SMASH: A Tool for Articulatory Data Processing and Analysis

Jordan R. Green [1], Jun Wang [2], David L. Wilson [3]

[1] MGH Institute of Health Professions, Boston, MA, United States
[2] Callier Center for Communication Disorders
University of Texas at Dallas, Dallas, TX, United States
[3] Waisman Center, University of Wisconsin-Madison, Madison, WI, United States

jgreen2@mghihp.edu, wangjun@utdallas.edu, dwilson@cae.wisc.edu

Abstract

Recent innovations in 3D motion capture technology such as electromagnetic articulography (EMA) are providing unprecedented access to the intricate movements of the articulators during speech production. Although these technological advances afford exciting opportunities for advancing the assessment and treatment of speech, they have presented new challenges associated with data collection, processing, and analysis. To address these challenges, we have standardized our EMA data collection protocols and developed a Matlab-based software tool, SMASH, for processing, visualizing, and analyzing speech movement data. The goal of the software is to advance research on speech production by improving the efficiency and reliability of speech movement analyses.

Index Terms: speech production, 3D motion capture, articulatory data processing, Electromagnetic Articulograph

1. Introduction

Speech movements have been studied using a variety of technologies over the past fifty years. The first comprehensive studies of speech movements were conducted in the 1960s and 1970s and were based on mid sagittal x-rays films of talkers [1, 2]. The next generation of speech movement research was based on a variety of very specialized technologies including strain-gauge [3], x-ray microbeam [4, 5], Magnetic Resonance Imaging [6], and ultrasound [7, 8]. These instruments were largely custom-built and used by only a small number of laboratories. Over the past decade, the number of studies on speech and swallowing motor control has increased steadily due to the recent availability of commercial 3D motion capture devices including 3D optical motion capture and 3D Electromagnetic Articulography (EMA) systems [9, 10, 11]. EMA is rapidly becoming one of the predominant technologies used to study speech movement [12, 13, 14].

To our knowledge, there are currently only two commercially available 3D EMA devices one manufactured by Cartens (Carstens Medizinelektronik GmbH, Bovenden, Germany) and the other by NDI (NDI, Waterloo, Ontario, Canada). The latest models by Carstens are the AG500 and AG501, which superseded their 2D predecessors AG100 and AG200 [15, 16] and the EMMA device developed by Perkell and colleagues at MIT [9]. The spatial accuracy of the AG500 device was reported to be approximately 0.5 mm at a 200 Hz sampling rate [17]. The accuracy of the recently released AG501 was reported by the manufacture to be 0.3 mm using a 250 Hz sampling rate [18]. The accuracy of the NDI Wave system has been reported to have the similar accuracy level as that of the Carstens AG500 [19].

These systems are now enabling a critical mass of investigators to study speech behaviors more comprehensively than previously (e.g., larger number of speech tasks and participants) and to study the speech movements of previously inaccessible participant populations including medically fragile patients and young children. Moreover, recent improvements in hardware portability and real-time data streaming are opening possibilities for the development of a large number of clinical applications including articulator movement control devices [20], biofeedback [21], speech recognition with articulatory information [22, 23], and silent speech interfaces [24, 25, 26]. The accessibility and large amount of articulatory data now afforded by this technology, however, is presenting new challenges associated with data collection, reduction, and analysis:

- Efficient and reliable methods are needed for data preprocessing and post-acquisition analysis.
- Standardization data collection protocols are needed to allow for across study comparisons.
- Efficient protocols are needed for medically fragile patients and young children.
- Efficient and reliable methods are needed to segment speech utterances from continuous streams of movement data.

In short, the scientific and clinical promise of these technologies will only be realized through efforts to develop standardized protocols for data collection, and robust computational tools for data reduction and analysis.

To address those challenges we have formalized our procedures for collecting and analyzing 3D speech movement data. Our data processing and analyses are performed in a customized MATLAB-based program called SMASH (Speech Movement Analysis for Speech and Hearing research). SMASH was developed to improve the efficiency and reliability of speech movement data visualization, processing, and analysis. This effort is in conjunction with those from several other labs who are also developing visualization and analysis software (e.g., MVIEW [27], EMATOOLS [28], TRAP [29], for visualization and analysis, Carstens JustView, and recently developed VisArtico [30] for visualization).

