Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT 7), pages 53–58
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Improved Facial Realism through an Enhanced Representation of Anatomical Behavior in Sign Language Avatars

Ronan Johnson
DePaul University
Chicago, USA
sjohn165@depaul.edu

Abstract

Facial movements and expressions are critical features of signed languages, yet are some of the most challenging to reproduce on signing avatars. Due to the relative lack of research efforts in this area, the facial capabilities of such avatars have yet to receive the approval of those in the Deaf community. This paper revisits the representations of the human face in signed avatars, specifically those based on parameterized muscle simulation such as FACS and the MPEG-4 file definition. An improved framework based on rotational pivots and pre-defined movements is capable of reproducing realistic, natural gestures and mouthings on sign language avatars. The new approach is more harmonious with the underlying construction of signed avatars, generates improved results, and allows for a more intuitive workflow for the artists and animators who interact with the system.

Keywords: signing avatars, sign language representation, computer animation

1. Introduction

The translation of spoken language to signed language is not only a translation of meaning, but also modality. It is therefore the place of the signing avatar to act as the intermediary between verbal and visual communication. In spoken language, most of the linguistic and syntactic information is conveyed by voice through the mouth while the hands provide secondary gesture and nuance. Signed languages are the opposite, with most of the lexical information occurring on the hands, allowing the face to supply grammatical and prosodic information. While research efforts have made progress on generating the primary hand and arm movements of signed languages, the processes on the face have not been examined so thoroughly, although the Deaf community has expressed their concerns on this matter (Verlinden, et al., 2001; Kipp, et al., 2011; Ebling, et al., 2015; Huenerfauth, et al., 2011).

Due to the complexity of the task, a perfect recreation of a real human is both unnecessary in practice and logistically untenable. Therefore, a major challenge in developing a representation of a human avatar is simplification. Any framework for a signed avatar must be complex enough to achieve the desired results while being simple enough to be workable by artists and procedural algorithms.

2. Previous Work

One of the primary descriptions of human facial movement is the Facial Action Coding System (FACS) (Ekman and Friesen, 1978). The basis for this system is a set of combined facial muscle movements first described by (Hjorstjo, 1970) and coded as a set of action units, each defining a specific motion on the face. These action units can be combined to classify all possible movements of the human face based on the underlying musculature. FACS has continued to be an on-going resource to industry professionals and academics studying the motion of the human face (Seymour, 2019).

FACS has been widely influential in the parameterization of human facial movements. One such example is the standardized facial representation in the MPEG-4 file description (Pandzic and Forshheimer, 2003). This attempted to define a minimal set of parameters necessary to recreate the facial actions observed by descriptive systems such as FACS. These parameters are conceptualized as a set of markers across key portions of the face. Each marker acts as a feature point for either an artist or procedural computer algorithm to control the shape and position the facial features. Figure 1 shows the control points defined for the mouth.

This implementation has been the foundation for previous developments in signed avatar technology such as the work of EMBR Virtual Human Animation System (Huenerfauth and Kacorri, 2015), the VSign sign synthesis web tool (Papadogiorgaki et.al., 2004), and the Paula avatar of DePaul University's American Sign Language Avatar Project (Wolfe, et al., 2018), the latter of which will be used by way of example. In Paula's case, the original underlying framework defines the landmarks as a set of joints that are skinned to the mesh, allowing the avatar's geometry to follow the movements of the joint.

Machine learning implementations for generating expressive facial animation, such as the Tacotron2 developed by Apple, yield promising results (Hussen Abdelaziz et.al., 2021). However, their major drawback is the sheer amount of data needed to adequately train an algorithm, especially a neural network. Tacotron2 used a dataset consisting of 10 hours of data captured from real human performance to train their convolutional neural network (CNN). Another research group based in the United Kingdom implemented a similar system using a temporal generative adversarial net (GAN) which used over 26 hours of video for its training data (Vougioiukas et.al., 2019). Even projects that have achieved success with far less training data such as the one developed by (Laine et.al., 2017) still require every desired facial movement be present in the training data. These restrictions make such models expensive to develop. They also require entirely separate data to properly model movements and gestures in other languages, limiting their generalizability.

