Relating Objective and Subjective Performance Measures for AAM-Based Visual Speech Synthesis

Barry-John Theobald* and Iain Matthews

***Abstract*—We compare two approaches for synthesizing visual speech using Active Appearance Models (AAMs): one that utilizes acoustic features as input, and one that utilizes a phonetic transcription as input. Both synthesizers are trained using the same data and the performance is measured using both objective and subjective testing. We investigate the impact of likely sources of error in the synthesized visual speech by introducing typical errors into real visual speech sequences and subjectively measuring the perceived degradation. When only a small region (e.g. a single syllable) of ground-truth visual speech is incorrect we find that the subjective score for the entire sequence is subjectively lower than sequences generated by our synthesizers. This observation motivates further consideration of an often ignored issue, which is to what extent are subjective measures correlated with objective measures of performance? Significantly, we find that the most commonly used objective measures of performance are not necessarily the best indicator of viewer perception of quality. We empirically evaluate alternatives and show that the cost of a dynamic time warp of synthesized visual speech parameters to the respective ground-truth parameters is a better indicator of subjective quality.**

***Index Terms*—Visual speech synthesis, visual speech evaluation, canonical correlation analysis, active appearance models.**

I. Introduction

VISUAL speech synthesis is the process of animating a face model to provide visual speech gestures that match an accompanying acoustic speech signal — see [1] for a broad overview. This is a sub-field of the broader topic of facial animation, and the typical graphics pipeline for animating faces usually involves four components: **1)** A geometric *model* of the face, **2)** an animation *rig* that parameterizes the deformation of the face model, **3**) a synthesis module that generates facial *animation* parameters, and **4**) an output module that *renders* the face model to produce the output visual speech animation.

1) Model: The face model typically represents the surface of a face as points which are connected to form the vertices of a mesh either in two or three-dimensions. The number of points varies from sparse sets containing only a few tens of points, to dense sets containing hundreds or more points.

2) Rig: The animation rig provides the mechanism for deforming the model either directly in terms of specific facial actions or gestures, where the parameters may be hand-crafted [2]–[4], derived using a data-driven approach [5]–[7], or computed indirectly via some form of underlying model, often based on facial muscles [8]–[10].

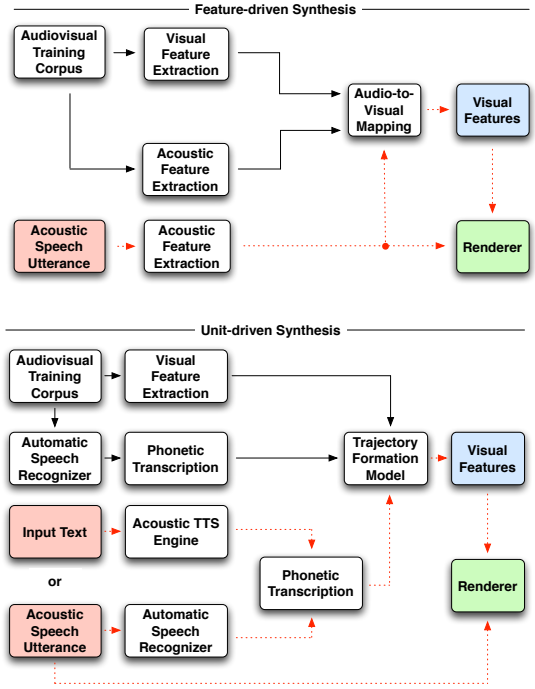

Fig. 1. An overview of: (top) feature-driven, and (bottom) unit-driven visual speech synthesis. Training is marked by the (solid) black arrows, whilst synthesis is marked by the (dotted) red arrows.

3) Animate: For visual speech synthesis, the animation parameter values are typically generated using either a feature-driven or a unit-driven approach. Feature-driven approaches [6], [11]–[21] generate animation parameter values as a direct mapping from parameterized acoustic speech on a video frame-by-frame basis. Unit-driven approaches [3], [22]–[30] utilize an indirect mapping of auditory to visual speech, where multi-frame animation curves are formed from typically phoneme, diphone or triphone-level representations of an utterance on a unit-by-unit basis. Both of these approaches are illustrated in Figure 1.

4) Render: For visual speech synthesis, generally the renderer will take one of two forms: an artist drawn computer graphics (CG) model [31], or an image-based technique [22], [25], [28], [32] that may also include a statistical model of the appearance variation [12], [26], [29], [30], [33]. The choice of renderer will largely be determined by the application.

B. Theobald is with the School of Computing Sciences, University of East Anglia, Norwich, UK. e-mail: b.theobald@uea.ac.uk. Iain Matthews is with Disney Research, Pittsburgh, USA. e-mail: iainm@disneyresearch.com

Manuscript received November 7, 2011, revised February 28, 2012.

For example, a speaker-independent or a real-time computer game system might adopt simple CG approaches as these offer efficiency and flexibility. However, real-time graphics methods generally lack videorealism — it is difficult to convince a viewer the video sequence contains a real person. The appearance of the face may be improved using texture-mapping and shading techniques, where an image of a face is texture-mapped onto the geometric model. However, the static nature of the texture becomes apparent as the model is animated. This can be overcome using complex and computationally expensive shading and lighting techniques [34], but these require specialist capture equipment or artist input and do not readily lend themselves to real-time synthesis of visual speech. Conversely, image-based rendering approaches can achieve close to videorealism, but they lack flexibility: often only the face region is animated, the range of facial gestures may be confined to examples seen during training, and unless complex retargeting methods are employed the identity of the talker is fixed. Hybrid approaches that utilize a statistical appearance model offer a convenient compromise between efficiency and realism [7], [29], [35], [36].

The Active Appearance Model (AAM) adopted in this paper (see Section: IV) encapsulates the model, the animation rig and the renderer in a single statistical model, meaning different synthesis strategies can be evaluated independently of the rest of the animation pipeline.

Despite the wealth of research into visual speech synthesis over the last two decades, there still are no common standards against which systems can be benchmarked. This is true for the nature of the tests that should be conducted, the environmental conditions in which the evaluation should be conducted, or the metrics against which systems are rated. The notion of evaluation of speech synthesizers (visual or otherwise) has been a "hot topic" for almost a decade [37]. There have been recent attempts to redress this issue, e.g. the introduction of the LIPS Challenge [38], but broad uptake by the community has been slow.

Part of the reason for the lack of standardized evaluation is that the particular metric used to quantify the performance of a synthesizer will largely depend on the application of the system. For example, realism might be measured in terms of the *appearance* of the face, the improvement in speech *intelligibility* provided by the synthesized talking face, or it might be measured more generally in terms of audiovisual *coherence*. It is possible that a system that scored highly in terms of one metric may score poorly with respect to another. For example, a system that is perceived as looking very natural (human-like) may not provide benefit in terms of intelligibility [39], yet a system that is less natural in terms of appearance might be significantly more intelligible [2].

