DOI: 10.17469/O2108AISV000021

CHIARA BERTINI, PAOLA NICOLI, NICCOLÒ ALBERTINI, CHIARA CELATA

A 3D model of linguopalatal contact for virtual reality biofeedback

Modelling the spatiotemporal dynamics of linguopalatal contact is important in the context of speech pathologies for both their diagnosis and rehabilitation. This paper describes a three-dimensional model of linguopalatal contact issued from real phonetic multilevel data produced by an Italian speaker. The model allows the simulation in a virtual reality environment of the mechanisms underlying the production of lingual consonants and vowels. We describe the procedures that allowed the development of the model as well as the outcome, which was an animation app to be experienced within a Unity 3D graphics engine, desktop or in an immersive environment.

Keywords: linguopalatal contact, visual biofeedback, 3D technologies, electropalatography (EPG), Ultrasound Tongue Imaging (UTI), Italian.

1. *Introduction. Background and goals*[1]

In this paper we describe a 3D model of the linguopalatal contact issued from real multilevel data for the simulation in a virtual reality (VR) environment of the mechanisms underlying the production of lingual sounds. The outcome is an animation that can be experienced within a Unity 3D graphics engine, desktop or in an immersive environment.

The model was developed within a project on speech motor disorders aimed at developing rehabilitation techniques based on VR visual biofeedback (Barone, 2017). Modelling the spatiotemporal dynamics of linguopalatal contact is important in the context of speech pathologies for both diagnosis and rehabilitation. Several biomechanical models of tongue movements exist that are based on mathematical models of muscular actions and their interactions (e.g. Gérard, Perrier & Payan, 2006, Moschos, Nikolaidis, Pitas & Lyroudia, 2011; Lloyd, Stavness & Fels, 2012; Wrench, Balch, 2015). The model developed in this project takes an opposite kinematic perspective: the goal is to produce a patient-specific model starting from

[1] Author contributions as follows. Conceptualization: C. Celata and C. Bertini; design of the articulatory experiment: C. Celata and C. Bertini; acoustic-articulatory recordings, data pre-processing and annotation: C. Bertini; model realization (digital tongue, palate and jaw): P. Nicoli and C. Bertini; visualization in the virtual environment: P. Nicoli and N. Albertini; writing – original draft preparation: C. Celata (§1-2, §7), C. Bertini and P. Nicoli (§§ 3-5), C. Bertini and N. Albertini (§6); writing – review and editing: C. Celata and C. Bertini; supervision: C. Celata. The authors would like to thank Irene Ricci and Vincenzo Barone for their collaboration and support and two anonymous reviewers for their constructive comments on a previous version of the paper.

real articulatory data issued from specific experimental settings. As a matter of fact, our model exploits the information about the positioning of the tongue with respect to the palate, and this information is obtained from real data acquired by means of a digital ultrasound device for tongue imaging (UTI) and an electropalatograph (EPG) in synchronized combination (Chen, Celata & Ricci, 2017). Furthermore, we do not model the movement of the tongue in general but, more specifically, the contact between the active (tongue) and passive (palate) articulator in the production of lingual sounds.

The goals of the current study are therefore the following:

- to develop a language-specific model of the linguopalatal contact in the production of Italian lingual sounds;
- to start from real articulatory data issued from a multi-level phonetic platform;
- to provide the model with a 3D visualization for a VR experience of speech production mechanisms.

