Weakly Supervised Automatic Transcription of Mouthings for Gloss-Based Sign Language Corpora

Oscar Koller[1,2], Hermann Ney[1] and Richard Bowden[2]

[1] Human Language Technology and Pattern Recognition - RWTH Aachen University, Germany
[2] Centre for Vision Speech and Signal Processing - University of Surrey, Guildford, UK
{koller,ney}@cs.rwth-aachen.de, r.bowden@surrey.ac.uk

Abstract

In this work we propose a method to automatically annotate mouthings in sign language corpora, requiring no more than a simple gloss annotation and a source of weak supervision, such as automatic speech transcripts. For a long time, research on automatic recognition of sign language has focused on the manual components. However, a full understanding of sign language is not possible without exploring its remaining parameters. Mouthings provide important information to disambiguate homophones with respect to the manuals. Nevertheless most corpora for pattern recognition purposes are lacking any mouthing annotations. To our knowledge no previous work exists that automatically annotates mouthings in the context of sign language. Our method produces a frame error rate of 39% for a single signer on the alignment task.

Keywords: Sign Language, Mouthing, Lip Reading, Unsupervised Automatic Annotation

1. Introduction

Sign languages consist of several information streams that convey meaning. Historically, research on automatic recognition of sign language has focused on the manual components of the signs, such as the hand shape, its orientation, position and movement (Starner et al., 1998; Vogler and Metaxas, 2004; Zaki and Shaheen, 2011). These manual parameters are widely considered to contain a large part of the information in sign language. However, it is clear that a full understanding of sign language, particularly with respect to idioms, grammatical structures and also semantics, is not possible without further exploring the remaining information channels, namely facial expressions (mouthing, eye gaze) and upper body posture (head nods/shakes and shoulder orientation). Mouthing can be observed in many European sign languages. Nevertheless, its linguistic status is still debated (Sandler, 2006). However, there is a lot of evidence that mouthings can discriminate homophones with respect to the manual parameters and thus constitute an important feature for automatic recognition of sign language, which has not been exploited in current approaches. This is due to the fact that sign language corpora intended for pattern recognition and machine learning usually do not have any mouthing annotations.

This work aims to automatically annotate mouthings for gloss-based sign language corpora when annotations are not available. The employed corpus is recorded from broadcast news and constitutes a translation from German speech to sign language performed by hearing interpreters. We use the automatic transcriptions of the speech and exploit this as weak supervision through the fact that mouthings in sign language often correspond to parts of orally pronounced words.

In Section 2. related work in viseme recognition and linguistics is shown. In Section 3. we present the corpus and the manual annotation used for evaluation. Section 4. presents the approach. Finally, results are given in Section 5. and Section 6. draws conclusions with future work.

2. Related Work

Two types of mouth actions can be observed in sign languages: mouthings and mouth gestures. While mouthings are silently pronounced parts of spoken words that originate from speech contact, mouth gestures constitute patterns unrelated to spoken language. Mouthings occur often with nouns and with morphologically simple signs (Crasborn et al., 2008). Furthermore, they are often related to lexical items (Sutton-Spence, 2007), while mouth gestures have a morphological role (Horst Ebbinghaus and Jens Hessmann, 2001). The status of mouthings in sign language is highly debated in the linguistic community. Some researchers understand it as part of sign language, while others see it as separate entity. Refer to (Crasborn et al., 2008) for details on this debate. However, in German Sign Language (DGS) mouthings play an important role. DGS contains many signs with identical manual parameters that have related meanings and seem to be only disambiguated by combination with different, though semantically related, mouthings (Horst Ebbinghaus and Jens Hessmann, 1994; Kutscher, 2010). In terms of synthesis, (Kipp et al., 2011) have analysed the perception of sign language avatar systems and found that the absence of mouthings strongly disturbs the Deaf evaluators. Movement of cheeks and lips, but also teeth and tongue were determined crucial for understanding certain mouthings.

In this paper, we deal with signing of sign language interpreters. The question arises, if their mouthings differ from native Deaf mouthings. However, not much literature has systematically researched this question. (Weisenberg, 2009) found that sign language interpreters adjust their mouthing with respect to their target audience. However the study only evaluates four interpreters and the influence of Deaf family members is neglected. In a study comparing three native and two non-native signer, (Lisa Monschein, 2011) reports that the non-native (hearing) signers do not use more mouthings than the native Deaf signers.