Motion-tracking based frameworks such as the ARTUS project (Bailly et.al., 2006) present an alternative that is more extendible and can be used in broader real-time applications such television broadcasts. Their use of marker-less tracking also allows their system to function on a variety of video clips in order to generate clearer lip movements for Deaf and hearing-impaired viewers to follow, as opposed to traditional subtitles. This methodology has proven to be effective in generating realistic facial movements, but is reliant on the underlying video. Further research would be beneficial to evaluate its performance in generating original movements in the absence of human video.

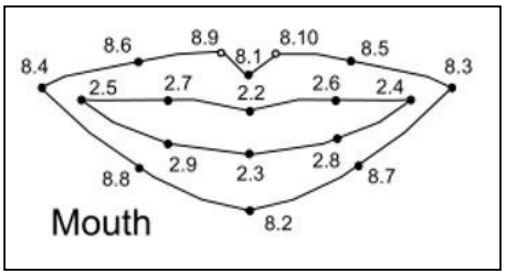

Figure 1: The mouth landmarks from MPEG-4 (Ekman and Friesen, 1978).

3. Revisiting Avatar Facial Representations

While they must appear similar in their final renderings, humans and avatars have little in common in terms of underlying structure. Humans are made up of layers of skin, fat, muscle, and bone. Mobility is achieved via the contraction of various muscles that pull on the underlying bones and ligaments. In contrast, signing avatars are defined primarily by geometric positioning and color information. Any movement is caused by some sequence of matrix operations on the avatar's positional data. These two highly contrasting modalities must nevertheless facilitate the same results: realistic and believable phonemes, visemes, and gestures.

The previous implementation of these facial processes on the Paula avatar utilized a FACS-based approach using the MPEG-4 facial marker definition. Although FACS is good at describing the process of observed actions on real human faces, an avatar framework instead needs to mathematically manipulate geometry to produce a final effect. The MPEG-4 representation attempts to define a complex series of muscle contractions with 28 points of positional control in two dimensions. Not only is there no strong structural connection between these modalities, but insistence on anatomical accuracy can distract from the ultimate goal of rendering expressive movement that garners the approval of the Deaf community. Ultimately, the underlying structure is only as useful as its ability to generate results. An improved model will be more congruent with the medium of avatar technology while allowing for greater artistic freedom and expediency.

Probably the biggest shortcoming of the MPEG-4 modality is its reliance on positional movement while ignoring rotation. For example, when the muscles around the sides of the mouth are activated, they pull the corners of the lips out towards the sides of the face. However, instead of simply shifting all the muscle and fat farther to the side, the lips are pulled around the curvature of the teeth in an arc. This kind of curved movement path is so fundamental to animation and recreating naturalistic motion, it is one of the twelve foundational principles of animation as defined by the original Disney animators (Johnston and Thomas, 1995).

This lack of rotation also creates an inability to reproduce several of Ekman's action units, in particular, the Lip Funneler (AU 22) and the Lip Suck (AU 28) as seen in Figure 2. These two actions are particularly challenging to recreate with positional movement because of the way the lips curl over the teeth and push away from the face towards the camera. These limitations have led to undesirable results on the avatar.

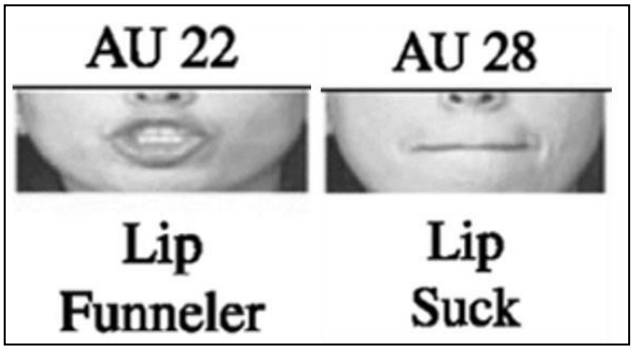

Figure 2: Two action units from FACS (Ekman and Friesen, 1978) that are difficult to recreate using only positional facial markers.

