3D VISUAL SPEECH ANIMATION USING 2D VIDEOS

Rabab Algadhy, Yoshihiko Gotoh, Steve Maddock

Department of Computer Science
University of Sheffield, United Kingdom

ABSTRACT

In visual speech animation, lip motion accuracy is of paramount importance for speech intelligibility, especially for the hard of hearing or foreign language learners. We present an approach for visual speech animation that uses tracked lip motion in front-view 2D videos of a real speaker to drive the lip motion of a synthetic 3D head. This makes use of a 3D morphable model (3DMM), built using 3D synthetic head poses, with corresponding landmarks identified in the 2D videos and the 3DMM. We show that using a wider range of synthetic head poses for different phoneme intensities to create a 3DMM, as well as a combination of front and side photographs of the real speakers rather than just front photographs to produce initial neutral 3D synthetic head poses, gives better animation results when compared to ground truth data consisting of front-view 2D videos of real speakers.

Index Terms— visual speech animation, lip motion, 3D morphable model.

1. INTRODUCTION

Humans notice any slight flaws in visual speech animation. This increases the focus on creating realistic mouth animation that is synchronized with a real speaker's utterance. Various approaches to visual speech animation have been proposed, which can be broadly classified into two categories: viseme-driven approaches and data-driven approaches. Viseme-driven approaches involve segmenting speech into phonemes, which are then classified into visual units called visemes, which are poses matched to the main visual appearance of the phoneme, e.g. a closed mouth shape for a bilabial phoneme such as /m/ [1]. Viseme parameters are then interpolated with co-articulation rules incorporated [2, 3]. Data driven approaches involve motion capturing data (video or 3D) from a real speaker to produce a synthesized talking head [4, 5] or to reanimate faces in images and videos [6, 7, 8]. The captured data is either organised based on phonetic information (sample-based approaches) [9], or processed using statistical models to control the facial motion that is learned from the training data (learning-based approaches) [10, 11, 12, 13, 14].

In this paper we present a data driven approach to fit a 3D morphable model (3DMM) [15] to images and video streams. We use a 3DMM that is based on 3D synthetic head poses generated using commercial software (FaceGen[1]) to train the model, rather than 3D scans of real speakers. The FaceGen models have corresponding vertex data in each head, making model training easier. In order to generate the neutral head pose for a face, photographs of a real speaker are used. Our 3DMM is fitted to 2D video of a real speaker (from [16]) using the method in Huber et al's work [17], which involves reconstructing 3D faces from images using a 3DMM.

We conducted a series of experiments to investigate how using different amounts of data in different stages of the process influences the final animation results. We show that using different versions of visemes for a phoneme, showing different strengths of viseme shape (e.g. different amounts of mouth openness for the same viseme), when constructing the 3DMM, gives better results in the final animation. We also show that using both a front- and side-view photograph in the initial 3D head construction further improves the final animation results. We use ground-truth data (the front-view videos of a speaker [16]) to compare the final synthetic 3D animation results against.

The rest of the paper is structured as follows. Section 2 describes the 3DMM and presents the approach to fit the 3DMM to 2D video. Section 3 describes the experiments and the results. Finally, Section 4 concludes the paper.

2. METHODS

Two separate steps can be identified: the construction of a 3DMM and the mapping of 2D video data to a 3D synthetic head.

2.1. A 3DMM

A 3DMM requires a set of head poses for training. These are often generated by taking scans of real people. Instead, we use FaceGen to produce synthetic head poses. An initial neutral head pose can be generated using photographs of a real person, either a front-view only photograph or front and side

The first author is supported by a scholarship from the Libyan Embassy.

[1] https://facegen.com

views. The software can then be used to deform the face into a range of poses. The software includes 16 default viseme poses, which are parameterised so that different intensities of each viseme can be generated, i.e. different amounts of openness for an open-mouthed viseme. We make use of this functionality in generating our datasets (see Section 3). Each head pose created using FaceGen automatically has vertex correspondence, something which is more complex to achieve with scanned data. FaceGen also generates tongue and teeth poses, but we exclude this since we are concentrating on lip shape.

