Facial Expressions for Sign Language Synthesis using FACSHuman and AZee

Paritosh Sharma, Camille Challant, Michael Filhol

Facial Expressions for Sign Language Synthesis using FACSHuman and AZee

Paritosh Sharma, Camille Challant, Michael Filhol
Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,
91400, Orsay, France
{paritosh.sharma, camille.challant}@universite-paris-saclay.fr, michael.filhol@cnrs.fr

Abstract

This paper presents an approach to synthesising facial expressions on signing avatars. We implement those generated by a recently proposed set of rules formalised in the AZee framework for French Sign Language. Our methodology combines computer vision, linguistic insights, and morph target animation to address the challenges posed by the synthesis of nuanced facial expressions, which are pivotal for conveying emotions and grammatical cues in Sign Language. By implementing a set of universally applicable morphs and incorporating these advancements into our animation system, we aim to improve the realism and expressiveness of signing avatars. Our findings suggest an enhancement in the synthesis of non-manual signals, which extends to multiple avatars. This work opens new avenues for future research, including the exploration of more sophisticated facial modelling techniques and the potential integration of facial motion capture data to refine the animation of facial expressions further.

Keywords: Sign Language, Avatars, Facial expressions, AZee

1. Introduction

Signing avatars represent a crucial development in facilitating accessible communication for the Deaf and hard of hearing communities, enabling the visualization of Sign Language (SL) through computer-generated figures. The AZee model is instrumental in this, allowing for the synthesis of detailed multi-track animation timelines that specify the entirety of an utterance for rendering, thus enabling the creation of new SL content without the need for pre-existing animations.

Facial expressions, essential for conveying nuanced meanings in SL, pose significant challenges in the synthesis process for signing avatars. These non-manual features not only add depth and emotion to the communication but are also key in identifying meaning. The integration of facial expressions into signing avatars requires sophisticated modeling to accurately capture the wide range of emotions and grammatical cues that are communicated through subtle facial movements. This complexity makes the synthesis of facial expressions a vital area of focus to enhance the realism and effectiveness of signing avatars, ensuring they can serve as true representatives of SL communication.

This paper introduces an approach to formalizing the modeling and synthesis of facial expressions. We propose a methodology that combines computer vision, linguistic intervention, and morph target animation to improve the expressiveness and realism of signing avatars. This methodology integrates these advancements into our animation system to enhance the synthesis of non-manual signals based on a recent corpus (Challant and Filhol, 2024).

The paper is structured into sections discussing the background research on facial expressions in SL, the methodology for creating and implementing these expressions, the results of applying this methodology, and concludes with the key findings, implications, and potential future research directions.

2. Background Research

Although this has not always been recognised, we now know that the use of non-manual articulators in SLs is essential: it conveys linguistic information as much as hands activity (Pfau and Quer, 2010; Crasborn, 2006; Liddell, 2003). In this paper, we only focus on facial expressions: movements of the lips, eyebrows, cheeks and the tongue. The key role of facial expressions in SLs can be clearly seen when animating avatars: the presence of facial expressions on a virtual signer considerably helps Deaf people to better understand the generated discourse (Huenerfauth et al., 2011).

In linguistic studies, face articulators are most of the time studied separately: we can find studies on eyebrows (Kimmelman et al., 2020; De Vos et al., 2009) or on mouth gestures (Lewin and Schembri, 2011). We do not account for the particular case of mouthing, which consists in articulating lips following words from a spoken language. Indeed, the phenomenon is not observed on all signers, so we decided not to give it priority in our work.

We can also note that a facial articulator is often linked to a particular grammatical phenomenon such as questions (Schalber, 2006), conditional clauses (Reilly et al., 1990) or negation (Zeshan, 2004) and recognised as belonging to a defined linguistic level: phonological, lexical or syntactic.

Proceedings of the 11th Workshop on the Representation and Processing of Sign Languages, pages 354–360
25 May 2024.

Nevertheless, considering articulators together (rather than separately as in traditional approaches) seems relevant. The meaning conveyed by a set of articulators is not the same as that carried by an articulator studied on its own.

