Deep Learning of Mouth Shapes for Sign Language

Oscar Koller, Hermann Ney
Human Language Technology & Pattern Recog.
RWTH Aachen University, Germany
{koller,ney}@cs.rwth-aachen.de

Richard Bowden
Centre for Vision Speech & Signal Processing
University of Surrey, UK
r.bowden@surrey.ac.uk

This paper deals with robust modelling of mouth shapes in the context of sign language recognition using deep convolutional neural networks. Sign language mouth shapes are difficult to annotate and thus hardly any publicly available annotations exist. As such, this work exploits related information sources as weak supervision. Humans mainly look at the face during sign language communication, where mouth shapes play an important role and constitute natural patterns with large variability. However, most scientific research on sign language recognition still disregards the face. Hardly any works explicitly focus on mouth shapes. This paper presents our advances in the field of sign language recognition. We contribute in following areas: We present a scheme to learn a convolutional neural network in a weakly supervised fashion without explicit frame labels. We propose a way to incorporate neural network classifier outputs into a HMM approach. Finally, we achieve a significant improvement in classification performance of mouth shapes over the current state of the art.

1. Introduction

Automatic Sign Language Recognition (ASLR) helps to bridge the communication gap between deaf and hearing people who are not able to communicate in the same language. It is well known that sign languages (SLs) make use of the hands but also of the face to convey information [23]. Recently, among SL linguists, there has been increasing interest in so-called mouthings, *i.e*. sequences of mouth shapes or visemes visible on the signer's face, that convey a substantial part of signs' meaning [2, 24, 22]. However, in ASLR mouth shapes have not received much attention, despite their promise to distinguish semantically close signs that otherwise share the same hand shapes and movements.

The aim of this work is to learn a classifier that can recognise mouth shapes as part of sign language. These mouth shapes, linguistically termed as mouthings, accompany the manual parts of a sign (*e.g*. the hand movement and shape) and are often related to the corresponding spoken word. The visible mouth shapes can add information to the meaning of a sign and make it distinguishable from a semantically related sign. But in other cases, mouthings simply provide additional information that is redundant to the hand's information channel and not explicitly needed to understand the sign.

Mouth shapes are difficult to model and to recognise as their presence seems to be based on personal and situational preferences of the signers. Currently, no publicly available data set exist that contain manual mouthing annotations that would allow for supervised training. The reason lies in the difficulty to annotate mouth shapes. Sometimes, they represent the mouth pattern of shortened parts of a related spoken word, sometimes the whole spoken word or sometimes no mouth shape at all may be visible.

This paper will therefore explore a way to train mouth shapes with additional information sources used as weak supervision. Todate, sign linguists have not found any grammatical rules that constrain the usage of mouthings within sign languages, turning it into a challenging test bed for algorithms that are able to learn without explicit annotations.

Our approach relies on spoken language transcripts as a source of weak supervision. The data we are dealing with features natural signing that originates from interpreters making the daily broadcast news accessible to the Deaf. The signing represents the interpreted spoken German news and thus, we have access to the spoken German's news transcripts. However, the reader has to note that the transcripts (being in spoken German) and the sign language footage (being in German Sign Language) constitute two different languages with different vocabulary, syntax and word order. We will accommodate this fact in the way we design our learning algorithm.

This paper presents contributions in following areas: We present a scheme to learn a Convolutional Neural Network (CNN) in a weakly supervised fashion without explicit frame labels. We demonstrate how to incorporate CNN classifier outputs into an HMM approach allowing forced temporal alignment and iterative learning of CNNs on video. By doing so, we achieve a significant improvement in classification performance over the state of the art.

This paper is organised as follows: In the next section

a brief overview of the state of the art is given. Section 3 describes the approach presented in this paper. In Section 4 the data set and statistical details are provided, while Section 5 gives experimental validation and discussion. Finally the paper closes with the conclusions that can be drawn.

