# Synthesizing a Talking Mouth

Ziheng Zhou[*]
Machine Vision Group,
University of Oulu
Oulu, Finland
ziheng.zhou@ee.oulu.fi

Guoying Zhao
Machine Vision Group,
University of Oulu
Oulu, Finland
guoying.zhao@ee.oulu.fi

Matti Pietikäinen
Machine Vision Group,
University of Oulu
Oulu, Finland
mkp@ee.oulu.fi

## ABSTRACT

This paper presents a visually realistic animation system for synthesizing a talking mouth. Video synthesis is achieved by first learning generative models from the recorded speech videos and then using the learned models to generate videos for novel utterances. A generative model considers the whole utterance contained in a video as a continuous process and represents it using a set of trigonometric functions embedded within a path graph. The transformation that projects the values of the functions to the image space is found through graph embedding. Such a model allows us to synthesize mouth images at arbitrary positions in the utterance. To synthesize a video for a novel utterance, the utterance is first compared with the existing ones from which we find the phoneme combinations that best approximate the utterance. Based on the learned models, dense videos are synthesized, concatenated and downsampled. A new generative model is then built on the remaining image samples for the final video synthesis.

## 1. INTRODUCTION

Realistic speech animation has been an active topic in recent years for its wide range of potential real-world applications [17]. In human-computer interaction, for instance, instead of outputting text or audio only to communicate with users, an animated talking head synchronized with the audio/text may attract more attentions from users and make such applications more engaging. A visually realistic talking head could make users feel comfortable and natural and hence improve the quality of the human-machine interactions. Moreover, such animation techniques may be used to generate visual cues for audio clips so as to help hearing-impaired people understand machine responses better.

The key for such an animation task is to realize a visually realistic talking mouth since most of the dynamic shape and texture changes on face appear in the mouth area. In general, animating such a mouth consists of two steps: analysis and synthesis. In the analysis step, recorded video corpus is processed to learn models that capture the dynamics (*e.g.*, co-articulations) of a talking mouth. After that, new videos are synthesized for novel utterances based on the learned models.

In this paper, we present a speech animation system for synthesizing a talking mouth. Instead of modelling single phonemes, we consider an utterance (*e.g.*, a word, phrase or short sentence) as a continuous dynamic process and modelled by a novel generative model learned from the video footage. The model consists of two components: a continuous low-dimensional curve that characterizes the temporal relations between both the seen (video frames) and the unseen (to be interpolated) mouth images occurring during the process of uttering, and a linear transformation that maps a point on the curve to a mouth image, resulting in a fast image-synthesis process based on the model. Through sampling points on the curve and projecting them into the image space, the generative model allows us to stretch or compress (or, in other words, adjust the speaking speed for) any part of the utterance in the synthesized video.

To synthesize a video for a new utterance, we borrow the idea of concatenating video sequences corresponding to different phonetic transcripts in [5]. In our work, however, the number of phonemes in the transcripts is not limited to, for example, three [5]. The transcripts can contain phonemes representing any proportion of an existing utterance and the generative models learned from the training video footage are responsible for generating videos for concatenation. Instead of directly concatenating videos, we learn a new model for the whole utterance to diminish discontinuities in the synthesized video and to time-align the video to the new utterance.

The rest of this paper is organized in the following way. In Section 2, we review the literature related to our work. Section 3 describes how to construct a generative model. Section 4 provides details of how to synthesize videos for unknown utterances. Experiments and results are presented in Section 5. Finally, Section 6 concludes our work and gives future work.

## 2. BACKGROUND

Audio-visual speech synthesis [1, 19] has long been an active topic in the communities of computer graphics and vision. Generally speaking, most of the existing methods can be categorized as either model-based or image-based. The model-based methods [6, 15, 11] attempt to produce a 3D

[*]Corresponding author

*ICVGIP* '10, December 12-15, 2010, Chennai, India
Copyright 2010 ACM 978-1-4503-0060-5 ...$10.00.

model, either geometric or biomechanics-driven, to simulate a talking head. The image-based approaches, on the other hand, construct a talking-head model directly from video footage of the human subject, achieving relatively high levels of video-realism. We consider our mouth synthesizer as an image-based method and will give a brief review on some of the established research work.

Bregler *et al.* [5] developed a pioneering system called Video Rewrite. The model for animation was simply a collection of short triphone video segments of a talking mouth. An error function was defined to measure the similarity between two triphones based on the phoneme-context distance and the distance between lip shapes extracted from the video segments. Given a transcript of phonemes, a collection of existing triphone were chosen to approximate the utterance using dynamic programming. The corresponding video segments were then picked from the database and concatenated to produce the videos.

