Spatiotemporal Convolutional Features for Lipreading

Karel Paleček(✉)

Institute of Information Technology and Electronics,
Technical University of Liberec, 461 17 Liberec, Czech Republic
karel.palecek@tul.cz

Abstract. We propose a visual parametrization method for the task of lipreading and audiovisual speech recognition from frontal face videos. The presented features utilize learned spatiotemporal convolutions in a deep neural network that is trained to predict phonemes on a frame level. The network is trained on a manually transcribed moderate size dataset of Czech television broadcast, but we show that the resulting features generalize well to other languages as well. On a publicly available OuluVS dataset, a result of 91% word accuracy was achieved using vanilla convolutional features, and 97.2% after fine tuning – substantial state of the art improvements in this popular benchmark. Contrary to most of the work on lipreading, we also demonstrate usefulness of the proposed parametrization in the task of continuous audiovisual speech recognition.

Keywords: Audiovisual speech recognition · Deep learning · Spatiotemporal convolutional network · Lipreading

1 Introduction

Automatic lipreading is a task of recognizing speech purely from a video of a talking face. It is an inherently difficult task due to limited information content, speech ambiguities and high speaker-dependence. Majority of the work in this area has therefore traditionally concentrated on simplified applications such as phrase or continuous digit recognition. An area closely related to lipreading is audiovisual speech recognition (AVSR), where the main role of lipreading lies in enhancing performance of acoustic decoders, typically under noisy conditions. However, mainly due to the lack of publicly available data, large vocabulary lipreading is scarcely investigated even when coupled with acoustic decoders. For an overview of pre-deep learning advances in lipreading and AVSR see [14].

With its rapid advancement over the past decade, deep learning has gradually found its way into visual speech recognition as well as other applied research areas. One of its first applications in lipreading was in [6], where Ngiam et al. employed deep autoencoder trained via layer-by-layer stacked Restricted Boltzmann Machines. Purpose of the resulting deep belief network was to create joint audio-video bottleneck parametrization, which Ngiam et al. subsequently

K. Ekštein and V. Matoušek (Eds.): TSD 2017, LNAI 10415, pp. 438–446, 2017.
DOI: 10.1007/978-3-319-64206-2_49

used in a support vector machine (SVM) to classify videos of isolated letters and digits on two popular datasets AVletters and CUAVE. Another example is [7], where Noda et al. trained a convolutional neural network (CNN) on still images of mouths to predict phonemes and classified 300 Japanese isolated words using a hidden Markov model (HMM) trained on deep bottleneck features. A fully end-to-end approach was taken by Wand et al. [12], where 51 isolated words were classified using long short term memory (LSTM) recurrent network. Most recently, two advanced end-to-end deep learning systems for lipreading sentences were presented in [1,2]. Assael et al. [1] trained the system to recognize structured sentences of the GRID corpus by optimizing connectionist temporal classification (CTC) criterion and significantly improved state of the art word error rate (WER) from 13.6% to 4.8% in a multi-speaker split, albeit with still only 51 word vocabulary. Chung et al. [2] designed a first end-to-end trained truly large vocabulary deep learning system for lipreading sentences in the wild. To this end, they utilized watch, listen, attend, and spell framework instead of CTC, and were able to push the results on GRID even further down to 3.3%. Their system was, however, pre-trained on a large proprietary dataset of BBC television broadcast with over 100 thousands audiovisual utterances, not available to other researchers.

In this work, we exploit the predictive power of deep learning methods from a perspective of feature extraction for lipreading, similarly to e.g. [7]. We employ spatiotemporal convolutional network in order to predict frame-level labels, therefore also utilizing speech dynamics. As a basic visual speech unit, we choose phonemes, as it has repeatedly been shown, see e.g. [11], that even for purely visual tasks, visemes suffer from several key problems such as low ratio of between to inner class variance and strong context dependence. We then utilize the unnormalized log-probability, the network last layer's output, as a visual parametrization for a traditional HMM-based decoding system. In the experiments section we show that despite the network being trained to predict Czech phoneme set (PACZ) [8], the features achieve state of the art results on English data too, demonstrating their generalization ability and robustness. Advantages of this approach mainly include easy integration into existing frameworks, where the visual information can be plugged into as an additional modality. The proposed visual parametrization is evaluated in both lipreading and continuous speech AVSR.

