Utilizing Lipreading in Large Vocabulary Continuous Speech Recognition

Karel Paleček(✉)

The Institute of Information Technology and Electronics,
Technical University of Liberec, Studentská 2/1402, 46117 Liberec, Czech Republic
karel.palecek@tul.cz

Abstract. Vast majority of current research in the area of audiovisual speech recognition via lipreading from frontal face videos focuses on simple cases such as isolated phrase recognition or structured speech, where the vocabulary is limited to several tens of units. In this paper, we diverge from these traditional applications and investigate the effect of incorporating the visual information in the task of continuous speech recognition with vocabulary size ranging from several hundred to half a million words. To this end, we evaluate various visual speech parametrizations, both existing and novel, that are designed to capture different kind of information in the video signal. The experiments are conducted on a moderate sized dataset of 54 speakers, each uttering 100 sentences in Czech language. We show that even for large vocabularies the visual signal contains enough information to improve the word accuracy up to 15% relatively to the acoustic-only recognition.

Keywords: Audiovisual speech recognition · Lipreading · LVCSR

1 Introduction

It has been well established that visual cues extracted from lip movement can help the automatic speech recognition process mainly in noisy acoustic conditions. With sufficiently small vocabulary, frontal face videos provide enough information for reliable recognition even without acoustic data. Large variety of methods for visual parametrization, feature post-processing and modality integration have been proposed to date. For a comprehensive overview of recent advances in lipreading and audiovisual speech recognition see e.g. work by Zhou et al. [15].

Utilization of automatic lipreading techniques for large vocabulary continuous speech recognition (LVCSR) is rarely explored in the current literature. One of the main obstacles is the lack of freely available datasets, with AVICAR [8] probably being the only option. In [7] Lan et al. used proprietary corpus of 12 speakers and 1000 word vocabulary in order to classify individual visemes, but they did not report the word-level accuracy. Much of the important work on audiovisual LVCSR via frontal face lipreading was conducted in IBM laboratories

A. Karpov et al. (Eds.): SPECOM 2017, LNAI 10458, pp. 767–776, 2017.
DOI: 10.1007/978-3-319-66429-3_77

during the early 2000s [6,11]. The experiments were performed on IBM's proprietary large audiovisual dataset ViaVoice containing 290 speakers and vocabulary size of 10403 words and found the integration of visual features beneficial only for noisy acoustic conditions. Recently, two papers [1,3] using end-to-end trained deep learning systems improved state of the art in lipreading of sentences. Assael et al. [1] trained the system to recognize structured sentences of the GRID corpus [5] 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. [3] 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, pretrained on a large proprietary dataset of BBC television broadcast with over 100 thousands audiovisual utterances, not available to other researchers.

In this work, we tackle the problem from the traditional feature extraction and classification paradigm, which allows for easier integration and straightforward comparison with existing acoustic-only systems based on hidden Markov Model (HMM) decoding. We evaluate several popular state of the art visual speech parametrizations in the task of audiovisual LVCSR and experimentally investigate their impact on the word error rate. To this end, we utilize moderate sized dataset with 54 speakers and simulate various vocabularies of up to 500 k words. Moreover somewhat non-traditionally, since our dataset is recorded using Kinect, we also evaluate the lipreading performance when depth data is incorporated. Interestingly enough, recognition from the depth stream sometimes yields better results than from RGB, with the advantage of partial complementarity, which makes it suitable for integration with RGB.

The rest of the paper is organized as follows. We describe our dataset in Sect. 2. The visual parametrizations along with our modifications are explained in Sect. 3. System overview is presented in Sect. 4. Finally, the performed experiments and the discussion are described in Sect. 5.

2 Data

TULAVD is our own dataset recorded at the Technical University of Liberec containing 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. The sentences were divided into two groups with the first 50 being common to all speakers and the other 50 speaker-specific. The dataset also contains 583 manually annotated images of all speakers in various poses, expressions and face occlusions, which constitute a training dataset for the ESR detector. The audiovisual utterances were recorded in an office environment using Genius lavalier microphone, two Logitech C920 FullHD webcams, and Microsoft Kinect, which also offers depth stream that is fully synchronized with the video.

