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

Towards Lipreading Sentences with Active Appearance Models

George Sterpu, Naomi Harte

Sigmedia, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
sterpug@tcd.ie, nharte@tcd.ie

Abstract

Automatic lipreading has major potential impact for speech recognition, supplementing and complementing the acoustic modality. Most attempts at lipreading have been performed on small vocabulary tasks, due to a shortfall of appropriate audio-visual datasets. In this work we use the publicly available TCD-TIMIT database, designed for large vocabulary continuous audio-visual speech recognition. We compare the viseme recognition performance of the most widely used features for lipreading, Discrete Cosine Transform (DCT) and Active Appearance Models (AAM), in a traditional Hidden Markov Model (HMM) framework. We also exploit recent advances in AAM fitting. We found the DCT to outperform AAM by more than 6% for a viseme recognition task with 56 speakers. The overall accuracy of the DCT is quite low (32-34%). We conclude that a fundamental rethink of the modelling of visual features may be needed for this task.

Index Terms: Visual Speech Recognition, DCT, AAM, Large Vocabulary, TCD-TIMIT

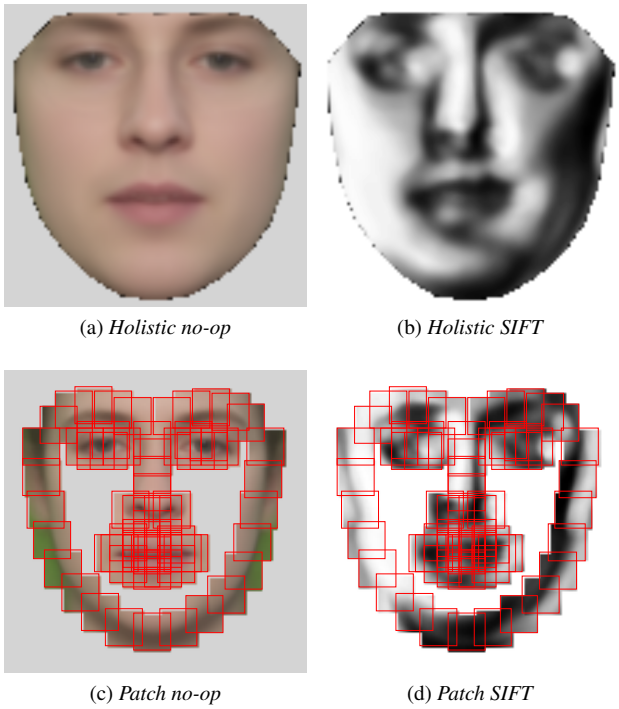
1. Introduction

Lipreading is the process of inferring someone's speech by analyzing the movement of their lips. Humans use lipreading to assist their auditory perception in tasks such as speaker localization, voice activity detection and ultimately speech recognition [1]. This skill allows a robust perception of speech in noisy acoustic environments, or when the hearing abilities have been partially or completely lost.

An open research problem in this area is finding the right representation of visual speech. As outlined by previous reviews [2, 3], most attempts demonstrate an improvement of the audio-visual fusion over the auditory-only modality, yet these results are generally valid for restricted tasks given the known limitations of the used datasets [4]. The main challenges come from speaker variation, pose variation and adequate exploitation of the temporal correlations [3], in addition to the context variation that causes co-articulation. Humans also rely heavily on their language skills when guessing difficult words or long sentences, so a proper integration of language, video and audio is required to reach human-level recognition performance.

Active Appearance Models (AAM), introduced in [5] and streamlined in [6], are state-of-the-art techniques for deformable object modeling. The robustness of AAMs has greatly improved since these early publications via several factors: better fitting algorithms [7], feature-based image descriptors [8] and patch models [9] (portrayed in Figurge 1). These improvements are fairly recent, yet remarkable efforts have been invested to make them available in an open-source project [10].