Cosatto and Graf [9] also proposed a speech animation system following the idea of choosing images from the existing footage to synthesize new videos. In their work, the face was decomposed into different facial parts. These parts were located, parameterized, normalized and stored in the database. Given a phonetic transcript for synthesis, candidates of mouth images were selected from the database for each phoneme. Distances between these images were defined and the best image sequence was found using the Viterbi algorithm. After that, bitmaps of facial parts were projected to the base face for the final synthesis.

Ezzat *et al.* [13] introduced a multidimensional morphable model (MMM) to model a talking mouth (or face). Such a model was comprised of a set of prototype images that represent various lip texture and a set of prototype flows that represent the correspondences between a reference image and other prototype images. Given MMM parameters, the target image was synthesized as a linear combination of images warped from the prototype images. During the training stage, a Gaussian was trained for each phoneme in the MMM space. These distributions were then used to find the MMM parameters for the given phonetic transcript through regularization. A video was then generated by projecting the MMM parameters back into the image space.

Theobald *et al.* [20] used the active appearance model (AAM) [8] to model the shape and appearance of a talking head. Instead of triphone video segments, they stored the parameter trajectories corresponding to various triphones in AAM space. When synthesizing videos for the input phonetic transcript, triphone trajectories were selected based on some phoneme-context errors, concatenated, smoothed by fitting cubic splines and temporally warped to the desired duration. Face images were synthesized from the final trajectory representing the target utterance.

It can be seen that video realism of the synthesized face may be achieved by storing and re-organizing the original facial images in the training video corpus [5, 9] or by a generative model [10, 13, 20] that is learned from the training images and provides a way to project its model parameters back into the image space. The former methods have the advantage of being able to preserve the fine facial textures and natural dynamics (if video segments are stored) in synthesized videos, but typically require the storage of a large number of sample images (*e.g.*, the triphone model [5]) to show correct mouth motion, dynamics and coarticulation effects because of their inability to generate novel images.

On the other hand, the generative models (MMM [13] and AAM [10, 20]) produce new faces from any valid data points in the model parameter space, which turns the problem of video synthesis in the image domain into the problem in the domain of model parameters, that is, to synthesize a trajectory best representing the target utterance. The parameterization of facial images allows the use of different model to capture video dynamics (*e.g.*, the hidden Markov model [4, 21] and the Gaussian representation of phonemes [13]). The above generative models also have their disadvantages. For instance, to control the dimension of the parameter space, the principle component analysis (PCA) [3] is implemented to compress original images, resulting in, for instance, the smoothed skin textures or the blurring of tongues and teeth. Moreover, subtle dynamics of a talking mouth may be lost after the smoothing of the synthesized trajectories.

In this work, we present a novel generative model to address the issues mentioned above. Unlike the MMM and AAM, which rely PCA to reduce the number of model parameters, our model is built on the original video frames of an utterance. After training, the model is controlled by a single variable which determine the temporal position of the image to be interpolated within the utterance. In this way, the model is capable of reflect the fine facial textures in the synthesized videos. We can then produce any part of an existing utterance easily, allowing us to preserve the true dynamics of a talking mouth without exhaustive sampling of, for example, triphone segments [5]. Moreover, inside the model, the utterance is represented by a deterministic and analytic continuous curve, instead of a synthesized and smoothed one, from which we are able to reproduce the original training images.