The long-term goal of the project is that of developing speech learning and rehabilitation paradigms based on the developed VR biofeedback system. Various studies have suggested the validity of treatment and rehabilitation protocols based on the visualization of the articulatory organs, in addition to the auditory or spectrographic feedback traditionally used (e.g. Berhanrdt, Bacsfalvi, Gick, Radanov & Williams, 2005; Bacsfalvi, 2007; Katz, McNeil & Gast, 2010; Katz, Campbell, Wang, Farrar, Coleman Eubanks, Balasubramanian, Prabhakaran & Rennaker, 2014; Sebkhi, Desai, Islam, Lu, Wilson & Ghovanloo, 2017). In particular, it has been shown that patients who have to repair for incorrect postures during speech or coordination dynamics in the production of language greatly benefit from the possibility of accessing the visualization of their own vocal organs in movement, alongside repeated listening to their own voice, which only returns indirect evidence of articulatory activity. Moreover, the use of simple technologies (with the assistance of the speech therapist and the clinician) makes the rehabilitation environment much more interactive, playful and therefore motivating for the patient herself/himself, who receives additional knowledge from her/his own rehabilitation effort.

2. *Articulatory data and data acquisition*

The dataset for the modelisation was composed of 11 disyllabic pseudo-words + 1 real Italian word, which had previously been uttered by 1 Tuscan Italian female speaker. The stimuli included alveolar stops, liquids and sibilants and velar stops in different vocalic contexts. Bilabial consonants were also included to allow a more direct visualization of the different articulatory gestures needed to produce vowels at different places. Also, the real word *aiuole* 'flowerbeds' was included for its particular phonetic form, showing a sequence of 5 different vowels/glides. The 12 stimuli were therefore the following: [ˈata] [ˈiti] [ˈutu] [ˈaka] [ˈiki] [ˈal:a] [ˈar:a] [ˈas:a] [ˈam:a] [ˈum:u] [ajˈwɔle]. The stimulus set was purposely kept small to be

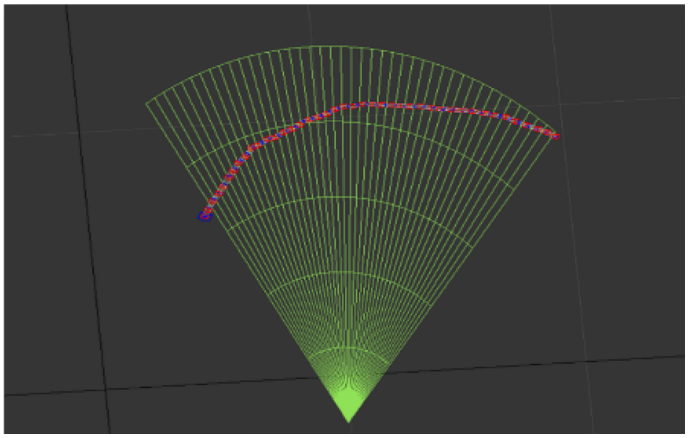
easily managed in the context of the model implementation and will hopefully be enlarged for future and more elaborated versions of the model.

The data source for the articualtory model were ultrasound tongue imaging (UTI) and electropalatographic (EPG) data acquired simultaneously through *SynchroLing* (Chen et al., 2017). *SynchroLing* is a multi-level phonetic platform that allows the real-time automatic synchronisation of three different channels, namely, the audio, the UTI and the EPG channel; the synchronization is controlled by the Articulate Assistant Advanced software (AAA). A previous version of the platform was developed for the multi-level phonetic and phonological study of Italian rhotics and is described in Celata, Vietti & Spreafico (2019) and Spreafico, Celata, Vietti, Bertini & Ricci (2015).

The data were collected in the sound-proof studio of the linguistics laboratory of Scuola Normale Superiore, Pisa.

Audio and tongue profiles were recorded via a Micro Speech Research US system associated to a Shure unidirectional microphone (44kH sampling frequency). Audio synchronization was automated by means of TTL pulse on completion of every frame. UTI data consisted in the discrete sampling over time of the mid-sagittal profile of the tongue during speech production. UTI data were collected at 100 Hz via the Micro Speech Research US system (Articulate Instruments Ltd), using a micro-convex probe (10mm; 5-8MHz; max FOV 150°). At each ultrasound frame, a maximum of 42 discrete points, corresponding to the lines of sight of the ultrasound probe, were used to reconstruct the tongue midsagittal upper contour. The reading of the tongue profile data was therefore n ≤42 positions supplied as coordinates (in mm) in the x-y plane at each ultrasound frame (Fig. 1).