2. Related Work

In 1968 Fisher [7] was the first to mention differences between spoken phonemes and corresponding visemes as mouth shapes. Nowadays, 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 in 1984 by Petajan [20] who distinguished letters of the alphabet and numbers from zero to nine and achieved 80% accuracy on that task. Since then the field has advanced in terms of recognition vocabulary, features and modelling approaches [28, 16, 29, 18].

Facial expression recognition is another related field for algorithms that relate to mouth shapes. For instance Tian [26] recognises action units and models the mouth with only three different states: open, closed, very closed.

When it comes to sign language recognition, the mouth has mostly been disregarded in recognition pipelines. But recently, the community has started developing some interest in this area [1]. Recovering the mouth shapes using a distance measure based on tracked Active Appearance Model (AAM) landmarks with an iterative Expectation Maximization (EM) learning has been proposed by [14], whereas other work focuses on automatically annotating mouth shapes [15]. Pfister *et al*. [21] employ lip-movement to distinguish signing from silence by inferring the state of mouth openness. This is used to reduce the candidate sequences in multiple instance learning, which is also supported by a single SIFT descriptor of the mouth region.

In the general scope of automatic sign language recognition several works exist that exploit weak supervision to learn hand-based sign models [4, 12, 5, 13]. Facial features have also been used before. Michael *et al*. [17] employs spatial pyramids of Histogram of Oriented Graphs (HOG) and SIFT features together with 3D head pose and its first order derivative to distinguish three grammatical functions trained on isolated American Sign Language (ASL) data of three signers. Vogler and Goldstein [27] present a facial tracker specifically for ASL. Ong *et al*. [19] use boosted hierarchical sequential pattern trees to combine partly parallel, not perfectly synchronous features through feature selection by the trees.

3. Method

This work provides a solution to the problem of learning a CNN model on sequence data without having manual class annotations for supervised training available. The underlying idea is to iteratively 1) learn a mouth shape CNN model and 2) to find the most likely frame-model-state-alignment in a Hidden-Markov-Model (HMM) framework, while constraining the overall sequence of mouth shapes to concatenations of valid mouth shape patterns defined by a mouth shape lexicon. This lexicon constitutes the source of weak supervision applied to solve the problem jointly with a language model providing the order of signs following each other.

Thus, we consider the weakly supervised mouth shape training to be a search problem of finding the sequence of mouth shapes $v_1^Z := v_1, \ldots, v_Z$ belonging to a sequence of silently pronounced 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 mouth shape models. We maximise the posterior probability $p(v_1^N | x_1^T)$ over all possible viseme sequences for the given sequence of glosses.

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

where $p(m_1^N)$ denotes the prior probability defined by the pronunciation lexicon.

In a first step 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 model has independent transition probabilities. We initialise the viseme models by linearly partitioning the video frames (flat start). Following [14] we use distance measurements of facial landmarks based on signer-dependent AAMs. We then use the EM algorithm 1) to iteratively estimate the best alignment based on the current models and 2) to accumulate updated viseme models.

After the alignment has converged, we use it as class labels to learn a new CNN model. To overcome the signer-dependency of the AAMs and the requirement of manually annotated landmarks to train them, the aim and the contribution of this work is to replace the AAM feature extraction and the Gaussian Mixture Models (GMMs) by learnt convolutional Deep Neural Networks (DNNs).

3.1. Convolutional Neural Network Architecture

Knowing that we can only solve our task in a weakly supervised fashion, where the actual training samples constitute noisy samples of each class, we base our work on a CNN model previously trained in a supervised fashion for the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2014 . We choose a 22 layers deep network architecture following [25] which achieved a top-1 accuracy of 68.7% and a top-5 accuracy 88.9% in the ILSVRC. The

network involves an inception architecture, which helps to reduce the numbers of free parameters while allowing for a very deep structure. Our model has about 6 million free parameters. All convolutional layers and the last fully connected layer use rectified linear units as non-linearity. Additionally, a dropout layer with 70% ratio of dropouts is used to prevent overfitting. We base our CNN implementation on [11], which is an efficient C++ implementation using the NVIDIA CUDA Deep Neural Network GPU-accellerated library.

