# On Evaluating Synthesised Visual Speech

*Barry-John Theobald*[1]*, Nicholas Wilkinson*[1]*, and Iain Matthews*[2]

[1]School of Computing Sciences University of East Anglia, Norwich, UK.
[2]Weta Digital Limited, Wellington, New Zealand

{bjt,nw}@cmp.uea.ac.uk, iainm@wetafx.co.nz

## Abstract

This paper describes issues relating to the subjective evaluation of synthesised visual speech. Two approaches to synthesis are compared: a text-driven synthesiser and a speech-driven synthesiser. Both synthesisers are trained using the same data and both use the same model for rendering the synthesised visual speech. Naturalness is used as a performance metric, and the naturalness of real visual speech re-rendered on the same model is used as a benchmark. The naturalness of the text-driven synthesiser is significantly better than the speech-driven synthesiser, but neither synthesiser can yet achieve the naturalness of real visual speech. The impact of likely sources of error apparent in the synthesised visual speech is investigated. Similar forms of error are introduced into real visual speech sequences and the degradation in naturalness is measured using the same naturalness ratings used to evaluate the performance of the synthesisers. We find that the overall perception of sentence-level utterances is severely degraded when only a small region of an otherwise perfect rendering of the visual sequence is incorrect. For example, if the visual gesture for only a single syllable in an utterance is incorrect, the overall naturalness of this real sequence is rated lower than the text-based synthesiser.
**Index Terms**: evaluation, visual speech synthesis, active appearance models

## 1. Introduction

A visual speech synthesiser has the goal of mapping a representation of an utterance to the associated movements of the visible articulators. This can be a direct mapping from encoded acoustic speech, i.e., speech-driven synthesis [1–5], or an indirect mapping from a phonetic transcription of an utterance, i.e., text-driven synthesis [6–13]. The advantage of mapping directly from acoustic speech is the articulators are positioned to form speech sounds, so the relationship between the auditory and the visual modalities can be learned. The main disadvantage is that only short-term information is exploited — typically frames are considered in isolation, or immediately surrounding frames are concatenated to provide minimal temporal context. Indirectly mapping from a phonetic transcription has the advantage that longer-term coarticulation effects can be estimated from the phonetic context. For example, knowledge of the phonetic context allows the "best" candidate samples to be selected from a codebook of real data. Thus subtle variation apparent in natural speech production can be retained in the synthesised visual speech. The main disadvantage is that a phonetic transcription is required, so these approaches do not readily lend themselves to real-time applications. For an overview of audio-visual speech synthesis see [14, 15].

In this paper we focus on the subjective evaluation of synthesised visual speech using both a speech-driven and a text-driven synthesiser. In particular we are interested in contrasting the two approaches using the same model and the same training data, and we seek to quantify the effect of likely sources of error on the perceived naturalness of the synthesised visual speech.

## 2. Active Appearance Models

Active Appearance Models (AAMs) [16] belong to a class of linear, generative, parametric model. An AAM is comprised of two components: a model of shape variation and a model of appearance variation. This makes the use of such models in speech animation attractive as both the facial geometry and the texture is represented using a single model.

The *shape*, $\mathbf{s}$, of an AAM is defined by the concatenation of the $x$ and $y$-coordinates of $n$ vertices that form a two-dimensional (2D) triangulated mesh: $\mathbf{s} = (x_1, y_1, \ldots, x_n, y_n)^T$. To construct the model, a set of training images are annotated manually by aligning the shape vertices with the corresponding facial features as they undergo non-rigid variation. PCA is then applied to the shapes to a provide compact model that allows linear variation:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i, \tag{1}$$

where the coefficients $p_i$ are the shape parameters. The base shape $\mathbf{s}_0$ is the mean shape and the vectors $\mathbf{s}_i$ are the (reshaped) eigenvectors corresponding to the $m$ largest eigenvalues.