# 3. LEARNING GENERATIVE MODELS

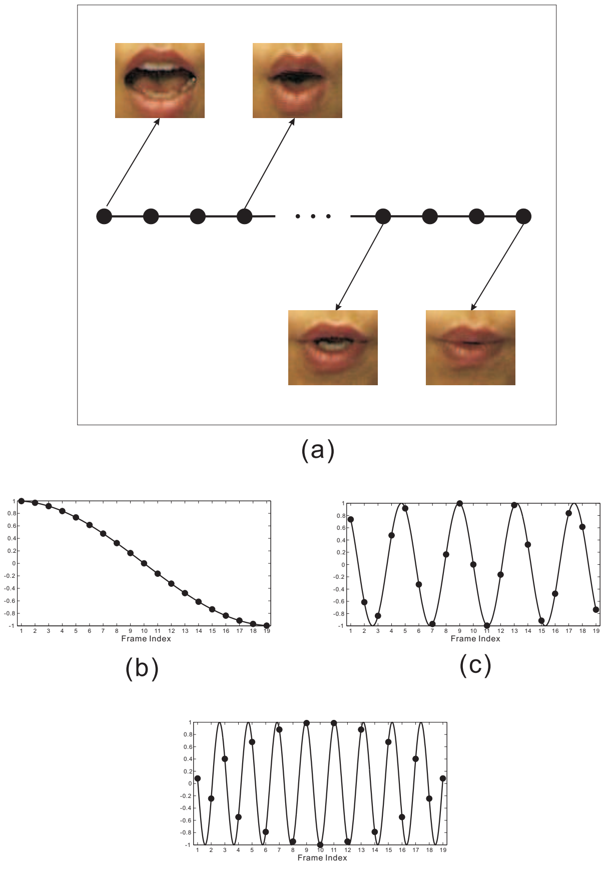
## 3.1 Graph Representation

As mentioned above, we consider the movement of a talking mouth as a continuous process. Therefore, an input video of an utterance can be viewed as a set of image samples sampled at a fixed pace along the curve that represents the utterance in the image space. Typically, the image space has a high dimension and we may assume that there exists a low-dimensional manifold within which the continuous process of uttering can be characterized by a determined and continuous function.

In our work, we reveal such a function through representing the input video as a path graph $P_n$ [12] where $n$ is the number of vertices. An example of such a graph representation is given in Figure 1(a). As shown in the figure, each vertex corresponds to a frame in the video and the connections between the vertices can be represented by an adjacency matrix $\boldsymbol{W}\in\{0,1\}^{n\times n}$ where $W_{ij}=1$ if $|i-j|=1$, $i,j=1,2,\ldots,n$ and 0 otherwise. As described in [2], to get the manifold embedded in the graph, we can consider the problem of mapping $P_n$ to a line so that connected vertices stay as close as possible. Let $\boldsymbol{y}=(y_1,y_2,\ldots,y_n)^{\mathrm{T}}$ be such a map and we can obtain $\boldsymbol{y}$ by minimizing

$$\sum_{i,j}(y_i-y_j)^2\,W_{ij},\quad i,j=1,2,\ldots,n. \tag{1}$$

It is equivalent to calculate the eigenvectors of the graph Laplacian $\boldsymbol{L}$ [7] of $P_n$. The matrix $\boldsymbol{L}$ is defined as: $\boldsymbol{L}=$



(a)

(b)

(c)

(d)

**Figure 1: (a) Graph ($P_{19}$) representation of an input video of the utterance 'How are you' with 19 frames in total and (b)-(d) the 1st, 9th and 18th eigenvectors of the Laplacian of the graph. Each eigenvector has a dimension of 19 and the value of its $i$th element is marked by the dot at the frame index $i$. The dash lines show the curves of trigonometric functions $f_1^{19}$, $f_9^{19}$ and $f_{18}^{19}$ on which the eigenvectors lie.**

$\boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D}$ is a diagonal matrix with the $i$th diagonal entry computed as $D_{ii} = \sum_{j=1}^{n} W_{ij}$. According to the definition of $\boldsymbol{L}$, it is not difficult to verify that it has $n-1$ eigenvectors $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n-1}\}$ with non-zero eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_{n-1}$ and the $u$th element ($u = 1, 2, \ldots, n$) of $\boldsymbol{y}_k$ ($k = 1, 2, \ldots, n-1$) is determined by:

$$\boldsymbol{y}_k(u) = \sin\left(\pi k u/n + \pi(n-k)/n\right). \tag{2}$$

If we replace $u$ by $t = u/n$ in Equation 2, $\boldsymbol{y}_k$ can be viewed as a set of points on the curve described by functions $f_k^n(t) = \sin\left(\pi k t + \pi(n-k)/n\right)$, $t \in [1/n, 1]$ sampled at $t = 1/n, 2/n, \ldots, n/n$. Figures 1(b)-(d) illustrate the 1st, 9th and 18th eigenvectors (black dots) of path graph $P_{19}$ and functions $f_1^{19}$, $f_9^{19}$ and $f_{18}^{19}$ (dashed curves). It can be seen that the temporal relations between the video frames are governed by the curve, which motivates us to make an assumption that the unseen mouth images occurring in the continuous process of uttering can also be characterized by the $n-1$ dimensional curve defined by function $\mathcal{F}^n$ : $[1/n, 1] \to \mathbb{R}^{n-1}$:

$$\mathcal{F}^n(t) = \begin{bmatrix} f_1^n(t) \\ f_2^n(t) \\ \vdots \\ f_{n-1}^n(t) \end{bmatrix}, \tag{3}$$

that is, we are to use function $\mathcal{F}^n$ to temporally interpolate images at arbitrary positions within the utterance.