We replace the pretrained output layer with a 40 dimensional fully connected layer. As a preprocessing step, we apply a global mean normalisation to the images prior to fine-tuning the CNN model with Stochastic Gradient Descent (SGD) and a softmax based cross-entropy classification loss E

$$E = -\frac{1}{N}\sum_{n=1}^{N} log(p(v|x_n)). \quad (2)$$

The DNN learns the output probabilities in terms of 40 posteriors for each of the originating phoneme classes plus a garbage model. The overall learning rate, the learning rates across different parts of the network and the shuffling of the input data are studied in Section 5. Based on the validation data, the best performing training iteration is chosen.

3.2. Sequential HMM-Decoding

After a successful CNN training, the model's softmax outputs $p(v|x_n)$ are used in a HMM framework to add temporal information and re-align the data in order to measure the final alignment error. We decided not to retrain the GMMs with features from the CNN in the so-called tandem approach [10]. We opted for the cleaner hybrid approach [3] known from Automatic Speech Recognition (ASR) for this procedure. The latter usually shows equal or superior performance in comparative automatic speech or handwriting recognition experiments, is faster as it does not require retraining of Gaussian mixtures and is thus more appropriate to evaluate the direct impact of the CNN [9].

To successfully decode a mouth shape sequence, we perform a maximum likelihood search on the visual model as given in Equation 1 and therefore convert the CNN's posterior output to likelihoods using the Bayes' rule as follows:

$$p(x_n|v) \propto p(v|x_n)/p(c)^{\alpha}. \quad (3)$$

This allows us to add prior knowledge from the pronunciation lexicon and language model. Equation 1 then becomes

$$\arg\max_{v_1^Z} \{p(m_1^N)\frac{p(v_1^Z|x_1^Z)}{p(c)^{\alpha}}\}, \quad (4)$$

Figure 1. Example of cropped and up-scaled images used as input for the DNN.

where $p(v_1^Z|x_1^Z)$ constitutes the mouth shape probabilities given by the CNN and $p(c)$ represents the class counts used to train the CNN. The scaling factor α is empirically set to 0.3 in our experiments.

4. Data Set

The proposed approach uses the publicly available RWTH-PHOENIX-Weather corpus [8], which contains 7 hearing interpreter's continuous signing in German Sign Language (DGS). The corpus consists of a total of 190 TV broadcasts of weather forecast recorded on German public TV. It provides a total of 2137 manual sentence segmentations and 14717 gloss annotations, summing up to 189.363 image frames. Glosses constitute a less labour intense way of annotating sign language corpora. They can be seen as an approximate semantic description of a sign, usually annotated w.r.t. the manual components (*i.e.* the hand shape, orientation, movement and position), neglecting many details. For instance, 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.

The interpreters' original images containing the whole upper body and measuring 210x260 pixels are cropped based on the smallest crop covering all AAM tracked points (being the whole face) and scaled to 227x227 pixels. This is shown in Fig. 1. To add more variation to the training data we further crop the images randomly to 224x224 pixels (central cropping to the same dimension for test data).

4.1. Manual Ground Truth

The ground truth, which is made available by the authors of [14], constitutes annotations of 5 sentences per signer on the frame level with 39 phoneme labels (using the SAMPA phoneme inventory) and one garbage label (used when encountered unclear or non-mouth shapes). The annotations cover a total of 3687 video frames. As stated by [14], each frame may contain more than a single label (being a total of

