# Dynamic Lip Animation from a Limited number of Control Points: Towards an Effective Audiovisual Spoken Communication

Slim Ouni, Guillaume Gris

# Dynamic Lip Animation from a Limited number of Control Points: Towards an Effective Audiovisual Spoken Communication

Slim Ouni

*LORIA - Université de Lorraine, LORIA, UMR7503, Villers-lès-Nancy, F-54600, France*

Guillaume Gris

*Ecole Polytechnique, Palaiseau, France*

**Abstract**

In audiovisual speech communication, the lower part of the face (mainly lips and jaw) actively participates during speech production. Modeling well lip motion and deformation in audiovisual speech synthesis is important to achieve realism and effective communication. This is essential for challenged population as hard-of-hearing people or new language learners. In this scope, we propose a technique that allows for animation of a human face with realistic lips using a limited number of control points. We have used an articulograph that provides high temporal and spatial precision, allowing tracking the positions of small electromagnetic sensors, even when occluded, which is often the case when tracking the lip movement. In our work, the control point data are first acquired, then fitted to a 3D face model of a human speaker, i.e., each control point is associated with a region of the face by minimizing the distance between the control points and the surface of the face model. Finally, we apply an interpolation scheme of the displacement field between the control points. This displacement field describes the deformation of a surface. In the case of the face, this method is well adapted to animating the region of the face that is highly correlated with speech, specifically the lips and the lower part of the face, even with a very limited number of control points.

*Keywords:* audiovisual speech, lips, animation, articulograph, speech articulation

## 1. Introduction

Speech communication is usually understood as the process of sending and receiving oral messages between people. We consider human-produced speech as a bimodal signal with two channels: acoustic and visual. The acoustic signal is the acoustic consequence of the deformation of the vocal tract under the effects of jaw, lips and tongue movements; and the visual signal is the consequence of this same deformation, which affects the shape of the face. Several research showed that acoustics and facial movements are correlated [1, 2]. Additionally, research in audiovisual speech intelligibility has shown the importance of the information provided by the face especially when audio is degraded [3, 4, 5]. Moreover, Le Goff et al. [4] have shown that when audio is degraded, the natural face can provide two thirds of the missing auditory intelligibility, their synthetic face without the inner mouth (without the tongue) provided half of the missing intelligibility and the lips restored a third of it. For audiovisual synthesis and animation, this suggests that one should pay careful attention to model the part of the face that participates actively during speech. In fact, a facial animation system [6] needs extremely good lip motion and deformation in order to achieve realism and effective communication. This is essential for challenged population as hard-of-hearing people or new language learners. As a matter of fact, hard-of-hearing people perceive speech signals in a distorted way and audiovisual speech becomes important to them, as it provides additional visual information which can tremendously enhance speech perception. Second language learners cannot easily perceive nor produce the phonemes that are absent in their native language. Visualizing accurately the articulation of the new sounds can help in learning their pronunciation in addition to a better perception of their acoustic consequences.

A large body of research has been conducted to model the lips for speech communication purpose to attempt to create a convincing talking head. In some works, parametric models, based on geometric shapes to represent the lips, have been used. The model-based approaches adjust the control parameters of a geometric model of the lips to fit the inner and outer contours of images of lips. Several works have used 3D polynomial surface or geometric shapes to model the lips [7, 8, 9]. For instance, King et al. [9] represented the lip model with a B-spline surface and parameters which define the deformation of the surface. The B-spline surface specifies the external and internal portions of the lips. The geometric representation replaces the orig-

inal lip geometry of the facial model. Recently, Kuratate and Riley [10] used a B-spline representation for each cross-section of their lip model. The cross-section information was extracted from magnetic resonance imaging (MRI) midsagittal data to help in defining the basic representation of the polynomial curves. They generated an approximate lip model from the contours using parameters suggested by the MRI data; then they fitted each cross-sectional curve to the lip surface of the original 3D scan data, and generate a lip model closely matching the observed lip data [10]. All these methods did not take finely into consideration the dynamics, which we consider crucial to build effective communication-oriented lip model in the context of speech communication. These methods are mainly parametric and are not based on human data, or based on very limited static human data.

Markereless techniques as those based on structural lights [11] can provide high-quality 3D results, however, they do not provide high temporal resolution which can be critical to replicate the finite articulatory speech gestures, and they are sensitive to lighting variation. We have shown in dedicated study that markerless techniques can provide a reasonable result during normal speech. However the quality is uneven during fast articulated speech, and the quality of the data is dependent on the temporal resolution of the markerless system [12]. For these reasons, it is appropriate to consider marker-based technique [13, 14, 15] to overcome the markerless-technique problems acquiring a better temporal resolution. Nevertheless, the acquired data in this case is sparse and lacking details.

Bhat et *al.* [15] presented a motion capture method that focuses on improving the quality of the contour features on the face. In particular, they manually traced the contour of the inner lip and they reconstructed a mesh that matches the contour. Although this method adds useful information on the inner shape of the mouth, the technique is to some extent manual and this method does not fully resolve the occluding of markers in the case of bilabial or labiodental speech sounds for instance.

There is another body of research that does not model the lips separately, but instead the whole face data are acquired. These methods animate the whole face at once [16, 17, 18, 19, 20]. The main strategy is to acquire the facial data using motion-capture technique that can be with markers [21, 22, 23, 13] or without markers [11, 24, 25]. In the first case, several markers can be painted or glued on the face to track the movement of the skin of the face using stereovision-based techniques [26]. In the second case, some methods extract the facial features from 2D images (using cameras) or 3D

images (based on the motion sensing).

Although these methods can give globally satisfactory results with regard to static realism of the face rendering, some of them fail when dealing with speech related animation. Very often, the quality of the lip animation during speech is not realistic and, more importantly not intelligible. Currently, a limiting factor is the complexity. In fact, whatever the motion-capture technique used, the lips are not well captured. Whether using marker-based or markerless-based methods, only the outer lip contour is captured. Thus, it is difficult to ensure that the lips are completely closed or not, which is important for the realization of bilabial sounds, for instance. When trying to resolve this problem by attaching markers to the inner lips, some of them can be occluded during protrusion or complete mouth closure. For instance, in [23, 20], a strereovision technique has been used to acquire the motion data of the face. A large number of markers has been painted on the lips. As can be seen in Figure 1, for some phonemes, several markers are occluded and it is very difficult to retrieve the missing markers (longer processing is needed, and heavy human intervention may be needed to assist the tracking system to interpolate the hidden markers).