Given a set of head poses, Principal Component Analysis (PCA) can be applied to the vertices to generate a 3DMM. Only shape needs to be considered, since every head pose shares the same texture. The geometry of the head is represented by a shape vector $S = (X_1, Y_1, Z_1, \ldots, X_n, Y_n, Z_n)^\top$, containing the X, Y, Z coordinates of the vertices, where n is the number of FaceGen poses used to build the 3DMM. The 3DMM consists of a PCA model of the shape, which is represented as:

$$M := \{\overline{F}, \sigma, V\} \tag{1}$$

where $\overline{F} \in R^{3N}$ is the mean vector of the example meshes (mean pose) with N being the number of mesh vertices, and $\sigma \in R^{n-1}$ denotes the standard deviation, where $V = [v_1, \ldots, v_{n-1}] \in R^{3N \times n-1}$ is a set of principal components in the model.

A new pose can be generated as follows:

$$S = \overline{F} + \sum_{i=1}^{K} \alpha_i \sigma_i v_i \tag{2}$$

where $K \leq n-1$ is the number of principal components and $\alpha_i \in R^K$ is the shape coefficient [17].

2.2. Mapping 2D to 3D

To generate the 3D animation, 2D video of a speaker needs to be mapped to the 3DMM. This is achieved using the camera matrix method presented by Huber et al [17]. This section briefly explains the 2D facial landmarks tracking process and how the pose of the 3DMM is estimated and fitted to the mouth shape of a real speaker.

In order to track the facial features of a real speaker in a video, the random cascaded-regression copse (R-CR-C) approach presented by Feng et al [18] is used, which regresses a set of facial feature landmarks to fit a predictive shape model to the true shape. Based on that, when a video is run, a learned landmark detection model using the Ibug-Helen test set [19] will detect and track the facial features of the real speaker.

Given 51 2D landmarks and the corresponding 3D landmarks a pose of the face is estimated using the Gold Standard Algorithm (more details in [17]). It computes the camera matrix that is used to reconstruct the 3D shape. Figure 1 shows the facial landmarks labelled on a video frame of a real speaker (left) and on the corresponding 3D head model (right) that correspond to a set of Ibug[2] facial landmarks (the contour landmarks were excluded).

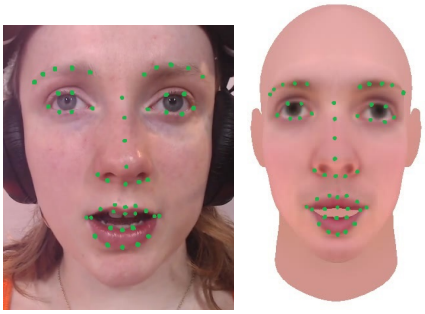

Fig. 1: The facial landmark points labelled on a real speaker (left) and the corresponding 3D head model (right).

The most likely vector of PCA shape coefficients, α, is found by minimising the following cost function:

$$E = \sum_{i=1}^{3L} \frac{(y_{3D,i} - y_{2D,i})^2}{2\sigma_{2D}^2} + \|\alpha\|_2^2 \tag{3}$$

where L is the number of landmarks, $y_{2D,i}$ is the 2D landmarks represented in homogeneous coordinates, σ_{2D}^2 is an ad hoc variance of these landmarks, and $y_{3D,i}$ is the projected 3D landmarks to a 2D plane using the camera matrix [17].

3. EXPERIMENTS AND RESULTS

The experiments address two main questions: (i) would using different intensities of the same viseme shape (e.g. different amounts of mouth openness for the same viseme) when constructing the 3DMM produce better animation results? (ii) would using both front- and side-view photographs, rather than just a front-view photograph, in the construction of the initial 3D head pose produce better animation results?

3.1. Data sets

Four data sets were used to build different 3DMMs for a speaker. Table 1 summarises the data sets. The differentiating factors are whether 17 (16 visemes and a neutral pose) or 161 poses (10 intensity variations of 16 visemes and a neutral pose) are used for a 3DMM and whether a front-view photo only or front- and side-view photos are used in constructing the neutral head pose. Figure 2 shows the front and side photographs of a real speaker (ID: S32) and the corresponding 3D heads that were generated using a front-view photograph only (left), and front- and side-view photographs (right). Each of the data sets was used in producing a 3DMM, which was subsequently used in the process described in Section 2.

