Journal on Multimodal User Interfaces (2018) 12:309–318
https://doi.org/10.1007/s12193-018-0266-2

ORIGINAL PAPER

Experimenting with lipreading for large vocabulary continuous speech recognition

Karel Paleček[1]

Received: 31 October 2017 / Accepted: 15 June 2018 / Published online: 16 July 2018

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 and also depth 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 and depth signals. The experiments are conducted on a moderate sized dataset of 54 speakers, each uttering 100 sentences in Czech language. Both the video and depth data was captured by the Microsoft Kinect device. We show that even for large vocabularies the visual signal contains enough information to improve the word accuracy up to 22% relatively to the acoustic-only recognition. Somewhat surprisingly, a relative improvement of up to 16% has also been reached using the interpolated depth data.

Keywords Audiovisual speech recognition · Lipreading · LVCSR

1 Introduction

It has been repeatedly shown that in humans, understanding speech is a multi-modal process. Probably the most famous example of this fact is the well known McGurk effect [13]. It illustrates how the apparent movement of speakers lips might influence the actual acoustic perception. It was later explained by Summerfield [28], who suggested that the visual component carries information about the *place* of articulation, e.g. whether the sound is labial, dental, alveolar, etc., while the auditory one mainly determines the *manner*, e.g. voice, voiceless, nasal, fricative, etc. This hypothesis is known as VPAM (visual: place, auditory: manner).

Due to the complementary rather than redundant character of the audio and video modalities, automatic lipreading purely from the visual channel presents a challenging task. Without the acoustic signal, the video alone carries limited information content, displays speech ambiguities and its intelligibility highly depends on the speaker. However, even despite these difficulties, it has been well established that visual cues extracted from lip movement can help the automatic speech recognition process, especially in noisy acoustic conditions. Such supplementary role of lipreading then leads to a closely related area of audio-visual automatic speech recognition (AVASR).

With sufficiently small vocabulary, frontal face videos provide enough information for reliable lipreading even without the 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. [27,31].

During the last decade, algorithms based on boosting, graph embedding and manifold learning have proven quite successful for tackling the lipreading problem, see e.g. [17,21,32]. Such systems exploit sophisticated feature selection and modeling techniques to project the high dimensional input to a more discriminative subspace better suited for classification. However, their main disadvantage lies in the inapplicability to recognition based on sub-word units. Usually,

This paper is an extended version of [20] that was presented at the SPECOM 2017 conference.

✉ Karel Paleček
karel.palecek@tul.cz

[1] Institute of Information Technology and Electronics, Technical University of Liberec, 461 17 Liberec, Czech Republic

the projection algorithms behave essentially as static classifiers, i.e. the whole utterance must first be normalized to a specific length before it can be classified as a single feature vector. This makes the systems closely tied to the target application, for example isolated phrase or digit recognition, and not easily generalizable to e.g. continuous speech recognition.

With its rapid advancement, the ubiquitous deep learning has gradually found its way into visual speech recognition, and seems like the next major trend in the area. One approach is to utilize the bottleneck features, which can be learned either in a supervised [15] or a unsupervised manner [14]. An end-to-end trained system with long short term memory (LSTM) recurrent network (RNN) was used e.g. in [29]. A bidirectional LSTM together with convolutional features (CNN) was proposed for lipreading in [22]. Most recently, two advanced end-to-end trained lipreading networks were developed independently by Assael et al. [1] and Chung et al. [3]. The former used a connectionist temporal classification (CTC) and the latter was based on the watch, listen, and spell approach used often seen in machine translation. While especially the end-to-end trained systems achieve impressive results, a drawback is that their utilization in existing systems that are often based on hidden Markov model (HMM) is not straightforward, because these systems are essentially monolithic and serve as both feature extractors and as speech decoders, i.e. also handling vocabulary, language modeling etc.