Figure 1 - *Two-dimensional reconstruction of the 42 UTI fan radii and tongue spline*



EPG data consisted of binary information about presence/absence of contact (value 1 or 0) between the tongue and the 62 sensors arranged on the artificial palate worn by the speaker. EPG data were collected at 100 Hz via the WinEPG™ (SPI 1.0) system by Articulate Instruments Ltd.

The synchronized UTI and EPG data were then arranged in tabular form so as to have, for each row, the reference time point and the sequence of positions (for the UTI data) and contacts (for the EPG data).

The digital elements necessary for the creation of the interface were the tongue, the palate and the jaw. We describe the procedures adopted for the realization of each of them in the following sections.

3. *Digital tongue: rigging and skinning*

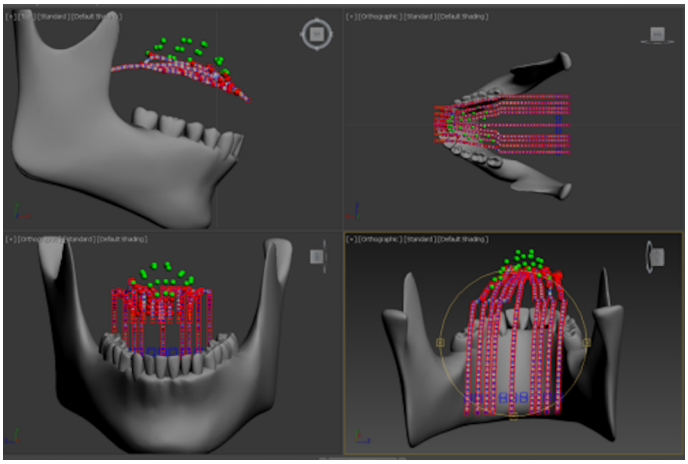
The tongue model was created by using chains of bones (i.e., chains of movement units for a 3D object during an animation), whose sum represents the skeleton of the virtual tongue.

The skeleton was made up of 9 chains, each of which consisted of 42 bones (Fig. 2). Each bone was positioned in the virtual space according to the spatial coordinates recorded by the articulatory instruments, and was directed towards the bone immediately in front of it. The central chain was the one responsible for receiving the positional information coming from the UTI data. When the acquired UTI data were < 42 (e.g. when the tongue was somewhat retracted and did not intersect the front rays), the missing data were filled through the Bézier interpolation algorithm. The central chain was therefore animated at each frame by the acquired UTI positional data; then, it transmitted this positional information to the side chains. The side chains corresponded each to a different sensor line of the artificial palate. When a sensor was contacted at a given time, the closest bone of the corresponding chain was detected and the position of the sensor was attributed to that bone. Since the same sensor was not always contacted by the same area of the tongue, a function was created that evaluate which part of the tongue surface was most likely to contact a given sensor based on proximity.

The skeleton was first tested on a very simple polygonal skeleton (mesh), and subsequently incorporated into that of the definitive virtual model of the tongue. The skinning process was the definition of which vertices were influenced by which bone and with what weight. The automatic association provided by 3ds Max 2019 was manually corrected and validated: a custom script in Maxscript was implemented in order to automate the process of data acquisition from files and animation creation.

The geometric structure of the model for the final animation had a low polygon density in order to improve the performance of the virtual application. After its definition, the visualisation was optimised through the application of smoothing, generation of the UVW map (a technique for associating 2D textures with 3D objects) and, finally, the texturing process.