Figure 3: The best result achieved for AU 28 Lip Suck.

4. An Improved Framework

4.1 Geometric Marker Placement

In light of these considerations, the new representation is based not on positional translation, but rather the rotation of 44 individual mouth landmarks about a series of local pivot points. These landmarks lie along the surface of the geometry, centered on significant underlying geometry, and following the curvature of the lips. The original MPEG-4 landmarks are based on a general model of the movement of human facial anatomy, following the underlying muscles that pull on the lips. However, in a geometric representation, the landmarks should follow the underlying geometry that they will be transforming. This allows the model to work with the structural form of the avatar rather than retro-fitting a technique developed for an entirely separate modality. While this does technically increase the absolute number of control points from 28 to 44, through the use of rotational movement, the final structure allows maximum control to the artist with far fewer controls.

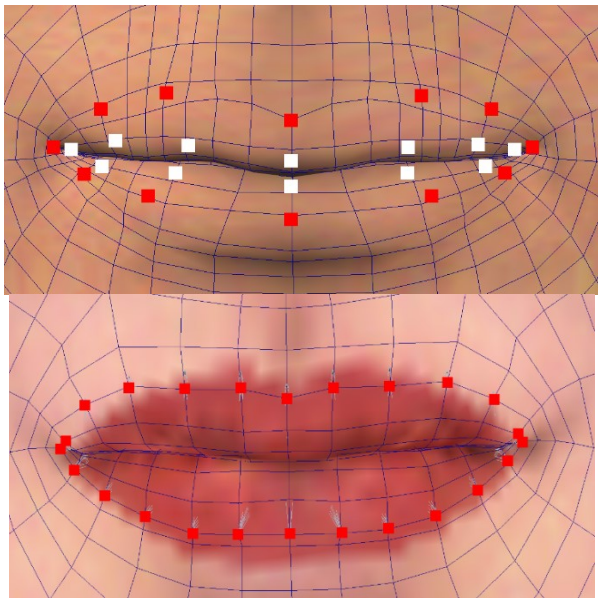

Figure 4: The original set of 28 control points (top); Red markers are the outer lip controls, white markers are the inner lip controls. Compare to the new set of control points and their geometric positioning (bottom).

It should be noted that there are a number of approaches to determining optimal marker placement. The work of (Le et.al., 2013) attempts to find a minimal layout that is effective for motion capture retargeting based on their effectiveness at recreating a given series of deformations in motion. Additionally, (Reverdy et.al., 2015) and (Will et.al., 2018) find compelling results on their motion tracker placement by using clustering methods to identify areas of the mesh with the strongest deformations while performing a series of expressions. This research does find marker placements that appear to perform more efficiently than the empirical placements such as the ones presented here. However, the primary goal of this new approach on the Paula avatar is to reduce the complexity of the work required of skilled artists, not necessarily the underlying computation.

Of particular concern is the number of control points surrounding the lips. The proposed optimization methods take the entire face into account when evaluating performance, which may mask underlying issues with localized performance in certain deformations. With the use of parameterized script controls as described in section 4.2, there can be greater flexibility in the absolute number of markers without placing undue strain on the artists' workflow. This yields the additional advantage of allowing the more complex control to be exposed to the artist if necessitated by a specific situation.

