The 14th International Conference on Auditory-Visual Speech Processing
25-26 August 2017, Stockholm, Sweden

Combining Multiple Views for Visual Speech Recognition

Marina Zimmermann[1], *Mostafa Mehdipour Ghazi*[2], *Hazım Kemal Ekenel*[3], *Jean-Philippe Thiran*[1]

[1] Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[2] Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey
[3] Department of Computer Engineering, Istanbul Technical University (ITU), Istanbul, Turkey

marina.zimmermann@epfl.ch, mehdipour@sabanciuniv.edu, ekenel@itu.edu.tr, jean-philippe.thiran@epfl.ch

Abstract

Visual speech recognition is a challenging research problem with a particular practical application of aiding audio speech recognition in noisy scenarios. Multiple camera setups can be beneficial for the visual speech recognition systems in terms of improved performance and robustness. In this paper, we explore this aspect and provide a comprehensive study on combining multiple views for visual speech recognition. The thorough analysis covers fusion of all possible view angle combinations both at feature level and decision level. The employed visual speech recognition system in this study extracts features through a PCA-based convolutional neural network, followed by an LSTM network. Finally, these features are processed in a tandem system, being fed into a GMM-HMM scheme. The decision fusion acts after this point by combining the Viterbi path log-likelihoods. The results show that the complementary information contained in recordings from different view angles improves the results significantly. For example, the sentence correctness on the test set is increased from 76% for the highest performing single view (30°) to up to 83% when combining this view with the frontal and 60° view angles.

Index Terms: visual speech recognition, automatic lip reading, PCA network, multi-view processing

1. Introduction

Visual speech recognition, or automatic lip-reading, has been a topic of research for several decades and is the subject of a renewed research interest in recent years with diverse applications ranging from audio-visual speech recognition [1] to pure visual speech recognition (VSR), cybersecurity [2], mouthings in sign language [3], speech production [4] or general human machine interfaces.

Pure audio-based speech recognition has seen significant improvements over the last decades and has been tested on and applied to many real-life datasets and scenarios. However, visual speech recognition still focuses mainly on in-lab databases. To overcome some of the shortcomings that need to be addressed to work on real-world data, several studies have focused on VSR for various head poses [5, 6]. Moreover, a few databases have included recorded sentences from cameras at various view angles [7, 8]. Following up with this perspective, the current work continues the use of the recent OuluVS2 dataset [9] which includes simultaneous recordings from cameras placed at five different angles.

Recent studies on visual speech recognition have seen VSR moving further and further towards deep learning, a technique widely employed in audio speech recognition as well as computer vision. Deep neural networks (DNNs) have set a new baseline in speech recognition [10] and are now the standard in computer vision tasks [11, 12]. Since one of the main requirements for deep learning is having a huge amount of data, the need for larger databases is increasing. There exist several such datasets publicly available to the research community such as the BBC-Oxford 'Lip Reading in the Wild' (LRW) Dataset [13] and the GRID audio-visual sentence corpus [14] containing up to 1000 repetitions of 500 different words and sentences based on an artificial grammar, respectively. These databases have been widely used in the recent VSR literature using deep learning. However, a few other fairly large audio-visual databases have been developed and published that contain more diverse sentences and various view angles, namely TCD-TIMIT and OuluVS2 [8, 9]. However, the size of these databases still remains rather small compared to audio-only databases or image datasets.

In this paper, we employ a PCA-based convolutional network [12] in combination with long short-term memory (LSTM) cells and a GMM-HMM scheme to model temporal evolution between words [15]. This work focuses on the combination possibilities between various view angles and the complementary information that can thus be exploited. We show that different views indeed complement each other and thus produce better overall sentences recognition results of up to 83% for the combination of the frontal, 30°, 60° and 90° side views. In general, combinations with views such as the 30° and 60° views showed good improvements, highlighting the complementary nature of the information between various view angles. On the contrary, although the 90° pose does not seem to contain as many relevant features to correctly recognise a sentence on its own, combined with other views, especially other lower performing angles such as 45°, it improves recognition rates. This observation implies that there exists a certain amount of complementary information between these views as well.