Visemes, the visual representations of phonemes in the mouth area, were first mentioned by (Fisher, 1968). Nowa-

days lipreading and viseme recognition is a well established, yet challenging research field in the context of audio-visual speech recognition. The first system was reported by (Petajan, 1984) who distinguished letters from the alphabet and numbers from zero to nine and achieved 20% error rate on that task. Since then, the field has advanced in terms of recognition vocabulary, features and modelling approaches. (Ong and Bowden, 2011) achieved an error rate of 13.2% using sequential patterns for lipreading. A good overview is given in (Potamianos et al., 2003). Previous applications of viseme recognition specifically to automatic sign language recognition are very rare. The state of mouth openness has been used to distinguish signing from silence (Pfister et al., 2013). However, little work has been done in training viseme models in an unsupervised or weakly supervised fashion. Most deal with the problem of clustering visemes in order to find an optimal phoneme to viseme mapping (Luca Cappelletta and Naomi Harte, 2012) and to our knowledge no previous application of dedicated viseme recognition to sign language recognition exists.

3. Corpora

The proposed approach uses the publicly available RWTH-PHOENIX-Weather corpus, containing continuous signing in DGS of 7 hearing interpreters. The corpus consists of 190 TV broadcasts (weather forecast) recorded on public TV. It provides a total of 2137 manual sentence segmentations and 14717 gloss annotations. Glosses constitute an economical way of annotating sign language corpora. They represent an approximate semantic description of a sign, usually annotated w.r.t. the manual components. The same gloss 'MOUNTAIN' denotes the sign alps but also any other mountain, as they share the same hand configuration and differ only in mouthing. Moreover, the RWTH-PHOENIX-Weather corpus contains 22604 automatically transcribed and manually corrected German speech word transcriptions. The boundaries of the signing sentences are matched to the speech sentences. It is worth noting that the sentence structures for spoken German and DGS do not correlate. This is a translation rather than a transcript. Furthermore, it has to be noted that the corpus contains signing of professional hearing interpreters. Some have Deaf family members and grew up with sign language as mother tongue, others did not. As the interpreters translate live, they face very tight time constraints. Due to the direct interpretation task, it can be expected that the interpreter's mouthings are partly closer to speech, than they usually would be. Nevertheless, this remains to be proven.

To evaluate this work, we annotated 3 sentences per signer on the frame level with viseme labels totalling 2082 labelled frames. The annotation was performed three times by a competent non-native signer. While annotating, the annotator had access to the video sequence of signing interpreters showing their whole body (not just the mouth), the gloss annotations and the German speech transcriptions. In each of the three annotation iterations, the frame labels varied slightly due to the the complexity and ambiguity of labelling visemes. See (Yuxuan Lan et al., 2012) for a human evaluation. We consider each annotation to be valid, yielding 1.6 labels per frame (see Table 4).

4. Weakly Supervised Mouthing Alignment

The approach exploits the fact that mouthings are related to spoken language and its words, for which automatic spoken language transcripts are part of the RWTH-PHOENIX-Weather corpus. However, the relation between speech and mouthings is loose and holds for some signs only.

Visual features of the mouth region are extracted. These consist of ten continuous distance measurements around the signers mouth and the average colour intensity of three areas inside the mouth (to capture tongue and teeth presence), as shown in Fig 1. The distance measurements are based on salient point locations on the interpreter's face tracked using the deformable model registration method known as Active-Appearance-Models (AAMs). For details refer to (Schmidt et al., 2013).

