INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Video-based tracking of jaw movements during speech: Preliminary results and future directions

Andrea Bandini[1], *Aravind Namasivayam*[1,2], *Yana Yunusova*[1,2,3]

[1]University Health Network: Toronto Rehabilitation Institute, Toronto, Canada
[2]Department of Speech-Language Pathology, University of Toronto, Canada
[3]Brain Sciences, Sunnybrook Research Institute, Toronto, Canada

andrea.bandini@uhn.ca, a.namasivayam@utoronto.ca, yana.yunusova@utoronto.ca

Abstract

Facial (e.g., lips and jaw) movements can provide important information for the assessment, diagnosis and treatment of motor speech disorders. However, due to the high costs of the instrumentation used to record speech movements, such information is typically limited to research studies. With the recent development of depth sensors and efficient algorithms for facial tracking, clinical applications of this technology may be possible. Although lip tracking methods have been validated in the past, jaw tracking remains a challenge. In this study, we assessed the accuracy of tracking jaw movements with a video-based system composed of a face tracker and a depth sensor, specifically developed for short range applications (Intel® RealSense™ SR300). The assessment was performed on healthy subjects during speech and non-speech tasks. Preliminary results showed that jaw movements can be tracked with reasonable accuracy (RMSE≈2mm), with better performance for slow movements. Further tests are needed in order to improve the performance of these systems and develop accurate methodologies that can reveal subtle changes in jaw movements for the assessment and treatment of motor speech disorders.

Index Terms: jaw kinematics, depth sensors, face tracking, markerless

1. Introduction

An accurate evaluation of the articulatory movements during speech and non-speech tasks is important for the assessment, diagnosis and treatment of motor speech disorders [1-3]. Changes in the movements of the lips and jaw have been associated with early signs of oromotor decline, disease progression and speech intelligibility decline in amyotrophic lateral sclerosis (ALS), Parkinson's disease (PD) and post stroke [1-5]. Changes in the motor control of the jaw have been salient not only in the adult-onset conditions such as ALS and PD, but also in childhood speech disorders such as Childhood Apraxia of Speech (CAS) [6]. Despite their importance, the assessment of facial (lip and jaw) movements has not been incorporated into clinical practice due to the complexity and high cost of the current motion tracking technologies (e.g., electromagnetic articulography, optoelectronic systems, video-based motion tracking) [7, 8].

The development of relatively cheap and non-intrusive methods for tracking speech movements is of high interest not only for non-clinical fields such as animation and gaming but also for clinical purposes. Feng et al., [9] developed a system for tracking jaw and lip movements, composed of two consumer-grade cameras. Despite its good accuracy, with tracking error < 0.5 mm, this method required calibration of the cameras and the attachment of reflective markers. The recent development of 3D depth cameras [10] and powerful face tracking algorithm [11] has offered a unique opportunity to develop systems for studying facial movements without using sensors and markers.

Bandini et al., 2015 [8] tested the accuracy of a depth sensor (Primesense Carmine 1.09) along with a face tracking algorithm (Intraface [11]) for tracking 3D movements of lips during various speech tasks (e.g., syllables, words and sentences). The average root mean square error (RMSE) of the lip points with respect to an optoelectronic marker-based method ranged between 1 and 4 mm. The validity of the above video-based method was also demonstrated in a clinical population of individuals diagnosed with PD [12]. We were able to discriminate patients with PD from healthy controls, using lower lip movements during syllable repetitions. Ouni and Dahmani, 2016 [13] provided further evidence for suitability of markerless methods by comparing the performance of two depth sensors (Primesense and Intel® RealSense™ camera) with respect to an optoelectronic marker-based reference during sentence and syllable productions. The authors demonstrated that lips and chin could be tracked with good accuracy (overall RMSE < 2 mm). However, separate results for lips and chin points were not provided. The RealSense™ camera proved to be more accurate than the Primesense sensor. This improvement was likely due to the higher temporal resolution of the RealSense™ camera (nearly 50 Hz vs 30 Hz of the Primesense). The first attempt to develop a fully markerless system for studying jaw movements was proposed by Tanaka et al., 2016 [14]. The authors used the Microsoft Kinect to track the jaw during masticatory movements from the chin. Results indicated the range of tracking errors between 2.4 and 9 mm. Only chewing was investigated in this paper.

