Eurographics/ ACM SIGGRAPH Symposium on Computer Animation (2012)
P. Kry and J. Lee (Editors)

Dynamic Units of Visual Speech

Sarah L. Taylor [1], Moshe Mahler [2], Barry-John Theobald [1] and Iain Matthews [2]

[1] University of East Anglia, Norwich, England
[2] Disney Research, Pittsburgh, USA

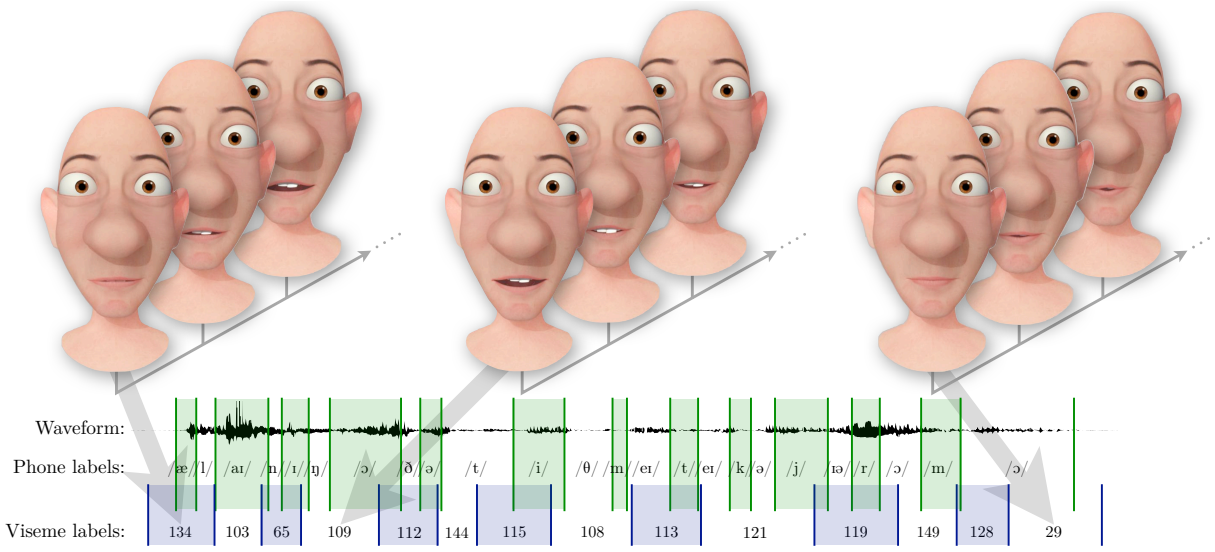

Figure 1: *We learn new* dynamic *units for visual speech (dynamic visemes) by considering the video portion of speech training data. Dynamic visemes naturally capture visual coarticulation and the inherent asynchrony between visual and acoustic speech. The training data is also used to learn a mapping from phonemes to dynamic visemes to create any text-to-speech animation. Our approach can be applied to any style of character animation.*

Abstract

We present a new method for generating a dynamic, concatenative, unit of visual speech that can generate realistic visual speech animation. We redefine visemes *as temporal units that describe distinctive speech movements of the visual speech articulators. Traditionally visemes have been surmized as the set of static mouth shapes representing clusters of contrastive phonemes (e.g. /p, b, m/, and /f, v/). In this work, the motion of the visual speech articulators are used to generate discrete, dynamic visual speech* gestures. *These gestures are clustered, providing a finite set of movements that describe visual speech, the visemes. Dynamic visemes are applied to speech animation by simply concatenating viseme units. We compare to static visemes using subjective evaluation. We find that dynamic visemes are able to produce more accurate and visually pleasing speech animation given phonetically annotated audio, reducing the amount of time that an animator needs to spend manually refining the animation.*

Categories and Subject Descriptors (according to ACM CCS): I.2.7 [Artificial Intelligence]: Natural Language Processing—Speech recognition and synthesis I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Time-varying imagery I.5.4 [Image Processing and Computer Vision]: Applications—Computer vision I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Computer vision

1. Introduction

Realistic facial animation requires care and painstaking manual effort. This is especially true during speech as viewers are extremely sensitive to any discrepancy between the sounds and the accompanying facial movements [SP54]. As the visual quality of computer graphics facial models improves, one might also expect the quality of animated facial behavior to follow. However, this generally has not been the case and practical applications of speech animation in games and movies is achieved using hand-crafted animation or using expensive and time consuming motion capture.

There are a number of factors that compound the difficulty of synthesizing realistic facial movements during speech. Firstly, the biomechanics of the face are complex and it is not clear how these should best be modeled for speech or other facial expressions, especially for character animation. Secondly, it is not clear how the underlying visual speech signal should be represented at a segmental level for synthesis. Previously this was done using visemes (visual phonemes [Fis68]), but whilst defining a visual unit based on speech acoustics may be convenient, this simple approach has a number of problems. Firstly, the number of phonemes and visemes in an utterance transcription are considered to be the same — phoneme labels are simply substituted for viseme labels. Coarticulation effects, where neighboring sounds influence one another, must be modeled as part of a post process [Mas98], but there is no well defined model of coarticulation in the phoneme-to-viseme mapping. Secondly, the boundaries between the units in the acoustic and visual modalities must align and in general this is not true as coarticulation on the heavy visible articulators may differ from that of the inner articulators. In short, context matters and it may differ acoustically and visually. In a standard phoneme-to-viseme mapping there is no accounting for this natural asynchrony of audiovisual speech. Thirdly, and more importantly, a phoneme is by *definition* a group of related **sounds** that are perceived to have the same function. Phonemes serve to represent meaningful contrasts between utterances. Different realizations of the same phoneme can, and often do, appear very different visually, for example see Figure 2. Clearly these different lip poses should not all belong to the same visual class.