The features are clustered using Gaussian clustering and Expectation Maximization (EM) while constraining the sequence of features to the sequence of automatically transcribed German words in a Hidden-Markov-Model (HMM) framework. Thus, we consider the weakly supervised viseme training to be a search problem of finding the sequence of visemes $v_1^Z := v_1, \ldots, v_Z$ belonging to a sequence of mouthings (or silently pronounced partial words) $m_1^N := m_1, \ldots, m_N$, where the sequence of features $x_1^T := x_1, \ldots, x_T$ best matches the viseme models. We maximise the posterior probability $p(v_1^N|x_1^T)$ over all possible viseme sequences for the given sequence of glosses.

$$x_1^T \rightarrow \hat{v}_1^Z(x_1^T) = \arg\max_{v_1^Z} \left\{ p(m_1^N) p(x_1^T|v_1^Z) \right\}, \quad (1)$$

where $p(m_1^N)$ denotes the pronunciation probability for a chosen mouthing. We model each viseme by a 3 state HMM and a garbage model having a single state. The emission probability of a HMM state is represented by a single Gaussian density with a diagonal covariance matrix. The HMM states have a strict left to right structure. Global transition probabilities are used for the visemes. The garbage or 'no-mouthing' model has independent transition probabilities. We initialise the viseme models by linearly partitioning the data.

The given word sequence that stems from the Automatic Speech Recognition (ASR) transcripts is reordered to better match the syntax present in DGS. This is done by aligning the manual gloss annotations and the speech transcripts with the GIZA++ toolkit (Och and Ney, 2003) commonly used in statistical machine translation to align source and target language. Furthermore, a lexicon is built that includes a finite set of possible pronunciations for each German word. This lexicon consists of different phoneme sequences for each word and an entry for 'no-mouthing'. However, the mouthings produced by signers often do not constitute fully pronounced words, but rather discriminative bits of words. Thus, for each full pronunciation we add multiple shorter pronunciations to our lexicon ψ by truncating the word w which consists of a sequence of phonemes $s_1^N = s_1, \ldots, s_N$, such that $\psi = \left\{ w' : s_1^{N-\phi} | \phi \in \{0, \ldots, \phi_{trunc}\} \wedge N - \phi \geq \phi_{min} \right\}$, where we empirically set $\phi_{trunc} = 10$ and $\phi_{min} = 3$.

Finally, to account for the difference in articulatory phonemes and visual visemes, we need to map phonemes

	∑	A	E	F	I	L	O	Q	P	S	U	T	gb	ratio
Signer 1	275	13	25	8	27	8	30	28	19	19	18	54	143	1.43
Signer 2	266	25	40	24	18	8	27	29	18	25	16	58	147	1.64
Signer 3	318	35	18	23	51	15	39	70	34	21	16	83	185	1.86
Signer 4	236	43	35	38	27	8	12	33	14	20	15	46	63	1.50
Signer 5	320	36	32	23	56	8	19	48	22	14	39	44	103	1.39
Signer 6	366	65	39	38	28	6	44	43	28	12	36	98	191	1.72
Signer 7	301	28	21	23	56	18	40	42	32	2	14	79	136	1.63
∑	2082	11.8	10.1	8.5	12.7	3.4	10.1	14.1	8.0	5.3	7.4	22.2	46.5	1.60
ratio	1.60	1.76	1.80	1.78	1.99	2.04	1.79	1.90	1.75	1.88	1.77	1.90	1.43	

Table 1: Frame annotation statistics for each of the employed 11 visemes (Eeva A. Elliott, 2013) on the RWTH-PHOENIX-Weather corpus. The last line shows relative annotation per viseme in [%]. 'gb' denotes frames labelled as non-mouthings/garbage. 'ratio' refers to the average labels per frame, which reflect the uncertainty of the annotator.

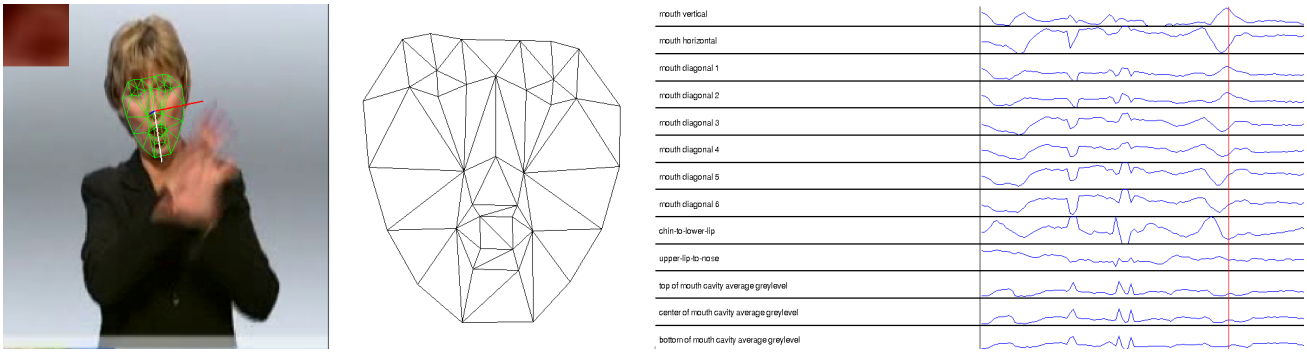

Figure 1: Feature extraction, left: fitted AAM grid and inner mouth cavity patch, center: rotated and normalised AAM grid, right: high-level feature values over time

