Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT 7), pages 125–130
Language Resources and Evaluation Conference (LREC 2022), Marseille, 20-25 June 2022

Supporting Mouthing in Signed Languages: New innovations and a proposal for future corpus building

Rosalee Wolfe[1], John McDonald[2], Ronan Johnson[2], Ben Sturr[2], Syd Klinghoffer[2], Anthony Bonzani[2], Andrew Alexander[2], Nicole Barnekow[2]
[1]Institute for Language and Speech Processing, Athens Greece
[2]DePaul University, Chicago USA
Rosalee.Wolfe@athenarc.gr,
jmcdonald@cs.depaul.edu, {sjohn165, bsturr, sklingho, abonzan1, aalexa47, nbarneko}@depaul.edu

Abstract

A recurring concern, oft repeated, regarding the quality of signing avatars is the lack of proper facial movements, particularly in actions that involve mouthing. An analysis uncovered three challenges contributing to the problem. The first is a difficulty in devising an algorithmic strategy for generating mouthing due to the rich variety of mouthings in sign language. For example, part or all of a spoken word may be mouthed depending on the sign language, the syllabic structure of the mouthed word, as well as the register of address and discourse setting. The second challenge was technological. Previous efforts to create avatar mouthing have failed to model the timing present in mouthing or have failed to properly model the mouth's appearance. The third challenge is one of usability. Previous editing systems, when they existed, were time-consuming to use. This paper describes efforts to improve avatar mouthing by addressing these challenges, resulting in a new approach for mouthing animation. The paper concludes by proposing an experiment in corpus building using the new approach.

Keywords: sign language avatar, sign language display, computer animation, sign language linguistics, sign language translation, mouthing

1. Introduction

For nearly 25 years, researchers have been working toward the goal of an avatar that can produce grammatically correct signing that is easy to read. In the late 1990s Kuroda et al. (Kuroda, Sato, & Chihara, 1998) reported on an avatar-based system for Japanese Sign Language, and shortly thereafter, the ViSiCAST project began in Europe (Elliott, Glauert, Kennaway, & Marshall, 2000). A vital component of the project was evaluation by user communities, and Verlinden (2001) reported on avatar signing quality. The conclusion was "The main aspects that need further attention are the mouthing and to a lesser extent the mimicry."

In a subsequent evaluation ten years later, researchers organized two focus groups comprised of members from deaf communities in Germany to assess the potential use of signing avatars. They examined and evaluated videos of existing avatars. The researchers reported, "The absence of mouth patterns, especially mouthings (i.e., mouth patterns derived from the spoken language), seemed to be one of the most disturbing factors for the participants since this is an important element of DGS [German Sign Language]" (Kipp, Nguyen, Heloir, & Matthes, 2011). Even with an improved avatar, the same researchers noted, "In many cases, the lack of mouthing simply introduces irritation (Kipp, Heloir, & Nguyen, 2011) ."

In 2016 (Ebling & Glauert) and again in 2019 (Brumm, Johnson, Hanke, Grigat, & Wolfe) feedback indicated that mouthing on signing avatars is an aspect that still requires improvement. This paper analyses the causes for this deficit and presents an innovative strategy that incorporates important, but previously neglected considerations and proposes a methodology that capitalizes on these considerations to provide an additional tool for corpus building.

2. Related work

Of the myriad challenges to creating convincing mouthing on an avatar, we found three to be quite substantive. The first is the linguistic consideration of the diversity of mouthing within a sign language, and the second consideration entails the exacting requirements of mouth animation. The third consideration is a lack of usability in the current editing tools.

2.1 Linguistic considerations

"Mouthings," quoting Pfau, (2010) "are silent articulations of (a part of) a corresponding spoken word of the surrounding language." The usage of mouthing varies by language, ranging from occurring on virtually every sign, as in DGS, (Ebbinghaus & Heßmann, 1996) to virtually none at all as in Kata Kolok (De Vos & Zeshan, 2012). Within a single sign language, mouthing the whole word or part of a word also varies. In NGT (Sign Language of the Netherlands), signs can be accompanied by full or partial mouthing, but in the partial case, it is usually the stressed syllable (Bank, Crasborn, & Van Hout, 2011). However, the temporal reduction in a mouthed word may be even more extreme, up to the "mere onset consonant of the stressed syllable." Further, for a particular sign within a specified language, mouthing variations can occur based on register and the discourse setting.