In this paper we define a new dynamic unit for visual speech. This unit represents contrastive movements of the speech articulators that are derived by analyzing real visual speech. Dynamic visemes better represent the visual speech signal in that each viseme serves a particular function, and so substituting one dynamic viseme for another changes the visual appearance of the utterance, a true visual analog to a phoneme. The dynamic nature of our unit means that coarticulation effects are explicit in our model and the boundaries between visemes are not tied to the boundaries of the underlying phones. Indeed, as our model represents the **movements** of the visible articulators, we find that a single viseme often extends over several acoustic phones. We learn dynamic units by clustering visual gestures in a large corpus of real speech video data. The cluster of visual gestures assigned to a viseme in our model represents the visual equivalent of the allophones of a phoneme. We demonstrate this by representing a visual speech utterance simply by concatenating the median sample from the corresponding dynamic visemes forming the utterance and animating a 3D face model. We apply dynamic visemes to animation more generally by mapping any given phoneme string to a dynamic viseme sequence and show that the speech is more visually pleasing than a common interpolated static mouth shape approach.

Figure 2: *Example video frames showing differences in lip-shape at the onset of the sound /t/. In the sense of traditional visemes, each of these poses represents the same unit.*

2. Related Work

2.1. Modeling Visemes

Traditionally visemes have been defined as groups of phonemes that are expected to appear visually the same on the lips. There are two broad approaches for obtaining this grouping: using either subjective assessment with human viewers [MJ83, OB86] or using a data-driven approach [HSLG04]. In the case of subjective assessment, phonemes are grouped if the within-group responses in a stimulus-response confusion matrix account for a significant proportion (usually 75%) of all viewer responses. However, limitations of subjective assessment necessitate that the stimuli be simple, so phonemes are presented in the context of isolated mono- or bi-syllabic words. This does not reflect the longer-term coarticulation effects found in natural speech production. Also, phonemes need only be confused 75% of the time, which suggests that up to 25% of the realizations of the phonemes within a viseme group are visually distinct.

Data-driven approaches for identifying visemes typically use some form of unsupervised clustering, where phonemes that are clustered together frequently are said to form a viseme. Phoneme-to-viseme mappings obtained using ob-

jective methods tend to be less reliable than those defined using subjective methods both for computer facial animation of speech [MLV11] and for visual speech recognition [CH11].

A simple many-to-one mapping is not sufficient to model the complex relationship between the visual gestures and the underlying sounds. As a result there is no definitive agreement regarding either the number of visemic classes or how the set of phonemes map to visemes. Thus, the definition of a traditional viseme is only informal and as a unit of speech for computer facial animation it is poorly defined.

2.2. Animating Speech

The goal of speech animation is to present the correct articulatory dynamics on a face model. One approach is to transfer real motion data from a talker to a model [CFP04, Wil90, XCXH03], which has the advantage of capturing the liveliness and subtleties of facial gestures produced by the performer. Bilinear or multilinear models can separate identity, speech and expression such that the characteristics of the transferred speech can be manipulated so the animation can be presented in different emotional contexts or on different faces [CB05, NN01, VBPP05, WSZP07]. Whilst these performance-driven approaches are effective for generating realistic animation, motion transfer lacks the flexibility of true animation in that an actor is always required.

Discretizing speech into a string of phonemic targets, which are mapped to viseme targets using a lookup table is another approach. An interpolation function is required to generate animated sequences by computing the in-between frames. The interpolation function may be based either on static targets [EP98] or a more complex function that attempts to model visual coarticulation. In [CM94] the interpolation function is hand-crafted based on exponentially decaying dominance functions, which was extended in [WCH02] to personalize the animation to a particular speaking style. A limitation of this approach is the hand-tuning of the interpolation function is extremely time consuming. In [EGP02] the interpolation function is generated automatically and is based on the distribution of multidimensional morphable model (MMM) parameters belonging to specific phonemes. However, although the MMM was shown to transfer to different faces [CE05] the animation parameters are limited to that specific form of model.

Rather than interpolating between static targets, an alternative approach is to stitch together sequences of visual speech based on some animation unit [BCS97, CTFP05, CG00, MCP*06], which typically are selected from a training corpus by minimizing a cost function that trades off a measure of similarity between candidate and desired phonetic contexts and the smoothness at the concatenation boundaries. The units selected from the corpus might be fixed length, e.g. [BCS97] or variable length [CG00, MCP*06]. The advantage of variable length units is that the longest possible sequences of real data are extracted from the corpus so there are fewer concatenation discontinuities. However, the quality of the speech animation depends on the amount of data available. The visual appearance of the underlying units (phonemes/visemes) is highly context dependent, see Figure 2, so examples of each unit in as many different contexts as possible is required so the correct sample can be selected. Our dynamic units have the advantage that sufficient data is required only to learn the viseme clusters and because the units typically extend over multiple phones, coarticulation effects are explicitly captured by the unit.

Generative statistical models can be employed to model the joint distribution of acoustic and visual speech [Bra99, ECR07]. These distributions can be sampled given new acoustic speech as input to estimate the maximum likelihood facial animation parameters, which can then be applied to the visual model. Disadvantages of this approach are that mapping from acoustic speech to the resulting facial movements is highly speaker-specific, and so this approach lacks flexibility, and not all of the facial motion during speech can be determined from the acoustics. It is important to ensure that the choice of features and mapping function capture the perceptually significant variance, and that any error in the mapped facial motion is not in perceptually important speech regions (e.g. lip closure).

A major limitation of all previous approaches is that the representation of speech is ultimately either a phonemic or a static visemic transcription, and information related to the dynamics of the speech is lost. A dynamic gesture, referred to as an *anime*, was presented in [CTFP05], but this unit was also tied to the underlying phone structure of the speech. Our dynamic unit for speech animation is derived from analysis of real visual speech, and is *not* tied directly to the underlying phones. Rather we seek a series of canonical speech gestures that, after clustering, form sets of related gestures that we refer to as *dynamic visemes*. These visemes each serve a particular visual function — they represent a specific action on the visible speech articulators. To generate speech we search *all* the possible dynamic viseme sequences that are plausible from the entire phoneme string. This is much more powerful and expressive than assuming a fixed mapping, and is a more accurate model of speech production.