Research in the audio-visual speech recognition focuses mainly on different approaches to audio-video combination. One may broadly classify the methods into three groups: early, late, and hybrid (middle) integration. Early integration combines the modalities on the lowest level, typically by concatenation of their feature vectors, see e.g. [7]. More complex ways of early integration are also possible; for example Ngiam et al. [14] combined them by creating a joint audio-visual bottleneck features within an autoencoder. Late fusion combines the modalities on the classifier level, after classifying each stream separately. There are various rules of output combinations, see e.g. [12] for an overview. The typical and probably the most popular example of hybrid integration, or sometimes also referred to as middle fusion, is multi-stream synchronous hidden Markov model (MSHMM) [12], in which every state has multiple emissions, one for each modality. See e.g. [6] for its application in lipreading with dynamically adapted weights. More complex variants of dynamic graphs can also be used to partially correct asynchrony between the streams, see e.g. [25].

2 Related work

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 [11] being one of the few options. Recently, TCD-TIMIT corpus [9] has also been released for research.

Research therefore often use their own proprietary datasets, not available to others, and that makes the results difficualt to compare. For example, in [10] 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 during the early 2000s [8,23]. 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, 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, which is an extended version of our previous conference paper [20], 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 both existing state of the art visual speech parametrizations as well as novel ones 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 500k 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.

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

3 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. 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 our work, we use linearly interpolated depth maps instead of raw data to ensure there are no zero "holes" that correspond to the black spots in the Fig. 1.

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.

4 Visual speech parametrization

4.1 Discrete cosine transform

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 a 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.

4.2 Active appearance model

The Active Appearance Model (AAM) is a well-known method for describing appearance of a deformable object by a hierarchical application of Principal Component Analysis (PCA). The appearance is represented by shape and texture that are both modeled linearly using PCA. The shape is a concatenation of (x, y) coordinates of facial landmarks into a vector. Similarly, the texture is formed by concatenating pixels from the hull defined by the landmarks. In hte second stage, these modality-specific representations are normalized and combined into a single vector, and then decorrelated by another PCA.

When utilizing AAM as a visual parametrization, one can consider various subsets of the facial landmarks, such as depicted in Fig. 2. Moreover, one can also choose to utilize only subset of the AAM features, i.e. shape, texture, or both. We empirically evaluated these different configurations in the task of purely visual isolated word lipreading on several datasets and have found out that the AAM-r variant consistently achieves the highest performance. We can also observe that for all landmark subset variants, the combined feature vector performs best, as it benefits from both the shape and texture information and their partially complementary nature. See the Table 1 for the results on the TULAVD dataset.

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. In other words, the only difference from the standard AAM is that three (shape, video, depth) instead two (only shape and video) modalities are concatenated to form the input to the second stage PCA. We denote this case as **DAAM**. The number of AAM coefficients constitutes a hyperparameter that is optimized w.r.t. the recognition accuracy.

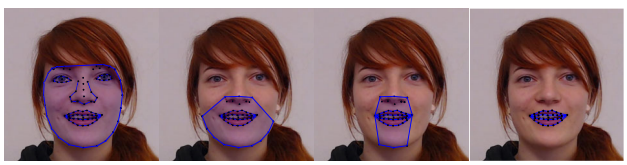

Fig. 2 Possible landmark configurations. We empirically found out that the second configuration (AAM-r) performs best in most experiments. Values in parentheses denote the total number of facial landmarks of each respective configuration

Table 1 Isolated word recognition accuracy (%) achieved using different AAM configurations on the TULAVD dataset

AAM type	P	Video		Depth	
		L	C	L	C
AAM-f	48.4	48.6	54.9	42.6	54.4
AAM-r	**54.6**	**52.7**	**58.1**	**48.9**	**59.4**
AAM-s	50.1	49.9	56.4	27.4	46.0
AAM-m	47.1	51.5	54.2	28.8	42.3

P, L and C denote the shape, texture and combined parameters, respectively
The highest values are emphasized in bold

4.3 Spatiotemporal local binary patterns

For our experiments we also utilize the popular **Spatiotemporal Local Binary Patterns (LBPTOP)** introduced in [30]. 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. 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 [30], we extract the LBPTOP 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).