For this reason, an alternative way to improve tracking accurately and dynamically the shape of the lips is to use a tracking system that can provide the appropriate information even though the markers are hidden. Electromagnetic Articulography (EMA) can be a robust technique to provide such information. In fact, EMA captures articulatory movements in three dimensions (3D) with a high temporal resolution ($250Hz$) and high spatial resolution ($0.3mm$ RMSE), by tracking tiny sensors attached to speech articulators such as the tongue, teeth, and lips or any part of the face. The positions and orientations of these sensors are calculated by measuring the electrical currents produced within multiple low-intensity electromagnetic fields [27]. This technique is known to present no risk to the health of the speaker, and was used in articulatory speech production researches for more a decade [28, 29, 27].

In our work, we are interested in modeling the lip movement accurately during speech. From this aspect, our purpose is to animate the lower part of the face accurately where the result is as intelligible as possible. Thus, we propose to use the EMA technique to capture dynamically the shape of the lips. The EMA sensors will be glued mainly on the lips. It should be noticed that some research has used EMA sensors to control the lips, by using mainly 4 sensors to animate a 3D face [30] and a 2D geometric model [31]. The mod-

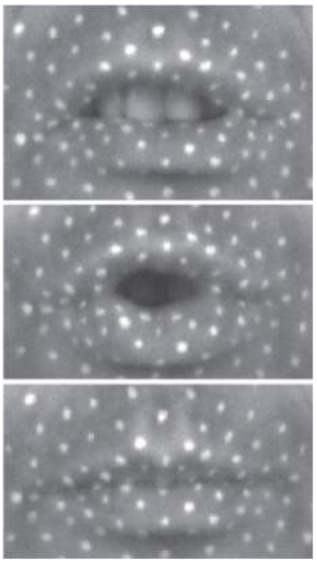

Figure 1: Positions of painted markers on the lips (from the top) for sibilant, protruded, and bilabial phonemes. Some markers get occluded in some context and it is difficult to track. [20]

eling of the lips in these cases is simplistic and does not reflect the complexity and the flexibility of the lip deformation.The articulograph AG501 can track at maximum 24 sensors simultaneously. This may be enough to control the lips, but it is not sufficient to animate reasonably well the face. It is more likely that this technique needs to be combined with other motion-capture technique to be able to animate the whole face (e.g., the EMA sensors for the lips, and other motion-capture technique for the rest of the face, to capture facial expressions). In our work, instead, we try to resolve this problem by defining the problem differently: given the limited number of sensors provided by the articulograph, is it possible to control the face accurately? In other words, we propose to glue the EMA sensors on the lips and on a very limited region of the face, and then we propose a method to animate the face reasonably well, mainly the lower part of the face that is highly correlated with speech articulation. The purpose is not to minimize the number of sensors to be used, but in the case where the number of sensors is limited, how to be able to accurately control finely the lips and the lower part of the face without the need of covering all the face with sensors. Even with

sparse sensors on the face, not necessarily dense, neither well distributed, the proposed method allows animating accurately the lower part of the face (lips, jaw, chin, cheeks).

In the following paragraphs, we present how to acquire the data: the 3D human model and 3D EMA data. Then, we present how to apply an interpolation scheme of the displacement field between the control points. This displacement describes the deformation of a surface or more generally a solid. In the case of the face, this method is well adapted to animate the part of the face that is highly correlated with speech, in particular the lips and the lower part of the face.

## 2. Data acquisition technique

The first steps of our work are : (1) acquiring a 3D model of the face of the speaker (static) and (2) acquiring the motion of the face (dynamic) using EMA data. The tracked sensors on the face are the control points that deform the shape of the face. These control points are first fitted to the face model. The surface of the 3D face model and the motion data are acquired separately. In particular, these data are expressed in two separate referential systems. It is possible to have a mapping from one referential system to another by means of successive transformations. The acquisition systems that have been used in our work provide models on the same scale, thus, we can have a mapping from one referential system to another by means of a rigid transformation.

### *2.1. Motion Data: EMA Data*

The motion data consist of the position of sensors of an articulograph (3D or 5D coordinates in space of electromagnetic sensors), reflective markers, painted markers or any control points on a given surface. As in this study we have used EMA data, for the sake of clarity, in the following we describe mainly the case where the motion data is the position of sensors of an articulograph, and the surface is that of the face. Nevertheless, the method can be applied to other solid or surface that have similar behavior.

The articulograph (presented in Figure 2) allows tracking the positions of small electromagnetic sensors attached to key points on the face mainly around the lips and the lower part of the face (see Fig. 3). This technique allows tracking sensors even though they are completely hidden, which is adequate to track the movement of the lips.

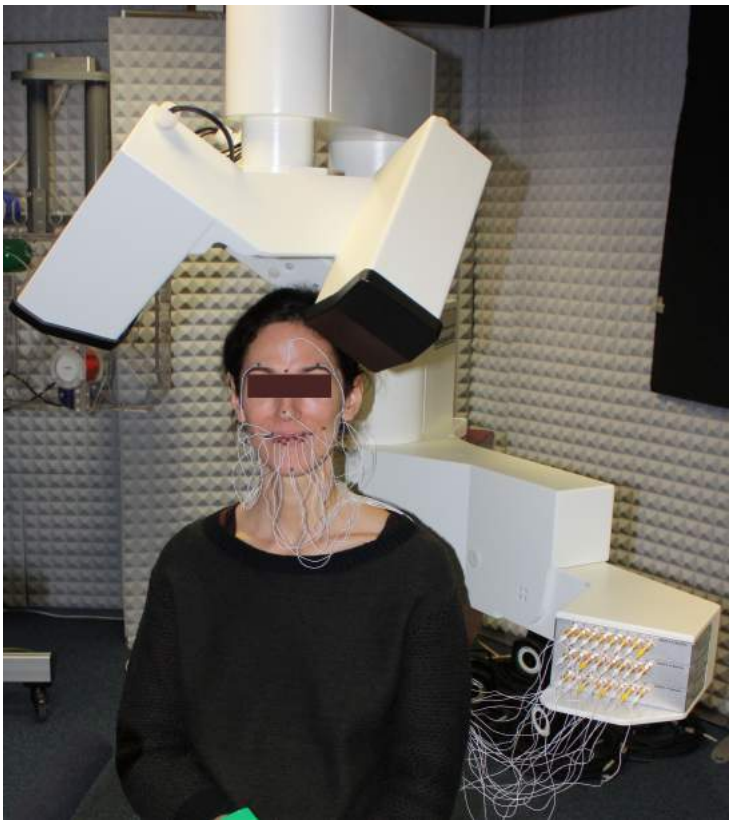

Figure 2: The articulograph AG501. A typical setting: The speaker is sitting under a block generating a low-intensity magnetic field. The sensors are glued on the face of the speaker. A unidirectional microphone is set in the front of the human speaker. The sensor movements and the speech signal are acquired synchronously.