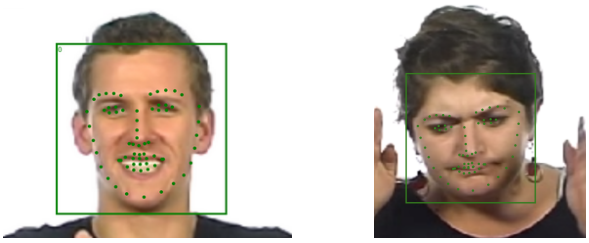
We are thus interested in AZee, a formal model which allows a wholistic approach of facial expressions (Filhol, 2021; Filhol et al., 2014). Indeed, the AZee approach is based on the notion of production rule, which associates a meaning to a set of observable forms. These can be movements of the hands, arms, chest, or any part of the face: there is no hierarchy between all these articulators.

A study has just been published on facial expressions in AZee (Challant and Filhol, 2024), based on a corpus called *40 brèves* (Filhol and Tannier, 2014). It consists of 40 news items in written French, each translated into French Sign Language by three deaf translators, for a total of one hour of SL. A new set of 22 AZee production rules producing facial expressions was found (for instance `big-threatening`, `closer-look` or `with-surprise`). This covers all expressions of the corpus, which to us constitutes a substantial subset to start with for LSF animation.

While the meanings are clearly identified for the rules concerning facial expressions, a problem is that the forms have only been approximately described or captured with still shots of signers producing them. It is now necessary to describe the forms of these facial expressions more precisely in order to animate them on virtual signers.

The methods for synthesizing facial expressions in SL animations encompass a variety of techniques. These methods include manual animation based on linguistic insights, automated techniques using motion capture data, and computer vision approaches for feature extraction. Sims (2000) offer unique approaches with varying degrees of success in capturing and animating nuanced facial expressions critical to SL communication. Kennaway et al. (2007) create blend shapes for the face which map to HamNoSys. However, they group various facial parts such as eyebrows, eyelids, and nose in same tier which complicates the modeling of facial expressions where these parts of the face are not moving in parallel and could pose restrictions in co-occurring facial expressions that share some of the parts of the face. Gibet et al. (2011) utilizes motion capture for more naturalistic expressions, facing challenges in data capture and representation granularity. A set of blend shapes were rigged on the Paula avatar (McDonald et al., 2022), which can also directly map to some AZee blendshapes. However, a bigger, more comprehensive mapping is still missing.

The FACS (Ekman and Friesen, 1978) breaks down facial expressions into individual components



(a) `big-threatening` (b) `closer-look`

Figure 1: AU detection for two productions rules

called action units (AUs), each of which corresponds to the principle muscles responsible for that movement. FACS is used in various fields, including psychology, cognitive science, and animation, to analyze and understand emotions, intentions, and reactions through facial expressions. Gilbert et al. (2021) developed a set of blendshapes which map directly to a subset of the FACS based on a template mesh. Thus, defining our facial expressions in terms of FACS AUs and mapping the FACSHuman blendshapes as AZee morphs would allow us to create a comprehensive set of facial expressions for any avatar which is based on this template mesh (Sharma and Filhol, 2023a).

3. Methodology

3.1. Modeling

To begin with, the first step was to model the 22 facial expressions AZee production rules using the software MakeHuman. For this, we used the FACSHuman plugin, which allows to model a human face thanks to different sliders, which are divided as follows:

- upper face (movements of the eyebrows, the lids and the cheeks);
- lower face (movements of the nose and the lips);
- head position;
- eye positions;
- lip parting and jaw opening;
- miscellaneous (e.g. cheek puff, tongue out, movements of the nostrils or the pupils).

Within each of these categories, there are different AUs for which the cursor can be placed between 0 (rest) and 100 (extreme position for this AU).

We worked with pictures extracted from the *40 brèves* corpus, which allowed us to find the new AZee facial expressions production rules. These