2. Data Acquisition using EMA

2.1. Procedure

EMA tracks the translation and rotation of small electromagnetic sensors that are attached to target articulators. The sensors are attached to the surface of each articulator using oral tissue adhesive (i.e., PeriAcryl). Prior to attachment, the tongue surface is dried using sterilized gauze and a dental air compressor. Figure 1 shows the typical placement of the sensors. Four tongue sensors (T1, T2, T3, and T4) are placed on the midline tongue approximately 10 mm apart. Placement

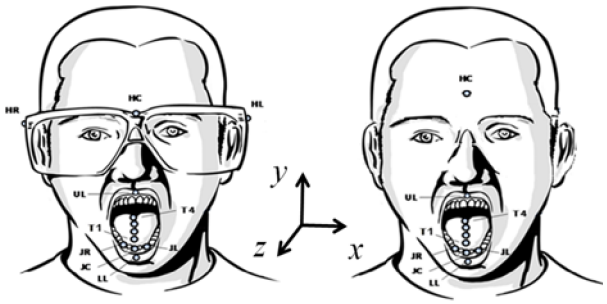

Figure 1. *Sensor positions and the coordinate system for Carstens EMA (left) and NDI Wave system (right). Sensor labels are described in text.*

is referenced to T1, which is placed 10 mm posterior to the tongue apex [13, 14]. The lip sensors are attached midline to the vermilion borders of the upper (UL) and lower (LL) lips. Jaw sensors are attached to the mandibular teeth rather than externally on the chin to avoid skin movement artifact [31]. These sensors are attached just above the mandibular gum line between (1) the right first premolar and the right canine (JR), (2) the left first premolar and the left canine (JL), and (3) the central incisors (JC). Although only two sensors are required to describe the translation and rotation of the mandible, the third sensor is used if one sensor becomes detached during data collection.

Of course, sensor placement varies depending on the purpose of the study and other constraints on data collection; for example, medically fragile and young participants may tolerate only one or two tongue sensors [32]. Our recent work has shown that two tongue sensors (i.e., tongue tip and tongue body back) plus lips are sufficient for distinguishing the major English phonemes [32].

Head movement needs to be tracked during data collection to derive head-independent articulatory movements. Two sensors are needed to account for the translation and rotation of the head using the Carstens 5 DOF (degrees of freedom) sensors (three sensors are required if using *x*, *y*, and *z* coordinates only) and one using the NDI 6 DOF sensor. For the Carstens system, we attached the head sensors to a pair of glasses to avoid skin motion artifacts [33]. As labeled in Figure 1, HC (Head Center) is placed on the bridge of the glasses; HL (Head Left) and HR (Head Right) are placed about 2.5 cm posterior, but on the left and right outside edge of each lens, respectively. For the NDI system, the single 6 DOF head sensor is mounted to a small stiff piece of cardboard and then attached to the center of the forehead using double-sided tape.

Figure 1 shows the orientation of the coordinate system we use. The default orientations of the Carstens and NDI systems are rotated to match the convention used in our lab: *x* is lateral (left-right); *y* is vertical (up-down); and *z* is protrusive (anterior-posterior).

After the sensors are attached, the relevant speech stimuli are displayed orthographically on a large computer screen in front of the participants while pre-recorded speech samples are played at approximately 60 dB. Participants are typically asked to rest shortly (about 0.5 second) between productions to minimize co-articulation effects and to facilitate segmenting the stimuli prior to analysis [14]. We have successful recorded tongue movement data using this approach in persons with advanced neuromuscular disease and in children as young as four years of age.

2.2. Preprocessing

Both EMA systems (Carstens and NDI) have their own built-in procedures for head-correction. Please refer to the manuals of the two systems for details.

3. Data Processing and Analysis using SMASH

3.1. Overview

SMASH is a GUI (graphical user interface) driven software program that we designed to provide a full solution for articulatory data processing, visualization, and analysis. It supports a variety of data formats, including articulatory data generated by the Cartens and NDI systems, motion capture systems that use the c3d format, .wav audio files, ASCII files, and Microsoft Excel files. SMASH automates many aspects of data preprocessing and provides a menu driven approach for accessing speech movement visualization and analysis tools.