Fig. 1. Sample frame of RGB image and corresponding depth map

Only the microphone and Kinect RGBD data with resolution of 640×480 pixels is used in this work. See Fig. 1 for a sample frame from a frontal face video of a talking speaker. In order to build the language models, we also collected more than 60 GB of texts mostly consisting of online journals and manual transcriptions of television and radio broadcast.

3 Visual Speech Parametrization

In audio visual speech literature, **discrete cosine transform (DCT)** represents a widely used method for visual speech parametrization, and often the first choice. The visual speech features are usually selected as a subset of the full 2D DCT transform computed over the ROI.) Number of feature selection methods have been proposed to date, e.g. zig-zag ordering or selection by mutual information. In this work, we treat the coefficient selection as hyperparameter optimization problem. We sort the DCT coefficients based on an average energy obtained on a training set and then select their optimal number according to validation score.

The Active Appearance Model (AAM) is a well-known method for describing appearance of a deformable object by a hierarchical application of PCA. The appearance is represented by shape and texture that are both modeled linearly using PCA. These modality-specific representations are normalized and concatenated into a single vector, and then subjected to a second-level PCA. In this work, we extract the AAM features using 46 landmarks from the lower part of the speakers face, see the AAM-r in Fig. 2. In addition to the standard AAM, we also evaluate a variant with both video and depth texture included as a form of early feature integration. We denote this case as **DAAM**. The number of AAM coefficients constitutes a hyperparameter that is optimized w.r.t. the recognition accuracy.

For our experiments we also utilize the popular **Spatiotemporal Local Binary Patterns (LBPTOP)** introduced in [14]. Local Binary Pattern (LBP) describes the texture in terms of a histogram of binary numbers that are formed by comparing each pixel of the image to its close neighborhood. Zhao et al. extended the static LBP by considering the neighborhood not only in the spatial domain, but also in the time axis, in order to capture the speech dynamics.

Fig. 2. Possible landmark configurations. We empirically found out that the second configuration (AAM-r) performs best in most experiments

Thus, LBPs are effectively extracted from three orthogonal planes (TOP): xy, xt, and yt. These are then concatenated into a single vector forming the visual speech parametrization. Contrary to the original work [14], we extract the LBP-TOP densely for every frame. We cross validate the parameters of the LBP, i.e. the number of histogram bins and the aggregation method (standard, rotation invariant, uniform, non-rotation invariant uniform).

The last considered parametrization is the **Spatiotemporal Histogram of Oriented Gradients (HOGTOP)**. We proposed this parametrization in [10] inspired by the LBPTOP as a dynamization technique of the standard Histogram of Oriented Gradients (HOG). Normally, the histograms are formed by counting and weighting the gradient orientations in the xy plane. Here, we also add orientations from the xt and yt planes, process them independently, concatenate, and reduce the resulting HOG hypervector by PCA into the final parametrization. Extraction of the HOGTOP features is illustrated in Fig. 3. The only hyperparameter to be cross-validated is the final PCA dimension.

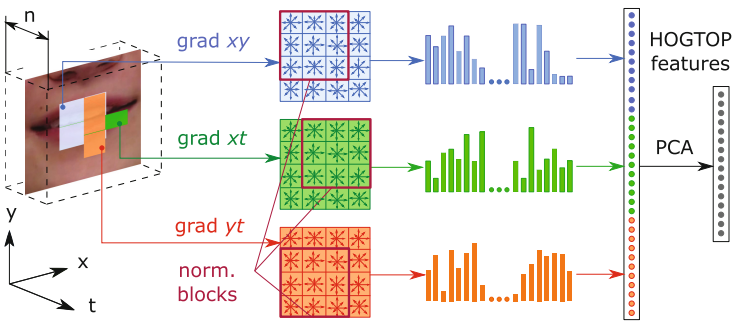

Fig. 3. Extraction of spatiotemporal histogram of oriented gradients

4 System Overview

4.1 Visual Front-end

We pre-process the image in several stages with progressing level of precision. First, an approximate position of the face is estimated using the well known Viola-Jones algorithm. We use the pre-trained model that ships with the OpenCV library. Second, to estimate the facial shape precise positions of 93 facial landmarks are obtained by utilizing the Explicit Shape Regression method (ESR) [2]. The ESR is a discriminative method that iteratively refines the joint landmark configuration (i.e. the face shape) based on the value of only few pixel differences and thus is very efficient (i.e. hundreds of frames per second on regular PC). However, since there is no objective to be optimized, the final landmark positions are slightly different in each frame, which introduces an inter-frame jitter. We reduce it by running the detector from different starting positions 10 times and then taking the median of the fit shapes.