The *appearance*, $A(\mathbf{x})$, of an AAM formed of the pixels that lie inside the base mesh, $\mathbf{x} = (x, y)^{\mathrm{T}} \in \mathbf{s}_0$. The annotated training images from which the shape component was constructed are shape-normalised by warping each from the hand-labels to the base shape using a piece-wise affine warp [16]. PCA is then applied to the resulting images to provide a compact model that allows linear variation in appearance:

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{l} \lambda_i A_i(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbf{s}_0, \tag{2}$$

where the coefficients $\lambda_i$ are the appearance parameters. The base appearance $A_0$ and appearance images $A_i$ are again the (reshaped) mean and eigenvectors corresponding to the $l$ largest eigenvalues respectively.

Face images are synthesised from a set of AAM parameters by first applying the shape parameters, $\mathbf{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$, to generate the shape, $\mathbf{s}$, of the AAM using Eq. (1). Next the appearance parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_l)^{\mathrm{T}}$ are used to generate the AAM appearance image, $A(\mathbf{x})$, using Eq. (2). Finally a piece-wise affine warp is used to warp $A(\mathbf{x})$ from $\mathbf{s}_0$ to $\mathbf{s}$.

# 3. Synthesising Visual Speech

This paper considers two approaches to synthesising visual speech: one driven from speech acoustics and one driven from text.

### 3.1. Speech-Driven Synthesis

Acoustic speech from the Messiah corpus, see [17] for details, is encoded as MFCCs at 10ms intervals and the corresponding 25Hz AAM parameters are up-sampled to match the audio frame-rate. This up-sampling is achieved by fitting cubic splines through parameters, then resampling the splines at four times the original frame-rate. At each time-step eleven frames of AAM parameters are concatenated (five frames either side of each) to provide temporal context.

An artificial neural network (ANN) is trained to learn the mapping from MFCCs to AAM parameters. To maximise the limited training data available, a leave-one-out training strategy is adopted. The corpus is formed of 279 sentences, and a separate network is trained for each sentence. The parameters for the sentence of interest are not included in training. Each network has three-layers: an input layer, a 50-node hidden layer, and the output layer.

Given a trained network, visual speech is synthesised by first computing the MFCCs from the novel acoustic speech. These are then input to the network to generate a sequence of AAM parameters. At each time step, the output feature vector contains AAM parameters for 11 frames, the centre frame itself and five frames either side. The final trajectory of parameters is generated by cross-blending the neighbouring regions of the feature vectors. The resulting parameters are smoothed using a cubic smoothing spline, Eq. 5, with a smoothing parameter of 0.9 before being applied to the model to generate the sequence of synthesised visual speech synchronised to the novel audio.

### 3.2. Text-Driven Synthesis

The text-based approach [13] measures the similarity between phoneme pairs in terms of AAM parameters using:

$$S_{ij} = e^{-\gamma\left(\sum_{k=1}^{m+l}\sum_{n=1}^{5}\left[\left(v_i P_{kn}^{i} - v_j P_{kn}^{j}\right) w_k\right]^2\right)}. \qquad (3)$$

$P^i$ and $P^j$ are the mean of phonemes $i$ and $j$ computed from examples in the corpus. The first summation is over the dimensions of the AAM and the second over samples equally spaced over the phoneme sub-trajectories. The parameters $v_i$ are inversely proportional to the variance of the $i^{th}$ phoneme, and $w_k$ reflects the significance of the $k^{th}$ AAM parameter. The approach is fully described in [13]. The similarities obtained using this measure match intuitive expectation. For example, {/b/, /p/, /m/}, {f/, /v/}, {/tʃ/, /dʒ/, /ʃ/, /ʒ/}, etc., are all considered most similar to one another.

Sequences are synthesised from text by selecting parameter sub-trajectories for each phoneme from the original corpus based on the distance between the desired context and those available in the corpus. The distance itself is based on the similarities measured during training, and is given by

$$\delta_j = \sum_{i=1}^{C} \frac{\mathbf{S}_{l_{ij}}}{i+1} + \sum_{i=1}^{C} \frac{\mathbf{S}_{r_{ij}}}{i+1}, \qquad (4)$$

where $C$ is the context width and $\mathbf{S}_{l_{ij}}$ and $\mathbf{S}_{r_{ij}}$ are the similarity between the left and right contexts respectively. The selected sub-trajectories for the *best* examples are temporally normalised to the desired duration, concatenated, smoothed using a cubic smoothing spline (with a smoothing parameter of 0.5) and applied to the model to create the synthesised sequences. Note, the text-based synthesiser requires a higher degree of smoothing since there are no smoothness constraints in the unit selection. We evaluate the affect of the difference in smoothing in Section 4.2.