Prior to analysis, users are required to create a template that specifies information such as the type of signals in each channel (e.g., movement, audio, EMG, force, aerodynamic), sensor channel labels, sampling rates, and desired filter settings. The program aligns data channels that are obtained using different sampling rates (see Figure 3 for an example of acoustic and kinematic time-series alignment). Sub-routines are provided to read the header of some known formats (AG500, C3D) to automatically obtain information regarding sensor labels and sampling rates. This feature significantly minimizes the time required to label each column.

3.2. Interfaces and Visualization

The Spatial Analysis and Time-series Analysis windows are the used for data visualization. The Spatial Analysis displays the 2D or 3D motion paths and the Time-series analysis displays the positional time-histories.

Spatial Analysis. Figure 2 shows the Spatial Analysis GUI in SMASH, which supports the data visualization in 3D view or any combination of 2D views (i.e., *x*-*y*, *x*-*z*, or *y*-*z*). The 3D motion paths of multiple data files (e.g., 0001.txt, 0002.txt… in Figure 2) can be displayed simultaneously and in different colors. The displayed data can be animated. The frequency parameter on the left of the button "animate" is specified to control the animation speed. The checkbox "Trace" is used to display a history of the motion path during animation.

Time-series Analysis. Figure 3 shows the Time-series Analysis window in SMASH, which displays the time-series data from each spatial dimension (*x*, *y*, and *z*) for each selected sensors. The horizontal axis is time (in seconds); the vertical axis is the *x*, *y*, and *z* coordinates (in mm) of the selected sensors. The associated time-aligned acoustic signal is plotted at the top of the window for reference.

A data trimming function is provided to allow users to analyze regions within the data file. The buttons "Set Trim Left" and "Set Trim Right" are used to mark the beginning and ending of a selected segment of data (e.g., the cyan vertical lines in Figure 3). The onset and offset time (in seconds and in frame number) of the trimmed segment are displayed above the buttons "Set Trim Left" and "Set Trim Right". The radio

button “View Trimmed” is for displaying the trimmed segment only (between the cyan vertical lines in Figure 3); the button “View All” is used to restore the view of the whole file.

To assist with data trimming, peaks and troughs in the data can be visualized. If the checkbox “Mark Peaks” and “Mark Troughs” are selected, a hash mark will appear on each trace identifying the associated landmarks in each time-series. These landmarks are identified based on zero-crossings in the first derivative of each signal. Alternatively, the user can identify landmarks in the time-series based on a threshold such as +/- 2.5 standard deviations above or below the average value in the signal.