The contributions and main outcomes of this paper can be listed as below:

- A thorough analysis on fusion of all possible view angle combinations in a multiple camera setup is provided and their influence on the performance is assessed. It has been found that combining multiple views results in significant performance improvement.
- Feature level and decision level fusion is compared. In order to facilitate decision level fusion, Viterbi path log-likelihoods are extracted from a GMM-HMM framework. For decision fusion, two different weighting schemes are employed. Decision fusion has been found to be more useful than feature fusion.
- Different view angles contributions to the performance

10.21437/AVSP.2017-10

are analysed. 0° and 30° are found to be more useful for visual speech recognition.

The rest of this paper is organized as follows. Section 2 gives a short overview of the state-of-the-art approaches. Section 3 reviews the proposed method for VSR utilising a PCA network, LSTMs, and the GMM-HMM system and introduces the decision fusion scheme. Section 4 presents the dataset, experiments, and obtained results. Finally, Section 5 concludes the paper with a summary and discussions.

2. Related work

Feature extraction for visual speech recognition is usually split into two steps: first a region of interest (ROI) is extracted [1], typically around the mouth which contains most of the visual information for an utterance, then specific features are computed based on this ROI. The current trend for ROI extraction is based on using a face tracker and occasionally a face model, even though some work might still use manual annotations or corrections.

Three different types of features are generally extracted from the ROI: texture-based features, shape-based features, or a combination of these two [1, 16]. Texture-based features are computed directly on the pixel values of the ROI. Traditionally, transformations such as the discrete cosine transform (DCT), possibly combined with dimensionality reduction techniques like the linear discriminant analysis (LDA), principle component analysis (PCA), or a maximum-likelihood transform (MLLT) have been employed [1].

Shape-based features attempt to take into account the actual shape of the mouth by extracting the contours or computing geometrical distances between certain points of interest around the mouth [1]. Nowadays, many studies that make use of these types of features directly extract these points and shapes with a face tracker and apply PCA to them, as performed by active appearance models (AAMs) [16].

These features then need to be processed in a classification scheme. Traditionally, this is performed by a combination of Gaussian mixture models (GMMs) and hidden Markov models (HMMs), where the acoustics (the phonemes or visemes) are modelled by the GMMs and the HMM describes the time evolution with states modelling the phonemes and the larger scale transitions within and between words [1].

Recent VSR research has slowly moved away from these traditional approaches and has replaced these by deep learning based approaches which consistently show higher performance. The first step involved the use of autoencoders either for feature extraction using, for example, deep Boltzmann machines [17], convolutional neural networks (CNNs) [18, 19] or deep belief networks (DBNs) [20], or for feature post-processing [21]. Some of these then either use a simple classification system like support vector machines (SVM) for normalised-length utterances [17], or pass the features into a so-called tandem system [22] made up of a GMM-HMM recogniser [20, 18, 21]. Another way of combining neural networks with HMMs is the hybrid approach [23]. This method passes the posterior probability outputs of the neural network directly as inputs to the HMM, using the network as an acoustic model. Finally, more recent work replaces the recognition system by a bilinear network [24] or substitutes the whole recognition pipeline by recurrent neural networks such as LSTMs [25, 26, 27].

Varying head poses and, therefore, the combination of various view angles are very important topics in VSR. Early literature on the topic addressed this particularly with the goal of using a single view for training and for testing each. This led to research on cross-view training/testing, i.e. training on one and testing on another view, as well as projections of images or features from one view angle to another one [5, 6, 28]. [28] also establishes that the 30° view angle provides the best recognition performance, even over the frontal view. In [29] a synchronous HMM was built to include four different views (centre left, centre right, side left, side right). The weights for this multi-stream HMM are determined empirically by comparing the training performance between the centre and the side views and the left and right views for varying weights. The final individual weights are a combination of these coarser weights. Finally, some recent research has applied cross-view analysis to 3D-AAMs [30] and used channel, image and feature fusion for multiple- and cross-view analysis [31].