Different measures of performance have been proposed that are either objective in nature, where a synthesized sequence is compared numerically in some way to a ground-truth representation, or subjective in nature, which uses human viewers to score the quality against some criteria. 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. The advantage of subjective measures is that they measure the perceived quality of the synthesizer.

In this paper we compare measures of performance for two different synthesis approaches: one that maps acoustic features to visual features, and one that utilizes only a phonetic transcription of the acoustic speech. In addition, we also seek to quantify the effect of likely sources of error on the perceived naturalness of the synthesized sequences, and we consider the relationship between objective performance measures and the perceived naturalness, an issue that has largely been ignored.

II. Related Work

Acoustic features used in feature-driven synthesis of visual speech have included Mel frequency cepstral coefficients (MFCCs) [14]–[16], [18]–[20]; filter-bank outputs [6]; line spectral pairs/frequencies (LSPs/LSFs) [17], [40]; formant frequencies [21]; linear prediction coefficients (LPCs) [12], [13] or perceptual LPCs (RASTA-PLP) [11]; and several forms of mapping function have been proposed, including vector-quantisation or a nearest neighbour look up [6]; regression [17], [19]; artificial neural networks [12], [13], [15], [18], [41]; hidden Markov models (HMMs) [11], [42]; and switching linear dynamical systems [14].

Rather than generating parameter values on a frame-by-frame basis, unit-driven synthesis instead forms trajectories of parameter values corresponding to a unit of interest (typically phones, diphones or triphones). The required input information might be derived from an automatic speech recognition (ASR) system if the corresponding acoustic speech is available, or existing (acoustic) text-to-speech synthesis rules can generate the required phoneme and timing sequence. Trajectory formation models have included concatenation [22], [24], [28]–[30], [43]–[46]; interpolation [3], [25], [26], [47]–[49]; probabilistic approaches [20], [23], [50]; and hybrid approaches [27], [51].

An advantage of unit-driven synthesis is that longer-term coarticulation effects can be estimated using phonetic context. For example, knowledge of the context might allow the best candidate sample to be selected from a corpus of real data. This enables subtle variation in natural speech production to be retained in the synthesized visual speech. The main disadvantages are that the corpus from which speech units are selected will not include all possible contexts, a phonetic transcription is required *a priori*, and a large corpus may result in a lengthy search, making a real-time system difficult to implement. The advantage of mapping directly from acoustic speech features is that the speech articulators are physically positioned to form the speech sounds, so the underlying relationship between the acoustic features and the articulatory movements can be learned and exploited. However, only short-term information is exploited — typically frames are considered in isolation, or immediately surrounding frames are concatenated to provide minimal temporal context. It is thus difficult to model the longer-term coarticulation effects that are apparent in natural speech production. Also, learning the mapping from acoustic features to visual features is not trivial.

A. Evaluating Visual Speech Synthesizers

One approach for objectively measuring the performance of a synthesizer is to re-synthesize a set of test sentences for which the original visual speech is available and measure the distance between key points located about the face [6], [17], [44], [46], [52], [53], within the parameters used to model the visual speech [20], [23], [43], [54]–[56], or in the image pixels [12]. Although this approach is intuitive and simple to compute, there are two main limitations. Firstly, the error typically is computed assuming all components of the model are equally significant and it might be appropriate to take into account the natural range of variation [57], and secondly only the magnitude of the difference is usually considered, and differences in direction might also need to be penalized [11].

A more general criticism of measuring performance using the *error* is that it involves computing the distance between individual frames, then averaging over time. Instead one might consider the extent to which the real and synthesized visual speech signals covary [6], [18], [27], [52], [54]–[56], [58]–[60]. However, this is usually measured by computing the correlation coefficient for each parameter (and possibly averaging over parameters), so this perhaps does not provide the full picture since the visual speech signal itself is the variation in the *combination* of these parameters over time.

Other objective measures include the smoothness and the synchronicity [28], which penalize a concatenative synthesizer when selecting either non-consecutive frames from a training sequence, or selecting frames from an incorrect phoneme. Limitations of these measures are that there might be several equally valid paths through the training data that do not require consecutive frames (equally smooth), but these would be scored differently, secondly it is the realisation of the sequence that is important, which does not depend only on the associated phone labels, and thirdly they are applicable only to concatenative synthesizers. ASR has also been used to measure synthesis quality by comparing the differences in phone transcriptions output by a recognizer for both real and synthesized visual speech [61]. Problems with this approach are that the types of error are important, and that phone-based recognition is notoriously difficult for visual-only data. Instead, one might measure the degree of synchrony between a real acoustic signal and an accompanying synthesized visual signal [62]. However, there is a natural degree of asynchrony between the acoustic and visual speech modalities, and this asynchrony must be accounted for in the distance measure. Furthermore, reliable acoustic features are required and previously pitch was used [62]. A limitation is that regions of speech that are unvoiced do not have accompanying pitch, and this affects many significant regions of visual speech, e.g. the lip closure in a bilabial stop.

Care is required in the interpretation of objective measures to ensure that relatively small errors in perceptually significant regions of the visual speech are not overshadowed by larger errors in perceptually less important regions [40]. Furthermore, since the score is measured using a reference signal and humans cannot repeat the same utterance in exactly the same way, it is important to determine if any differences between the ground-truth and synthesized sequences are *perceptually* significant. They might result from natural variation observed in speech production, so are not perceived by viewers.

Ultimately it is viewer perception of quality that is important, so subjective assessment using human viewers is preferable. This could be measured indirectly by estimating the cognitive load on the user by measuring the time taken to perform a task [63], or it could be measured more directly either as the improvement in the intelligibility of noisy acoustic speech provided by the synthesized visual speech, or by scoring viewer opinion of particular aspects of the system.

For intelligibility assessment, the type of stimuli used varies from isolated bisyllabic words [42], [64], multi-syllable nonsense words [65], isolated real words [39], [66], or sentences [2], [39]. Accuracy can be measured using the word, syllable or phone recognition rate, or it might involve keyword spotting in synthesized sentences [55]. The advantage of shorter stimuli for intelligibility testing is that the accuracy of particular speech gestures can be measured, but it is difficult to gauge the accuracy of modelling the longer term aspects of speech articulation. Sentence-level stimuli overcome this somewhat, but care is required to ensure that any gain in intelligibility arises from the speech model and not knowledge of the language. One way to overcome this is to use semantically unpredictable sentences [67]. Alternative intelligibility measures include comparing viewer responses to McGurk stimuli [68] for both real and synthesized sequences [69], or using a modified rhyme test to measure errors in the discrimination of the articulation of words [70]–[72].

