

An Extended Pose-Invariant Lipreading System

Patrick Lucey,¹ Gerasimos Potamianos,² Sridha Sridharan¹

¹ Speech, Audio, Image and Video Technology Laboratory,
Queensland University of Technology, Brisbane, Australia

² IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

p.lucey@qut.edu.au, gpotam@us.ibm.com, s.sridharan@qut.edu.au

Abstract

In recent work, we have concentrated on the problem of lipreading from non-frontal views (poses). In particular, we have focused on the use of profile views, and proposed two approaches for lipreading on basis of visual features extracted from such views: (a) Direct statistical modeling of the features, namely use of view-dependent statistical models; and (b) Normalization of such features by their projection onto the “space” of frontal-view visual features, which allows employing one set of statistical models for all available views. The latter approach has been considered for two only poses (frontal and profile views), and for visual features of a specific dimensionality. In this paper, we further extend this work, by investigating its applicability to the case where data from three views are available (frontal, left- and right-profile). In addition, we examine the effect of visual feature dimensionality on the pose-normalization approach. Our experiments demonstrate that results generalize well to three views, but also that feature dimensionality is crucial to the effectiveness of the approach. In particular, feature dimensionality larger than 30 is detrimental to multi-pose visual speech recognition performance.

Index Terms: Audio-visual automatic speech recognition (AVASR), pose invariance, profile and frontal views, lipreading

1. Introduction

In the past decade, significant progress has been achieved in the area of audio-visual automatic speech recognition (AVASR) [1]. However, practical deployment of AVASR systems has yet to emerge. This is mainly due to the fact that most research on the subject has neglected addressing robustness of the AVASR visual front end component to variations such as head pose (view-point). Indeed, nearly all work has concentrated on the case where the speaker’s face is captured in a fully frontal pose – a rather restrictive human-computer interaction scenario. This has also been dictated by the lack of large corpora that allow addressing pose / view-point effects on lipreading performance. Recently however, interest in the subject has been increasing, especially with work focusing on meetings and lectures inside smart rooms [2, 3]. One such effort has been taking place within the framework of integrated project CHIL, “Computers in the Human Interaction Loop” [4]. As part of work in this project, we have collected an audio-visual database that contains synchronized multi-view videos of subjects, and is suitable for research on lipreading from non-frontal views [5, 6].

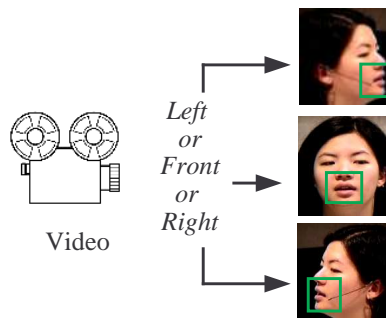


Figure 1: The developed lipreading system in this paper is able to recognize visual speech from either frontal, left profile, or right profile poses using a single classifier.

In previous work based on this database [5], experiments were constrained to each view-point having its own dedicated lipreading system (i.e. two separate systems were developed, one dedicated for frontal views and another for the profile). A different approach has been considered in our follow-up work on the subject [6]. There, we have tried to make our AVASR system more “real-world”, by having one unified lipreading system using a single camera, but allowing it to lipread from both frontal and right profile views using a single statistical model. In this paper, we extend this work by also including the left profile view as an additional view-point (see also Fig. 1).

The implications of such a system are significant for practical AVASR deployment. By loosening the constraint on the speaker’s pose, we allow a more pervasive or “real-world” technology to develop, which would be of major benefit to in-vehicle AVASR, for example. However, by allowing more flexibility in the system, we also introduce more complexity. As already suggested, a possible solution to this would be to model and recognize each view independently of each other, thus minimizing the train/test mismatch; this approach has been followed in [5]. Unfortunately, this is complicated to achieve in a continuous pose setting. A “one model for all” approach could be much more viable. However, having one model which can generalize over all views is also problematic, as it may “over-generalize”, causing large train/test mismatch. This over-generalization can be particularly costly, if one view is more prevalent than the other. This scenario is expected in lipreading systems, where the speaker could be predominantly (but not always!) in the frontal pose. In our recent work [6], we

