997). Figure 2 shows an example of results for the vertical displacement of the tongue tip.

4. Results

Figure 3 shows the computed indices for the four speakers. Our expectation was that the articulators contributing to the required constriction/closure would show the least variability. In fact, the indices show that, during the /k/ closure, vertical variability in the tongue root (TRY) is relatively low compared to other articulators. This suggests that movements of the posterior tongue are tightly constrained during production of velar closures. Variability in the jaw (JAWY)

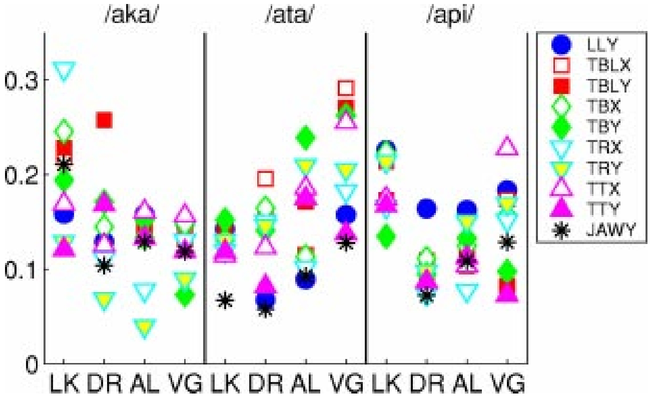

Fig. 3. Indices of variability for all subjects. Vertical dimensions in cm.

is among the lowest of all articulators during the alveolar closure for /t/, which is consistent with past reports indicating that the jaw is a major contributor to raising and anchoring the tongue for alveolars (Stone and Vatikiotis-Bateson, 1995). On the other hand, the results for the bilabial stop /p/ show a different pattern. The indices indicate that variability for the active articulator (lower lip, LLY) is among the highest of all articulators. The jaw shows a low variability, although not the lowest relative to other articulators. We hypothesize that, in bilabial closures, tissue compression upon lip contact may yield higher variability than in cases of an articulator meeting a rigid surface such as the hard palate (Löfqvist and Gracco, 1997).

5. Conclusions

This letter has presented a technique for computing variability of a set of multidimensional records. We believe that such use of FDA to quantify speech production variability has wide applicability in studying questions of speech motor control. Our results indicate that articulatory variability in normal adults varies as a function of both the linguistic constraints on speech movements and the biomechanical characteristics of the articulatory structures involved. We are currently expanding our data base of speech movement trajectories, to perform further analyses.

This technique may also have clinical utility by providing detailed profiles of an individual speaker's production variability, which could be used to assess the efficacy of treatment. It might also be easily extended to deal with other physiological signals, and thus find relevant applications in other areas.

Acknowledgments

This work was supported by NIH Grant No. DC-0008665 to Haskins Laboratories, and by CNPq and Finatec (Brazil).

References and Links

Adams, S. G., Weismer, G., and Kent, R. D. 1993. "Speaking rate and speech movement velocity profiles," J. Speech Hear. Res. **36**, 41–54.

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

Koenig, L. L., and Lucero, J. C. (**2002**). "Oral-laryngeal control patterns for fricatives in 5-year olds and adults," Proceedings of the 7th International Conference on Spoken Language Processing, pp. 49–52.

Löfqvist, A. 2004. "Lip kinematcs in long and short stop and fricative consonants," J. Acoust. Soc. Am. in press.

Löfqvist, A., and Gracco, V. L. 1997. "Lip and jaw kinematics in bilabial stop consonant production," J. Speech Lang. Hear. Res. **40**, 877–893.

Lucero, J. C. 2002. "Identifying a differential equation for lip motion signals," Med. Eng. Phys. **24**, 521–528.

Lucero, J. C. (**2004**). "Comparison of measures of variability of speech movement trajectories using synthetic records," J. Speech Lang. Hear. Res., accepted for publication.

Lucero, J. C., and Koenig, L. L. 2000. "Time normalization of voice signals using functional data analysis," J. Acoust. Soc. Am. **108**, 1408–1420.

Lucero, J. C., Munhall, K. G., Gracco, V. L., and Ramsay, J. O. 1997. "On the registration of time and the patterning of speech movements," J. Speech Lang. Hear. Res. **40**, 1111–1117.

Ramsay, J. O., and Silverman, B. W. (**1997**). *Functional Data Analysis* (Springer-Verlag, New York).

Ramsay, J. O., Munhall, K. G., Gracco, V. L., and Ostry, D. J. 1996. "Functional data analyses of lip motion," J. Acoust. Soc. Am. **99**, 3707–3717.

Sharkey, S. G., and Folkins, J. W. 1985. "Variability in lip and jaw movements in children and adults: Implications for the development for speech motor control," J. Speech Hear. Res. **28**, 8–15.

Smith, A., and Goffman, L. 1998. "Stability and patterning of speech movement sequences in children and adults," J. Speech Lang. Hear. Res. **41**, 18–30.

Smith, A., Goffman, L., Zelaznik, H. N., Ying, G., and McGillem, C. 1995. "Spatiotemporal stability and patterning of speech movement sequences," Exp. Brain Res. **104**, 439–501.

Smith, B. L. 1995. "Variability of lip and jaw movements in the speech of children and adults," Phonetica **52**, 307–316.

Stone, M., and Vatikiotis-Bateson, E. 1995. "Trade-offs in tongue, jaw, and palate contributions to speech production," J. Phonetics **23**, 81–100.

Ward, D., and Arnfield, S. 2001. "Linear and nonlinear analysis of the stability of gestural organization of speech movements sequences," J. Speech Lang. Hear. Res. **44**, 108–117.