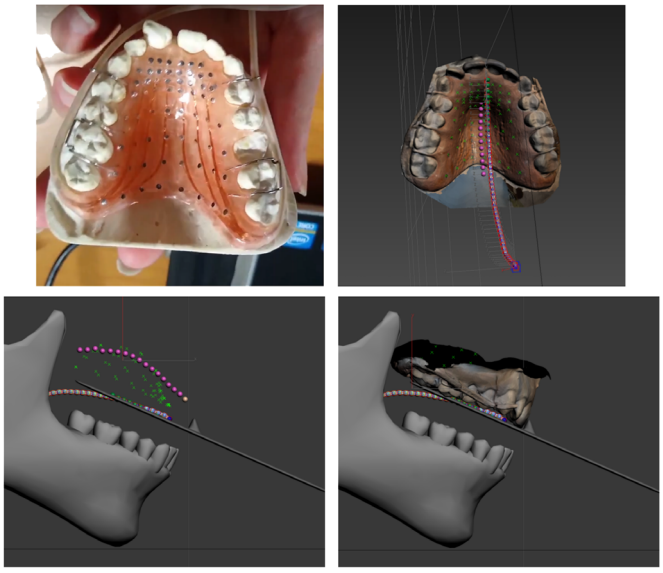
Figure 2 - *Transposition of the 42 UTI fan radii into 9 chains of 42 bones each (four different visualization angles). The 9 chains are in red. Each dot of the chain represents one bone. Green dots represent the electrodes on palate surface, here provided for reference*



4. *Digital palate*

The palate was derived from the acquisition of a real cast of the speaker's palate by 3D laser scanning, and subsequent production of a model for Unity 3D, a cross-platform game engine supporting a variety of desktop, mobile, console and virtual reality platforms (Fig. 3).

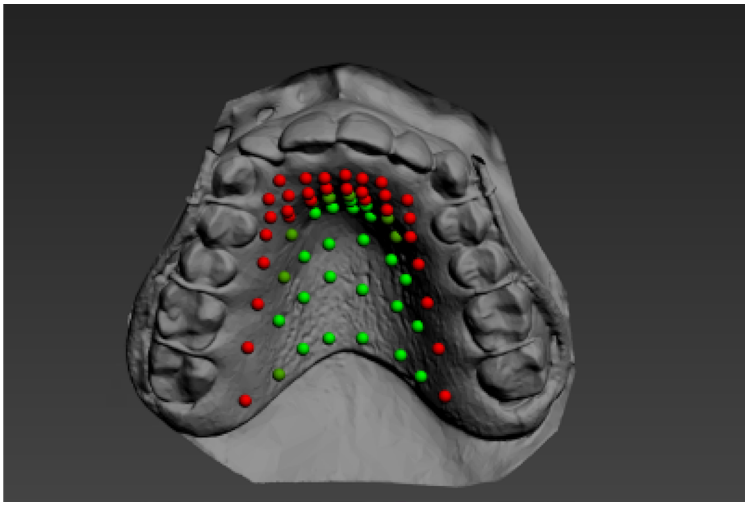
Figure 3 - *Top left: the EPG artificial palate and its chalk cast. Top right: Virtual palate modeling with positioning in the midsagittal plane. Bottom left: Ultrasound trajectory of the palate. Bottom right: superposition of the artificial palate*



More specifically, the 3D reconstruction of the palate was obtained by acquiring and processing midsagittal and transversal ultrasound images of the speaker's palate; subsequently, a 3D laser scanning of the plaster cast of the palate and of the artificial palate were made in order to realize a mesh of the as much detailed as possible, thus allowing to minimize the tolerances of reproduction and calibration data to the aim of the simulation. By superimposing the two spatial information, the 3D anatomy of the speaker's upper oral cavity was reconstructed (Fig. 4).

An echogenic object of known size and shape (biteplane) was used as reference for both the alignment of the virtual structures of the oral cavity and the analysis of the multilevel data obtained from different experimental sessions.