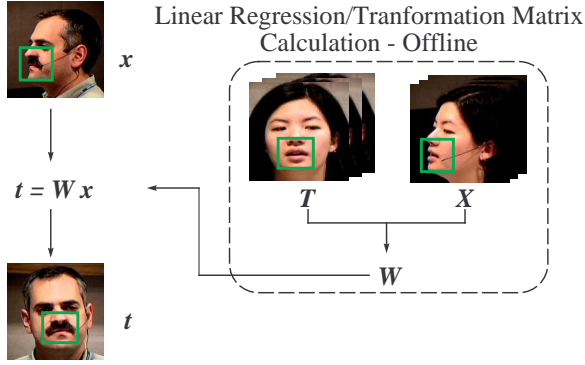


Figure 2: Schematic of the proposed multi-pose lipreading scheme: Visual speech features \mathbf{x} extracted from an undesired view-point (e.g. profile) are transformed into visual features \mathbf{t} in the target view-point space (e.g. frontal) via a linear regression matrix \mathbf{W} , calculated offline based on synchronized multi-view training data \mathbf{T} and \mathbf{X} of features extracted from the different poses.

showed that when a speaker is in one particular pose (such as frontal) more than another (right profile), it is advantageous to project the visual speech features in the undesirable view-point (right profile) into the desirable (frontal) view-point. This projection is performed via a “pose-invariant” technique based on linear regression, and was motivated by work on face recognition [7]. Even though the improvement over the one-model approach was slight, this work clearly demonstrated that by utilizing this “pose-invariant” or “pose normalizing” step, the train/test mismatch between the visual speech features of the different view-points was reduced. A caveat of this approach is the number of effective parameters that can be used. The issue was not investigated in [6], but will be examined in detail here.

The remainder of the paper is structured as follows: In Section 2, the pose-invariant technique based on linear regression is described. Following that, Section 3 focuses on the lipreading system description. Section 4 presents our experimental results, and finally, Section 5 concludes the paper with a summary and a few remarks.

2. Pose-Invariant Lipreading

Blanz et al. [7] cites two possible ways of performing pose-invariant face recognition, either via a viewpoint-transformed or a coefficient-based approach. The viewpoint-transform approach acts in a pre-processing manner to transform/warp an image of an undesirable view-point into the desired view-point. Coefficient-based recognition on the other hand attempts to estimate the face under all view-points given a single view (i.e. frontal and profile in this case), otherwise called the “lightfield” of the face [8].

Although it is not clear which approach is superior, for the purposes of this paper, we used the viewpoint-transform approach. We chose this approach because our frontal-only system is optimized for frontal mouth regions-of-interest only, which was a similar motivation cited by Blanz et al. [7] for their face recognition system. The most common way to perform this is to find the linear regression/transformation matrix \mathbf{W} between a training set consisting of N offline input examples

of the undesirable view-point \mathbf{X} , and their synchronized target examples in the preferred view-point \mathbf{T} [9]. Matrix \mathbf{W} is then determined by minimizing

$$\text{tr}[(\mathbf{W}\mathbf{T} - \mathbf{X})^T(\mathbf{W}\mathbf{T} - \mathbf{X})] + \lambda \cdot \text{tr}[\mathbf{W}^T\mathbf{W}], \quad (1)$$

where $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\mathbf{T} = \{[\mathbf{t}_1, 1]^T, \dots, [\mathbf{t}_n, 1]^T\}$, and data vectors \mathbf{x}_n , \mathbf{t}_n are of dimension D . In (1), a unit bias has been added to \mathbf{T} to allow for any fixed offset in the data. The regularization term, λ , was also introduced into this equation to avoid over-fitting [9]. Over-fitting was not an issue in these experiments due to the large number of training samples ($>100k$), and therefore the value of λ was not significant. From (1), the solution to \mathbf{W} is

$$\mathbf{W} = \mathbf{T}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}. \quad (2)$$

For these experiments, the transformation matrix \mathbf{W} was found using the input visual speech features of a particular view-point \mathbf{X} and their synchronized counterparts \mathbf{T} , instead of the raw mouth image data. By just mapping in the feature domain, we found that keeping the dimensionality low improved performance, as we will show in more detail later in this paper. The matrix \mathbf{W} was used to project all visual speech features of an undesirable view-point (\mathbf{x}), into the desired view-point (\mathbf{t}). The whole process is depicted in Fig. 2.