A problem with intelligibility as a measure of performance is that over-articulated speech can emphasize the place of articulation and improve the intelligibility, but the resultant sequences will look less natural. Instead, one might use a form of Turing test, where a viewer is asked to classify sequences as either real or synthesized [11], [39], [73], or they might be asked to state their preference between pairs of stimuli [14], [74]. Perhaps the most common method for obtaining a numerical measure of performance using subjective assessment involves asking viewers to score their opinion of a particular aspect of the system on a (typically) five point Likert scale, then reporting the mean opinion score averaged across viewers. Specific aspects that have been tested in this way include the coherency between the audio and visual signals [59], [75], [76], and, more commonly, the naturalness of the talking face [15], [20], [29], [42], [46], [56], [70], [76], [77].

The rest of this paper is organized as follows: Section III describes the data capture and pre-processing, Section IV discusses the model that forms the basis of our visual speech representation, Sections V considers the choice of acoustic feature for predicting visual speech features, Section VI discusses the synthesizers used in this work, and Section VII discusses the evaluation of these synthesizers and investigates the impact of different forms of synthesis error on the perceived quality.

III. Data Capture

The training data used in this work consists of a single speaker reciting the 279 sentences forming the Messiah corpus,

see [78] for a transcript. To ensure that the pose of the head is as constant as possible, the training sentences were collected using a head mounted camera. The resolution of the video frames was 360x288 pixels (one quarter DV-PAL) at 25 Hz. The audio was digitized at 11,025 Hz and 16 bits/sample stereo. The video sequences were recorded in a single sitting to ensure constant lighting throughout the training video. The speaker was instructed to maintain a neutral facial expression (no emotion) to confine, as far as possible, the variation of the facial features to only speech gestures.

IV. Active Appearance Models

The model used throughout this work is an Active Appearance Model (AAM) [36] trained on the face of the speaker in the training video. An AAM is a generative parametric model and is commonly used to track and synthesize faces in video sequences. The model is comprised of two components: a model of shape variation and a model of appearance variation. This makes the use of such models attractive in visual speech synthesis as both the geometry and the texture of the face are captured jointly.

The *shape*, $\mathbf{s}$, of an AAM is defined by the concatenation of the x and y-coordinates of n vertices: $\mathbf{s} = (x_1, y_1, \ldots, x_n, y_n)^T$. A compact model that allows a linear variation in the shape is given by,

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

where the coefficients p_i are the shape parameters. Such a model is usually computed by applying principal component analysis (PCA) to a set of shapes hand-labelled in a corresponding set of images [36]. 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 is defined by the pixels that lie inside the base mesh, $\mathbf{x} = (x, y)^{\mathrm{T}} \in \mathbf{s}_0$. AAMs allow linear appearance variation, so $A(\mathbf{x})$ can be expressed as a base appearance $A_0(\mathbf{x})$ plus a linear combination of l appearance images $A_i(\mathbf{x})$:

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

where the coefficients λ_i are the appearance parameters. The base appearance A_0 and appearance images A_i are usually computed by applying PCA to the shape-normalized training images [36]. A_0 is the mean and the vectors A_i are the (reshaped) eigenvectors corresponding to the largest eigenvalues.

To render a face image from a set of AAM parameters, first the shape parameters, $\mathbf{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$, are used 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 $\mathbf{s}$. In the context of AAMs, a single image maps to a point in AAM space at location $\{\mathbf{p};\ \boldsymbol{\lambda}\}$, thus the task of the synthesizers is to generate the time-varying *trajectory* of model parameters through AAM space such that the resultant *image sequence* contains the correct movement of the facial features given a novel utterance. Example images rendered using an AAM are shown in Figure 2.

Fig. 2. Face images rendered using an AAM.

V. Acoustic Features for Synthesis

We first investigate the degree of correlation between the acoustic features commonly used previously in synthesis and the corresponding AAM features. It follows that the better correlated the features, the easier it will be to accurately predict AAM features for novel speech.

The acoustic speech from the training video is divided into non-overlapping frames of 40 ms duration to match the frame-rate of the video. Empirical evidence [19], [23] suggests that this over smoothes the acoustic features, but improves the correlation with the slower sampled AAM parameters. The acoustic speech in each frame is parameterized as formant frequencies, MFCCs, LPCs and LSFs, and the video is encoded in terms of the AAM parameters. The acoustic training data are thus represented as four $r \times 33,575$ matrices, where $r = \{3, 13, 16, 15\}$ for formant frequencies, MFCCs, LPCs and LSFs respectively and $33,575$ is the number of frames. The visual data are represented as two $s \times 33,575$ matrices, where $s = \{5, 23\}$ for shape and appearance respectively. It is customary to append the first and second derivatives to the audio and visual features to incorporate temporal information, but empirical evidence suggests that this does not significantly improve the mapping between these spaces [19], [79].

Corresponding auditory and visual feature vectors are sampled randomly from the training data 100 times for sample sizes of $N = \{50, 100, 150, \ldots, 2500\}$ and canonical correlation analysis (CCA), as described in [19], is used to measure the extent to which the feature subspaces are correlated. The correlation for each N is then averaged over the 100 trials.

A. Results

The mean correlation for each acoustic feature type and the AAM features is shown in Figure 3. The shape and appearance components of the AAM are reasonably strongly correlated with MFCCs and LSFs ($\rho_1 > 0.8$, where ρ_1 is the correlation captured by the first canonical basis vector), but this quickly falls off in the higher dimensions [19]. LPCs are less well correlated ($\rho_1 \approx 0.75$) and formant frequencies are poorly correlated ($\rho_1 \approx 0.5$). For small sample sizes ($N \leq 50$) the correlation for both the shape and appearance between the MFCCs, LSFs and LPCs is very strong ($\rho_1 > 0.97$), but in this instance a model generated from these features will not generalize well to unseen examples. Increasing the sample size decreases the canonical correlation, but allows the model to

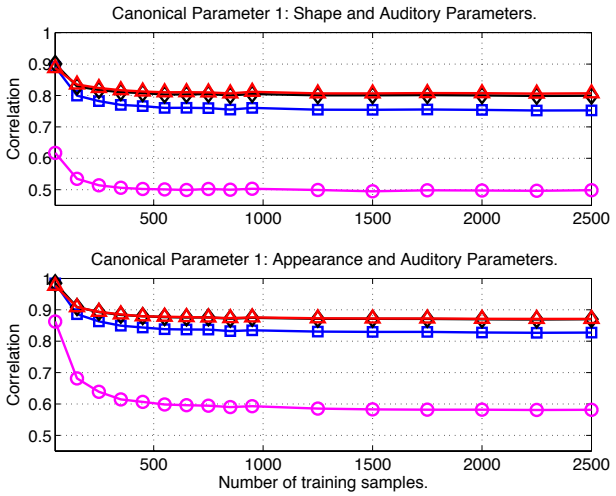

Fig. 3. The effect of the number of training frames on the first canonical correlation coefficient for shape (top) or appearance (bottom) and the acoustic features. The acoustic features are MFCCs (red), LSFs (black), LPCs (blue) and F0 (magenta). Each point is the mean correlation averaged over 100 trails.

better generalize. The estimate of the correlation appears to stabilize after training on approximately 200 (non-consecutive) frames.