Such rich diversity is intrinsic to natural language; however, from the standpoint of a software developer, this diversity renders it difficult to find a reliable pattern to automate. Previous efforts to create avatar mouthing have either relied on SAMPA or a speech generator to create complete mouthings. SAMPA is a set of computer-readable characters based on the International Phonetic Alphabet and is part of the SiGML standard (Jennings, Elliott, Kennaway, & Glauert, 2010). It was used in several projects, including ViSiCAST (Zwitserlood, 2005), eSIGN

(Hanke, Popescu, & Schmaling, 2003), DICTA-SIGN (Efthimiou, et al., 2012) and Trainslate (Ebling, 2013). Efforts utilizing speech generators included an extension to the EMBR (Kipp, Heloir, & Nguyen, 2011) and Paula avatars (Wolfe, et al., 2018).

2.2 Animation considerations

By stepping back from the challenge of mouthing in sign language to examine the closely related process of *lip sync* (lip synchronization) in animation (Williams, 2009), we (temporarily) remove the issue of choosing partial or full mouthing, and we focus on the basic steps:

1. Generate the phonemes corresponding to a spoken word.
2. Map each phoneme to a viseme, which is the visual appearance of the phoneme. There is a reasonable amount of consensus for the mappings, but the visemes are language dependent.
3. Retrieve the visemes from a library of facial poses.
4. Apply the viseme poses to the avatar as animation keys.

Previous avatars suffered from either a lack of realism in their visemes, a lack of realistic timing, or a combination of both. They relied on the MPEG4 H-Anim standard which did not allow for sufficient precision for naturally appearing visemes, but recent developments (Johnson, 2022), (McDonald, Wolfe, & Johnson, 2022) have created more responsive rigs that facilitate better realism.

In the end, visemes are simply poses. How effective they are in conveying mouthing depends on the timing and intensity of their appearance. A string of equally spaced visemes at equal intensities will not correspond to speech production – vowels and consonants have different durations, and these vary based on their location within a word. Therefore, the use of text-to-speech software is preferred to relying solely on SAMPA. Such software will produce timing information that an avatar can utilize for placing visemes as animation keys. Although the EMBR and Paula projects did use text-to-speech software, the EMBR project did not make use of any timing information for individual visemes, and the Paula project's visemes were not adequate for mouthing beyond American Sign Language and exhibited a curious 'lip snap', where the mouth occasionally moved too abruptly from one viseme to the next.

2.3 Usability considerations in editing

Several software packages offer tiers for annotation, (Neidle, Opoku, Dimitriadis, & Metaxas, 2018), (Max Planck Institute for Psycholinguistics, 2022), (Hanke, 2002) but there are only two systems that offer editing capabilities for mouthing, namely iLex/eSIGN (Hanke, 2014) and Paula. The interface in iLex supports three conventions for storing mouthings as text: orthography, IPA and SAMPA. The pronunciation data in iLex allows for the generation of visemes from orthography (Figure 1). iLex does not include functionality for adding timing or intensity to individual visemes.

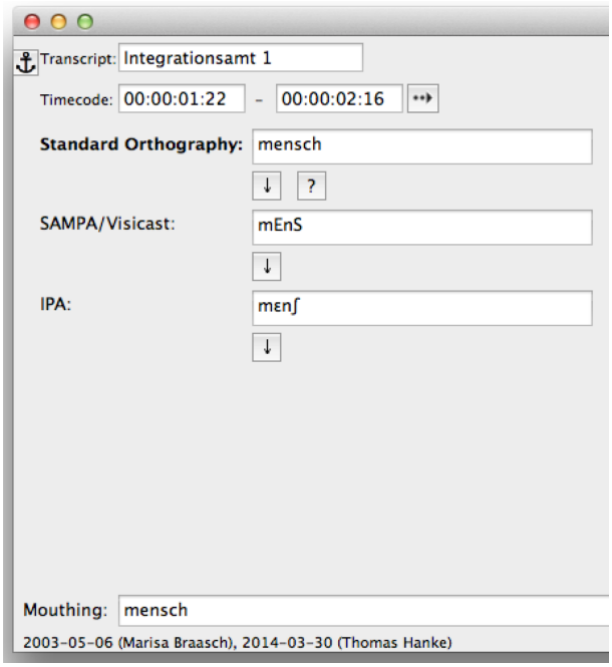


Figure 1: Mouthing dialog from iLex

The previous Paula mouthing interface did provide a rudimentary interface to edit mouthing. After the system generated the visemes via a speech generator, a user could edit the results if the animation was not convincing, or it contained an error such as a lip snap. However, the editing dialog was primitive, consisting of a data grid where each row contained a viseme, its start time relative to the beginning of the word, and its intensity (Figure 2). The editing process was cumbersome, as the user was forced to use the mouse to change the input focus to the cell needing modification before typing a new value and was required to rely on a sequence of numbers rather than a graphical interface for timing; a mode of interaction that is not artist friendly. This required continual switching from keyboard to mouse, when the preference is to use the mouse exclusively. Further, if a user wanted to increase the duration of a viseme, it was necessary to modify the starting times of all subsequent visemes. The editing experience proved sufficiently awkward that it was rarely used.