Our proposed method builds upon previous work making use of a PCA network and a subsequent LSTM network inside a tandem GMM-HMM scheme [15]. We explore the influence of multi-view fusion on the recognition results, showing that using more than one view angle for visual speech recognition is indeed beneficially, since it allows to exploit the complementary nature of the features contained in the different camera views. Unlike previous work, where either multi-stream HMMs or feature fusion are employed, in our method, the decision fusion takes place at the end of the recognition pipeline by weighting the log-likelihoods of the paths of the Viterbi algorithm for several views.

3. The proposed method

In this work we use a PCA network and LSTM framework developed in [15] to extract robust features for visual speech recognition. The two-stage PCA network is a type of convolutional neural network where the filters are replaced by the first eight eigenvectors obtained through a PCA on concatenations of normalised square patches of the ROI. After passing through two stages of these filters, the outputs are binarized, stacked and their block-wise histograms are calculated.

The output of the PCA network is then processed in a layer of LSTMs. These cells have the possibility to accept new input values or to forget the previous values and output values depending on the activation level of the cells' input, forget and output gates [32].

Finally, the logarithm of the LSTM output, together with its delta and acceleration components, is processed as spatiotemporal feature input in a GMM-HMM system in the Hidden Markov Model Toolkit (HTK) [33]. This tandem system is made up of 15 Gaussian mixtures per observation and uses four states per word, rather than modelling visemes separately due to the rather small amount of training data.

This study extends the previous work by offering further analyses of results using decision fusion techniques. The following fusion scheme of the likelihoods for two views v_a and v_b is used

$$p(o_{v_a}, o_{v_b}|q = q_i) = p(o_{v_a}|q = q_i)^{\lambda_{v_a}} p(o_{v_b}|q = q_i)^{\lambda_{v_b}} \tag{1}$$

where the weights λ_{v_a} and λ_{v_b} are constrained by

$$\lambda_{v_a} + \lambda_{v_b} = 1 \tag{2}$$

Taking the logarithm of equation 1 to obtain the log-

likelihoods like in HTK's Viterbi algorithm we have

$$\log(p(o_{v_a}, o_{v_b}|q = q_i)) = \lambda_{v_a} \log(p(o_{v_a}|q = q_i)) + \lambda_{v_b} \log(p(o_{v_b}|q = q_i)) \quad (3)$$

To finally fuse the results, the top-5 Viterbi output sequences for each view and utterance are retained together with their log-likelihoods. The weighted log-likelihoods of the two views are summed up and the Viterbi sequence with the highest weighted sum is then selected to obtain the final joint log-likelihood. This approach was similarly extended to multiple views by summing up the weighted log-likelihoods and restricting the sum of their weights to one.

The optimal weights were obtained through two different approaches. The first method determines the weighting based on the recognition performance during training. The second method performs a grid search and applies all possible weights on a cross-validation of the training set, then choosing the weights with the best recognition results.

Feature fusion, on the contrary, is the simple concatenation of the spatiotemporal features for several view angles at the output of the LSTMs which are then fed into a single GMM-HMM system [15].

4. Performance analysis

This section presents the data evaluated in this study, the experiments conducted, and the results obtained when performing decision fusion across multiple views.

4.1. The dataset

The dataset used in this paper is the OuluVS2 database [9]. It consists of audio recordings and videos of 52 subjects pronouncing various English sentences from 5 different view angles: 0° (frontal), 30°, 45°, 60° and 90° (profile). The videos were recorded at a resolution of 1920×1080 pixels at a framerate of 30 fps in an ordinary office environment with varying lighting conditions.