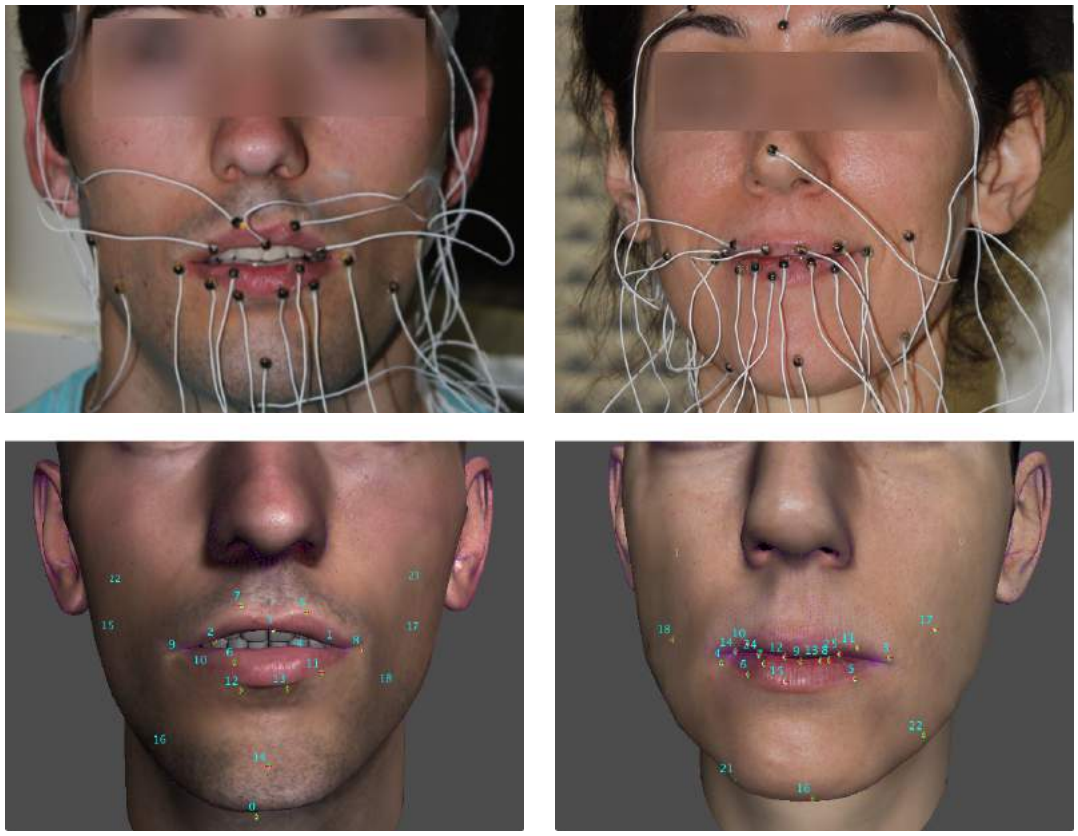

Figure 3: EMA Sensor configuration for two human speakers. The sensors are glued on the face: mainly on the lips and to a lesser extent on the face. The lower panel shows the projection of the sensors on the 3D model after fitting sensor positions to the model.

*2.2. 3D Face Model*

The 3D face model is a three-dimensional representation of the recorded speaker. Thus, we acquire a relatively high-resolution static model of the face using one of the several existing techniques to model the 3D face of the speaker. In our case, we chose to use a low-cost acquisition kinect-like device [32]. This technique is based on a markerless system, using mainly RGB and depth sensors, to acquire the 3D scan of the face from the observed 2D and 3D data. This solution allows acquiring easily and quickly an acceptable 3d model. First, we adapt a generic 3D model of a human face, composed of 7500 vertexes, to the speaker's face anatomy [33, 25] that was acquired using the kinect-like device as a 3D scanner. This technique allows having a 3D model of a given speaker without the use of face markers, intrusive lighting, or complex scanning hardware. Based on our own experience, we noticed that using textured surface for the lips can enhance the perception of the lip movement. For this reason, we have applied a high resolution texture map to the 3D model using a UV texturing mapping technique implemented in Blender. Figure 4, shows the two 3D models used in our study with the high-resolution texture. We recall that in our work, we focus on the quality of the animation and the dynamics of speech movements. It is very likely possible to get a better static facial model (a better modeling of the eyes and the hair, for instance), but for our work, we consider the quality sufficient for now to illustrate the animation method proposed in this study, to control the lower part of the face (the mouth and the jaw).

## 3. Deforming the Face Using Control Points

The problem of deforming a surface using control points has been addressed in previous research, where the problem can be formulated as transferring the dynamics of a sparse mesh onto a dense mesh, or deforming a high-resolution mesh from motion capture data (that can be considered as a low-resolution mesh). For instance, Bickel et al. [13] presented a method where motion-capture data deforms the high-resolution geometry using a linear thin-shell deformation model. In [34], a method that uses Radial Basis Functions (RBFs) which provides a global deformation while optimizing smoothness of the face mesh has been used to produce real-time facial expressions. The RBFs of the face mesh were created by a set of control points located on the mesh and their displacement values that describe the movements of the controls. Berger et al. [14] proposed a densification method

Figure 4: 3D models of two speakers. High-resolution texture of the real speakers have been applied to the models.

where the dense mesh vertexes were projected onto the sparse mesh, which constitutes the reference points. Each point of the dense mesh is associated with a sparse facet. In the simplest case, this can be considered as an affine interpolation of the displacement of the vertexes of the associated facet. However, the distribution of the control points must be sufficiently dense to reflect accurately the surface topology. In addition, this method is sensitive to the choice and distribution of the control points.