### 3.2 From $\mathcal{F}^n$ to Images

To find the correspondences for the curve $\mathcal{F}^n$ in the mouth-image space, we start from mapping the image frames of the input video to the points defined by $\mathcal{F}^n(1/n), \mathcal{F}^n(2/n), \ldots, \mathcal{F}^n(1)$. Given a video with $n$ frame, we first vectorize images and denote them as $\{\boldsymbol{\xi}_i \in \mathbb{R}^m\}_{i=1}^n$. Here $m$ is the dimension of the image space. Typically, $n \ll m$ and we assume that image frames $\boldsymbol{\xi}_i$ are linearly independent. The mean image $\bar{\boldsymbol{\xi}}$ is calculated and removed from $\boldsymbol{\xi}_i$. The reason for doing that will be described shortly. The mean-removed vectors are denoted as $\boldsymbol{x}_i = \boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}}$. Based on the assumption on $\boldsymbol{\xi}_i$ and the mean-removal operation, we have a rank $n-1$ matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$.

Recall that we represent the video by graph $P_n$ with adjacency matrix $\boldsymbol{W}$. By using the linear extension of graph embedding [22], we can learn a transformation vector $\boldsymbol{w}$ that minimizes

$$\sum_{i,j} \left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j\right)^2 W_{ij}, \quad i, j = 1, 2, \ldots, n. \tag{4}$$

Vector $\boldsymbol{w}$ can be computed as the eigenvector of the following generalized eigenvalue problem:

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w} = \lambda^{'}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}. \tag{5}$$

He *et al.* [16] solved the above problem using the singular value decomposition [14] on $\boldsymbol{X}$, *i.e.*, $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ and then turned the problem into a normal eigenvalue problem

$$\begin{aligned} \boldsymbol{A}\boldsymbol{v} &= \lambda^{'}\boldsymbol{v} \\ \boldsymbol{A} &= (\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}})^{-1}(\boldsymbol{Q}\boldsymbol{L}\boldsymbol{Q}^{\mathrm{T}}) \\ \boldsymbol{Q} &= \boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}. \end{aligned} \tag{6}$$

such that $\boldsymbol{w} = \boldsymbol{U}\boldsymbol{v}$. Since we remove the mean from all $\boldsymbol{x}_i$ which makes $\mathrm{rank}(\boldsymbol{X}) = n-1$, $\boldsymbol{Q} \in \mathbb{R}^{(n-1)\times n}$, $\boldsymbol{A} \in \mathbb{R}^{(n-1)\times(n-1)}$, and they are both of full rank.

Let $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{n-1}$ be the eigenvectors of $\boldsymbol{A}$ with their eigenvalues $\lambda_1^{'} \leq \lambda_2^{'} \leq \cdots \leq \lambda_{n-1}^{'}$. From Equation 6, for any of its eigenvectors, $\boldsymbol{v}_k$ ($k = 1, 2, \ldots, n-1$) we have:

$$\begin{aligned} (\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}})^{-1}(\boldsymbol{Q}\boldsymbol{L}\boldsymbol{Q}^{\mathrm{T}})\boldsymbol{v}_k &= \lambda_k^{'}\boldsymbol{v}_k \\ \Rightarrow \boldsymbol{L}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{v}_k &= \lambda_k^{'}\boldsymbol{Q}\boldsymbol{v}_k \end{aligned} \tag{7}$$

It can be seen that vectors $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{v}_k$ are eigenvectors of $\boldsymbol{L}$. Therefore, we have

$$\begin{aligned} \lambda_k^{'} &= \lambda_k \\ \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{v}_k &= m_k\boldsymbol{y}_k \end{aligned} \tag{8}$$

where $m_k$ is a scaling constant. Without loss of generality, $m_k$ can be evaluated as the ratio of the first element of vector $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{v}_k$ to the first element of $\boldsymbol{y}_k$:

$$m_k = \frac{\sum_{i=1}^{n-1} \boldsymbol{Q}_{i1}\boldsymbol{v}_k(i)}{\boldsymbol{y}_k(1)}. \tag{9}$$

Let $\boldsymbol{M}$ be a diagonal matrix with $M_{kk} = m_k$, $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n-1}]$ and $\boldsymbol{\Upsilon} = [\boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2, \ldots, \boldsymbol{\upsilon}_{n-1}]$. From Equation 8 and $\boldsymbol{Q} = \boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}$, we have

$$\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{\Upsilon} = \left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}\right)^{\mathrm{T}}\boldsymbol{\Upsilon} = \boldsymbol{Y}\boldsymbol{M}. \qquad (10)$$

