Accuracy of a markerless acquisition technique for studying speech articulators

Andrea Bandini[1,2], *Slim Ouni*[3], *Piero Cosi*[4], *Silvia Orlandi*[1], *Claudia Manfredi*[1]

[1] DINFO-Università degli Studi di Firenze, Firenze, Italy
[2] DEI-Università di Bologna, Bologna, Italy
[3] Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[4] ISTC-CNR, UOS Padova, Italy

{andrea.bandini, silvia.orlandi, claudia.manfredi}@unifi.it, slim.ouni@loria.fr,
piero.cosi@pd.istc.cnr.it

Abstract

The main disadvantages of the existing methods for studying speech articulators (such as electromagnetic and optoelectronic systems) are the high cost and the discomfort to participants or patients. The aim of this work is to introduce a completely markerless low-cost 3D tracking technique in the context of speech articulation, and then compare it with a well-established marker-based one to evaluate the performances. A Kinect-like device was used in conjunction with an existing face tracking algorithm to track lips movements in 3D without markers. The method was tested on two subjects uttering 200 words and 100 sentences. For most of points of the lips the RMSE ranged between 1 and 3 mm. Although the image resolution used in this experiment was low, these results are very promising. Nevertheless, further studies should consider higher video resolutions in order to obtain better results.

Index Terms: speech articulation, markerless, Kinect sensor, comparison

1. Introduction

In the past decades, several techniques were proposed and used for studying the movements of speech articulators (lips, tongue and jaw) as x-ray imaging, magnetic resonance imaging (MRI), ultrasound technique, electromagnetic articulography (EMA) and optoelectronic systems [1]. Applications of these methods may include: the study of speech disorders in neurological illnesses (Parkinson's disease, amyotrophic lateral sclerosis, etc.) using optoelectronic techniques [2], [3], EMA [4] and x-ray [5]; the use of EMA to estimate the parameters of an articulatory model [6]; the study of tongue movements for speech therapy applications [7], [8].

The main disadvantages of these methods are the high cost and its discomfort to participants or patients. Moreover, the above techniques needs a lengthy preparation protocol. To our knowledge, one of the few attempts to go beyond this limit was a marker-based system composed by 2 consumer-grade cameras in conjunction with a tracking software to study lips and jaw movements, presented in Feng *et al.* (2014) [9]. Nevertheless, it needs some preparation, as gluing markers, setting and calibrating the cameras, and can still present discomfort to patients.

The spread of 3D low-cost structured light 3D sensors (like Microsoft Kinect, Asus Xtion, Primesense Carmine, etc.), providing 3D information of the observed scene without markers, could be used to extract trajectories and kinematic parameters in the 3D space, and to analyze some fundamental articulatory parameters like lip protrusion. Moreover, these devices could be integrated in speech therapy applications since most of tasks consist of tracking facial and articulatory movements and providing some feedback [10].

Speech therapy can address the slowdown of speech disorders related to neurological illnesses, as hypokinetic dysarthria associated with Parkinson's disease or stroke [7], [10], [11]. Although speech disorders concern to some extent a large population, speech therapy is often applied to a small number of patients. This is due to the following factors:

- During group sessions, the speech therapist has difficulty to give exactly the same rigorous attention to each patient, in order to evaluate the therapy exercises and provide a valuable feedback to patients;
- Due to neurodegenerative diseases, most of the patients with hypokinetic dysarthria are elderly people, who could encounter physical difficulties to visit specialized centers;
- In several cases, patients should continue the therapy exercises at home. They might not be sufficiently motivated without the direct supervision of the therapist.

For these reasons, it arises the need for a system, which could automatically provide a feedback about the articulatory movements and that could also be integrated to home environment. Therefore, this system should be as much as possible at reasonable price and using a contact-less technique.