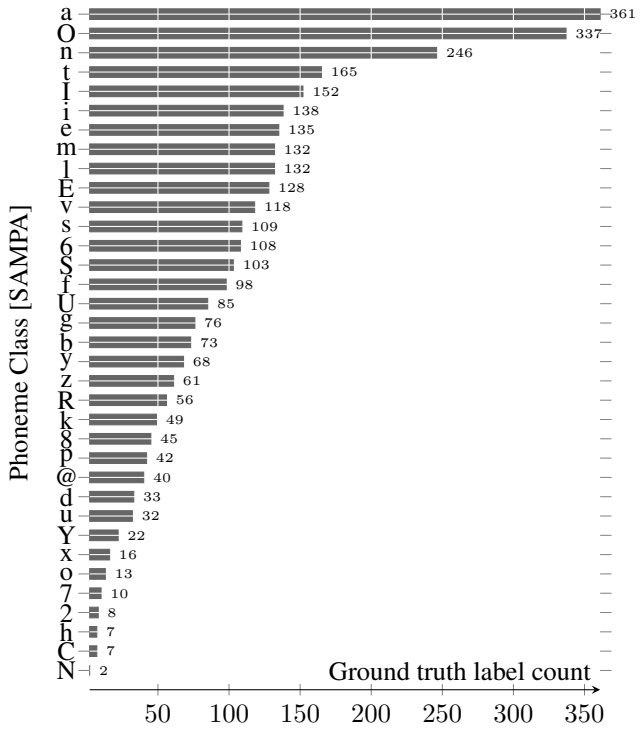

Figure 2. Ground truth label count, including only clear frames.

Visemes	Phonemes
A	a a~ a:
E	e: E E:
F	f v
I	i: I j
L	l
O	2: 9 o: O
P	b m p
Q	6 C g h k N @ R x
S	S tS
T	d n s ts t z
U	u: U y: Y
A I	aI
A U	aU
O I	OY
P F	pf

Table 1. Phoneme to viseme mappings in SAMPA notation, Elliott [6]. This mapping is used for evaluation in the experimental results in Section 5.

	1) Alignment		2) Classification	
	precision	recall	precision	recall
AAMs [14]	55.1	36.9	47.1	48.2
our CNNs	64.5	44.2	55.7	55.6

Table 2. Results for the alignment task and the single frame classification task compared to state of the art in [%].

6226 labels for all frames, covering 36 of all 40 phoneme classes applicable to the training data). About 56% of all annotated frames (2071 frames with a total of 3207 labels) contain clear mouth shapes, where none of the given labels contains a garbage tag. The frequency of the ground truth labels is summarised in Fig. 2, where only clear mouth shapes have been considered. Fig. 3 shows some examples of the manual labels. The reader may note the large intra-class variability paired with a high inter-class similarity, which are due to co-articulation effects and the fact that the human annotators had access to the whole signed sequences rather than isolated images during annotation. To have a very strict evaluation criterion, we evaluate our approach both on the whole sequence and in a per frame fashion (see Section 5).

5. Experiments

In this section, we present our experimental results. We evaluate our approach on the RWTH-PHOENIX-Weather data set [8]. We compare ourselves to previously established state-of-the art results from [14]. The task involves two different sub tasks: 1) An alignment-task: predicting a sequence of mouth shapes given a language model that provides knowledge about the sequence of signs and 2) A classification-task: predicting a single image's class label without any additional information resources, such as a grammar or any sequence constraints. We provide evaluation results on the 40 modelled phoneme classes and test the final classifier on 12 reduced viseme (mouth shape) classes after applying the mapping presented in Tab: 1, which makes results comparable to [14].

As shown in Table 2, we are able to outperform the state of the art results both in the alignment task (Task 1) and in the single frame classification task (Task 2) by around 8% absolute.

In Table 5 a deeper insight in the class confusion of our classifier is possible. It can be seen that the vowel classes "A", "O" and "U" achieve above 60% precision. Which is not surprising, as the underlying phonemes tend to produce more visible and more easily distinguishable mouth shapes. "S" (being spread out lips coming from the /sh/ sound) is partially confused with "U", which clearly share a lot of visual similarity. "P" (a closed mouth) is often confused with the background class ("GB"), which may be explained due to the fact that non-mouth-patterns often involve a closed mouth, as the signer does not make use of it during these periods. Furthermore, class "I" and "E" are mutually confused, as they also share visually similar characteristics.