Recall that vectors $\boldsymbol{y}_k$ are determined by a set of trigonometric functions $f_k^n$ (see Equation 2). We can write matrix $\boldsymbol{Y}$ as:

$$\begin{aligned}\boldsymbol{Y} &= \left[\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n-1}\right] \\ &= \begin{pmatrix} f_1^n(1/n) & f_2^n(1/n) & \cdots & f_{n-1}^n(1/n) \\ f_1^n(2/n) & f_2^n(2/n) & \cdots & f_{n-1}^n(2/n) \\ \vdots & \vdots & \ddots & \vdots \\ f_1^n(n/n) & f_2^n(n/n) & \cdots & f_{n-1}^n(n/n) \end{pmatrix} \qquad (11)\end{aligned}$$

From Equation 3, $\boldsymbol{Y}^{\mathrm{T}} = [\mathcal{F}^n(1/n), \mathcal{F}^n(2/n), \ldots, \mathcal{F}^n(1)]$. We then have

$$\left(\boldsymbol{M}^{-1}\boldsymbol{\Upsilon}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{x}_i = \mathcal{F}^n(i/n), \quad i = 1, 2, \ldots, n. \qquad (12)$$

So far, we have found the map from the image frames to their correspondences on the curve defined by $\mathcal{F}^n$ through Equation 12. Now the question is raised that whether such a map is reversible. Once again, since the mean $\bar{\boldsymbol{\xi}}$ is removed from $\boldsymbol{\xi}_i$, resulting in $\mathrm{rank}(\boldsymbol{X}) = n-1$, $\boldsymbol{\Upsilon}$ is a $(n-1)\times(n-1)$ square matrix of full rank and hence, $\boldsymbol{\Upsilon}^{-1}$ exists. From Equation 12, we have

$$\boldsymbol{x}_i = \boldsymbol{U}\left(\boldsymbol{\Upsilon}^{-1}\right)^{\mathrm{T}}\boldsymbol{M}\mathcal{F}^n(i/n). \qquad (13)$$

It can be seen that the map is reversible and therefore, given any $t \in [1/n, 1]$, we can synthesize an image $\boldsymbol{\xi}^{\mathrm{syn}}$ by:

$$\boldsymbol{\xi}^{\mathrm{syn}} = \boldsymbol{U}\left(\boldsymbol{\Upsilon}^{-1}\right)^{\mathrm{T}}\boldsymbol{M}\mathcal{F}^n(t) + \bar{\boldsymbol{\xi}}. \qquad (14)$$

For color images, the synthesis should be carried out in each of the color channels. In this work, we use the RGB color model.

Very often, the dimension of the image space is much larger than the overall number of images contained in the video corpus used for training. In such a case, we can perform the PCA to represent images of the same speaker in a more compact way. To preserve all the fine facial textures, we keep all the eigenvectors of the covariance matrix with non-zero eigenvalues. Let $\boldsymbol{W}_{\mathrm{pca}}$ be the transformation matrix whose columns are the eigenvectors. After performing PCA, we can obtain a set of vectors $\tilde{\boldsymbol{\xi}}_i = \boldsymbol{W}_{\mathrm{pca}}^{\mathrm{T}}(\boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}})$ $(i = 1, \ldots, n)$ in the PCA domain. We then redefine vectors $\boldsymbol{x}_i$ as $\boldsymbol{x}_i = \tilde{\boldsymbol{\xi}}_i - \bar{\tilde{\boldsymbol{\xi}}}$ where $\bar{\tilde{\boldsymbol{\xi}}}$ is the mean of $\tilde{\boldsymbol{\xi}}_i$. Equation 14 can then be rewritten as

$$\boldsymbol{\xi}^{\mathrm{syn}} = \boldsymbol{W}_{\mathrm{pca}}\left(\boldsymbol{U}\left(\boldsymbol{\Upsilon}^{-1}\right)^{\mathrm{T}}\boldsymbol{M}\mathcal{F}^n(t) + \bar{\tilde{\boldsymbol{\xi}}}\right) + \bar{\boldsymbol{\xi}}. \qquad (15)$$

## 3.3 Usage of the Learned Generative Model

Equation 14 allows us to synthesize a mouth image from $t$ that locates the image within the utterance spoken in the input video. Since any proportion of the utterance can be represented by an interval $[t_1, t_2] \subseteq [1/n, 1]$, to synthesize an $n'$-frame video, we only need to sample $n'$ values of $t$ within the interval and generate frames using Equation 14. Given any frame rate, we can simply change $n'$ to control the duration of the synthesized video. Moreover, the sampled values do not have to be equally spaced in $[t_1, t_2]$. In the synthesized video, we can prolong/shorten any part of the utterance by sampling more/less densely in the interval corresponding to the part. It can be seen that the learned model provides an effective and efficient way to manipulate utterances in the training corpus.