4.2 Major controls

Instead of the artist directly manipulating all 44 control points, the new system defines twelve major lip movements based on industry best practices (Osipa, 2010):

1. Lip spread
2. Jaw drop
3. Upper lip roll
4. Lower lip roll
5. Left lip corner
6. Right lip corner
7. Show upper teeth
8. Show lower teeth
9. Left upper snarl
10. Right upper snarl
11. Left lower snarl
12. Right lower snarl

The artist control structure for this system is presented as a set of sliders, each one dictating the intensity of each of these twelve movements. Here, 'intensity' refers to how extreme the movement appears on the face and is defined by a set of positional and rotational values for each relevant marker. These values are obtained by artist-generated extreme poses, intended to represent the most intense form of the movement an animator is likely to need. The slider values are normalized to lie between 0 and 100. This abstracts the complexities of generating the final shape to a single number, easily understood and manipulated by artists. When used in conjunction with one another, it is possible to recreate a wider range of action units than Paula's previous MPEG-4 framework with only a dozen single values for the artist to manage.

Each slider is connected to its relevant landmarks on the face with a script. These short pieces of code contain the needed positions and rotations of the landmarks to generate the most extreme form of the movement. They are also responsible for managing the intensity of the pose by interpolating between the neutral and the extreme. The slider value dictates the proportion by which this interpolation should occur. For example, when a user moves the jaw drop slider to open the mouth and sets the value to 50, the markers will move from their neutral values to 50% of their most extreme positions.

Further implementation details concerning the technologic connection between the landmarks and the sliders is presented in (McDonald, et al., 2022).

This interface gives artists complex control over the geometry with a minimal number of controls to manage. Furthermore, not all controls must be used to produce every individual mouth movement, reducing the complexity of the animators' work. Extended controls can be revealed to the user as needed should smaller corrections be needed. Other potential uses for this slider interface could include

connecting the slider values to motion tracking markers, allowing for the retargeting of motion capture data.

4.3 Marker Pivot Placement

A rotation is defined by movement about some axis and centered around a pivot point. These define the local deformations of the geometry by the facial landmarks to portray the desired shape. For the lip landmarks, pivot points are derived from the sweep of the arc that the final movement must follow. For example, in the case of lip spread, the control points need to follow a curved path to simulate the pull of the lips across the teeth in a real human. Figure 5 shows the derivation of such a path with a simple circle following the curvature of the teeth as a guide. The circumference of the circle should extend past the teeth just enough to account for the mass of the lips sitting on top. The center of this circle is the pivot point for each lip landmark during any movement that spreads the lips wide. The arrow shows the connection between the circle center and the position of the landmark.

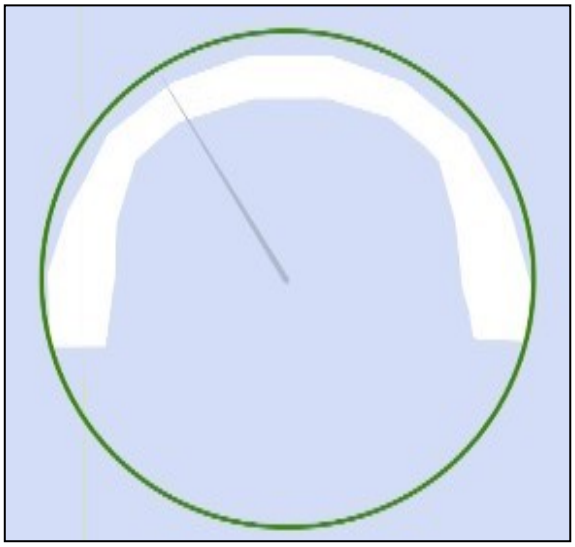

Figure 5: The guide circle for determining the appropriate pivot point of the lip spread control markers.

This same principle can be applied on an orthogonal plane to achieve the rotation necessary for AU 22 and AU 28. The difference in these movements is the location of the pivot. Instead of sweeping across the teeth, the lips in AU 28 need to curl under the teeth. Additionally, each landmark needs its own custom pivot point based on its exact location on the lips. This is because in order to avoid collisions with the teeth, the amount of rotation will be variable depending on the thickness of the lip at that location. This inward rotation must also account for the naturally curved orientation of the landmarks as the lips follow the curvature of the teeth, even when in a neutral pose. The same guiding curve of the previous example can determine the precise locations of these pivots as well. Instead of following the curvature of the teeth, this guide curve follows the thickness of the lips along the orientation of the geometry defining that section.