Although the latter two works [13, 14] attempted to track jaw movements, the points of interest were located on the chin, right below the lower lip. As demonstrated by Green et al., 2007 [15], this region (albeit easily accessible from images) is not ideal to represent jaw movements, since it is strongly affected by lower lip movements, especially during speech.

The aim of this study was to test the accuracy of a markerless method for tracking jaw movements during speech, using the guidelines provided by [15], focusing on the inferior border of the jaw. Here we reported preliminary results on two participants and discussed findings with respect to method optimization and future directions.

http://dx.doi.org/10.21437/Interspeech.2017-1371

2. Materials and methods

2.1. Participants and speech task

Two healthy young male volunteers, native speakers of English, were recruited for the experiment. Subjects had to perform the following speech and non-speech tasks:

- Sentence repetition task (SRT) – The sentence "Buy Bobby a puppy" was repeated 20 times at normal comfortable speaking rate and loudness, briefly pausing between each repetition.
- Oral diadochokinetic task (DDK) - Repetitions of the syllables /pa/ and /pataka/, as fast and clear as possible on one breath without stopping were obtained.
- Non-speech movements (NSM) – The maximum mouth opening task was repeated 5 times starting from the rest position (mouth closed). Participants were asked to pause at the maximum opening for a few seconds.

A total of 40 sentences, 50 /pa/, 40 /pataka/ and 10 opening/closing movements were considered for the analysis. The acquisitions were performed in a quiet room. Subjects were seated during the experiment, looking at the camera and avoiding large head movements during the tasks. The Wave field generator was positioned near the right side of the head.

2.2. Data acquisition

2.2.1. Wave system

The Wave Speech Research System (NDI - Waterloo, Ontario, Canada) was used to track jaw movements. Wave, an electromagnetic articulography system, is commonly used in speech research [3, 16]. We used this device to build the ground truth for comparison with the video-based system, since Wave can track sensors with accuracy < 0.5 mm within near magnetic field (< 200 mm) [16]. Four sensors were attached to the face in the following positions (Figure 1):

- N1 - Sensor placed in correspondence to the nasion (intersection between the frontal bone and nasal bones);
- N2 - Sensor placed on the tip of the nose;
- JR (JL) - These sensors were placed on the inferior border of the jaw, about one quarter of the distance from the gnathion and the right (left) gonion. These positions were chosen following a previous study [15] which demonstrated that these points were less susceptible to soft tissue deformation during movement than the central part of the chin.

The three-dimensional trajectories of the four sensors were collected with a sampling frequency (Fs) of 100 Hz. Synchronized audio signals (Fs = 22 kHz, 16 bits per sample) were collected during the experiments. Data were saved using the Wavefront software v. 1.1.

2.2.2. Video-based system: Intel® RealSense™ and face tracker

The Intel® RealSense™ SR300 (Intel Corp., Santa Clara, CA, USA) camera was used to acquire color and depth videos. The SR300 is a structured light sensor composed of a color camera, an infrared (IR) emitter and an IR camera. The SR300 was chosen because it was specifically designed for short range applications (such as face and hand tracking), and previous works demonstrated the higher performance for the analysis of facial movements during speaking with respect to other depth sensors [13].

The Intel® RealSense™ SR300 camera was placed in front of the subject's face at a distance between 0.4-0.5 m, according to manufacturer's specifications (0.3-1 m for face tracking [17]). The angle between the optical axis and the camera was nearly 20°, in order to optimize tracking of the inferior border of the chin [15]. Color and depth videos were collected with a resolution of 640x480 pixels at approximately 50 frames per second (fps) [13].