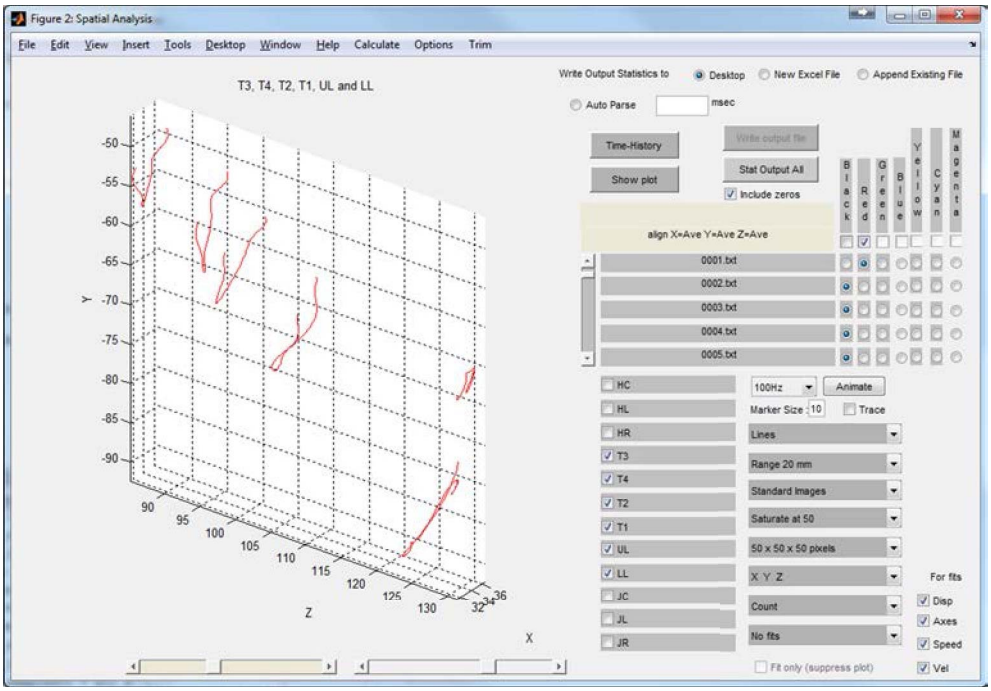

Figure 2. *Spatial Analysis window in SMASH. The Spatial Analysis window shows the motion path of six articulators T1, T2, T3, T4, UL, and LL in selected view (3D).*

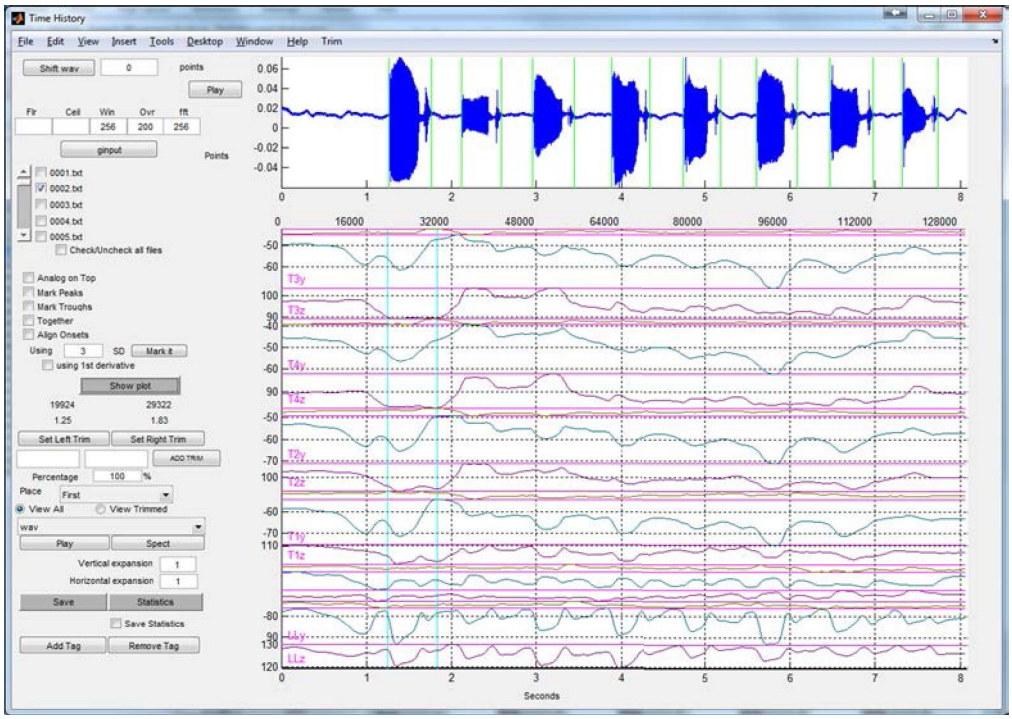

Figure 3. *Time-series Analysis window in SMASH displays the time-series data of each spatial dimension. In this plot, the time-series of six articulators' are displayed, i.e., T1, T2, T3, T4, UL, and LL. The Time-series Analysis also provides articulatory-acoustic alignment for data segmentation.*

Trimming can also be achieved based on events in the acoustic data. The button "Add Tag" and "Remove Tag" in the bottom left are for marking segments of interest in the acoustic data. In the example in Figure 3, the vertical green lines on the audio data are the onset and offset for each stimulus. The tag values (in time) are written to a text file and stored for later use. Tagging only marks the onset and offset of each segment, and does not perform trimming. The button "Save" in the bottom left is used to save the trimmed data to a separate file.

3.3. Data Processing

SMASH has built-in preprocessing and analyses routines commonly used in speech production research.

Filtering. Speech movement data are typically low-pass filtered prior to additional processing. Users can specify different cut off frequencies or filtering functions (e.g., high pass, low pass, band pass) for different types of data (e.g., movement, acoustic).