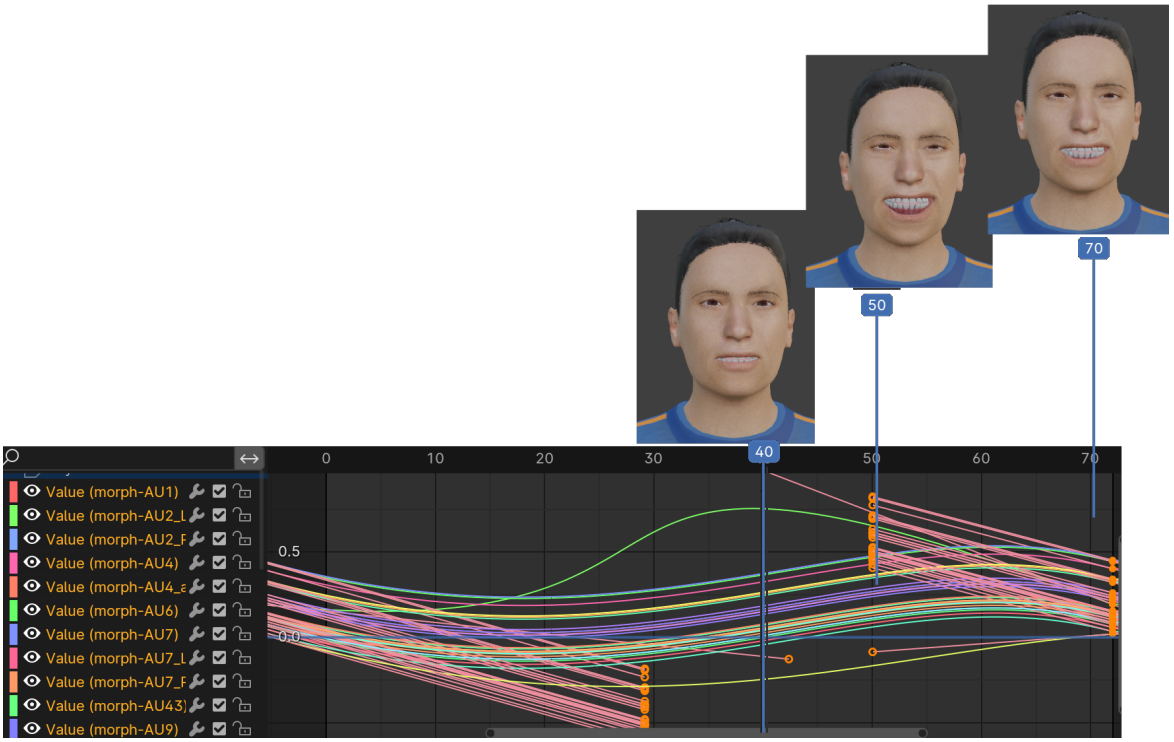


Figure 2: `big-threatening` based on a motion template (higher acceleration for jaw-drop)

pictures are easier to use when we try to model the face on an avatar than videos.

To start with, and to avoid modeling expressions from scratch manually, we tried to use automatic detection with FaceTorch (Figure 1), an AU detector based on work by Luo et al. (2022).

The detector models AU relationships and deep learns a unique graph to explicitly describe the relationship between each pair of AUs of the target face and thus detects composant Facial AUs from single RGB images. We realised that when AUs were detected, they were most of the time correct and gave us good clues to create the blendshapes for the face. But some activations were missing and the method is anyway constrained by the lack of AU intensity specification. Thus, linguist intervention was important at every iteration during the modeling process. For example, for the rule `big-threatening`, Luo et al. (2022) detects the following AUs (see figure 1): Brow Lowerer, Cheek Raiser, Lid Tightener, Upper Lip Raiser, Lip Corner Puller, Lips Part. When we tried to model the rule using FACSHuman, we used more AUs than what was detected initially : Inner Brow Raise, Outer Brow Raise (Left and Right), Eye Closure, Nose Wrinkle, Sharp Lip Puller (Left and Right), Dimpler, Lip Stretch (Left and Right), Lip Funneler (Bottom Lip and Both Lip), Lips Suck (Lower lip), Jaw Drop Bottom Lip Down.

Most of 22 AZee rules were therefore modelled manually with FACSHuman, without using FaceTorch (Luo et al., 2022).

3.2. Creating Shape keys

We model all FACSHuman AUs as Blender *shape keys*, using the *target* specification to define the bending of mesh at extreme positions. For example, the target file specifies vertex adjustments for facial movements, such as "`4 0.002 1`" to move vertex 4 by .002 units along the Y axis (labelled "`1`") for the extreme configuration.

During synthesis, these shape keys are modified as parts of *Facial Morph constraints* based on the AZee expression being synthesized. The avatar is then constrained based on these shape keys for the particular block.

3.3. Intermediate blocks

We extend out intermediate block generator (Sharma and Filhol, 2023b) algorithm to create interpolations for facial morphs as well. For this, we add additional *motion curves* (curves defining the displacement of vertices effected by the AU with respect to time) in the intermediate blocks based on the motion template.

This gives us a controllable motion curve profile for every AU for the facial morphs (Figure 2).

