SMASH: A Tool for Articulatory Data Processing and Analysis

Jordan R. Green¹, Jun Wang², David L. Wilson³

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

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 movements [12, 13, 14].

To our knowledge, there are currently only two commercially available 3D EMA devices one manufactured by Cartens (Cartens Medizinelektronik GmbH, Bovenden, Germany) and the other by NDI (NDI, Waterloo, Ontario, Canada). The latest models by Cartens are the AG500 and AG501, which superseded their 2D predecessors AG100 and AG200 [15, 16] and the EMMA device developed by Perkell and colleagues at MIT [9].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 manufacturer 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 Cartens 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
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 the NDI 6 DOF sensor). For the Carstens system, we attached the head sensors to a pair of glasses to avoid skin motion artifacts. The horizontal axis is time (in seconds); the vertical axis is the \( x \)-coordinate (in mm) of the selected \( x \)-coordinate (in mm) for each selected sensor. The associated time-aligned acoustic signal is plotted in the Time-series Analysis window in SMASH, which displays the time-series data from each spatial dimension (\( x \), \( y \), and \( z \)) for each selected sensor. 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 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 artifacts [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].

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 neuro muscular disease and in children as young as four years of age.
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 capture 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 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.

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 the 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 improved (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