Figure 4 - *The artificial palate with sensors in the 3D model. Red dots: contacted sensors. Green dots: no tongue contact*



5. *Digital jaw*

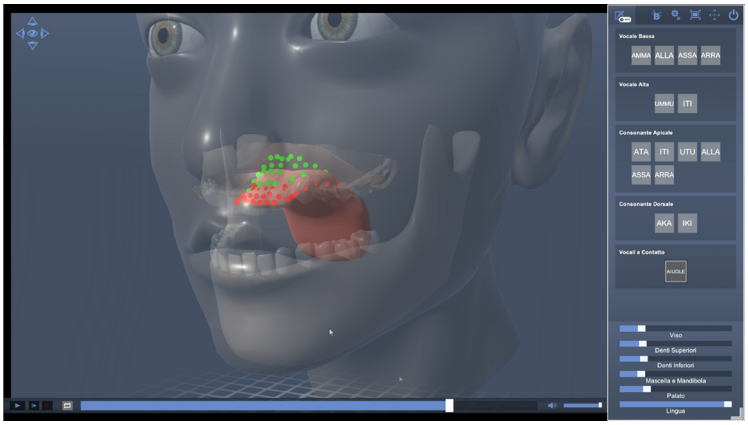
For the modelisation of the jaw, the mesh available from Artisynth (Lloyd et al., 2012) was used.

Taking into account jaw movements is crucial to correctly simulate the tongue movements within the oral cavity. *SynchroLing* is not equipped with a synchronized camera for video recording: therefore, for the current version of the 3D model, we opted for the following procedure.

Light sensors were positioned at specific points of the speaker's face on a separate session of speech recording. Facial recognition was then performed through Intel® RealSense™ D400 camera. This allowed an estimate of the angle of rotation of the jaw. The measures obtained were then inserted as correction values directly in the virtual editing software, by assigning average values for the duration of each articulatory gesture of the stimulus and applying smoothing functions for the transition between different degrees of opening and closing.

6. *Visualization in the virtual environment*

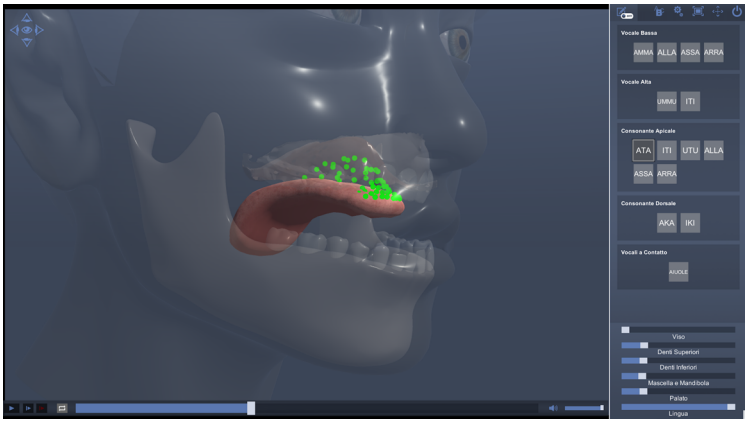
A 3D interactive animation app was therefore developed in Unity3D. The media file which accompanies this paper shows the basic functions of the app. The media file which accompanies this paper shows the basic functions of the app: please click on the image below to watch it.



The software used to virtually create the oral cavity and animate the tongue was 3ds Max 2019 (Harper, 2012). A script was implemented to automate the process of acquiring data from files and creating the animation. The script is executable within the 3ds Max 2019 program via .mcr file so that it can be called up directly from the user interface.

The app (Fig. 5 and 6) allows the user to visualize the oral cavity within the virtual head – the latter having a customizable degree of transparency in order to allow the visualization of the interior articulators. The user can also choose which anatomical structures to visualize and which ones to make completely transparent.

