A Unified Approach to Multi-Pose Audio-Visual ASR

Patrick Lucey[1]*, Gerasimos Potamianos*[2]*, Sridha Sridharan*[1]

[1]Speech, Audio, Image and Video Technology Laboratory, Queensland University of Technology, Brisbane, Australia
[2]Human Language Technologies Department, IBM T.J. Watson Research Center, Yorktown Heights, NY, 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. In contrast, our recent research efforts have concentrated on extracting visual speech information from profile views. 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. For our experiments for the task of multi-speaker lipreading, we show that this "pose-invariant" technique reduces the 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 ASR (AVASR) [1]. However, practical deployment of an AVASR system which will be useful in a variety of real-world applications, has not yet emerged. A reason for this is that most research conducted has neglected addressing variabilities in the visual domain such as viewpoint, with nearly all of the present work being conducted on video of a speaker's fully frontal face. This is mainly due to the lack of any large corpora which can accommodate poses other than frontal. But as more work is being concentrated within the confines of a "meeting room" [2] or "smart room" [3] environment, data is becoming available that allows visual speech recognition or *lipreading* from multiple views to become a viable research avenue. This last point has motivated our recent research efforts in AVASR from multiple views [4].

In our previous work, our experiments were constrained with each viewpoint having its own dedicated AVASR system (i.e. two systems, one dedicated for frontal views and another for profile views). In this paper, we make our AVASR system more "real-world", by having one camera but allowing it to lipread from both frontal and profile views. An example of this is shown in Figure 1.

The implications of such a system is of major benefit to AVASR. 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-car AVASR, for example.

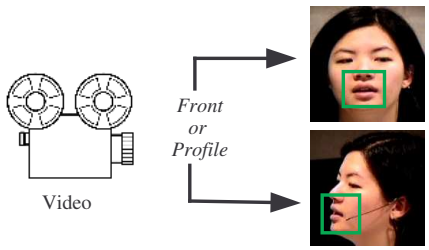

Figure 1: *Given one camera, the AVASR system is to be able to lipread either frontal or profile visual speech.*

Conversely, by allowing more flexibility in the system, we also introduce more complexity. A possible solution to this would be to model and recognize each view independently of each other, thus minimizing the train/test mismatch. Unfortunately, this is complicated to achieve in a continuous setting so a *one model for all* approach is usually employed. Having one model which can generalize over all views is also problematic, as it may *over generalize*, causing large train/test mismatch.

Train/test mismatch can drastically affect the performance of a classifier. Given that only one model is used, if some sort of invariance in the feature space of an input signal is provided then the entire system will benefit. A number of approaches have been devised in the acoustic speech domain to lessen the train/test mismatch caused by channel conditions and noise, such as cepstral mean subtraction (CMS) [5] and RASTA processing [6]. This type of approach has been used similarly in the visual domain for face recognition, where techniques such as linear regression have been used to project the unwanted non-frontal view face image into a frontal face image. Blantz et al. [7] cite the major advantage of doing this is because most state-of-the-art face recognition systems are optimized for frontal views of faces only, and their performance drops significantly if the faces in the input images are shown from non-frontal viewpoints due to large variation in train/test mismatch.

Motivated by these works, in this paper we describe our "pose-invariant" AVASR system, which makes use of linear regression to normalize the visual speech features into a single viewpoint. In this paper, we show by using this type of viewpoint normalization technique, we can make the system more robust to viewpoint change. We describe this pose-invariant technique in the next section (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.

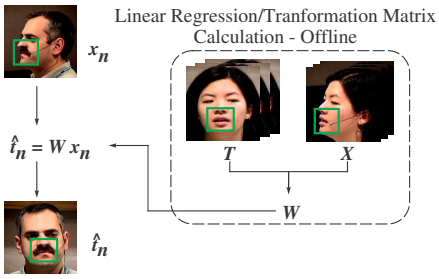

Figure 2: *The visual speech features in an undesired viewpoint (e.g. profile)* $\mathbf{x}_n$ *can be transformed into the wanted target features (e.g. frontal)* $\hat{\mathbf{t}}_n$ *via the transformation matrix* $\mathbf{W}$. $\mathbf{W}$ *is calculated offline through a supervised approach with the target training speech features* $\mathbf{T}$ *and their synchronized input features* $\mathbf{X}$.