In this work, the analysis was restricted to videos of the short phrase section of the database, containing three repetitions of 10 sentences, such as "Excuse me" and "How are you", per subject. Furthermore, these videos were pre-cropped to the mouth region by the authors of the dataset, and the training set was designed so as to contain videos of 40 out of the total 52 subjects.

4.2. Experimental results

Parameters for the spatiotemporal feature extraction by PCA network and LSTM from the given cropped mouth videos were selected similar to the approach in [15]. Subsequent experiments were conducted to measure the improvements in performance when combining classifier outputs. These results are compared to the previous recognition rates for single-view and multiple-view experiments through feature concatenation.

Three measures of word accuracy, word correctness, and sentence recognition per cent are used for reporting the results:

$$\text{Accuracy} = \frac{H - I}{N} \cdot 100\% \quad (4)$$

$$\text{Correctness} = \frac{H}{N} \cdot 100\% \quad (5)$$

where N is the total number of words, I, D, and S are the number of inserted, deleted, and substituted words, respectively, and $H = N - D - S$ represents the number of correct words.

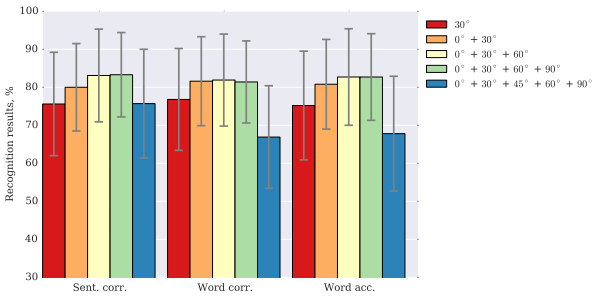

Figure 1: *Best mean phrase recognition results (in %) across test subjects for each number of views per combination of the OuluVS2 multi-view dataset using the proposed method.*

To adjust our system parameters, we use a leave-one-out cross-validation scheme on the given training set. Next, we apply the trained system on the test set for the final recognition at the word or phrase levels. We present the best test results obtained for phrase recognition on the OuluVS2 dataset for varying numbers of views combined in Figure 1, and the complete test results in Table 2.

4.2.1. Weighting schemes

The aim of these experiments is to investigate whether the complementary information contained in the different views can be exploited by combining the top-5 Viterbi decoder HMM outputs for several angles. To do so, we summed the weighted log-likelihoods for two views to determine which path would have the highest combined log-likelihood.

The optimal weights (see Table 1) were obtained through a leave-one-out cross-validation scheme applied only on the training set using the train-test splits of the data as provided by the authors of the dataset. These weights are determined in two ways. In the first method, the weights are based on the sentence correctness of a leave-one-out cross-validation in the training set and are later normalised by the total sum of weights. This measure is referred to as "Training recognition" or "Rec" in Tables 1 and 2. In the second approach, optimal weights were obtained by iterating over all possible values between 0 and 1 in 0.1 steps and choosing the highest performing one on the cross-validation of the training set, while observing the constraint of equation 2. These weights are referred to as "Grid search" or "Grid" in Tables 1 and 2. Afterwards, these weights were used on the test set.

Evaluating the weights obtained for various cases in Table 1, we can already see that for grid search the weights tend to be higher especially for the frontal and 30° side view. This does not seem very surprising, since these views also have the highest performance when taken separately. However, since the best weights are determined based on the results of the cross-validation of the training set, there are some variations on the performance of the test set. For example, the optimal combination of all views results in non-zero weights only for the 0° and 30° view angles. In this case, another weight combination is optimal on the training set than for the combination of these two views alone, so that the final test results for all views are poorer than for the weighting of those two views alone.