Figure 5 - *A screenshot showing lateral visualization of the 3D linguopalatal contact dynamics in the Unity3D-based app*

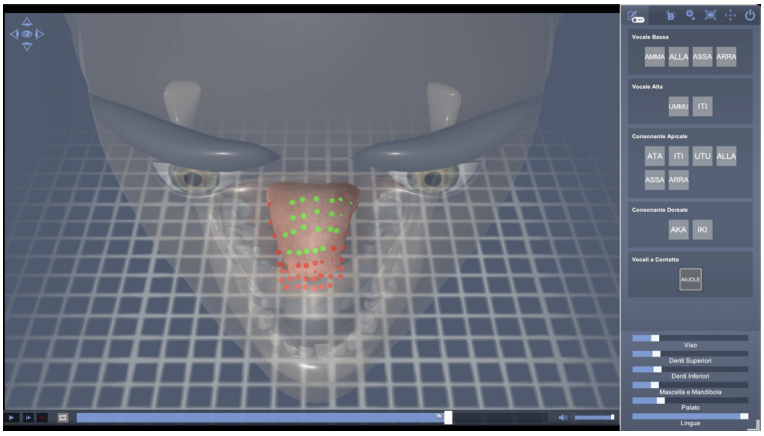


The interface allows 360° rotation of the head and, as a consequence, the articulators can be observed from any possible perspective. Specific keys also allow zooming in and out.

Electrodes on the palate are indicated as green dots. When linguopalatal contact occurs, the contacted electrodes become red, thus allowing a direct and extremely detailed appreciation of where contact occurs on the palate.

The user can also select the individual stimulus to visualize, and choose between different animation rates. The animation can be played frame-by-frame or cycle.

Figure 6 - *A screenshot showing visualization from above of the 3D linguopalatal contact dynamics in the Unity3D-based app*



7. *Future directions*

The model presented here is to be considered an initial test version and more elaborated versions need to be developed in order to fully achieve the goals of the research.

In particular, we believe that at least the following improvements will be needed to make the model fully exploitable.

One necessary improvement concerns contact visualization features; in particular, avoiding palatal 'fading-out' when the electrodes are contacted would make the visualization of the linguopalatal contact easier.

Second, the general visualization features will have to be improved, starting from the tongue shape and texturing, up to the humanoid face, to make the overall experience more enjoyable.

Third, it would be important to improve articulatory data pre-processing, particularly by developing automatic procedures that can reduce the amount of experimental noise necessarily brought about by manual data processing.

Fourth, we envisage testing the model with sensitive groups of users, primarily children, purposely tutored by speech clinicians. As a matter of fact, it would be important to have feedback about the experience with the 3D vocal tract animation in order to define further areas of improvement.

Fifth, once the model will have reached a more elaborated technical specification, the speech corpus will have to be enlarged and a subset of speech phenomena will have to be targeted which could be of major interest with respect to relevant speech disorders / speech rehabilitation needs.

Sixth, the model is currently intended to work with a reference speaker compared to which patients and clinicians may evaluate individual speech production characteristics. This is related to the fact that EPG palates are currently custom-made for each user and therefore relatively expensive. The elaboration of second generation, smart and cheap EPG palates (Surace, 2016; Mat Zin, Md Rasib, Suhaimi & Mariatti, 2021) is expected to enlarge the applicability of the model also to individual patients' production.

Acknowledgements

This work was supported by Fondazione Pisa [grant 2016 "Disturbi motori nel parlato e biofeedback visivo: Simulare i movimenti articolatori in 3D" to Vincenzo Barone].

Bibliography

Bacsfalvi, P.C.E. (2007). *Visual feedback technology with a focus on ultrasound: the effects of speech habilitation for adolescents with sensorineural hearing loss.* PhD Dissertation, University of British Columbia.

Barone, V. (2017-2020). *Disturbi motori nel parlato e biofeedback visivo: Simulare i movimenti articolatori in 3D*. Scuola Normale Superiore di Pisa, research project funded by Fondazione Pisa.