A	E	F	I	L	O	P	Q	S	T	U	GB
64.4	5.3	3.5	1.2	6.7	2.1	2.5	7.4	0.0	3.4	0.0	3.6
6.8	42.0	4.3	34.5	0.0	0.5	0.2	12.3	0.0	2.2	0.0	7.1
0.2	0.5	32.2	0.0	0.0	1.0	2.4	1.7	1.8	3.2	0.0	10.7
1.1	19.8	5.1	28.6	3.3	2.4	1.3	5.4	0.6	6.8	0.0	14.3
8.5	1.5	0.2	0.0	33.3	1.0	0.3	3.1	0.0	1.9	0.0	0.0
4.1	0.0	0.9	0.0	22.2	60.2	3.8	13.4	1.2	3.9	6.8	0.0
0.9	0.0	1.8	0.0	0.0	0.8	51.8	1.7	0.0	2.2	9.9	3.6
5.4	11.4	4.7	4.2	0.0	3.8	2.9	22.2	0.6	7.2	0.5	3.6
1.2	0.0	1.1	0.0	12.2	6.6	0.1	0.8	50.0	2.2	0.7	0.0
2.2	9.9	11.6	10.1	6.7	7.7	2.8	1.1	9.7	37.0	2.8	7.1
1.2	0.0	0.2	0.0	0.0	8.3	1.8	13.7	16.1	3.8	60.6	0.0
4.1	9.5	34.6	21.4	15.6	5.4	30.2	17.1	20.0	26.0	18.8	50.0

Table 3. Class confusion in [%], showing the per class precision on the diagonal. On the y-axis are the true classes, whereas on the x-axis are the predicted classes. "GB" corresponds to the garbage class.

5.1. Learning Rate

Figure 4 shows the impact of the most important hyper parameter, the overall learning rate, which is kept stable during the DNN training iterations. We notice that a higher learning rate yields more instability during training, but is also capable of learning a stronger classifier. Moreover, we note that the DNN's evaluation accuracy on the 40 phoneme classes increases and then worsens after reaching a maximum. This effect, possibly due to increased influence of falsely aligned training samples and over fitting, makes a careful selection of the number of training epochs necessary.

The employed CNN-architecture is based on a pretrained net, as stated in Section 3.1. Originally, it has been trained with a large amount of out-of-domain supervised data. Under the point of view that we are fine-tuning these learnt parameters to match a different task with different data, the question arises if the previously learnt parameters should be altered at all (learning just the new fully connected output layer, which starts off a random initialisation) or if everything needs adaptation. We want to answer this question by attributing different weights to the learning rate across different layers in our net. We compare an equal learning rate across all layers with emphasising the last layer's learning rate to just learning the last layer, while keeping all other layers constant. The learning rate has been optimised for each of the experiments separately. In Figure 5 we see that an equal learning rate across all layers yields an accuracy increase of about 1% to 2% absolute compared to weighting the last layer with factor 10 or factor 50. Exclusively fine-tuning the last layer and keeping all pretrained parameters fixed is about 8% worse. These findings suggest that it is crucial to adapt the weights of all layers similarly. This may be mainly related to the very different task the initialising network was trained on (ILSVRC vs. mouth shape images).

Moreover, Figure 5 shows a more thorough look at the training behaviour, as we performed 4 measurements per epoch (as opposed to Figure 4, where we just measure the accuracy once per epoch). We notice frequent variation of classification accuracy around a mean value. It seems that parts of the training data contribute negatively while others improve the evaluation performance, which may be attributed to large visual dissimilarity of different signers or wrongly aligned frames.

5.2. Impact of Shuffling

In Figure 6 we analyse the impact of shuffling the training data. Two main points can be observed: 1) if the data is unshuffled we see a regular oscillation of the evaluation accuracy. This backs the hypothesis that parts of the data (maybe even some specific signers) influence the training negatively. 2) the unshuffled training takes much longer to convergence and does not reach the same accuracy as the shuffled data.