4. Evaluation

This section presents the evaluation of our methodology in synthesizing facial expressions for signing avatars using the AZee model. We evaluate the

Figure 3: Synthesis of `closer-look` (bottom) for male, female and neutral gender and their neutral expression for reference (top)

avatar's ability to perform a wide range of AUs and the synthesis of the modeled facial expressions. The accompanying videos of this research can be found at https://doi.org/10.5281/zenodo.10912305. Video "all_action_units" demonstrates the full range of AUs synthesized by our avatar. Video "all_expressions" shows all the synthesized expressions based on the French Sign Language corpus (Challant and Filhol, 2024). Figure 3 illustrates the synthesis of the expression `closer-look` across avatars of different genders, showcasing our method's adaptability to various avatar designs. Additionally, video "big_threatening_hot" demonstrates the expression `big-threatening(hot())` and `hot()` alone without non-manuals illustrating the added depth and meaning when non-manual signals are incorporated.

Our observations confirm that the avatars can perform a substantial range of recognizable expressions. The ability to apply these expressions across different avatars with no limitations underscores the universality of our methodology. However, we feel that the current model can be further improved to capture more nuanced expressions. We have indeed encountered a few limitations when we tried to model the different productions rules. All the limits are detailed in Table 1.

5. Conclusion

Integrating FACSHuman and AZee, our methodology overcomes challenges in facial expression synthesis for signing avatars, applicable on a series of avatars based on the same template, and descriptively rich expressions, enhancing both realism and communicative clarity. Our approach represents a significant advancement in the animation of facial expressions for sign languages, utilizing state-of-the-art methods in sign language representation (AZee) and animation (building face shapes from recognition). By combining these techniques, we ensure that facial animations not only accurately represent the intended expressions but also maintain fidelity to the intricacies of sign language communication, thereby enhancing the overall user experience and effectiveness of sign language avatars.

The natural next step now is to include these expressions on sample utterances (e.g. AZee sub-expressions from the attested data), and run them by LSF users for a more systematic evaluation. This approach will facilitate a deeper understanding of how well the expressions are understood and received within the LSF community and also give us insights on potential improvements(range of action units, acceleration information, etcetera).

Another potential improvement on our system could be a better facial model. Recent works such as Li et al. (2017) and Qin et al. (2023) use similar

Expression	Limitations
almost-reaching	Mouth modeling unconvincing.
continuously	"Pffff" air and cheek puff difficult, neutral eyebrows.
do-you-realise	Thick eyebrow issue.
it-is-a-shame	Mouth expression not quite real.
most-probably	Less visible teeth preferred, thick eyebrow issue.
much-almost-too-much	Frowning eyebrows and lack of eye wrinkles not convincing.
nothing-sticks-out	Tucked lips difficult to model.
something-sticks-out	Interpreted as confusion, mouth modeling limitation.
trouble-disturbance	Frowning eyebrows difficult, mouth "rising" hard to model, result not convincing.
uneasy-awkward	Tongue tip out with slightly open mouth hard to model, unconvincing.
with-chaos	Single cheek blow/puff and alternating eye blinks hard without animation.
with-no-precision	Upper lip over lower and mouth near nose unmodellable.
with-surprise	Cannot lower lower eyelid fully, thick eyebrow issue.
with-uncertainty	Appears sadder than uncertain, thick eyebrow issue.
with-worry	Lack of wrinkles around nose/forehead.

Table 1: Limitations for each facial expression rule.

Figure 4: Better facial expressions achieved using FLAME (Li et al., 2017)

philosophy of using a template mesh but generate better facial expressions since their models also account for other parameters such as stretching of skin and underlying muscles. This is demonstrated in figure 4 where the expression was generated manually using the first 100 principle components of the FLAME model. However, this generation can be automated using a flame-compatible recognition technique such as EMOCA Daněček et al. (2022). Another potential area of improvement could be the automatic creation of motion templates from retargeted facial motion capture data, thus adding much more detail to the interpolations.

The potential impact of having facial expressions with signing avatars is substantial. It enhances the capabilities of signing avatars making them much more expressive and realistic and opens new pathways for research and development in SL synthesis.

6. Acknowledgements

This work has been funded by the Bpifrance investment “Structuring Projects for Competitiveness” (PSPC), as part of the Serveur Gestuel project (IVès et 4Dviews Companies, LISN — University Paris-Saclay, and Gipsa-Lab — Grenoble Alpes University).

7. Bibliographical References

Anastasia Bauer and Masha Kyuseva. 2022. New insights into mouthings: Evidence from a corpus-based study of russian sign language. *Frontiers in Psychology*, 12.