Bernhardt, B., Bacsfalvi, P.C.E., Gick, B., Radanov, B. & Williams, R. (2005). Exploring the use of electropalatography and ultrasound in speech habilitation. In *Journal of speech Language Pathology and Audiology*, 29, 169-182.

Celata, C., Vietti, A. & Spreafico, L. (2019). An articulatory account of rhotic variation in Tuscan Italian: Synchonized UTI and EPG data. In Gibson, M., Gil, J. (Eds.). *Romance Phonetics and Phonology.* Oxford (UK): Oxford University Press, 91-117.

Chen, C., Celata, C. & Ricci, I. (2017). An EPG + UTI study of syllable onset and coda coordination and coarticulation in Italian. In Bertini, C., Celata, C., Meluzzi, C., Lenoci, G. & Ricci, I. (Eds.). *Fattori sociali e biologici nella variazione fonetica – Social and biological factors in speech variation*. Milano: Officinaventuno, 151-172.

Gerard, J.-M., Perrier, P. & Payan, Y. (2006). 3D biomechanical tongue modeling to study speech production. In Harrington, J., Tabain, M. (Eds.). *Speech Production: Models, Phonetic Processes, and Techniques*. New York: Psychology Press, 85-102.

Harper, J. (2012). *Mastering Autodesk 3ds Max 2013*. Hoboken (New Jersey): John Wiley & Sons.

Katz, W., Campbell, T., Wang, J., Farrar, E., Coleman Eubanks, J., Balasubramanian, A., Prabhakaran, B. & Rennaker, R. (2014). Opti-Speech: A

real-time, 3D visual feedback system for speech training. In *Proceedings of Interspeech 2014*, Singapore, 1174-1178.

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

Lloyd, J.E., Stavness, I. & Fels, S. (2012). ArtiSynth: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation. In Payan, Y. (Ed.) (2012). *Soft Tissue Biomechanical Modeling for Computer Assisted Surgery. Studies in Mechanobiology, Tissue Engineering and Biomaterials*, vol 11. Berlin/Heidelberg: Springer, 355-394.

Mat Zin, S., Md Rasib, S.Z., Suhaimi, F.M. & Mariatti, M. (2021). The technology of tongue and hard palate contact detection: a review. In *BioMed Eng OnLine*, 20, #17.

Moschos, G., Nikolaidis, N., Pitas, I. & Lyroudia, K. (2011). A Virtual Anatomical 3D Head, Oral Cavity and Teeth Model for Dental and Medical Applications. In Czachórski, T., Kozielski, S. & Stańczyk, U. (Eds.) (2011). *Man-Machine Interactions 2. Advances in Intelligent and Soft Computing*, vol 103. Berlin/ Heidelberg: Springer, 197-206.

Sebkhi, N., Desai, D., Islam, M., Lu, J., Wilson, K. & Ghovanloo, M. (2017). Multimodal Speech Capture System for Speech Rehabilitation and Learning. *IEEE transactions on bio-medical engineering*, 64(11), 2639–2649.

Spreafico, L., Celata, C., Vietti, A., Bertini, C. & Ricci, I. (2015). An EPG+UTI study of Italian /r/. In The Scottish Consortium for ICPhS 2015 (Ed.) (2015). *Proceedings of the 18th International Congress of Phonetic Sciences*. The University of Glasgow, paper no. 775.

Surace, E. (2016). *Design e sviluppo di un elettropalatografo per la valutazione dell'articolazione della parola realizzato con materiali piezoresistivi flessibili*. Tesi di laurea magistrale, Università di Pisa & Scuola Superiore Sant'Anna.

Wrench, A.A., Balch, P. (2015). Towards a 3D Tongue model for parameterising ultrasound data. In The Scottish Consortium for ICPhS 2015 (Ed.) (2015). *Proceedings of the 18th International Congress of Phonetic Sciences*. The University of Glasgow.