# 4. Evaluation

An obvious method for measuring the quality of synthesised visual speech is to re-synthesise a set of test sentences for which the original speech is available and use some objective measure of similarity between the original and synthesised parameters. Typical measures include the correlation between original and synthesised parameters, the RMS error between image pixels in the two sequences, or the error in coordinates for key points located about the face (jaw, lip position, etc.). However, people never repeat the same utterance in exactly the same way, so undoubtedly there will be differences in the parameters representing repetitions of the same utterance. It follows therefore that a synthesiser will unlikely be able to exactly recreate an occurrence of an utterance. The question is: *are the differences between the reference and synthesised parameters significant*? The differences might be manifested in the variation observed in natural speech production, so are not perceived by viewers. Conversely the difference might result from errors in the synthesised speech. It is therefore the *perception* of the quality of the speech that is important, and this is difficult to measure using only objective scores.

In evaluating visual speech generated by our synthesisers we consider both objective and subjective quality measures and we are particularly interested in the relationship between the two. For example, if an objective measure exists that relates directly to subjective opinion, this objective measure need only be used in the future. The advantage of objective measures is they are repeatable, they can be computed automatically, they are much less time consuming, and the experiments are cheaper to conduct than subjective tests.

Test data for the experiments described in the following sections are generated by re-synthesising each sentence in the training corpus using a leave-one-out strategy, providing maximal use of the limited training data available.

### 4.1. Comparing real and model-generated faces

To provide a benchmark for quantifying the perceived naturalness of AAM-based visual speech synthesisers, we first run an experiment to investigate: the effect of the form of the presentation of the stimuli, and the effect on the perceived naturalness of visual speech rendered using an AAM.

#### *4.1.1. Method*

Fifteen sentences were selected randomly from the Messiah corpus [17]. The video sequences for these sentences were then processed by firstly masking the face in the original video and displaying against a constant (white) background, and secondly projecting the video onto an AAM and re-rendering the video from the resulting parameters. An example frame for each condition is shown in Figure 1. In both cases the visual speech presented to viewers is real, so any perceived degradation in naturalness cannot be as a result of synthesis.

Five graduate students were recruited to participate in the test. Each was briefed to inform them of the test procedure, but

(A) (B)

Figure 1: The same video frame displayed in two conditions: (A) the face in the original video masked from the background, and (B) the same video frame reconstructed using an AAM.

they were not provided with information about the treatment of the video or the purpose of the test. They were told they would see only the face in the video, and that in some sequences the face *may* have undergone some form of processing. They were instructed to ignore the image quality of the video and that they should focus their attention to rate only the *naturalness of the talking face*. The video was synchronised to accompanying acoustic speech.

The audiovisual sequences were presented to participants in a randomised order using a GUI. This interface used a slider to collect naturalness ratings and two buttons that allowed a sequence to be repeated and the naturalness score to be registered. Participants were told that one end of the slider equated to unnatural visual speech, while the other equated to perfectly natural video. Participants were free to repeat sequences as many times as required.

#### *4.1.2. Results*

The naturalness scores for a given sequence are averaged over all participants, and the mean score for the sentence pairs (video/AAM) subject to a Kruskal-Wallis test to determine if the differences between treatments is significant. The responses are summarised in Table 1 and Figure 2.