2 Convolutional Features

Spatiotemporal convolution has been proven successful in video classification tasks, see e.g. [3]. It generalizes classic 2D convolution by also considering the time dimension, i.e.

$$(\boldsymbol{x} \circledast \boldsymbol{w})_{t_0,u_0,v_0,c_0} = \sum_{t=0}^{l} \sum_{u=0}^{m} \sum_{v=0}^{n} \sum_{c=1}^{C} w_{c_0,t,u,v,c} x_{t+t_0,u+u_0,v+v_0,c} + b_{c_0} \tag{1}$$

where $\boldsymbol{x}$ represents a 4D input video sequence of l frames with $m \times n$ pixels and C channels (e.g. RGB images), $\boldsymbol{w}$ is a bank of d 4D convolution kernels, and $\boldsymbol{b}$ is an added bias.

Here, we exploit its representational power to classify short video chunks $\boldsymbol{x}$ into one of 48 Czech phoneme classes including silences. The chunks consist of 7 or 11 frames of 64×64 RGB region of interest (ROI) that cover the speaker's mouth and its closest surroundings. We stack four blocks of spatiotemporal convolutions (1), batch normalization, spatiotemporal max-pooling and rectified linear unit, with each new layer having twice more convolution kernels than the previous one. In order to produce probability for each class, a linear layer with output dimension equal to the number of phonemes is added after the last convolution. See Fig. 1 for the details.

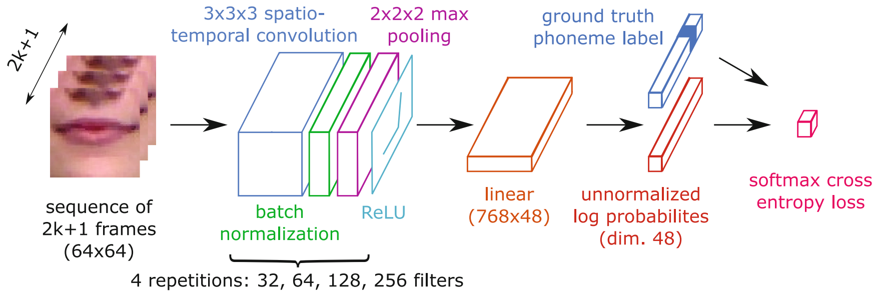
2k+1
3x3x3 spatio-temporal convolution
2x2x2 max pooling
ground truth phoneme label
sequence of 2k+1 frames (64x64)
batch normalization
ReLU
linear (768x48)
unnormalized log probabilites (dim. 48)
softmax cross entropy loss
4 repetitions: 32, 64, 128, 256 filters

Fig. 1. Spatiotemporal convolutional network architecture

The optimal parameters $\boldsymbol{\theta}^*$ (i.e. weights and biases) of the network are estimated by minimizing a softmax cross entropy loss of the network output vector $f(\boldsymbol{x};\boldsymbol{\theta})$ against ground truth labels r, i.e.

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \sum_{\boldsymbol{x},r \in \mathcal{X}} -f(\boldsymbol{x};\boldsymbol{\theta})_r + \log \sum_j e^{f(\boldsymbol{x};\boldsymbol{\theta})_j} \tag{2}$$

where $\mathcal{X}$ is a training set consisting of labeled pairs $(\boldsymbol{x}, r)$. For the optimization, we employ the mini-batch stochastic gradient descent algorithm with a momentum of 0.9.

After the network is trained, we use its output vector $f(\boldsymbol{x}_i;\boldsymbol{\theta}^*)$, whose j-th element represents an unnormalized logarithmic probability of the j-th phoneme, as a robust visual parametrization for the i-th frame. In order to deal with borderline cases, the input video is padded with the first and last frames on its respective ends.

The features are post-processed by concatenating parametrization vectors for several consecutive frames and reducing their dimensionality via linear discriminant analysis (LDA). Delta features computed as a difference between two frame feature vectors may be appended to the resulting parametrization. The length of the neighborhood for concatenation or inclusion of delta features are treated as hyperparameters and are cross validated as per splits in Table 1.