In order to use marker-less methods for studying articulatory movements, it is necessary to test their accuracy during the tracking of facial movements. Thus, the aim of this study is to compare the performance of a low-cost marker-less technique against a well-established marker-based one, to track articulatory movements, mainly focusing on lip movements during speech. The use of commercially available 3D sensors would help for this aim. In this case we used a Kinect-like sensor in conjunction with an existing face tracking algorithm called *Intraface* [12], that fits a face model to video frames, on the basis of SIFT texture descriptors [13]. We have used this algorithm for its robustness against illumination changes, for its performances and for its capability to generalize to cases never seen during training [12].

In this paper we present our method as well as to provide guidelines for using Kinect-like sensors to study articulatory movements.

2. Materials and Methods

To test the performances of the marker-less system (based on the Primesense Carmine 1.09 sensor, which is a kinect-like system, and the *Intraface* tracking algorithm) to track articulatory movements, we used an optoelectronic technique

(Vicon Motion Systems Ltd.) widely used as accurate marker-based motion capture system. Both systems (Primesense and Vicon) were used simultaneously and the different streams were acquired synchronously. In the following, we present the systems used in our study and we describe the method.

2.1. Marker-based stream

A Vicon system based on 4 cameras (MX3+ model) with special optics for near range applications has been used. This system allows tracking 3mm-diameter reflective markers, well adapted to facial movements [9]. The camera location, shown in Fig. 1, was chosen to cover as much as possible the lower part of the face. The acquisition of the 3D positions of the markers were provided by the Vicon Nexus software at sampling rate of 100Hz.

2.2. Video stream (Color and Depth)

We have used a 3D sensor, the Primsense Carmine 1.09, suitable for near range applications (0.4-1.5 m), as for facial movements. As other structured-light sensors, it provides two video streams: color (RGB) and depth, where each pixel codes the distance in mm of a particular point in the scene from the camera plane. The video acquisitions were performed using OpenNI and OpenCV libraries. The video resolution was 320 x 240 pixels at 30 frames per second for both streams. We placed the Primsense sensor as close as possible toward the subject's face, without interfering with the field of view of the Vicon cameras. The device was fixed on a boom at a distance from the subject's head between 0.7 and 0.8 m (Fig. 1). The audio stream was simultaneously acquired from the two built-in microphones of the Primsense.

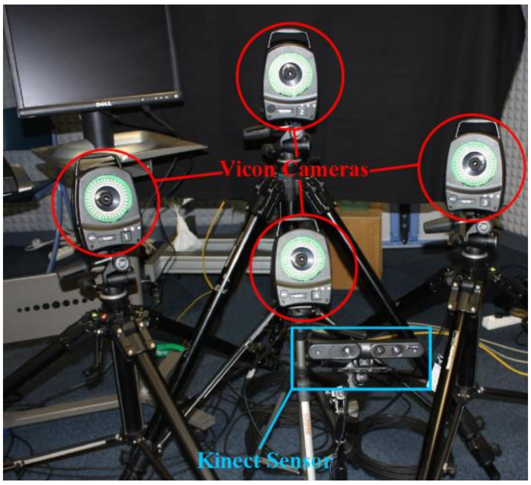

Figure 1: *Experimental setting: marker-based system and Kinect-sensor displacement.*

2.3. Speech corpus and acquisitions

The acquisitions were performed in a reduced noise. Two healthy subjects were recruited for the experiment: an Italian native speaker and a French one. Subjects were seated in front of the camera at a distance between 0.7 and 0.8 m from the Primesense sensor. This range is a tradeoff between the device characteristics and its distance from the subject's face (as close as possible) without interfering with the field of view of the Vicon cameras.

Before each acquisition, 16 reflective markers were accurately glued on the subject's face in precise anatomical points: 4 on the eyebrows, 3 on the noset, 7 on the external contour of the lips and 2 on the chin, as shown in Fig. 2. These locations were accurately chosen in correspondence to selected points of the face model used by the *Intraface* tracker (presented in the next section), as shown in Fig. 2. We verified that the markers did not alter the acquisition quality of the *Intraface* tracker.

Each subject was asked to read and pronounce the corpus (displayed on screen in front of the subject) without any excess of head-movement. The face was kept under a constant and uniform illumination during the whole acquisition.

