Refinement of Lip Shape in Sign Speech Synthesis

Zdeněk Krňoul

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Pilsen, Czech Republic
zdkrnoul@kky.zcu.cz

Abstract

This paper deals with an analysis of lip shapes during speech that accompanies sign language, referred to as *sign speech*. A new sign speech database is collected and a new framework for the analysis of mouth patterns is introduced. Using a shape model restricted to the outer lip contour, we show that the articulatory parameters for visual speech alone are not sufficient for representing sign speech. The errors occur mainly for the mouth opening. A correction to the standard articulatory parameters and additional articulatory parameters are investigated to cover the observed mouth patterns and thus refine the synthesised sign speech.

Index Terms: visual speech synthesis, talking head, sign speech synthesis, articulatory parameters

1 Introduction

A sign language synthesis system that translates from spoken words to sign language, including signed speech, provides a more user-friendly interface for the deaf community. There is a strong (negative) correlation between reading ability and degree of hearing loss, especially in deaf child [1]. The Czech grammar is a significant barrier for Czech deaf children because the Czech language is inflected and has a freer word order. Although they may understand all of the individual words in a written text, the sentence does not make sense [2].

Two communication forms have to be distinguished: the Czech Sign Language (CSE) and Signed Czech (SC). The CSE is the primary communication means of the deaf or hearing-impaired people in the Czech Republic. CSE uses the specific visual-spatial resources, i.e. hand shapes and movements (the manual component), facial expressions, mouth patterns, head and upper body positions (the non-manual component). CSE is not derived from or based on spoken Czech. There are basic language attributes, i.e. system of signs, double articulation, and it has its own lexical and grammatical structure [3]. On the other hand, SC was introduced as an artificial language system derived from spoken Czech language to facilitate communication between deaf and hearing people. SC uses grammatical and lexical resources of the Czech language. The non-manual component of signed utterances incorporates audibly or inaudibly articulated individual words spoken simultaneously with the hand articulation of CSE signs [4].

In the scope of this paper, we deal with the non-manual component of CSE. The non-manual component is an important factor for the perception by humans of sign speech [5] as it distinguishes the phonetic content of the speech, but does not provide information regarding the emotional state of the speaker. The mouth patterns used in non-manual component of CSL include lip shapes corresponding the words taken from the spoken language and also artificial combinations of phonemes (visemes) of the spoken language. Further, we find special mouth patterns (mouth gestures) that do not correspond to any well-known viseme used in spoken language. It should be noted too that an inventory of the mouth patterns has to be considered for particular sign language [6].

From point of view of data collection for speech synthesis, it is necessary to distinguish mouth patterns produced by deaf people who have only partial speaking skills (an analysis has been done for example in the ECHO project) and people who are professional sign language interpreters make the effort to produce intelligible sign speech.

This paper is structured as follows. Section 2 describes the acquisition of the sign speech database. Section 3 deals with the selection of appropriate data for the proposed refinement of synthesis process. Section 4 introduces a framework for analysis of mouth patterns. Section 5 describes the experiment using the proposed framework and the shape model designed for the of talking head system.

2 Sign Speech Database

Visual databases capturing speech accompanying sign language are not yet very widespread. From the perspective of sign language recognition, databases are available that capture more signers recorded in studio conditions, but include a lexical of only a few tens of signs [7]. From the perspective of 3D synthesis of continuous sign speech, the recording of sign speech data is a more complex problem. Data-driven synthesis systems uses motion capture to record the manual and the non-manual components, which include the movements of the head, the facial expressions and the lip articulations simultaneously [8]. These data are primarily obtained as the 3D coordinates and are specific to the particular speaker. On the other hand, we can distinguish this from rule-based approaches [9]. The lexical signs are captured in a form of rules and symbols, which are manually entered by annotators. In the original meaning, the control model used for this approach is purely deterministic, for example a small forward movement, circular movement or eyebrow raising. The symbolic notation allows a user to specify for example three sizes (intensity) of these actions.

