EVALUATING A MARKERLESS METHOD FOR STUDYING ARTICULATORY MOVEMENTS: APPLICATION TO A SYLLABLE REPETITION TASK

A. Bandini[1,2], S. Ouni[3], S. Orlandi[1], C. Manfredi[1]

[1]Department of Information Engineering, Università degli Studi di Firenze, Firenze, Italy
[2]Department of Electrical, Electronic and Information Engineering (DEI) "Guglielmo Marconi", Università di Bologna, Bologna, Italy
[3]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France

Abstract: **The analysis of the articulatory movements allows investigating the kinematic characteristics of some speech disorders. However, the methodologies most used until now, as electromagnetic articulography and optoelectronic systems, are expensive and intrusive which limit their use to specialized laboratories. In this work, we use a completely markerless and low-cost technique to study lip movements during a syllable repetition task. By means of a Kinect-like and an existing face tracking algorithm, we are able to track the movements of the lower lip, testing the performances against a reference method (marker-based optoelectronic system). Good results were obtained in terms of RMSE for the tracking of the lower lip during the repetitions. Some kinematic measures, as opening and closing velocities and accelerations, were also computed. Despite the limitations in terms of image resolution, these results are very promising in the optic of developing a new markerless system for studying speech articulation.**

Keywords : **speech articulation, markerless, Kinect, contactless, accuracy evaluation**

I. INTRODUCTION

Kinematic analysis of the articulatory movements (i.e., the movements of tongue, lips and jaw) allows investigating the characteristics of some speech disorders, like hypokinetic dysarthria. Walsh *et al.*, 2012 [1] studied jaw and lower lip movements in patients with Parkinson's disease (PD) using an optoelectronic system. They demonstrated that these patients exhibit reduced ranges of movements and velocities of jaw and lips during the pronunciation of plosive consonants. Yunusova *et al*., 2008 [2] studied the articulatory movements in patients with hypokinetic dysarthria due to PD and amyotrophic lateral sclerosis (ALS) by means of the X-ray microbeam technique, tracking the position of several markers located on tongue, lips and jaw. In that work, they showed that tongue movements in PD and ALS patients could be more discriminative in the comparison with healthy subjects, although there are also alterations in lips and jaw movements.

Wong *et al.*, 2011 [3] has also investigated articulatory movements in patients with speech disorders, where the tongue kinematics is studied by means of electromagnetic articulography (EMA), in dysarthric and non-dysarthric PD patients. They demonstrated that both categories exhibit different patterns of tongue movements with respect to healthy subjects.

These researches show clearly that techniques for studying movements are useful to describe the kinematic characteristic of the articulatory organs in dysarthric patients. However, the methodologies most used until now (EMA, optoelectronic systems, X-ray techniques, etc.), which are actually very accurate, have the big disadvantage of being expensive which limit their use to specialized laboratories [4]. Moreover, some of these techniques need long and tedious preparation protocols, resulting in a discomfort for patients. Thus, the use of these methodologies for studying speech articulation is limited to the research field. In order to broaden the kinematic studies of speech articulation, (e.g., for speech therapy purposes, or to track the disease progression), the use of a low-cost and fully contactless system would be desirable.

In the last five years the spreading of 3D video sensors (like Microsoft Kinect), has revolutionized the world of videogames and not only, providing new possibilities to study body movements without any sensor attached to the subject. These devices, unlike a normal camera, provide a 3D information about the observed scene. Even for speech therapy purposes, some applications with the Kinect sensor has been proposed, in order to study and automatically identify the therapeutic exercises that involve facial movements [5]. To our knowledge, no existing work has tested the accuracy of a fully markerless technique to study speech articulation. For these reasons, our aim is to test the performance of a system composed by a 3D depth sensor and a face tracking algorithm in order to track

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

lip movements during speech. In this study the accuracy is verified against an established optoelectronic method.

II. Methods

The markerless system proposed in this study is composed of a 3D structured light sensor (Primesense Carmine 1.09) and an existing face tracking algorithm [6], in order to study lips movements in the 3D space without any sensor attached to the subject's skin.