Camille Challant and Michael Filhol. 2024. Extending AZee with Non-manual Gesture Rules for French Sign Language. In *Proceedings of the 14th Language Resources and Evaluation Conference (LREC)*, Torino, Italy.

Onno Crasborn. 2006. Nonmanual structures in sign language. In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, second edition edition, pages 668–672. Elsevier, Oxford.

Radek Daněček, Michael J Black, and Timo Bolkart. 2022. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322.

Connie De Vos, Els Van Der Kooij, and Onno Crasborn. 2009. Mixed signals: Combining linguistic and affective functions of eyebrows in questions in sign language of the netherlands. *Language and Speech*, 52(2–3):315–339.

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Michael Filhol. 2021. *Modélisation, traitement automatique et outillage logiciel des langues des signes*. Habilitation à diriger des recherches, Université Paris-Saclay.

Michael Filhol, Mohamed Hadjadj, and Annick Choisier. 2014. Non-Manual Features: The Right to Indifference. In *International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Michael Filhol and Xavier Tannier. 2014. Construction of a French–Lsf Corpus. In *Building and Using Comparable Corpora, Language Resource and Evaluation Conference (LREC)*, page 4, Reykjavik, Iceland.

Sylvie Gibet, Nicolas Courty, Kyle Duarte, and Thibaut Le Naour. 2011. The SignCom system for data-driven animation of interactive virtual signers: Methodology and Evaluation. *ACM Trans. Interact. Intell. Syst.*, 1(1).

Michaël Gilbert, Samuel Demarchi, and Isabel Urdapilleta. 2021. FACSHuman, a software program for creating experimental material by modeling 3D facial expressions. *Behavior Research Methods*, 53(5):2252–2272.

Thomas Hanke. HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts.

Matt Huenerfauth, Pengfei Lu, and Andrew Rosenberg. 2011. Evaluating importance of facial expression in American Sign Language and pidgin signed English animations. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, page 99–106, Dundee Scotland, UK. ACM.

Hernisa Kacorri. 2015. *TR-2015001: A Survey and Critique of Facial Expression Synthesis in Sign Language Animation*. CUNY Graduate Center.

J. R. Kennaway, J. R. W. Glauert, and I. Zwitserlood. 2007. Providing signed content on the internet by synthesized animation. *ACM Trans. Comput.-Hum. Interact.*, 14(3):15–es.

Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study. *PLOS ONE*, 15(6):e0233731.

Donna Lewin and Adam C. Schembri. 2011. Mouth gestures in British Sign Language: A case study of tongue protrusion in BSL narratives. *Sign Language & Linguistics*, 14(1):94–114.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17.

Scott K. Liddell. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge University Press.

Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. 2022. Learning multi-dimensional edge feature-based au relation graph for facial Action Unit recognition. *arXiv preprint arXiv:2205.01782*.

John McDonald, Ronan Johnson, and Rosalee Wolfe. 2022. A Novel Approach to Managing Lower Face Complexity in Signing Avatars. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 67–72, Marseille, France. European Language Resources Association.

Roland Pfau and Josep Quer. 2010. *Nonmanuals: their Grammatical and Prosodic Roles*, pages 381–402. Cambridge University Press.

Dafei Qin, Jun Saito, Noam Aigerman, Thibault Groueix, and Taku Komura. 2023. Neural Face Rigging for Animating and Retargeting Facial Meshes in the Wild. *arXiv preprint arXiv:2305.08296*.

Judy Snitzer Reilly, Marina Mcintire, and Ursula Bellugi. 1990. The acquisition of conditionals in American Sign Language: Grammaticized facial expressions. *Applied Psycholinguistics*, 11(4):369–392.

Katharina Schalber. 2006. What is the chin doing?: An analysis of interrogatives in Austrian Sign Language. *Sign Language & Linguistics*, 9(1-2):133–150.

Paritosh Sharma and Michael Filhol. 2023a. Extending Morphs in AZee Using Pose Space Deformations. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

Paritosh Sharma and Michael Filhol. 2023b. Intermediate block generation for multi-track sign language synthesis. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '23, New York, NY, USA. Association for Computing Machinery.

Ed Sims. 2000. Virtual communicator characters. *SIGGRAPH Comput. Graph.*, 34(2):44.

Ulrike Zeshan. 2004. Hand, head, and face: Negative constructions in Sign Languages. *Linguistic Typology*, 8(1):1–58.