3. Speech Corpus

The data used in this work are drawn from an audio-visual corpus containing an actor reciting 2542 sentences from the TIMIT sentence list in a neutral speaking style. The video was recorded at 29.97 frames per second at a resolution of 1920 by 1080 progressive scan and runs to approximately 8 hours. Both a frontal view and side view of the actor were captured, but only the frontal view was used for this work (see Figure 3). All sentences were manually annotated to obtain a phonetic alignment using the Arpabet phonetic transcription code.

4. Discovering Dynamic Visemes

4.1. Parameterising Visual Speech

Active Appearance Models (AAMs) [CET01] provide a means for tracking the speech articulators in a video. The shape of an AAM is defined by the two-dimensional vertex locations of a mesh that delineates the inner and outer lip contours and the jaw (Figure 3):

$$\mathbf{s} = \{x_1, y_1, x_2, y_2, ..., x_N, y_N\}^T.$$

The model is built by hand-labelling a small number of training images with the vertices that define the mesh. These training meshes are then normalised for similarity, and principal component analysis (PCA) gives a compact linear model of the form:

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

where $\mathbf{s}_0$ is the mean shape and $\mathbf{s}_i$ are the shape basis vectors. The coefficients p_i are the shape parameters, which define the encoding of $\mathbf{s}$.

The appearance of an AAM is defined over the pixels within the shape mesh, $\mathbf{x} = (x,y)^T \in \mathbf{s}_0$. The appearance is constructed by warping each training image to the mean shape and then applying PCA to give a compact linear model of appearance variation:

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

where the coefficients λ_i are the appearance parameters, $A_0(\mathbf{x})$ is the mean appearance, and $A_i(\mathbf{x})$, are the appearance basis vectors.

In this work we use the *inverse compositional project-out* AAM algorithm [MB04] to track the facial features. For analysis, rather than building an AAM with a single appearance component (i.e. performing a single PCA on all of the pixels within the shape mesh), we use a multi-segment AAM [TMS*09], where different regions of the face are modelled as independent appearance components. This allows a model of the inner mouth that is not required to be linearly related to the surrounding appearance.

To construct a multi-segment AAM the images are segmented into two sub-regions, one containing the pixels of the inner-lip area and the other containing the remainder of the face pixels. Independent appearance models are then constructed for each sub-region and the corresponding appearance parameters are concatenated and normalised as follows:

$$\mathbf{b} = \begin{pmatrix} \mathbf{W}_p \mathbf{p} \\ \mathbf{W}_\lambda \lambda_1 \\ \lambda_2 \end{pmatrix} = U \mathbf{c} V^T = \sum_{i=1}^{q} \mathbf{j}_i c_i \tag{3}$$

where

$$\mathbf{W}_p = \sqrt{\frac{\sum_{i=1}^{n_2} \sigma_{\lambda_{2_i}}^2}{\sum_{i=1}^{m} \sigma_{p_i}^2}}, \quad \mathbf{W}_\lambda = \sqrt{\frac{\sum_{i=1}^{n_2} \sigma_{\lambda_{2_i}}^2}{\sum_{i=1}^{n_1} \sigma_{\lambda_{1_i}}^2}} \tag{4}$$

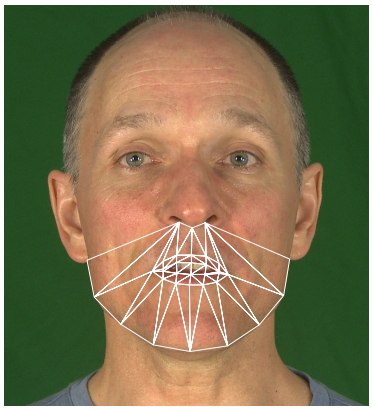

Figure 3: *The 34 vertex locations of the active appearance model mesh designed to capture the shape and appearance variation of the visible speech articulators.*

where $\mathbf{p}$ is a vector of shape parameters, and λ_1 and λ_2 are vectors of appearance parameters for the two segments of the multi-segment model. The number of dimensions of the respective appearance and shape models are n_1, n_2 and m, and $\sigma_{\lambda_{1_i}}^2$, $\sigma_{\lambda_{2_i}}^2$ and $\sigma_{p_i}^2$ represent the variance captured by each dimension of the respective model, $\mathbf{j}_i$ are the basis vectors spanning the combined shape and appearance space, and $\mathbf{c}$ is a 20-dimensional vector that compactly describes the combined shape and appearance variation of the lips and jaw during speech. The dimensionality of each of the respective models is selected such that a given proportion (we used 95%) of the total variation is captured.

4.2. Identifying Visual Gestures

Following [HTH10], we segment the AAM parameter trajectories corresponding to sentences into sequences of non-overlapping visual **gestures**, where the i^{th} gesture in a sequence, $\mathbf{G}_i$, is a sequence of AAM feature vectors that map a trajectory in AAM space representing a movement of the visible speech articulators. The boundaries between gestures are defined as salient points along the trajectory, which are identified by differentiating the gradient magnitude in 20D AAM parameter space, and locating the zero-crossings.

The motivation for identifying gesture boundaries in this way is that during speech the articulators do not move at a constant rate. Rather, they tend to accelerate away from articulatory targets and then decelerate as they approach the next target. This generates a visually intuitive and compelling segmentation, marking boundaries where the articulators change direction, or where they hit extreme poses, such as the lip closure during a bilabial. Figure 4 shows the gesture boundaries for an utterance and Figure 1 illustrates the asynchrony between phone and gesture boundaries. Further examples are given in the accompanying video.

