

A Unified Approach to Multi-Pose Audio-Visual ASR

Patrick Lucey¹, Gerasimos Potamianos², 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

The vast majority of studies in the field of audio-visual automatic speech recognition (AVASR) assumes frontal images of a speaker’s face, but this cannot always be guaranteed in practice. Hence our recent research efforts have concentrated on extracting visual speech information from non-frontal faces, in particular the profile view. The introduction of additional views to an AVASR system increases the complexity of the system, as it has to deal with the different visual features associated with the various views. In this paper, we propose the use of linear regression to find a transformation matrix based on synchronous frontal and profile visual speech data, which is used to normalize the visual speech in each viewpoint into a single uniform view. In our experiments for the task of multi-speaker lipreading, we show that this “pose-invariant” technique reduces train/test mismatch between visual speech features of different views, and is of particular benefit when there is more training data for one viewpoint over another (e.g. frontal over profile).

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

1. Introduction

Recently, a great deal of progress has been achieved in audio-visual automatic speech recognition (AVASR) [1]. However, practical deployment of AVASR systems, useful in a variety of real-world applications, has not yet emerged. A reason for this is that most research has neglected addressing variabilities in the visual domain such as viewpoint, with nearly all current work concentrating on frontal videos of the speaker’s face. This is mainly due to the lack of large corpora that can accommodate poses other than frontal. But as research has started addressing the meeting scenario inside smart rooms [2, 3, 4], data is becoming available that makes work on visual speech recognition or lipreading from multiple views feasible. This last point has motivated our recent research efforts in AVASR from multiple views [5].

In our previous work [5], experiments were constrained to each viewpoint having its own dedicated AVASR system. Namely, two separate systems were developed, one dedicated to frontal views, another to profile views. In contrast, in this paper we build a more “realistic” AVASR system, by having one unified system use single-camera video, but allowing it to lipread from both frontal and profile views. This is schematically depicted in Fig. 1.

The implications of such a system to practical AVASR are significant: By loosening the constraint on the speaker’s pose, we allow a more pervasive or “real-world” technology to develop, which could be beneficial to in-car AVASR, for example.

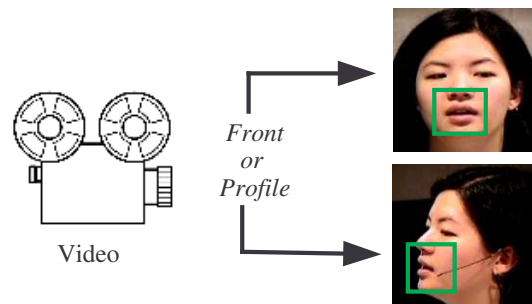


Figure 1: The developed AVASR system in this paper is able to recognize visual speech from both frontal and profile views using a single classifier.

Conversely, by allowing more flexibility in the system, we also increase its complexity. A possible solution to this would be to model and recognize each view independently of each other, thus minimizing training and testing mismatch. Unfortunately, this is complicated to achieve in a continuous setting. On the other hand, having one model trained over all views can also be problematic, as it may over-generalize, causing large train/test mismatch. Such mismatch can drastically degrade classification performance. However, if some sort of invariance in the feature space can be determined, then the entire system will benefit. A number of such approaches have been devised in the acoustic speech domain to lessen train/test mismatch caused by channel conditions and noise; for example, cepstral mean subtraction (CMS) [6] and RASTA processing [7]. This type of approach has also been considered in the visual domain for face recognition, where techniques such as linear regression have been employed to project undesirable non-frontal face images onto a frontal face image. Blanz et al. [8] cite that the major advantage of doing so is the fact that most state-of-the-art face recognition systems are optimized for frontal faces only, and their performance drops significantly for non-frontal data due to the train/test mismatch.

Motivated by these works, in this paper we propose a “pose-invariant” AVASR system that employs linear regression to normalize the visual speech features into a single viewpoint (frontal). We demonstrate that by using this type of viewpoint normalization technique, we can make the system more robust to viewpoint changes. We detail this pose-invariant technique next (Section 2). Following that, Section 3 focuses on the AVASR 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. [8] cite two possible ways of performing pose-invariant face recognition, either via a viewpoint-transformed or a coefficient-based approach. The former acts as a pre-processing stage to transform (warp) an image of an undesirable viewpoint into the preferred viewpoint. Coefficient-based recognition attempts to estimate the face under all viewpoints given a single view (i.e. frontal and profile in this case), otherwise called the “lightfield” of the face [9].