For this paper, the undesirable view-points were two: the right and left profile views. The frontal view-point was the desired one (see also Fig. 1). Therefore, two separate transformation matrices \mathbf{W} were calculated for projecting the right and left profile visual features into the frontal feature space.

3. The Lipreading System

There exist three main components in our lipreading system: (a) multi-view mouth detection; (b) visual feature extraction; and (c) the speech recognition system. Each will be discussed in the following subsections.

3.1. Multi-View Mouth Detection and Tracking

In these experiments, we used the Adaboost framework of Viola and Jones [10], later extended by Leinhardt and Maydt [11], to perform the mouth region-of-interest (ROI) detection and extraction. This framework allowed us to generate face and facial feature detectors specific for each view-point. As we assumed that we had prior knowledge of the speaker’s pose, detection and tracking of the mouth ROIs was relatively simple, and was accomplished by means of pose-specific face and facial feature detection classifiers. These classifiers were generated using OpenCV libraries [12]. Of course, in a real-world scenario, the speaker’s pose would have to be first estimated.

The actual task of mouth detection and ROI extraction was performed as follows: Given the video of a spoken utterance, the face detector of the specific pose was applied to estimate the location of the speaker’s face. For the frontal scenario, once the face was found, the two eyes were detected and then a coarse mouth region was detected. From this estimate, we applied detectors to find the corners of the mouth. From these detected lip corners, a normalized 32×32 -pixel ROI was then extracted for use in our lipreading system. For the right profile case, once the face was found, the left eye and the nose were detected. From these located features, a coarse mouth detector was applied to give an estimate of the mouth region. From there, we

detected the mouth center and the left mouth corner. A normalized 32×32 -pixel profile mouth ROI was then extracted, based on the distance from the left mouth corner to the left eye. These two points were used as reference points, as they were the most reliable to detect. More information can be found in [5]. As the Adaboost framework allows for extremely quick detection, we were able to perform detection on every frame and used median filtering to allow for smooth tracking.

For the left profile view-point, we used the extracted right profile ROIs, and then just mirrored these images to obtain the left profile ROIs.

3.2. Visual Feature Extraction

For all view-points, the same visual feature extraction process was applied. Following ROI extraction, the mean ROI over the utterance was removed. This approach is very similar to cepstral mean subtraction (CMS) in the audio domain and is known as feature mean normalization [1]. Our implementation is similar to that of Potamianos et al. [1], however in our approach we performed normalization in the image domain instead of the feature domain. A two-dimensional, separable, discrete cosine transform (DCT) was then applied to the resulting mean-removed ROI, with the 100 top DCT coefficients according to the zig-zag pattern retained. An intra-frame linear discriminant analysis (LDA) step was then used to project the features down to 30, resulting in a “static” visual feature vector. Subsequently, in order to incorporate dynamic speech information, five of these neighboring static feature vectors over ± 2 adjacent frames were concatenated, and were projected via an inter-frame LDA step to yield a D -dimensional “dynamic” visual feature vector, extracted at the video frame rate of 30 Hz. For the experiments in the next section, dynamic features of size D ranging from 10 to 60 will be analyzed to examine the effect on the transformation approach. The classes used for LDA matrix calculation were the hidden Markov model (HMM) states (see below), based on forced alignment using an audio-only HMM. This visual feature extraction system, is similar to the state-of-the-art process of Potamianos et al. [1], with the exception of the maximum likelihood linear transform (MLLT) step that is not used here.

3.3. Speech Recognition Systems

In our experiments, we trained three visual-only speech recognition systems:

- (1) A frontal view-point one, trained on 100% of the available frontal pose data (see also Section 4.1);
- (2) A combined two-pose system (frontal and right profile view-point), using 80% of the frontal and 20% of the right profile data;
- (3) A combined three-pose system (frontal, right profile, and left profile) using data at a ratio of 80%, 10%, and 10%, among the three views, respectively.