[2]https://ibug.doc.ic.ac.uk/resources/facial-point-annotations

Data set	17 poses	161 poses	front-view photo	front- and side-view photo
1	y		y	
2		y	y	
3	y			y
4		y		y

Table 1: The data sets

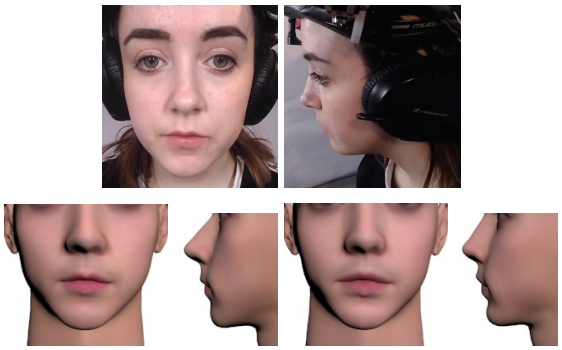

Fig. 2: First row: Front (left) and side (right) photographs of a real speaker (ID: S32); Second row: front and side view of the corresponding 3D heads generated using front photograph only (left) and front and side photographs (right) – the lips are more protruded in the image on the right.

3.2. Evaluation

In order to validate the process, videos of four female speakers (IDs: S15, S17, S24 and S32) and two male speakers (IDs: S20 and S48) from the Audiovisual Lombard Grid Speech corpus [16] were used. The corpus consists of both front- and side-view video of 54 speakers (30 female and 24 male) uttering sentences from the GRID corpus [20] in both plain and Lombard conditions. We used only the front-view videos of plain sentences.

For each real speaker, four plain sentences from the front-view video files were chosen to be mapped to each corresponding 3D synthetic head (built using FaceGen). The resulting 3D head animation was then compared to the original ground-truth 2D videos. This was done for each of the 3DMMs built for the 4 data sets summarised in Table 1.

For the comparison, Faceware Analyser[3] was used to track the facial features in the ground-truth 2D video and the front-view (2D) of the corresponding 3D animation. Two geometric articulatory measurements were calculated from the extracted facial features. The first was a width measurement defined by the horizontal distance between the right and left inner corners of the lips. The second was a height measurement defined by the distance between the top and the bottom middle of the inner mouth contour. In order to correct for the distance between the camera and the real speaker or the talking 3D head, all the landmarks were normalised by using the Euclidean distance between the midpoint of the inner corners of the eyes and the nose tip's point, since these were not affected by the articulations. All visual articulatory features for the real speakers and their corresponding 3D heads were normalised by their corresponding maximum and minimum mouth measurements in the videos. This gives all the articulatory measurements on a [0-1] scale. Given the height and width values for each frame of animation, for both the real video for a speaker and the corresponding 3D animation, the root mean square error (RMSE) over a sentence was used to evaluate the effectiveness of each 3DMM.

[3]http://facewaretech.com/products/software/analyzer

3.3. Results and Discussion

Figure 3 shows an example of consecutive frames of the phoneme /w/ during utterance of the letter y for a real speaker (ID: S17) and the corresponding 3D head for each data set. This figure shows that the performance of the animated 3D lips improves when a larger number of 3D head poses (i.e. different viseme intensities) are used to train the 3DMM, and further improves when front- and side-view photos are used to generate the initial neutral head pose in FaceGen.

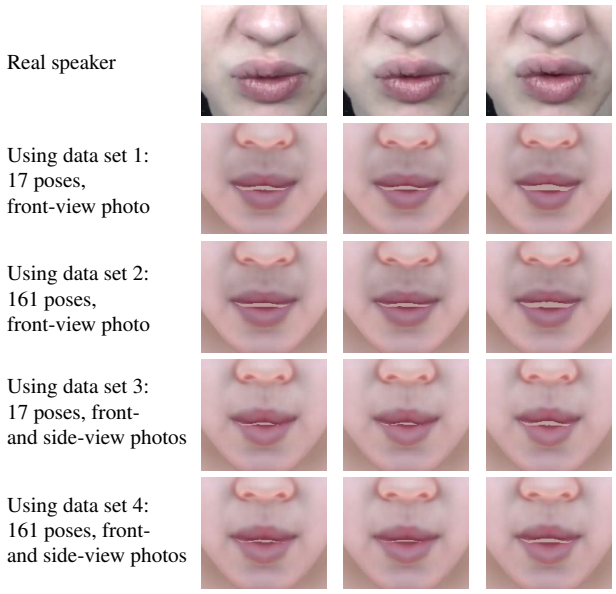

Fig. 3: Video frames of a real speaker (ID: S17) and the 3D head produced for each data set.