The video-based jaw tracking was performed by using the face tracking algorithm of the RealSense SDK R3 [17]. This algorithm fits a face model composed by 78 points to the image streams. 19 out of 78 points are dedicated for the contour of the face. The location of these points on the face, with their original index as retrieved from the algorithm, are reported in Figure 1. The algorithm provided the 3D positions (in mm) of each point in the camera coordinate system, with origin corresponding to the IR camera center. These points were extracted with a customized code written in C++ language, using the RealSense SDK R3.

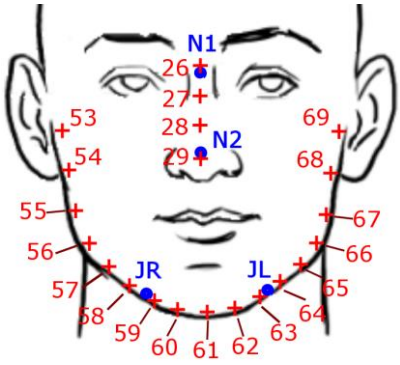

Figure 1: *Placement of Wave sensors (blue dots) and face tracker points (red crosses) on the nose and chin. Only selected face tracking points of face contour (see text) and nose are reported. For an exhaustive layout of facial points used by this face tracking algorithm please refer to [17].*

2.3. Data analysis

2.3.1. Pre-processing

Any gaps associated with missing data in the Wave trajectories were filled by using a spline interpolation. The 3D trajectories of the points of interest extracted with both methods were low-pass filtered (8-pole Butterworth, cut-off frequency 15 Hz) in order to remove high frequency noise. Subsequently, the trajectories of the Wave sensors were resampled to the sampling frequency of the SR300 camera (nearly 50 Hz). The downsampling was performed in order to make the trajectories from both systems comparable.

2.3.2. Kinematic measures

Jaw movements were assessed with respect to a reference point located on the nose tip. Thus, two trajectory signals were extracted from both data streams (Wave and video-based):

- NJR – Euclidean distance between the reference point on the nose and JR (Wave), and Euclidean distance between the reference point on the nose and midpoint 58-60 (video-based, Figure 1).
- NJL – Euclidean distance between the reference point on the nose and JL (Wave), and Euclidean distance between

the reference point on the nose and midpoint 62-64 (video-based, Figure 1).

For the video-based analysis, we used the average of the points 58-60 for the right side and 62-64 for the left side instead of using single nearest points to JR and JL because the midpoints appeared to be more robust.

The velocity signals vNJR and vNJL (in mm/s) were calculated as the first derivative of NJR and NJL with time, respectively.

2.3.3. Accuracy assessment

The RMSE between the trajectories (NJR, NJL) obtained with the two methods was computed in order to evaluate the average error of the video-based method from the ground truth (i.e., Wave). Moreover, the cross-correlation with zero lag (CC0) of trajectories and velocities (NJR, NJL, vNJR, vNJL) obtained with both techniques was computed in order to provide a measure of similarity. These measures were calculated for each repetition and then averaged across the total number of speech and non-speech productions. Bland-Altman plots [18] were used to compare values of distance (maximum, minimum and average values of each repetition) and velocity (maximum – opening phase, minimum – closing phase, of each repetition), in order to assess the agreement between the two methods. Data analysis steps were performed in Matlab 2016b.

3. Results

An example of NJL and NJR estimated with the two methods during the SRT is shown in Figure 2. Mean values and standard deviations of RMSE and CC0 are reported in Table 1.

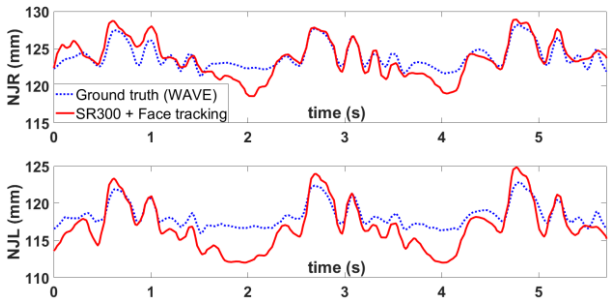

Figure 2: *Trajectory signals NJR (top) and NJL (bottom) during sentence "Buy Bobby a puppy".*