In the following, we present a method that describes how to deform a high resolution mesh from a very limited number of control points that does not have to be well distributed or well covering the considered region. The animation algorithm allows deforming the whole face by exploiting the influence of the different control points on the surface composed of hundreds of vertexes. The effect of the deformation decreases when the points of the surface are far away from the control points. The topology of the face is captured using a geodesic distance to calculate the influence weight of each control point on the vertexes in its vicinity.

Our method is a general interpolation scheme of the displacement field between the control points in an isotrope and a uniform space region.

For the sake of clarity, we will assume that the control points and the surface are in the same referential, and they correspond to the same real

surface, i.e., the face. We will denote the surface points corresponding to the control points as *reference points*. We should note that this correspondence is an approximation. Our purpose is to provide a robust reconstruction of the displacement field.

### *3.1. Displacement field*

To animate the surface, we need to determine a displacement field $\boldsymbol{u}(X) : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ for each point $X$ of the surface at each time-stamp.

As illustrated in Figure 5, we determine the current position $x$ of each point $X$ of the surface of the face. In practice, the points $X_i$ correspond to points on the face where are attached the articulograph sensors while the other points $X$ are other points of the face for which we have no direct measurement.

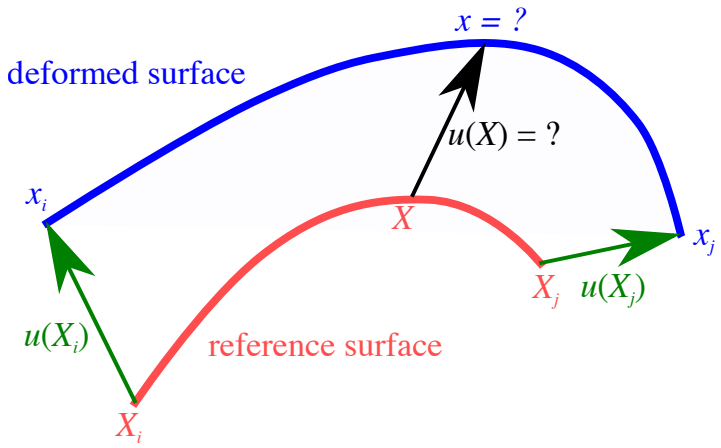


Figure 5: Interpolation problem: for each point $X$ of the reference surface, the displacement field $\boldsymbol{u}(X)$ at the point $X$ is determined in order to build the deformed surface, from the knowledge of the displacement of a set of points $X_i$

We have to determine a displacement field $\boldsymbol{u}$, which is at least $C^1$ – to be able to deform the face skin without the appearance of any angle (see Figure 5). We define $\boldsymbol{u}(X_i) = \boldsymbol{u_i}$ and $\nabla\boldsymbol{u}(X_i) \cdot \boldsymbol{V_i} = \nabla\boldsymbol{u}_i \cdot \boldsymbol{V_i} = \boldsymbol{v_i}$ in $p$ points $X_i$ where $\boldsymbol{V_i}$ and $\boldsymbol{v_i}$ are known directions – points where sensors are placed (i.e., control points).

We define $\boldsymbol{u}$ in the vicinity of each $X_i$ as follows:

$$\boldsymbol{u}(X) = \sum_i w_i(X)(\boldsymbol{u}_i + \nabla \boldsymbol{u}_i \cdot (X - X_i)) \tag{1}$$

where we consider $p$ weight functions $w_i : \mathbb{R}^3 \longrightarrow [0, 1]$ where $\sum_i w_i = \mathbb{1}$
The general form of $w_i$ is :

$$w_i(X) = \frac{f(X_i, X)}{\sum_j f(X_j, X)} \tag{2}$$

This general form is based on an influence function $f$ that expresses the influence between each couple of points. This formulation characterizes the propagation of the deformation of the surface.

It is possible to rewrite $w_i$ of the equation (2), in a similar way, where we define the functions $f_i$ at the control points $X_i$ as follows:

$$w_i(X) = \frac{f_i(X)}{\sum_j f_j(X)} \tag{3}$$

This candidate is quite valid locally, but it diverges as one moves away from the set of reference points. In fact, at a *large distance* of the set of reference points are considered as a single point where the influence is averaged uniformly.

we have:

$$\boldsymbol{u}(X) \sim \sum_i \frac{1}{p}(\boldsymbol{u}_i + \nabla \boldsymbol{u}_i \cdot (X - X_i)) \tag{4}$$

Which diverges because of the terms $\nabla \boldsymbol{u}_i \cdot (X - X_i)$. However, it can compensate for this difference while keeping the local properties by exploiting the $o(||X - X_i||)$. For instance, if we consider:

$$\boldsymbol{u}(X) = \sum_i \frac{f_i(X)}{\sum_j f_j(X)}(\boldsymbol{u}_i + \frac{\nabla \boldsymbol{u}_i \cdot (X - X_i)}{1 + \frac{||X - X_i||^\alpha}{R^\alpha}}) \tag{5}$$

Where $\alpha \geq 1$ characterizing the convergence at infinity of the second term. For $\alpha = 1$ the limit is a constant, for $\alpha > 1$ the limit is 0 and convergence is even faster when $\alpha$ is large. $R$ characterize the resistance of the surface to bend.

*3.2. Weight and Distance Choice Considerations*

It is possible to consider different types of functions $f_i$ for the definition of $w_i$ in the equation (3). For instance, if the surface/solid is homogeneous and isotropic, which is the case of the face, we can choose a function $f_i$ defined as follows:

$$f_i(X) = ||X - X_i||^{-2} \tag{6}$$

To sum up, equation (1) describes the displacement field at a given point $X$ relatively to the control points $X_i$.The values of $w_i$ describing the influence of a given control point $X_i$ on the deformation of another point. The $w_i$ uses a function that describes the rules of influence that are well adapted to the topology of the surface.

To apply this method, it is possible to use any metric space. However, in the case of the lips and more generally the lower part of the face, it is more appropriate to use a distance that reflects the topology of the face, as a geodesic distance or any derived one, for example. The topology of the face is captured using a geodesic distance to calculate the influence weight of each control point. Thus, even though the control points are not well distributed on the mesh and very limited in number, this method allows deforming the face mesh correctly as the used distance reflects well the topology of the face.

## 4. Lower Face and Lip Animation

In the following, we present the application of our method to control the animation of the face by using control points (the EMA sensor positions). We have recorded two speakers: a male native speaker of French and a female native speaker of American English. We have applied the method to examine the animation results on each speaker model and to compare the variability between subjects.