Two healthy volunteers (an Italian native speaker and a French one) were recruited for the experiment. The speech task consists in the repetition of the syllable /pa/ for at least 30 times with a single breath. The acquisitions were performed in a room with reduced environmental noise. Each subject had to repeat the syllables avoiding large head movements. The subjects' face was kept under constant and uniform illumination during the whole acquisition time.

Markerless system: during the experiments the subjects' faces were acquired by means of the depth sensor Primesense Carmine 1.09. This device was chosen for its ability to work at short distances (0.4-1.5 m), thus appropriate for face movements. As classical structured-light sensors, it provides two video streams: the color video (like a normal webcam) and the depth stream, where the pixels of each frame code the distance of a point in the scene from the camera plane. The image resolution of both streams was set at 320 x 240 pixels. Both videos were acquired synchronously at 30 frames per second, and stored as avi files by means of the OpenNI (ver. 2.2) and OpenCV (ver. 2.4.9) libraries.

The device was located in front of the subject's face (at the height of the mouth) at a distance around 0.7-0.8 m from the lips, according to the specifications provided by the manufacturer.

For the automatic identification of the facial features, the tracking algorithm *Intraface* was used. This algorithm fits to the video frames a face model composed of 49 points, on the basis of texture descriptors like SIFT (Scale-Invariant Feature Transform) [6, 7]. This algorithm was chosen for its robustness against illumination changes, for its ability to describe asymmetrical face movements (very important in the context of speech therapy applications) and for its efficiency [6]. In particular, lips are modeled as a set of 18 points: 12 on the outer border and 6 on the inner border. In the case of our study, only 7 points on the outer border were considered for the analysis (Fig. 1) to compare the performance of the system against the marker-based method.

Since the coordinates of the points that were computed with the *Intraface* tracker are on the image plane, a further step to extract the 3D locations of the points of interest is required, as this algorithm works only on the color image. Starting from the coordinates on the image plane and using the depth values (Z) retrieved from the depth image, it was possible to calculate the 3D coordinates in mm. In fact, before each acquisition the color and depth frames were aligned and synchronized, then, we just sampled the depth image in the same pixel coordinates of the model points provided by the tracker.

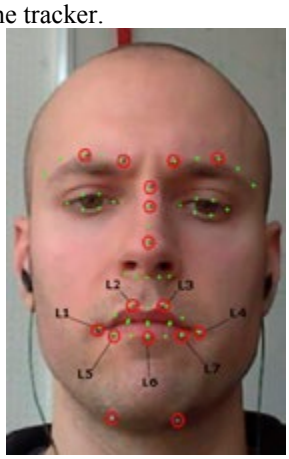

Fig. 1: *Intraface* tracker model points (green dots) and optical markers locations (red circles). The markers were located in the same position of some model points, in order to estimate the 3D rigid transformation to register the two sets of points.

According to the scheme in Fig. 2, we calculated the X and Y coordinates with the following formulas [8]:

$$X = Z\,\frac{(x-c_x)}{f} \qquad with\ f = \frac{W}{2}\left[\tan\left(\frac{FOV_h}{2}\right)\right]^{-1} \quad (1)$$

$$Y = Z\,\frac{(y-c_y)}{f} \qquad with\ f = \frac{H}{2}\left[\tan\left(\frac{FOV_v}{2}\right)\right]^{-1} \quad (2)$$

Where x and y are the coordinates on the image plane (in pixels) of the 3D point $[X\ Y\ Z]^{\mathrm{T}}$, (c_x, c_y) are the coordinates (in pixels) of the principal point (i.e. the point where the optical axis intersects the image plane) of the color camera, f is the focal length (in pixels) of the camera, W and H are the dimensions of the image in pixels (width and height, respectively), FOV_h and FOV_v are the horizontal and vertical field of view of the camera (58° and 45° respectively).

Marker-based system: To compare the performance of the aforementioned markerless method, we used an optoelectronic system (Vicon Motion Systems Ltd., UK) as a reference. This system was composed by four cameras (MX3+ model) with special optics for near range applications. Sixteen reflective markers of 3mm diameter were glued on the faces of the subjects. This size is suitable to study facial movements without interfering with the face tracker.