All systems were designed to recognize connected-digit sequences (ten-word vocabulary with no grammar), and they were based on single-stream HMMs using visual feature vectors of dimension D ranging from 10 to 60. In our experiments, each digit was modeled using nine states with seven Gaussian mixtures per state. A silence and short-pause model were also employed. All models were bootstrapped from a segmentation of the parallel audio channel, obtained by an audio-only HMM

with identical topology. The HTK toolkit was utilized for training and testing [13].

4. Experimental Results

4.1. Database

A total of 38 subjects uttering connected digit strings have been recorded inside the IBM smart room, using two microphones (head-mounted and far-field) and three pan-tilt-zoom (PTZ) cameras (one frontal and two side views of subject). For these lipreading experiments, we utilized the two video views: the frontal and the right profile view. As mentioned previously, the left profile view was obtained by mirroring the right profile extracted ROIs. A total of 1440 utterances were used in our experiments, partitioned using a multi-speaker paradigm into 1198 sequences for training, and 242 for testing. As this data was synchronous, all training and test sequences were available for all view-points. More details can be found in [5].

4.2. Experimental Framework

The training set for system (2) was made up of 80% of frontal features (958 sequences) and 20% right profile features (240 utterances). System (3) was trained on 80% of frontal features (958 sequences), 10% on left profile (120), and 10% on right profile features (120). For systems (2) and (3), all of the different 1198 sequences were accounted for, by randomly substituting the frontal sequences with their synchronously recorded left or right profile utterance.

As mentioned previously, for this paper two experiments were conducted. The first was performed to examine the effect of the feature dimensionality D on the projected profile data. For this purpose, systems (1) and (2) were utilized. These systems were tested on frontal, profile, and projected profile data sets, as well as two “combined” test sets. The latter were made up similarly to the training set of system (2): Set “Comb2” consisted of 80% frontal and 20% right profile data, whereas “Comb2-Proj” of 80% frontal and 20% right profile data projected to the frontal view space. In the second experiment, we considered the addition of the third view-point. In this, all three trained systems were used, and were tested on frontal, as well as combined-view test sets. In this case, in addition to the “Comb2” and “Comb2-Proj” sets, two test sets consisting of three view-points were also considered, consisting of 80% frontal, 10% right profile, and 10% left profile data (original features as well as projected ones), which will be denoted by “Comb3” and “Comb3-Proj”, similarly to the two-view sets. It is worth noting that since the left profile view was just the mirror of the right profile view, the results were identical when testing on the left or right profile views. As such, results are just termed as “profile”.

The projected profile features of system (2) were projected into the frontal view via \mathbf{W} , by having the training frontal features as the target variable \mathbf{T} and the training profile features as the input variable \mathbf{X} . For system (3), each view-point had its own transformation matrix \mathbf{W} , to project the respective profile features into the frontal domain via the above process. It is also worth noting that the regression training sets remained the same (the whole training set was used) due to the limited number of synchronized examples.