Table 1. Overview of datasets used in this work

Dataset	# speakers	Voc. size	Training set	Validation set
TV-data	∼50	26197	6.47 h	2.2 h
OuluVS	20	10	19 speakers (∼29 min)	1 speaker (∼1.5 min)
TULAVD-isol	54	366	45 speakers (∼1 h)	9 speakers (∼12 min)
TULAVD-cont	54	366	45 speakers (∼3.1 h)	9 speakers (∼40 min)

3 Data

Deep convolutional networks usually require large amount of data to train from scratch. In areas such as image classification, the preferred way often is to use a pre-trained model and then fine-tune it for specific purposes. In our case, this is not an option, since there are no publicly available spatiotemporal models that would be close to our application.

3.1 TV-data

In order to train the network, we utilize a manually transcribed dataset of Czech television broadcast. The complete dataset contains over 800 h of video content, most of which, however, is not suitable for training of our lipreading system. We process it in following way. Faces in each frame of every video are detected using histogram of oriented gradients (HOG) based deformable part model as implemented in the dlib library [5]. Also, for every frame precise locations of 68 landmarks describing facial shape of each face are predicted using the Ensemble of Regression Trees (ERT) algorithm [4]. We then pick a small subset of the complete data such that the resulting sequences are at least 3 s long and have only a single frontal facing talking speaker for the whole length. In order to achieve scale invariance we define the size of the region of interest (ROI) relative to the normalized mean facial shape. The coordinates of the ROI in the input image are then found by computing Euclidean transformation between the normalized shape and the detected one via least squares minimization.

All videos have been manually checked whether the visible face really is the speaker, or, for example, only serves for illustration purposes, such as when the studio host actually talks over the phone. The complete clean subset then contains approximately 11 h of audiovisual data, which have been split into training, validation, and test set in 0.6 : 0.2 : 0.2 respective ratio. See Table 1 for details. The phoneme class labels for each frame are produced by force-aligning a pre-trained acoustic model on mel frequency cepstral (MFCC) parametrization of the synchronized audio stream.

3.2 OuluVS

We demonstrate the performance of our proposed visual features on more restricted datasets that are better suited to video-only lipreading. OuluVS [13] is a popular and widely used publicly available dataset containing 20 speakers (17 male, 3 female), each of which utters 10 different short phrases five times. Examples of such phrases are for instance “Hello!” or “How are you?”. The videos were recorded at 25 fps with resolution of 720 × 576 pixels in an interlaced mode. Even though OuluVS ships with four different kinds of pre-extracted ROIs, we use our own extraction procedure similar of Sect. 3.1.

3.3 TULAVD

TULAVD [9] contains data from 54 speakers, of which 23 are female and 31 male with age ranging from 20 to 70 years. Each speaker uttered 50 isolated words and 100 sentences in Czech language, which were automatically selected according to phonetic balance. First 50 sentences are common for all speakers, whereas the second 50 differ. Audiovisual utterances were captured by two Logitech C920 FullHD webcams and Microsoft Kinect, which also offers depth stream that is fully synchronized with the video. In this work, we only consider 640×480 pixels RGB data from the Kinect. The visual pre-processing pipeline is similar to the other two considered datasets.

4 Experiments

First we evaluate our proposed spatiotemporal convolutional features for lipreading in simpler and more restricted task of purely visual isolated unit recognition. To this end, videos are parametrized using the 48-dimensional output of our spatiotemporal convolution network, as trained on the TV-data dataset. On top of these features, a hidden Markov model with Gaussian mixture emissions (HMM/GMM) is constructed for each possible unit, i.e. a word on the TULAVD dataset, or a phrase in case of OuluVS. In case of TULAVD, we perform a 6-fold cross validation and report the average score. With OuluVS, we follow a speaker-independent leave-one-out cross validation scheme that is compatible with most existing research on this dataset.