Before each acquisition, the markers were accurately located in some precise facial points defined by the

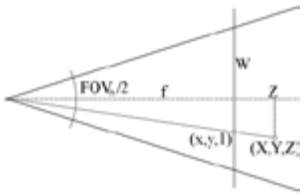

Fig. 2: Pinhole camera model. This model was used to retrieve the 3D coordinates of the face points (estimated with the markerless system), starting from the image coordinates plus the depth information Z (according to equations 1 and 2).

Intraface model: two for each eyebrow, three on the nose, seven on the outer border of the lips (one for each corner – L1 and L4, two on the upper lip – L2 and L3 – and three on the lower lip – L5-L7) and two on the chin (Fig. 1). The 3D trajectories of these markers were acquired synchronously using the markerless system at 100 Hz and reconstructed using the Vicon Nexus software.

Data Processing: To compare the trajectories of the points of interest extracted with the markerless system with those of reference, the two sets of points must be aligned in the space, since the two reference frames are different. To do this, since we paid a lot of attention to locate the markers in the same position as some *Intraface* points, the 3D rigid transformation that allows mapping the markerless points in the marker-based reference frame can be estimated. Using pairs of corresponding points provided by the two systems, the rotation matrix R and the translation vector T were estimated through a least squares solution and this transformation was applied to each point extracted from the markerless system.

Once the two sets of points were aligned, the trajectories extracted with the markerless system were resampled at 100 Hz using a spline interpolation technique. In this way, the comparison with the reference trajectories was possible by means of the calculation of the root-mean-square error (RMSE) in mm, according to the following formula:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (3)$$

Where N is the number of samples of the trajectories during a single repetition, y_i is the *i-th* sample of the marker-based trajectory and $\hat{y}_i$ is the corresponding sample extracted from the markerless trajectory.

Afterwards, for each syllable repetition the following kinematic parameters (for both systems) were computed: the maximum velocity (V_{open}) and acceleration (A_{open}) during the opening phase, the maximum velocity (V_{close}) and acceleration (A_{close}) during the closing phase. These parameters were calculated differentiating in time the trajectory on the vertical axis of the central point of the lower lip (point L6 in Fig. 1). V_{open} was calculated as the minimum speed value during the first half of the repetition, while V_{close} was identified as the maximum speed value from the time instant of V_{close} up to the end of the utterance (Fig. 3). The same criteria were adopted to extract A_{open} and A_{close} from the acceleration values of the same lip point (Fig. 3). Moreover, for each syllable repetition the Pearson's correlation coefficient between trajectories, velocities and accelerations extracted with both systems was computed. Correlation values close to 1 indicate that the trends of displacement, speed and acceleration calculated with the proposed method are very similar to the ground truth, as shown in Fig. 3.

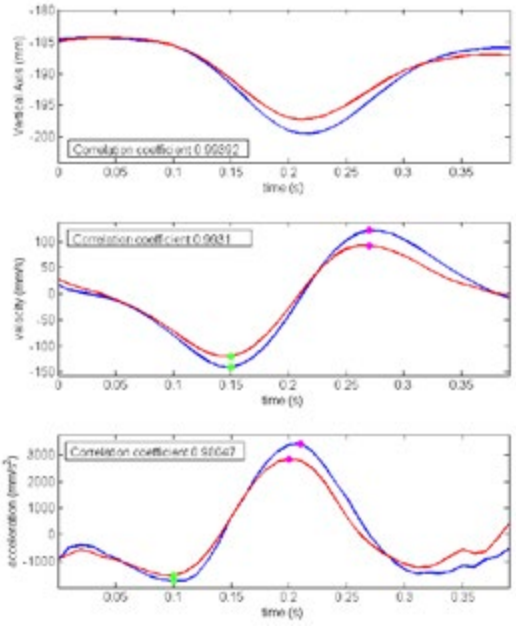

Fig. 3: vertical trajectory of the central point of the lower lip (upper plot) during the repetition of the syllable /pa/; speed (central plot) and acceleration (lower plot) on the vertical axis. The blue lines are relative to the reference method (marker-based), while the red lines are estimated with the markerless technique. The green points indicate the maximum velocities and accelerations during the opening phase, while the magenta points are the maximum velocities and accelerations during the closing phase.