to visemes. Two different mappings are compared in this work. A mapping to 16 visemes by (Weiss and Aschenberner, 2005) (compare left side of Table 2) and a mapping to 12 visemes by (Eeva A. Elliott, 2013) (see right part of Table 2). Furthermore, we propose a mapping ourselves that considers a many-to-many relationship depending on the context of a viseme, i.e. the preceding and succeeding viseme. The mapping consists of 29 visemes and one 'no-mouthing' entry and is displayed in Table 3. It has been created using a phonetic decision trees (Beulen, 1999). All visemes are clustered based on their feature representation, while considering visual properties (roundness or openness).

5. Results

In the scope of this paper we provide a solution to automatically annotate mouthings in sign language corpora with not more than gloss annotations and speech transcripts as source of weak supervision given. With this in mind, we perform a forced alignment on the RWTH-PHOENIX-Weather data using different phoneme-to-viseme mappings to assess how suitable each is for the task of modelling sign language mouthings. See Figures 3 and 4 for qualitative examples of some alignments on our data.

We can determine the alignment error per frame based on the 2082 manually annotated frames (see Section 3.) for each of the seven signers. We compare the case of not using any viseme mapping and modelling 40 phonemes of the spoken language instead ('Phonemes'), a viseme mapping with 16 visemes by (Weiss and Aschenberner, 2005)

(Weiss and Aschenberner, 2005)		(Eeva A. Elliott, 2013)	
Visemes	Phonemes	Visemes	Phonemes
A	a a~ a:	A	a a~ a:
C	j C	E	e: E E:
E	i: I e: E: E	F	f v
F	f v	I	i: I j
M	m	L	l
N	n l	O	2: 9 o: O
O	o: O	P	b m p
P	p b	Q	6 C g h k N @ R x
Q	@ 6	S	S tS
R	h r x N	T	d n s ts t z
S	s z	U	u: U y: Y
T	t d k g	A I	aI
U	u: U	A U	aU
Y	y: Y 2: 9	O I	OY
Z	S tS	P F	pf
A E	aI		
A U	aU		
O E	OY		
P F	pf		

Table 2: Tested phoneme to viseme mappings in SAMPA.

('Weiss'), a mapping with 12 units by (Eeva A. Elliott, 2013) ('Elliott') and our proposed many-to-many viseme mapping with 30 context dependent visemes ('Proposed'). Results are given in Table 4, with the frame error rate per signer and averaged across the 7 signers given. It has to be noted that depending on the number of visemes, a cer-

	Visemes	SAMPA Phonemes	Context	
			Left	Right
Open/Round	A_1	a a~ a:	#	l
	A_2	a a~ a: aI		
	aU	aU		
	L_1	l		#
	L_2	l		
Semi-Open/Round	S_1	S	#	
	S_2	S tS		
	O	OY 2: O o:		
	U	U u: y:		
	Y	Y	t k	# s 6
	@	@		
Tense	E	e: E i I		
	F	f v pf		not #
Semi-Closed	LT_1	d l t	y:	#
	LT_2	b d l t	y:	not #
	LT_3	R l t ts	f	e: E i I
	CON_1	R d g h k l n t ts z	#	e: E i I
Strong Context	CON_2	6 C R b d f g k m n p s t ts v x z		#
	CON_3	6 N R f g k l m n s t v x z	U u: aU	
	CON_4	6 ts k n t		f v
	$CON_{5\text{-}11}$	different consonants+context		
Closed	M_1	b m p	#	
	M_2	b m p		

Table 3: Proposed many-to-many phoneme to viseme mapping depending on context. '#' refers to word boundaries.