6. Conclusion

This work shows a promising way to model mouth shapes with convolutional DNNs without having explicitly labelled data available. We search for the most likely sequence of mouth shapes using loosely related information as guidance in a weakly supervised fashion. The approach relies on sequence data in which the target classes occur multiple times in varying contexts. The image-class alignment is then used to fine-tune a CNN that has been pretrained on out-of-task data. Therefore, just the new output layer starts with a random initialisation.

We propose a method to include the CNN's output into a HMM framework that outperforms state of the art results of mouth shape classification in the context of sign language recognition by an absolute improvement of around 8%. Besides the higher accuracy, the new approach does not require any feature preprocessing, such as expensive signer-dependent AAMs, but it recognises mouthings directly from a single image.

We analyse the impact of the learning rate on different layers of the pretrained net, finding it benificial to set it equal on both the pretrained and the new output layer. Furthermore, we experimentally confirm the need for shuffling the training samples.

In terms of future work, it seems interesting to investigate the cause for the variability in evaluation accuracy during the CNN training with the weakly supervised frame labels. Furthermore, it seems promising to look at the loss function to make the learning more robust to the outliers inherent in training with weak supervision.

References

[1] E. Antonakos, A. Roussos, and S. Zafeiriou. A survey on mouth modeling and analysis for sign language recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–7, Lubljana, Slovenia, May 2015. 2

[2] R. Bank, O. Crasborn, and R. van Hout. Variation in mouth actions with manual signs in sign language of the netherlands (NGT). *Sign Language & Linguistics*, 14(2), 2011. 1

[3] H. A. Bourlard and N. Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media, 2012. 3

[4] P. Buehler, M. Everingham, and A. Zisserman. Employing signed TV broadcasts for automated learning of british sign language. In *Proceedings of 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 22–23, 2010. 2

[5] H. Cooper, B. Holt, and R. Bowden. Sign language recognition. In T. B. Moeslund, A. Hilton, V. Krüger, and L. Sigal, editors, *Visual Analysis of Humans*, pages 539–562. Springer London, Jan. 2011. 2

[6] Eeva A. Elliott. *Phonological Functions of Facial Movements: Evidence from deaf users of German Sign Language*. Thesis, Freie Universität, Berlin, Germany, 2013. 4

[7] C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech, Language and Hearing Research*, 11(4):796, 1968. 2

[8] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney. RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus. In *International Conference on Language Resources and Evaluation*, pages 3785–3789, Istanbul, Turkey, 2012. 3, 4

[9] P. Golik, P. Doetsch, and H. Ney. Cross-entropy vs. squared error training: a theoretical and experimental comparison. In *INTERSPEECH*, pages 1756–1760, 2013. 3

[10] H. Hermansky, D. W. Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional HMM systems. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, pages 1635–1638. IEEE, 2000. 3

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 3

[12] D. Kelly, J. McDonald, and C. Markham. Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 41(2):526–541, Apr. 2011. 2

[13] O. Koller, H. Ney, and R. Bowden. May the force be with you: Force-aligned SignWriting for automatic subunit annotation of corpora. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–6, Shanghai, PRC, Apr. 2013. 2

[14] O. Koller, H. Ney, and R. Bowden. Read my lips: Continuous signer independent weakly supervised viseme recognition. In *Proceedings of the 13th European Conference on Computer Vision*, Zurich, Switzerland, Sept. 2014. 2, 3, 4

[15] O. Koller, H. Ney, and R. Bowden. Weakly supervised automatic transcription of mouthings for gloss-based sign language corpora. In *LREC Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, pages 98–94, Reykjavik, Iceland, May 2014. 2

[16] Y. Lan, B.-J. Theobald, R. Harvey, E.-J. Ong, and R. Bowden. Improving visual features for lip-reading. In *Proceedings of International Conference on Auditory-Visual Speech Processing*, volume 201, 2010. 2

[17] N. Michael, C. Neidle, and D. Metaxas. Computer-based recognition of facial expressions in ASL: from face tracking to linguistic interpretation. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, LREC, Malta*, 2010. 2

