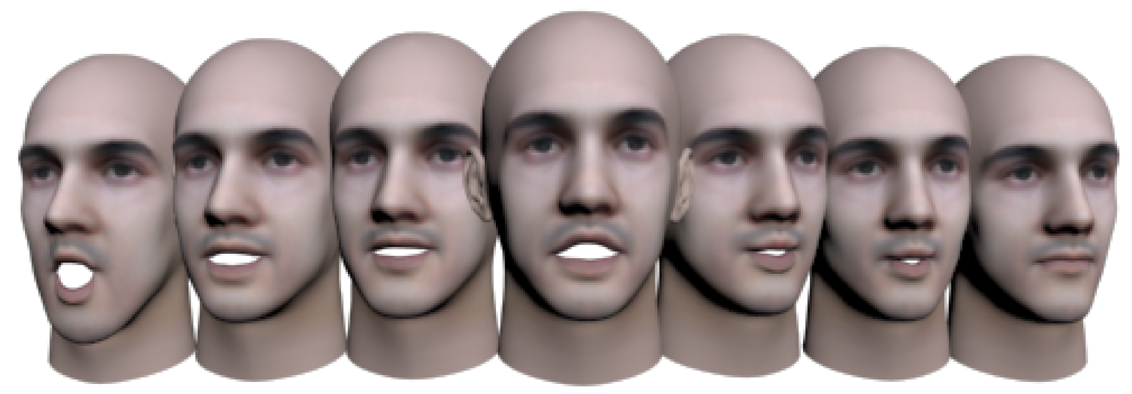
3D VISUAL SPEECH ANIMATION USING 2D VIDEOS

Rabab Algadhy, Yoshihiko Gotoh, Steve Maddock
Department of Computer Science, University of Sheffield, United Kingdom



Introduction

- 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.
- The experiments address two main questions:
  - Q1. Would using different intensities of the same viseme shape, when constructing the 3DMM produce better animation results?
  - Q2. 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?
- We use ground-truth data (the front-view videos of a speaker [1]) to compare the final synthetic 3D animation results against.

Method

3D Morphable Model (3DMM)

- FaceGen software is used to produce synthetic head poses.
- Principal Component Analysis (PCA) can be applied to the vertices to generate a 3DMM.
- A new pose can be generated as follows:

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

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

Mapping 2D to 3D

- Mapping 2D video of a speaker to the 3DMM uses Huber et al's method [2].
- Facial features of a real speaker in a video are tracked using the random cascaded-regression copse (R-CR-C) approach [3].
- Given 51 2D landmarks and the corresponding 3D landmarks (figure 1) a pose of the face is estimated using the Gold Standard Algorithm [2].

Figure 1: The facial landmark points.

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

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

where N is the number of landmarks, y is the 2D landmarks represented in homogeneous coordinates, σ_{2D}^2 is an ad hoc variance of these landmarks, and $y_{m2D,i}$ is the projected 3D landmarks to a 2D plane using the camera matrix.

Contact Details

Web http://staffwww.dcs.shef.ac.uk/people/R.Algadhy/
Email rralkasah1@sheffield.ac.uk

Experiments and Results

Data sets

- Four data sets were used to build different 3DMMs for a speaker - see Table 1.

Number of poses	Front-view	Front- & side-views
17 poses	Data set 1	Data set 3
161 poses	Data set 2	Data set 4

Table 1: Four data sets that were used to build different 3DMMs for a speaker.