*4.1. Data Acquisition*

In Figure 2, a typical setting for an EMA acquisition is presented. The head of the speaker has been emerged in a low-intensity magnetic field. A set of 24 sensors were glued on the face of the speaker. Three of these sensors were used to remove the head movement. In the setting of the acquisition of Figure 3, 13 sensors were glued on the lips of the two speakers. The layout of the sensors on the lips was chosen arbitrarily by the experimenter, but in such a way that it covered reasonably well the possible deformations of the lips.

However, the exact positions themselves are not important and may change from one acquisition to another. The algorithm is capable of deforming the mesh, as explained in section 3. We have also recorded concurrently the acoustic speech. During each acquisition session, we have recorded several sentences in French (for the French speaker) and in English (for the English speaker).

As mentioned in section 2.2, we have acquired the 3D model using a kinect-based technique. This allowed acquiring a good quality 3D model. We have added the eyes and teeth to the model. The facial textures of both speakers were created based on high-resolution photos from different views of the speakers. The obtained result can be improved by using further enhancements. Nevertheless, our current focus is the quality of the dynamics, and we will deal with improving the quality of the texture, eyes and teeth in our future work.

### *4.2. Face Animation*

The deformation of the mesh was computed using our animation method. First of all, each sensor is associated to a region of the face by minimizing the distance between the sensors and the surface of the face model. The result of this fitting process was to associate each control point (a vertex on the surface) to a given articulograph sensor. This process was applied once. Then, the animation method deformed the facial mesh from a limited number of control points. The method was applied for each frame. As the motion capture data was acquired at $250Hz$, the output animation result consisted of displaying all the frames and there is no need to interpolate between the frames. The recorded speech associated with the motion capture data was played as it was, and it was used to keep the animation synchronous with audio (to avoid any possible lag).

To be able to animate the jaw, we have attached its rotation to the movement of the chin. Green et al. [35] have shown that the movements of the chin and mandibular were highly correlated and this suggests that it is possible to extract the movement of the jaw from a control point on the chin.

## 5. Results

### *5.1. Visual output*

Figure 6 and 7 show an example of the mouth opening animation from front and side view. Although the number of control points is very limited,

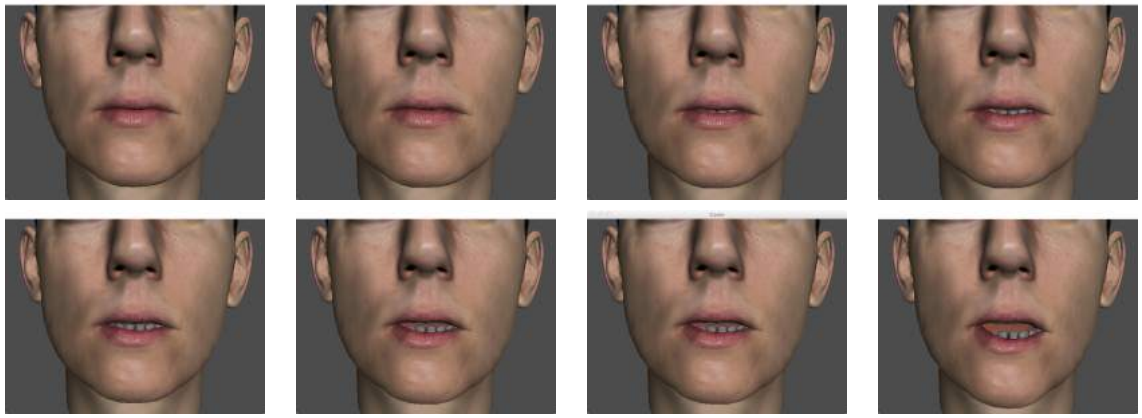

Figure 6: Animation of an open vowel : frontal view

the animation of the lips, the jaw, and all the lower part of the face is visually very precise and the transitions are smooth. It is even possible to notice the rotation of the lips, natural behavior observed during human speech. The animation algorithm allows deforming the whole face by exploiting the influence of the different control points on the surface. The effect of the deformation decreases when the points of the surface are far away from the control points.

Figure 8 shows how the vowel-consonant-vowel transition /ipu/ was reproduced realistically. It is even possible to notice the compression of the lips during the bilabial sound /p/. The quality of the texture helps to get a better perception of this aspect. Most interesting in this example is that the animation method is able to reproduce the protrusion of the lips and the explosion during the stop consonant /p/. As it is difficult to judge the quality of dynamics from static images, we highly recommend watching the accompanying videos to assess the visual and perceptual quality of the animation.

One advantage of this method is its fast execution time. In fact, it is possible to animate the face in real time. In the current implementation of the method, the computation of how the control points animate the face, is done on the fly. For example, we have tested the implementation of this method on a MacBook Pro - Intel i7 2.7 GHz - NVIDIA GeForce GT650M. The data has been acquired using the articulograph AG501, which allows controlling the face with 24 control points. The mesh has 7500 vertexes and

Figure 7: Animation of an open vowel : Side view

we applied a high resolution texture (8192x8192). In this configuration, the frame rate was between 100 and 140 FPS. Naturally, the computation time may increase if the number of control points increases excessively.

*5.2. Evaluation and Comparison*

To evaluate the accuracy of the proposed deformation method we made two experimental evaluations: (1) Evaluate the accuracy of the animation result compared to ground truth; and (2) Compare the animation of critical articulatory realization obtained by our method to a robust commercial software used for animation.

*Accuracy.* In this experiment, we applied our method to deform the face model using all the sensors, but one. The removed sensor in the computation was used as ground truth. In Figure 9, the EMA sensor $b$ (red circle) was the sensor that was not considered in the deformation computation. The red cross was the vertex on the surface that was observed to compare its trajectory to that of the sensor b. The animation of this vertex was obtained by applying the deformation method using all the EMA sensors but the sensor b. The results showed that the RMSE was $1.11mm$ on the x-axis, $0.61mm$ on the y-axis and $1.20mm$ on the z-axis. In addition, we have measured the mean distance between the sensor and the associated vertex that was $1.20mm$ with the standard deviation of $0.46mm$. The result shows that the animation of this vertex is highly accurate, the precision is very good and corresponds to the real observed behavior of the sensor.