[18] E.-J. Ong and R. Bowden. Learning sequential patterns for lipreading. In *Proceedings of the British Machine Vision Conference*, pages 55.1–55.10. BMVA Press, 2011. 2

[19] E.-J. Ong, O. Koller, N. Pugeault, and R. Bowden. Sign spotting using hierarchical sequential patterns with temporal intervals. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1931–1938, Columbus, OH, USA, June 2014. 2

[20] E. D. Petajan. *Automatic Lipreading to Enhance Speech Recognition (Speech Reading)*. PhD thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 1984. 2

[21] T. Pfister, J. Charles, and A. Zisserman. Large-scale learning of sign language by watching TV (using co-occurrences). In *Proceedings of the British machine vision conference*, U. K. Leeds, 2013. 2

[22] A. Schembri. Mouth gestures in british sign language (BSL): A case study of tongue protrusion in BSL narratives, 2011. 1

[23] W. C. Stokoe. *American Sign Language Structure*. Silver Spring, Md.: Linstok Press, 1960. 1

[24] R. Sutton-Spence. Mouthings and simultaneity in british sign language. In M. Vermeerbergen, L. Leeson, and O. A. Crasborn, editors, *Simultaneity in Signed Languages: Form and Function*, page 147. John Benjamins Publishing, 2007. 1

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv:1409.4842 [cs]*, Sept. 2014. 2

[26] Y.-L. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, Feb. 2001. 2

[27] C. Vogler and S. Goldenstein. Facial movement analysis in ASL. *Universal Access in the Information Society*, 6(4):363–374, 2008. 2

[28] G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254–1265, Nov. 2009. 2

[29] Z. Zhou, G. Zhao, and M. Pietikainen. Towards a practical lipreading system. In *Computer Vision and Pattern Recognition*, pages 137–144, 2011. 2