2. Pose Invariant Lipreading

Blantz et al. [7] cites two possible ways of performing pose-invariant face recognition, either via a viewpoint-transformed or a coefficient-based approach. Viewpoint-transform approach acts in a pre-processing manner to transform/warp an image of an unwanted viewpoint into the desired 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 [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 mouths only, which was a similar motivation cited by Blantz et al. [7] for their face recognition system. According to Blantz et al., the most common way to perform this approach is the find the linear regression/transformation matrix $\mathbf{W}$ between a training set consisting of N offline input examples of the unwanted viewpoint $\mathbf{X}$, and their synchronized target examples in the wanted viewpoint $\mathbf{T}$. The matrix $\mathbf{W}$ is then found by minimizing

$$tr[(\mathbf{WT} - \mathbf{X})^T(\mathbf{WT} - \mathbf{X})] + \lambda \cdot tr[\mathbf{W}^T\mathbf{W}] \qquad (1)$$

where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{T} = \left[[\mathbf{t}_1, 1]^T, \ldots, [\mathbf{t}_n, 1]^T\right]$, and $\mathbf{x}_n$ and $\mathbf{t}_n$ are of dimension D. A unit bias has been added to $\mathbf{T}$ to allow for any fixed offset in the data. No such bias was given to the input matrix $\mathbf{X}$. The regularization term, λ, was also introduced into this equation and is used 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 this, the solution to $\mathbf{W}$ is

$$\mathbf{W} = \mathbf{TX}^T(\mathbf{XX}^T + \lambda\mathbf{I})^{-1} \qquad (2)$$

For these experiments, the transformation matrix $\mathbf{W}$ was found using the input visual speech features of a particular viewpoint $\mathbf{X}$ and their synchronized counterparts $\mathbf{T}$, and not the raw mouth image data. By just mapping in the feature domain, we found that keeping the dimensionality low ($D = 20$) compared to the image domain ($D = 32 \times 32 = 1024$), improved performance. The matrix $\mathbf{W}$, was used to project all visual speech features of an unwanted viewpoint ($\mathbf{x}_n$), into the wanted viewpoint ($\hat{\mathbf{t}}_n$). This full process is shown in Figure 2.

3. The AVASR System

There are four main components in our AVASR system: (a) multi-view mouth detection; (b) feature extraction (both visual and audio); (c) the audio-visual integration step; and (d) 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 Leinhart and Maydt [11] to perform the mouth *region-of-interest* (ROI) detection and extraction. This framework allowed us to generate 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, as we just had to apply the specific face and facial feature detection classifiers to the respective poses. These classifiers were generated using OpenCV libraries [12].

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 the mouth region was estimated. From this estimate, we applied lip corner detectors to find the mouth. A normalized 32×32 ROI based on these lip corners was then extracted for use in our AVASR system. For the profile case, the left eye and the nose were detected. From these points we were able to estimate where the mouth region was. From there we detected the mouth center and the left mouth corner. A normalized 32×32 profile mouth ROI was then extracted based on the distance from the left mouth corner to left eye. These two points were used as reference points, as they were the most reliable to detect. For more information, see [4]. As the Adaboost framework allows for extremely quick detection (quicker than real-time), we were able to do detection on every frame and used median filtering to allow for smooth tracking.

3.2. Feature Extraction

Following the 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 called *feature mean normalization* for visual feature extraction [1]. Our feature mean normalization is similar to that of Potamianos et al. [1], however in our approach we perform 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 30 top DCT coefficients according to the zig-zag pattern retained, resulting in a "static" visual feature vector. Subsequently, to incorporate dynamic speech information, 7 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 the HMM states. The delta and acceleration coefficients of this final "dynamic" vector were incorporated, resulting in a 60 dimensional visual feature vector at the video frame rate of 30Hz. This visual feature extraction system, is similar to the state-of-the-art process of Potamianos et al. [1]. In the next section, we show that these two systems have comparable results on the same data.

In parallel to the visual feature extraction, 39-dimensional perceptual linear prediction (PLP) based cepstral features including the zeroth, and the first and second time derivatives were extracted to represent the acoustic signal at the audio rate of 100Hz using a 25ms Hamming window.

3.3. Audio-Visual Integration

Following feature extraction, the visual features were upsampled to 100Hz using nearest neighbor interpolation to make it time-synchronous with the audio signal. In this preliminary version of the paper, we will be reporting results using a *feature fusion* approach. In this technique, the bimodal feature vectors are concatenated, resulting to 99-dimensional features that are subsequently projected onto 60 dimensions using LDA cascade. Similarly to the above subsection, the HMM states were used as classes for the LDA matrix calculation. The reduction in the number of features assists in overcoming the curse of dimensionality which allows convergence of the HMMs.

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
- (2) profile
- (3) combined frontal and profile
- (4) combined frontal and projected profile (into frontal)
- (5) combined profile and projected frontal (into profile)

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 which is 39). For both the audio and visual signals in these experiments, each of the digits were modeled using 9 states and 7 Gaussian mixtures per state using HTK [13]. This topology was used as experimental and heuristic evidence showed that this was the optimal configuration. A silence and short-pause model were also employed. All models were bootstraped from time labeled transcriptions.