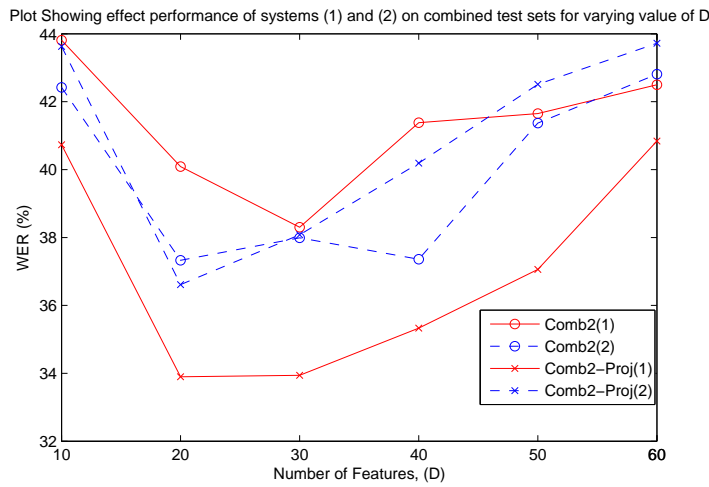
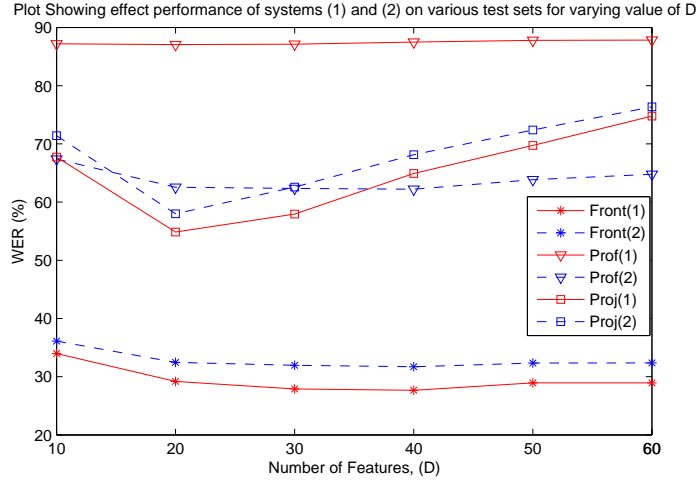


Figure 3: The top plot depicts the lipreading performance in word error rate (WER), %, of systems (1) and (2) on the frontal, profile, and projected test sets as a function of visual feature dimensionality (D). The bottom plot shows the performance on these systems, when they are tested on two multi-view test sets: The “Comb2” test set consists of 80% frontal data and 20% profile data. In contrast, the “Comb2-Proj” set consists of 80% frontal and 20% profile features projected onto the frontal feature space.

4.3. Recognition Results

In the upper plot of Fig. 3 it can be seen that when system (1) is tested on profile views, performance becomes extremely poor in comparison to frontal view data, for all values of D , due to the train/test mismatch. As expected, when these view-points are tested on trained system (2), performance is much better for all values of D . This improvement in the profile view-point test set though has come at a cost, as performance on frontal view data has degraded. This is due to models over-generalizing. This can be somewhat alleviated by projecting the profile features into the frontal domain. Again, for all values of D we see that the projected features outperform the profile features for system (1). However, this improvement is only obtained in system (2) using features with $D = 20$ or 30 , with the performance peaking at $D = 20$ with a word error rate (WER) of 54.85%. This is because once $D > 30$, the lipreading performance of the projected features steadily drops off. As noted by Bishop in [9], linear regression has certain limitations, one of them being the

number of effective parameters which can be used. In this case, it appears that constraining the number of features to $D = 20$ is necessary. Unfortunately, the best performance for the frontal and profile views is obtained using features with $D = 40$ – the WERs are 27.66% and 62.12%, respectively.

In the lower plot of Fig. 3, the lipreading results of systems (1) and (2) are shown for the combined test sets “Comb2” and “Comb2-Proj”. From this plot, it can be observed that the best performance was achieved by system (1) on set “Comb2-Proj”, for all feature dimensionalities considered. The best overall performance achieved was with $D = 20$ – a WER of 33.90%. But as it can be seen, there is not much variation in performance with D over the range of 20 to 40.

Table 1 summarizes results in the case that $D = 20$, as results from the previous experiment demonstrated that the optimal performance was obtained for this feature dimensionality. It can be observed from this table that when data from a third pose are added, the benefit of normalizing the pose via the pro-

| trained system | test set | | | | | | |
|----------------|----------|---------|-----------|--------------|--------------|--------------|--------------|
| | Frontal | Profile | Projected | “Comb2” | “Comb2-Proj” | “Comb3” | “Comb3-Proj” |
| (1) | 29.18 | 87.07 | 54.85 | 40.09 | 33.90 | 40.07 | 33.81 |
| (2) | 32.46 | 62.55 | 57.98 | 37.33 | 36.61 | 41.23 | 40.76 |
| (3) | 32.51 | 69.74 | 58.02 | 38.19 | 37.31 | 39.96 | 36.82 |