The weighting scheme based on the training recognition results shows very balanced weights, since the training results are fairly similar so that rounding the weights to one digit after the

Table 1: *Optimal weights obtained via grid search and by training performance normalisation*

View combination	Grid search					Training recognition				
$v_a + v_b + v_c + v_d + v_e$	λ_{v_a}	λ_{v_b}	λ_{v_c}	λ_{v_d}	λ_{v_e}	λ_{v_a}	λ_{v_b}	λ_{v_c}	λ_{v_d}	λ_{v_e}
0°	-	-	-	-	-	-	-	-	-	-
30°	-	-	-	-	-	-	-	-	-	-
45°	-	-	-	-	-	-	-	-	-	-
60°	-	-	-	-	-	-	-	-	-	-
90°	-	-	-	-	-	-	-	-	-	-
$0^\circ + 30^\circ$	0.4	0.6	-	-	-	0.5	0.5	-	-	-
$0^\circ + 45^\circ$	0.6	0.4	-	-	-	0.6	0.4	-	-	-
$0^\circ + 60^\circ$	0.9	0.1	-	-	-	0.5	0.5	-	-	-
$0^\circ + 90^\circ$	0.7	0.3	-	-	-	0.6	0.4	-	-	-
$30^\circ + 45^\circ$	0.8	0.2	-	-	-	0.6	0.4	-	-	-
$30^\circ + 60^\circ$	0.6	0.4	-	-	-	0.5	0.5	-	-	-
$30^\circ + 90^\circ$	0.9	0.1	-	-	-	0.6	0.4	-	-	-
$45^\circ + 60^\circ$	0.4	0.6	-	-	-	0.5	0.5	-	-	-
$45^\circ + 90^\circ$	0.8	0.2	-	-	-	0.5	0.5	-	-	-
$60^\circ + 90^\circ$	0.7	0.3	-	-	-	0.5	0.5	-	-	-
$0^\circ + 30^\circ + 45^\circ$	0.4	0.5	0.1	-	-	0.4	0.4	0.2	-	-
$0^\circ + 30^\circ + 60^\circ$	0.4	0.4	0.2	-	-	0.3	0.3	0.4	-	-
$0^\circ + 30^\circ + 90^\circ$	0.4	0.6	0.0	-	-	0.4	0.4	0.2	-	-
$0^\circ + 45^\circ + 60^\circ$	0.8	0.1	0.1	-	-	0.4	0.3	0.3	-	-
$0^\circ + 45^\circ + 90^\circ$	0.8	0.1	0.1	-	-	0.4	0.3	0.3	-	-
$0^\circ + 60^\circ + 90^\circ$	0.8	0.1	0.1	-	-	0.4	0.3	0.3	-	-
$30^\circ + 45^\circ + 60^\circ$	0.6	0.0	0.4	-	-	0.4	0.3	0.3	-	-
$30^\circ + 45^\circ + 90^\circ$	0.8	0.1	0.1	-	-	0.4	0.3	0.3	-	-
$30^\circ + 60^\circ + 90^\circ$	0.8	0.1	0.1	-	-	0.4	0.3	0.3	-	-
$45^\circ + 60^\circ + 90^\circ$	0.1	0.8	0.1	-	-	0.3	0.4	0.3	-	-
$0^\circ + 30^\circ + 45^\circ + 60^\circ$	0.3	0.4	0.1	0.2	-	0.3	0.3	0.2	0.2	-
$0^\circ + 30^\circ + 45^\circ + 90^\circ$	0.4	0.4	0.1	0.1	-	0.3	0.3	0.2	0.2	-
$0^\circ + 30^\circ + 60^\circ + 90^\circ$	0.4	0.4	0.1	0.1	-	0.3	0.3	0.2	0.2	-
$0^\circ + 45^\circ + 60^\circ + 90^\circ$	0.8	0.1	0.0	0.1	-	0.3	0.2	0.3	0.2	-
$30^\circ + 45^\circ + 60^\circ + 90^\circ$	0.7	0.1	0.1	0.1	-	0.3	0.2	0.3	0.2	-
$0^\circ + 30^\circ + 45^\circ + 60^\circ + 90^\circ$	0.9	0.1	0.0	0.0	0.0	0.2	0.2	0.2	0.2	0.2

comma results in very close values. Therefore, for the same example of the combination of all views, all views are weighted equally. In the end, the test results in sentence correctness are similar to the other weighting scheme, however, word accuracy and correctness are much lower.

