Exploring Localization for Mouthings in Sign Language Avatars

Rosalee Wolfe[1], Thomas Hanke[2], Gabriele Langer[2], Elena Jahn[2], Satu Worseck[2], Julian Bleicken[2], John C. McDonald[1], Sarah Johnson[1]

[1]School of Computing, DePaul University, 243 S. Wabash Ave., Chicago IL 60604, USA
[2]Institute of German Sign Language and Communication of the Deaf (IDGS), University of Hamburg, Binderstr. 34, 20146 Hamburg, Germany

rwolfe@depaul.edu
{thomas.hanke,gabriel.langer,elena.jahn,satu.worseck,julia.bleicken}@uni-hamburg.de
jmcdonald@cs.depaul.edu, sarahej101@gmail.com

Abstract

According to the World Wide Web Consortium (W3C), localization is "the adaptation of a product, application or document content to meet the language, cultural and other requirements of a specific target market" (Ishida & Miller, 2015). One aspect of localizing a sign language avatar is creating a capability to produce convincing *mouthing.* For purposes of this inquiry we make a distinction between *mouthings* and *mouth gesture* (Crasborn et al., 2008). The term 'mouthings' refers to mouth movements derived from words of a spoken language while 'mouth gesture' refers to mouth movements not derived from a spoken language. This paper reports on a first step to identify the requirements for an avatar to be capable of mouthings in multiple signed languages.

Keywords: avatar technology, mouthing, German Sign Language (DGS), lip sync

1. Mouthings in signed languages

The occurrence of mouthings has been reported for many signed languages (cf. for example (Boyes-Braem & Sutton-Spence, 2001) and their origin in spoken languages are self-evident. The prevalence of mouthings varies across different sign languages and individual signers. In German Sign Language (DGS) the occurrence of mouthings is very common. In (Ebbinghaus & Heßmann, 1996) researchers report that the mouthings may be changed and adapted when used in signing. Examples of this include dropping grammatical endings, exaggerating selected elements, holding end position as in the case of word-final L in Apfel or adapting rhythm of mouthed syllables to the rhythm of the manual singing.

While mouthings occur regularly in most sign languages, their significance and status have been a matter of sometimes heated discussions among sign linguists. For example, there is no consensus on the role of mouthings in American Sign Language (ASL) (Lucas & Valli, 1989; Nadolske & Rosenstock, 2007).

However, no matter the theoretical viewpoint one takes on the issue of mouthing, one must acknowledge that for most if not all sign languages mouthings do occur. If an avatar purports to fully express any signed language, it must have the capacity to express all aspects of the language which likely will include mouthings. Without mouthings, avatar signing would not only look unnatural for most sign languages and could also omit important information, resulting in utterances that could be incomprehensible. However, the avatar should also have sufficient flexibility to omit mouthings all together and limit its production exclusively to mouth gestures.

2. History of lip sync technology

Portraying mouthings requires animating an avatar's mouth. Animating an avatar's mouth originates with the technique of creating speaking characters. These first appeared with the advent of sound cartoons in the 1920s (Fleischer, 2005). A believable speaking character requires lip motion that moves in synchrony with a pre-recorded sound track (Johnson & Thomas, 1995), hence the name *lip sync*. Animators drew images to portray *visemes,* or the shape that the lips take while producing the phonemes of a spoken dialog (Fisher, 1968). Because some phonemes appear identical on the face even though they have different sounds, lip sync requires fewer visemes than there are phonemes in a language. For example, for lip sync of English dialog, animation artists typically use between seven and 12 visemes to represent the 44 phonemes of the spoken language (Halas & Manvell, 1971 ; Johnson & Thomas, 1995). However, for extremely simple animation, animators reduce this number to four (Atkinson, 2017) or even two (Hess, 2016). This was a manual, time-consuming process, with the artist being responsible for viseme selection and timing.

The turn of the century witnessed the rise of multimodal technology, which integrated audio and video output in intelligent agents for an enhanced interactive user experiences (Kshirsagar et al., 2002). The intelligent agents were embodied as *avatars* which are representations of human figures. To enhance its human-like qualities, the avatar must move its lips in synchrony with its speech output. This requires the automation of the lip sync animation.