The sign speech database captured here is continuous sign speech and is collected with one primary aim: To explore mouth patterns and to improve the animation of the non-manual component of CSL. The sign speech database contains video records of television news interpreted in CSL. Each video record of the database contains continuous sign speech interpreted by one speaker and takes an average of 10 minutes. The database con-

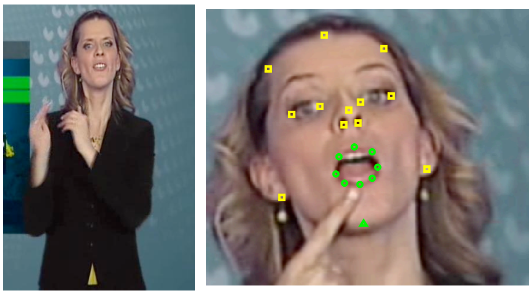

Figure 1: An example of image data collected in the sign speech database. On the left: the bounding box around the signer, on the right: the region of interest and the feature points used in the proposal (data source www.ceskatelevize.cz/ivysilani).

tains five different speakers (three female and two male). Each video record is composed of the three separate parts: general news, sport and weather forecast. The signer is always in the right half of the image. He or she does not have any auxiliary marks, and clothing and hair style may vary across records. The video of the signer is recorded in a TV studio against a chroma-key background — Figure 1 shows an example. The format of the video is 720x576 pixels, 25 fps, and RGB colour. The number of signers and recordings are summarized in Table 1. The bounding box around the signer is approximately 340x540 pixels and a region of interest around the head is approximately 300x250 pixels.

Table 1: A statistic of the sign speech database.

Speaker	F1	F2	F3	M1	M2
#Records	7	6	4	10	7

3 Mouth Patterns and Templates

An automatic parameterization of the complete non-manual component of sign speech from standard video capturing the speaker's head without auxiliary marks and with sufficing accuracy for the synthesis process is currently unknown. For small speaker's head rotations, the active appearance model (AAM) seems to be sufficient [10]. In next section we introduce an alternative approach that enables the 3D structure of the speaker's head recorded in a 2D video sequence to be reconstructed.

The first step of this investigation is a selection of several key frames (templates) from the collected sign speech database. For this purpose a female speaker, identified in Table 1 as F3, and a segment of general TV news is considered. The region of interest is limited to speaker's head only, see Figure 1 (right). Approximately 6000 frames were manually analyzed corresponding to 4 min of video. These frames are chosen to include different poses of the speaker's head and different lip shapes. The selected images form *templates*. In total 127 templates are identified — nine for the 3D shape of the head and 118 for lip gestures. The lip shapes are selected to cover all Czech vowels (a, e, i, o, u) and the consonants from the viseme groups (v,f) (c,s,z), (S Z) and (p,b,m). For /l/ there is visible tongue movement from the top to bottom of the mouth, but the lip shape is generally incorporated into adjacent lip shapes in the speech context. For non-articulatory lip shapes we use two mouth patterns defined in the ECHO project for the NGT annotation: "lip closed and stretched-up or down" and lip shape in "tongue extra feature", see [6].

4 Analysis of Mouth Patterns in Templates

4.1 Feature Points

For the analysis of mouth patterns 21 manually placed feature points in each template are sufficient, see Figure1 (right). This basic parameterization captures both the actual pose of the speaker's head and the lip shape. The pose of the head is captured by 12 feature points, W_{head}, which include the ear lobes, the eye corners, the nose tip, the nostrils, and three points along the forehead-hair line (illustrated as rectangles in Figure 1). These feature points are considered projections of a 3D rigid body. To capture non-rigid variation, eight feature points, W_{lip}, are marked around the outer lip boundary (illustrated as circles in Figure 1) plus an additional point on the chin (the triangle in Figure 1).

4.2 Structure from Motion

Methods for recovering 3D structure from motion (SFM) can be applied to two or more video frames: either a moving object in one view at multiple time steps, or a static object from different view points. The framework described here assumes an orthogonal rather than a perspective camera model. Whilst the perspective model is more promising for real data, it lacks robustness. The justification for selecting the orthogonal camera model can be confirmed given the type of images stored in the sign speech database and the parameterization of the templates: The image resolution and the number of feature points are relatively low. The type of data also ensures that errors of pose estimations caused by movements of the object along the camera optical axis are eliminated.

For this work we employ the factorization method of [11], which uses an affine camera matrix and 2D calibrated measurements. It is a linear design using a fixed number of the feature points detected in all images (the templates). The image positions of N feature points and F templates form a $2F \times N$ matrix, W. This matrix includes sub-matrices, $W_f = [w^u_f w^v_f]^T$ $f = 1 \dots F$, where u, v indicate the image plane coordinates of f-th template and N feature points. The $3 \times N$ matrix S of 3D point positions approximating the reconstructed structure is projected to W by the projection model:

$$W = MS + T, \tag{1}$$

where M is the $2F \times 3$ matrix composed from the 2×3 matrices $M_f = [m^u_f m^v_f]^T$ modeling rotations of S in the image planes. The $2F \times N$ matrix T contains relevant translations of S. The factorization assumes an elimination of the translation T. We can compute the factorization of $\tilde{W} = M\hat{S}$ using Singular Value Decomposition (SVD):

$$\tilde{W} = UDV^T, \tag{2}$$

where $\tilde{W} = W - [\mu_W \dots \mu_W]_{2F \times N}$ are normalized positions of feature points, the $2F \times 1$ vector μ_W contains mean u and v coordinates of N feature points in all templates and $\hat{S}$ contains coordinates of the S structure centered around the origin.