Although it is not clear which approach is superior, in this paper we employ the viewpoint-transform scheme. We choose this method because our frontal-only system is optimized for frontal mouth regions-of-interest (ROI) only, a motivation that is similar to the one cited by Blanz et al. for their face recognition system [8]. The most common way to perform this approach is to estimate a linear regression (transformation) matrix \mathbf{W} between a training set consisting of N examples of the undesirable viewpoint \mathbf{X} , and their synchronized target examples in the preferred viewpoint \mathbf{T} [10]. Matrix \mathbf{W} is then computed 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 \mathbf{x}_n and \mathbf{t}_n are synchronous data vectors of the two views with dimension D (see below). A unit bias has been added to \mathbf{T} to allow for any fixed offset in the data. In addition, regularization term λ has been introduced into (1), as a means to avoid over-fitting [10]. In our experiments, over-fitting was determined not to be an issue, due to the large number of training samples (over 100k), and therefore the value of λ was not significant. Based on (1), the solution for \mathbf{W} becomes

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

In our AVASR experiments, transformation matrix \mathbf{W} was estimated using the input visual speech features of a particular viewpoint, \mathbf{X} , obtained in parallel with synchronized features extracted from the desirable viewpoint, \mathbf{T} (feature extraction is discussed in Section 3). This was deemed preferable to transforming the entire raw image patch of mouth data (ROI), as it reduced dimensionality ($D = 20$, compared to the image domain $D = 32 \times 32 = 1024$), and it also improved performance. Matrix \mathbf{W} , therefore, was used to project visual speech features (\mathbf{x}_n) of an unwanted viewpoint into estimates of desirable viewpoint features ($\hat{\mathbf{t}}_n$). This process is depicted in Fig. 2.

3. The AVASR System

There are four main components in the proposed AVASR system: (a) multi-view mouth detection; (b) feature extraction (both visual and audio); (c) audio-visual integration; and (d) the speech recognition system. Each subsystem will be briefly discussed in the following.

3.1. Multi-View Mouth Detection and Tracking

In this work we used the Adaboost framework of Viola and Jones [11], later extended by Leinhardt and Maydt [12], to perform mouth region-of-interest (ROI) detection and extraction. This framework allowed us to obtain generic face and facial feature detectors specific for each viewpoint. As we assumed that we had prior knowledge of the pose of the speaker, detection and tracking of the mouth ROIs was relatively simple: We just had to apply the specific face and facial feature detection

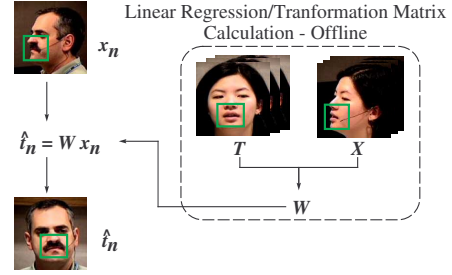


Figure 2: Schematic of the proposed multi-pose AVASR scheme: Visual speech features \mathbf{x}_n extracted from an undesired viewpoint (e.g. profile) are transformed into visual features $\hat{\mathbf{t}}_n$ in the target viewpoint space (e.g. frontal) via a linear regression matrix \mathbf{W} , calculated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

classifiers to the respective pose. The classifiers were generated using OpenCV libraries [13].

In more detail, 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 the mouth region was estimated. Based on this estimate, we applied lip corner detectors to locate the mouth. A normalized 32×32 -pixel frontal ROI based on the lip corners was then extracted for use in our AVASR system. For the profile case, the left eye and the nose were detected first. From these points we were able to estimate where the mouth region was. Next, 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, since they were the most reliable to detect. More information can be found in [5]. The whole process allowed extremely quick detection (faster than real-time). We therefore performed detection at every frame, followed by median filtering to result in smoother tracking.

3.2. Feature Extraction

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, as well as the feature mean normalization for visual feature extraction employed in [1]. Subsequently, a two-dimensional, separable, discrete cosine transform (DCT) was applied to the resulting mean-removed ROI, and the 30 top DCT coefficients according to a zig-zag pattern were retained, resulting in a “static” visual feature vector. In order to incorporate dynamic speech information, seven of these neighboring static feature vectors over ± 3 adjacent frames were concatenated, and were projected via an inter-frame linear discriminant analysis (LDA) cascade to a 20-dimensional “dynamic” visual feature vector. The classes used for the LDA matrix calculation were states of the hidden Markov model (HMM) employed in speech recognition (see Section 3.4), with the necessary data labels obtained by a “forced” state alignment using the audio-only HMM. In addition, the delta and acceleration coefficients were appended to the feature vector, resulting in 60-dimensional visual features available at the video frame rate of 30 Hz. The visual feature extraction system is similar to the state-of-the-art process of Potamianos et al. [1], achieving comparable performance, as discussed in Section 4.2.

In parallel to visual features, 39-dimensional acoustic features were extracted to represent the acoustic signal at the rate of 100 Hz. These were perceptual linear prediction (PLP) based cepstral features, obtained using a 25 ms Hamming window, and augmented by their first and second derivatives.