III. Results

The analysis was conducted on a total of 80 utterances. RMSE values of the central point of the lower lip (point L6, Fig. 1) were around 2 mm on the three axes, respectively (1.89 ± 0.82) mm on the lateral axis, (1.61 ± 0.45) mm on the frontal axis and (2.16 ± 0.70) mm on the vertical axis. The kinematic parameters (mean values and standard deviations) for speed and acceleration during opening and closing phases relative to the same point were reported in Tab. 1.

The correlation coefficient for the trajectory on the vertical axis of the point L6 was (0.96 ± 0.03), while those for speed and acceleration were respectively (0.95 ± 0.05) and (0.88 ± 0.10).

Tab.1: Mean values and standard deviations of the kinematic parameters during the opening and closing phases of the syllable repetition

	Marker-based	Markerless
V_{open} (mm/s)	-114.37 ± 30.55	-96.39 ± 23.48
V_{close} (mm/s)	100.01 ± 45.82	79.77 ± 26.97
A_{open} (mm/s^2)	-1689.63 ± 559.86	-1759.10 ± 665.73
A_{close} (mm/s^2)	2619.02 ± 1068.82	2141.91 ± 910.86

IV. Discussion

Good results were obtained in term of accuracy with RMSE for the point L6 around 2 mm on the three axes. The kinematic parameters reported in Tab. 1 show a tendency to underestimate the module of the maximum and the minimum speed values (closing and opening phases) with differences around 20 mm/s. An underestimation is visible also for the closing acceleration, while during the opening phase the two estimates seem to be closer.

Although the results on kinematic parameters seem to be inconsistent, from the plot in Fig. 3 and from the correlation values between the two systems, it is possible to observe that the trajectories, the velocities and the accelerations extracted with the markerless technique were very similar when compared with the reference. This suggests that a bias is present in the estimation of the kinematic parameters.

This bias might be due to the distance from the face at which the device was located (about 0.8 m), or to the different framerate of the systems (30 Hz for the depth sensor, 100 Hz for the marker-based method). This distance was a trade-off between the need to move the sensor as close as possible to the subject's face and its characteristic (range of work: 0.4-1.5 m), without interfering with the field of view of the Vicon cameras. The distance, in conjunction with the low image resolution (320 x 240 pixels) probably explain these differences. However, further experiments with structured light sensors should consider an experimental design with higher frame resolutions (at least 640 x 480 pixels) and smaller distances from the subject's face (i.e., 0.5-0.6 m, according to the specification provided by the manufacturer).

V. Conclusion

In this work, we have introduced a fully contactless and low-cost method to track the articulatory movements, in particular those relative to the lower lip during a syllable repetition task. We demonstrated that good accuracies could be reached in terms of RMSE with respect to a marker-based reference methods. These results are very promising in the optic of developing new systems to study speech articulation, that could be implemented also in domestic environments. This would allow enlarging the number of patients who undergoes to speech therapy, in particular elderly people who suffer from hypokinetic dysarthria due to Parkinson's disease.

Further developments will be oriented to test the performance of this system with different configurations (image resolution, distance from the camera), as well as to use this contactless technique with PD patients (but not only), in order to check if it is possible to highlight kinematic differences due to the dysarthria with respect to healthy control subjects, as demonstrated with other more expensive techniques [1-3].

References

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

[2] 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.

[3] M.N. Wong, B.E. Murdoch, B. Whelan, "Lingual kinematics during rapid syllable repetition in Parkinson's disease", *International Journal of Language and Communication Disorders*, vol. 47, no. 5, pp. 578-588, 2012.

[4] 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.

[5] C. Lanz, B.S. Olgay, J. Denzler, H-M. Gross, "Automated classification of therapeutic face exercises using the kinect", *Proceedings of the 8th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2013), Barcelona, Spain,* pp. 556-565, 2013.

[6] 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.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[8] R. Szeliski, "Computer vision: algorithms and applications", Springer London, 2010.