## 3.4 Discussions on Linear-Independence Assumption

The validity of the synthesis formula, Equations 14 and 15, depends on the reliability of the assumption that all the frames of an input video, $\boldsymbol{\xi}_i$ are linearly independent. In this work, the utterances included in the video corpus are either single words, phrases or short sentences. The videos contain no more than 30 frames and are no longer than a couple of seconds. It has been tested that the assumption holds on the data. In case of the video frames being linearly dependent, we suggest two ways to tune the input video. Firstly, we may downsample the video (*e.g.*, using only the odd frames) to make the images linearly independent. The other way we may try is to divide the input video into subsequences somewhere, for instance, the speaker does not utter, such that the assumption holds for each subsequence. We can then learn a model for each of the subsequences.

# 4. SYNTHESIZING VIDEOS FOR NOVEL UTTERANCES

In [5], Bregler *et al.* proposed to concatenate triphone videos to synthesize a video for a novel utterance. The synthesis was done in the following way: 1) Previously stored triphone videos were selected based on the distances between the tripones and those in the utterance. 2) The selected videos were then stitched together and time-aligned to the utterance.

In this work, we borrow the idea of synthesizing videos through concatenating various video sequences. However, we implement this idea in a very different way. Firstly, the video to be concatenated are not stored, but generated by the previously learned models. The generated videos do not have to be a triphone video, but can be arbitrary proportions of the existing utterances. Moreover, we do not simply concatenate various video sequences, but learn a new model for the mouth-appearance continuity and time alignment.

## 4.1 Generating Videos for Concatenation

A novel utterance $\mathcal{U}$ can be labelled by a sequence of phoneme or a phonetic transcript. Our goal is to synthesize a video that describes the transitions within the phoneme sequence. If the exact phoneme transitions were found in the existing utterances, we would be able to synthesize a video using the models learned for those utterances. However, very often, the exact transitions cannot be found and we will have to search for subsequences of phonemes of the existing utterances that best approximate $\mathcal{U}$.

To maximally preserve the dynamics in the recorded video corpus, we match $\mathcal{U}$ against every existing utterance to find out the maximum overlap between their phoneme sequences. An $l$-phoneme overlap is defined as follows. Let $p_i, p_{i+1}, \ldots, p_{i+l-1}$ be the phoneme subsequence found from $\mathcal{U}$ and $\mathcal{Q} = q_j, q_{j+1}, \ldots, q_{j+l-1}$ the counterpart from an existing utterance where $i$ and $j$ are phoneme indices. It is required that $p_i$ and $q_j$ are in the same viseme category. Same are $p_{i+l-1}$

and $q_{j+l-1}$. For the rest of the phonemes, $p_{i+k} = q_{j+k}$, $k = 2, \ldots, l-1$.

Having had a match for $p_i, p_{i+1}, \ldots, p_{i+l-1}$, the rest of the $L$-long phoneme sequence of $\mathcal{U}$, $p_1, p_2, \ldots, p_i$ and $p_{i+l-1}, p_{i+l}, \ldots, p_L$ are matched again against all the existing utterances to search for the maximum overlap. Such an operation is carried on until a series of subsequences, $\mathcal{Q}_1, \mathcal{Q}_2, \ldots$ being found to approximate $\mathcal{U}$. After that, we will be able to use those already learned models to generate a video for each $\mathcal{Q}_i$ for concatenation.

## 4.2 Stitching Videos

Given $\mathcal{Q}_i$, we synthesize a dense video $\mathcal{V}_i$ from the learned model. By dense, we mean that its frame rate is relatively high (*e.g.*, one frame per millisecond) such that there are a number of frames marking every phoneme in $\mathcal{Q}_i$. To concatenate two videos $\mathcal{V}_i$ and $\mathcal{V}_{i+1}$, we calculate the Euclidean distances between images corresponding to the last phoneme of $\mathcal{V}_i$ and those corresponding to the first phoneme of $\mathcal{V}_{i+1}$. Based on the distances, we search for a pair of frames that look most similar and then concatenate $\mathcal{V}_i$ and $\mathcal{V}_{i+1}$ based on the found frames.