4.2.2. Multiple-view experiments

Table 2 shows the results of various combinations on the test set. It contains the results obtained with the two weighting schemes, as well as the baseline results from [15] under "Feat". These are the single view test results and the combinations through feature concatenation. When evaluating the results, it should be taken into account that we do not perform a simple 10-class classification, but rather make use of a typical speech evolution process with HMMs, modelling word sequences. This includes silence as a possible utterance, which is only removed for evaluation purposes, in order not to distort the results.

Looking at the results in Table 2 we can see that there are various improvements over the baseline results for the 30°-side view (the highest performing single view). While this view on its own has a sentence correctness of around 76% on the test set, we can see that both through feature concatenation, and by combining classifier results with the frontal view we can achieve a sentence correctness of around 80%. For other types of combinations, the impact of combining classifiers over features becomes more apparent: all the combinations involving either the frontal or the 30° side view achieve a sentence correctness of at least 76%, while most of the feature concatenation schemes stay around 70%. Similar trends are also observed for the word accuracy and word correctness.

It is interesting to note that, aside from the frontal and 30° views, the combination of the 30° and 60° side views give particularly nice improvements, showing the degree of complementarity between these views. In general, it is evident that the frontal, 30° and 60° side views contain the most information and only some complementary information can be exploited in the 45° and 90° views.

Comparing the combinations of more than just two views, we can see that they improve the results further. This is true for almost all combinations, but especially in combining the frontal and 30° views with the 60° we can reach sentence recognition results of 83%. A slightly better correctness is still achieved when in addition using the 90° side view as well.

When finally combining all views, we see a drop in performance again. This is probably due to the effect described regarding the weighting procedure: since the weighting scheme is determined on the cross-validation of the training set, certain

Table 2: *Mean phrase recognition results (in %) across test subjects on the combination of different views of the OuluVS2 multi-view dataset using the proposed method*

View combination	Sentence Correctness			Word Accuracy			Word Correctness		
	Grid	Rec	Feat	Grid	Rec	Feat	Grid	Rec	Feat
0°	-	-	73.1	-	-	73.0	-	-	74.1
30°	-	-	75.6	-	-	75.2	-	-	76.8
45°	-	-	67.2	-	-	66.6	-	-	68.7
60°	-	-	63.3	-	-	60.6	-	-	63.7
90°	-	-	59.2	-	-	56.8	-	-	63.1
0° + 30°	79.4	80.0	79.2	79.9	80.8	82.9	80.6	81.6	80.9
0° + 45°	76.1	76.1	72.8	76.4	76.4	73.9	77.3	77.3	72.4
0° + 60°	77.2	74.7	72.2	77.9	74.6	73.1	78.9	75.7	71.9
0° + 90°	75.6	76.4	69.7	76.4	76.9	72.7	77.7	77.8	70.4
30° + 45°	77.2	76.9	-	77.4	77.2	-	78.8	78.6	-
30° + 60°	78.1	74.7	-	77.7	73.8	-	79.0	75.7	-
30° + 90°	76.7	76.9	-	77.7	77.9	-	79.2	79.4	-
45° + 60°	69.7	72.2	-	67.6	70.4	-	69.8	72.6	-
45° + 90°	72.5	71.9	-	71.9	71.3	-	73.9	73.8	-
60° + 90°	66.7	67.5	-	65.8	66.2	-	68.4	69.9	-
0° + 30° + 45°	82.3	80.4	-	81.3	79.6	-	81.1	79.7	-
0° + 30° + 60°	82.0	83.1	-	81.4	82.7	-	80.6	81.9	-
0° + 30° + 90°	80.6	82.3	-	79.9	81.3	-	79.4	80.3	-
0° + 45° + 60°	80.9	79.8	-	80.1	79.0	-	79.4	78.9	-
0° + 45° + 90°	78.0	79.4	-	77.2	78.4	-	76.7	78.1	-
0° + 60° + 90°	78.3	78.2	-	77.3	77.0	-	76.1	76.1	-
30° + 45° + 60°	79.0	78.3	-	77.7	77.1	-	78.1	77.2	-
30° + 45° + 90°	80.1	79.9	-	78.8	78.9	-	78.1	78.1	-
30° + 60° + 90°	80.1	79.0	-	78.8	77.4	-	78.1	76.9	-
45° + 60° + 90°	70.6	72.7	-	68.1	70.8	-	69.2	71.4	-
0° + 30° + 45° + 60°	82.7	81.1	-	82.1	80.2	-	81.7	79.4	-
0° + 30° + 45° + 90°	82.7	82.7	-	81.7	81.6	-	80.6	80.3	-
0° + 30° + 60° + 90°	82.1	83.3	-	81.4	82.7	-	80.3	81.4	-
0° + 45° + 60° + 90°	78.0	80.2	-	77.2	79.2	-	76.7	78.1	-
30° + 45° + 60° + 90°	80.0	78.3	-	78.7	76.4	-	77.5	76.1	-
0° + 30° + 45° + 60° + 90°	75.0	75.7	65.0	72.6	67.8	66.4	72.8	66.9	64.2