Other preprocessing routines. In addition to filtering, users can specify several other preprocessing routines in the project template that will automatically be applied to each signal including interpolation, demeaning, rectification, and temporal and amplitude normalization.

Head movement correction. Most optical motion capture systems require custom routines for extracting lip and jaw movements that are independent from that of the head. This is achieved in SMASH by labeling the head markers as "reference markers" in the project template. Once the reference markers are specified, new head-corrected versions of the signals are computed and saved as additional columns in a data matrix that stores the entire data set; thus, the original data columns are preserved. At least three reference markers are needed for 3D head-correction when using optical motion captures systems. Because the quality of the speech movement data is dependent on the quality of the reference markers, SMASH allows users to specify more than three reference markers for head movement correction. A subroutine automatically determines the best three reference markers based on the assumption that the reference markers define a rigid object.

Signal re-expression. SMASH provides several routines for reducing the dimensionality of the 3D data prior to analysis. These "calculated" signals can be defined in the template (that describes data formats) and then appended to the data matrix that stores the entire data set. Current options include expressing the data as a single dimension along the principal axis of motion, or as the 3D Euclidean distance from a predefined origin or between two sensors. In addition, multiple sensors can be used to create a single signal that represents the change in area defined by three or more sensors over time. For example, the upper lip, lower lip, and right and left corners of the mouth can be used to derive a signal that represents time-varying changes in mouth area.

3.4. Data Analysis

Spatial analysis. Figure 2 illustrates the Spatial Analysis GUI where several kinematic metrics can be obtained from each motion path including the total path distance, the orientation of the principal axis of motion, and the peak and average 3D speed. The 3D working space (volume in mm^3) can also be derived based on a convex hull (see Figure 4 and [33]) or an ellipsoid fit [34]. Once a "fit" (e.g., 2 SD or 3 SD) is selected, pressing the "Show Plot" button outputs the values for each kinematic metric in the MATLAB command window.

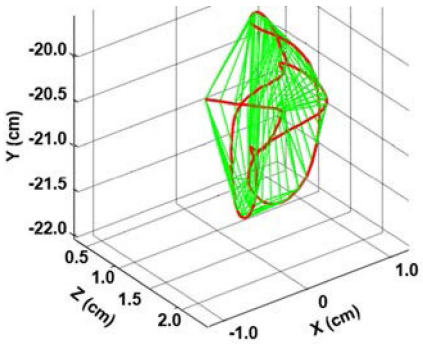

Figure 4. *Jaw working space in mm3 as defined by a convex hull fit (green) of the 3D movement path (red). The convex hull represents the tightest polygon mesh that contains all of the data points.*

The kinematic metrics of time-series data are obtained using the *Time-series Analysis* window (Figure 3). The button "Statistics" on the bottom left generates a variety of kinematic metrics for data displayed in the Time-series Analysis window. The metrics include mean, median, min, max, standard deviation (SD), duration, cumulative distance, slope, peak speed, average speed, and normalized jerk-cost (an index for measuring movement smoothness [35]), and area under curve. These values are displayed as a table in the MATLAB command window or are saved directly to a text file if the checkbox "Save Statistics" is selected.

Other additional analyses are provided in SMASH, for example, auto-correlation [36], cross-correlation [37], and measures of spatiotemporal movement instability [38, 39]. The spectral components of each time series can also be estimated using the Fast Fourier Transformation (FFT) routine [40].

4. Conclusions and Future Work

In this paper, we have described our lab's protocol for articulatory data collection using EMA, and a Matlab-based software program called SMASH for articulatory data processing and analysis. The purpose of the software is to accelerate the pace of speech production research by standardizing and automatizing many aspects of speech movement analysis. SMASH has been used by our lab for over a decade and more recently by a small number of speech production labs in North America, Europe, Asia, and Australia. SMASH is currently actively maintained and improvement (e.g., new features, more user friendly GUI) will be continued in the future.