Figure 4 shows the trajectories of the width and the height parameters for the same speaker (ID: S17) and the corresponding 3D heads whilst uttering the sentence "place green in y zero again". Whilst all the trajectories generated using the animation pipeline generally follow the real speaker's trajectory, the trajectories of the 3D heads that contain 161 poses

ID	Front photo				Front+side photo			
	17 poses		161 poses		17 poses		161 poses	
	W	H	W	H	W	H	W	H
S15	0.152	0.120	0.154	0.117	**0.129**	0.102	0.131	**0.087**
S17	0.121	0.137	0.115	0.128	0.120	0.109	**0.092**	**0.095**
S20	0.239	0.166	0.247	0.158	**0.229**	0.156	0.244	**0.155**
S24	0.287	0.141	0.223	0.151	0.260	0.142	**0.219**	**0.123**
S32	0.117	0.067	0.115	0.075	0.210	0.067	**0.111**	**0.056**
S48	0.199	0.086	0.175	0.080	0.203	0.075	**0.149**	**0.071**

Table 2: The RMS error averaged over 4 sentences for width (W) and height (H) of the mouth of the real speakers and their corresponding 3D heads. Values in bold means the decreased RMS error. Width and height error=$\pm$0.001.

and which are generated using front- and side-view photos (i.e. data set 4) are much closer to the ground truth trajectory. Thus, using different intensities of viseme data in the construction of the 3DMM, as well as one extra photograph in the construction of the 3D head, improves the performance of the resulting 3D lip motions.

Table 2 shows the RMSE results averaged over 4 sentences for width and height of the mouth of the real speakers and their corresponding 3D heads. The 3DMMs that contain 161 poses and which are generated using both front- and side-view photographs give the lowest RMSE scores for height for all the speakers and width for four out of six of the speakers. For the 3D heads that contain 161 poses, a t-test suggests a significant difference in RMSE results for the 3D heads that use front- and side-view photos versus front-view photos only (p=0.0292 for width and p=0.0009 for height). Also, there is a significant difference for height between the 3D heads containing 161 poses and 17 poses that are generated using front- and side-view photos (p=0.0135), although there is no significant difference for the width (p=0.0967).

4. CONCLUSIONS AND FUTURE WORK

This paper has presented a 3D talking head based on fitting a 3DMM, created using synthetic data, to 2D video frames of a real speaker. The experiments show that increasing the number of 3D head poses (different viseme intensities) to train the 3DMM improves the performance of the 3D lip motions. In addition, using both a front- and side-view photo in the construction of the neutral pose 3D head further improves the results in comparison to just using the front-view photo. Future evaluation work will make use of side-view videos (from [16]) and a subjective evaluation of the resulting animation will also be conducted.