Similar to manually-produced animation, automated lip sync requires a sound track and visemes to generate a talking figure, but it differs in its representation of visemes. The automation strategies fall into two categories, based on the avatar's representation of visemes. In video-based 2D approaches, computer vision techniques analyze frames of pre-existing video recordings and extract the visemes as groups of pixels (Theobald et al., 2003). To accommodate a new sound track, the software identifies and changes the visemes in existing frames of the video to synchronize the lips with the new phoneme stream. When dubbing a movie in a foreign (spoken) language, this technique helps with synchronizing an actor's lip movements with the translated dialog.

In synthetic 3D approaches, the visemes are not sets of pixels, but collections of 3D data. An avatar system can rely directly on a set of artist-created models of an avatar's lip positions to depict visemes. These can use *blend shapes* expressed as polygon meshes, or they can utilize a muscle-based system (Deng & Noh, 2008). Alternatively, it can utilize a high-level animation standard such as MPEG-4 Face and Body Animation which consists of a set of predefined Facial Animation Parameters (FAPs) including 14 static visemes (Pandzic & Forchheimer, 2003).

In 3D strategies, the technique used to generate the animation depends on the source of the dialog. In the case where there is a prerecorded voice track, a speech-recognition module can detect the phonemes and select the corresponding viseme (Zorić & Pandžić, 2005). The viseme choice and timing become part of the data that the animation system interpolates to create the individual frames of the animation. In the case where there is no prerecorded voice track, but only a text containing the dialog, this approach can still be effective if there is a text-to-speech (TTS) service available. Many TTS services provide an option to produce phonemes and timing information as text, which can easily be converted into a stream of viseme choices with timing.

No matter the strategy, there is a question of how best to choose the visemes to match the spoken phonemes for automatic lip sync. (Chen & Rao, 1998) suggested that the possibility of using data-analysis techniques to analyze video recording with the goal of identifying the visemes. However, (Cappelletta & Harte, 2012) examined five phoneme-to-viseme mappings for visual speech recognition, four of which were developed through data analysis and one which was created by linguists. They found that the linguistically-motivated viseme mapping performed the best on visual-only recognition of continuous speech.

3. Lip synch technology for enhanced accessibility

Although most interactive lip sync systems were created for hearing communities, several technologies emerged to improve speech recognition for those who are hard-of-hearing or who find themselves in noisy environments. An early example was a multimedia telephone to assist the hard-of-hearing which used a simple "2-1/2D" head that portrayed lip sync to accompany the voice data (Lavagetto, 1995). A similar project (Oviatt & Cohen, 2000) strove to enhance speech recognition for hearing people located in noisy environments.

One of the distinguishing characteristics between a person who is hard-of-hearing and a person who is Deaf is their language preference. A person who is hard-of-hearing prefers a spoken language, but will use assistive technology such as hearing aids or closed captioning to gain better access to the content of spoken messages. In contrast a person who identifies as Deaf will use a signed language, such as ASL or DGS as their preferred language (Padden & Humphries, 1988).

For the Deaf community, access to spoken language, or to the written form of a spoken language requires translation to the preferred signed language. An essential part of any automatic spoken-to-sign translation system is an avatar capable of producing all aspects of the language, including mouthings. The earliest avatar designed specifically to provide improved access to the Deaf community was part of the ViSiCAST project (Elliott et al., 2000). This project included the development of the Signing Gesture Markup Language (SiGML), based on HamNoSys (Hanke, 2004). It specifies a *mouth picture* or viseme for each letter of the International Phonetic Alphabet (IPA) (Glauert et al., 2004). Strings of visemes are expressed using SAMPA encoding conventions and the mapping of SAMPA symbols to visemes is part of an avatar-specific configuration file. The mapping was subsequently revised, and the current pronunciation dictionary used for DGS is from IKP Bonn (Aschenberner & Weiss, 2005).

The sign annotation software iLex uses the same system for annotating mouthings in lexical items (Hanke, 2002). (Elliott et al., 2008) describe the continuation of this research as part of the eSIGN project, and gives a complete example of SiGML notation, including mouthings, for the DGS sign HAUS, as well as selected frames from an avatar signing HAUS. (Jennings, Elliott, Kennaway, & Glauert, 2010) give an in-depth discussion of the implementation details for Animgen, the animation engine used to create the avatar. To implement mouthings, they use a set of blend shapes, one for each mouth picture.