For all parameter types, it is the *appearance* information rather than *shape* that is better correlated with the auditory parameters. Generally, the correlation with the appearance is at least 0.1 higher. The shape-free appearance captures detail such as tongue and teeth visibility, whilst the shape model captures information that relates to mouth opening, lip-rounding, etc.

VI. Synthesising Visual Speech

We consider two approaches to synthesising visual speech on an AAM: mapping directly from acoustic features, and a concatenative system that uses a phonetic transcription.

A. Unit-driven synthesis

Our concatenative unit-driven synthesizer [30] selects AAM parameter subsequences from a training corpus by maximising for each phoneme:

$$\kappa_j = \sum_{i=1}^{C} \frac{\mathbf{S}_{l_i j}}{i+1} + \sum_{i=1}^{C} \frac{\mathbf{S}_{r_i j}}{i+1}, \tag{3}$$

where κ_j is the similarity between the desired context and the j^{th} context previously seen for the target phoneme, C is the context width and $\mathbf{S}_{l_{ij}}$ and $\mathbf{S}_{r_{ij}}$ are the similarity between phonemes forming the desired and observed left and right contexts. The similarity between phonemes is computed using:

$$S_{ij} = e^{-\gamma\left(\sum_{k=1}^{m+l}\sum_{n=1}^{5}\left[\left(v_i P^i_{kn} - v_j P^j_{kn}\right)\right]^2\right)}. \tag{4}$$

P^i and P^j are the mean AAM parameter trajectories for phonemes i and j. The first summation is over the dimensions of the AAM and the second over (five) samples equally spaced over the phoneme. v_i is inversely proportional to the variance of the AAM parameters for the i^{th} phoneme, which penalizes poorly represented phonemes. The parameter γ controls the spread (not the order) of the similarities over the range $(0-1)$. 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.

The selected sub-trajectories for the best examples (largest κ) for each phoneme are temporally normalized to the desired duration, concatenated, smoothed using a cubic smoothing spline [80] and applied to the model (Eqs 1 and 2).

B. Feature-driven synthesis

Following [18], a feed-forward artificial neural network (ANN) is used to map from acoustic to visual parameters. The difference here is that we are mapping to AAM features rather than the 37 parameters used to animate Baldi [3].

The acoustic speech from the training corpus is encoded as MFCCs and the corresponding AAM parameters are up-sampled to match the audio frame-rate using cubic (interpolating) splines. Each network has three-layers: an input layer, a 50-node hidden layer, and the output layer. The number of nodes in the hidden layer is selected such that on average the best performance is achieved. As described in [18] at each time step the MFCC feature vectors for 11 consecutive frames (five either side of the current frame) provide the input, and a separate network is used to map from MFCCs to the shape and appearance parameters.

Given a trained network, visual speech is synthesized by first computing the MFCCs from novel acoustic speech and providing these as input to the network to generate a sequence of AAM parameters, which are smoothed using a cubic smoothing spline before being applied to the model. Smoothing splines rather than simple temporal averaging, described in [18], are used to ensure consistency with the unit-driven synthesizer. Thus the two approaches we will compare *differ only in the parameter synthesis*. All other aspects are identical.

VII. Evaluation

We consider both objective and subjective quality measures for our synthesizers, and we are interested in the relationship between the two. If an objective measure exists that relates directly to subjective opinion, this objective measure need only be used in the future. We test the objective measures used most commonly within the community, and extend the measures we reported in [81]. Somewhat surprisingly, the relationship between objective and subjective measures has largely been ignored, and we demonstrate that objective quality is not necessarily a good indicator of subjective opinion.

A. Subjective evaluation

The aim of this experiment was to compare the perceived naturalness of a feature-driven and unit-driven synthesizer. Naturalness is one of the most widely reported subjective measures [15], [20], [29], [42], [46], [56], [70], [76], [77], and it has the advantage that viewers can take all aspects of the synthesizer into account and form an opinion of the overall

sense of the synthesis quality. A limitation is that it does not inform about specific aspects of the system (e.g. synchrony), but these can be later investigated with further subjective tests.

1) Methodology: Videos for 15 sentences not included in training were generated using both synthesizers, and all were synchronized to real acoustic speech. In addition, the original sequences were re-rendered using the AAM to provide a benchmark against which to compare the synthesizer scores. The audiovisual sequences were presented in a randomized order to 18 participants, who were asked to use a slider to score $(0-50)$ the naturalness of each sequence. Participants were told that they would see only the face in the video, as illustrated in Figure 2, and that in some sequences the face *may* have undergone some form of processing. They were told to ignore image quality and focus their attention only on the *naturalness of the talking face*. Participants were free to repeat the sequences as many times as required.

2) Results: The naturalness scores for a sequence are averaged over participants and the resulting scores are subject to a Kruskal-Wallis test [82] to determine if the difference between conditions is statistically significant. The responses are summarized in Table I. We see that AAM re-rendered video is perceived as more natural than both synthesis methods ($p < 0.005$) and the unit-driven synthesizer is perceived as more natural than the feature-driven synthesizer ($p < 0.015$). It is difficult to draw broad conclusions about feature-driven versus unit-driven synthesis given that we have considered only one of each synthesizer type, but a similar trend has been observed elsewhere for different visual parameters and different synthesis methods [54]. This, we believe, is because unit-driven synthesizers are better able to incorporate longer-term influences of the surrounding gestures, whereas feature-driven approaches tend to have only minimal context available.

TABLE I
MEDIAN AND MEDIAN ABSOLUTE DEVIATION FROM THE MEDIAN NATURALNESS SCORES FOR RE-RENDERED VIDEO, A FEATURE-DRIVEN SYNTHESIZER AND A UNIT-DRIVEN SYNTHESIZER. H IS THE KRUSKAL-WALLIS TEST STATISTIC, AND p THE SIGNIFICANCE LEVEL.

Treatment	n	Median	MAD
Video-Driven	15	38.71	1.12
Unit-Driven	15	27.47	3.99
Feature-Driven	15	22.35	2.25
H = 31.97		$p < 0.005$	

The overall perception of the naturalness of the synthesizers is disappointingly low. We note that the synthesized sequences are synchronized to real acoustic speech, and it has been noted that viewers will therefore expect higher quality synthesized visual speech to match the quality of the acoustic speech [59]. Inspection of the video sequences generated by the synthesizers suggests two possible causes for the low scores. Firstly, the parameter trajectories generated by the synthesizers require smoothing, so the resulting synthesized visual speech appears somewhat under-articulated. Secondly, the visual gestures are, on the whole, well re-produced, but occasionally there are isolated gestures that appear to stand-out as being obviously incorrect. A typical example is shown in Figure 4, where the syllable over frames 60–68 is very under-articulated. We next describe experiments carried out to determine the significance of these forms of error.

