A probabilistic trajectory synthesis system for synthesising visual speech

Barry-John Theobald and Nicholas Wilkinson

School of Computing Sciences
University of East Anglia, Norwich, UK.
{bjt, nw}@cmp.uea.ac.uk

Abstract

We describe an unsupervised probabilistic approach for synthesising visual speech from audio. Acoustic features representing a training corpus are clustered and the probability density function (PDF) of each cluster is modelled as a Gaussian mixture model (GMM). A visual target in the form of a short-term parameter trajectory is generated for each cluster. Synthesis involves combining the cluster targets based on the likelihood of novel acoustic feature vectors, then cross-blending neighbouring regions of the synthesised short-term trajectories.

The advantage of our approach is that it explicitly captures coarticulation effects and yields naturally increase and decrease with the likelihood of the acoustic feature vectors.

Index Terms: visual speech synthesis, speech-driven talking heads

1. Introduction

Speech is multimodal in nature. The sounds of speech are generated, in part, by the movement and configuration of the speech articulators, some of which are visible. A listener uses these visual cues to assist in the perception of spoken words whenever the face of the speaker is visible. A powerful demonstration of this is the well studied McGurk Effect [1]. These visual cues associated with speech production are referred to as visual speech.

Approaches for synthesising visual speech are either speech-driven [2, 3, 4, 5, 6, 7], or text-driven [8, 9, 10, 11, 12]. Speech-driven approaches map directly from parameterised auditory speech to the corresponding visual gestures, whilst text-driven approaches perform the mapping indirectly via a string of phonemes. Usually text-driven systems adopt either a trajectory synthesis or a concatenative approach, but a recent development [13] combines both. The trajectory synthesis component syntheses smoothly varying parameter trajectories, and the concatenative component uses the synthesised trajectory to index segments from a training corpus. The advantage is that idiosyncratic gestures that are apparent in natural speech but usually missing in traditional trajectory synthesis systems are captured by the concatenative component, and the trajectory synthesis component provides a good basis for the concatenative indexing.

The main advantage of speech-driven synthesizers is they generally are unsupervised, so do not require phonetic labels. The main disadvantage is mapping from speech acoustics to visual speech is non-trivial and most approaches perform only a frame-wise mapping, ignoring the temporal relationship between nearby frames. The advantage of text-based approaches is the phonetic labels are available during synthesis, so the influence of neighbouring speech segments can be predicted.

In this paper we present our new speech-driven approach that maps from speech acoustics to visual speech parameters using a probabilistic model. For each acoustic feature vector in the training corpus a short-term trajectory of visual parameters is generated centred on the acoustic feature. The acoustic features are then clustered and the probability density function (PDF) of each cluster modelled as a Gaussian mixture model (GMM). A target (visual) trajectory is generated for each cluster as a weighted average of the corpus examples, where the weights reflect the likelihood of the acoustic features given the cluster. Synthesis involves combining the cluster targets based on the likelihood of novel acoustic feature vectors, then cross-blending neighbouring regions of the synthesised short-term trajectories.

The advantage of our approach over other speech-driven approaches is that instead of learning the mapping from an individual acoustic feature vector to the corresponding visual feature vector, we instead use all of the visual targets at any one time to generate a visual feature vector, where the influence of cluster targets naturally increase and decrease with the likelihood of the corresponding acoustic features. Mapping in this way also explicitly captures coarticulation effects.

2. Active Appearance Models

Active Appearance Models (AAMs) [14] are generative parametric models commonly used to track [15] and synthesise [3, 12] 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 speech animation as both the geometry and the texture of the face is captured in a single model, which can later be used to synthesise near-photorealistic images of faces.

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

$$ s = s_0 + \sum_{i=1}^{m} s_i p_i, $$

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 [14]. The base shape $s_0$ is the mean shape and the vectors $s_i$ are the (reshaped) eigenvectors corresponding to the $m$ largest eigenvalues. An example shape model is shown in Figure 1.

The appearance, $A(x)$, of an AAM is defined as the pixels that lie inside the base mesh, $x = (x, y)^T \in s_0$. AAMs allow linear appearance variation, so $A(x)$ can be expressed as a base appearance $A_0(x)$ plus a linear combination of $l$ appearance

$$ A(x) = A_0(x) + \sum_{i=1}^{l} \alpha_i A_i(x), $$

where $\alpha_i$ are the appearance parameters.
The training data used in this work is a recording of the Messiah sentences, see [16] for a list of the specific sentences. A single speaker recited all of the sentences in a single session, and the video was recorded using a camera mounted on a helmet worn by the speaker. The resolution of the captured video is 360x288 pixels (one quarter DV-PAL), and the audio sampled at 11025 Hz, 16 bits/sample stereo. The speaker maintained, as far as possible, a neutral speaking style (no emotion) to confine the variation of the facial features to those associated with speech production.