Contemporaneous with the ViSiCAST/eSIGN projects, other groups explored the possibility of incorporating mouthings in sign language technology. These include projects at the German Research Center for Artificial Intelligence (DFKI) (Heloir, Nguyen, & Kipp, 2011) (Kipp, Heloir, & Nguyen, 2011) and DePaul University (Wolfe et al., 2009). The primary goal of the avatar developed at DFKI is to synthesize DGS, and it uses the OpenMARY speech synthesis system to generate the viseme specification and timing, but no mention was made of the underlying technology for representing individual visemes. In contrast, the avatar "Paula" developed at DePaul generates ASL, and uses a Microsoft.NET Text-to-Speech (TTS) service to generate the viseme selection and timing. Because the face is represented by a muscle system, Paula's mouth animation is not limited to linear combinations of selected visemes.

4. Extending a muscle-based avatar

In a past study (Schnepp, Wolfe, McDonald, & Toro, 2012), members of the Deaf community in the US viewed the Paula avatar with and without mouthings and consistently indicated a preference for animations with mouthings. Encouraged by this feedback, we are exploring the feasibility of adding localization to Paula. As a first step we attempted to teach Paula to sign DGS, for which mouthings are an important feature. For this first inquiry, we chose six signs from a previously existing vocabulary for Swiss German Sign Language (Ebling, et al., 2017) whose manual channel match signs in DGS. See Figure 1.

Creating the mouthings posed several challenges. The TTS library was specific to English, and had occasional difficulties in synthesizing spoken German words. Correcting these instances required manual editing of several of the generated viseme streams.

APFEL
BALL
BÄR
JUNGE
UHR
ZEIT

Figure 1: DGS signs under investigation

Another challenge was the style of enunciation. In general, native US speakers of English demonstrate an economy of lip motion in conversation. This, coupled with the lack of consensus on the role of mouthings in ASL lead to a previous design decision to keep Paula's lip movement to a minimum and to provide an option to omit it all together.

In contrast, mouthings often appear in DGS. Furthermore, there are important differences between German and English bases of articulation (Hall, 2003). Spoken German has a greater articulatory tension; muscles in the articulators such as the tongue are tenser, resulting in pronunciations that are more forceful. The tongue takes on positions that are more extreme and more prominent. Lip movements are more vigorous in German. Vowels such as /u:/ and /y:/ are articulated with strongly protruding and rounded lips.

These differences in the basis of articulation required adjustment of the viseme weights. Instead of the standard 30% of maximum viseme strength that had been used to accompany ASL, the DGS settings ranged from 50% to 120% of the (original) maximum. Figure 1 demonstrates the difference in the spoken English viseme and DGS mouth shape for the /s/ phoneme. We were motivated by the feedback of one of our German colleagues, who said, "I need to see more teeth!"

US-English viseme — Prototype DGS mouth shape

Figure 2: Mouth shapes for /s/

Informed by data from the DGS-Korpus (Blanck, et al., 2010) we also adjusted the timing of the viseme onset. Instead of coinciding with the onset of the manual channel of a lexical item, the mouthings in DGS tend to start earlier. Based on this finding, we set the onset of the lip motion to begin 0.2 seconds before the onset of the manual channel.

5. A first feedback session

A group of six linguistically aware native signers of the German Deaf community participated in a first feedback session. The session began with a brief introduction to avatar technology and its possible applications. Then, to familiarize the group with the current capabilities of avatar technology, a moderator presented three short animations that demonstrate the state of the art in sign language technology (Jordaan, 2014 ; Brun, 2014 ; The ASL Avatar Project Team at DePaul University, 2012), and conducted a discussion that compared the three animations.

A second moderator presented the newly-created DGS signs complete with a typical mouthing. Each sign was presented as a series of three slides. See Figure 3. The first slide simply gave an identification number for the sign. The next two slides contained the same identification number and a video frame. The first video used a medium shot showing the avatar from the waist up. The second video showed the same sign, but used a close-up shot, to show the mouth in extreme detail. The moderator played the videos as many times as group members requested.