4.4 Spatiotemporal histogram of oriented gradients

Another parametrization considered in the experiments is the **Spatiotemporal Histogram of Oriented Gradients (HOGTOP)**. Inspired by the LBPTOP, it has been proposed in [18] 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. HOGTOP also adds orientations from the xt and yt planes, processes them independently, concatenates, and reduces 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.

4.5 Spatiotemporal convolutional network

For the experiments we also utilize a spatiotemporal convolutional network that is learned to classify short video chunks $\boldsymbol{x}$ into one of 48 Czech phoneme classes including silences [19]. The chunks consist of 7 frames of 64 × 64 RGB region of

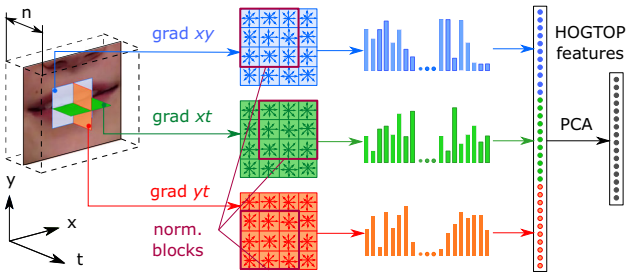

Fig. 3 Extraction of spatiotemporal histogram of oriented gradients

interest (ROI) that cover the speaker's mouth and its closest surroundings. Four blocks of spatiotemporal convolutions, batch normalization, spatiotemporal max-pooling and rectified linear unit are stacked, 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. 4 for the details. After the network is trained, we use its output vector $\boldsymbol{f}$, whose j-th element f_j 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. Wo do not cross-validate any hyperparameters of the network in the experiments.

We compare two versions of the spatiotemporal convolutional DNN features. In the first version, the network is only pre-trained for three epochs on an external dataset consisting of about 6.5 h of frame-labeled videos, see [19] for details. As for the second version, we also fine-tune the net on the TULAVD dataset for another three epochs using a ten times smaller learning rate. Note that the depth-based spatiotemporal network has only a single version – trained from scratch on the TULAVD dataset.

5 System overview

5.1 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 objec-

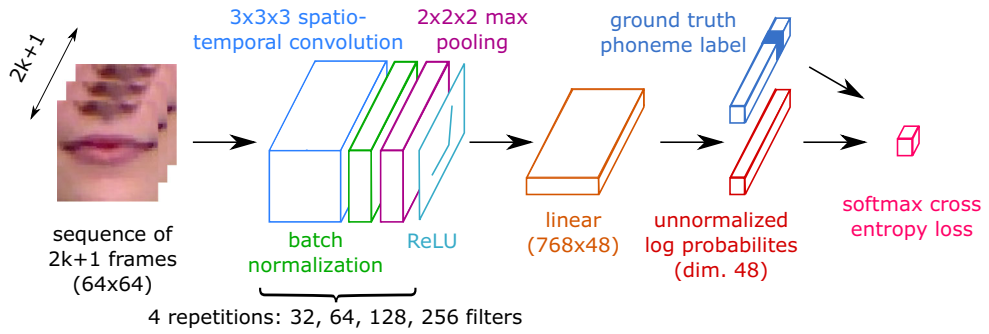

Fig. 4 Architecture of the spatiotemporal convolutional network

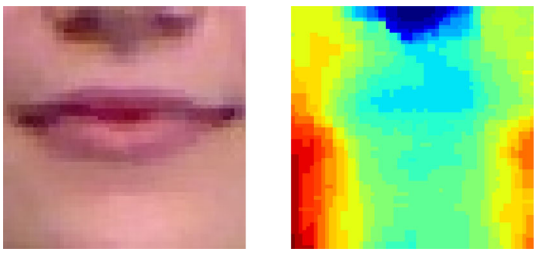

Fig. 5 Example of an interpolated depth map

tive 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. Figure 5 shows an example ROI extracted from the video and depth streams.