As will be shown in Figures 3(a)-(c), simply concatenation of all the synthesized videos $\mathcal{V}_i$ may cause discontinuity in the image or PCA domain. In this work, we evaluate our work in the PCA domain. To model $\mathcal{U}$ as a continuous process, we downsample the concatenated video at a frame rate (*e.g.*, one frame per 25 milliseconds) that makes the sampled frames linearly independent. After that, a model is learned using the method described in Section 3.2. In this way, we are able to synthesize images around the places where discontinuities occur.

## 4.3 Time Alignment

Another advantage of learning a generative model from downsampled image frames is that we can do time alignment very easily for the final output video. As mentioned in Section 3.3, we can represent any part of the utterance using an interval $[t_1, t_2] \subseteq [1/n, 1]$ where $n$ is the number of frames downsampled from the concatenated dense videos. For each phoneme $p_i$ in the phonetic transcript, we can use an interval $[t_1^i, t_2^i]$ to bound the phoneme. When synthesizing images for $p_i$, based on the length of $p_i$ in the phonetic transcript and the frame rate of the output video, we can calculate the number of images, $n_i$, to be synthesized for $p_i$. We then sample $n_i$ points that even spaced in $[t_1^i, t_2^i]$ and generate video frames using Equation 14 or 15.

When the duration of phoneme $p_i$ in the synthesized video is significantly (e.g., twice) longer than that in the original video, we have found that the linear sampling in $[t_1^i, t_2^i]$ could produce unnatural mouth movement. It is because the transition between $p_i$ and its previous/next phoneme is also significantly prolonged by such a sampling operation. To avoid the unnaturalness, we keep a short period (e.g., 30ms) at the beginning and end of $[t_1^i, t_2^i]$ unstretched to preserve the natural phoneme transitions.

# 5. EXPERIMENTS AND RESULTS

## 5.1 Video Corpus

We used the OuluVS database [23] as the training corpus. The database was originally designed for audio-visual speech recognition [18]. Twenty subjects were included in the database and each of them was asked to speak ten different utterances comprised of single words, phrases and short sentences. Videos in the database were recorded with a frame rate of 25 fps. The resolution of video frames was set as 720×576 pixels. Further detailed information about the database can be found in [23].

In our experiments, we preprocessed the original videos to get videos of a talking mouth. To do that, the eyes of the speaker were first located in video frames and moved to some fixed places in the image space through rotating and scaling the images. An 84×70 mouth region was then cropped from the frames and patched into videos which were later on used for training.

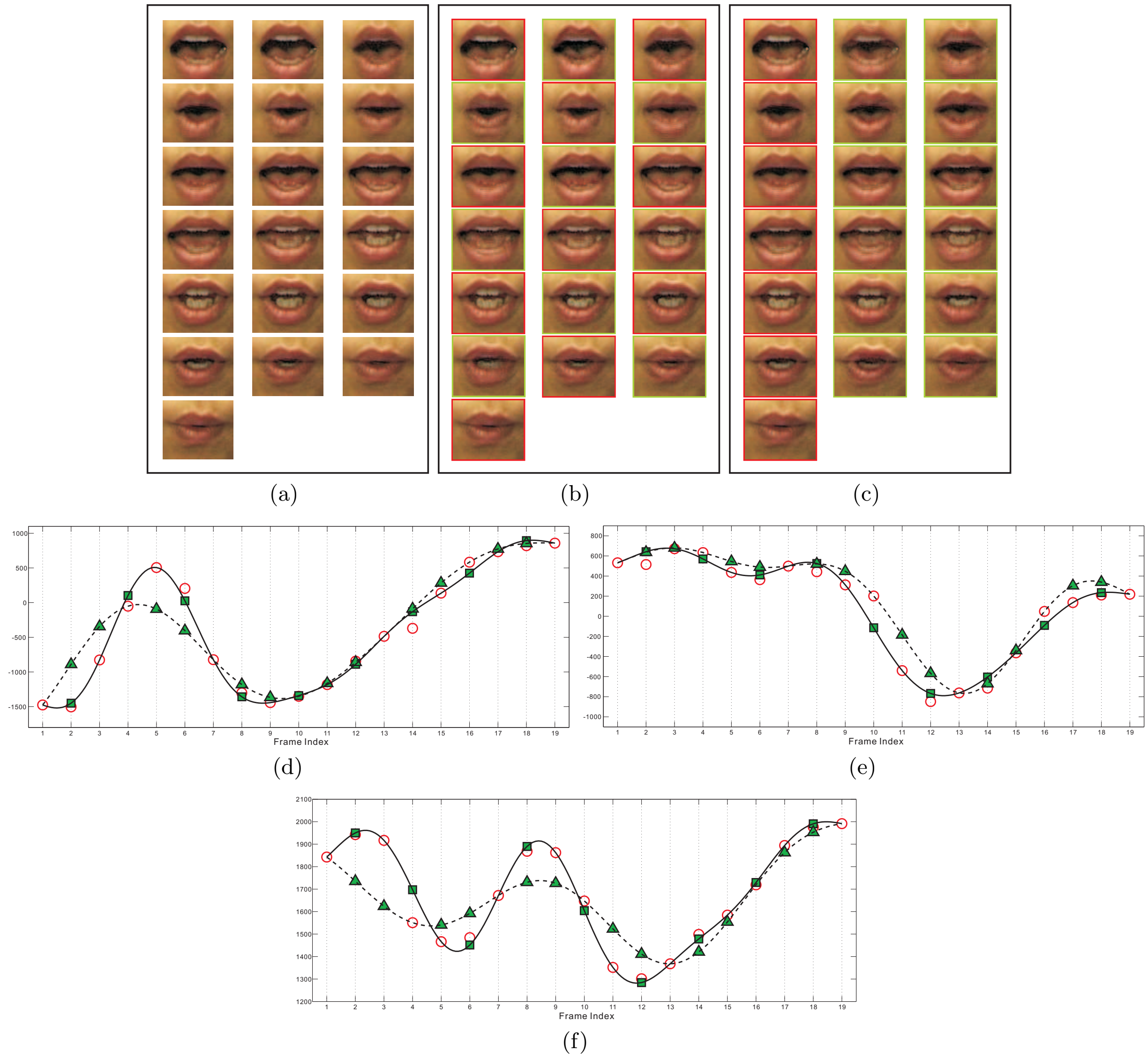
## 5.2 Evaluation of Using $\mathcal{F}^n$ to Model Utterances