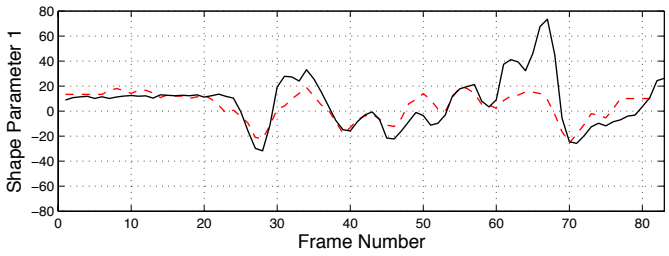

Fig. 4. The first shape parameter for a sentence: (A) as measured in the video (black, solid line), and (B) synthesized by the feature-driven synthesizer (red, dashed line). Overall the synthesized trajectory is well reproduced, except the visual gesture between frames 60–68 (mouth closure).

B. Effect of smoothing on perceived naturalness

Parameter trajectories generated using both the unit-driven and feature-driven synthesizers require smoothing. The unit-based system has no smoothness constraints in the unit selection, and the feature-driven system has no knowledge of past/future (visual) frames. Consequently the parameter trajectories are noisy, which results in *jitter* in the facial features in the synthesized video sequences. To help overcome this the parameters are smoothed before being applied to the model. Several methods have been proposed for smoothing synthesized sequences, which include using a cost term based on the distance to a prediction of the next frame [23], using local blending functions between key shapes [43], using some form of low-pass filter based on the geometric mean [6], triangular averaging windows [41] or the spectral energy in synthesized sequences [74], or finally smoothing using cubic spline filters [46]. The approach adopted here is to smooth by fitting a cubic smoothing spline to the parameters [80] that minimizes 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, ζ, trades-off a natural cubic spline interpolation of the data, or no smoothing ($\zeta = 1$) and the least squares fit, or maximally smoothed trajectory ($\zeta = 0$). Informal comparisons with low-pass filtering using a Gaussian suggest the smoothing spline is preferred by viewers [78].

1) Methodology: The unit-driven synthesizer requires greater smoothing than the feature-driven approach, and typical values for the smoothing parameter are $\zeta = 0.5$ and $\zeta = 0.9$ for the unit-driven and feature-driven synthesizers respectively. To test the significance of the affect of smoothing on synthesis quality, videos for 30 sentences were generated in three conditions for $\zeta = 1.0$ (no smoothing), $\zeta = 0.9$ and $\zeta = 0.5$. Thus, the sequences were derived from real parameters (as measured in video), the parameters had just undergone different degrees of smoothing for the different conditions. The same 18 participants from the previous experiment (Section VII-A) were given the same instructions regarding rating the naturalness of the talking face.

2) Results: The naturalness scores for a sequence are averaged over participants and the resulting scores are subject to a Kruskal-Wallis test to determine if the difference between conditions is statistically significant. The responses are summarized in Table II. There is no significant effect on naturalness when smoothing using $\zeta = 0.9$ ($p > 0.86$), although smoothing using $\zeta = 0.5$ does have a significant impact on naturalness ($p < 0.005$). Thus, we conclude that even if the unit-driven synthesizer was to generate *exactly* the required visual parameters, the perceived realism of the sequences would be severely impacted after smoothing. Note that on average the scores for the sequences generated by the unit-driven synthesizer in Section VII-A are comparable to the smoothed ($\zeta = 0.5$) sequences presented here, yet the sequences here were derived from real data measured directly from video. This suggests that future effort for improving the quality of the synthesizer should focus on the unit selection (Eq. 3) — generating smoother parameter trajectories from the outset will require less post-processing, which in turn will be less likely to impact on realism.

TABLE II
Naturalness results for re-rendered video before and after smoothing.

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
H = 0.04		$p < 0.005$	

C. Effect of errors in isolated visual gestures

To determine the impact on naturalness of the type of error highlighted in Figure 4, the effect of other potential sources of error must be removed. For example, although the overall shape of the trajectories in Figure 4 are broadly similar, some gestures are slightly under-articulated whilst others are slightly over-articulated (e.g. at frames 34 and 50 respectively). These subtle differences must be removed so that the only error is an isolated erroneous visual speech gesture.

1) Methodology: Videos for 10 sentences were generated in two conditions: Firstly by re-rendering using the AAM parameters measured from the original video, and secondly using the same parameters after substituting the parameters for one syllable with those of another chosen randomly from elsewhere in the corpus. Note that prior to rendering, the parameters of the selected syllable were normalized to the duration of the original syllable and smoothed appropriately at the boundary to ensure a seamless blend with the background video. The impact of this will of course depend on the class of phonemes forming the two syllables, but the aim of this experiment is not to determine the specific impact of substituting different classes of phonemes for one another, rather it is designed to obtain a broad estimate of the typical synthesis errors highlighted in Figure 4, where the mouth gesture is obviously wrong. That is, if only a single speech gesture appears incorrect, how much impact can this have on the perception of realism of an entire utterance?

2) Results: The naturalness scores for a given sequence are averaged over all participants and a Kruskal-Wallis test is used to determine if the differences between treatments is significant. The responses are summarized in Table III. Introducing only a single erroneous visual speech gesture *does* significantly degrade the perceived naturalness of the *entire* 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 feature-driven synthesizer ($p > 0.34$), and the sequences generated by the unit-driven synthesizer are perceived as significantly more natural than processed sequences presented here. Again, we note that the type of error is important and some are perceptually more significant than others. But what this result shows is that an entire sequence can be judged as poor when only a small section is bad even if the rest is otherwise perfect.

TABLE III
NATURALNESS RESULTS FOR: (T1) AAM RE-RENDERED VIDEO, AND (T2) THE **SAME** SEQUENCES WITH AN INCORRECTLY RENDERED SYLLABLE.

Treatment	n	Median	MAD
T1	10	42.6	1.48
T2	10	23.5	4.15
H = 41.32		$p < 0.0002$	

D. Objective evaluation

To maximize the use of the limited available training data in the objective evaluation, leave-one-out cross validation was used. A separate synthesizer was trained for each of the 279 sentences, where each sentence under test was not included in the training data for the respective synthesizer. Five objective measures of performance were computed between the synthesized and ground-truth parameters. Two of these measures were selected because of their common use elsewhere. Namely: 1) **Correlation** (ρ) computed between individual parameters, then averaged over parameters [6], [18], [27], [52], [54]–[56], [58]–[60], which gives a single score per sentence rather than multiple scores, as presented in [81]. 2) **Normalized RMS error** (ϵ) at each frame averaged over the utterance [20], [23], [43], [54]–[56]. Following [54] the error is normalized to be a percentage of the total variation in the parameters. This allows a more meaningful comparison between different systems as the absolute value of the error is dependent on the unit and the scale of the features.