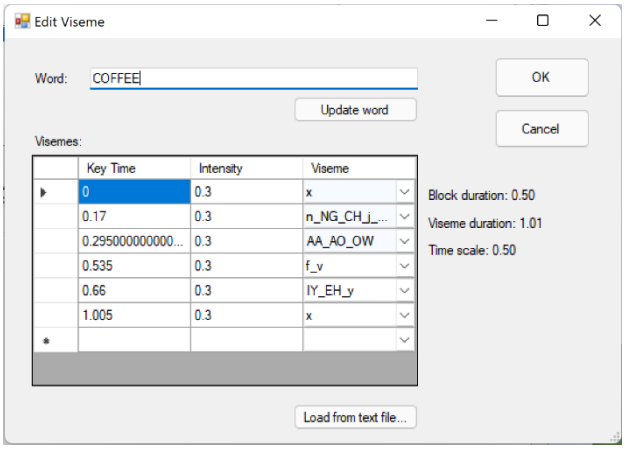


Figure 2: Mouthing dialog from circa 2019 Paula.

3. An improved approach

We introduce a new approach which consists of an improved set of automation heuristics for mouthing animation combined with a more artist friendly interface. The new approach produces better mouthing without lip snap errors through an integrated interface that combines better timing strategies informed by linguistics and offers an editing dialog that facilitates quicker, easier modifications. Additionally, it supports the fine-tuning of mouthing on a case-by-case basis depending on the context of the utterance being produced.

Animators can choose one of four automation hyperparameters: mouthing an entire word, mouthing the first syllable, mouthing the first viseme, or no mouthing, and then view the resulting animation. The new automation incorporates several strategies from traditional character animation, most notably that visemes need to be at least two frames long to avoid the dreaded lip snap. This is a duration of 0.083 seconds in conventional 24 frames-per-second movie technology. When using a frame rate of 30 frames-per-second, a duration of 0.083 seconds is the equivalent of a duration of 2.5 frames. In our experience, using a minimum of three frames works well when displaying on video playing at 30 fps.

Although this approach is working well, particularly for one- and two-syllable words, there will always be a need for possible revisions. The new mouthing dialog (Figure 3) offers multiple modification options, some at the word level, and some at the viseme level.

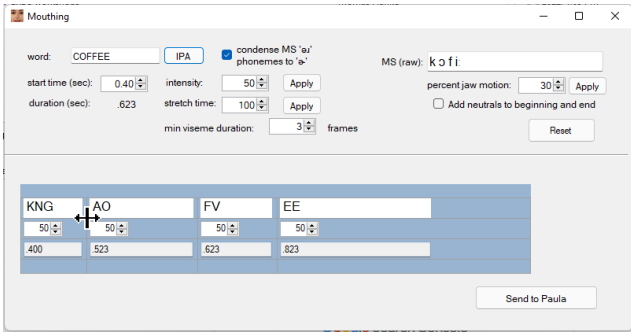


Figure 3: An improved editing interface. The animator is currently adjusting the duration of the initial viseme.

Animators can change the overall intensity, duration, and the start time relative to the manual channel[1]. They have two ways to change visemes – they can either edit the IPA characters in the upper right text box, or they can use the viseme editor. The viseme editor is contained in the blue rectangle in the lower half of the dialog. Each viseme has its own block whose width corresponds to the duration of the viseme. An animator can change the duration of the viseme by using the mouse to change the width of the block. All subsequent blocks are automatically realigned to visualize the new timing. Further, animators can add or delete visemes, change their individual intensities as well as modify the viseme selection through a context menu, as demonstrated in Figure 4. After making changes through the interface, an animator can tell Paula to display the modified animation through the "Send to Paula" button.