Table 1: *Measures computed to assess the accuracy of the video-based method for tracking jaw movements.*

	RMSE NJR (mm)	RMSE NJL (mm)	CC0 NJR	CC0 NJL	CC0 vNJR	CC0 vNJL
Total	2.27 ± 1.36	1.72 ± 1.48	0.58 ± 0.38	0.77 ± 0.29	0.51 ± 0.31	0.67 ± 0.24
SRT	2.11 ± 0.97	1.75 ± 0.85	0.70 ± 0.25	0.90 ± 0.08	0.53 ± 0.21	0.75 ± 0.13
DDK	2.30 ± 0.97	1.23 ± 0.81	0.49 ± 0.41	0.68 ± 0.33	0.47 ± 0.35	0.61 ± 0.28
NSM	2.70 ± 0.51	5.95 ± 1.51	0.99 ± 0.01	0.99 ± 0.01	0.80 ± 0.07	0.87 ± 0.04

Figures 3 and 4 show the Bland-Altman plots for the distance and velocity values obtained across all of the tasks for both participants. The mean difference for NJR was -0.94 mm (95% CI [-1.13, -0.75] mm), revealing the distance overestimation of nearly 1 mm for the video-based system. The limits of agreement for NJR were -5.62 mm (95% CI [-5.95, -5.29] mm) and 3.74 mm (95% CI [3.41, 4.07] mm). The mean difference for NJL was 0.57 mm (95% CI [-1.13, -0.75] mm), underestimating the distance NJL by about 0.5 mm. The limits of agreement for NJL were -3.64 mm (95% CI [-3.94, -3.34] mm) and 4.78 mm (95% CI [4.48, 5.08] mm).

The mean difference for vNJR was -2.45 mm/s (95% CI [-8.89, 3.98] mm/s), with limits of agreement for vNJR between -131.24 mm/s (95% CI [-142.39, -120.10] mm/s) and 126.34 mm/s (95% CI [115.19, 137.48] mm/s). The mean difference for vNJL was -0.44 mm/s (95% CI [-4.47, 3.58] mm/s), with limits of agreement for vNJ between -81.00 mm/s (95% CI [-87.97, -74.03] mm/s) and 80.12 mm/s (95% CI [73.14, 87.09] mm/s).

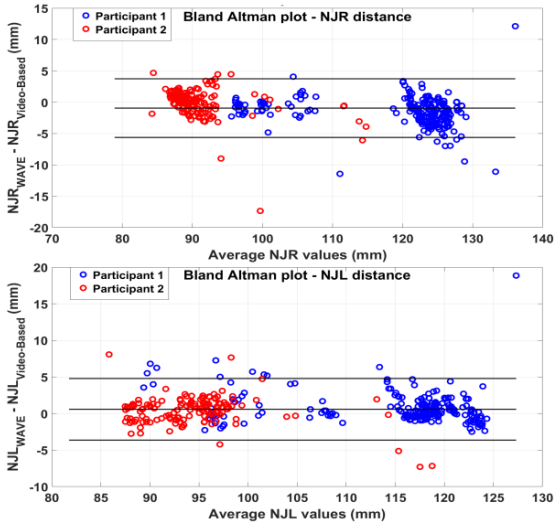

Figure 3: *Bland-Altman plot for the distance measures (NJR – top, NJL - bottom) from both participants.*

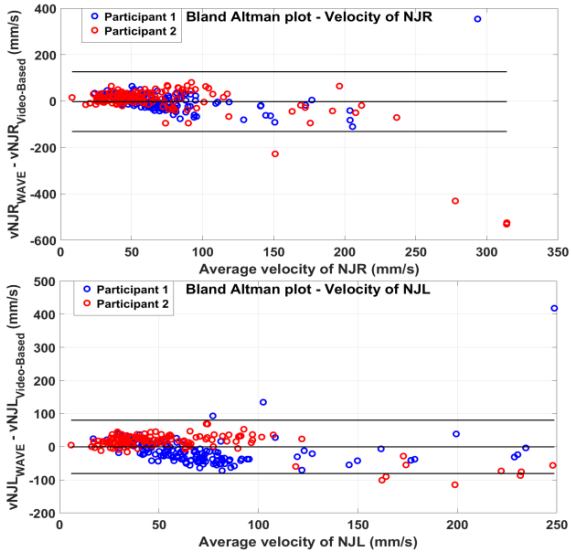

Figure 4: *Bland-Altman plot for the velocity measures (vNJR – top, vNJL - bottom) from both participants.*