The orthonormal matrix $2F \times 2F$ U and diagonal matrix D with non-negative entries are used to set the projection $\tilde{M} = UD$.

Since the rank of the matrix $\tilde{W}$ is three, the first three eigenvectors are sufficient to describe the orthogonal projection of $\hat{S}$ to the image plane. Since the matrix $\tilde{M} = UD$ does not include rotation matrices, a rectification of $\tilde{M}$ to matrix M is determined to impose the orthonormal constrains [12]. The final step is the bundle adjustment of M to calculate the affine matrices M_f. The structure $\hat{S}$ is a solution of the overdetermined system $M\hat{S} = \tilde{W}$ in the sense of least squares. The shape $\hat{S}$ is projected to f-th frame as:

$$W_f = M_f \hat{S} + [\mu_{fW} \ldots \mu_{fW}]_{2\times N}, \quad (3)$$

where $\mu_{fW} = [\mu_f^u \mu_f^v]^T$ is the 2×1 vector of mean feature point coordinates of f-th template.

4.3 3D Lip Structure and Template Projections

The first nine templates, which captured the speaker's head position with relaxed lips, are used to reconstruct the 3D shape of the outer lip contour. We apply the factorization (3) for F = 9 and N = 21 ($W_{head} + W_{lip}$, see Section 4.2) to obtain the structure S_{relax} including the 3D shape of W_{head} together with the 3D shape of the outer lip contour, see Figure 2.

SFM also applied to all 127 templates, but without the feature points of the lips and chin. The rigid body condition here is F = 127 and N = 12 (W_{head} only). This results in the 3D structure S_{tem} and the matrix of rotations M. Thus, we have pose estimations that enable us to project the S_{tem} structure to all templates, $f = 1 \ldots 127$. Next, an intuitive step is to find the 3D affine transformation between the structures S_{tem} without lips and S_{relax} with lips — i.e., so we can normalize the templates for rigid 3D head pose. For this purpose, the structure S_{relax} is divided to the 3×12 structure $S_{relaxhead}$ (features marked with squares in Figure 1) and the 3×8 structure $S_{relaxlip}$. Assuming the 3D shapes of S_{tem} and $S_{relaxhead}$ are sufficiently similar, the pose from $S_{relaxhead}$ to S_{tem} is estimated using:

$$S_{tem} = R_{tem} S_{relaxhead} + [t_{tem} \ldots t_{tem}]_{3\times 12}. \quad (4)$$

Furthermore, we apply the affine transformation, R_{tem}, t_{tem}, to the structure $S_{relaxlip}$ to get the 3D structure of the outer lip contour of the speaker's lip that can be already projected to all templates as:

$$S_{im} = R_{tem} S_{relaxlip} + [t_{tem} \ldots t_{tem}]_{3\times 8}. \quad (5)$$

4.4 Projection of Shape Model

For the proposed refinement of the sign speech synthesis system, we consider a shape model of the talking head system restricted to the outer lip contour. The shape model has the form:

$$S_{th} = S_{threlax} + \sum_{k=1}^{K} (PC_k x_k). \quad (6)$$

The structure $S_{th} = [\mathbf{s}_{th1} \ldots \mathbf{s}_{th8}]$ is the 3×8 matrix involving the 3×1 vector s_{th-i} of the actual 3D position of i-th control point. $S_{threlax}$ are 3D positions of 8 control points in the neutral (relaxed) position. The 3×8 matrices $PC_k = [\mathbf{pc}_{k1} \ldots \mathbf{pc}_{k8}]$, $k = 1 \ldots K$ are used to model differences between $S_{threlax}$ and k-th key lip shape, see Figure 4. The 3×1 vector pc_{ij} is one difference of i-th control point and j-th key lip shape. The vector $\mathbf{x} = [x_1 \ldots x_K]^T$ stores weight x_k of k-th PC (the articulatory parameters).