Table 1: *Pooled naturalness scores for AAM rendered and original video sequences.* Median *is the median score over all sequences within a treatment, and* MAD *is the median of absolute deviations. The possible range of scores is zero (unnatural) to 50 (entirely natural).*

| Treatment | $n$ | Median | MAD |
|---|---|---|---|
| Video | 15 | 42.4 | 1.4 |
| AAM | 15 | 39.6 | 1.6 |
| $\chi^2 = 8.2$ | | $p < 0.005$ | |

The median score of 42.4 (out of 50) for the original video sequence (with the face masked from the background) suggests that presenting a *floating* face, as illustrated in Figure 1, *does* have an impact on the perceived naturalness of visual speech. The video sequences presented to participants were in no way processed, other than masking non-face pixels, so the visual speech was exactly that produced while speaking the accompanying acoustic speech. The re-synthesised face could be composed with a background sequence, e.g. [6, 7, 11], however we are working towards a real-time system to animate head pose in addition to visual speech. Composing into a background

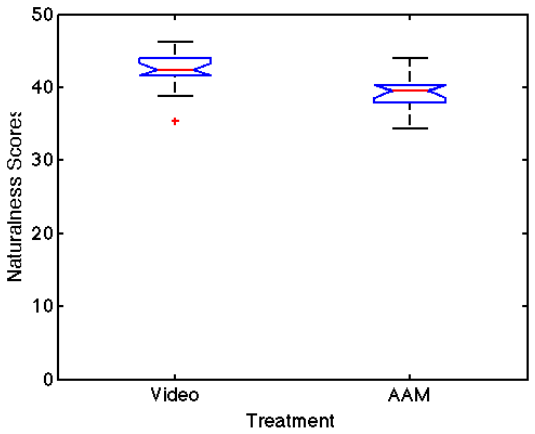


Figure 2: The naturalness scores for video sequences where the face pixels are masked and displayed against a white background, and the same video sequences re-rendered using an AAM. The boxes show the upper and lower quartiles, the red horizontal bar marks the median, and the error bars denote the extent of the data for each treatment. The difference in naturalness between an AAM encoded sequence and a video sequence **is** significant ($p < 0.005$).

sequence then becomes difficult in real-time as the most suitable background frame must be selected from the video, and the resulting image must be post-processed to ensure a seamless blend between the face and the background images. A second conclusion that can be drawn from these results is that the difference in naturalness between AAM generated and original video sequences is significant ($p < 0.005$). The truncation of the model likely loses very subtle information, without which the sequences appear smoothed. In addition, although instructed to ignore image quality and focus only on the speech dynamics, blurring artefacts might also influence viewer perception of naturalness.

The scores obtained from this experiment can be used as an upper bound for the expected performance of the synthesisers. We cannot realistically expect a synthesiser to generate speech that is perceived as perfectly natural (a score of 50/50) when original video sequences encoded using an AAM score only around 40.

### 4.2. Effect of smoothing on perceived naturalness

Parameter trajectories generated using both the text-driven and speech-driven synthesisers require smoothing. The text-based system has no smoothness constraints in the unit selection, and the speech-driven system has no knowledge of past/future frames. Consequently the parameter trajectories are noisy, which results in *jitter* in the facial features in the synthesised video sequences. To combat this the parameter sequences are smoothed before being applied to the model by fitting a cubic smoothing spline to the sequences, then re-sampling the smoothed trajectories.

The aim of this experiment is to determine the significance of this parameter smoothing. We note here the smoothing parameter is different for both systems. In the case of the speech-driven synthesiser, temporally-adjacent AAM features are concatenated and the neighbouring regions cross-blended, thus there is some degree of smoothing during synthesis. For

the text-driven synthesiser discontinues at the segment boundaries result in abrupt changes in the facial features. Thus more smoothing is required for the text-based synthesiser.

The smoothing spline used in both synthesis approaches minimises the functional:

$$L = \zeta \sum_{i=0}^{k} (p_i - S(s_i))^2 + (1-\zeta) \sum_{i=0}^{k-1} \int_{s_i}^{s_{i+1}} \left( \frac{d^2}{ds^2} S_i(s) \right)^2 ds, \quad (5)$$

where the smoothing parameter, $\zeta$, specifies the trade-off between a natural cubic spline interpolation of the data, or no smoothing ($\zeta = 1$) and the least squares fit, or maximally smoothed trajectory ($\zeta = 0$). For the text-based synthesiser the parameters are smoothed using $\zeta = 0.5$, and for the speech-driven synthesiser the parameters are smoothed using $\zeta = 0.9$. The impact of smoothing on the perceived naturalness of real visual speech is again evaluated using a subjective test.