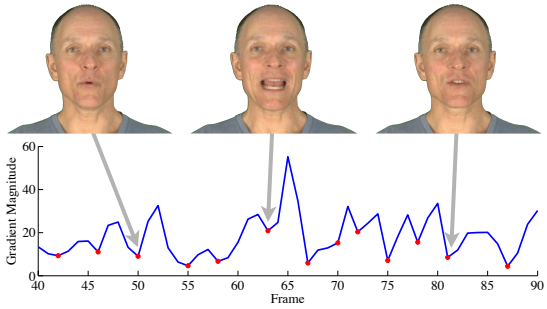

Figure 4: *Bottom: the gradient magnitude in 20D AAM parameter space over a sentence. The automatically derived gesture boundaries are highlighted as red dots. Top: the video frames corresponding to the segment boundaries.*

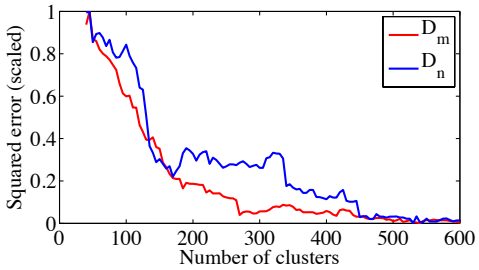

Figure 5: *The mean squared difference between the super-features and the respective cluster median for each gesture (D_m) and the nearest-neighbour from a different cluster (D_n). The number of clusters is varied over $k = \{40,45,50,\ldots,600\}$ and the errors have been scaled to the range $0\ldots1$ for visualization. The trade-off value for k is around 150 clusters.*

Our goal is to cluster a collection of segmented gestures from the training video into visually similar, variable length dynamic units. Rather than referring to visemes as the visually contrastive phonemes, we instead define a viseme as the **gestures** that have the same function visually. These gesture groups represent meaningful contrasts between visual speech utterances, and are the visual analog to the allophones of a phoneme.

4.3. Clustering Visual Gestures

Measuring the distance between multivariate time series data of arbitrary lengths is a non-trivial problem. Previous approaches include linearly resampling the data to a fixed length and then computing point-wise distances, calculating the cost of dynamically warping one sample to another using DTW [KR05] and modeling sequences as hidden Markov model (HMM) super-features [CSR06]. In this work we adopt the latter as this tends to generate better clusters.

To generate HMM super-features, a universal background model (UBM) in the form of a HMM is first trained using the AAM parameters for all of the gestures in the training video. Each state of the HMM, $\zeta_j(x)$, is represented as a multivariate Gaussian mixture model:

$$\zeta_j(x) = \sum_{k=1}^{M} w_{jk} N(x;\mu_{jk},\Sigma_{jk}), \tag{5}$$

where $N(x;\mu_{jk},\Sigma_{jk})$ denotes a multivariate Gaussian with mean μ_{jk} and covariance Σ_{jk}, M is the number of mixture components and w_{jk} is the weight of the k^{th} mixture component. For each gesture, the UBM is then updated using MAP adaptation [HM07], where the means of the mixture components are updated using:

$$\hat{\mu}_{jk} = \frac{N_{jk}}{N_{jk}+\tau}\bar{\mu}_{jk} + \frac{\tau}{N_{jk}+\tau}\mu_{jk}, \tag{6}$$

where μ_{jk} is the mean of the j^{th} state and k^{th} mixture component in the UBM, $\bar{\mu}_{jk}$ is the mean of the adaptation data, τ is the weight of a priori knowledge to the adaptation data, N_{jk} is the occupation likelihood of the adaptation data and $\hat{\mu}_{jk}$ is the updated mean.

The HMM super-features for each gesture are the vector difference between the UBM mean vectors and the MAP adapted mean vectors. The dimensionality of the super-features is $N \times M \times D$, where N is the number of states, M is the number of Gaussian mixture components and D is the dimension of the training features. In our case, the models are trained using the AAM parameters appended with the velocity and acceleration coefficients making $D = 20 \times 3$. We use an HMM with three emitting states, each with a single Gaussian mixture component in a left-to-right model with self-looping allowed, but no state skipping. The HMMs are trained using the EM algorithm from the Hidden Markov Model Toolkit (HTK) [YEG*06].

To generate the dynamic visemes, the super-features are clustered using a graph-based clustering algorithm [Kar02]. We find that this generates more visually appealing clusters than simple k-means clustering. The distance measure between gestures is the Euclidean distance between the respective super-features.

4.4. Determining the Number of Dynamic Visemes

To determine the number of dynamic visemes required to cover the visual speech space, two goodness-of-fit measures were computed for each of $k = \{40,45,50,\ldots,600\}$:

D_m the mean distance of the super-features to their respective cluster median.

D_n the mean distance of the super-features to the nearest sample that does not belong to the same cluster.

D_m will be large for ill-formed clusters because visually distinct gestures will be assigned to the same viseme. Thus D_m

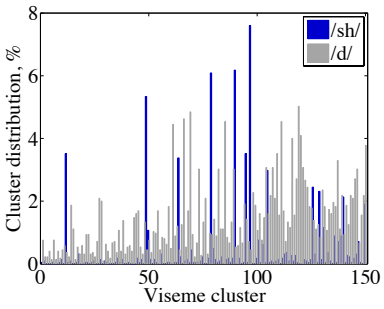

Figure 6: *The distribution of two phonemes across the dynamic visemes. The lack of (visual) consistency for the phonemes demonstrates that simply substituting viseme labels is a poor way of representing visual speech.*

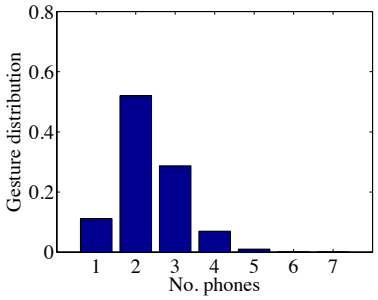

Figure 7: *The distribution of the number of phones forming a dynamic unit measured from our training corpus. Over 90% of the dynamic units span two or more phonemes.*

indicates if there are too few visemes. Conversely, D_n will be low when there are too many visemes as visually similar gestures are assigned to different visemes. These measures are plotted in Figure 5 as a function of the number of visemes. To determine the appropriate number of visemes, k, we locate the (approximate) *knee* of the curves [JMF99], which in this case is approximately 150 visemes.