Additional objective measures were investigated in light of specific issues with the two more common measures above. In particular: 3) **Normalized Peak RMS Error** (ϵ_p) computed by selecting the largest magnitude frame-wise error over the sequence (rather than averaging over frames). Section VII-C showed that isolated errors can significantly affect the perceived naturalness, so the peak error is included in these tests to determine how well this score relates to perceived naturalness. 4) **Dynamic time warp cost** ($\mathcal{D}$) measured as the cost of warping the synthesized parameters onto the ground-truth parameters, where the cost is normalized for path length. This score was used as it incorporates the notion of the temporal relationship between the two signals and not just the frame-wise error. Also, this measure should be less sensitive to isolated errors than is, say, the correlation. 5) **Phone-based**

Mahalonabis distance computed using:

$$d_k = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{p}_i - \mathbf{v}_i)\mathbf{S}_j^{-1}(\mathbf{p}_i - \mathbf{p}_i)' \quad (6)$$

where d_k is the score for the k^{th} utterance of length n video frames. The parameters generated by the synthesizer are denoted as $\mathbf{p}$ and those measured from the video as $\mathbf{v}$. The matrix $\mathbf{S}_j$ is the scatter matrix for the phoneme to which the i^{th} video frame belongs. The arguments for this form of error are that the higher variance (back of mouth) sounds are down-weighted and considered less important than sounds with lower variance visual parameters, and the score is invariant to scale.

The scores for our systems in terms of both the shape and the appearance parameters are presented in Table IV.

TABLE IV
Mean ($\pm$ standard deviation) objective measures computed between original (held-out) AAM parameters and the corresponding synthesized parameters.

Measure	Shape		Appearance	
	Feat.	Unit	Feat	Unit
ρ	0.74 ± 0.07	0.70 ± 0.07	0.77 ± 0.05	0.72 ± 0.06
ϵ	9.70 ± 1.46	9.66 ± 1.30	24.9 ± 6.72	11.6 ± 8.20
ϵ_p	22.4 ± 2.88	23.0 ± 3.59	39.3 ± 6.45	26.7 ± 3.85
$\mathcal{D}$	0.32 ± 0.18	0.33 ± 0.15	0.42 ± 0.08	0.04 ± 0.04
d	1.66 ± 0.09	1.48 ± 0.09	5.84 ± 1.60	1.47 ± 0.09

To place these objective scores into context, typical correlation values reported elsewhere vary from reasonable correlation ($\rho \leq 0.70$) [18], [23], [54], [56], [58], to higher correlation ($\rho > 0.70$) [27], [52], [59]. It is worth noting that different systems use different parameters and it has been shown that a different representation of the same visual speech sequences can result in a different (correlation) score [59]. Thus without direct access to the underlying systems, any comparison is only indirect and should be regarded with some degree of caution. Interestingly, the performance of both the feature-driven and the unit-driven synthesis methods here are better than those reported in [54] in terms of the correlation, but are at about the same level in terms of the normalized error. This raises an interesting question about the relationship between objective and subjective scores: is an objective measure a good indicator of naturalness? This is an issue that surprisingly has been ignored. Since naturalness is the overall sense of realism in terms of viewer perception, then we argue that an objective score that more closely relates to the naturalness is a more meaningful measure. It is difficult to comment on the remaining objective scores (ϵ_p, $\mathcal{D}$ and d) relative to other systems as these have not been used in the context of synthesis evaluation — they are included here to determine if they are likely better indicators of subjective performance than the two more typical measures.

E. Comparing Objective and Subjective Scores

To quantify the relationship between the subjective and objective scores for the unit-driven synthesizer, the absolute value of the correlation coefficient between the respective scores is computed. These are shown in Table V.

TABLE V
The absolute value of the correlation between the subjective ratings of naturalness for the unit-driven synthesizer and the respective objective measures of performance.

Feature	Measure				
	ρ	ϵ	ϵ_p	$\mathcal{D}$	d
Shape	0.55	0.55	0.67	0.76	0.36
Appearance	0.20	0.63	0.60	0.72	0.67

The objective scores for both the shape and the appearance are, on average, equally well correlated with perceived naturalness. However, the correlation is different for specific parameters. Significantly, we find that the two most commonly used objective measures of performance (ρ and ϵ) are not necessarily the most informative in terms of expected subjective quality. In fact, the objective score that correlates best with the subjective scores is the DTW cost. This is likely because it incorporates the notion of both a 'distance' (as does ϵ) *and* the temporal relationship between the parameters (as does ρ). Also, what is particularly interesting is that the peak error correlates better with subjective quality than the average error. This supports the finding in Section VII-C, where it was shown that a single synthesis error can significantly degrade the perception of the naturalness of an otherwise correct utterance: the larger the peak error, the lower the perceived quality.

VIII. Summary and Conclusions

In this paper we first considered the choice of feature type for input to a feature-driven visual speech synthesizer. We used CCA to compute the degree of correlation between the (visual) AAM space and several acoustic subspaces and found that, on average, MFCCs generally covary most with the AAM parameters. MFCCs were then used as input features to a feed forward ANN-based visual speech synthesizer and the performance was compared with that of a unit-driven synthesizer using both subjective and objective tests. In terms of objective measures both systems appear to perform equally well. However, using formal subjective testing we found that the unit-driven approach was perceived as significantly more natural than the feature-driven approach. A similar finding was outlined in [54] although their subjective performance was measured in terms of intelligibility. It is difficult to compare more formally the results reported by others without direct access to the underlying systems themselves, but the objective scores that we report here are similar to those that others report in terms of correlation and (normalized) RMS error.

We conducted a series of experiments designed to quantify the likely reasons for the low naturalness ratings obtained by both synthesizers. In particular we investigated the errors introduced by both smoothing and by parameter trajectory formation. The results of these experiments show 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 impacted when only a single syllable is erroneous, even if the rest of the sequence is perfect (as produced by the speaker). 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. The accuracy of short-term properties of the production of speech gestures can also be measured to give a more localized measure of performance.

The objective measures used in this work included two that are commonly used within the community (correlation and RMS error) and three that (to the best of our knowledge) have not been used elsewhere for this task (peak error, DTW cost, phone-based Mahalonabis distance). A surprisingly overlooked question has been *how reliably do objective measures of performance relate to subjective quality*? By measuring the correlation between the various objective scores and the subjective scores we have found that the DTW cost correlates better with subjective opinion of naturalness than do the two most common objective measures, and so this potentially is a better indicator of performance. Comparing the ordering of the sequences in terms of the two common objective measures suggests that one objective measure is not necessarily a good indicator of others — so the *measured* quality of the visual speech in an objective sense is dependent on the score that is chosen. An open issue is how the objective measures might be combined to give a yet more reliable indication of likely perceived quality. With a larger sample size of naturalness scores, this could be done by re-weighting scores for segments of utterances where they are more reliable. This is an issue currently under investigation.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Nick Wilkinson for his assistance with the neural network based synthesis system.

REFERENCES

[1] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," in *International Journal of Speech Technology*, vol. 6, 2003, pp. 331–346.