We chose two corpora (one for each language), both composed of 50 meaningful sentences and 100 meaningful words. The French sentences were extracted from of the Comberscure corpus [14], while the words were chosen from the Lafon lists [15]. The Italian sentences and words were chosen from the corpus defined by Bocca and Pellegrini [16].

2.4. Data processing

2.4.1. Face tracking

As mentioned above, to identify some facial feature points (as the lips) a tracking algorithm capable of detecting and tracking these points, is needed. The face tracker *Intraface* used in this work fits a model of the face to the color image using texture descriptors (SIFT) to resolve this optimization problem [12], [13]. Unlike other face trackers based on *a priori* learned face model, as active appearance models [17], [18], each landmark position is directly optimized to the current frame based on texture descriptors. This involves a better ability to generalize situations never seen in the training set, like asymmetrical face movements, leading to a higher flexibility. Moreover, since the fitting is based on SIFT descriptors, this algorithm is robust against illumination changes [12]. The *Intraface* tracker fits to the scene a model composed of 49 points: 10 for the eyebrows, 12 for the eyes, 9 for the nose and 18 for the lips (12 on the outer contour, 8 on the inner contour) as shown in Fig. 2. For this study we considered the points of the eyebrows, nose and of the outer lips contour.

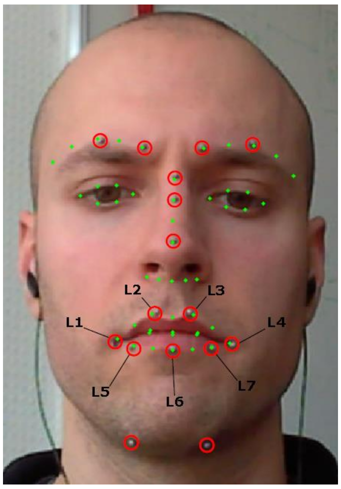

Figure 2: *Intraface model points (green dots) and optical markers located on the face (red circles).*

Since the face tracker works on 2D color images, it was necessary to extract the 3D coordinates of the points of interest. Thus, for each of these points we computed the coordinates on the lateral axis (X) and on the vertical axis (Y) starting from the depth information (frontal axis – Z). Since we registered and synchronized the depth frames with the color frames, to extract the Z value in mm for each point we just needed to sample the depth image in the same pixel coordinates provided by the face tracker. To calculate the other two coordinates of the points we used the following formulas [19]:

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

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

where x and y are the coordinates on the image plane (in pixels) of a point of coordinates $[X\ Y\ Z]^{\mathrm{T}}$ in the 3D space, f is the focal length (in pixels) of the camera, (c_x, c_y) are the coordinates (in pixels) of the principal point, W and H are the dimensions of the image (width and height, respectively) in pixels, FOV_h and FOV_v are the horizontal and vertical field of view (equal to 57.5° and 45°, respectively).

After computing the 3D coordinates of the points tracked with the marker-less system, it is necessary to align them with those acquired with the marker-based method, since the coordinates of the frame references are different. For this reason, since we paid a lot of attention in the positioning of the markers on the subject's face, it was possible to estimate the rotations and translation parameters which allow mapping the 3D points extracted with the face tracker in the Vicon coordinate system:

$$P'_k = \boldsymbol{R}\, P_k + T \tag{3}$$

where P_k is a generic point of coordinates $[X_{\mathrm{k}}\ Y_k\ Z_k]^{\mathrm{T}}$ in the marker-less reference frame mapped to the Vicon reference frame (P'_k) through the 3x3 rotation matrix $\boldsymbol{R}$ and the translation vector T. Thus, knowing couples of corresponding points in the two reference frames, it was possible to estimate the transformation parameters. Using a least squares solution and making use of more pairs of points than those required by the number of unknowns, we overestimated the system reducing the effect of noise on the estimation.

For this work we used 7 pairs of points, respectively: two for each eyebrow, two for the nose and one for the lips (midpoint on the lower lip – L6, Fig. 2).