As the recent AAM developments have been mostly oriented on fitting performance, the recognition performance of the AAM-based features on lipreading tasks once more becomes uncharted territory. AAMs have been applied to lipreading of simple tasks, such as isolated words [11, 12, 13] or small-vocabulary command sentences [14], while the very few attempts on large-vocabulary speech are performed on the IBM ViaVoice dataset which is not publicly available. [2].



(a) *Holistic no-op* (b) *Holistic SIFT*

(c) *Patch no-op* (d) *Patch SIFT*

Figure 1: *Overview of AAM types by warp and feature used. The* Patch *models are evaluating local neighborhoods of the landmarks instead of the entire appearance. The SIFT descriptors are robust alternatives to raw pixel intensities, where no additional operation is applied (no-op).*

The main contribution of this paper is a direct comparison between AAM and Discrete Cosine Transform (DCT)-based visual features on TCD-TIMIT [4], a publicly available audio-visual dataset aimed at large vocabulary continuous speech recognition (LVCSR). We also present an automatic procedure to train AAMs from estimates of pre-trained models, eliminating the need for manual annotations and making it applicable on any dataset. To encourage reproducibility, we make our code publicly available [1].

The rest of the paper is organized as follows. In Section 2 we present the mathematical formulation of our visual feature processing front-ends. In Section 3 we describe the steps taken to train AAMs and fit them to the data. Section 4 presents our experiments, and we draw the conclusions in Section 5.

[1] Note for reviewers: code will be available on GitHub upon paper acceptance

10.21437/AVSP.2017-14

2. Visual features

2.1. DCT

The Discrete Cosine Transform (DCT) represents a standard choice for visual feature extraction in many lipreading tasks [2, 3]. Although aimed at compressing the energy of a signal, it often outperformed algorithms tuned to maximize the classification accuracy, so it is used here as a baseline method.

To obtain a DCT-based feature in our framework, a region of interest (ROI) has to be first localized and isolated from the full-sized image. As the initial work [4] provided extracted mouth ROIs, we obtained their coordinates through cross-correlation-based template matching, so we could apply different post-processing steps. The extracted ROI is converted to gray-scale, then downsampled to 36x36 pixels using cubic interpolation, and finally a 2D DCT transform is applied. The feature vector is made of the first 44 coefficients (without the DC coefficient) chosen in a zig-zag pattern and is concatenated with the first and the second derivatives. The derivatives are computed using a central finite differences scheme that is fourth order accurate, and the same order is preserved at the boundaries by using forward and backward schemes.

Since we are keeping the feature size constant, there is a trade-off between the frequency range captured by the selected DCT coefficients and the granularity of the representation. The choice for the window size was made experimentally, after trying values of 24, 28, 32, 36 and 40 pixels per side.

2.2. AAM

An AAM is a deformable statistical model of shape and appearance that learns the variance of an annotated set of training images. The shape consists of a set of landmarks $\mathbf{s} = [x_1, y_1, ..., x_N, y_N]$ placed on the object to be modeled, which are a priori aligned using Generalized Procrustes Analysis to reduce the effect of translation, rotation and scaling. Applying Principal Component Analysis (PCA) on the set of aligned training shapes leads to a shape model expressed as:

$$\mathbf{s} = \bar{\mathbf{s}} + \sum_{i=1}^{n} p_i s_i = \bar{\mathbf{s}} + \mathbf{Sp} \qquad (1)$$

where any shape s is a linear combination of the shape eigenvectors s_i with the weights p_i also known as shape parameters, plus the mean shape $\bar{s}$.

To construct the appearance model, the pixels within the training shapes are first warped to their corresponding locations in a common reference shape (typically the mean shape $\bar{s}$) using techniques such as piecewise affine warping or thin plate splines. PCA is applied again on the serialized warped image, such that any appearance $A(x)$ could be expressed as a mean appearance $\bar{A}(x)$ plus a linear combination of the appearance eigenvectors $A_i(x)$:

$$\mathbf{A}(x) = \bar{\mathbf{A}}(x) + \sum_{i=1}^{m} c_i A_i(x) = \bar{\mathbf{A}}(x) + \mathbf{Ac} \qquad (2)$$

where the weights c_i denote the appearance parameters.