#### *4.2.1. Method*

Thirty sentences were selected randomly from the Messiah corpus [17]. All were encoded using an AAM and the original and smoothed sequences were re-synthesised. The audiovisual sequences were then played to participants in a randomised order and participants were asked to rate the naturalness of the speech dynamics using the same interface described previously.

Eighteen undergraduate and postgraduate students took part in the test and all were paid for their participation. Participants were briefed to inform them as to what was required and they were told they would see only the face in the video, as illustrated in Figure 1, and that in some sequences the face *may* have undergone some form of processing. They were instructed to ignore the image quality of the video and that they should focus their attention to rate only the *naturalness of the talking face*. The video was synchronised to accompanying acoustic speech.

#### *4.2.2. Results*

The naturalness scores for a given sequence are averaged over all participants, and the mean score for the sentence triple subject to a Kruskal-Wallis test to determine if the differences between treatments is significant. The responses are summarised in Table 2 and Figure 3.

Table 2: *Pooled participant ratings for AAM rendered video before and after smoothing. There is no significant effect of smoothing using $\zeta = 0.9$ ($p<0.85$), although smoothing using $\zeta = 0.5$ does have a significant impact on naturalness ($p<0.005$).*

| Treatment | $n$ | Median | MAD |
|---|---|---|---|
| $\zeta = 1$ | 30 | 34.1 | 2.6 |
| $\zeta = 0.9$ | 30 | 33.4 | 2.1 |
| $\zeta = 0.5$ | 30 | 27.9 | 1.3 |
| $\chi^2 = 0.04$ | | $p < 0.005$ | |

The scores obtained from this experiment again reflect the expected upper bound on the performance of the two synthesis techniques. The stimuli presented to participants were derived from smoothed AAM parameters measured in original video sequences. Thus, parameter trajectories generated *exactly* by the synthesiser, would still suffer the impact on naturalness as a result of smoothing.

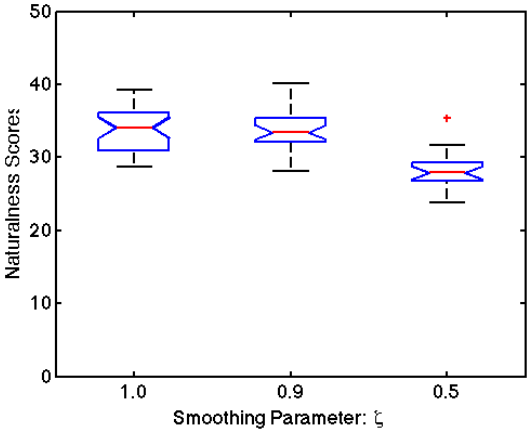


Figure 3: The naturalness scores for each of the three smoothing treatments ($\zeta = \{1.0, 0.9, 0.5\}$. There is no significant affect of smoothing using $\zeta = 0.9$ ($p<0.85$), although smoothing using $\zeta = 0.5$ does have a significant impact on naturalness ($p<0.005$).

### 4.3. Contrasting speech-driven and text-driven synthesis

The aim of this experiment is to contrast speech-driven synthesis and text-driven synthesis. The advantage of the approaches adopted here is that both synthesisers are trained *using the same data* and both are evaluated using *the same participants*.

#### *4.3.1. Method*

Fifteen sentences were selected randomly from the Messiah corpus [17] and synthesised using both the speech- and the text-driven synthesisers. In addition, the original sequences were re-rendered using the AAM. All video sequences were synchronised to the accompanying acoustic speech. The sequences were then played to participants in a randomised order and participants were asked to rate the naturalness of the speech dynamics using the same interface described previously.