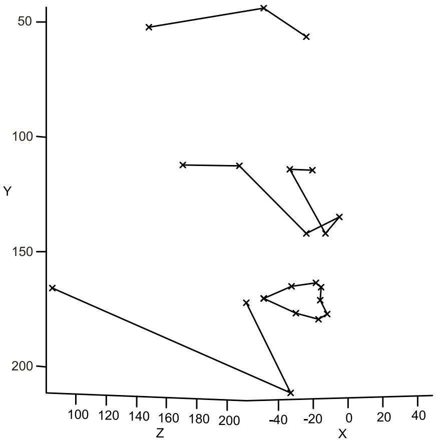

Figure 2: 21 3D positions of the feature points define the structure S_{relax}. The auxiliary lines in Figure emphasize the 3D shape of face only.

The pose estimation determines the transformation of the shape model S_{th} for $\mathbf{x} = \mathbf{0}$ to the image space of the structure S_{im} as:

$$S_{im} = sRS_{th} + [t \ldots t]_{3\times 8}. \quad (7)$$

An assumption, that the structure S_{th} for $\mathbf{x} = \mathbf{0}$ and S_{im} share the similar 3D shape, enables us to use the transformation s, R and t for the shape model (6) and substitute it to (3) as:

$$W_{fth} = M_f sR(S_{threlax} + \sum_{k=1}^{K} (PC_k x_k)) + t) + [\mu_{fW} \ldots \mu_{fW}]_{2\times N}, \quad (8)$$

where W_{fth} are the 2D image coordinates of the control points projected to f-th template.

An estimation error $||W_{fth} - W_f||$ is evaluated for all templates, $f = 1 \ldots 127$. The estimation error can be converted to an overdetermined linear system $A_f \mathbf{x}_f = \mathbf{b}_f$ of unknown values of the articulatory parameter vector $\mathbf{x}_f$. The matrix A_f and the vector $\mathbf{b}_f$ are then given as:

$$A_f = \begin{bmatrix} \mathbf{m}_f^{uT} sR\mathbf{pc}_{11} & \ldots & \mathbf{m}_f^{uT} sR\mathbf{pc}_{1K} \\ \ldots & & \\ \mathbf{m}_f^{uT} sR\mathbf{pc}_{81} & \ldots & \mathbf{m}_f^{uT} sR\mathbf{pc}_{8K} \\ \mathbf{m}_f^{vT} sR\mathbf{pc}_{11} & \ldots & \mathbf{m}_f^{vT} sR\mathbf{pc}_{1K} \\ \ldots & & \ldots \\ \mathbf{m}_f^{vT} sR\mathbf{pc}_{81} & \ldots & \mathbf{m}_f^{vT} sR\mathbf{pc}_{8K} \end{bmatrix} \quad (9)$$

$$\mathbf{b}_f = \begin{bmatrix} w_{f1}^u - \mathbf{m}_f^{uT} sR\mathbf{s}_{th1} - \mu_f^u \\ \ldots \\ w_{f8}^u - \mathbf{m}_f^{uT} sR\mathbf{s}_{th8} - \mu_f^u \\ w_{f1}^v - \mathbf{m}_f^{vT} sR\mathbf{s}_{th1} - \mu_f^v \\ \ldots \\ w_{f8}^v - \mathbf{m}_f^{vT} sR\mathbf{s}_{th8} - \mu_f^v \end{bmatrix} \quad (10)$$

K unknown values of articulatory parameters $\mathbf{x_f}$ and the template f are a least squares solution that minimizes the norm $||A_f \mathbf{x}_f - \mathbf{b}_f||$.

5 Experiment

The proposed experiment investigates settings of the shape model in accordance with the templates collected in Section 3 and the analysis introduced in Section 4. The list of all considered groups of articulatory parameters (key lip shapes) is in Table 2. All articulatory parameters are modeled manually in accordance with required mouth pattern. Firstly, we verify a standard model design. The PC1–4 parameters are the parameters commonly used for visual speech. These parameters are gradually added, K = 1,2,3,4, see the groups G1 to G4 in Table 2. The median projection errors are in the bar graph in Figure 3, bar G1 to G4, 95% bootstrap confidence intervals are added to the scores.

The median error for the standard design of the shape model is 6.2 pixels (the bar G4). Figure 5 separately shows the reconstruction errors of all templates. The reconstruction errors are depicted in the ascending order (solid thick line). It can be seen from the graph that the projection error of the last ten templates significantly increases.

Table 2: Groups of articulatory parameters used in the proposed experiment.