5.2 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. 4 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.

Table 2 Mapping between phonemes and visemes for the Czech language as proposed in [4]

Viseme	Phoneme (PAC-CZ)
A	a,
B	b, m, M, p
C	c, C, s, z
CH	č, Č, ř, Ř,
D	d, n, N, t
DJ	ď, j, ň, ť
E	e,
F	f, v
G	g, h, X, k
I	i,
L	l, r
O	o,
U	u,

5.3 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 [16] 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] as shown in Table 2.

Phonemes and visemes are modeled using a 3-state hidden Markov model (HMM) with Gaussian mixture (GMM) emission probability $p(x^{(s)}|q)$, where $x^{(s)}$ represent features of the stream s (i.e. audio, video, or depth) and q is the model state. Due to limited data and efficiency considerations, for the whole-word models we only consider single-component GMM, i.e. simple Gaussian distribution. In other cases, C is fixed to 8 or 16; see e.g. the Table 4. 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 $s = (1, \ldots, S)$:

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

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 most 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.

5.4 Language models

We evaluate our audiovisual recognition system for four different bigram language models with vocabulary size ranging from 366 up to 500k words, see Table 3 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. 3. Note that for the purpose of calculating the statistics, none of the TULAVD sentences are used. Inclusion of the words from the TULAVD corpus only ensures that the test data will not contain any previously unseen words. We employed the SRILM toolkit [26] with Knesser-Nay smoothing for the language model training.

Table 3 Vocabularies considered in the experiments

LM	tulavd	5k	50k	500k
# words	366	5182	50,056	499,993
# bigrams	48,338	9865 k	73,905 k	141,670 k

6 Experiments

6.1 Isolated word recognition

In order to tune the hyperparameters of the visual parametrizations described in Sect. 4, we first perform experiments with isolated word lipreading. For reasons of efficiency, these hyperparameters were optimized using 14-state whole-word models with a single component GMM (i.e. Gaussian distribution) that were trained only using isolated word data, i.e. without the continuous speech part of the dataset. The 54 speakers were split into 6 groups of 9 and we followed the k-fold cross-validation protocol, where $4 + 1$ groups constitute a training and validation sets and 1 is reserved for testing. To minimize data leakage, the validation rather than the test scores were maximized by the hyperparameter selection process in each respective fold of the cross validation. All the reported results are the average word accuracy (Wacc) achieved over the 6 different test sets.

The optimized parametrizations were then used for unimodal (single-stream) recognition of the 50 isolated words using phoneme and viseme models. These models were

Table 4 Word accuracy (%) of unimodal isolated word recognition and lipreading. Note that even in the case of DAAM, the model is still unimodal, since the combination is performed in the feature extraction stage

Param.:	Src.	Word	Phoneme		Viseme	
Mixtures		1	8	16	8	16
MFCC	a	99.8	99.5	99.8	97.4	98.0
DCT	v	67.4	42.6	42.8	42.4	43.9
	d	71.8	39.3	42.5	38.6	43.1
AAM	v	71.4	57.5	58.5	59.0	59.3
	d	73.0	54.1	55.0	55.3	56.6
DAAM	v ∘ d	73.9	62.0	64.6	63.0	64.7
LBPTOP	v	72.3	54.6	56.4	54.6	56.3
	d	64.6	48.7	47.4	45.3	48.2
DCT3	v	67.9	42.6	43.1	43.4	45.6
	d	65.9	45.1	47.0	45.4	47.6
HOGTOP	v	84.8	59.5	61.0	59.8	60.1
	d	84.9	56.6	58.3	56.6	57.7
DNN	v	89.1	54.1	53.5	54.3	55.4
DNN-tuned	v	**92.3**	**65.3**	**66.3**	**63.7**	**65.8**
	d	88.1	62.5	63.4	61.7	62.0