2.4.2. Articulatory parameters and error calculation

In addition to conducting the comparison on the 3D coordinates of the points of interest (points L1-7 in Fig.2), we computed some articulatory parameters:

- Lip width: distance on the lateral axis between the two corner points (points L1 and L4);
- Lip opening: distance on the vertical axis from the midpoint between points L2 and L3, and the central lower lip point (point L6);
- Lip protrusion: distance on the frontal axis from the midpoint between points L2, L3 and L6 and a fixed reference point, in this case the nose tip.

All these parameters were normalized with respect to head rotation angles. These angles were calculated from the markers located on the eyebrows and nose.

After extracting the trajectories and the articulatory parameters we calculated the Root Mean Square Error (RMSE) between the marker-less and the marker-based measures (points trajectories and articulatory parameters).

2.4.3. Depth accuracy

The manufacturer of the Primesense sensor provided only the spatial resolution at 0.5 m from the camera, equal to 1 mm for the depth and 0.9 mm for the other two axes. Since our experiments were performed at a distance between 0.7 and 0.8 m, we expected lower resolutions. To estimate the error introduced by the Kinect sensor in the estimation of the depth value of a point in the scene (and thus for the estimation of its 3D coordinates), we used a phantom object composed by a box on which 7 reflective markers (of the same type used during experiments) were glued on small squares of yellow paper (surface = 1 cm^2) on one surface of the box. These markers were located like a cross: the points on the horizontal axis were equally spaced of 25 mm, while those on the vertical axis were spaced of 50 mm. The Primesense sensor and the Vicon cameras were placed in front of a table on which the box translated at a constant speed from a distance of 900 mm to 500 mm from the depth device. This test was repeated twice.

We calculated the mean RMSE for the 7 points along the 3 coordinates for the entire range of movement covered by the box. Finally, dividing this range into intervals of 20 mm, within which the mean RMSE was computed, 20 error values were obtained.

In order to verify if the errors introduced by the Primesense sensor in the estimation of the depth values was comparable to the RMSE values obtained during speech acquisitions, we extracted the mean distance of the lips (in mm) from the camera.

3. Results

The acquisition relative to the French subject was carried out at a mean distance of the mouth from the Primesense sensor of (737.99 ± 41.88) mm, while those relative to the Italian subject were performed at a mean distance of (770.20 ± 14.14) mm.

The mean values and the standard deviations of the RMSE for the 3D trajectories of the 7 points of interest as well as the RMSE for the 3 articulatory parameters were reported in table 1. These results were computed on the whole corpus. Since the points extracted with the marker-less method were mapped to the Vicon reference frame, the notation used in this table refers to Vicon coordinates system: x is the lateral axis, y is the frontal axis and z is the vertical axis.

Considering the mean distances at which the experiments were performed, the mean errors introduced by the Primesense sensor, for the French corpus were 0.99 mm on the lateral axis, 0.83 mm on the vertical axis and 1.13 mm on the frontal axis. For the Italian corpus, the mean errors were 1.17 mm on the lateral axis, 0.81 mm on the vertical axis and 1.21 mm on the frontal axis.

As shown in tab. I, the mean values of RMSE relative to the French corpus ranged between 1 and 3 mm (except for the coordinate y of the points L5 and L7 and the coordinate x of the point L7). Even in the Italian corpus we noticed mean values under 3 mm, although errors higher than 3 and 4 mm are more frequent. In particular, for the coordinate y of the points L2, L3, L5 and L7 the RMSEs ranged between 4 and 5 mm.

Considering the depth accuracy of the device, the error for the depth values (z-axis) was approximately of 0.7 mm for z < 600 mm, ranged between 0.8 and 1.2 mm for z between 600 mm and 800 mm and is about 2 mm for z > 800 mm. Similar values were present on the x-axis, although the increase of the

RMSE values over 2 mm occurs after 850-870 mm of distance. On the y-axis (vertical), the error seems to be constant along the entire range covered by the object. In fact, we found that the RMSE ranged between 0.7 mm and 1.2 mm in the considered distance range.