4. Discussion

Our preliminary results showed acceptable accuracy of the video-based system in tracking jaw movements, with average RMSE of approximately 2 mm across tasks. The cross-correlations between the movement trajectories obtained with the two methods were high (> 0.90) for SRT and NSM (as compared to 0.68 in DDK) suggesting that a high similarity between the trajectories can be obtained with slow movements. In general, the video-based system showed higher performance in tracking the left side of the face as compared to the right. This discrepancy was likely due to the Wave field generator that created a shadow on the right side of the face. Since the face tracker showed susceptibility to light changes in our pilot study, our continued work will use constant and uniform light to prevent shadows and other illumination artifacts.

The NSM task was the only case where the RMSE of NJR was lower than that of NJL (Table 1). Although the correlation between the two methods was very high during NSM (0.99 for both JR and JL, and 0.87 for vJL), we noticed that during the maximum jaw opening the algorithm failed to track the lower contour of the face and great fluctuations of the facial points occurred around the inferior chin border when the mouth was maximally opened. This may indicate a difficulty of the algorithm in tracking large movements of the jaw, and the comparison of different face tracking/image processing algorithms will be explored to better understand and resolve this issue.

The Bland-Altman plots in Figure 3 showed relatively good agreement between the two methods for the distance measures. The limits of agreement were relatively small when compared to the average displacement. In general, the bias observed for the distance measures (~1 mm) was probably due to the non-exact matching between the Wave sensors and the face tracking points. We are exploring ways to improve on facial tracking point selection. In contrast, plots in Figure 4 reported poor agreement between the velocity measures, with limits of agreement comparable to the average velocity values. The velocity errors will be further explored, particularly with respect to task differences.

In general, the errors were in the same order of magnitude as those reported in previous studies that tracked lips during speech [8, 13]. Nevertheless, a similar video-based system showed higher performance in computing velocity measures of the lower lip [19], confirming that jaw tracking is a more challenging task than lip tracking, and still requires substantial improvements.

Green and colleagues [15] provided guidelines for choosing the most reliable points for tracking jaw externally via selected points on the chin. Following their findings, we chose the points at the inferior border of the jaw, one quarter of the distance from the gnathion and the gonion on the right and left, as reference. However, due to their position on the face, these points were not easy to track, especially from images, since small head rotations around the vertical axis may have occluded the region of interest. Thus, in the future we will also consider other reference points, such as the gnathion that, due to its lower position on the chin, may be less affected by lower lip movements than the central part of the chin [15]. Moreover, future works will focus on the comparison among various camera setups (in terms of distance, video resolution, etc.), different face tracking algorithms, and different head positions and rotations (a major problem in face tracking) in order to simulate real-world conditions. These factors, along with the recruitment of a larger sample of subjects, will allow us to find the optimal setting for improving the system's performance.

5. Conclusion

Tracking of the jaw is a very challenging task, which becomes even harder when attempting to track it without sensors or markers. In this study, we provided preliminary results regarding the accuracy of jaw tracking with a video-based method composed of a depth sensor and a face tracking algorithm. We also provided suggestions to improve the tracking method such as uniform lighting and point selection optimization. The constant development of novel face tracking algorithms in conjunction with more powerful depth sensors provides novel opportunities to track articulatory movements in clinical populations and clinical settings, in order to develop accurate systems that will help clinicians in the assessment and rehabilitations of motor speech disorders.