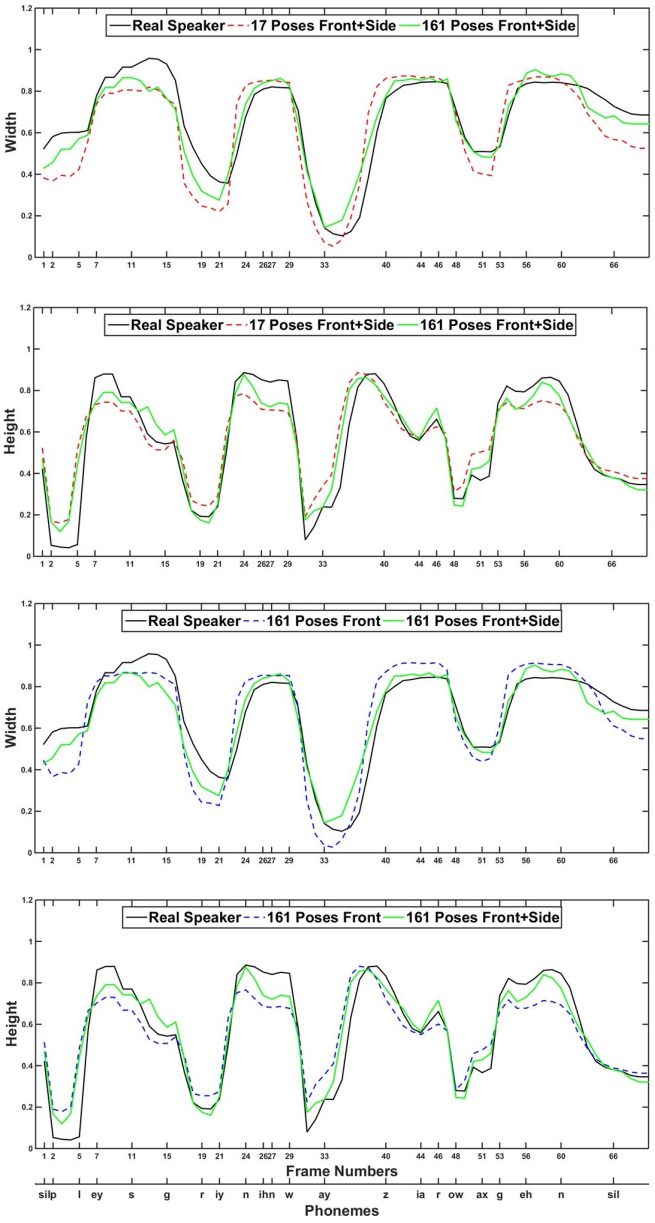

Fig. 4: Width and height of mouth trajectories of 2D frames of the real speaker (ID:S17) and the corresponding 3D heads. Top two compare height and width between 17 and 161 poses (both with front- and side-view photos), while the bottom two compare height and width between front- view photo only and front- and side-view photos (both with 161 poses).

5. REFERENCES

[1] Tsuhan Chen and Ram R Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852, 1998.

[2] Michael M Cohen, Dominic W Massaro, et al., "Modeling coarticulation in synthetic visual speech," *Models and techniques in computer animation*, vol. 92, pp. 139–156, 1993.

[3] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh, "Jali: an animator-centric viseme model for expressive lip synchronization," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 127, 2016.

[4] Pierre Badin, Frédéric Elisei, Gérard Bailly, and Yuliya Tarabalka, "An audiovisual talking head for augmented speech generation: models and animations based on a real speakers articulatory data," *Articulated Motion and Deformable Objects*, pp. 132–143, 2008.

[5] Slim Ouni and Guillaume Gris, "Dynamic lip animation from a limited number of control points: Towards an effective audiovisual spoken communication," *Speech Communication*, vol. 96, pp. 49–57, 2018.

[6] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter, "Reanimating faces in images and video," in *Computer graphics forum*. Wiley Online Library, 2003, vol. 22, pp. 641–650.

[7] Weicheng Xie, Linlin Shen, Meng Yang, and Jianmin Jiang, "Facial expression synthesis with direction field preservation based mesh deformation and lighting fitting based wrinkle mapping," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 7565–7593, 2018.

[8] Yong Zhao, Meshia Cédric Oveneke, Dongmei Jiang, and Hichem Sahli, "A video prediction approach for animating single face image," *Multimedia Tools and Applications*, pp. 1–22, 2018.

[9] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews, "Dynamic units of visual speech," in *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*. Eurographics Association, 2012, pp. 275–284.

[10] Pablo Garrido, Michael Zollhöfer, Chenglei Wu, Derek Bradley, Patrick Pérez, Thabo Beeler, and Christian Theobalt, "Corrective 3d reconstruction of lips from monocular video.," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 219–1, 2016.

[11] Jun Yu, Chen Jiang, Rui Li, Chang-Wei Luo, and Zeng-Fu Wang, "Real-time 3d facial animation: From appearance to internal articulators," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[12] Lingyun Yu, Jun Yu, and Qiang Ling, "Deep neural network based 3d articulatory movement prediction using both text and audio inputs," in *International Conference on Multimedia Modeling*. Springer, 2019, pp. 68–79.

[13] Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic, "End-to-end learning for 3d facial animation from speech," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 361–365.

[14] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 94, 2017.

[15] Volker Blanz and Thomas Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194.

[16] Najwa Alghamdi, Steve Maddock, Ricard Marxer, Jon Barker, and Guy J Brown, "A corpus of audio-visual lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[17] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, Willem P Koppen, William Christmas, Matthias Rätsch, and Josef Kittler, "A multiresolution 3d morphable face model and fitting framework," in *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[18] Zhen-Hua Feng, Patrik Huber, Josef Kittler, William Christmas, and Xiao-Jun Wu, "Random cascaded-regression copse for robust facial landmark detection," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 76–80, 2015.

[19] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[20] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu 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.