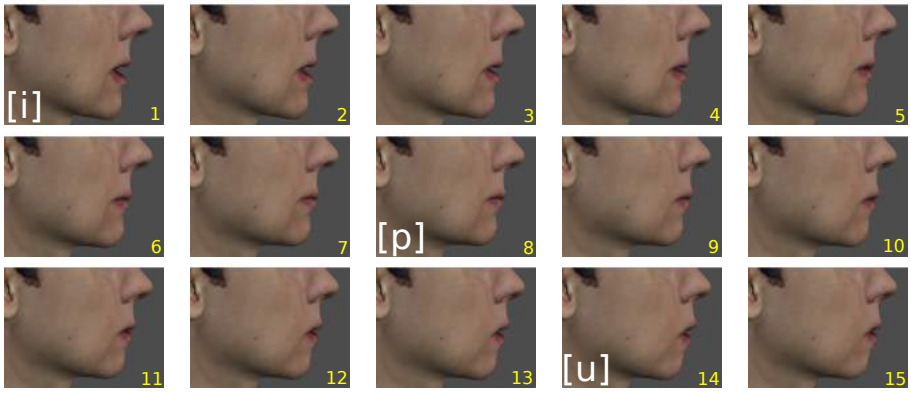


Figure 8: Animation of the transition [ipu]

*Comparaison: Articulatory Precision.* As our main goal is to use a talking head in the context of speech communication, where the articulation should be as accurate as possible, it is important to ensure the realization of some critical gestures. For instance, for the bilabial sounds as /b/ , /p/ and /m/, the complete closure of the lips is primordial to perceive these sounds correctly. Even if the lips are slightly open, the human perceiver will not see a bilabial sound (the McGurk effect). We have compared our animation method to the one used in the software FaceShift, which is a real time motion capture software using a kinect-like device [36]. Faceshift uses a markerless technique to acquire motion data. The output of the system is represented by a deformable facial model. Faceshift is considered robust and renowned in the field of expression and facial animation. We have conducted an experiment to compare the precision of the bilabial articulation provided by our technique and that of Faceshift. For this reason, we have used FaceShift simultaneously with the articulograph to acquire the motion capture data. Because of the presence of the sensor wires, we have cleaned the FaceShift data: during the postprocessing stage, we have corrected the tracking errors. As Faceshift allows getting virtual markers from their facial model, we have taken two makers on the upper and the lower lips at similar positions as EMA sensors. Figure 9 shows the position of the markers ($m_1$ and $m_2$). We have measured the opening of the lips, i.e., the distance between $m_1$ and $m_2$, on the whole set of bilabial sounds in the corpus. To extract the bilabial sounds, we did a

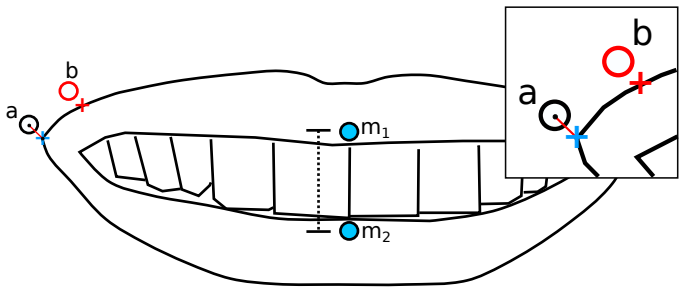


Figure 9: Markers positions for bilabial opening measurement. The opening is represented by the distance between $m_1$ and $m_2$. $a$ and $b$ represent two EMA sensors: $a$ is used in the computation of displacement fields in our method to animate the face, and $b$ is a sensor that represents the ground truth and is not used in the computation of the displacement fields to animate the face. The trajectories of the sensor $b$ and the vertex associated to $b$ (represented by a red cross) are used in the accuracy evaluation.

phonetic segmentation using classical speech recognition techniques [37]. The results show that on average the lip opening for bilabials using our method was $0.98mm$ with a standard deviation of 0.95 and the lip opening using FaceShift was $8.75mm$ with a standard deviation of 2.76. Figure 10 shows an example of such a result. The results of the experiment show clearly that our technique based on EMA outperforms Faceshift. In fact, Faceshift may miss easily lip closure during bilabials which may be problematic during speech articulation. This indicates as expected that EMA-based technique can provide a better quality of motion capture that is faithful to speech articulation of the human speaker, as the strength of the acquisition is to capture even the occluded gesture (i.e, the complete closure of lips).

## 6. Discussion

We have presented a technique that describes how to deform a high resolution mesh from a very limited number of control points that does not have to be well distributed or well covering the considered region. The animation algorithm allows deforming the whole lower part of the face by exploiting the influence of the different control points on the surface composed of hundreds of vertexes. The effect of the deformation decreases when the points of the surface are far away from the control points. The topology of the face is