The digitised auditory speech is parameterised as Mel-frequency cepstral coefficients (MFCCs), where for each 10ms window of speech the first 13 (0–12) MFCCs are computed, as is typical in speech recognition. Approximately thirty images were selected from the training video to cover a broad range of the facial gestures associated with speech production. These were hand-labelled with vertices to define the shape component of the AAM, and the model built and used to label automatically all frames in the training video — see [14] for a description of the algorithm used.

To overcome the asynchrony in the sampling of the audio and visual information (100Hz acoustic versus 25Hz visual), the AAM parameter trajectories representing each utterance are up-sampled from 25Hz to 100Hz by fitting a cubic spline to the parameters in each dimension, then re-sampling at four times the original frame-rate.

4. Mapping Audio-to-Visual Speech

The task of the synthesiser is to receive as input an acoustic (MFCC) feature vector and determine the corresponding configuration of the visible articulators (in terms of AAM parameters).

4.1. Training the Synthesiser

The MFCC feature vectors from the training corpus are clustered by first constructing a single component GMM. The mixture component is then divided into two by perturbing the mean by ±0.2σ and the model retrained. Next, the component of the GMM with the largest mixture-weight is again split, giving three components, and the model retrained. The process is repeated until some desired number of clusters, N, have been found — we are currently conducting subjective experiments to determine the affect of N. This partitioning of the acoustic feature space has the advantage that it does not require a phonetic transcription of the audio, so the acoustic features fall into a natural grouping rather than being forced into a cluster based on phonetic labels attached to the frames. We consider each cluster to represent an acoustic event rather than some linguistic unit of speech since the feature vectors assigned to a cluster are unlikely to all belong to a single allophone.

To allow anticipatory and perseverative coarticulation effects to be taken into account, the static visual features are converted into fixed-duration trajectories to model the short-term temporal behaviour about each MFCC vector. A trajectory is formed by concatenating the k AAM feature vectors either side of each point in the AAM parameter trajectory. More formally:

\[ \nu_k = \text{concat}(\{p_{i-k}; \lambda_{i-k}\}; \{p_{i-k+1}; \lambda_{i-k+1}\}; \ldots; \{p_{i}; \lambda_{i}\}; \ldots; \{p_{i+k-1}; \lambda_{i+k-1}\}; \{p_{i+k}; \lambda_{i+k}\}) \]

where \( \text{concat}(\bullet) \) denotes the concatenation operator and \( k \) signifies the number of frames either side of each feature vector to append. Thus assuming a q-dimensional AAM, the feature vector at each time step is of dimension \((2k + 1)q\). As with the choice of the number of clusters, the effect on the perceived quality of
the synthesised visual speech as $k$ varies is currently undergoing subjective evaluation. For this study we select $k = 5$ frames and generate $N = 45$ clusters.

Each acoustic event, $C_n$, is represented as a multi-dimensional Gaussian, $C_n \sim \mathcal{N}(\mu, \Sigma)$, where $\mu$ is the mean computed from the MFCC vectors assigned to the event and $\Sigma$ a diagonal covariance matrix. A (visual) target trajectory, $t_n$, and a weight, $w_n$, are generated for each event using:

$$t_n = \frac{\sum_{j=0}^{J} p(m_j) \nu_j}{\sum_{j=0}^{J} p(m_j)},$$

(4)

and

$$w_n = \frac{\sum_{j=0}^{J} p(m_j) (t - \nu_j) (t - \nu_j)' \nu_j}{\sum_{j=0}^{J} p(m_j)},$$

(5)

where the subscript $n$ denotes an index for the acoustic events, $J$ is the number of MFCC vectors forming the training data and $p(m_j)$ is the probability of $C_n$, emitting the MFCC vector in the $j^{th}$ frame. Thus the visual trajectory used to represent an event is a weighted average of the trajectories in the training corpus, where the weights reflect the likelihood of the corresponding MFCCs belonging to event $C_n$. The weights, $w_n$, reflect the dispersion of the data about the target trajectory. Those with a low weight correspond to more significant (i.e. stable) gestures than those with high weight. Example target trajectories corresponding to two events are shown in Figure 3.