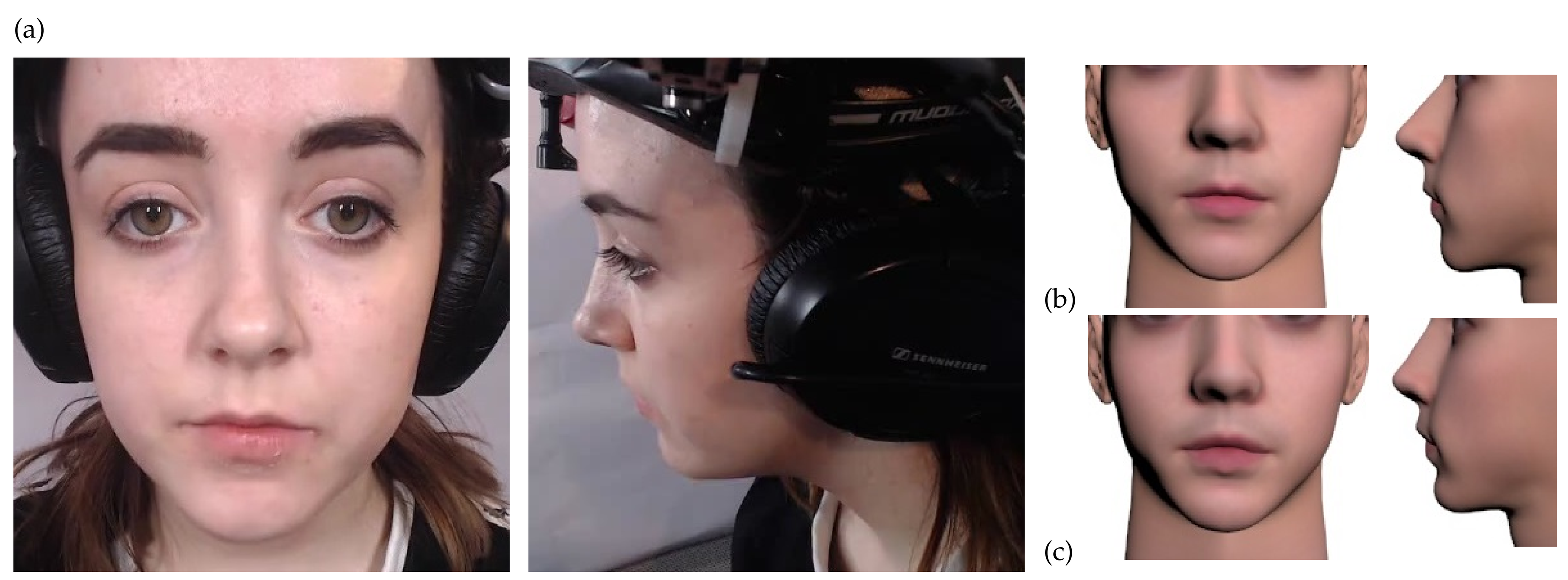


Figure 2: (a): Front (left) and side (right) photographs of a real speaker (ID: S32); (b and c): front and side view of the corresponding 3D heads generated using front photograph only (b), and front and side photographs (c). The lips are more protruded in (c).

Evaluation

- 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 [1] were used for validation.
- For the comparison, Faceware Analyser 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 (width and height of the mouth).
- Given the measurements values, the root mean square error (RMSE) over a sentence was used to evaluate the effectiveness of each 3DMM.

Figure 3: 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.

Results and discussion

- The performance of the animated 3D lips improves when a larger number of 3D head poses are used to train the 3DMM, and further improves when front- and side-view photos (figure 2) are used to generate the initial neutral head pose in FaceGen, as shown in figure 3.
- 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).

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 decreased RMS error. Width and height error=±0.001.

- Whilst all the trajectories generated using the animation pipeline generally follow the real speaker's trajectory, the trajectories of the 3D heads generated using data set 4 are much closer to the ground truth trajectory, as shown in figure 4.