tain error rate can be achieved by guessing a frame's label. In order to appropriately compare the mappings with different numbers of viseme models we define another error rate ('compensated ER') that removes all correct classifications achieved by chance. On average, over all signers 'Elliott' outperforms 'Weiss', which outperforms 'Phonemes' (56.84% to 60.21% to 74.16% respectively). Our proposed mapping lags 3% behind with 59.49%. However, if we consider the 'compensated ER' our proposed mapping outperforms all others by between 4% and 17%. Apart from the averaged results, we note that the alignment error rates differ among all signers. This can be explained by the fact that each signer's mouthing differs slightly. It manifests itself in different sets of preferred visemes by each signer, whereas not all visemes can be equally well modelled. Table 5 shows the alignment statistics of the whole data set using the 'Elliott' viseme mapping. Relative frame alignments per viseme are reported for all 180000 frames present in the data set. This allows us to observe the signers' mouthing preferences. As such, Signer 1 pronounces 'A' and 'O' more frequently than average. Our models represent these two visemes particularly well, which might explain why the viseme alignments for Signer 1 perform better than on other signers. The last line in Table 5 shows empirically determined occurrence frequencies reported in (Eeva A. Elliott, 2013) for reference. We see that 'T' and 'Q' are as reported (as well as in our paper) the two most frequently occurring visemes. The same similarity holds for the least frequently occurring viseme 'S'. On average our method aligns 44% of all frames to no-mouthings, which have been excluded

	'Phonemes'	'Weiss'	'Elliott'	'Proposed'
Signer 1	74.45	39.66	39.09	49.57
Signer 2	77.25	59.87	57.96	63.8
Signer 3	82.80	76.82	69.14	68.39
Signer 4	61.17	54.29	40.66	40.74
Signer 5	73.83	58.43	55.68	56.6
Signer 6	71.51	63.05	62.12	68.65
Signer 7	74.17	62.96	60.26	60.51
Total	74.16	60.21	56.84	59.49
#visemes	40	16	12	30
chance ER	96.12	90.36	87.5	94.18
compensated ER	78.03	69.85	82.87	65.31

Table 4: Frame error rates (ER) per signer in [%] for no viseme mapping ('phonemes'), a mapping by (Weiss and Aschenberner, 2005) ('Weiss'), (Eeva A. Elliott, 2013) ('Elliott') and our proposed mapping ('Proposed'). Lower is better.

in the linguistic reference. All viseme alignments seem to roughly correspond to the linguistic reference, however, we note that viseme 'Q' is only aligned 15.50%, whereas Elliott reports over 25%.

Figure 2 shows the top 15 glosses with the most frequently aligned mouthings in the corpus. We see that sensible mouthings have been chosen by our proposed weakly supervised alignment scheme. Furthermore it is shown that the approach is able to spot the different mouthings that specify signs with the same manuals and thus the same gloss annotation but with different mouthings. For example, the gloss REGEN (RAIN) has been found to occur with mouthings /R e g/ (rain) and /S aU 6/ (shower). Moreover, it is apparent that the weak supervision allows to spot mouthings that only share a semantic relation to the employed gloss, but actually constitute different words. Such an example is the mouthing /g R a t/ (degree) belonging to the gloss TEMPERATUR (TEMPERATURE), which represents an information stemming from the audio transcripts.

By showing the most commonly aligned mouthings, Figure 2 also contains information about pronunciation reductions. The type of reduction that we allowed (see Section 4.), was truncating the ends of the pronunciations. We see that apparently shorter pronunciations are preferred, as most of the most frequently aligned viseme sequences in the red bars consist of only 3 visemes (e.g./R e g/, /m O 6/, /n a x/). This coincides with the expectation that mouthings in sign language are more context cues than full silently pronounced words. Among the displayed mouthings there is only one that is very unlikely to actually have occurred (AUCH: /d a b/), which most likely constitutes noise injected during the statistical reordering process. In terms of word types, the result follows linguistic findings that mouthings mainly occur with nouns, as 13 of the 15 glosses are nouns.