Group	Parameters
G1	PC1
G2	PC1 PC2
G3	PC1 PC2 PC3
G4	PC1 PC2 PC3 PC4
G5	PC5 PC2 PC3 PC4
G6	PC5 PC2 PC3 PC4 PC6
G7	PC5 PC2 PC3 PC4 PC6 PC7
G8	PC5 PC2 PC3 PC4 PC6 PC7 PC8

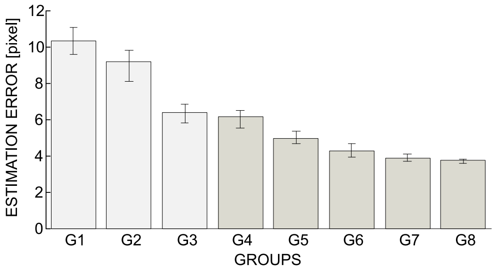

Figure 3: Median projection errors and confidence intervals determined for the groups of the articulatory parameters.

A detailed look at the templates with a projection error greater than 10 pixels results in seven templates with a very large mouth opening, e.g. the vowel /a/. This mouth pattern is not possible to set up by the standard design of the talking head system. The first step does not define new articulatory parameter, but rather tries to remodel the parameters PC1 (lip opening). The talking head system and manual tuning is used to create new key lip shape, PC5, that allows the shape model (6) to generate a deeper shift of lower lip. Thus, the PC1 parameter is replaced by the PC5 parameter and projection errors of templates are re-estimated, see group G5. The median projection error is reduced to 5.0 pixels. Once more, the ascending order of the projection error is in Figure 5 as thick doted line. Now, we can observe only four templates with largest projection errors. These templates capture mouth patterns with significant uncovering of the upper and lower teeth that is caused by wide range displacement of the lips. The key lip shape marked as PC6 is manually modeled using the talking head system and assigned to the G7 group, see Figure 4. The median projection error is significantly reduced to 4.3 pixel, see thick dashed line in Figure 5.

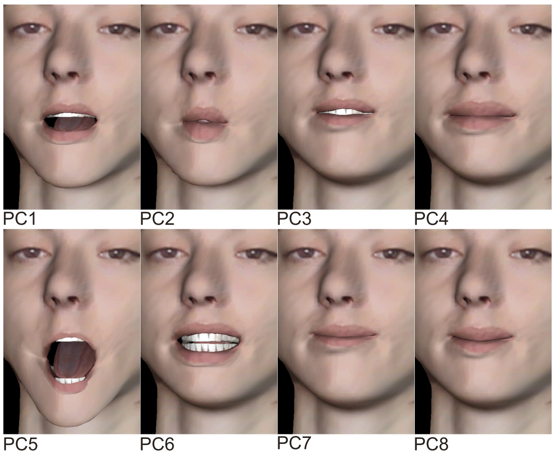

Figure 4: The key lip shapes considered in the proposal. The up four lip shapes show the standard articulatory parameters of the talking head system. The parameter, PC5, is correction of the parameter, PC1. PC6 to PC8 are experimentally modeled lip shapes considered in the proposed experiment.

In this stage of the experiment, the control points of the shape model projected to all templates are inspected in detail. Improvement is observed across all templates. The projection error less than 4 pixels indicates that the shape of the outer lip contour projected to the templates is satisfactory. The 3D lip shapes produced by the shape model (6) and the estimated parameters are correct as well. In particular, relaxed lip closing of viseme (p,b,m) and /o/ and /u/ vowels are very well approximated. The mean projection error around 6 pixels is mainly caused by noise caused by the manual placement of the feature points in the templates and associated estimation errors of the 3D position of the speaker's head.

However, small inaccuracies in the template projections could be observed. They are caused by raising or falling of lip corners or raising of entire mouth. These mouth patterns are experimentally added to the group G6. The median projection error decreases to 3.9 pixels for the G7 group and 3.7 pixels for the G8 group, see thin solid and doted line in Figure 5, .

6 Conclusion

This proposal investigates a standard design of the talking head system applied to the synthesis of the non-manual component of Czech sign language. For this purpose, the sign speech database containing both manual and non-manual component was collected and selected video frames were used in the experiment. The aim of the experiment is to verify the ability of the talking head system to produce the desired mouth patterns observed in the video frames.