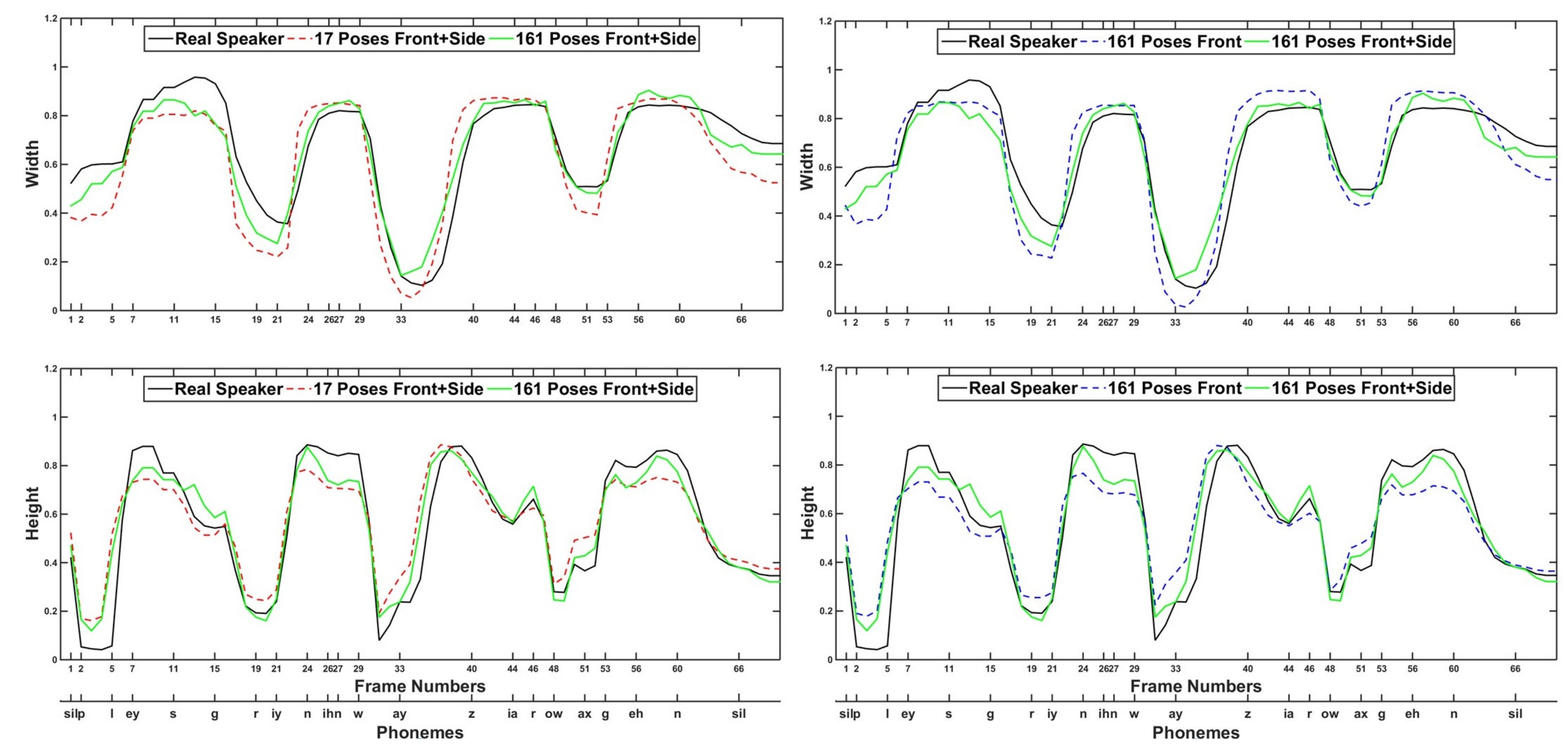


Figure 4: Width and height of mouth trajectories of 2D frames of the real speaker (ID:S17) and the corresponding 3D heads whilst uttering the sentence "place green in y zero again".

Conclusions

- The performance of the 3D lip motions is improved when the number of 3D head poses used to train the 3DMM is increased.
- It is also improved when a front- and side-view photo is used in the construction of the neutral pose 3D head.
- Future work: evaluation of lip motion from the side-view.

References

[1] Alghamdi, Maddock, Marxer, Barker, and Brown, "A corpus of audio-visual lombard speech with frontal and profile views," *JASA*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[2] Huber, Hu, Tena, Mortazavian, Koppen, Christmas, Rätsch, and Kittler, "A multiresolution 3d morphable face model and fitting framework," in *Proc. 11th International Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[3] Feng, and Kittler Huber, Christmas, and Wu, "Random cascaded-regression copse for robust facial landmark detection," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 76–80, 2015.