6. Conclusions

In this paper we show how to automatically annotate mouthings in sign language corpora with no more than gloss annotations needed and speech transcripts as source of weak supervision. We further compare the impact of

	frames	A	E	F	I	L	O	Q	P	S	U	T	gb
Signer 1	49753	6.07	4.74	3.76	5.74	1.97	5.11	9.33	3.09	0.97	2.83	12.38	44.01
Signer 2	7399	6.27	3.24	2.34	3.89	1.50	4.05	7.95	4.58	1.20	5.84	11.83	47.30
Signer 3	27381	3.42	6.32	3.82	6.52	1.50	4.24	8.21	3.76	1.41	3.02	13.02	44.77
Signer 4	33394	4.60	4.91	3.04	3.94	1.34	4.22	6.45	3.08	0.92	2.34	8.64	56.50
Signer 5	41845	5.34	7.99	4.68	7.10	2.54	4.70	10.42	4.57	1.13	4.84	14.72	31.97
Signer 6	9841	4.88	3.94	4.16	4.89	2.93	4.55	9.14	4.89	1.06	8.45	11.77	39.36
Signer 7	19750	4.38	2.62	3.89	4.81	2.14	4.66	7.30	4.85	1.00	3.44	9.40	51.51
$\sum$	189363	5.04	5.39	3.82	5.62	1.97	4.62	8.63	3.85	1.08	3.69	11.96	44.33
$\sum$ no gb	105414	9.05	9.69	6.87	10.10	3.53	8.30	15.50	6.91	1.93	6.63	21.49	-
comparison	-	8.57	5.05	4.59	8.18	4.97	3.83	25.66	6.79	2.60	5.31	24.39	-

Table 5: Frame alignment statistics in [%] for each of the employed 11 visemes on the RWTH-PHOENIX-Weather corpus.'gb' denotes non-mouthings/garbage. The last line shows comparative statistics from (Eeva A. Elliott, 2013).

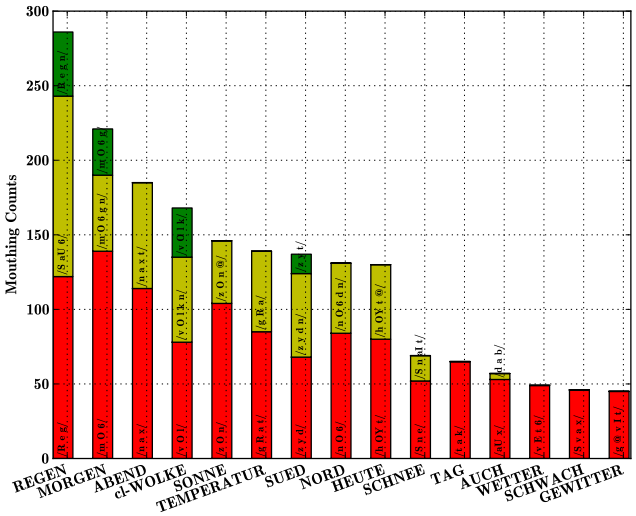

Figure 2: Top 15 glosses with the most frequent occurring mouthings shown in SAMPA annotation on the bars. Any mouthings occurring less than 20% w.r.t. all mouthings of a gloss have been filtered out for better readability.

four different schemes to map phonemes to visemes and find that a many to many mapping that relies on visemic context is best if one takes into account the complexity of the classification.

We achieve a frame error rate of 39.09% in the alignment task for a specific signer and 56.84% averaged over all signers. Furthermore, we show that our proposed method yields alignment statistics comparable to those in the linguistic literature. Finally, the mouthings are shown to further disambiguate gloss transcriptions of a sign. As expected, the mouthings represent reduced forms of German words.

In terms of future work, we plan to apply our method to native Deaf signing to separate influence from the German to DGS interpretation task and to include it into a sign language recognition pipeline. Furthermore, there is a need to find features that better represent tongue and inner mouth and modelling of mouth gestures remains untouched.

7. References

Klaus Beulen. 1999. *Phonetische Entscheidungsbäume für die automatische Spracherkennung mit großem Vokabular*. Mainz.

Onno Crasborn, Els Van Der Kooij, Dafydd Waters, Bencie Woll, and Johanna Mesch. 2008. Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language & Linguistics*, 11(1).

Eeva A. Elliott. 2013. *Phonological Functions of Facial Movements: Evidence from deaf users of German Sign Language*. Thesis, Freie Universität, Berlin, Germany.