3.3. Audio-Visual Integration

Following feature extraction, the visual features were upsampled to 100 Hz using nearest-neighbor interpolation to force them to become time-synchronous to the acoustic features to enable easier audio-visual integration. Using the feature fusion approach, the bimodal feature vectors were concatenated, producing 99-dimensional features. The combined features were subsequently projected onto 60 dimensions using another step of LDA. This process is also known as hierarchical LDA (HiLDA) [1]. Similarly to the previous subsection, HMM states were used as the LDA classes.

3.4. Speech Recognition System

In the experiments below we will be comparing five lipreading systems. These systems were trained on the following data:

- (1) Frontal view;
- (2) Profile view;
- (3) Combined frontal and profile views;
- (4) Combined frontal and projected profile (into frontal) views;
- (5) Combined profile and projected frontal (into profile) views.

In addition to these, we will be comparing audio-only and audio-visual systems. All systems are designed to recognize connected-digit sequences (10-word vocabulary with no grammar), and they are based on single-stream HMMs operating on sequences of 60-dimensional features (except the audio-only system that uses 39-dimensional features). For both the audio and visual signals in these experiments, each of the digits were modeled by 9 states and 7 Gaussian mixtures per state using HTK [14]. A silence and short-pause model were also employed.

4. Experimental Results

4.1. Database

A total of 38 subjects uttering connected digit strings have been recorded inside the IBM smart room [2], using two audio channels (head-mounted and far-field microphones) and three pan-tilt-zoom cameras (the frontal and two side views of the subjects). For these experiments, we utilize the far-field audio channel and two video views: the frontal and one of the two side views, namely the one that consistently provides views closest to the profile pose. A total of 1440 utterances are used, partitioned using a multi-speaker paradigm into 1198 sequences for training and 242 for testing. More details can be found in [5]. For this work, we treated both frontal and profile views independently. This means that systems (3), (4), and (5) were trained and tested on twice the amount of data (i.e. 2396 utterances for training and 484 for testing).

The projected profile features of system (4) 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

| trained system | test-set view | | | |
|----------------|---------------|--------------|---------------|---------------|
| | frontal | profile | proj. profile | proj. frontal |
| (1) | 31.42 | 81.65 | 51.09 | — |
| (2) | 78.19 | 37.60 | — | 46.01 |
| (3) | 35.17 | 41.35 | — | — |
| (4) | 33.56 | — | 41.40 | — |
| (5) | — | 40.26 | — | 35.65 |

Table 1: *Visual-only recognition in WER (%) of the five trained systems of Section 3.4, when tested on data of various views.*

the input variable \mathbf{X} . The projected frontal features were projected into the profile view by using the opposite configuration of system (4).

4.2. Recognition Results

Table 1 depicts visual-only recognition results for the various trained systems of Section 3.4. Before proceeding with discussing these results, it is worth noting that in system (2), our visual feature extraction technique gives comparable results to the visual feature extraction scheme in [5]. Indeed, the word error rate (WER) in this experiment is 37.60%, compared to 39.90% on the same profile dataset in [5]. It is clear from Table 1 that systems (1) and (2) give best case scenario results when they are tested on their own viewpoints (31.42% WER for frontal views in (1) and 37.60% for profile views in (2)). However, when they are tested on the other viewpoint, their performance degrades dramatically due to the train/test mismatch. It can be readily observed that the proposed linear regression technique described in Section 2 reduces this mismatch by effectively normalizing the different viewpoint features into a uniform mode (from 81.65% down to 51.09% for (1) and from 78.19% to 46.01% WER in (2)). However, this improvement is still not as good as the performance obtained by combined-view systems (3), (4) and (5). This is because the combined systems are trained on both viewpoints, and are effectively averaged across both views. This generalization does not seem to have affected performance significantly, although the performance of systems (3), (4) and (5) is still not as good as the best case scenarios of (1) and (2).

Over-generalization can be particularly costly, if one view is more prevalent than the other. As mentioned previously, most AVASR systems are set up for fully frontal faces. This is because the system typically expects the speaker to be predominantly in the frontal pose, rather than the profile pose. Consequently, it would be intuitive that the system be trained more on frontal examples than profile ones to cater for this bias. To examine what impact this has, we decided to conduct a secondary experiment which biased the various systems to the frontal scenario. To proceed, we assumed that a speaker would be in the frontal pose for approximately 80% of the time and in the profile pose for about 20%. This was reflected in the training of the various models for the systems, with systems (3) and (4) being trained on 100% of the available frontal data, but only 25% of the profile data (systems (2) and (5) were not used as they were biased towards the profile pose). These profile training sequences were randomly selected from the original training set. The testing sets remained the same. The results for this experiment are shown in Table 2. Note that the regression training sets remained the same due to the limited number of synchronized examples.