Table 1. *RMSE (mean values and standard deviations) for the 7 points of interest and for the 3 articulatory parameters*

Points and	French RMSE (mm)			Italian RMSE (mm)		
Parameters	x	y	z	x	y	z
L1	1.63 ± 0.99	2.15 ± 0.96	2.57 ± 1.53	2.91 ± 2.30	2.35 ± 1.54	1.61 ± 1.01
L2	1.39 ± 0.82	2.98 ± 1.44	1.40 ± 0.96	2.79 ± 1.85	4.17 ± 2.17	1.50 ± 0.99
L3	1.28 ± 0.95	2.99 ± 1.49	1.50 ± 1.04	5.62 ± 3.06	4.14 ± 1.68	1.57 ± 1.39
L4	2.10 ± 1.14	1.53 ± 0.85	2.23 ± 1.51	2.54 ± 2.08	1.85 ± 1.27	1.92 ± 1.21
L5	1.30 ± 0.80	3.43 ± 1.40	2.46 ± 1.52	2.93 ± 2.30	5.50 ± 2.81	2.36 ± 1.27
L6	1.96 ± 1.13	2.47 ± 1.39	2.49 ± 1.61	2.85 ± 2.49	2.92 ± 1.85	2.47 ± 1.31
L7	3.15 ± 1.54	6.50 ± 2.08	2.44 ± 1.59	3.29 ± 2.73	5.08 ± 2.62	2.63 ± 1.51
Width	2.07 ± 1.22			1.86 ± 1.07		
Opening	2.72 ± 1.66			3.81 ± 2.39		
Protrusion	1.36 ± 0.81			4.45 ± 2.08		

4. Discussion

For most of the points, the RMSE mean values ranged between 1 and 3 mm (Tab. I). Considering the low image resolution used for the experiment, this is a very promising result. Further information about acceptable error ranges could be provided by future works that might try to extract error measures with other algorithms, devices and configurations.

Concerning the articulatory parameters (Tab. I), we obtained good results for the width in both corpora (mean RMSE around 2 mm) and for opening and protrusion in the French corpus (mean values: 2.72 mm and 1.36 mm, respectively). Instead, the errors for opening and protrusion in the Italian corpus were higher, with values over 4 mm for protrusion. Since protrusion is computed from the y-coordinate of points L2, L3 and L6, bigger errors for these points could lead to a bigger protrusion error; in fact, the error on the y-coordinate for these 3 points is always bigger in the Italian corpus with respect to the French one (Tab. I).

These differences could be due to several factors. First of all, the markers were accurately positioned to match the *Intraface* points, but this positioning (as far as accurate may be), presents an intrinsic error due to the manual settings. For the opening, the higher RMSE could be due to the higher distance of the acquisition and different orientation of the Primesense with respect to the face that, in conjunction with the low resolution of the images, could lead to bigger errors.

Considering the depth accuracy of the device we found that the errors on the x (lateral) and z (frontal) axes exhibited similar behaviors, with higher values with increasing distance from the camera. Considering the distances at which the speech acquisitions were performed (0.7-0.8 m) the errors in the depth estimation reflected what has already been observed in the results of table I, namely for most of points the biggest error is on the frontal axis (that in the Vicon reference frame is the y-axis). However, this error is always smaller than those reported in Tab. I (only in the case of the point L4 might be similar, since it is between 1.5 mm and 1.8 mm). Even for the other two coordinates the device errors were lower than those computed during speech experiments. We believe that this is due to the low resolution of the video frames (320 x 240 pixels) and the distance of the Primesense to the subject. This means that the pixels of the color images (those on which the face tracker works) correspond to an area of the face larger than that which would be using a higher resolution and/or decreasing the distance between the face and the Primesense camera. Thus, some depth variations of the face (in particular the cavities in the corner of the mouth due to the lip anatomy) might be indistinguishable.

For these reasons for further experiments that will involve structured light cameras (Microsoft Kinect, Primesense, Asus Xtion, etc.) we strongly recommend to acquire images of at least 640 x 480 pixels of resolution for both streams.