Once the facial landmarks are localized, we define the region of interest (ROI) as a square area barely covering the mouth and its closest surrounding. In order to achieve scale invariance we set its size relative to the normalized mean shape. The geometric transformation for the extraction is estimated by aligning the normalized mean and the detected shapes. To further reduce the inter-frame landmark jitter and stabilize the ROI extraction, we average the fitting results over three neighboring frames in time.

4.2 Feature Extraction and Post-processing

The acoustic channel is parametrized by 39 Mel Frequency Cepstral Coefficients (MFCC) with a 25 ms window at a 100 Hz rate. The visual parametrizations described in Sect. 3 are extracted densely for each frame of the input utterance. Sequences $x_{t-k}, \ldots, x_{t+k}$ of $2k + 1$ feature vectors $x_{t'}$ are concatenated into hypervectors, where k represents the number of left and right adjacent frames, and then reduced by the linear discriminant analysis (LDA) with phonemes as class labels. The k is treated as a hyperparameter for each parametrization separately and therefore is subject to optimization of the validation score. Since visual features tend to be highly speaker dependent, we also perform feature mean subtraction (FMS) with the average computed over the whole utterance. Addition of delta (Δ) features is similarly to k also considered to be a hyperparameter and thus tuned for each parametrization separately. Finally, the video features are linearly interpolated from 30 Hz to 100 Hz frequency to match the acoustic parametrization.

4.3 Acoustic and Visual Models

Due to the limited amount of audiovisual data, we utilize only basic monophone models without context. There are 40 distinct phonemes of the PAC-CZ phonetic alphabet [9] and 13 corresponding visemes [4]. In order to obtain frame-level class

labels, we forced-aligned the audio recordings using a separate robust acoustic model that was trained on approximately 300 h of spoken data. The viseme labels were then obtained by a simple phoneme-viseme mapping proposed in [4] and shifted by approximately 0.023 s to synchronize the streams.

Phonemes and visemes are modeled using 3-state hidden Markov model (HMM) with Gaussian mixture emission probability. The main advantage of HMM in our context is that it allows for straightforward weighted combination of acoustic and visual channels via multi-stream synchronous variant of the model (MSHMM), in which each state q has an emission probability equal to the weighted product of the individual streams $s = (1, \ldots, S)$:

$$p\left(x^{(1)}, \ldots, x^{(S)}|q\right) = \prod_{s=1}^{S} p\left(x^{(s)}|q\right)^{\lambda^{(s)}}. \tag{1}$$

We treat the stream weights $\lambda^{(s)}$ as hyperparameters and therefore cross-validate them w. r. t. the recognition accuracy.

We utilized the HTK 3.4.1 toolkit to train the phoneme and viseme models. We followed a simplified procedure by first initializing the models with Viterbi training (`HInit`) and then reestimating with Baum-Welch in an isolated-unit manner (`HRest`). We have empirically found out that the more commonly used approach of embedded re-estimation using `HERest` only degrades the results in our case. This is due to the limited discriminative power of the visual parametrization that makes it unsuitable for alignment on the phonetic level, even when constrained by the acoustic features in the multi-stream model, and as a result, the re-estimation procedure fails to converge.

4.4 Language Models

We evaluate our audiovisual recognition system for four different bigram language models with vocabulary size ranging from 366 up to 500 k words, see Table 1 for the exact numbers. The smallest vocabulary contains only words from the corpus of our audiovisual dataset, whereas the other ones also include the most frequent words in Czech language. The word frequencies and language models are assessed using the 60 GB text corpus described in Sect. 2. We employed the SRILM toolkit [13] with Knesser-Nay smoothing for the language model training.

Table 1. Vocabularies considered in the experiments

LM	min	5 k	50 k	500 k
# words	366	5 182	50 056	499 993
# bigrams	48 338	9 865 k	73 905 k	141 670 k

5 Experiments

Throughout the experiments we follow the k-fold cross validation protocol. The 54 speakers are split into 6 groups of 9, where in each turn of the cross validation 5 groups constitute a training set and 1 is reserved for testing. We then report the average word accuracy (Wacc) achieved over the 6 different test sets.