Human artists determined the exact position and orientation of these curves based on the orientation of the underlying geometry, specifically the edge loops that define the shape of the lips. While these positions have yet to be determined analytically, the initial results of the new approach were promising enough to continue with development. Future improvements may include optimization of these orientations, especially for application in the general case of any humanoid avatar.

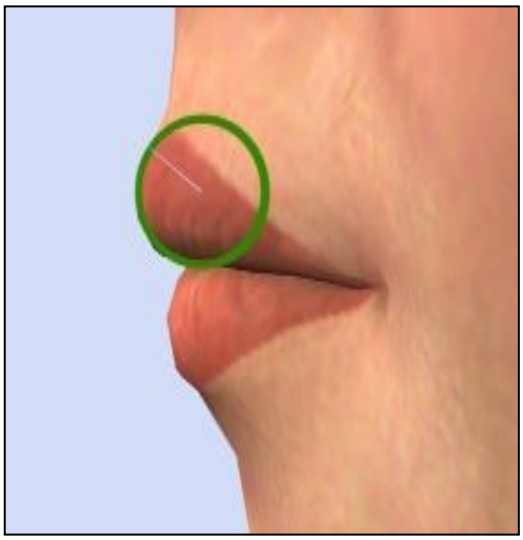

Figure 6: The guide circle for determining the appropriate pivots for the lip roll control markers.

By utilizing this rotational movement rather than relying exclusively on positional information, the markers naturally follow the curvatures of the face, yielding more realistic results. Additionally, rotational movement allows for the complex lip behaviors that were difficult to replicate with the previous MPEG-4 landmarks.

5. Results

The new system of empirically-placed facial markers driven by pre-set animation scripting is capable of reproducing all poses the original MPEG-4 framework could manage, while surpassing it in both control and flexibility. While each marker maintains only a small area of influence during a deformation, the combination of all markers working together gives more complex results with a far simpler interface for the artists. The results on the avatar are much improved in range of motion and expressivity.

One of the most compelling aspects of this design is its extendibility. The framework can accommodate any number of additions by simply defining another set of pivot points for each landmark. Figures 7 and 8 demonstrate the capabilities of the new parameterized framework. Artists are able to recreate subtle, intricate nuance in the shape of the mouth with relatively few controls. Further extensions may include generalized parameterization of the placement of the markers and their pivots for application to other avatars.

Figure 7: AU 22 (left) and AU 28 (right) created by an artist using the new framework. AU 28 can be generated by adjusting only two sliders.

Figure 8: Example expressions created on Paula using the new framework. The system is capable of generating a wide range of expressions and mouth postures.

6. Future Work

Due to the extendibility of the system, future work will include additional support for many signed languages including German Sign Language (DGS) and French Sign Language (LSF). Some expressive features of these languages require additional capabilities beyond those of both the new framework and the MPEG-4 description. For instance, there are several DGS mouth gestures that require interaction between the tongue and cheek. This complex deformation has yet to be recreated satisfactorily on a signing avatar.

Previous research on clustering-based facial marker placement may be of use in extending the expressivity of the Paula avatar. One area in need of improvement is the extent to which the cheeks and surrounding areas react to wide movements on the lips. While there are landmarks in areas such as the upper cheeks that are scripted to react to certain artist input, informal subjective assessment of the results indicates that additional naturalism might be possible without increasing the workload on the artists. These studies could inform the optimal locations of additional markers to allow more flexibility in these secondary movements.

Furthermore, a perceptual user study will be conducted to better assess the subjective quality of the final results compared to previous attempts on the Paula avatar.

7. Acknowledgements

Many thanks to Nicole Barnekow for her fantastic work creating facial expressions with the new framework.

8. Bibliographic References

Bailly, G., Attina, V., Baras, C., Bas, P., Baudry, S., Beautemps, D., ... & Nguyen, P. (2006). ARTUS: synthesis and audiovisual watermarking of the movements of a virtual agent interpreting subtitling using cued speech for deaf televiewers. *Modelling, measurement and control C, 67(2, supplement: handicap),* 177-187.