Table 2 then presents the achieved word accuracy and compares the proposed parametrization to other popular methods, namely discrete cosine transformation (DCT) of the ROI, principal component analysis (PCA), active appearance model (AAM), spatiotemporal local binary pattern descriptor (LBPTOP) [13], dynamic histogram of oriented gradients [9], and random forest manifold alignment [10]. Input chunks of length 7 ($k = 3$) and 11 ($k = 5$) were evaluated in the experiments. As we can see, the former performs much better than the latter. Although longer sequences offer more context and discriminative information, they also represent more specific cases, and training models on longer inputs therefore is more prone to overfitting and sensitive to mismatch between training and testing sets.

Table 2. Recognition of isolated words and phrases

Parametrization	TULAVD [%]	OuluVS [%]
	50 words	10 phrases
DCT	72.5	79.2
PCA	73.9	77.9
AAM	74.1	82.1
LBPTOP [13]	74.2	82.5
HOGTOP [9]	86.4	85.7
Random forest manifold alignment [10]	-	89.7
phoneme-3D-CNN-k3	90.6	91.0
phoneme-3D-CNN-k5	74.4	-
phoneme-3D-CNN-k3 (fine-tuned)	**91.0**	**97.2**

Interestingly, parametrization trained on Czech data also performs very well in recognition of English phrases. We achieved state of the art result on the OuluVS of 91% correctly recognized phrases in speaker-independent lipreading. Moreover, the model could be further fine-tuned to different phoneme set by appending additional linear layer on top of the original output to produce unnormalized log-probabilities of each of the 33 English phonemes that are uttered within this dataset. Note that 20 different models had to be fine-tuned for every respective training set of the leave-one-out protocol in order to preserve the speaker independence. By applying the fine-tuned model, the best score reached as high as 97.1%. It is fair to note however, that the result mainly is due to low language variability caused by a small vocabulary, i.e. phonemes only have limited context options, making the chunks easier to classify. Improvement of fine-tuning on TULAVD is marginal, as the score increased by a mere 0.4%.

We also performed experiments continuous speech recognition on the TULAVD dataset. To this end, spatiotemporal features were linearly up-sampled to 100 Hz and concatenated with 39 mel-frequency cepstral coefficients (MFCC) that had been extracted from the audio stream. An HMM model was trained for each phoneme of the PACZ set. Simple bi-gram language model (LM) pre-trained on an external Czech newspaper corpus was applied and filtered for the 366 words in the test utterances. Again the proposed convolutional features perform the best, albeit with much smaller margin than in previous experiments. This is to be expected, as generally visual signal contains much less information than the acoustic one, which will be exploited by the decoder.

Table 3 presents the results of recognition of 50 test sentences of the TULAVD dataset with vocabulary size of 366 words for selected visual features. Results are represented by word accuracy (WAcc) and word correctness (WCor), which differ in that WCor ignores insertions errors. We can observe that under good acoustic conditions, visual parametrization adds relatively little improvement, which is of no surprise. The only two parametrizations increasing the resulting score are LBPTOP and our proposed spatiotemporal convolutional features.

Table 3. Recognition of continuous speech on TULAVD dataset

Parametrization	WAcc [%]	WCor [%]
MFCC	76.7	84.0
MFCC+DCT	42.7	67.6
MFCC+AAM	74.4	83.8
MFCC+LBPTOP [13]	77.7	86.5
MFCC+HOGTOP [9]	75.4	86.2
MFCC+phoneme-3D-CNN-k3	**79.2**	**86.9**

4.1 Error Analysis

It is interesting to look at the errors the network makes on the character level. When taking the maximum of the log-probabilities, the average phoneme accuracy on the validation set of the TV-data reached 25.2%. But the correct label is among the top 5 scoring classes more than 60% of the time, and among the top 10 almost 80% of the time. This suggests that even though the correct class on average still achieves high log-probability, which is important for purposes of parametrization, there are lots mismatches between phonemes of similar corresponding visemic classes. Indeed, by examining the confusion matrix we observed that the most common errors occur e.g. between long and short 'o' (0.65), 'ts' (pronunciation of 'c' in Czech) and 's' (0.49), or 'f' and 'v' (0.35). Note that, however, switching to prediction of visemes does not help due to problems mentioned in Sect. 1.