[2] J. Beskow, I. Karlsson, J. Kewley, and G. Salvi, "SYNFACE - a talking head telephone for the hearing-impaired," in *Computers Helping People with Special Needs*, 2004, pp. 1178–1186.

[3] D. Massaro, *Perceiving Talking Faces*. The MIT Press, 1998.

[4] F. Parke, "Parametric models for facial animation," *Computer Graphics and Applications*, vol. 2, no. 9, pp. 61–68, 1982.

[5] G. Bailly, F. Elisei, P. Badin, and C. Savariaux, "Degrees of freedom of facial movements in face-to-face conversational speech," in *International Workshop on Multimodal Corpora*, 2006, pp. 33–36.

[6] R. Gutierrez-Osuna, P. Kakumanu, A. Esposito, O. Garcia, B. A., and I. Rudomin, "Speech-driven facial animation with realistic dynamics," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 33–42, 2005.

[7] B. Theobald, I. Matthews, M. Mangini, J. Spies, T. Brick, J. Cohn, and S. Boker, "Mapping and manipulating facial expression," *Language and Speech*, vol. 52, no. 2/3, pp. 369–386, 2009.

[8] K. Kähler, J. Haber, and H. Seidel, "Geometry-based muscle modelling for facial animation," in *Graphics Interface*, 2001, pp. 27–36.

[9] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic modeling for facial animation," in *Proceedings of SIGGRAPH*, 1995, pp. 55–62.

[10] L. Nedel and D. Thalmann, "Real time muscle deformations using mass-spring systems," in *Computer Grahpics International*, 1998, pp. 156–165.

[11] M. Brand, "Voice puppetry," in *Proceedings of SIGGRAPH*, Los Angeles, California, 1999, pp. 21–28.

[12] Y. Du and X. Lin, "Realistic mouth synthesis based on shape appearance dependence mapping," *Pattern Recognition Letters*, vol. 23, no. 14, pp. 1875–1885, 2002.

[13] P. Eisert, S. Chaudhuri, and B. Girod, "Speech driven synthesis of talking head sequences," in *Proceedings of the Workshop 3D Image Analysis and Synthesis*, 1997, pp. 51–56.

[14] G. Englebienne, T. Cootes, and M. Rattray, "A probabilistic model for generating realistic speech movements from speech," in *Proceedings of Advances in Neural Information Processing Systems*, 2007.

[15] G. Feldhoffer, A. Tihanyi, and O. Balázs, "A comparative study of direct and asr-based modular audio to visual speech systems," *The Phonetician*, vol. 97/98, pp. 15–24, 2008.

[16] P. Hong, Z. Wen, and T. Huang, "Real-time speech-driven expressive synthetic talking faces using neural networks," *IEEE Transaction on Neural Networks*, vol. 13, no. 4, pp. 916–927, 2002.

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

[18] D. Massaro, J. Beskow, M. Cohen, and T. Fry, C.and Rodriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks," in *Proceedings of the International Conference on Auditory Visual Speech Processing*, 1999.

[19] B. Theobald and N. Wilkinson, "Real-time visual speech synthesis using active appearance models," in *Proceedings of the International Conference on Auditory Visual Speech Processing*, 2007.

[20] L. Wang, W. Han, X. Qian, and F. Soong, "Synthesizing photo-real talking head via trajectory-guided sample selection," in *Proceedings of Interspeech*, 2010.

[21] Z. Wen, P. Hong, and T. Huang, "Real time speech driven facial animation using formant analysis," in *Proceedings of the International Conference on Multimedia and Expo*, 2001, pp. 817–820.

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

[23] S. Deena, S. Hou, and A. Galata, "Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model," in *International Conference on Multimodal Interfaces*, 2010, pp. 1–8.

[24] J. Edge and A. Hilton, "Visual speech synthesis from 3D video," in *IET European Conference on Visual Media Production*, 2006, pp. 174–179.

[25] T. Ezzat and T. Poggio, "Miketalk: A talking facial display based on morphing visemes," in *Proceedings of the Computer Animation Conference*, 1998, pp. 96–103.

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

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

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

[29] W. Mattheyses, L. Latacz, and W. Verhelst, "Active appearance models for photorealistic visual speech synthesis," in *Proceedings of Interspeech*, 2010.

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

[31] F. Parke and K. Waters, *Computer Facial Animation*. A K Peters, 1996.

[32] E. Cosatto and H. Graf, "Photo-realistic talking-heads from image samples," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 152–163, 2000.

[33] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," in *Proceedings of the International Conference on Auditory Visual Speech Processing*, 2001, pp. 90–97.

[34] T. Hawkins, A. Wenger, C. Tchou, A. Gardner, F. . Goransson, and P. Debevec, "Animatable facial reflectance fields," in *Eurographics Symposium on Rendering*, June 2004.

[35] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," in *Proceedings of Eurographics*, 2003, pp. 641–650.

[36] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[37] G. Bailly, N. Campbell, and M. Mbius, "ISCA special session: Hot topics in speech synthesis," in *Proceedings of Eurospeech*, 2003, pp. 37–40.

[38] B. Theobald, S. Fagel, F. Elsei, and G. Bailly, "LIPS2008: Visual speech synthesis challenge," in *Proceedings of Interspeech*, 2008, pp. 1875–1878.

[39] G. Geiger, T. Ezzat, and T. Poggio, "Perceptual evaluation of video-realistic speech," MIT, Cambrige, MA, Tech. Rep. CBCL Paper 224/AI Memo 2003-003, 2003.

[40] H. Yehia, R. P., and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behaviour," *Speech Communication*, vol. 26, pp. 23–43, 1998.

[41] P. Hong, Z. Wen, T. Huang, and H. Shum, "Real-time speech-driven 3D face animation," in *Proceedings of the 3D Data Processing Visualization and Transmission Symposium*, 2002, pp. 713–716.

[42] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip movement synthesis from speech based on hidden Markov models," in *Proceedings of International Conference on Face and Gesture*, 1998, pp. 154–159.

[43] Z. Deng, U. Neumann, J. Lewis, T. Kim, and S. Narayanan, "Expressive facial animation synthesis by learning speech coarticulation and expression spaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1523–1534, 2006.

[44] O. Engwall, "Evaluation of a system for concatenative articulatory visual speech synthesis," in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 665–668.

[45] K. Liu and J. Ostermann, "Realistic facial animation system for interactive services," in *Proceedings of Interspeech*, 2008, pp. 2330–2333.

[46] J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise, "Accurate visible speech synthesis based on concatenating variable length motion capture data," *IEEE Transactions on Visualization and Compuer Graphics*, vol. 12, no. 2, pp. 266–276, 2006.

[47] E. Bevacqua and C. Pelachaud, "Expressive audio-visual speech," *Computer Animation and Virtual Worlds*, vol. 15, pp. 297–304, 2004.