Table 1: Summary of the lipreading results in WER, %, for the three trained systems ((1)-(3)), evaluated on single-view data, as well as two- and three-view test sets. In all cases, visual features of dimensionality $D = 20$ are used.

posed pose-invariant technique is more substantial, as compared to the combined systems. When only two poses were used, the performance on set “Comb2” showed that system (2) obtained a WER of 37.33%. In contrast, in set “Comb3” system (3) obtained a WER of 39.96%, a degradation of approximately 2.6% absolute. When adding the third pose on the training data, performance on frontal and projected profile data does not vary much; however performance on profile data degrades from a WER of 62.55% for the two-pose system (2) to 69.74% for the three-pose system (3). This can be attributed to the lack of classification power the system possesses to accurately model features across the different poses. In comparison, projecting the features into a uniform pose did not alter the performance of the lipreading systems at all. It therefore appears that by utilizing the pose-invariant step, the degradation to the overall lipreading performance due to pose variation can be minimized.

5. Conclusions and Further Work

In this paper, we showed that there exists a limit on the number of effective parameters which can be used before the performance of the features generated by the linear regression matrix is affected. Once $D > 30$, the benefit of using this pose-invariant technique is diminished and better performance is gained through a combined model of the different view-points. We also extended our previous work in [6] by including an additional view-point to further illustrate the benefit of projecting all visual features into a single uniform view-point for the task of lipreading. From the results, it is clear that when one particular view-point is more frequent than another, better performance can be gained by using the model of the more prevalent view-point, rather than using a combined model of all the view-points. This is because the combined model has over-generalized exhibiting large train/test mismatch. It would be expected that this trend would continue when more view-points are added (i.e. $\pm 30^\circ$, $\pm 60^\circ$ etc). In future work, we plan to develop a continuous pose-invariant lipreading system that can deal with pose changes within the video sequence.

6. Acknowledgements

QUT work in this paper was supported by Australian Research Council Grant No. LP0562101. Some of this work was conducted as part of Patrick Lucey’s internship with the IBM T.J. Watson Research Center, and was partially supported by the European Commission under integrated project CHIL, “Computers in the Human Interaction Loop”, contract number 506909.

7. References

- [1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, “Recent advances in the automatic recognition of audio-visual speech,” *Proc. IEEE*, 91(9): 1306–1326, 2003.
- [2] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, “Multimodal multispeaker probabilistic tracking in meetings,” in *Proc. Int. Conf. Multimodal Interfaces (ICMI)*, 2005.
- [3] A. Pentland, “Smart rooms, smart clothes,” in *Proc. Int. Conf. Pattern Recog. (ICPR)*, 1998.
- [4] CHIL: Computers in the Human Interaction Loop. [Online]. Available: <http://chil.server.de>
- [5] P. Lucey and G. Potamianos, “Lipreading using profile versus frontal views,” in *Proc. Int. Works. Multimedia Signal Process. (MMSP)*, pp. 24–28, 2006.
- [6] P. Lucey, G. Potamianos, and S. Sridharan, “A unified approach to multi-pose audio-visual ASR,” (To Appear) in *Proc. Interspeech*, 2007.
- [7] V. Blanz, P. Grother, P. Phillips, and T. Vetter, “Face recognition based on frontal views generated from non-frontal images,” in *Proc. Int. Conf. Computer Vision Pattern Recog. (CVPR)*, vol. 2, pp. 454–461, 2005.
- [8] R. Gross, I. Matthews, and S. Baker, “Appearance-based face recognition and light-fields,” *IEEE Trans. Pattern Analysis Machine Intell.*, 26(4): 449–465, 2004.
- [9] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [10] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. Int. Conf. Computer Vision Pattern Recog. (CVPR)*, vol. 1, pp. 511–518, 2001.
- [11] R. Leinhardt and J. Maydt, “An extended set of Haar-like features,” in *Proc. Int. Conf. Image Process. (ICIP)*, pp. 900–903, 2002.
- [12] *Open Source Computer Vision Library*. [Online]. Available: <http://www.intel.com/research/mrl/research/opencv>
- [13] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Ltd., 1999.