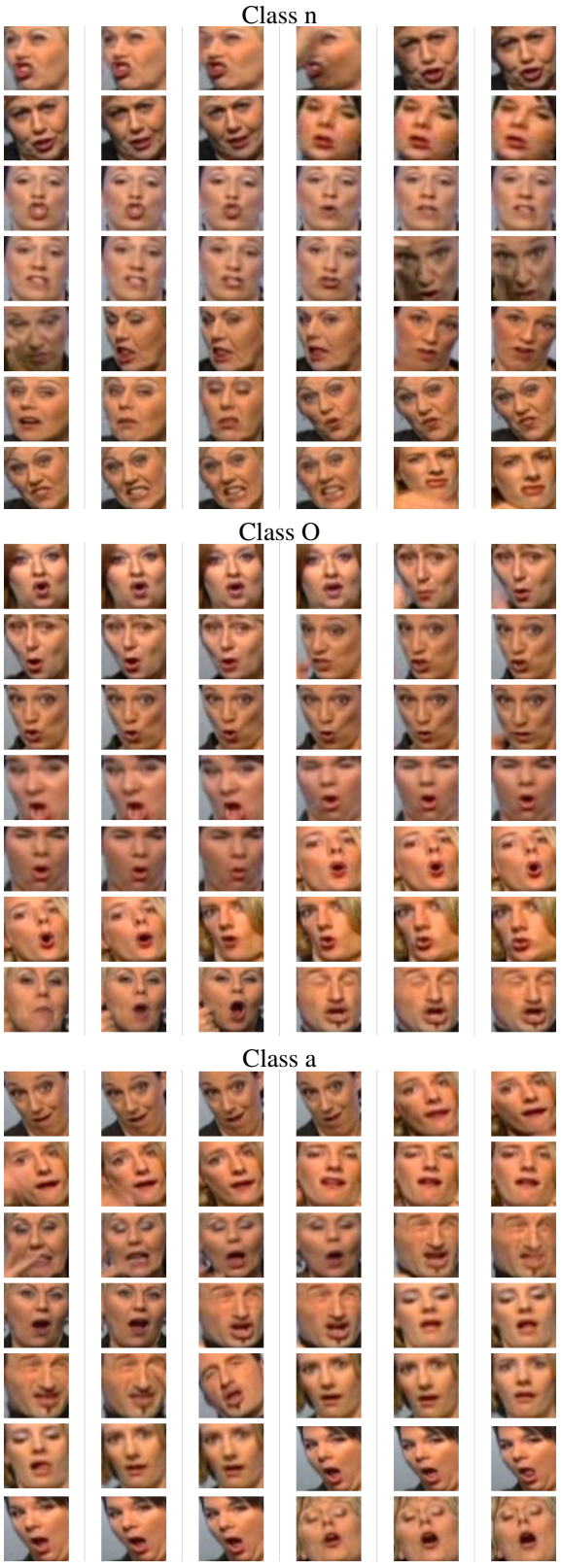

Figure 3. Examples of the three most frequent ground truth classes, showing the intra-class variability.

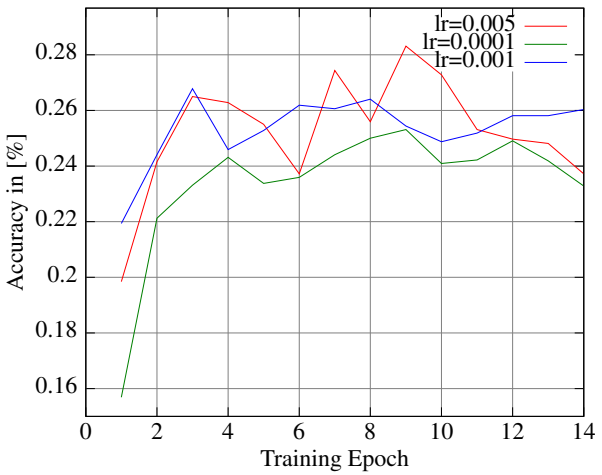

Figure 4. Different learning rates, evaluated on 40 classes.

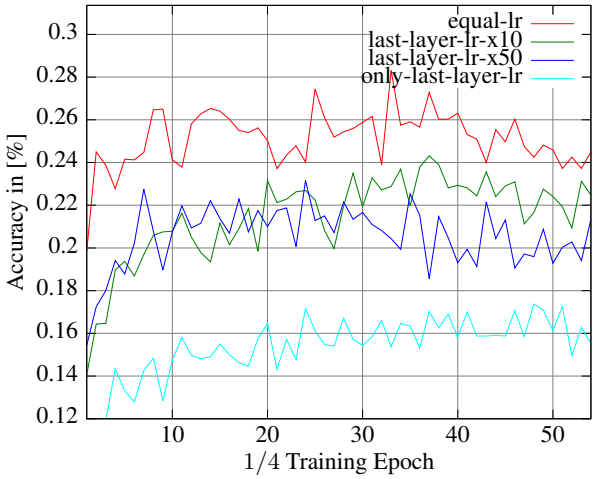

Figure 5. Learning rates between the pretrained layers and the new output layer have different weights. Evaluated on 40 phonemes.

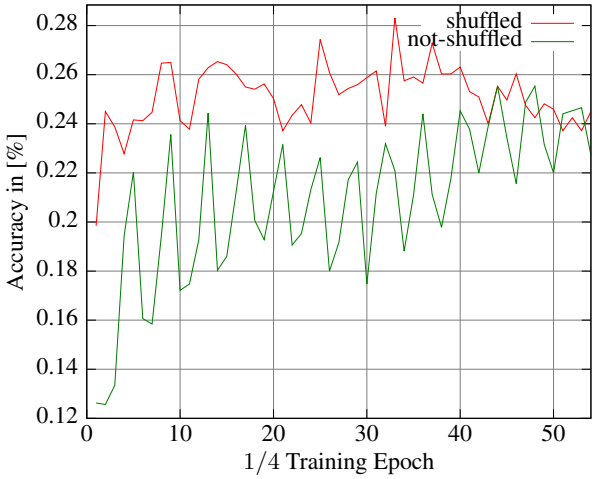

Figure 6. Shuffling the training data, evaluated on 40 phonemes.