[48] S. King and R. Parent, "Creating speech-synchronized animation," *IEEE Transactions on Visualization and Compuer Graphics*, vol. 11, no. 3, pp. 341–352, 2005.

[49] J. Melenchon, E. Martinez, F. De la Torre, and J. Montero, "Emphatic visual speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 3, pp. 459–468, 2009.

[50] B. Theobald and N. Wilkinson, "A probabilistic trajectory synthesis system for synthesising visual speech," in *Proceedings of Interspeech*, 2008, pp. 2310–2313.

[51] J. Tao, L. Xin, and Y. Panrong, "Realistic visual speech synthesis based on hybrid concatenation method," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 3, pp. 469–477, 2009.

[52] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. Kakumanu, and O. Garcia, "Audio/visual mapping with cross-modal hidden markov models," *IEEE Transactions on Multimedia*, vol. 7, no. 2, pp. 243–252, 2005.

[53] S. Nakamura and E. Yamamoto, "Speech-to-lip movement synthesis by maximizing audio-visual joint probability based on the EM algorithm," *Journal of VLSI Signal Processing*, vol. 27, pp. 119–126, 2001.

[54] J. Beskow, "Trainable articulatory control models for visual speech synthesis," *Journal of Speech Technology*, vol. 4, no. 7, pp. 335–349, 2004.

[55] R. Carlson and B. Granström, "Data-driven multimodal synthesis," *Speech Communication*, vol. 47, pp. 182–193, 2005.

[56] X. Zhuang, L. Wang, F. Soong, and M. Hasegawa-Johnson, "A minimum converted trajectory error (mcte) approach to high quality speech-to-lips conversion," in *Proceedings of Interspeech*, 2010.

[57] D. Cosker, D. Marshall, P. Rosin, and Y. Hicks, "Speech driven facial animation using a hidden markov coarticulation model," in *Proceedings of the International Conference on Pattern Recognition*, 2004, pp. 128–131.

[58] L. Arslan and D. Talkin, "3D face point trajectory synthesis using an automatically derived visual phoneme similarity matrix," in *Proceedings of the International Conference on Auditory-Visual Speech Processing*, 1998, pp. 175–180.

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

[60] T. Kuratate, K. Munhall, P. Rubin, E. Vatikiotis-Bateson, and H. Yehia, "Audio-visual synthesis of talking faces from speech production correlates," in *Proceedings of Eurospeech*, vol. 3, 1999, pp. 1279–1282.

[61] J. Dongmei, X. Lei, Z. Rongchun, W. Verhelst, I. Ravyse, and H. Sahli, "Acoustic viseme modelling for speech driven animation: A case study," in *Workshop on Model-based Processing and Coding of Audio*, 2002, pp. 49–52.

[62] S. Cadavid, M. Abdel-Mottaleb, D. Messinger, M. Mahoor, and L. Bahrick, "Detecting local audio-visual synchrony in monologues utilizing vocal pitch and facial landmark trajectories," in *Proceedings of the British Machine Vision Conference*, 2009.

[63] I. Pandzic, J. Ostermann, and D. Millen, "User evaluation: Synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, pp. 330–340, 1999.

[64] S. Fagel, G. Bailly, and F. Elisei, "Intelligibility of natural and 3D-cloned german speech," in *Proceedings of the International Conference on Auditory-Visual Speech Processing*, 2007.

[65] C. Benoît and B. Le Goff, "Audio-visual speech synthesis from french text: Eight years of models, designs and evaluation at the ICP," *Speech Communication*, vol. 26, pp. 117–129, 1998.

[66] M. Železný, K. Zdeněk, P. Císař, and M. Jindřich, "Design, implementation and evaluation of the czech realistic audio-visual speech synthesis," *Signal Processing*, vol. 86, pp. 3657–3673, 2006.

[67] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech Communication*, vol. 18, no. 4, pp. 381–392, 1996.

[68] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[69] D. Cosker, D. Marshall, P. Rosin, S. Paddock, and S. Rushton, "Towards perceptually realistic talking heads: Models, metrics and mcgurk," *ACM Transactions on Applied Perception*, vol. 2, no. 3, pp. 270–285, 2005.

[70] P. Dey, S. Maddock, and R. Nicolson, "Evaluation of a viseme-driven talking head," in *Proceedings of The Eighth Theory and Practice of Computer Graphics 2010 Conference*, 2010, pp. 139–142.

[71] S. Fagel and C. Clemens, "An articulation model for audiovisual speech synthesis — Determination, adjustment, evaluation," *Speech Communication*, vol. 44, pp. 141–154, 2004.

[72] S. Fagel, "MASSY speaks english: Adaptation and evaluation of a talking head," in *Proceedings of Interspeech*, 2008.

[73] K. Liu and J. Ostermann, "Optimization of an image-based talking head system," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, 2009.

[74] W. Mattheyses, L. Latacz, and W. Verhelst, "Optimized photorealistic audiovisual speech synthesis using active appearance modeling," in *Proceedings of the International Conference on Auditory-Visual Speech Processing*, 2010, pp. 148–153.

[75] K. Choi and J. Hwang, "Automatic creation of a talking head from a video sequence," *IEEE Transactions on Multimedia*, vol. 7, no. 4, pp. 628–637, 2005.

[76] W. Mattheyses, L. Latacz, and W. Verhelst, "On the importance of audiovisual coherence for the perceived quality of synthesized visual speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, 2009.

[77] G. Takacs, "Direct, modular and hybrid audio to visual speech conversion methods - a comparative study," in *Proceedings of Interspeech*, 2009.

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

[79] H. Bredin and G. Chollet, "Audio-visual speech synchrony measure for talking-face identity verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, april 2007, pp. 233–236.

[80] C. de Boor, "Calculation of the smoothing spline with weighted roughness measure," *Mathematical Models and Methods in Applied Sciences*, vol. 11, no. 1, pp. 33–41, 2001.

[81] B. Theobald, N. Wilkinson, and I. Matthews, "On evaluating synthesised visual speech," in *Proceedings of the International Conference on Auditory Visual Speech Processing*, 2008, pp. 7–12.

[82] D. Wakerly, W. Mendenhall, and R. Scheaffer, *Mathematical Statistics with Applications*. Duxbury Advanced Series, 2002.

Barry-John Theobald received the BEng degree in Electronic Engineering in 1999 and the PhD degree in Computer Science in 2003, both from the School of Computing Sciences, University of East Anglia, UK. He is a member of faculty in the Graphics, Vision and Speech Laboratory in the School of Computing Sciences, University of East Anglia.

Iain Matthews received the BEng degree in Electronic Engineering in 1995 and the PhD degree in Computer Science in 1999, both from the School of Computing Sciences, University of East Anglia, UK. He has worked as a senior systems scientist at the Robotics Institute, Carnegie Mellon University and as a consultant for Weta Digital, New Zealand. He is currently a Senior Research Scientist at Disney Research, Pittsburgh.