Cletus G. Fisher. 1968. Confusions among visually perceived consonants. *Journal of Speech, Language and Hearing Research*, 11(4):796.

Horst Ebbinghaus and Jens Hessmann. 1994. German words in german sign language: Do they tell us something new about sign languages? In Carol Erting, editor, *The Deaf Way: Perspectives from the International Conference on Deaf Culture*. Gallaudet University Press.

Horst Ebbinghaus and Jens Hessmann. 2001. Sign language as multidimensional communication - or: Why manual signs, mouthings, and mouth gestures are three different things. In P. Boyes Braem and R. L. Sutton-Spence, editors, *The Hands Are The Head of The Mouth. The Mouth as Articulator in Sign Languages*, pages 133–153. Signum Press, Hamburg.

Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *Proceedings of ACM Conference on Computers and Accessibility*, ASSETS '11, page 107–114, New York, NY, USA. ACM.

Silvia Kutscher. 2010. Ikonizität und indexikalität im gebärdensprachlichen lexikon – zur typologie sprachlicher zeichen. *Zeitschrift für Sprachwissenschaft*, 29(1), January.

Lisa Monschein. 2011. Empirical research on mouth patterns considering sociolinguistic factors: A comparison between the use of mouth patterns of deaf l1- and hearing l2-users of german sign language (DGS), August.

Luca Cappelletta and Naomi Harte. 2012. Phoneme-to-viseme mapping for visual speech recognition. In *ICPRAM*, page 322–329.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Eng-Jon Ong and Richard Bowden. 2011. Learning sequential patterns for lipreading. In *Proceedings of the British Machine Vision Conference*, page 55.1–55.10. BMVA Press.

Eric David Petajan. 1984. *Automatic Lipreading to En-*

Figure 3: Example showing the frame alignment of signer 1 following the phoneme to viseme mapping from (Eeva A. Elliott, 2013). Original gloss annotation: PLUS EINS BIS SIEBEN TEMPERATUR. Audio transcription: Und das ganze dann bei plus ein und sieben Grad.

Figure 4: Example showing the frame alignment of signer 5 following the phoneme to viseme mapping from (Eeva A. Elliott, 2013). Original gloss annotation: MONTAG WAHRSCHEINLICH SONNE cl-WOLKE. Audio transcription: Am Montag mal Sonne, mal Wolken.

hance Speech Recognition (Speech Reading). Ph.D. thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA. AAI8502266.

Tomas Pfister, James Charles, and Andrew Zisserman. 2013. Large-scale learning of sign language by watching TV (using co-occurrences). In *Proceedings of the British machine vision conference*, U. K. Leeds.

G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior. 2003. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, September.

Wendy Sandler. 2006. *Sign Language and Linguistic Universals*. Cambridge University Press, February.

Christoph Schmidt, Oscar Koller, Hermann Ney, Thomas Hoyoux, and Justus Piater. 2013. Enhancing gloss-based corpora with facial features using active appearance models. In *International Symposium on Sign Language Translation and Avatar Technology*, volume 2, Chicago, IL, USA.

Thad Starner, Joshua Weaver, and Alex Pentland. 1998. Real-time american sign language recognition using desk and wearable computer based video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12):1371–1375.

Rachel Sutton-Spence. 2007. Mouthings and simultaneity in british sign language. In Myriam Vermeerbergen, Lorraine Leeson, and Onno Alex Crasborn, editors, *Simultaneity in Signed Languages: Form and Function*, page 147. John Benjamins Publishing.

Christian Vogler and Dimitris Metaxas. 2004. Handshapes and movements: Multiple-channel ASL recognition. In *Lecture Notes in Computer Science*, page 247–258. Springer.

Julia Weisenberg. 2009. *Audience effects in American Sign Language interpretation*. Ph.D. thesis, State University of New York at Stony Brook.

Christian Weiss and Bianca Aschenberner. 2005. A german viseme-set for automatic transcription of input text used for audio-visual speech synthesis. In *Proc. Interspeech*, pages 2945–2948, Lisbon, Portugal.

Yuxuan Lan, Richard Harvey, and Barry-John Theobald. 2012. Insights into machine lip reading. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4825–4828, March.

Mahmoud M. Zaki and Samir I. Shaheen. 2011. Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters*, 32(4):572–577.