The phonetic models are learned on all the available training data from each respective fold of the cross validation, which amounts to approximately 5 h of spoken data on average. In order to minimize the number of sources of variability across different folds and to better control the vocabulary, the test data comprise only of the first 50 sentences that are common to all speakers instead of the full set of 100 sentences.

5.1 Isolated Word Recognition

In order to tune the hyperparameters of the visual parametrizations described in Sect. 3, we followed a slightly different approach. For reasons of efficiency, these hyperparameters were optimized using 14-state whole-word models with one or two components per GMM in the task of lipreading of 50 isolated words. The optimized parametrizations were then used for unimodal recognition of the 50 isolated words using phoneme and viseme models. In these experiments we employed the HTK `HVite` decoder. Table 2 summarizes the results of both whole-word and phonetic models.

Table 2. Word accuracy [%] of isolated word recognition and lipreading

Param	Src.	Word	Phoneme		Viseme	
Mixtures:		1/2	8	16	8	16
MFCC	a	99,8	99,5	99,8	97,4	98,0
DCT	v	72,5	42,6	42,8	42,4	43,9
	d	74,4	39,3	42,5	38,6	43,1
AAM	v	74,1	57,5	58,5	59,0	59,3
	d	75,2	54,1	55,0	55,3	56,6
LBPTOP	v	74,2	54,6	56,4	54,6	56,3
	d	64,3	48,7	47,4	45,3	48,2
HOGTOP	v	**86,4**	59,5	61,0	59,8	60,1
	d	84,4	56,6	58,3	56,6	57,7
DAAM	v ∘ d	74,9	**62,0**	**64,6**	**63,0**	**64,7**

The experiment is conducted for both video (a) and depth (d) streams, with v ∘ d denoting their early integration, i.e. concatenation of the feature vectors. Note that in the special case of DAAM, the concatenation of video and depth

textures is also followed by coupling via PCA. One can observe that in this simpler scenario, video-based and depth-based parametrizations perform roughly on par, with their combination in the form of DAAM achieving the best results overall.

While the phoneme and viseme models reach similar word accuracies, they perform much worse compared to the whole-word approach. This illustrates one of the issues with the current state of the art in lipreading, where the parametrization and classification algorithms mainly target isolated unit recognition, and the results do not necessarily apply to systems with larger vocabularies.

5.2 Continuous Speech

The results on isolated word recognition show that on average viseme-based models outperformed the phone-based ones. However, the results are inconsistent and the margin never exceeds 2%. This observation may be attributed to the viseme context dependency on the surrounding vowels [12]. For instance, the u-shaped lip protrusion when pronouncing "s" in the word "super" significantly differs from the horizontal extension when pronouncing "s" in "see". As a result, it seems that phonemes cannot be unambiguously mapped to visemes in a surjective many-to-one manner. Considering this issue and potential problems with the score combination, we employed only phone models in the following experiments on continuous speech recognition.

Table 3 presents the achieved results. Due to performance reasons we switched from `HVite` to the Julius[1] decoder, which is compatible with HTK model definitions. For example, a + v denotes a middle fusion of audio and video channels via MSHMM with optimally set weights $\lambda^{(s)}$ that are cross-validated on a dense grid of all possible combinations with the step of 0.1 and constraint $\sum_s \lambda^{(s)} = 1$.

As expected, with the increasing size of vocabulary, the performance in terms of accuracy and correctness degrades rather quickly, which is mostly due to the relatively small amount of training data. On the other hand, in all experiments the combined audiovisual representations achieved to some improvement over acoustic-only recognition, showing that the visual cues provide useful information even for very large vocabularies with 500 k words. This especially holds for LBPTOP and HOGTOP, as they manage to exploit some of the speech dynamics, which is crucial for phoneme discrimination. The best results overall were obtained by our proposed HOGTOP features extracted from both video and depth, although the difference from video-only LBPTOP is almost negligible.

In contrast to recognition of isolated words, integration of the depth channel does not seem to improve the word accuracy. The only exception to this rule was the HOGTOP parametrization, which in most cases achieved slightly better results in the three modality setting.

For all four vocabularies the highest improvement achieved over audio-only recognition ranged between 5–7% absolutely, i.e. 7–15% relatively. In most cases the optimal weight ratio of audio and video (or depth) channels, which indicates

[1] https://github.com/julius-speech/julius.