4.5. Properties of Dynamic Visemes

The visemes identified by clustering (Section 4.3) represent sets of visually similar gestures. Therefore, if mapping phonemes to static visemes is valid, we would expect phonemes to be assigned to the same visual clusters consistently. Figure 6 shows that this is far from the case. Occurrences of the phonemes /sh/ and /d/ are distributed widely over the 150 dynamic visemes because their visual appearance varies in different phonetic contexts. Figure 7 shows the distribution of the number of phones in the dynamic units segmented in our training corpus, and we find that $\approx 90\%$ of these extend over two or more phones. By spanning multiple phone segments dynamic visemes naturally capture the effects of coarticulation in speech production.

5. Animating Speech with Dynamic Visemes

We have implemented speech animation using dynamic visemes on a 3D model artistically rigged using surface deformers in Autodesk Maya 2011. This represents an industry standard modeling and rigging approach. All gestures belonging to a dynamic viseme serve the same visual function, so each viseme is represented using the median visual gesture of those assigned to it during clustering. Each of the 150 visemes are animated on our model. Each animation is short, on average four frames, and need only be defined once for any given character. Figure 8 shows examples.

To animate speech when the dynamic viseme sequence is known, e.g. reanimating a training sequence, we simply concatenate the required dynamic visemes in the sequence, then blend at the boundaries to create a smooth join. Figure 9 shows example frames from a sequence animated in this way. To blend two gestures, the adjoining frames at the boundary are interpolated using Maya's cubic two-dimensional Bezier curve fitting function through an interpolating mean point defined on the adjoining half frame. Figure 10 shows the effect of the boundary smoothing. Note that only the boundaries are affected by the join — the viseme dynamics remain the same.

5.1. Mapping Phonemes to Dynamic Visemes

Given an input sequence of N phoneme labels, $P = p_1, p_2, ..., p_N$ with the corresponding durations, an output sequence of M dynamic viseme labels, $V = v_1, v_2, ..., v_M$, that best corresponds to the desired speech movements is required. To find this mapping we exploit our knowledge of the phoneme strings that are associated with the viseme clusters during training. Specifically, each viseme, v_i, has a number of variable length phoneme strings associated with it, corresponding to the constituent gestures assigned during clustering. Using these phoneme strings, we perform an exhaustive search to locate all possible sequences of visemes that could have given rise to the input phoneme sequence P.

As an example, if the target phrase is 'word', we first search for the instances of the phoneme string, $P =$ /w/, /er/, d/, in the viseme clusters. Any clusters that contain this sequence are identified as candidate viseme sequences. Next, we search for the phoneme substrings {/w/, /er/} and {/d/} and all combinations of viseme clusters containing these sequences are added to the candidate viseme sequences. Finally, we search for the sequences {/w/} and {/er/, /d/}. To account for asynchrony between the phoneme and dynamic viseme boundaries, phonemes corresponding to the end of one gesture are also allowed to appear at the beginning of the next gesture — see Figure 11. This is required since the boundaries between the phonemes and the dynamic visemes tend not to align.

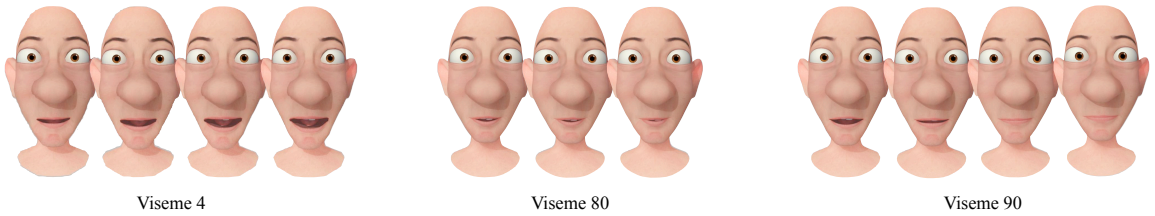

Figure 8: *Three example dynamic visemes animated by an artist on a surface-deformer model in Maya.*

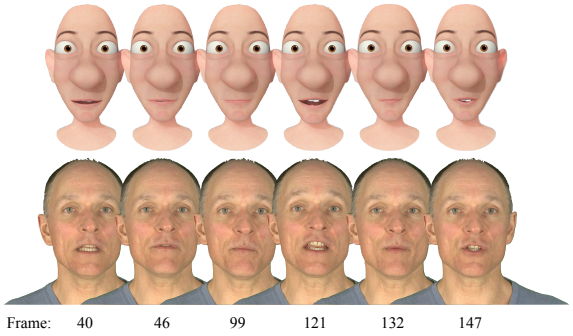

Figure 9: *Examples frames from an animation sequence for a face model (top row) and the corresponding video frames (bottom row).*

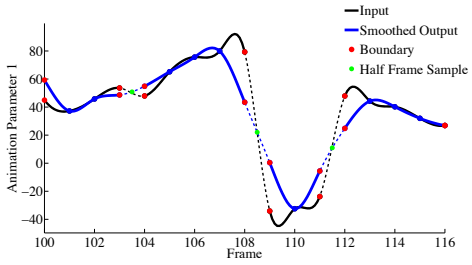

Figure 10: *Stitching dynamic sequences together requires only simple blending at the segment boundaries. The segment start and end key values (black curve, red points) are replaced with a half-frame, mid-value point (green). Default Maya anim curve interpolation computes new values (blue curve, red points) for the segment start and end key values without disrupting other key values.*

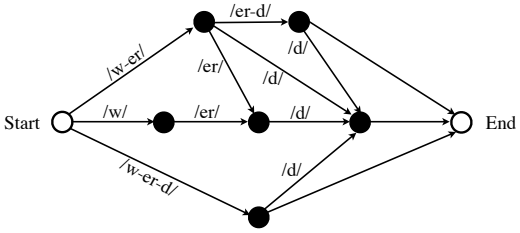

Figure 11: *Possible paths for mapping the phoneme string /w-er-d/ to visemes (black nodes).*