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 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 in our experiments, partitioned using a multi-speaker paradigm into 1198 sequences for training, and 242 for testing. For full description see [4]. For this work, we treated both the 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 the input variable $\mathbf{X}$. The projected frontal features were projected into the profile view by using the opposite configuration of system (4).

Table 1: *Visual-only results in WER (%) for the various systems tested on different data.*

	System Tested on			
System Trained	Frontal	Profile	Projected Profile	Projected Front
(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

4.2. Recognition Results

Table 1, gives the visual-only results for the experiments for the various systems. Before any analysis occurs, it is worth noting that in system (2), our visual feature extraction technique gives comparable results to the visual feature extraction scheme in [4], with the WER in this experiment 37.60% compared to 39.90% on the same profile dataset. Also from these results it can be seen that systems (1) and (2) give best case scenario results when they are tested on their own viewpoints (31.42% for frontal in (1) and 37.60% for profile in (2)). However, when they are tested on the other viewpoint, the performance severely degrades due to the train/test mismatch. It can be seen that our linear regression technique described in Section 2 reduces this train/test 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% down to 46.01% in (2)). However, this improvement is still not as good as the performance obtained by the combined systems of (3), (4) and (5). This is because the systems of (3), (4) and (5) are trained on both sets of data and are effectively averaged or generalized 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 to cater for this bias. To see what impact this has, we decided to conduct a secondary experiment which biased the various systems to the frontal scenario. To do this, we estimated 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 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 the regression training sets remained the same due to the limited number of synchronized examples.

From Table 2, it can be seen that system (1) outperforms system (3) for the frontal case. It can also be seen that system (1) obtains slightly better performance than system (3) for the

Table 2: *Visual-only results in WER (%) for the various systems biased towards the frontal pose tested on different data.*

	System Tested on		
System Trained	Frontal	Profile	Projected Profile
(1)	**31.42**	-	51.09
(3)	32.60	52.04	-
(4)	31.84	-	**47.29**

profile case (51.09% 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 strongly trained viewpoint. A possible reason for this could be that the train/test mismatch between the projected features and the 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 will be even more dramatic, as these non-dominant views increase the amount of variation in the train/test set. 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.

For the fusion experiments, we wanted to see how our pose-invariant AVASR system performed when it is 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 ease of comparison, 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% of them to be frontal and 20% of them to be 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 system 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 shown 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, even in the presence of pose variability, to an ASR system when operatining in noise.

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

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

5. Conclusions and Further Work

In this paper, we presented an AVASR system which is able to recognize speech from both frontal and profile views. We also presented a pose-invariant technique based on linear regression which 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. The topic of pose-invariant AVASR is central to the future deployment of an AVASR system in a "real-world" scenario as we showed that the train/test mismatch between the different viewpoints is large and severely degrades performance. By employing linear regression as our pose-invariant technique, we showed that we can reduce the train/test mismatch between the visual speech features of the different viewpoints. We showed that this is of particular benefit when an AVASR system is biased to one particular viewpoint (such as frontal).

In future work, we plan to develop our system across more poses (e.g. $\pm 90^o$,$\pm 60^o$,$\pm 30^o$ and frontal etc.) 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

The QUT portion of the research was supported by the Australian Research Council Grant No: LP0562101. Some of this work was also conducted as part of Patrick Lucey's internship with the IBM T.J. Watson Research Center.

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. of the IEEE*, vol. 91, no. 9, 2003.

[2] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan, "Multimodal multispeaker probabilistic tracking in meetings," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI)*, 2005.

[3] A. Pentland, "Smart rooms, smart clothes," in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, vol. 2, 1998.

[4] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing*, Victoria, BC, Canada, 2006, pp. 24–28.

[5] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature based approach," *IEEE Signal Processing Magazine*, vol. 13, pp. 58–70, September 1996.

[6] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, October 1994.

[7] V. Blanz, P. Grother, P. Phillips, and T. Vetter, "Face recognition based on frontal views generated from non-frontal images," in *International Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 454–461.

[8] R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light-fields," *IEEE Trans. PAMI*, vol. 26, no. 4, pp. 449–465, April 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. on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.

[11] R. Leinhart and J. Maydt, "An extended set of Haar-like features," in *Proc. Int. Conf. on Image Processing*, 2002, pp. 900–903.

[12] Open Source Computer Vision Library, `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 (for HTK version 2.2)*. Entropic Ltd., 1999.