After playing the first (medium shot) video, the moderator asked the group to identify the sign, and solicited comments on what they liked and what needed to be improved. After playing the second (close-up shot) video, the moderator again solicited comments on what needed improvement.

Other than the brief introduction, which was presented in written German and interpreted into DGS, the entire session was conducted by two Deaf moderators. To accommodate the hearing note takers, the discussions were voiced in German by an interpreter.

6. Feedback

In all cases, the signs were immediately identified, which was consistent with results previously received from a focus group fluent in DSGS (Ebling, et al., 2017). The color selections for the avatar clothing, hair and background made it easy to read the manual channel and it was well positioned in the signing space. However, the lighting on the face was too even and needs to reveal the contours of the lower face. Viewers wanted to the nasolabial folds (smile lines) to be clearly visible at all times.

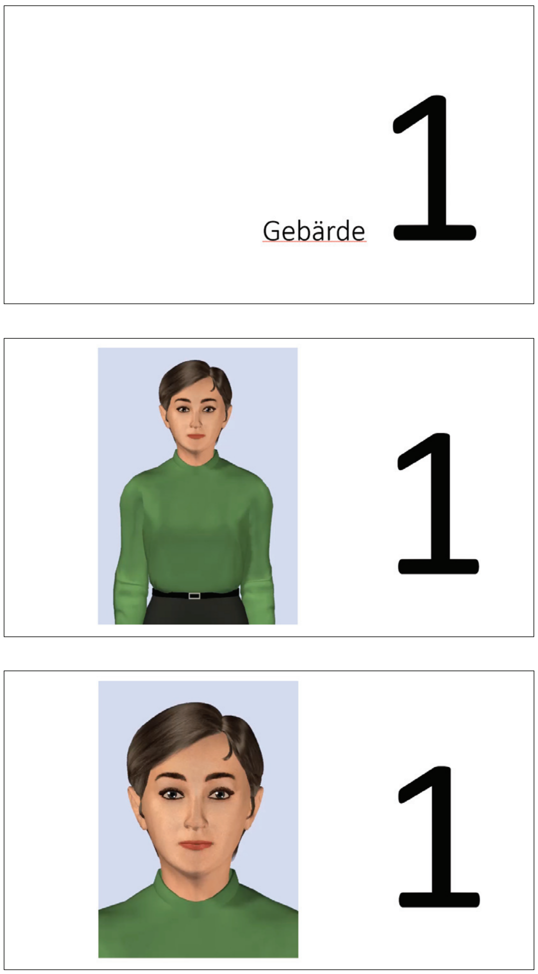

Figure 3: Slide format for presenting avatar signing DGS

There were several issues identified with the mouthings. There was general agreement that all of the visemes need to be more "pronounced". The teeth needed to be more prominent. The word-final viseme corresponding to /l/ in BALL and APFEL requires the tongue to be farther forward in the mouth, with the blade of the tongue at the alveolar ridge, and the tongue tip behind the upper teeth. This is consistent with the findings of (Ebbinghaus & Heßmann, 1994).

However, there was one aspect of creating a more vigorous pronunciation that had nothing to do with the lips themselves. Group members consistently pointed to a lack of cheek motion in the mouthings. They indicated that cheek movement is important for all visemes, but are particularly vital for the labial plosives /b/ and /p/, and demonstrated how the cheek movement is necessary for a mouthing that is easy to recognize.

7. Conclusions and future work

In this effort, we explored the viability of avatar localization, to identify challenges of adapting an avatar to produce sign languages from different geographic regions. We started with an avatar that was designed to produce ASL, and used it to create lexical items in DGS. The major change was to modify its capabilities to produce DGS mouthings. We then solicited feedback from linguistically aware native DGS signers. Although all of the lexical items were immediately identified, there were issues with several of the visemes comprising the mouthing.

Previously, focus of viseme development has been almost exclusively on mouth shape. Future avatar development will need to consider how to incorporate areas surrounding the mouth including cheeks and nose for improved legibility.