Eighteen undergraduate and postgraduate students took part in the test and all were paid for their participation. Participants were briefed to inform them as to what was required and they were told they would see only the face in the video, as illustrated in Figure 1, and that in some sequences the face *may* have undergone some form of processing. They were instructed to ignore the image quality of the video and that they should focus their attention to rate only the *naturalness of the talking face*.

#### *4.3.2. Results*

The naturalness scores for a given sequence are averaged over all participants, and the mean score for the sentence triples are subject to a Kruskal-Wallis test to determine if the differences between the synthesis methods and the AAM re-rendered video are significant. The responses are summarised in Table 3 and Figure 4.

The overall perception of the naturalness of the synthesised sequences is disappointingly low. However, we note the text-driven synthesiser achieves a naturalness score that is **not** significantly different from the corresponding smoothed ($\zeta = 0.5$) AAM re-rendered video. Furthermore, inspection of the video sequences generated using the speech-driven synthesiser suggest two possible causes of the low naturalness ratings.

Table 3: Pooled viewer ratings for both speech- and text-driven synthesis, and video re-rendered using an AAM. AAM re-rendered video is perceived as more natural than both synthesis methods ($p < 0.005$), and the text-driven synthesiser is perceived as more natural than the speech-driven synthesiser ($p < 0.015$).

| Treatment | $n$ | Median | MAD |
|---|---|---|---|
| Video | 15 | 38.71 | 1.12 |
| Text-Driven | 15 | 27.47 | 3.99 |
| Speech-Driven | 15 | 22.35 | 2.25 |
| $\chi^2 = 31.97$ | | $p < 0.005$ | |

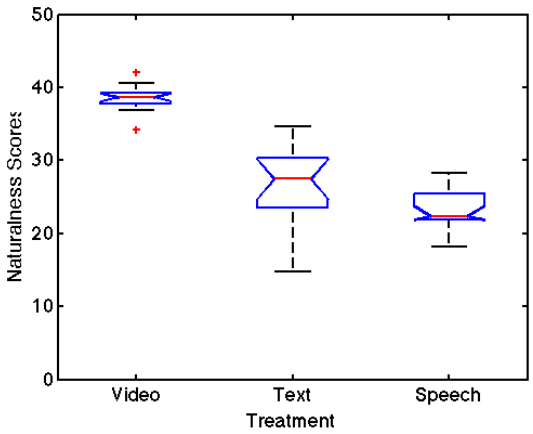


Figure 4: The naturalness scores for each of the three treatments: video re-rendered from AAM parameters, text-driven synthesis, and speech-driven synthesis. AAM re-rendered video is perceived as significantly better than both synthesis methods ($p < 0.005$), and the text-driven synthesiser is perceived as more natural than the speech-driven synthesiser ($p < 0.015$).

Firstly, the visual gestures are, on the whole, well re-produced. Occasionally however there are isolated gestures that appear to *stand-out* as being obviously incorrect. A typical example is shown in Figure 5, where the syllable over frames 60–68 is very under-articulated. We next describe an experiment carried out to determine the effect of this form of error.

### 4.4. Effect of errors in isolated visual gestures

To determine the impact on naturalness of (individual) gestures incorrectly re-synthesised (e.g., see Figure 5), the effect of other potential sources of error must be isolated. For example, although the overall shape of the trajectories in Figure 5 are broadly similar, some synthesised gestures are slightly under-articulated (e.g., at frame 34), whist others are slightly over-articulated (e.g., around frame 53–55). These subtle over- and under-articulation must be removed so the only errors present in the trajectories isolated gestures incorrectly produced.