The highest values are emphasized in bold

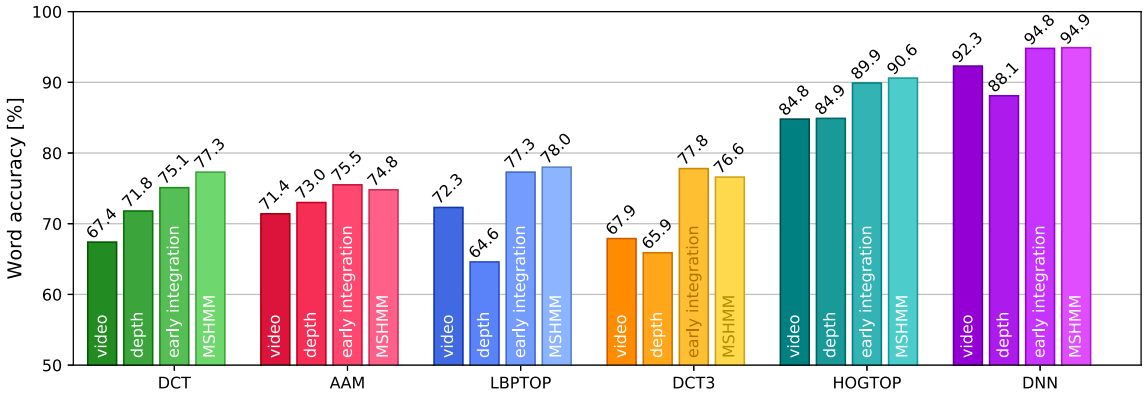

Fig. 6 Comparison of isolated word recognition using whole word models for video-only, depth-only, early integration, and middle fusion (MSHMM) models

learned on all the available training data from each respective fold of the cross validation, i.e. including continuous speech data, which amounts to approximately 5 h of spoken data on average. Table 4 summarizes the results of both whole-word and phonetic models. In these experiments we employed the HTK `HVite` decoder.

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 the simpler scenario with whole-based models, video and depth-based parametrizations perform roughly on par, with their combination in the form of DAAM achieving one of the best results overall. Contrary to our observation, Galatas et al.[7] achieved much worse results using the depth stream as compared to RGB. This discrepancy might be explained by the fact that we linearly interpolate the depth maps, as otherwise the random missing spots manifesting as regions of zeros adds too much of unpredictable variability.

The beneficial effect of adding the depth stream can also be demonstrated by comparing single-stream (video or depth) and two-stream (video and depth) results. Performance of isolated word recognition as displayed in Fig. 6 shows that for all considered features, combination of video and depth achieves higher accuracy then single-stream variant. In cases where the streams differ by their reliability, middle fusion takes advantage of setting the optimal stream weights and achieves higher scores than simple concatenation, i.e. early integration. This is also important for audio-visual recognition, where the disparity between relative importance of each channel much higher.

One of the crucial aspects determining the final accuracy seems to be whether or not the parametrization exploits speech dynamics. We can see that the highest rates were achieved by HOGTOP and DNN features that extract robust features from both spatial and temporal axes.

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 often target only isolated unit recognition, and the results do not necessarily apply to systems with larger vocabularies and/or recognition based on sub-word units.

6.2 Continuous speech

The results on isolated word lipreading show that on average viseme-based models do not outperform the phone-based ones. For the lower number of mixtures, the phoneme models even achieve a slightly higher average. This observation may be attributed to the viseme context dependency on the surrounding vowels [24]. 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 monophone models in the following experiments.

In order to minimize the number of sources of variability across different folds and to better control the vocabulary, the test data for the continuous speech recognition comprise only of the first 50 sentences that are common to all speakers instead of the full set of 100 sentences. Also, due to performance reasons we switched from `HVite` to the Julius[1] decoder, which is compatible with HTK model definitions.