Table 3. Word accuracy [%] of audiovisual speech recognition by middle fusion of acoustic and visual parametrizations for different vocabularies

Par.	Source	Vocabulary			
		min	5 k	50 k	500 k
MFCC	a	74,0	55,9	43,9	36,3
DCT	a + v	76,8	59,8	47,1	38,9
	a + d	74,3	55,5	43,4	38,3
	a + v + d	77,3	59,6	46,8	38,2
AAM	a + v	76,7	60,5	48,7	40,2
	a + d	76,8	60,0	48,0	39,5
	a + v + d	76,9	60,2	48,3	39,9
LBPTOP	a + v	79,2	62,7	**50,1**	**41,7**
	a + d	77,8	60,8	48,5	39,8
	a + v + d	79,3	62,6	50,0	41,4
HOGTOP	a + v	78,1	60,2	47,8	42,0
	a + d	77,2	58,3	46,2	40,7
	a + v + d	**79,4**	**62,9**	**50,1**	41,6
DAAM	A + v ∘ d	75,2	58,6	48,0	40,7

the relative importance of each modality, was 0.7 : 0.3 or 0.8 : 0.2, with the former being more common for the 500 k vocabulary. Note that the results hold for relatively clean data, i.e. without acoustic noise, and one might expect even higher relative improvement in worse conditions.

6 Conclusion

We have shown that given quality parametrization, the visual cues provided by the lip movement can improve the recognition accuracy even for very large vocabularies with hundreds of thousand words. The best results were achieved using the HOGTOP and LBPTOP features that are designed to exploit the speech dynamics as opposed to static features such as AAM. The relative improvement of audiovisual over audio-only recognition ranged between 7% and 15% when the channels were integrated via multi-stream hidden Markov model with optimally set weights. There might be a potential issue in that improvement observation could be somewhat influenced by the limited amount of data and it is uncertain if the same results would hold for more robust acoustic models trained on hundreds of hours data. In order to verify this, transfer learning techniques could potentially be employed to circumvent the lack of large audiovisual dataset availability.

References

1. Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N.: Lipnet: Sentence-level lipreading. CoRR abs/1611.01599 (2016)
2. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: CVPR (2012)
3. Chung, J.S., Senior, A.W., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. CoRR abs/1611.05358 (2016)
4. Císař, P.: Application of lipreading methods for speech recognition. Ph.D. thesis (2006)
5. Cooke, M., Barker, J., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**(5), 2421–2424 (2006)
6. Glotin, H., Vergyr, D., Neti, C., Potamianos, G., Luettin, J.: Weighting schemes for audio-visual fusion in speech recognition. In: Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001), vol. 1, pp. 173–176 (2001)
7. Lan, Y., Theobald, B., Harvey, R., Bowden, R.: Improving visual features for lipreading, pp. 142–147 (2010)
8. Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., Huang, T.S.: AVICAR: audio-visual speech corpus in a car environment. In: 8th International Conference on Spoken Language Processing, INTERSPEECH 2004 - ICSLP, Jeju Island, Korea, 4–8 October 2004
9. Nouza, J., Psutka, J., Uhlíř, J.: Phonetic alphabet for speech recognition of czech (1997)
10. Palecek, K.: Lipreading using spatiotemporal histogram of oriented gradients. In: EUSIPCO 2016, Budapest, Hungary, pp. 1882–1885 (2016)
11. Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W.: Recent advances in the automatic recognition of audio-visual speech. In: Proceedings of the IEEE, pp. 1306–1326 (2003)
12. Ramage, M.D.: Disproving Visemes as the Basic Visual Unit of Speech. Ph.D. thesis (2013)
13. Stolcke, A.: SRILM - an extensible language modeling toolkit. In: Proceedings of ICSLP, Denver, USA, vol. 2, pp. 901–904 (2002)
14. Zhao, G., Barnard, M., Pietikäinen, M.: Lipreading with local spatiotemporal descriptors. IEEE Trans. Multimedia **11**(7), 1254–1265 (2009)
15. Zhou, Z., Zhao, G., Hong, X., Pietikinen, M.: A review of recent advances in visual speech decoding. Image Vis. Comput. **32**(9), 590–605 (2014)