Ebling, S., Wolfe, R., Schnepp, J., Baowidan, S., McDonald, J., Moncrief, R., . . . Tissi, K. (2015). Synthesizing the finger alphabet of Swiss German Sign Language and evaluating the comprehensibility of the resulting animations. *Proceedings of SLTAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies,* 10-16.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior.*

Hjortsjö, C. H. (1969). *Man's face and mimic language.* Studentlitteratur.

Huenerfauth, M., Lu, P., & Rosenberg, A. (2011). Evaluating importance of facial expression in American Sign Language and pidgin signed English animations. *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility* (pp. 99-106).

Huenerfauth, M., & Kacorri, H. (2015). Augmenting EMBR virtual human animation system with MPEG-4 controls for producting ASL facial expressions. *International symposium on sign language translation and avatar technoloty* (Vol. 3, p. 94).

Hussen Abdelaziz, A., Kumar, A. P., Seivwright, C., Fanelli, G., Binder, J., Stylianou, Y., & Kajareker, S. (2021). Audiovisual Speech Synthesis using Tacotron2. *In Proceedings of the 2021 International Conference on Multimodal Interaction* (pp. 503-511).

Kipp, M., Heloir, A., & Matthes, S. (2011). Assessing the deaf user perspective on sign language avatars. *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility. ACM*, (pp. 107-114).

Laine, S., Karras, T., Aila, T., Herva, A., Saito, S., Yu, R., & Lehtinen, J. (2017). Production-level facial performance capture using deep convolutional neural networks. *In Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation* (pp. 1-10).

Le, B. H., Zhu, M., & Deng, Z. (2013). Marker optimization for facial motion acquisition and deformation. *IEEE transactions on visualization and computer graphics, 19(11),* 1859-1871.

McDonald, J., Wolfe, R., Johnson, R. (2022). A novel approach to managing lower face complexity in signing

avatars (submitted). *Seventh Sign Language Translation and Avatar Technology Workshop, Language resources and Evaluation Conference.* Marseilles: ELRA

Osipa, J. (2010). *Stop staring: facial modeling and animation done right*. John Wiley & Sons.

Pandzic, I. S., & Forchheimer, R. (Eds.). (2003). *MPEG4 facial animation: the standard, implementation and applications*. John Wiley & Sons.

Papadogiorgaki, M., Grammalidis, N., Sarris, N., & Strintzis, M. G. (2004, May). Synthesis of virtual reality animations from SWML using MPEG-4 body animation parameters. In *Workshop on the Representation and Processing of Sign Languages, 4th International Conference on Language Resources and Evaluation LREC 2004*.

Reverdy, C., Gibet, S., & Larboulette, C. (2015). Optimal marker set for motion capture of dynamical facial expressions. *In Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games* (pp. 31-36).

Seymour, M. (2019). FACS at 40: facial action coding system panel. In *ACM SIGGRAPH 2019 Panels* (pp. 1-2).

Thomas, F., Johnston, O., & Thomas, F. (1995). *The illusion of life: Disney animation* (p. 28). New York: Hyperion.

Verlinden, M., Tijsseling, C., & Frowein, H. (2001). A Signing Avatar on the WWW. *International Gesture Workshop,* (pp. 169-172).

Vougioukas, K., Petridis, S., & Pantic, M. (2019). End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs. *CVPR Workshops* (pp. 37-40).

Will, A. D., De Martino, J. M., & Bezerra, J. (2018). An Optimized Marker Layout for 3D Facial Motion Capture. *In STAG* (pp. 107-113).

Wolfe, R., Hanke, T., Langer, G., Jahn, E., worsek, S., Bleicken, J., …, Johnson, R. (2018). Exploring Localication for Muothings in Sign Language Avatars. *Language Resources and Evaluation Converence* (pp. 207-212). Miyazaki, Japan: European Language Resurces Association (ELRA).