Our preliminary findings seem to run counter to (Glauert 2004)'s supposition that if single set of visemes will suffice for mouthings in all signed language. The visemes we created to support spoken English are inadequate for DGS. It will be necessary to create a library of visemes, preferably by artists aware of the role of mouthings in DGS. However, for effective production of mouthings it will not suffice to use such a library with a simple surface mapping from audible phonemes produced by a TTS. Ultimately, it will require corpus data that contain instances consistent with the findings of (Elliott E. A., 2013) and (Ebbinghaus & Heßmann, Signs and words: Accounting for spoken language elements in German Sign Language, 1996) that demonstrate the adaptation of mouthings as produced in DGS.

However, it would be interesting to further explore the possibility of viseme reuse to support mouthings for multiple signed languages. Creating a set of language-independent visemes for spoken languages has been a topic of research for some time (Zorić & Pandžić, Real-time language independent lip synchronization method using a genetic algorithm, 2006). However, attempting to extend this idea to signed languages is an open question and an intriguing topic for future work.

8. Acknowledgments

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

9. References

Aschenberner, B., & Weiss, C. (2005). *Phoneme-Viseme Mapping for German Video-Realistic Audio-*

Visual-Speech-Synthesis IKP-Working Paper NF 11. Retrieved 2018, from CiteseerX: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.8839&rep=rep1&type=pdf

Atkinson, D. (2017, December 24). *Frank Hellard's mouth shapes.* Retrieved December 24, 2017, from Centre for Animation & Interactive Media, RMIT University: http://minyos.its.rmit.edu.au/aim/a_notes/mouth_shapes_02.html

Blanck, D., Hanke, T., Hofmann, I., Hong, S.-E., Jeziorski, O., Kleyboldt, T., . . . Wager, S. (2010). The DGS Corpus Project. Development of a Corpus Based Electronic Dictionary German Sign Language – German. *Theoretical Issues in Sign Language Research (TISLR) 10.* West Lafayette, Indiana.

Boyes-Braem, P., & Sutton-Spence, R. (2001). *The hands are the head of the mouth.* Hamburg: Signum-Verlag.

Brun, R. (2014, May 01). *Sign Language: Gallaudet University | Mocaplab.* Retrieved December 24, 2017, from mocaplab: http://www.mocaplab.com/projects/gallaudet-university/

Cappelletta, L., & Harte, N. (2012). Phoneme-to-viseme Mapping for Visual Speech Recognition. *ICPRAM (2)*, (pp. 322-329).

Chen, T., & Rao, R. R. (1998). Audio-visual integration in multimodal communication. *Proceedings of the IEEE, 86*, 837-852.

Crasborn, O. A., Van Der Kooij, E., Waters, D., Woll, B., & Mesch, J. (2008). Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language \& Linguistics, 11*, 45-67.

Deng, Z., & Noh, J. (2008). Computer facial animation: A survey. In *Data-driven 3D facial animation* (pp. 1-28). Springer.

Ebbinghaus, H., & Heßmann, J. (1994). Formen und Funktionen von Ablesewörtern in gebärdensprachlichen Äußerungen (Teil I). *Das Zeichen, 30*, 480-487.

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

Ebling, S., Johnson, S., Wolfe, R., Moncrief, R., McDonald, J., Baowidan, S., . . . Tissi, K. (2017). Evaluation of Animated Swiss German Sign Language Fingerspelling Sequences and Signs. *Universal Access in Human-Computer Interaction*, 3-13.

Elliott, E. A. (2013). Phonological Functions of Facial Movements: Evidence from deaf users of German Sign Language. Berlin: Doctoral dissertation, Freie Universität Berlin.

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).

Elliott, R., Glauert, J. R., Kennaway, J. R., Marshall, I., & Safar, E. (2008). Linguistic modelling and language-processing technologies for Avatar-based sign language presentation. *Universal Access in the Information Society, 6*, 375-391.

Fisher, C. (1968, December). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 1*(11), 796-804.

Fleischer, R. (2005). *Out of the Inkwell: Max Fleischer and the Animation Revolution.* Lexington: University Press of Kentucky.

Glauert, J. R., Kennaway, J. R., Elliott, R., & Theobald, B.-J. (2004). Virtual human signing as expressive animation. *Symposium on Language, Speech and Gesture for Expressive Characters, University of Leeds*, (pp. 98-106).

Halas, J., & Manvell, R. (1971). *The Technique of Film Animation.* Waltham: Focal Press.

Hall, C. (2003). *Modern German pronunciation: An introduction for speakers of English.* Manchester University Press.