![Figure 3: Example trajectories representing two acoustic events: (a) a trajectory of the second shape parameter (mouth opening/closing) capturing a stop-like action, where an increasingly positive value corresponds to mouth more closed, and (b) a trajectory of the fourth shape parameter (lip rounding) capturing a diphthong-like action of the form /æ/ , where an increasingly positive value corresponds to lips more rounded. For $k = 5$, the trajectories correspond to a duration of 0.11s.](image)

4.2. Synthesising Novel Utterances

The synthesiser first computes MFCC vectors from novel acoustic speech, and for each auditory feature vector, a visual trajectory is synthesised using:

$$\{p; \lambda\} = \frac{\sum_{n=0}^{N} p_n(m_j) t_n w_n^{-1}}{\sum_{n=0}^{N} p_n(m_j)},$$

(6)

where $m_j$ are the acoustic features representing the novel utterance at time-step $j$, $p_n(m_j)$ is the probability $m_j$ was emitted by event $C_n$, and $t_n$ and $w_n$ are the target (visual) trajectories and weights respectively. The synthesised trajectories are thus a weighted average of the $N$ target trajectories, where the overlapping regions, from Equation 4, are cross-blended, requiring $2k + 1$ frames to be buffered before synthesis can begin. Notice the generation of the synthesised trajectories uses only soft decision rules. This allows coarticulation effects to be considered, where the influence of target vectors increase and decrease as the likelihood of the corresponding MFCC feature vectors increase and decrease over time.

5. Results

We evaluate our approach by synthesising fifty sentences not included in training the synthesiser. The correlation between the original parameters measured from video and the synthesised parameter generated by our new approach is computed for the first three main modes of variation for the shape and appearance components of the AAM. They are summarised in Table 1, and example trajectories are shown in Figure 4 with examples frames shown in Figure 5. We are currently performing subjective assessment of the effect of $k$ and $N$ on the perceived quality of the synthesise visual speech generated by our approach.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Shape</th>
<th>Appearance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.74 ± 0.04</td>
<td>0.62 ± 0.10</td>
</tr>
<tr>
<td>2</td>
<td>0.67 ± 0.07</td>
<td>0.71 ± 0.09</td>
</tr>
<tr>
<td>3</td>
<td>0.56 ± 0.16</td>
<td>0.67 ± 0.09</td>
</tr>
</tbody>
</table>

6. Conclusions and Further Work

In this paper we have described a new approach for synthesising visual speech from speech acoustic. The approach is probabilistic in nature, using soft decision rules to generate the visual parameters. The approach is entirely unsupervised and is completely data-driven. MFCC vectors are first clustered and the probability density of each cluster modelled as a Gaussian mixture model. A short-term visual trajectory corresponding to each acoustic feature vector is generated, from which visual targets are assigned to each cluster. New parameter sequences are generated as a weighted combination of these visual targets, where the weights are likelihood that (novel) MFCC vectors belong to the corresponding cluster.

The approach described in this paper works well, but key speech gestures, such as the point of closure on a stop, is sometimes only approximated — as can be seen in Figure 5. In each iteration of the current clustering algorithm the mixture component with the largest mixture weight is divided, leaving short-duration stops being modelled as only a single event. To overcome this we will investigate three approaches. Firstly pruning the training data to remove outlier data samples (and all silence before and after each utterance) to provide a better representation of $C_n$ and $t_n$. Secondly, this trajectory synthesis approach could be integrated into our existing concatenative approach [12], in a similar fashion to [13], to better capture the short-term dynamics apparent in natural visual speech. Thirdly we will investigate methods for iteratively refining the cluster.
targets and weights using an analysis-by-synthesis strategy.

7. Acknowledgements

The authors gratefully acknowledge the support of EPSRC (EP/D049075/1) and Prof. Stephen Cox for his comments.

8. References


Figure 4: Example AAM parameter trajectories for the first three shape parameters over an utterance. For each parameter the three signals represent the parameter measured directly from video (solid line), synthesised using our probabilistic model (dashed line) and a mapping using artificial neural networks (ANN) (dotted line). The ANN approach is shown here as a baseline to compare our method against. This is a standard method for mapping from audio to visual speech. Note, for our new approach the general trend of parameters one and two is correct (jaw displacement and mouth opening respectively), but the result is somewhat under-articulated. The magnitude of the displacements about zero are lower than in the original case. We outline possible solutions to this in Section 6.


Figure 5: Video frames generated using our new probabilistic trajectory synthesis method. Note the subtle under-articulation apparent in the trajectories in Figure 4 result in the mouth not being fully closed for the point of closure in /p/.