Since the number of parameters is as large as the number of landmarks and the number of pixels respectively, a trade-off can be made between the representation power of the models and the size of the parameter vectors by analyzing the cumulative ratio of the corresponding eigenvalues.

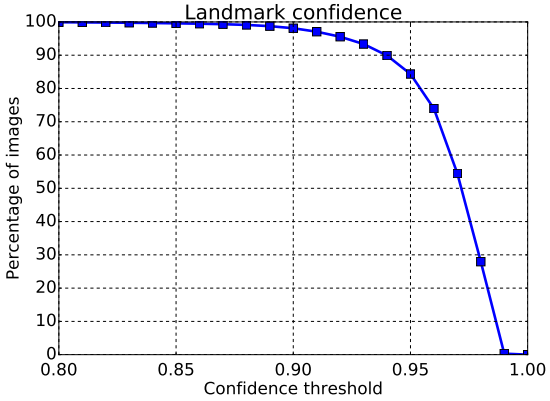


Figure 2: *OpenFace landmark confidence on TCD-TIMIT*

For unlabeled images, when a good initialization of the shape can be provided (e.g. the mean shape aligned on a face localized using a face detector), several fitting algorithms can be applied to iteratively update the parameters that minimize an error between the given image and the model instance. In [7], such algorithms are classified with respect to the cost function, type of composition and optimization method. The parameters obtained at the last iteration constitute the foundation of the AAM-based visual features.

3. Methodology

3.1. Dataset

We have used the TCD-TIMIT dataset [4] for our experiments. The sentences of TCD-TIMIT are designed for good coverage of phoneme pairs in English, implicitly providing realistic viseme contexts, thus well suited for a large vocabulary lipreading task.

To make our results comparable to [4], we have used an identical setup for the speaker-dependent scenario. Hence, we worked on the subset of 56 speakers with Irish accents, each speaker contributing with 67 sentences for training and 31 sentences for testing. For labels, we reused the transcription file made of sequences of the 12 viseme classes.

3.2. AAM training

An annotated set of images is required to train AAMs. Previously, this has been a time-consuming step for most datasets. In [14] and [15], a few frames per speaker are manually annotated, then person-specific AAMs are trained and fitted on the remaining frames. In addition, the final parameters are obtained by projecting the shapes and appearances onto the PCA subspace, which would roughly be equivalent to a Sum of Squared Differences (SSD) formulation of the cost function.

To eliminate the need for manual labor, we propose an automatic procedure to train our models. The open-source tool OpenFace [16] was used to get 68 facial landmark estimates for each frame, storing at the same time their confidence scores as returned by the tool. We then analyzed the cumulative distribution of these confidence scores on our dataset, shown in Fig. 2. This reveals an overall high confidence, which means that most frames have reliable labels. From a visual inspection we observed that most landmarks above a confidence score of 0.9 were very accurate, with the exception of the lips region.

Training generative models such as AAMs with a massive amount of similar data, such as consecutive video frames, leads to poor performance in practice, so we apply a sampling strategy. Taking the faces that get detected successfully and that have a high confidence score, we sort them by the amount of lip opening (distance between the upper and lower lips) and uniformly sample between 3 to 6% of them. For TCD-TIMIT we decided to use a confidence threshold of 0.94 to train our models, which kept 90% of the frames. In addition, we randomly selected only 5 training sentences per speaker from the available of 67, further reducing the training data size. We will refer to these models as *global*, since they use training data from each volunteer. The models built from the training samples of a single person will be coined *person-specific*.