Table 5 presents the achieved results. Note that a + v denotes a middle fusion of audio and video channels via MSHMM with optimally set weights $\lambda^{(s)}$ that are cross-

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

Table 5 Word accuracy (%) of audiovisual speech recognition with monophones by middle fusion of acoustic and visual parametrizations for different vocabularies

Par.	Source	Vocabulary			
		tulavd	5k	50k	500k
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
DAAM	a + v ∘ d	75.2	58.6	48.0	40.7
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
DNN	a + v	78.3	61.5	48.7	39.8
DNN-tuned	a + v	79.7	**64.8**	**52.4**	**44.3**
	a + d	78.7	62.5	50.1	42.0
	a + v + d	**80.0**	**64.8**	52.0	43.4

The highest values are emphasized in bold

Table 6 Word accuracy (Wacc) versus word correctness (Wcor) of purely visual continuous speech lipreading with monophones as a basic speech unit for the smallest (tulavd) vocabulary of 366 words

Par.	Video		Depth	
	Wacc (%)	Wcor (%)	Wacc (%)	Wcor (%)
DCT	− 17.3	9.3	− 2.6	6.3
AAM	12.3	16.2	9.3	12.4
LBPTOP	6.31	17.8	− 0.1	10.8
DCT3	− 17.3	10.0	− 8.3	7.9
HOGTOP	3.8	21.3	4.1	15.8
DNN-tuned	9.5	30.1	0.4	20.0

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 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 some improvement over acoustic-only recognition, showing that the visual cues provide useful information even for very large vocabularies with 500k words. This especially holds for the LBPTOP, HOGTOP, and DNN features, as they manage to exploit some of the speech dynamics, which is essential for phoneme discrimination. The best results overall were obtained by combination of MFCC and our proposed spatiotemporal DNN features extracted from video. Although the overall margins between different parametrizations are rather slight, they are consistent throughout all of the experiments, which suggests they are relevant and not just random flukes.

Contrary 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. The reason for this behavior will be subject to investigation in further research.

For all four vocabularies the highest improvement achieved over audio-only recognition ranged between 5–8% absolutely, i.e. 7–22% relatively. In most cases the optimal weight ratio of audio and video (or depth) channels, which indicates the relative importance of each modality, was 0.7:0.3 or 0.8:0.2, with the former being more common for the 500k 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.

We have also performed the above experiments utilizing only lipreading without any acoustic information, see Table 6 for results using the smallest vocabulary of 366 words. Not surprisingly, the results in form of the word accuracy were basically random even for such a small vocabulary, however by looking at the word correctness, which ignores insertion errors, the visual channel still seemed helpful as a word spotter. For example, Wacc for continuous speech lipreading with the smallest vocabulary using the DNN features was 9.5%, whereas the word correctness reached as high as 30.1%. Similar difference was observed for HOGTOP: 3.8 versus 21.3%. It therefore seems that one of the main sources of lipreading errors stems from word insertions, which might be corrected for by incorporating confidence into predictions and accordingly dynamically adapt the corresponding stream weights.

6.3 Noisy environments

We also performed experiments in simulated noisy environments. We additively mix white and babble noises from the NOISEX-92 dataset at various levels of signal-to-noise ratio (SNR). For each recording, the energies of speech and noise are calculated as average over the whole length, that is not adaptively. We fixed the weights of the acoustic and video (depth) to a 0.7:0.3 (or 0.7:0.2:0.1) ratio.