To select the best matching dynamic viseme sequence from the list of candidate samples each candidate is assigned a cost as follows:

$$c_i = \alpha(-\Pr(V_i|P)) + \beta(t_s(V_i,P)) + \gamma(d(V_i)), \qquad (7)$$

where V_i represents the i^{th} candidate viseme sequence. The first term in Equation 7 represents the probability of viseme sequences, V_i, given the phoneme string. This is calculated by summing the bigram log probabilities for the viseme pairs and the log probabilities of the respective phoneme substrings with respect to the viseme cluster:

$$\Pr(V|P) = \sum_{m=2}^{|V|} (\log(\Pr(v_m|v_{m-1}))) + \sum_{m=1}^{|V|} (\log(\Pr(p_m|v_m))), \qquad (8)$$

The second term in Equation 7 represents the cost of temporally aligning the dynamic visemes in V_i to the target sequence P in terms of duration. This term biases the viseme selection towards those that most closely match the speaking rate of the target sentence. The final term is a measure of discontinuity at the boundaries of the concatenated dynamic visemes in AAM space. The weights, α, β and γ are determined subjectively and are set to 0.699, 0.3 and 0.001 respectively. These parameters can be adjusted to vary properties of the output animation, but for all results in this paper these are the values used. On completion of the search algorithm the lowest cost viseme sequence is used to generate the output speech animation. See Figure 12.

An important point about the use of dynamic visemes is that the same phoneme string can map to a different sequence of dynamic visemes depending on the context in which phonemes appear, which is not the case for static visemes. As an example, instances of the word "another" are shown in Table 1. The center column shows the viseme sequence and the left and right columns show the context in which the word was spoken. Notice that the transcription of the word differs both in the number and in the composition of the dynamic visemes.

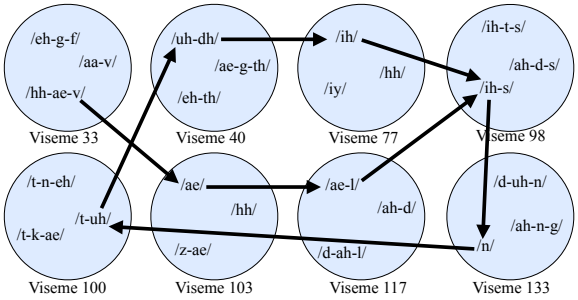

Figure 12: *Mapping variable length phoneme substrings for the sentence "Have a listen to this" to dynamic visemes.*

5.2. Viseme Alignment

Dynamic visemes are independent of phonemes, so the boundaries tend not to align. However, the viseme boundaries can be approximated from the (known) phone boundaries using:

$$v_i^e = \begin{cases} p_j^e, & \text{if } v_i^e \text{ does not intersect } p_j \\ \dfrac{p_{j-1}^e + p_j^e}{2}, & \text{otherwise} \end{cases} \quad (9)$$

where v_i^e represents the end frame of viseme i, and p_j^e represents the end frame of phoneme p_j. A viseme is assumed to intersect a phoneme if the phoneme label is split over two consecutive visemes, otherwise the boundaries are assumed to align. This exploits the phenomenon that humans do not perceive an offset of 80ms ($\approx$ 3 video frames) when the audio leads and 140 ms ($\approx$ 5 video frames) when the audio lags in speech [Sum92].

6. Results

6.1. Subjective Evaluation

Subjective evaluation was used to evaluate two aspects of our system: 1) The efficacy of dynamic visemes for modeling visual speech, where we reanimate 50 of the training sequences for which the correct viseme sequences are known, and 2) the quality of visual speech animated using 50 sentences held-out from training, which were generated using the phoneme-to-dynamic viseme lookup described in Section 5. In both cases we compare dynamic visemes with a phoneme-to-static key pose mapping taken from [PW96] where keyframes were placed at the midpoint of each phone segment and a cubic two-dimensional Bezier curve in Autodesk Maya 2011 was used to generate the intermediate frames. Diphthongs are not included in the phoneme-to-pose lookup described in [PW96], so we approximate them by concatenating the two corresponding vowels.

Thirty two participants took part in a pairwise preference test where, for each sentence, they were shown two movies side-by-side — one for each condition, dynamic viseme and static pose interpolation. They were played the left movie, followed by the right movie and finally both movies synchronously. After each sentence, viewers selected whether they preferred the left or the right movie. The order of the sequences and the left-right position on screen for each treatment were randomized for each participant. The results of the subjective study are shown in Figure 13.

Left context	Visemes (/ah-n-ah-dh-er/)	Right context
After	70-80-124	long pause...
(Silence)	134-80-101	memo for...
... one or	83-80-149	of the...
(Silence)	134-117-35	field had...
... can have	28-80-104	tunafish sandwich
(Silence)	145-45-145-148	longer strip...
(Silence)	145-80-69	brand of...
... pick up	123-80-5	pack on ...
(Silence)	145-80-1-137	put sex...
(Silence)	145-45-67-132	snarled close...
... ideas surfeit	117-80-133	sector of...
... progress,	145-45-80-134	is delineating...
... not try	75-80-134	club
(Silence)	145-45-67-125	stock vaudeville...

Table 1: *The center column shows the viseme sequences for the word "another" spoken in different contexts.*

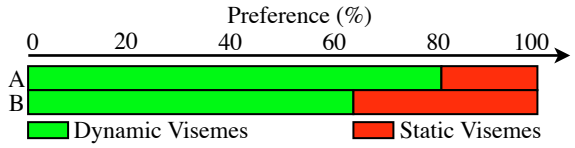

Figure 13: *Pairwise preferences averaged over thirty two participants for animations using: (A) the ground-truth dynamic visemes, and (B) dynamic viseme generated Equation 7.*

Case A: *Reanimating training data with known dynamic viseme sequences*. Viewers prefer ($p < 0.01$) animation generated using concatenated dynamic visemes to animation using a phoneme-to-static viseme lookup, on average, 80% of the time. This shows that these units are an effective visual analog of phonemes since a dynamic viseme is always the same example of the unit from the training video, and these are simply concatenated.