Although the working range used for this experiment led to reasonable errors in the estimation of the 3D coordinates another requirement to adopt for future experiments is to perform the acquisitions at distances lower than 0.7 m. In fact, according to our results and considering the range of working specified by the manufacturer (0.4-1.5 m), the resolution on the 3 axes should be better at distances between 0.4 and 0.6 m from the camera. This recommendation allows increasing indirectly the resolution. However, due to the technical design of this experiment, we could not bring the camera closer to the subject's face.

5. Conclusion

This work is a first attempt to test the performance of a completely low-cost marker-less technique against a well-established marker-based one, to track articulatory movements during speech. Using a depth sensor in conjunction with an efficient face-tracking algorithm, it is possible to obtain good accuracies to analyze lip movements during speech, despite the limitations in terms of low-resolution images and distance from the camera. Moreover, these promising results are encouraging in order to achieve marker-less low-cost techniques to study facial movements for speech therapy purposes. This would result in the increasing of the percentage of patients undergoing speech therapy, bringing benefits in particular to elderly patients who cannot move to specialized centers, and for children which could deal with the rehabilitation exercises in the form of interactive games. From the results of this study, any new implementation or experiment in this perspective should be performed at distances of the face between 0.5 m and 0.6 m from the camera, acquiring images of resolution greater than 640 x 480 pixels. Since this paper presents overall results, further developments will concern the analysis of errors at phoneme level, in order to find the phonemes for which the best and the worst results can be achieved with this technique.

6. References

[1] 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.

[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, Y. Yunusova, D. Thomas, J. R. Green, L. Zinman, "Compensatory articulation in amyotrophic lateral sclerosis: Tongue and jaw interactions", *The Journal of the Acoustical Society of America,* vol. 133, no. 5, pp 133, 2013.

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

[5] 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.

[6] A. Toutios, S. Ouni, Y. Laprie, "Estimating the control parameters of an articulatory model from electromagnetic articulograph data" *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3245-3257, 2011.

[7] M. Shtern, M. B. Haworth, Y. Yunusova, M. Baljko, P. Faloustsos, "A game system for speech rehabilitation" in M. Kallman, K. Bakris [Eds], *Motion in Games, Proceedings of the 5th International MiG Conference*, Rennes, France, pp. 43-54, Springer-Verlag Berlin Heidelberg, 2012.

[8] W. Katz, T. Campbell, J. Wang, E. Farrar, J. C. Eubanks, A. Balasubramanian, B. Prabhakaran, R. Rennaker, "Opti-speech: a real-time, 3D visual feedback system for speech training" in *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association. September 14-18, Singapore*, pp. 1174-1178, 2014.

[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] C. Lanz, J. Denzler, H. M. Gross, "Facial movement dysfyctions: Conceptual design of a therapy-accompanying training system" in *Ambient Assisted Living – Advanced Technologies and Societal Change*, pp. 123-141, Springer Heidelberg, 2013.

[11] J. A. Russell, M. R. Ciucci, N. P. Connor, T. Schallert, "Targeted exercise therapy for voice and swallow in persons with Parkinson's disease", *Brain Research*, vol. 1341, pp. 3-11, 2010.

[12] 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.

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

[14] J.-CH. Lafon, "Le test phonétique et la mesure de l'audition." Eindhoven, Dunod, 1964.

[15] P. Combescure, "Vingt listes de dix phrases phonéiquement équilibrées", Revue d'acoustique, vol. 14, no. 56,1981.

[16] E. Bocca, A. Pellegrini, "Studio statistic sulla composizione fonetica della lingua italiana e sua applicazione pratica all'audiometria con la parola", *Archivio Italiano di Otologia, Rinologia e Laringologia*, vol. 56, no. 5, pp. 116-141, 1950.

[17] T. F. Cootes, G. J. Edwards, C. J. Taylor, "Active appearance models", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681-685, 2001.

[18] I. Matthews, S. Baker, "Active appearance models revisited", *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135-164.

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