All the previous attempts at lipreading with AAMs have used the original formulation where the entire appearance texture within the landmark area was modeled. It has been shown that learning only small patches around the landmarks leads to robust models that outperform the state-of-the-art at fitting to unseen faces [9]. We considered both approaches, coined *Holistic* and *Patch* AAMs in [10], (and illustrated in Fig.1) in order to compare their fitting and classification performance. In addition to the traditional pixel intensities for appearance features (denoted in this work and in [10] as *no-op*), we also considered SIFT [17] image features, which were shown to largely outperform popular alternatives at fitting to unconstrained images, requiring at the same time fewer appearance components [8].

Modeling only a part of the face can be beneficial for lipreading [18, 13], since the PCA energy would better describe the subtle movements. Yet, as the area being modeled gets smaller, is it expected to see an increase in fitting error. We built two additional models, one for the lips area only, and another for the whole chin and mouth area (further denoted as *chin*), the latter being chosen as a trade-off between relevancy to lipreading and fitting performance. The face and the chin models use a pyramid of three resolution levels (25%, 50%, 100%), while the lip models only use the last two.

Some other important parameters for our models were the image rescaling to a diagonal of ≈ 150 pixels at full scale, 40 and 150 shape and appearance components respectively, and a patch size of 17x17 pixels around landmarks for the Patch models.

Table 1 shows how well our models were able to represent the appearance of the training data. High values of the kept variance imply that model is able to reconstruct accurately any given face, provided that the optimization algorithm finds the right parameters. More variance was kept using pixel intensities than SIFT features, as the color images have only three channels while SIFT has eight, thus more data is being modeled. The variance kept by the shape eigenvectors was close to 100% using 40 components, suggesting that there are strong correlations between the landmark locations.

3.3. AAM feature selection

The AAM fitting process consists in the optimization of a cost function (typically the error between a given image and the AAM reconstruction) with respect to the shape and appearance parameters, provided that a good initialization is available. The shape was initialized using the *dlib* face detector implemented in *menpo* [10] by aligning the mean shape with the face bounding box. The Wiberg Inverse Compositional (WIC) algorithm was chosen for the optimization problem, as it was shown to be an efficient alternative to state of the art algorithms [7]. We ran 10 iterations of WIC for the first two resolution scales and 5 more for the full resolution model, with an important exception of the Holistic no-op model that needed 20 iterations at the lowest scale in order to converge more often. For the *chin* and *lips* models, the shapes were initialized from a subset of the final *face* shape, iterating 10 more times per resolution scale to make room for corrections.

Table 1: *Percentage of kept variance for the appearance models using 150 appearance components*

Model →	**Holistic**		**Patch**		**Scale**
↓ Part	**no-op**	**SIFT**	**no-op**	**SIFT**	
face	96.6	78.7	83.1	63.0	**25%**
	96.8	79.2	87.6	71.1	**50%**
	93.2	76.9	82.8	74.7	**100%**
chin	97.9	75.9	82.8	56.4	**25%**
	97.1	73.4	87.4	65.0	**50%**
	93.6	70.1	83.9	69.6	**100%**
lips	95.4	72.2	89.2	61.6	**50%**
	91.6	68.5	90.9	65.9	**100%**

We considered the shape and the appearance parameters after the last iteration as feature vectors, either taken separately or concatenated, and we also considered the first derivative of the appearance alone or the concatenation. Among these five features, the highest performance was achieved by the latter, which was our default choice in the subsequent experiments. The first four shape parameters were discarded, as they represented the global similarity transform used for normalization.

It is worth mentioning that fitting is a slow process, taking a couple of days to process the files of a single speaker using the four face models alone in *menpo*. We ran the fitting process on a HPC cluster made of 16 nodes and 40 cores, achieving a theoretical speedup factor of 160.