Case B: *Speech animation for unknown dynamic viseme sequences*. Viewers again prefer ($p < 0.01$) animation generated using concatenated dynamic visemes to animation using a phoneme-to-viseme lookup, this time on average, 62% of the time. Feedback suggests that even a single error in the selection of the animation units can severely impact the perceived quality.

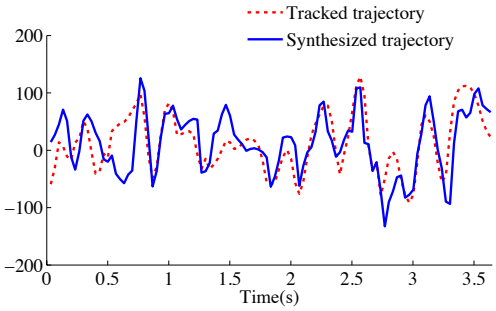

Figure 14: *Ground-truth parameter trajectory, c_1 over a sentence (red dotted line) and the corresponding parameter trajectory generated using the phoneme-to-dynamic viseme look up (blue solid line).*

	μ	σ
Training:	10.615	1.729
Testing:	13.578	2.211

Table 2: *The mean (μ) and standard deviation (σ) of the RMS error averaged over 50 sequences for AAM parameters generated both by re-synthesising known dynamic viseme sequences and sequences generated using Equation 7.*

6.2. Objective Evaluation

An example parameter trajectory for the first component of an AAM generated using the phoneme-to-dynamic viseme look up and the corresponding ground-truth trajectory as measured from a video sequence is shown in Figure 14. It is clear that the trajectory generated using the animation pipeline follows closely the desired trajectory. We have evaluated dynamic visemes numerically by comparing the generated $20D$ AAM parameter trajectory to the corresponding trajectory measured from a video sequence for 50 training sentences, for which the dynamic viseme sequence is known, and for 50 held-out sentences for which the phoneme-to-dynamic viseme search is used to generate the dynamic viseme sequence. The mean root-mean-square error averaged over the the 50 sentences for both the training and test cases are given in Table 2 with corresponding standard deviations. This also shows a slight increase in error in the test case, where viseme selection is based on Equation 7.

7. Summary

We have introduced a new dynamic unit for visual speech, which better represents the visual equivalent of phonemes than the traditional idea of static viseme shapes. A large corpus of video speech data is segmented into short sequences by compactly modeling the database using an AAM and then defining points of zero-crossing in parameter space acceleration as segment boundaries. This generates a set of reliable and visually intuitive speech gestures that cluster to 150 dynamic visemes that can be concatenated to animate speech.

We demonstrate a synthesizer that stitches together the dynamic viseme cluster centers using simple spline interpolation at the unit boundaries. The accompanying video shows several example animations including side-by-side static vs. dynamic viseme animation. The dynamic synthesizer has visual units that can, and often do, span more than one phoneme, so visual coarticulation across multiple phones is accounted for in the unit. We are able to map any phoneme string to dynamic visual speech animation by exhaustively searching the graph of viseme transitions to find possible viseme sequences that match the phoneme string.

An advantage of using dynamic visemes for speech animation is that they are applicable to any form of model or rigging. All that is required is that an animator must (creatively if desired) define the short dynamic viseme sequences once for the particular rig. An alternative is to define a mapping from the AAM parameterization to a new rig.

A subjective evaluation shows that animation resulting from dynamic visemes is more natural and plausible than animation generated using static pose interpolation. This reinforces the problems associated with traditional many-to-one mapping approaches, and suggests that dynamic visemes are a more suitable, if more complex, basis for visual speech.

7.1. Further Work

Future work will focus on improving the phoneme-to-dynamic viseme lookup. This is our initial attempt at such a mapping, and we note from feedback that if a single unit is used incorrectly in a sentence, the perceived quality of the entire sequence is significantly affected. Given our definition of a viseme, we expect this to be the case. This is supported by the findings in [Wit12] which considered the impact of different forms of error in animated speech sequences. We intend to investigate alternative cost functions (Equation 7).

We intend to build upon this work by extending the analysis to multiple speakers and considering other prosodic speech effects such as speaking rate, speaking volume, and emotion. We will also consider the relationship between dynamic visemes for different speakers, and in particular how animation can be transferred between speakers.

Acknowledgements

The authors would like to thank Dr. Graham Taylor for discussions regarding the use of HMM super-features, Valeria Reznitskaya for her help with scripting in Maya, and our recording actor Ken Bolden.

References

[BCS97] BREGLER C., COVELL M., SLANEY M.: Video rewrite: Driving visual speech with audio. In *Proceedings of SIGGRAPH* (1997), pp. 353–360. 3

[Bra99] BRAND M.: Voice puppetry. In *Proceedings of SIGGRAPH* (Los Angeles, California, 1999), pp. 21–28. 3

[CB05] CHUANG E., BREGLER C.: Mood swings: Expressive speech animation. *ACM Trans. on Graphics 24*, 2 (2005), 331–347. 3

[CE05] CHANG Y., EZZAT T.: Transferable videorealistic speech animation. In *ACM SIGGRAPH/ Eurographics Symposium on Computer Animation* (Los Angeles, July 2005), pp. 29–31. 3

[CET01] COOTES T., EDWARDS G., TAYLOR C.: Active appearance models. *IEEE PAMI 23*, 6 (June 2001), 681–685. 4

[CFP04] CAO Y., FALOUTSOS P. KOHLER E., PIGHIN F.: Real-time speech motion synthesis from recorded motions. In *Eurographics Symposium on Computer Animation* (2004), pp. 347–355. 3

[CG00] COSATTO E., GRAF H.: Photo-realistic talking-heads from image samples. *IEEE Transactions on Multimedia 2*, 3 (2000), 152–163. 3