5. Acknowledgements

This work was in part funded Barkley Trust, University of Nebraska-Lincoln, Excellence in Education Fund, University of Texas at Dallas, and a grant awarded by the National Institutes of Health (R01 DC009890/DC/NIDCD NIH HHS/United States). We would like to thank Yana Yunusova, Vasily Giannak, Antje Mefferd, Ignatius Nip, Erin Wilson, Mili Kuruvilla, Anusha Thomas, Nirmal Srinivasan, Cara Ullman, Toni Hoffer, Sara Benham, Cynthia Didion, Lori Synhorst and the many of students and users who have helped debug SMASH.

6. References

[1] K. L. Moll, "Cinefluorographic techniques in speech research", *Journal of Speech and Hearing Research*, vol. 3, no. 3, pp. 227-241, 1960.

[2] J. S. Perkell, *Physiology of speech production: Results and implications of a quantitative cineradiographic study*. Cambridge, MA: MIT Press, 1969.

[3] J. H. Abbs and B. N. Gilbert, "A strain gauge transducer system for lip and jaw motion in two dimensions", *Journal of Speech and Hearing Research*, vol.16, pp. 248-256, 1973.

[4] J. Westbury, "X-ray microbeam speech production database user's handbook", University of Wisconsin, 1994.

[5] G. Weismer, Y. Yunuvosa, J. R. Westbury, "Interarticulator coordination in dysarthria: An x-ray microbeam study", *Journal of Speech, Language, and Hearing Research*, vol. 46, pp. 1247-1261, 2003.

[6] S. S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y.-C. Kim, A. Lammert, M. I. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time MRI articulatory corpus for speech research", *Interspeech*, 2011.

[7] M. Stone, "A guide to analyzing tongue motion from ultrasound images", *Clinical Linguistics and Phonetics*, vol. 19, pp. 455-502, 2005.

[8] T. Bressmann, "Ultrasound imaging and its application in speech-language pathology and speech science", *American Speech-Language-Hearing Association Newsletter*, vol. 33, no. 4, pp. 204-211, 2007.

[9] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements", *Journal of Acoustical Society of America*, vol. 92, no. 6, pp. 3078–3096, 1992.

[10] P. Hoole and A. Zierdt, *Five-dimensional articulography*, in Speech Motor Control: New developments in basic and applied research, B. Maassen and P. van Lieshout, Eds. Oxford University Press, ch. 20, pp. 331–349, 2010.

[11] M. Hasegawa-Johnson, "Electromagnetic exposure safety of the Carstens Articulograph AG100", *Journal of Acoustical Society of America*, vol. 104, no. 4, pp. 2529–2532, 1998.

[12] C. M. Steele, and P. H. H. M. van Lieshout, "The use of electromagnetic midsagittal articulography in the study of swallowing", *Journal of Speech, Language, and Hearing Research*, vol. 47, pp. 342-352, 2004.

[13] J. R. Green and Y. Wang, "Tongue-surface movement patterns during speech and swallowing", *Journal of Acoustical Society of America*, vol. 113, pp. 2820-2833, 2003.

[14] J. Wang, J. R. Green, A. Samal, and Y. Yunusova, "Articulatory distinctiveness of vowels and consonants: A data-driven approach," *Journal of Speech, Language, and Hearing Research*, 2013 (In press).

[15] P. Hoole, "Issues in the acquisition, processing, reduction, and parameterization of articulographic data", *Instituts für Phonetik und Sprachliche Kommunikation*, München (FIPKM), vol. 34, pp. 158–173, 1996.

[16] B. Tuller, S. Shao, and J. A. S. Kelso, "An evaluation of an alternating magnetic field device for monitoring tongue movements", *Journal of Acoustical Society of America*, vol. 88, pp. 674–679, 1990.

[17] Y. Yunusova, J. R. Green, and A. Mefferd, "Accuracy assessment for AG500 electromagnetic articulograph", *Journal of Speech, Language, and Hearing Research*, vol. 52, pp. 547-555, 2009.

[18] Carstens Medizinelektronik GmbH, "Articulograph AG501 Flyer", online: http://www.articulograph.de/wp-content/uploads/2013/01/Articulograph-AG501-brochure-2012.pdf, accessed on Mar 08, 2013.

[19] J. Berry, "Accuracy of the NDI wave speech research system", *Journal of Speech, Language, and Hearing Research*, vol. 54, pp. 1295-1301, 2011.