3.4. Viseme recognition

Our monoviseme recognizer was implemented in HTK 3.5 [19], following the procedure described in [4] as close as possible. For each of the 12 viseme classes we have built 3-state left-to-right Hidden Markov Models (HMMs) with mixtures of 20 diagonal covariance Gaussian densities per state, initialized in flat start mode with *HCompV*. Additionally, for the silence state viseme we have added backward and skip transitions. Finally, we have applied 5 runs of embedded training using *HERest* for every increment of the mixture components.

The reported correctness and accuracy results are computed using *HResult* between the ground-truth transcriptions provided with TCD-TIMIT and the predicted ones.

No language model was used. This allows a comparison of the raw lipreading ability of the feature sets.

4. Experiments

4.1. Fitting performance

The overall system performance relies first on the accuracy of landmark localization on unseen faces. In this experiment we compare the performance of our face AAMs in terms of face-normalized point-to-point Euclidean error between the WIC fitter prediction and the ground-truth shapes.

Although the ground-truth labels are not perfect, having a high confidence rate as in Figure 2 leaves little room for noise.

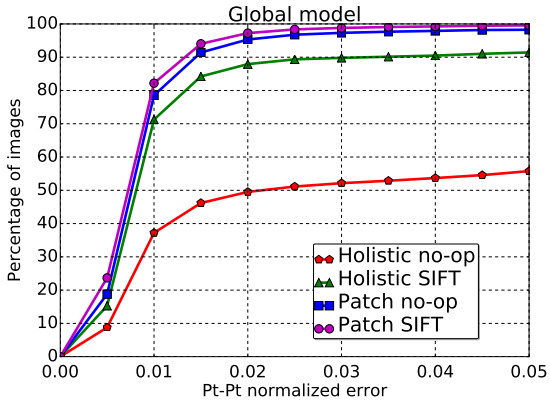


Figure 3: *AAM fitting convergence using global face models (trained on the full set of volunteers)*

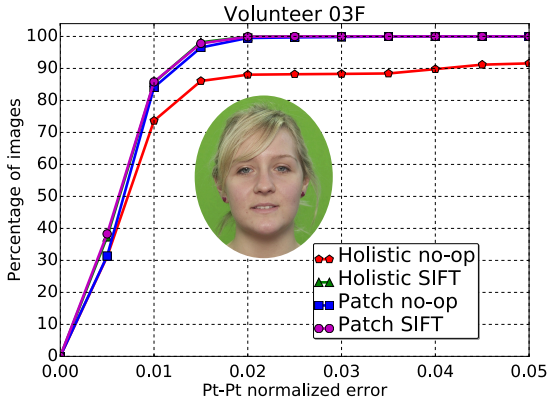


Figure 4: *AAM fitting convergence using a person-specific model for volunteer 03F*

We obtained almost identical results when considered the fitting performance only on the frames above 0.94 confidence.

Figure 3 shows the proportion of frames fitted with an error lower than a certain value, using the global face models, while Figures 4-5 show the same information using person-specific AAMs of two volunteers. The two speakers modeled individually were drawn from the top/bottom 10 performers in [4], where volunteer *03F* was considered easier to lipread than *34M*, which had a full beard and moustache.

The Holistic models were outperformed by the Patch models in almost all cases, with the exception of volunteer *03F* where *Holistic SIFT* managed to match them, although for volunteer *34M* it couldn't cope well with the facial hair. Both Patch models achieved a convergence rate above 95% for an error of 0.02 and were almost indistinguishable in performance, demonstrating their robustness not only for fitting to unseen frames, but also when trained from less perfect landmarks.

In most cases, AAMs were able to improve the pre-trained OpenFace estimates where the confidence score was low. One such example is shown in Figure 6, where the eyes and the eyebrows landmarks were corrected for volunteer *05F* wearing glasses with the eyebrows not visible. This leads to a better parametrization of the fitter for faces that are otherwise more challenging to landmark.