#### *4.4.1. Method*

The original parameters for ten utterances were selected at random from the Messiah corpus [17] and re-rendered using an AAM. For each utterance, a single syllable was selected and the parameters representing that syllable were replaced with parameters from another syllable from elsewhere in the corpus, where

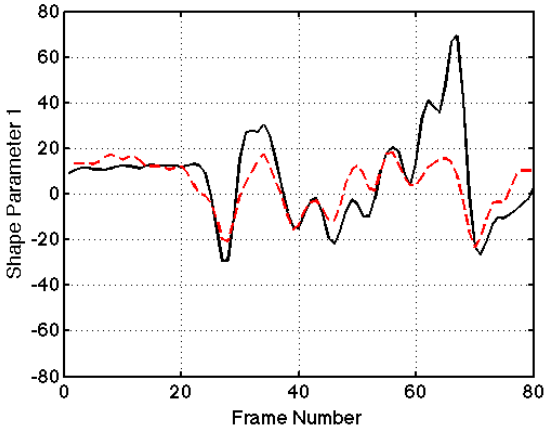


Figure 5: An example trajectory of the first shape parameter: (A) as measured in the video (black, solid line), and (B) synthesised by the speech-driven synthesiser (red, dashed line). Both trajectories correspond to the same sentence. Overall the trajectory is well reproduced, except the visual gesture between frames 60–68 (mouth closure).

the parameters were normalised to the duration of the original syllable. To ensure a seamless blend, the parameters near the concatenation boundaries were smoothed. Thus, for each utterance we have two identical (real visual speech) sequences: one is the original parameters replayed on the model, and the second is same sequence except one syllable is incorrectly rendered. All sequences were synchronised to the original auditory speech, and sequences were played to participants in a randomised order and participants were asked to rate the naturalness of the speech dynamics using the same interface described previously.

Seven participants, both undergraduate and graduate students, were recruited and paid for their participation. All were briefed to inform them as to what was required and they were told they would see only the face in the video, as illustrated in Figure 1, and that in some sequences the face *may* have undergone some form of processing. They were instructed to ignore the image quality of the video and that they should focus their attention to rate only the *naturalness of the talking face*.

#### *4.4.2. Results*

The naturalness scores for a given sequence are averaged over all participants, and the mean score for the sentence pairs are subject to a Kruskal-Wallis test to determine the impact of the single error on the naturalness of the overall utterance. The responses are summarised in Table 4 and Figure 6.

Table 4: Pooled viewer ratings for (T1) AAM re-rendered video, and (T2) the **same** sequence after introducing an error in the visual gesture for a single syllable.This error does significantly impact on the perceived naturalness of the *whole* sequence ($p < 0.0002$).

| Treatment | $n$ | Median | MAD |
|---|---|---|---|
| T1 | 10 | 42.6 | 1.48 |
| T2 | 10 | 23.5 | 4.15 |
| $\chi^2 = 41.32$ | | $p < 0.0002$ | |

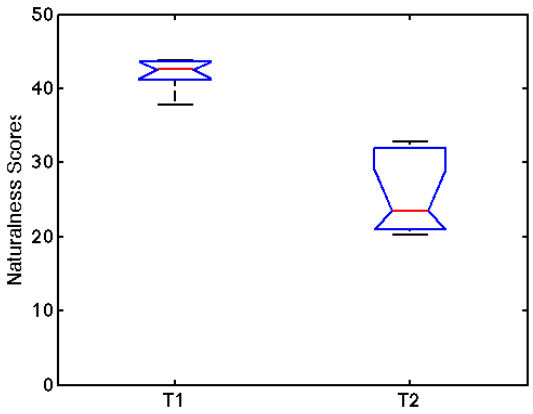


Figure 6: The naturalness scores for original video (*T1*), and original video re-rendered after introducing an error in a single syllable toward the end of the sentence (T2). The differences between the treatments is significant ($p < 0.0002$).

Representing the visual gesture for only a single syllable *does* significantly degrade the perceived naturalness of the *overall* sequence ($p < 0.0002$). Note, there is no significant difference in the perceived naturalness of the processed sequences presented here and the perceived naturalness ratings of the speech-driven synthesiser ($p < 0.35$), and the sequences generated by text-driven synthesiser are perceived as significantly more natural than original AAM parameters with a single gesture incorrectly rendered.