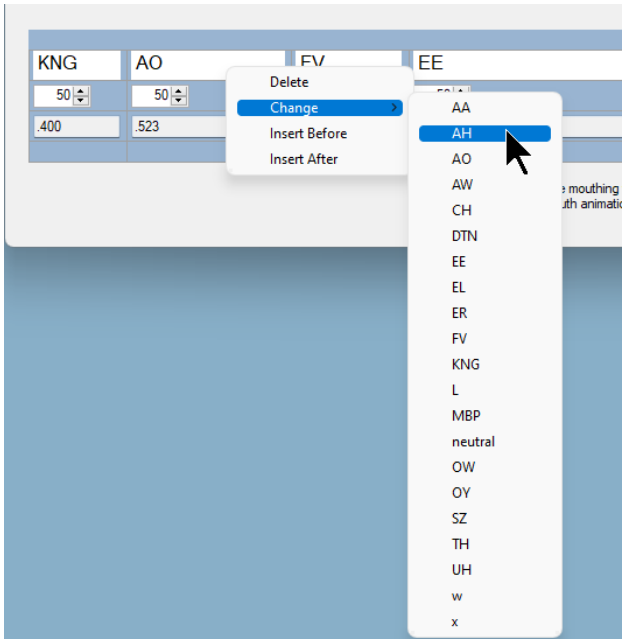


Figure 4: Context menu for adding, deleting, or changing a viseme.

Although there are still options for text input, our animators use mouse gestures exclusively, obviating the need to switch from mouse to keyboard. The result is quicker and more intuitive editing, particularly for visually oriented animation artists.

Although there are options to create complete or partial mouthing at the lexical level in our Sign Transcriber, the proclivity is towards recording complete mouthing for individual signs, because there are also mouthing options available when building complete sentences.

Our Sentence Generator builds sentences by retrieving lexical items from a database and applying modifications to them. In a mouthing track (tier) separate from the gloss, or lexical item track, an animator has the option to do nothing and use the mouthing associated with the basic lexical item, or to activate the same mouthing dialog as seen in Figure 3. Changes made to the mouthing in a sentence do not change the mouthings of a basic lexical item but are stored separately.

4. Results

The new interface supports multiple languages, including LSF, GSL, DGS, DSGS and ASL. For examples demonstrating the different styles of mouthing, including complete words, partial words and single viseme, please refer to Table 1. Although Figure 5 includes several sample images from the mouthings, the reader is encouraged to

[1] In our Sign Transcriber, a preparatory stroke precedes the sign, and the sign starts at 0.5 seconds. Thus, a start time in the mouthing editor of 0.4 seconds means that the mouth will begin moving 0.1 seconds before the start of the manual portion of the sign.

view the full animations via the link http://asl.cs.depaul.edu/video/wolfe2022/Mouthing.mp4 .

5. Conclusions and a proposal for future work

A future avenue to explore is the applicability of the methods described here to portray mouth gestures, which are commonly found across sign languages. In contrast to mouthings, mouth gestures do not arise from the surrounding ambient spoken language, and for this reason, there will be no need for a speech generator. However, effective application would require careful investigation of the postures of the lower face that are created when a signer produces mouth gestures. What descriptive/corpus work is needed to feed such an extension?

However, even for the case of mouthing, there are many refinements and questions remaining, because there is a tradeoff between automated and manual animation. Manual animation quality is superior but expensive, whereas automated animation is awkward but cheap.

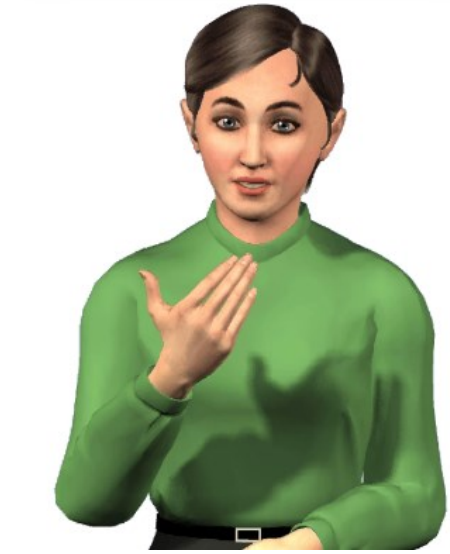
We propose an experiment in corpus acquisition for mouthing. This would involve using the automatic generation of full mouthing for all lexical items. Then, to create sentences, apply the hyperparameter (full, partial, single viseme, none) most appropriate for the sign language being produced. Then, in consultation with deaf communities, we customize those mouthings that are not acceptable via the new editing options and store those modifications. We anticipate that this will provide a promising resource for future study of mouthing synthesis.

English translation	Language mouthed	Mouthing type
EASIER project (name sign)	English	full
Hello, I'm ready to begin.	Swiss German	full
Please wait – response is pending	Greek	viseme
Thank you for using our service. Goodbye!	French	partial

Table 1: Examples of avatar mouthing found in http://asl.cs.depaul.edu/video/wolfe2022/Mouthing.mp4

ɪ in German 'ich' ('ɪç)

i in English 'easier' ('i:zi:ɚ)



i in French 'merci' (mɛʁsi)

Figure 5: Mouthing samples from various languages.