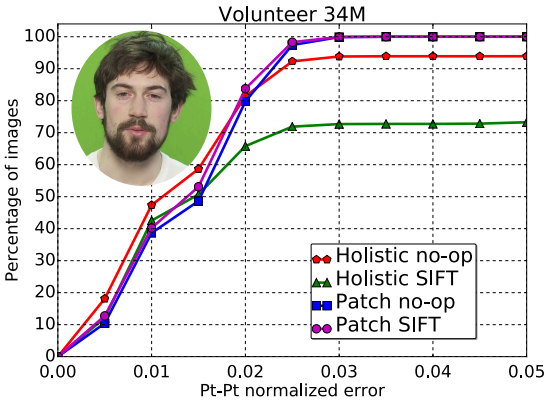


Figure 5: *AAM fitting convergence using a person-specific model for volunteer 34M*

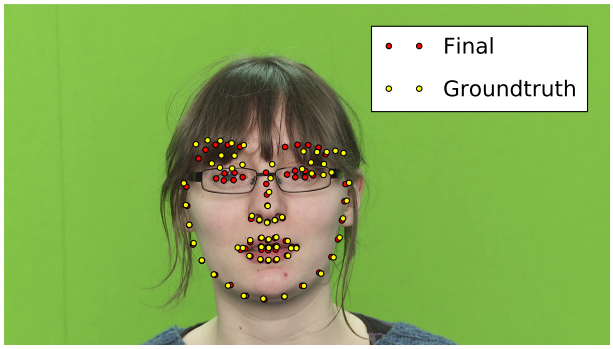


Figure 6: *Landmark correction for volunteer 05F wearing glasses and with the eyebrows occluded*

4.2. Recognition performance

We now focus on the recognition results obtained by training HMMs in a speaker-dependent scenario, thus using 67 training sentences from each volunteer and testing on their remaining 31 unseen sentences. The predicted viseme sequence is computed using the HTK tool *HVite*.

In Figure 7 we plot the correctness and accuracy scores returned by *HResults* for an increasing number of volunteers added to the system (ordered by their alphanumeric IDs). The accuracy on the entire set of volunteers (31.59%) is 3% below the one obtained in [4]. An increase of 1-2% was possible when we interpolated the features to double the rate and used 4-state HMMs, but we reverted to the original settings to have a fair comparison with the AAM features.

In Figure 8 we show the accuracy obtained using AAM-based features and an identical HMM recognition framework. As anticipated, the *Holistic no-op* model has the lowest accuracy, since less than 60% frames converged on average. The other three models perform similarly, yet reaching an accuracy of $\approx 25\%$ on the entire set, significantly lower than DCT.

We repeated the experiment with features extracted using the two part models, *chin* and *lips*, on a subset of the first 33 volunteers, following the process described in Section 3.3. The results are displayed in Figure 9, showing the *chin* model to perform only marginally better, although the decreasing trend remains. This small increase comes with the cost of doubling the processing time, as it requires a cascade of two fittings.