## 5. Conclusions

This paper has contrasted two approaches for synthesising visual speech: one driven by text and one driven by voice. In terms of the naturalness of synthesised sentences, the text-driven synthesiser performs significantly better than the speech-driven synthesiser. We have also described experiments carried out to quantify likely reasons for the low naturalness ratings of both synthesisers compared with AAM re-rendered video. These have shown that taken in isolation, these naturalness measures are not entirely informative of *absolute* performance. For example, the perceived naturalness of an entire utterance (sentence) is significantly degraded when only a single syllable is erroneous, even if the remainder of the sequence is perfect. While evaluating naturalness using sentence level units is useful, after all the longer-term properties of the visual speech must ultimately be considered, they should not be used in isolation. Other tests of the short-term properties of the accuracy of the production of speech gestures can be measured, e.g., [2], to give a more localised measure of performance, and individual components of the synthesiser could be isolated and tested using, for example, the point-light technique [18].

## 6. Acknowledgements

The authors gratefully acknowledge the support of EPSRC (EP/D049075/1) for funding.

## 7. References

[1] Brand, M., "Voice puppetry," In Proceedings of SIGGRAPH, 21–28, 1999.

[2] Cosker, D., Marshall, D., Rosin, S., Paddock, P., and Rushton, S., "Towards perceptually realistic talking heads: Models, metrics and McGurk," in Proceedings of the Symposium on Applied Perception in Graphics and Visualization, 2004.

[3] Gutierrez-Osuna, R., Kakumanu, P., Esposito, A., Garcia, O., Bojorquez, A., and Rudomin. I., "Speech-driven facial animation with realistic dynamics", IEEE Transactions on Multimedia, 7(1):33–42, February, 2005.

[4] Hsieh, C., and Chen, Y., "Partial linear regression for speech-driven talking head application," Signal Processing: Image Communication, 21:1–12, 2006.

[5] Theobald, B., and Wilkinson, N., "A Real-time Speech-Driven Talking Head using Active Appearance Models", In Proceedings of Auditory Visual Speech Processing, 264–269, 2007.

[6] Bregler, C., Covell, M., and Slaney, M., "Video rewrite: Driving visual speech with audio," in Proceedings of SIGGRAPH, 353–360, 1997.

[7] Ezzat, T., Geiger, G., and Poggio, T., "Trainable videorealistic speech animation," in Proceedings of SIGGRAPH, 388–398, 2002.

[8] Fagel, S., "Joint audio-visual unit selection — The JAVUS speech synthesizer", In Proceedings of the International Conference on Speech and Computer, 2006

[9] Govokhina, O., Bailly, G., Breton, G., and Bagshaw, P., "TDA: A new trainable trajectory formation system for facial animation," In Proceedings of Interspeech, 2474–2477, 2006.

[10] Grauwinkel, K., Dewitt, B., and Fagel, S., "Visualization of internal articulator dynamics and its intelligibility in synthetic audiovisual speech", In International Congress on Phonetic Sciences, 2007.

[11] Huang, F., Cosatto, E., and Graf, H., "Triphone based unit selection for concatenative visual speech synthesis," in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2:2037–2040, 2002.

[12] Massaro, D., "Perceiving Talking Faces", The MIT Press, 1998.

[13] Theobald, B., Bangham, J., Matthews, I., and Cawley, G., "Near-videorealistic synthetic talking faces: Implementation and evaluation", Speech Communication, 44:127–140, 2004.

[14] Bailly, G., Bérar, M., Elisei, F., and Odisio, M., "Audiovisual speech synthesis", International Journal of Speech Technology, 6:331–346, 2003.

[15] Theobald, B., "Audiovisual Speech Synthesis", International Congress on Phonetic Sciences, 285–290, 2007.

[16] Cootes, T., Edwards, G., and Taylor, C., "Active appearance models", In IEEE Transactions on Pattern Analysis and Machine Intelligence, 23:681–685, June, 2001.

[17] Theobald, B., "Visual speech synthesis using shape and appearance models", Ph.D. dissertation, University of East Anglia, Norwich, UK, 2003.

[18] Bailly, G., Gibert, G., and Odisio, M., "Evaluation of movement generation systems using the point-light technique", IEEE Workshop on Speech Synthesis, 27–30, 2002.