6. Acknowledgements

The authors are grateful to Rebecca Shaftman and Kathryn Willis Wolfe for sharing their linguistic knowledge, for their keen observational skills and for their articulate, constructive feedback.

This work is supported in part by the EASIER (Intelligent Automatic Sign Language Translation) Project. EASIER has received funding from the European Union's Horizon 2020 research and innovation programme, grant agreement n° 101016982.

7. Bibliographical References

Bank, R., Crasborn, O. A., & Van Hout, R. (2011). Variation in mouth actions with manual signs in Sign Language of the Netherlands (NGT). *Sign Language & Linguistics, 14*, 248–270.

Brumm, M., Johnson, R., Hanke, T., Grigat, R. R., & Wolfe, R. (2019). Use of avatar technology for automatic mouth gesture recognition. *SignNonmanuals Workshop*, *2.*

De Vos, C., & Zeshan, U. (2012). Introduction: Demographic, sociocultural, and linguistic variation across rural signing communities. *Sign languages in village communities: Anthropological and linguistic insights*, 2–23.

Ebbinghaus, H., & Heßmann, J. E. (1996). Signs and words: Accounting for spoken language elements in German Sign Language. *International review of sign linguistics, 1*, 23–56.

Ebling, S. (2013). Evaluating a swiss german sign language avatar among the deaf community.

Ebling, S., & Glauert, J. (2016). Building a Swiss German Sign Language avatar with JASigning and evaluating it among the Deaf community. *Universal Access in the Information Society, 15*, 577–587.

Efthimiou, E., Fotinea, S.-E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., . . . Lefebvre-Albaret, F. (2012). The dicta-sign wiki: Enabling web communication for the deaf. *International Conference on Computers for Handicapped Persons*, (pp. 205–212).

Elliott, R., Glauert, J. R., Kennaway, J. R., & Marshall, I. (2000). The development of language processing support for the ViSiCAST project. *Proceedings of the fourth international ACM conference on Assistive technologies*, (pp. 101–108).

Hanke, T. (2002). iLex-A tool for Sign Language Lexicography and Corpus Analysis. *LREC 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora.*

Hanke, T. (2014). Annotation of Mouth Activities with iLex. *6 th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel Language Resources and Evaluation Conference (LREC)* (pp. 67-70). Reykjavik, Iceland: ELRA.

Hanke, T., Popescu, H., & Schmaling, C. (2003). eSIGN–HPSG-assisted Sign Language Composition. *Gesture Workshop.*

Jennings, V., Elliott, R., Kennaway, R., & Glauert, J. (2010). Requirements for a signing avatar. *sign-lang@ LREC 2010*, (pp. 133–136).

Johnson, R. (2022). Improved facial realism through an enhanced representation of anatomical behavior for signing avatars (submitted). *SLTAT@ LREC 2022.*

Kipp, M., Heloir, A., & Nguyen, Q. (2011). Sign language avatars: Animation and comprehensibility. *International Workshop on Intelligent Virtual Agents*, (pp. 113–126).

Kipp, M., Nguyen, Q., 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*, (pp. 107–114).

Kuroda, T., Sato, K., & Chihara, K. (1998). S-TEL: An avatar based sign language telecommunication system. *International journal of virtual reality, 3*, 20–26.

Max Planck Institute for Psycholinguistics. (2022). ELAN (Version 6.3) [Computer software]. The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

McDonald, J., Wolfe, R., & Johnson, R. (2022). A novel approach to managing the complexity of the lower face in signing avatars (submitted). *SLTAT @ LREC 2022.*

Neidle, C., Opoku, A., Dimitriadis, G., & Metaxas, D. (2018). New shared & interconnected asl resources: Signstream® 3 software; dai 2 for web access to linguistically annotated video corpora; and a sign bank. *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, Miyazaki, Language Resources and Evaluation Conference 2018.*

Pfau, R., & Quer, J. (2010). Nonmanuals: Their prosodic and grammatical roles. *Sign languages*, 381–402.

Verlinden, M., Tijsseling, C., & Frowein, H. (2001). Sign language on the WWW. *Proceedings of 18th Int. Symposium on Human Factors in Telecommunication.*

Williams, R. (2009). *The animator's survival kit.* Farrar, Strauss and Giroux.

Wolfe, R., Hanke, T., Langer, G., Jahn, E., Worseck, S., Bleicken, J., . . . Johnson, S. (2018). Exploring Localization for Mouthings in Sign Language Avatars. *sign-lang@ LREC 2018*, (pp. 207–212).

Zwitserlood, I. (2005). Synthetic signing. *The World of Content Creation, Management, and Delivery*, 352–357.