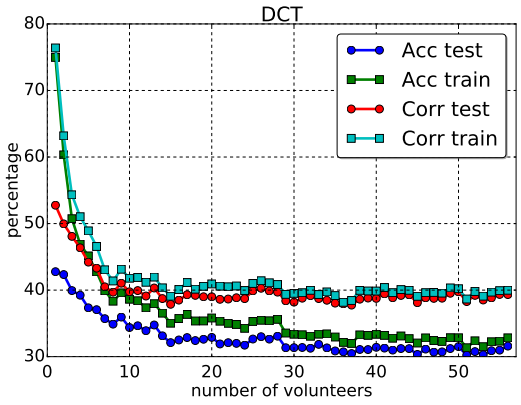


Figure 7: *Correction and Accuracy scores for DCT features*

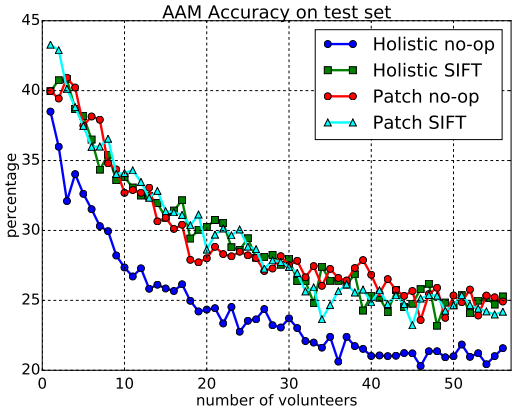


Figure 8: *Accuracy scores for AAM features*

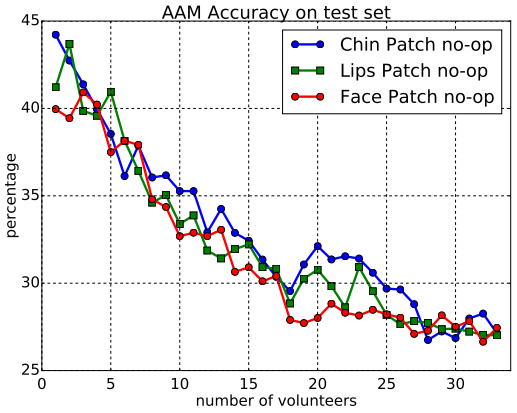


Figure 9: *Performance of the* chin *and* lips *models*

Table 2: *Recognition performance for person-specific models versus the global models. Since there was less data available for individual speakers, the highest values were obtained on average with 14 Gaussian mixture densities*

Speaker →	**03F**		**34M**	
↓ Part	**Corr**	**Acc**	**Corr**	**Acc**
Specific AAM				
face	51.84	42.62	51.63	43.24
chin	52.62	45.24	53.11	40.18
lips	50.87	43.98	52.62	43.14
Global AAM				
face	53.11	44.66	51.04	41.76
chin	53.88	43.50	51.53	41.95
lips	52.43	44.27	53.70	42.15
DCT	54.66	46.80	47.88	39.68

4.3. Speaker-specific models

In order to see how much the quality of the AAM impacts the viseme recognition accuracy, we tested the case of person-specific AAMs for the two volunteers described in Section 4.1. If there was a problem with the global model, we should notice a significant increase in accuracy when switching to person-specific models. Table 2 shows the viseme recognition results obtained with both specific and global models for these two speakers, along with the DCT baseline. We could not find a significant advantage of the person-specific models, hence at this stage it would not be useful to attempt adapting a global AAM to particular faces in order to gain a performance boost.

5. Discussion

In this paper we have explored the performance of hand-crafted visual features for a LVCSR lipreading task in a traditional HMM framework. We first computed DCT-based features for a baseline, reaching a similar result as in [4]. Then we trained several AAMs using an automatic procedure and fitted them to each video frame to obtain the AAM-based features.

A first finding is that AAM features do not outperform the DCT ones in an identical recognition framework. This has been reported before on IBM ViaVoice [2]. This dataset has 290 subjects and over 50 hours of speech. However their approach was to rescore audio-only lattices with visual unit HMMs. Their scenario therefore bypassed the issue of using visual features to find the viseme boundaries. On the other hand, the study of [14] found AAM better than DCT on a lipreading task with a small vocabulary of 51 words, where word-level HMMs were used. Later work from the same authors in [20] reported results on a corpus of 12 speakers, each speaking 200 sentences from a vocabulary totalling 1000 words. Again AAM outperformed DCT, but the approach made use of Linear Discriminant Analysis requiring frame-aligned viseme labels, while the facial landmarks were obtained semi-automatically from person-specific trackers. Another study [11] used speaker-specific normalization that makes the results less comparable. This is the most comprehensive comparison between DCT and state-of-the-art AAM that we are aware of.