[CH11] CAPPELLETTA L., HARTE N.: Viseme definitions comparison for visual-only speech recognition. In *Proceedings of the European Signal Processing Conference* (2011). 3

[CM94] COHEN M., MASSARO D.: Modeling coarticualtion in synthetic visual speech. In *Models and Techniques in Computer Animation*, Thalmann N., D T., (Eds.). Springer-Verlag, 1994, pp. 141–155. 3

[CSR06] CAMPBELL W., STURIM D., REYNOLDS D.: Support vector machines using GMM supervectors for speaker verification. *Signal Processing Letters 13*, 5 (2006), 308 – 311. 5

[CTFP05] CAO Y., TIEN W., FALOUTSOS P., PIGHIN F.: Expressive speech-driven facial animation. *ACM Transaction on Graphics 24*, 4 (2005), 1283 – 1302. 3

[ECR07] ENGLEBIENNE G., COOTES T., RATTRAY M.: A probabilistic model for generating realistic speech movements from speech. In *Proceedings of Advances in Neural Information Processing Systems* (2007). 3

[EGP02] EZZAT T., GEIGER G., POGGIO T.: Trainable videorealistic speech animation. In *Proceedings of SIGGRAPH* (2002), pp. 388–398. 3

[EP98] EZZAT T., POGGIO T.: Miketalk: A talking facial display based on morphing visemes. In *Proceedings of the Computer Animation Conference* (1998), pp. 96–103. 3

[Fis68] FISHER C.: Confusions among visually perceived consonants. *Journal of Speech and Hearing Research 11* (1968), 796–804. 2

[HM07] HOWARD L., MIRGHAFORI N.: Word-conditioned HMM supervectors for speaker recognition. In *INTERSPEECH-2007* (2007), pp. 746–749. 5

[HSLG04] HAZEN T., SAENKO K., LA C., GLASS J.: A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In *Proceedings of the International Conference on Multimodal Interfaces* (2004). 2

[HTH10] HILDER S., THEOBALD B., HARVEY R.: In pursuit of visemes. In *Proceedings of the International Conference on Auditory-Visual Speech Processing* (2010), pp. 154–159. 4

[JMF99] JAIN A., MURTY M., FLYNN P.: Data clustering: a review. *ACM Computing Surveys 31*, 3 (1999), 264–323. 6

[Kar02] KARYPIS G.: *CLUTO - A Clustering Toolkit*. Tech. rep., University of Minnesota, Department of Computer Science, Minneapolis, April 2002. 5

[KR05] KEOGH E., RATANAMAHATANA C. A.: Exact indexing of dynamic time warping. *Knowledge and Information Systems 7* (2005), 358–386. 5

[Mas98] MASSARO D.: *Perceiving Talking Faces*. The MIT Press, 1998. 2

[MB04] MATTHEWS I., BAKER S.: Active appearance models revisited. *International Journal of Computer Vision 60*, 2 (2004), 135–164. 4

[MCP*06] MA J., COLE R., PELLOM B., WARD W., WISE B.: Accurate visible speech synthesis based on concatenating variable length motion capture data. *IEEE Transactions on Visualization and Compuer Graphics 12*, 2 (2006), 266–276. 3

[MJ83] MONTGOMERY A., JACKSON P.: Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America 73*, 6 (1983), 2134–2144. 2

[MLV11] MATTHEYSES W., LATACZ L., VERHELST W.: Automatic viseme clustering for audiovisual speech synthesis. In *Proceedings of Interspeech* (2011). 3

[NN01] NOH J., NEUMANN U.: Expression cloning. In *ACM SIGGRAPH* (2001), pp. 277–288. 3

[OB86] OWENS E., BLAZEK B.: Visemes observed by the hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research 28* (1986), 381–393. 2

[PW96] PARKE F. I., WATERS K.: *Computer facial animation*. A. K. Peters, Ltd., Natick, MA, USA, 1996. 8

[SP54] SUMBY W., POLLACK I.: Visual contribution to speech intelligibility in noise. *Journal of Speech and Hearing Research 26*, 2 (March 1954), 212–215. 2

[Sum92] SUMMERFIELD Q.: Lipreading and audio-visual speech perception. *Phil. Trans. of the Roy. Soc. Series B: Biological Sciences 335*, 1273 (1992), 71–78. 8

[TMS*09] THEOBALD B. J., MATTHEWS I., SPIES J. R., BRICK T. R., COHN J. F., BOKER S. M., MANGINI M.: Mapping and manipulating facial expression. *Language and Speech 52*, 2/3 (2009), 369 – 386. 4

[VBPP05] VLASIC D., BRAND M., PFISTER H., POPOVIC J.: Face transfer with multilinear models. *ACM Transactions on Graphics 24*, 3 (2005), 426–433. 3

[WCH02] WANG Z., CAI L., HAZHOU A.: A dynamic viseme model for personalizing a talking head. In *International Conference on Signal Processing* (2002), pp. 1015–1018. 3

[Wil90] WILLIAMS L.: Performance driven facial animation. *Computer Graphics 24*, 2 (1990), 235–242. 3

[Wit12] WITHHELD: Relating objective and subjective performance measures for AAM-based visual speech synthesis. IEEE Trans. on Audio, Speech and Language Processing (accepted for publication), 2012. 9

[WSZP07] WAMPLER K., SASAKI D., ZHANG L., POPOVIĆ Z.: Dynamic, expressive speech animation from a single mesh. In *Eurographics Symposium on Computer Animation* (2007), pp. 53–62. 3

[XCXH03] XIANG CHAI J., XIAO J., HODGINS J.: Vision-based control of 3D facial animation. In *Eurographics Symposium on Computer Animation* (2003), pp. 193–206. 3

[YEG*06] YOUNG S. J., EVERMANN G., GALES M. J. F., HAIN T., KERSHAW D., MOORE G., ODELL J., OLLASON D., POVEY D., VALTCHEV V., WOODLAND P. C.: *The HTK Book, version 3.4*, 2006. 5