From Table 2 it can be observed that system (1) outperforms system (3) for the frontal case. It is also clear that system (1) ob-

| trained system | test-set view | | |
|----------------|---------------|---------|---------------|
| | frontal | profile | proj. profile |
| (1) | 31.42 | — | 51.09 |
| (3) | 32.60 | 52.04 | — |
| (4) | 31.84 | — | 47.29 |

Table 2: *Visual-only WER (%) for various trained systems biased towards the frontal pose, when tested on various views.*

tains slightly better performance than system (3) for the profile case (51.09% WER compared to 52.04%). This result suggests that when the models are biased towards one particular viewpoint, such as the frontal one, it is advantageous to normalize all viewpoints into the better trained viewpoint. A possible reason for this could be that the train/test mismatch between the projected and frontal features is less or comparable to the train/test mismatch between the profile features and the frontally biased combined features, due to the increased importance placed on the frontal viewpoint. It would be expected that when the number of non-dominant viewpoints is increased, this result would be even more dramatic, as non-dominant views increase the amount of variation. As expected, system (4) achieved better performance than (1) and (3) for recognizing profile speech. However, this small improvement in the profile performance may be of little consequence if the majority of visual speech is in the frontal domain.

In our fusion experiments, we wanted to determine how our pose-invariant AVASR system performed when it was biased towards the frontal pose. We chose this scenario, as we believe this would be more likely in a “real-world” situation (i.e. speaker in frontal pose more than profile). For easier comparisons, we selected system (1), as it had the same training set as the audio-only system. It also achieved the best performance for the frontal scenario and gave comparable results for the profile view. We compared this AVASR system to the audio-only system and the visual-only system of (1). Of the original test set, we randomly selected 80% to be frontal and 20% profile (this did not affect the audio test set, as the audio-only signal does not depend on pose). For the clean acoustic case, the audio-only and AVASR systems achieved similar performance (3.80% WER). However, their difference becomes more pronounced if we corrupt the audio channel by “speech babble” noise. The results are depicted in Table 3. As expected, in high noise environments, the visual modality benefit to the audio-only system is dramatic. This once again highlights the importance of the visual modality to an ASR system when operating in noise, even in the presence of pose variability.

5. Conclusions and Further Work

In this paper, we presented an AVASR system able to recognize speech from both frontal and profile views. The system employs a pose-invariant technique based on linear regression that effectively normalizes visual speech features into a single uniform viewpoint. To our knowledge, this is the first work conducted on the topic of pose-invariant AVASR. Such topic is crucial to the deployment of AVASR systems in “real-world” scenarios, as we showed that the train/test mismatch between different viewpoints is large and severely degrades AVASR performance. By employing linear regression as our pose-invariant technique, we demonstrated that we can reduce the mismatch between the visual speech features of the different viewpoints. We also showed that this is of particular benefit when an AVASR system is biased towards one viewpoint (such as frontal).

| SNR | Audio-only | Visual-only (1) | AVASR (1) |
|------|------------|-----------------|-----------|
| 12dB | 5.75 | 35.36 | 5.77 |
| 6dB | 33.46 | 35.36 | 16.38 |
| 0dB | 79.82 | 35.36 | 33.22 |

Table 3: *Comparison of audio-only, visual-only, and audio-visual WERs (%), when the audio signal is corrupted by additive noise to the specified signal to noise ratio (SNR). Both visual-only and audio-visual systems were tested on a 80%-20% mixture of frontal and projected profile data.*

In future work, we plan to develop our system across more poses (e.g. $\pm 90^\circ$, $\pm 60^\circ$, $\pm 30^\circ$, and frontal) and benchmark the pose variation effect on performance. Also, we plan to develop a continuous pose-invariant AVASR system that can deal with pose change within video sequences.

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] CHIL: Computers in the Human Interaction Loop. [Online]. Available: <http://chil.server.de>
- [3] 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.
- [4] A. Pentland, “Smart rooms, smart clothes,” in *Proc. Int. Conf. Pattern Recog. (ICPR)*, 1998.
- [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] R.J. Mammone, X. Zhang, and R.P. Ramachandran, “Robust speaker recognition: A feature based approach,” *IEEE Signal Process. Mag.*, 13(5): 58–70, 1996.
- [7] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE Trans. Speech Audio Process.*, 2(4): 578–589, 1994.
- [8] 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.
- [9] 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.
- [10] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [11] 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.
- [12] R. Leinhardt and J. Maydt, “An extended set of Haar-like features,” in *Proc. Int. Conf. Image Process. (ICIP)*, pp. 900–903, 2002.
- [13] *Open Source Computer Vision Library*. [Online]. Available: www.intel.com/research/mrl/research/opencv
- [14] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Ltd., 1999.