Both AAM and DCT perform a basis decomposition of the image, although the first considers the eigenvectors specific to a training set of images, while the latter uses a fixed frequency decomposition. Since both transforms are not optimized for classification, e.g. maximizing the separability between classes, this suggests that the raw parameters are not necessarily ideal features, requiring further processing to find person-independent cues. This is reinforced by the fact that *Patch AAMs* obtained a high convergence rate at fitting to unseen images, so the parameters should contain meaningful information.

Modeling a subset of the face has only shown minor im-

provements of the recognition accuracy. The *chin* model seems to have a slightly better advantage versus the *lips* one, and this could be explained by two factors. The extra iterations of the part model ensured a more accurate fitting where there were more control points available. Also, the chin area contains additional visemic information, as speech articulators are not limited to the lips region.

In this context, we could also question the suitability of HMMs for LVCSR lipreading. We plan to reproduce our experiments on simpler datasets such as GRID [21], CUAVE [22], and also on OuluVS2 [23] which is similar to TCD-TIMIT. A thorough analysis of the HMM suitability would require debugging down to the Baum-Welch and Viterbi algorithms to understand the fail cases. Another informative experiment would be to replace the HMM framework for pattern recognition with a Long short-term Memory (LSTM) one, while reusing exactly the same features, as it could reveal a bottleneck at the recognition level and not the feature one.

An important conclusion about AAMs is that the *Patch* models, especially when combined with SIFT image descriptors, are able to achieve a much higher fitting and implicitly recognition accuracy than the traditional *Holistic* ones that have been used so far in lipreading. As shown in [9], their robustness is conspicuous when trained on unconstrained in-the-wild faces, making them more suitable candidates for realistic lipreading scenarios.

6. Acknowledgements

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. This work was also supported by TCHPC.

7. References

[1] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by eye: The psychology of lip-reading*, Dodd, Ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1987.

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, Sept 2003.

[3] Z. Zhou, G. Zhao, X. Hong, and M. Pietikinen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, no. 9, pp. 590 – 605, 2014.

[4] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," 1998, pp. 484–498.

[6] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[7] J. Alabort-i Medina and S. Zafeiriou, "A unified framework for compositional fitting of active appearance models," *International Journal of Computer Vision*, vol. 121, no. 1, pp. 26–64, 2017.

[8] E. Antonakos, J. A. i Medina, G. Tzimiropoulos, and S. P. Zafeiriou, "Feature-based lucas-kanade and active appearance models," *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2617–2632, Sept 2015.

[9] G. Tzimiropoulos and M. Pantic, "Gauss-newton deformable part models for face alignment in-the-wild," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1851–1858.

[10] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou, "Menpo: A comprehensive platform for parametric image alignment and visual deformable models," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 679–682.

[11] K. Paleček and J. Chaloupka, "Audio-visual speech recognition in noisy audio environments," in *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*, July 2013, pp. 484–487.

[12] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, Feb 2002.

[13] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, March 2009.

[14] Y. Lan, R. Harvey, B. Theobald, E. Ong, and R. Bowden, "Comparing visual features for lipreading," in *AVSP 2009*, 2009, pp. 102 – 106.

[15] H. L. Bear, R. Harvey, B. J. Theobald, and Y. Lan, "Resolution limits on visual speech recognition," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 1371–1375.

[16] T. Baltrušaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–10.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[18] C. Berry, A. Kokaram, and N. Harte, "An extended multiresolution approach to mouth specific aam fitting for speech recognition," in *2011 19th European Signal Processing Conference*, Aug 2011, pp. 1959–1963.

[19] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.5*. Cambridge, UK: Cambridge University Engineering Department, 2015.

[20] Y. Lan, B.-J. Theobald, R. W. Harvey, E.-J. Ong, and R. Bowden, "Improving visual features for lip-reading." in *AVSP*, 2010, pp. 7–3.

[21] 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, pp. 2421–2424, 2006.

[22] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, May 2002, pp. II–2017–II–2020.

[23] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "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)*, vol. 1, May 2015, pp. 1–5.