Hanke, T. (2002). iLex-A tool for Sign Language Lexicography and Corpus Analysis. *LREC.* Las Palmas: ELRA.

Hanke, T. (2004). HamNoSys -- Representing sign language data in language resources and language processing contexts. *Fourth International Conference on Language Resources and Evaluation (LREC 2004). Representation and Processing of Sign Languages Workshop* (pp. 1-6). Paris: European Language Resources Association.

Heloir, A., Nguyen, Q., & Kipp, M. (2011). Signing avatars: A feasibility study. *The Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT), Dundee, Scotland, United Kingdom.*

Hess, D. (2016, August 17). *Lip Semi-Sync: The Incomplete Guide to Mouth Flapping.* Retrieved December 24, 2017, from DIY Animation Club: https://diyanimation.club/animation-lip-sync-mouth-flapping/

Ishida, R., & Miller, S. K. (2015, April 9). *Localization vs. Internationalization.* Retrieved December 23, 2017, from World Wide Web Consortium: https://www.w3.org/International/questions/qa-i18n

Jennings, V., Elliott, R., Kennaway, R., & Glauert, J. (2010). Requirements for a signing avatar. *Proceedings of the 4th LREC Workshop on the Representation and Processing of Sign Languages*, (pp. 133-136).

Johnson, O., & Thomas, F. (1995). *The Illusion of Life: Disney Animation.* New York: Random House (Disney Press).

Jordaan, B. (2014, November 13). *Introductory video of myself at "Festmeny" show.* Retrieved December 24, 2017, from Facebook: https://www.facebook.com/braam3D/videos/10153283146689409/

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

Kshirsagar, S., Magnenat-Thalmann, N., Guye-Vuilleme, A., Thalmann, D., Kamyab, K., & Mamdani, E. (2002). Avatar markup language. *ACM International Conference Proceeding Series*, *23*, pp. 169-177.

Lavagetto, F. (1995). Converting speech into lip movements: A multimedia telephone for hard of hearing people. *IEEE Transactions on Rehabilitation Engineering, 3*, 90-102.

Lucas, C., & Valli, C. (1989). Language contact in the American deaf community. In C. Lucas (Ed.), *The Sociolinguistics of the Deaf Community* (pp. 11-40). San Diego, CA: Academic Press.

Nadolske, M. A., & Rosenstock, R. (2007). Occurrence of mouthings in American Sign Language: A preliminary study. *Trends in linguistics studies and monographs*, 35.

Oviatt, S., & Cohen, P. (2000). Perceptual user interfaces: multimodal interfaces that process what comes naturally. *Communications of the ACM, 43*, 45-53.

Padden, C., & Humphries, T. (1988). *Deaf in America: Voices from a Culture.* Cambridge: Harvard University Press.

Pandzic, I., & Forchheimer, R. (Eds.). (2003). *MPEG-4 facial animation: the standard, implementation and applications.* . Hoboken, NJ: John Wiley & Sons.

Schnepp, J. C., Wolfe, R. J., McDonald, J. C., & Toro, J. A. (2012). Combining emotion and facial nonmanual signals in synthesized american sign languag. *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility* (pp. 249-250). Dundee, UK: ACM.

The ASL Avatar Project Team at DePaul University. (2012, May 31). *Demonstration video of our ASL Avatar.* Retrieved December 24, 2017, from The American Sign Language Avatar Project: http://asl.cs.depaul.edu/demo.html

Theobald, B.-J., Bangham, J. A., Matthews, I., & Cawley, G. (2003). Evaluation of a talking head based on appearance models. *AVSP 2003-International Conference on Audio-Visual Speech Processing.*

Wolfe, R., Cook, P., McDonald, J., & Schnepp, J. (2009). *Toward a Better Understanding of Nonmanual Signals through Acquisition and Synthesis.* Frankfurt: Presented at Nonmanuals in Sign Languages (NISL).

Zorić, G., & Pandžić, I. (2005). A real-time lip sync system using a genetic algorithm for automatic neural network configuration. *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, (pp. 1366-1369).

Zorić, G., & Pandžić, I. (2006). Real-time language independent lip synchronization method using a genetic algorithm. *Signal processing, 86*(12), 3644-3656.