Figure 2 shows per-phoneme accuracy for 28 phonemes (excluding noise) that attained at least 1% frame-wise relative frequency in the validation set of the TV-data. It can be seen that clearly visible and on average longer vowels such as 'aa', 'o', or 'u', or visually distinctive consonants such as 'm', 'p', 's', or 'v' are

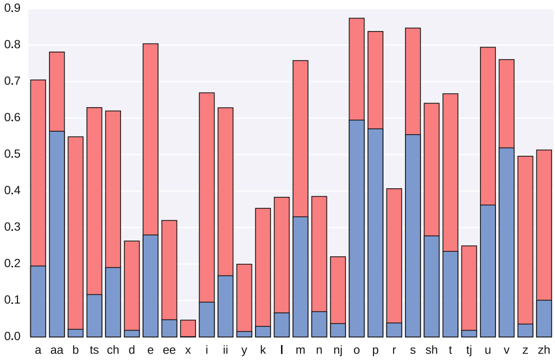

Fig. 2. Per-phoneme accuracy of the spatiotemporal network. The phoneme label convention follows PACZ [8]. Blue: top-1, red: top-5. (Color figure online)

predicted by the network with higher accuracy than e.g. ‘x’ (PACZ [8] label for Czech digraph ‘ch’, pronounced as voiceless velar fricative), ‘nj’, or ‘tj’. As many of the easily distinguishable phonemes are shared between Czech and English, by capturing characteristic lip motion in its output the network generalizes well to the latter language as well.

5 Conclusion

We have presented a new method of feature extraction for lipreading suitable for both recognition of isolated unit as well as continuous speech. The features are based on learned spatiotemporal convolutions in a deep network trained to predict phonemes on a frame level. The proposed parametrization method can be utilized in traditional or hybrid decoders based on hidden Markov model as well as end-to-end trained deep learning systems. Compared to traditional parametrization methods, we have achieved superior performance on two datasets with different languages, demonstrating feature representativeness and robustness against input variability. On the popular OuluVS speaker-independent benchmark, state of the art word error rate was substantially reduced from 10.3% to 2.8%. Also, unlike many other popular parametrization methods, our features also generalize well to recognizing continuous speech jointly with audio, even in clean environment without acoustic noise. Our efforts in the nearest future will focus on developing end-to-end learned lipreading system for spontaneous speech using spatiotemporal features and attention mechanism.

References

1. Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N.: LipNet: sentence-level lipreading. CoRR abs/1611.01599 (2016). http://arxiv.org/abs/1611.01599
2. Chung, J.S., Senior, A.W., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. CoRR abs/1611.05358 (2016). http://arxiv.org/abs/1611.05358
3. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, pp. 1725–1732. IEEE Computer Society, Washington, DC (2014)
4. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014, pp. 1867–1874 (2014)
5. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
6. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, 28 June–2 July 2011, pp. 689–696 (2011)
7. Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H., Ogata, T.: Lipreading using convolutional neural network. In: International Speech and Communication Association, pp. 1149–1153 (2014)

8. Nouza, J., Psutka, J., Uhlíř, J.: Phonetic alphabet for speech recognition of Czech (1997)
9. Palecek, K.: Lipreading using spatiotemporal histogram of oriented gradients. In: EUSIPCO 2016, Budapest, Hungary, pp. 1882–1885 (2016)
10. Pei, Y., Kim, T., Zha, H.: Unsupervised random forest manifold alignment for lipreading. In: IEEE International Conference on Computer Vision, Sydney, Australia, pp. 129–136 (2013)
11. Ramage, M.D.: Disproving visemes as the basic visual unit of speech (2013). http://www.mramage.id.au/phd
12. Wand, M., Koutník, J., Schmidhuber, J.: Lipreading with long short-term memory. CoRR abs/1601.08188 (2016). http://arxiv.org/abs/1601.08188
13. Zhao, G., Barnard, M., Pietikäinen, M.: Lipreading with local spatiotemporal descriptors. IEEE Trans. Multimedia **11**(7), 1254–1265 (2009)
14. Zhou, Z., Zhao, G., Hong, X., Pietikinen, M.: A review of recent advances in visual speech decoding. Image Vision Comput. **32**(9), 590–605 (2014)