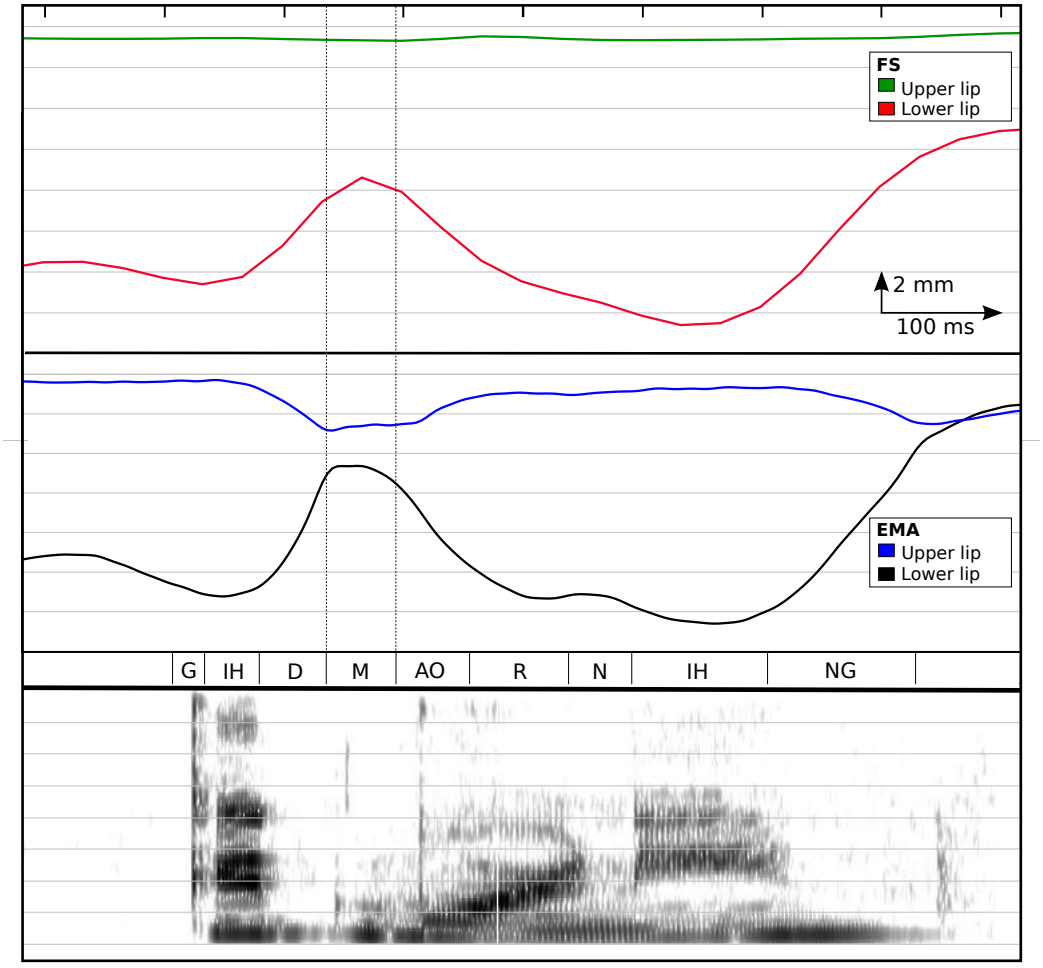


Figure 10: Bilabial realization comparing Faceshift and EMA-based technique. The lower panel shows the speech spectrum of the utterance 'good morning'. The two other panels show the trajectories of the lips using Faceshift and EMA outputs.

captured using a geodesic distance to calculate the influence weight of each control point on the vertexes in its vicinity. We have used EMA as motion-capture acquisition technique, as it is very useful for lip tracking to overcome the classical occlusion problem when using optical motion-capture technique. This animation method can be applied to other kinds of motion-capture data, as it mainly needs as input some control points (spatial positions of markers on the face) to animate a facial model.

Regarding the animation method, we apply an interpolation scheme of the displacement field between the control points. This displacement describes the deformation of the surface. The algorithm describes the behavior of the surface under the influence of the different control points. Although at first glance, the method seems based on classical interpolation techniques, the combination of the choice of the displacement field computation, the technique of computing the weights and the way how the propagation of displacement is defined, all together with the motion-capture technique based on EMA make the method well adapted to the problem of animating of the lips and the face. As the EMA data is highly dependent on the anatomy of each speaker face, it is difficult to reuse the EMA data directly for another speaker without considering advanced mapping transfer techniques.

The proposed animation method can be applied to any surface or solid that is continuous and constitute a metric space. This is the case of the face, and we showed that it was well adapted to animate the part of the face that is highly correlated with speech, in particular the lips and the lower part of the face. Moreover, it is possible to apply this method to other objects defined by a surface or solid as soon as they can be controlled by reference points. For instance, in the field of articulatory speech production, we are conducting studies to animate a human tongue, where the data are acquired with an articulograph, using exactly the same technique. It is even feasible to combine the animation of the face with that of the tongue. This is a part of our ongoing work.

The evaluation results showed the good quality of the animation and the accuracy of the method. The results presented in the video shows that the quality of the realism is quite good and the overall animation is realistic.

In our future work, we are planning to evaluate the intelligibility by conducting perceptual experiments [38, 5], that unfortunately take longer to prepare. In these experiments, the audio channel of audiovisual stimuli is degraded (speech in noise conditions, for instance) to see what perceivers can extract as information from the face.

The proposed animation method can be used to improve the quality of the animation in the field of entertainment, gaming or language learning. Our main purpose in proposing this work is to apply the animation technique to develop a hyper-realistic (dynamics and graphics) audiovisual speech synthesis (i.e., generate the animation and the acoustics of a talking head from any given text), which provides very intelligible speech that can be used by challenged population as hard-of-hearing people or learners of a new language. The essential step to build a text-to-audiovisual speech system is to have a better and accurate audiovisual data that reflect speech articulation faithfully as observed in real speech (see [20], for an example of the whole process to build an audiovisual speech synthesis system). The quality of the animation data would have a tremendous impact on the final result of the audiovisual speech synthesis. This is the case, for instance, when using a text-to-audiovisual speech synthesis based on a unit-selection technique or when using more recent techniques used in deep-learning speech synthesis [39, 40].

The proposed animation method is well adapted to animate the part of the face that is highly correlated with speech, in particular the lips and the lower part of the face. It is also possible to combine the EMA data with another optical motion-capture technique (using reflective markers for instance) to capture the full face : EMA for the lips, and optical motion-capture system for the other part of the face. The animation method can be applied to the merged data and it is possible to animate the full face. This allows getting complete data of the full face and allows addressing expressive speech synthesis where we deal not only with animation of the lips but also with the synthesis of facial expressions (eyes, eyebrows, cheeks, etc.)

## Acknowledgements

This work was supported in part by Equipex Ortolang.

## Reference

[1] J. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," in *ICPhS*, San Francisco, USA, 1999.

[2] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.

[3] W. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, p. 212, 1954.

[4] Le Goff, T. Guiard-marigny, M. Cohen, and C. Benoit, "Real-time analysis-synthesis and intelligibility of talking faces," in *In 2nd International conference on Speech Synthesis*, 1994, pp. 53–56.

[5] S. Ouni, M. M. Cohen, H. Ishak, and D. W. Massaro, "Visual contribution to speech perception: measuring the intelligibility of animated talking heads," *EURASIP J. Audio Speech Music Process.*, vol. 2007, no. 1, pp. 3–3, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1155/2007/47891

[6] Z. Deng and J. Noh, *Computer Facial Animation: A Survey.* London: Springer London, 2008, pp. 1–28.

[7] T. Guiard-Marigny, N. Tsingos, A. Adjoudani, C. Benoit, and M.-P. Gascuel, "3d models of the lips for realistic speech animation," in *Computer Animation'96. Proceedings.* IEEE, 1996, pp. 80–89.

[8] C. Pelachaud, E. Magno-Caldognetto, C. Zmarich, and P. Cosi, "Modelling an italian talking head," in *AVSP 2001-International Conference on Auditory-Visual Speech Processing*, 2001.