Figure 7 compares the results of continuous speech recognition on the TULAVD dataset with the smallest vocabulary of 366 words (tulavd) for audio, audio-video, and audio-video-depth middle fusions for the fined-tuned DNN features. As expected, the biggest improvements of up to 25% absolutely have been reached for lower SNRs. Similarly to recognition on the clean data, the depth stream did

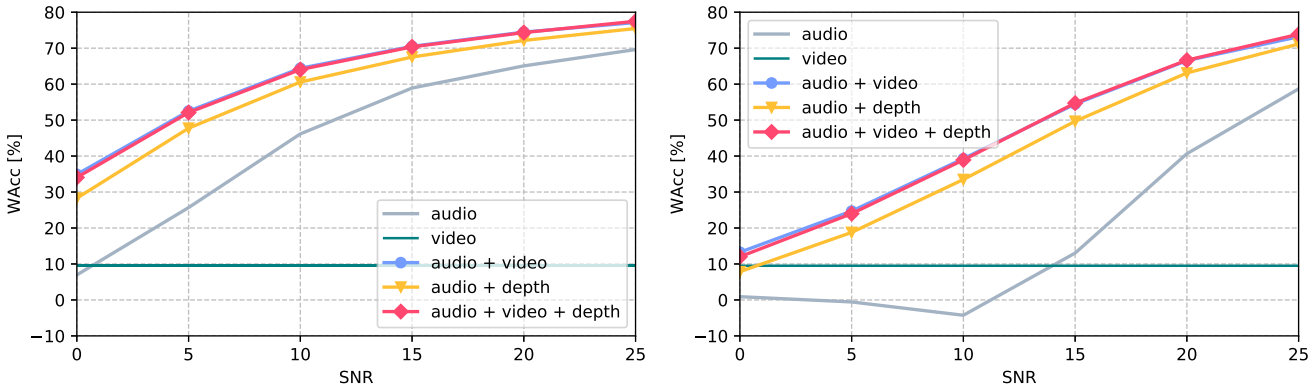

Fig. 7 Continuous speech recognition in noisy environments using the middle fusion of the MFCC and spatiotemporal convolutional features for the smallest vocabulary of 366 words. Left: babble noise, right: white noise

not improve the result over audio-video combination. One of the problems here is too coarse grid for various weight configurations, as the step was 0.1. It is possible that with finer weight search, one may find configuration that will slightly improve the results in the three modality settings.

7 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 spatiotemporal convolutional DNN, 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 22% when the channels were integrated via multi-stream hidden Markov model with optimally set weights. We have also shown that with careful preprocessing of the data, the depth maps can serve as an additional modality to RGB with only slightly lower overall performance with the deficiency seemingly manifesting only in continuous speech recognition.

There might be an issue that the observed improvement when adding visual component to the process could be influenced by the limited amount of acoustic data and it is uncertain if the same results would hold for more robust acoustic models trained on hundreds of hours of speech. In order to verify this, transfer learning techniques could potentially be employed to circumvent the lack of large audiovisual dataset availability.

References

1. Assael YM, Shillingford B, Whiteson S, de Freitas N (2016) Lipnet: sentence-level lipreading. In: CoRR abs/1611.01599
2. Cao X, Wei Y, Wen F, Sun J(2012) Face alignment by explicit shape regression. In: CVPR
3. Chung JS, Senior AW, Vinyals O, Zisserman A (2016) Lip reading sentences in the wild. In: CoRR
4. Císař P (2006) Application of lipreading methods for speech recognition. Ph.D. thesis
5. Cooke M, Barker J, Cunningham S, Shao X (2006) An audio-visual corpus for speech perception and automatic speech recognition. J Acoust Soc Am 120(5):2421–2424
6. Estellers V, Gurban M, Thiran J (2012) On dynamic stream weighting for audio-visual speech recognition. IEEE Trans Audio Speech Lang Process 20(4):1145–1157
7. Galatas G, Potamianos G, Makedon F (2012) Audio-visual speech recognition incorporating facial depth information captured by the kinect. In: Proceedings of the 20th European signal processing conference (EUSIPCO), pp 2714–2717
8. Glotin H, Vergyr D, Neti C, Potamianos G, Luettin J (2001) Weighting schemes for audio-visual fusion in speech recognition. In: 2001 IEEE international conference on acoustics, speech, and signal processing (ICASSP '01), vol 1, pp 173–176
9. Harte N, Gillen E (2015) Tcd-timit: an audio-visual corpus of continuous speech. IEEE Trans Multimed 17(5):603–615
10. Lan Y, Theobald B, Harvey R, Bowden R (2010) Improving visual features for lip-reading. In: Proceedings of the international conference on auditory-visual speech processing, 2010, pp 142–147
11. Lee B, Hasegawa-Johnson M, Goudeseune C, Kamdar S, Borys S, Liu M, Huang TS (2004) AVICAR: audio-visual speech corpus in a car environment. In: INTERSPEECH, pp 2489–2492
12. Lucey S, Chen T, Sridharan S, Chandran V (2005) Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition. IEEE Trans Multimed 7(3):495–506
13. McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748
14. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011)Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning, ICML 2011, Bellevue, Washington, USA, June 28-July 2, 2011, pp 689–696
15. Noda K, Yamaguchi Y, Nakadai K, Okuno, H, Ogata, T (2014) Lipreading using convolutional neural network. In: International speech and communication association, pp 1149–1153
16. Nouza, J, Psutka J, Uhlíř (1997) Phonetic alphabet for speech recognition of czech. Radioengineering 6(4):16–20
17. Ong E, Bowden, R (2011) Learning sequential patterns for lipreading. In: Proceedings of the British machine vision conference, BMVC 2011, Dundee, UK, August 29-September 2, 2011, pp 1–10