[20] X. Huo, J. Wang, and M. Ghovanloo, "A magneto-inductive sensor based wireless tongue-computer interface", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 16, no. 5, pp. 497-504, 2008.

[21] W. Katz, M. McNeil, and D. Garst, "Treating apraxia of speech (AOS) with EMA-supplied visual augmented feedback", *Aphasiology*, 24, 826-837, 2010.

[22] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition", *Journal of Acoustical Society of America*, vol. 121, no. 2, 723-742, 2007.

[23] F. Rudzicz, G. Hirst, P. Van Lieshout, "Vocal tract representation in the recognition of cerebral palsied speech", *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 4, 1190-1207, 2012.

[24] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces", *Speech Communication*, vol. 52, no. 4, pp. 270-287, 2010.

[25] J. Wang, A. Samal, J. R. Green, and F. Rudzicz, "Sentence recognition from articulatory movements for silent speech interfaces", Proc. *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pp. 4985-4988, Kyoto, Japan, 2012.

[26] J. Wang, A. Samal, J. R. Green, and F. Rudzicz, "Whole-word recognition from articulatory movements for silent speech interfaces", *Interspeech*, Portland, OR, 2012.

[27] M. K. Tiede. MVIEW: Multi-channel visualization application for displaying dynamic sensor movements, unpublished.

[28] N. Nguyen, "A MATLAB toolbox for the analysis of articulatory data in the production of speech", *Behavior Research Methods, Instruments, & Computers*, vol. 32, no. 3, pp. 464–467, 2000.

[29] Savariaux, C. "TRAP, un logiciel d'analyse de signaux multi formats sour Matlab", Online: http://www.gipsa-lab.grenoble-inp.fr/~christophe.savariaux/recherches.html, accessed on May 22, 2013.

[30] S. Ouni, L. Mangeonjean, I. Steiner, "VisArtico: a visualization tool for articulatory data", *Interspeech*, Portland, OR, 2012.

[31] J. R. Green, E. M. Wilson, Y. Wang, and C. A. Moore, "Estimating mandibular motion based on chin surface targets during speech", *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 4, pp. 928-939, 2007.

[32] J. Wang, J. R. Green, and A. Samal, "Individual articulator's contribution to phoneme production", Proc. *IEEE Intl. Conf. on Acoustics, Speech, Signal Processing*, Vancouver, Canada, 2013.

[33] J. R. Green and E. M. Wilson. "Spontaneous facial motility in infancy: A 3D kinematic analysis." *Developmental Psychobiology,* vol. 48, pp. 16-28, 2006.

[34] E. M. Wilson and J. R. Green. "The development of jaw motion for mastication." *Early Human Development,* vol. 85, pp. 303-311, 2009.

[35] K. Yashiro, M. Fujii, O. Hidaka, K. Takada, "Kinematic modeling of jaw-closing movement during food breakage", *Journal of Dental Research*, vol. 80, pp. 2030–2034, 2001.

[36] J. R. Green, C. A. Moore, J. L. Ruark, P. R. Rodda, W. T. Morvée, and M. J. Vanwitzenburg. "Development of chewing in children from 12 to 48 months: Longitudinal study of EMG patterns." *Journal of Neurophysiology*, vol. 77, pp. 2704-2716, 1997.

[37] J. R. Green, C. A. Moore, M. Higashikawa, and R. W. Steeve. "The physiologic development of speech motor control: Lip and jaw coordination." *Journal of Speech, Language and Hearing Research,* vol. 43, pp. 239-256, 2000

[38] A. Smith, L. Goffman, H. Zelaznik, G. Ying, and C. McGillem, "Spatiotemporal stability and pattering of speech movement sequences", *Experimental Brain Research*, vol. 104, pp. 493–501, 1995.

[39] J. R. Green, C. A. Moore, M. Higashikawa, and K. J. Reilly. "The sequential development of jaw and lip control for speech." *Journal of Speech, Language and Hearing Research,* vol. 45, pp. 66-79, 2002.

[40] E. M. Wilson, J. R. Green, and G. Weismer, "A kinematic description of the temporal characteristics of jaw motion for early chewing: Preliminary findings." *Journal of Speech, Language, and Hearing Research*, vol. 55, pp. 626-638, 2012.