seemingly similar combinations result in different weights and thus a very different final performance on the test set.

The above discussions only take into account the sentence correctness. However, similar trends can be observed looking at both word accuracy and word correctness.

5. Conclusions

In this paper, we have explored the influence of multi-view fusion on visual speech recognition. The results have shown that exploiting multiple-view data can improve the recognition results significantly. This is particularly true for combinations involving the frontal view, the 30° and 60° view angles. The other views do provide additional information, however, the improvements are not as noteworthy.

In this work we have extended our previous experiments to include a more in-depth study of various combinations of different view angles. However, this study still has several limitations: First, it only takes into account a simple decision fusion scheme of the log-likelihoods of various Viterbi paths. Furthermore, the dataset is limited to simple phrase recognition. Future work should further extend this effort to test further fusion schemes and to evaluate them on larger databases.

6. References

[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-Visual Automatic Speech Recognition : An Overview," in *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds. MIT Press, 2004, ch. 10, pp. 1–30.

[2] A. Hassanat, "Visual Passwords Using Automatic Lip Reading," *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, vol. 13, no. 1, 2014.

[3] C. Schmidt and O. Koller, "Using viseme recognition to improve a sign language translation system," in *International Workshop on Spoken Language Translation*, Heidelberg, Germany, 2013, pp. 197–203.

[4] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue lips and face, based on MRI and video images," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, jul 2002.

[5] P. Lucey, G. Potamianos, and S. Sridharan, "An Extended Pose-Invariant Lipreading System," in *Proceedings of AVSP'07: International Conference on Auditory-Visual Speech Processing*. International Speech Communication Association, 2007.

[6] V. Estellers and J.-P. Thiran, "Multi-pose lipreading and audio-visual speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 51, 2012.

[7] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR : Audio-Visual Speech Corpus in a Car Environment," in *8th International Conference on Spoken Language Processing*, 2004.

[8] N. Harte and E. Gillen, "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 603–615, may 2015.

[9] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Institute of Electrical & Electronics Engineers (IEEE), may 2015.

[10] A. Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, T. Jebara and E. P. Xing, Eds. JMLR Workshop and Conference Proceedings, 2014, pp. 1764–1772.

[11] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Institute of Electrical & Electronics Engineers (IEEE), jun 2015.

[12] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "PCANet: A Simple Deep Learning Baseline for Image Classification?" *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, dec 2015.