[9] S. King, R. Parent, and B. Olsafsky, "A muscle-based 3d parametric lip model for speech-synchronized facial animation," in *Deformable Avatars*, ser. IFIP The International Federation for Information Processing, N. Magnenat-Thalmann and D. Thalmann, Eds. Springer US, 2001, vol. 68, pp. 12–23.

[10] T. Kuratate and M. Riley, "Building speaker-specific lip models for talking heads from 3d face data," in *AVSP 2010 - International Conference on Auditory-Visual Speech Processing*, 2010.

[11] S. Zhang and P. Huang, "High-resolution, real-time 3d shape acquisition," in *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on*, June 2004, pp. 28–28.

[12] S. Ouni and 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. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/139/6/10.1121/1.4954497

[13] B. Bickel, M. Botsch, R. Angst, W. Matusik, M. Otaduy, H. Pfister, and M. Gross, "Multi-scale capture of facial geometry and motion," *ACM Trans. Graph.*, vol. 26, no. 3, Jul. 2007. [Online]. Available: http://doi.acm.org/10.1145/1276377.1276419

[14] M.-O. Berger, J. Ponroy, and B. Wrobel-Dautcourt, "Realistic face animation for audiovisual speech applications: A densification approach driven by sparse stereo meshes," in *Computer Vision/Computer Graphics CollaborationTechniques*, ser. Lecture Notes in Computer Science, A. Gagalowicz and W. Philips, Eds. Springer Berlin Heidelberg, 2009, vol. 5496, pp. 297–307.

[15] K. S. Bhat, R. Goldenthal, Y. Ye, R. Mallet, and M. Koperwas, "High fidelity facial animation capture and retargeting with contours," in *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '13. New York, NY, USA: ACM, 2013, pp. 7–14. [Online]. Available: http://doi.acm.org/10.1145/2485895.2485915

[16] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 4, pp. 1283–1302, 2005.

[17] K. Wampler, D. Sasaki, L. Zhang, and Z. Popović, "Dynamic, expressive speech animation from a single mesh," in *Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '07. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2007, pp. 53–62. [Online]. Available: http://dl.acm.org/citation.cfm?id=1272690.1272698

[18] Z. Deng and X. Ma, "Perceptually guided expressive facial animation," in *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '08. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2008, pp. 67–76. [Online]. Available: http://dl.acm.org/citation.cfm?id=1632592.1632603

[19] G. Bailly, O. Govokhina, F. Elisei, and G. Breton, "Lip-synching using speaker-specific articulation, shape and appearance models," *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[20] S. Ouni, V. Colotte, U. Musti, A. Toutios, B. Wrobel-Dautcourt, M.-O. Berger, and C. Lavecchia, "Acoustic-visual synthesis technique using bimodal unit-selection," *EURASIP Journal on Audio, Speech, and Music Processing*, 2013.

[21] L. Williams, "Performance-driven facial animation," in *Proceedings of the 17th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '90. New York, NY, USA: ACM, 1990, pp. 235–242. [Online]. Available: http://doi.acm.org/10.1145/97879.97906

[22] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," in *AVSP 2001-International Conference on Auditory-Visual Speech Processing*, 2001.

[23] B. Wrobel-Dautcourt, M.-O. Berger, B. Potard, Y. Laprie, and S. Ouni, "A low-cost stereovision based system for acquisition of visible articulatory data," in *5th Conference on Auditory-Visual Speech Processing - AVSP'2005*, Vancouver Island (BC), Canada, Jul. 2005. [Online]. Available: https://hal.inria.fr/inria-00000432

[24] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer, "High resolution passive facial performance capture," *ACM Trans. on Graphics (Proc. SIGGRAPH)*, vol. 29, no. 3, 2010.

[25] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," in *ACM SIGGRAPH 2011 Papers*, ser. SIGGRAPH '11. New York, NY, USA: ACM, 2011, pp. 77:1–77:10. [Online]. Available: http://doi.acm.org/10.1145/1964921.1964972

[26] D. B. Gennery, *Stereo vision for the acquisition and tracking of moving three-dimensional objects*. North-Holland, 1986.

[27] P. Hoole and A. Zierdt, "Five-dimensional articulography," in *Speech Motor Control: New developments in basic and applied research*, B. Maassen and P. van Lieshout, Eds. Oxford University Press, 2010, ch. 20, pp. 331–349.

[28] P. Hoole and S. Gfoerer, "Electromagnetic articulography as a tool in the study of lingual coarticulation," *The Journal of the Acoustical Society of America*, vol. 87, no. S1, pp. S123–S123, 1990. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/87/S1/10.1121/1.2027899

[29] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3078–3096, Dec. 1992.

[30] L. Wang, H. Chen, S. Li, and H. M. Meng, "Phoneme-level articulatory animation in pronunciation training," *Speech Communication*, vol. 54, no. 7, pp. 845–856, 2012.

[31] H. Li, M. Yang, and J. Tao, "Speaker-independent lips and tongue visualization of vowels," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.* IEEE, 2013, pp. 8106–8110.

[32] Y. Arieli, B. Freedman, M. Machline, and A. Shpunt, "Depth mapping using projected patterns," Apr. 3 2012, uS Patent 8,150,142.

[33] T. Weise, B. Leibe, and L. J. V. Gool, "Accurate and robust registration for in-hand modeling." in *CVPR*. IEEE Computer Society, 2008.

[34] T. Rhee, Y. Hwang, J. D. Kim, and C. Kim, "Real-time facial animation from live video tracking," in *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '11. New York, NY, USA: ACM, 2011, pp. 215–224. [Online]. Available: http://doi.acm.org/10.1145/2019406.2019435

[35] J. R. Green, E. M. Wilson, Y.-T. Wang, and 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. [Online]. Available: + http://dx.doi.org/10.1044/1092-4388(2007/066)

[36] Faceshift. (2015) http://www.faceshift.com. Faceshift is not available anymore since September 2015. Last accessed: 2015-08-05. [Online]. Available: http://www.faceshift.com

[37] X. Huang, F. Alleva, H. Hon, M. Hwang, and R. Rosenfeld, "The sphinx-ii speech recognition system: An overview," School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, Pittsburgh, PA, USA, Tech. Rep., 1992.

[38] D. W. Massaro, *Perceiving talking faces: From speech perception to a behavioral principle*. Mit Press, 1998, vol. 1.

[39] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7962–7966.

[40] B. Fan, L. Xie, S. Yang, L. Wang, and F. K. Soong, "A deep bidirectional lstm approach for video-realistic talking head," *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5287–5309, 2016.