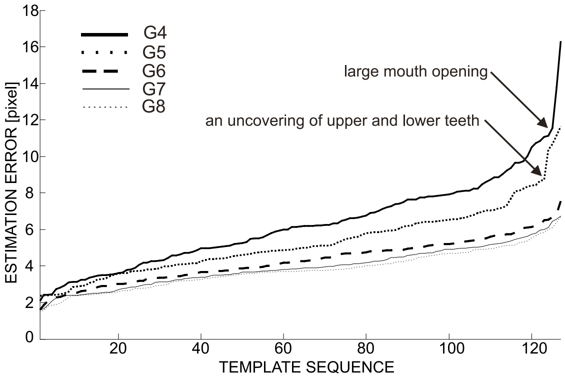

Figure 5: Projection errors of the templates on each occasion sorted in the increasing order.

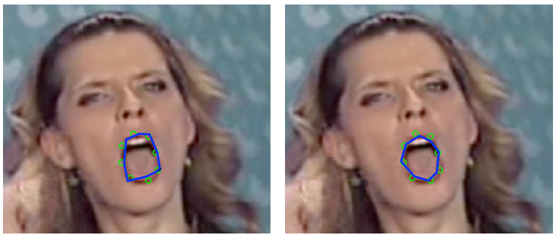

Figure 6: An estimation of large mouth opening by the standard design (the G4 group) and the refined approach (the G5 group, data source, www.ceskatelevize.cz/ivysilani).

For this purpose, a framework for orthogonal projections of the 3D model to 2D images is designed. The proposed linear solution is able to estimate unknown values of pre-defined articulatory parameters from speech data including movements of speaker's face. The proposed framework is verified on 127 mouth patterns (the templates) and eight groups of the articulatory parameters.

The result of the experiment indicates the correctness of the current articulatory parameter "lip opening" in order to animate mouth patterns involving large mouth openings. Furthermore, the experiment shows that the significant decrease of the projection error is caused by supplying new articulatory parameter modeling the wide ranged upper and lower teeth uncovering. In addition, next two articulatory parameters are considered. Nevertheless, this extension does not cause a significant decrease of the projection error.

The advantage in terms of synthesis is that the proposed refinement of the synthesis system is done using speech data disseminated by public television and accepted by the deaf community. Future application of the sign speech database can be found for an automatic segmentation process of sign speech to lexical signs or smaller units. Transcription, segmentation and an extension of the symbolic notation system will be investigated as well.

7 Acknowledgments

This research was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/09/P609, and by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416.

References

[1] R. Conrad, *The deaf school child.* London: Harper & Row, 1979.

[2] O. Velehradská and K. Kuchler, "Průzkum čtenářských dovedností na školách pro děti s vadami sluchu," *INFO-Zpravodaj FRPSP*, vol. 6, 1998.

[3] A. Macurová, "Poznáváme český znakový jazyk (úvodní poznámky)," *Speciální pedagogika*, vol. 11, 2001.

[4] Z. Krňoul and M. Železný, "Evaluation of synthesized sign and visual speech by deaf," in *Proceedings of AVSP 2008*, in press, 2008.

[5] L. Namir and M. Schlesinger, *Sign Language of the Deaf.* New York: Academic Press, 1978.

[6] A. Nonhebel, O. Crasborn, and E. van der Kooij, "Sign language transcription conventions for the ECHO project," University of Nijmegen, Tech. Rep. 6, 2003. [Online]. Available: http://www.let.kun.nl/sign-lang/echo/docs/ECHO_transcr_conv.pdf

[7] P. Campr, M. Hrúz, J. Trojanová, and M. Železný, "Collection and preprocessing of czech sign language corpus for sign language recognition," in *LREC 2008*, Workshop proceedings: Construction and Exploitation of Sign Language Corpora. ELRA, 2008.

[8] J. Kennaway, "Experience with and requirements for a gesture description language for synthetic animation," in *Lecture Notes in Artificial Intelligence*, A. Camurri and G. Volpe, Eds. Genova, Italy, 2003, pp. 300–311.

[9] Z. Krňoul, J. Kanis, M. Železný, and L. Müller, "Czech text-to-sign speech synthesizer," *Machine Learning for Multimodal Interaction, Series Lecture Notes in Computer Science*, vol. 4892, pp. 180–191, 2008.

[10] B.-J. Theobald, I. Matthews, and S. Baker, "Evaluating error functions for robust active appearance models," in *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. Southampton, UK, 2006, pp. 149–154.

[11] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *Int. J. Comput. Vision*, vol. 9, no. 2, pp. 137–154, November 1992. [Online]. Available: http://dx.doi.org/10.1007/BF00129684

[12] T. Morita and T. Kanade, "A sequential factorization method for recovering shape and motion from image streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 858–867, 1997.