[13] J. S. Chung and A. Zisserman, "Out of Time: Automated Lip Sync in the Wild," in *Computer Vision – ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, 2017, no. 1, pp. 251–263.

[14] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition." *The Journal of the Acoustical Society of America*, vol. 120, no. 5 Pt 1, pp. 2421–2424, 2006.

[15] M. Zimmermann, M. Mehdipour Ghazi, H. K. Ekenel, and J.-P. Thiran, "Visual Speech Recognition Using PCA Networks and LSTMs in a Tandem GMM-HMM System," in *Computer Vision – ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, 2017, no. 1, pp. 264–276.

[16] R. Bowden, S. Cox, R. Harvey, Y. Lan, E.-J. Ong, G. Owen, and B.-J. Theobald, "Recent developments in automated lip-reading," in *Optics and Photonics for Counterterrorism Crime Fighting and Defence IX and Optical Materials and Biomaterials in Security and Defence Systems Technology X*, R. Zamboni, F. Kajzar, A. A. Szep, D. Burgess, and G. Owen, Eds. SPIE-Intl Soc Optical Eng, oct 2013.

[17] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," *Proceedings of The 28th International Conference on Machine Learning (ICML)*, pp. 689–696, 2011.

[18] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, dec 2014.

[19] O. Koller, H. Ney, and R. Bowden, "Deep Learning of Mouth Shapes for Sign Language," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Institute of Electrical & Electronics Engineers (IEEE), dec 2015.

[20] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *2013 IEEE International Conference on Acoustics Speech and Signal Processing*. Institute of Electrical & Electronics Engineers (IEEE), may 2013.

[21] C. Sui, M. Bennamoun, and R. Togneri, "Listening with Your Eyes: Towards a Practical Visual Speech Recognition System Using Deep Boltzmann Machines," in *2015 IEEE International Conference on Computer Vision (ICCV)*. Institute of Electrical & Electronics Engineers (IEEE), dec 2015.

[22] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *2000 IEEE International Conference on Acoustics Speech, and Signal Processing. Proceedings*. Institute of Electrical & Electronics Engineers (IEEE), 2000.

[23] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition*. Springer Nature, 1994.

[24] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for Audio-Visual Speech Recognition," in *2015 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Institute of Electrical & Electronics Engineers (IEEE), apr 2015.

[25] M. Wand, J. Koutnik, and J. Schmidhuber, "Lipreading with long short-term memory," in *2016 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Institute of Electrical & Electronics Engineers (IEEE), mar 2016.

[26] J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *CoRR*, vol. abs/1611.05358, 2016. [Online]. Available: http://arxiv.org/abs/1611.05358

[27] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with lstms," *CoRR*, vol. abs/1701.05847, 2017. [Online]. Available: http://arxiv.org/abs/1701.05847

[28] Y. Lan, B.-J. Theobald, and R. W. Harvey, "View independent computer lip-reading," in *Proceedings - IEEE International Conference on Multimedia and Expo*, 2012, pp. 432–437.

[29] R. Navarathna, D. Dean, S. Sridharan, and P. Lucey, "Multiple cameras for audio-visual speech recognition in an automotive environment," *Computer Speech & Language*, vol. 27, no. 4, pp. 911–927, jun 2013.

[30] T. Watanabe, K. Katsurada, and Y. Kanazawa, "Lip Reading from Multi View Facial Images Using 3D-AAM," in *Computer Vision – ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, 2017, vol. 1, pp. 303–316.

[31] D. Lee, J. Lee, and K.-e. Kim, "Multi-view Automatic Lip-Reading Using Neural Network," in *Computer Vision – ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II*, 2017, vol. 1, pp. 290–302.

[32] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics Speech and Signal Processing*. Institute of Electrical & Electronics Engineers (IEEE), may 2013.

[33] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book," Tech. Rep., 2002.