6. Acknowledgements

The funding for this work was provided by the Canadian Partnership for Stroke Recovery (CPSR) Collaborative Catalyst Grant and the CPSR Trainee Award. The work was also supported by the University Health Network: Toronto Rehabilitation Institute.

7. References

[1] Y. Yunusova, J. R. Green, M. J. Lindstrom, L. J. Ball, G. L. Pattee, L. Zinman, "Kinematic of disease progression in bulbar ALS," *Journal of Communication Disorders*, vol.43, no. 1, pp. 6-20, 2010.

[2] B. Walsh, A. Smith, "Basic parameters of articulatory movements and acoustics in individuals with Parkinson's disease," *Movement Disorders*, vol. 27, no. 7, pp. 843-850, 2012.

[3] S. Shellikeri, J. R. Green, M. Kulkarni, P. Rong, R. Martino, L. Zinman, Y. Yunusova, " Speech movement measures as markers of bulbar disease in amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 5, pp. 887-899, 2016.

[4] Y Yunusova, G. Weismer, J. R. Westbury, M. J. Lindstrom "Articulatory movements during vowels in speakers with dysarthria and healthy controls," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 3, pp. 596-611, 2008.

[5] D. A. Robin, C. Bean, J. W. Folkins, "Lip movement in apraxia of speech," *Journal of Speech and Hearing Research*, vol. 32, no. (3), pp. 512-523, 1989.

[6] D. A. Hayden, P. A. Square, "Motor speech treatment hierarchy: a system approach," *Clinics in Communication Disorders*, vol. 4, no. 3, pp. 162-174, 1994.

[7] M. M. Earnest, L. Max, "En route to the three-dimensional registration and analysis of speech movements: Instrumental techniques for the study of articulatory kinematics," *Contemporary Issues in Communication Science and Disorders*, vol. 30, pp. 5-25, 2003.

[8] A. Bandini, S. Ouni, P. Cosi, S. Orlandi, C. Manfredi, "Accuracy of a markerless acquisition technique for studying speech articulators," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, Proceedings*, 2015, pp. 2162-2166.

[9] Y. Feng, L. Max, "Accuracy and precision of a custom camera-based system for 2-D and 3-D motion tracking during speech and nonspeech motor tasks," *Journal of Speech, Language, and Hearing Research*, vol. 57, pp. 426-438, 2014.

[10] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia*, vol.19, no. 2, pp. 4-10, 2012.

[11] X. Xiong, F. De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, *June 23-28, Portland-OR, USA*, pp. 532-539, 2013.
[12] A. Bandini, S. Orlandi, F. Giovannelli, A. Felici, M. Cincotta, D. Clemente, P. Vanni, G. Zaccara, C. Manfredi, "Markerless analysis of articulatory movements in patients with Parkinson's disease," *Journal of Voice*, vol. 30, no. 6, pp. 766.e1-766.e11, 2016.
[13] S. Ouni, S. Dahmani, "Is markerless acquisition technique adequate for speech production?," *The Journal of the Acoustical Society of America*, vol. 139, no. 6, pp. EL234-EL239, 2016.
[14] Y. Tanaka, T. Yamada, Y. Maeda, K. Ikebe, "Markerless three-dimensional tracking of masticatory movements," *Journal of Biomechanics*, vol. 49, pp. 442-449, 2016.
[15] J. R. Green, E. M. Wilson, Y. T. Wang, 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.
[16] J. Berry, "Accuracy of the NDI Wave Speech Research System," *Journal of Speech, Language, and Hearing Research*, vol. 54, pp. 1295-1301, 2011.
[17] https://software.intel.com/en-us/intel-realsense-sdk/documentation Accessed: March 14, 2017.
[18] J. M. Bland, D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol 327, no. 8476, pp. 307-310, 1986.
[19] A. Bandini, S. Ouni, S. Orlandi, C. Manfredi, "Evaluating a markerless method for studying articualtory movements: application to a syllable repetition task," in *9th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA), September 2-4, Firenze, Italy, Proceedings*, 2015, pp. 99-102.