18. Palecek K (2016) Lipreading using spatiotemporal histogram of oriented gradients. In: EUSIPCO 2016, Budapest, Hungary, 2016, pp 1882–1885
19. Paleček K (2017) Spatiotemporal convolutional features for lipreading. Springer, Cham, pp 438–446
20. Paleček K (2017) Utilizing lipreading in large vocabulary continuous speech recognition. In: Karpov A, Potapova R, Mporas I (eds) Speech and computer. Springer, Cham, pp 767–776
21. Pei Y, Kim T, Zha H (2013) Unsupervised random forest manifold alignment for lipreading. In: IEEE international conference on computer vision, ICCV 2013, Sydney, Australia, December 1–8, 2013, pp 129–136
22. Petridis S, Li Z, Pantic M (2017) End-to-end visual speech recognition with LSTMS. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) pp 2592–2596
23. Potamianos G, Neti C, Gravier G, Garg A, Senior AW (2003) Recent advances in the automatic recognition of audio-visual speech. In: Proceedings of the IEEE, pp 1306–1326
24. Ramage MD (2013) Disproving visemes as the basic visual unit of speech. Ph.D. thesis
25. Saenko K, Livescu K, Siracusa M, Wilson K, Glass J, Darrell T (2005) Visual speech recognition with loosely synchronized feature streams. In: Proceedings of the tenth IEEE international conference on computer vision, ICCV '05, vol 2. IEEE Computer Society, Washington, DC, USA, pp 1424–1431
26. Stolcke A (2002) SRILM: an extensible language modeling toolkit. In: Proceedings of ICSLP, vol 2. Denver, USA, pp 901–904
27. Sui C, Bennamoun M, Togneri R (2016) Visual speech feature representations: recent advances. Springer, Cham, pp 377–396
28. Summerfield Q (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd B (ed) Hearing by eye: the psychology of lip-reading. Lawrence Erlbaum Associates, Hillsdale
29. Wand M, Koutník J Schmidhuber J (2016) Lipreading with long short-term memory. In: CoRR
30. Zhao G, Barnard M, Pietikäinen M (2009) Lipreading with local spatiotemporal descriptors. IEEE Trans Multimed 11(7):1254–1265
31. Zhou Z, Zhao G, Hong X, Pietikinen M (2014) A review of recent advances in visual speech decoding. Image Vis Comput 32(9):590–605
32. Zhou Z, Zhao G, Pietikainen M (2011)Towards a practical lipreading system. In: Proceedings of the 2011 IEEE conference on computer vision and pattern recognition, CVPR '11. IEEE Computer Society, Washington, DC, USA, pp 137–144