Recall that we consider the mouth movement when speaking an utterance as a continuous process and assumed that the utterance can be modelled by function $\mathcal{F}^n$ which is formed by a set of trigonometric functions (see Equation 3). It is crucial that we evaluate such an assumption since it forms the foundation of the construction of the generative models.

In this experiment, we took a video as an example to demonstrate our method. To do that, we used part of its frames as training data and the rest as the ground truth for comparison. The original video contained 19 frames of a talking mouth speaking the utterance 'How are you'. Figure 2(a) shows all the 19 frames which are placed from left to right and top to bottom. We first trained a model on the odd frames (10 frames in total) and synthesized a new 19-frame video. Figure 2(b) shows the synthesized video where the red boxes mark the positions of the frames used for training and the green ones for testing. As guaranteed by our method, the images in the red boxes are duplicates of their counterparts in the original video. By comparing the green-boxed images with the original video frames, it can be seen that the synthesized images well catch the intermediate appearance (shape and texture) of the mouth between two red-boxed frames.

We further challenged our method by using fewer frames for training. At this time, we chose the first frame every three frames as illustrated by the red boxes in Figure 2(c). By doing so, we intentionally increase the dissimilarity between any two consecutive frames used for training. Once again, we can see that the synthesized images (green boxed) approximate the dynamics of the mouth well. Moreover, it can be seen that the blurring in one image used for training may cause blurring in its nearby synthesized images. Both experiments have demonstrated the capability of function $\mathcal{F}^n$ for modelling the dynamics of a talking mouth.

To quantify our visual findings, we project the original and synthesized videos along eigenvectors in the PCA domain. Here, we choose the eigenvectors with the largest eigenvalue in R channel, second largest eigenvalue in G channel and third largest eigenvalue in B channel. The corresponding eigenvalues are $1.68\times10^9$, $9.52\times10^8$ and $4.63\times10^8$, respectively. Figures 2(d), (e) and (f) show the projections from the original (circles) and synthesized images (squares and triangles) as well as the synthesized trajectories (the solid and dash lines). From the figures, we can see that quantitatively, the synthesized images are close to the original frames, which is consistent with their visual similarities.



(a) (b) (c)

(d) (e)

(f)

**Figure 2: (a) The original mouth video of a subject uttering 'How are you', (b) the synthesized video using odd frames for training and (c) the synthesized video using the first frame every three frames for training. In (b) and (c), the red boxes locate the frames used for training in the original video and the green ones mark the synthesized novel mouth images for comparison. Figures (d), (e) and (f) show the values of the above images along the eigenvectors with the largest eigenvalue in R channel, second largest eigenvalue in G channel and third largest eigenvalue in B channel. The circles mark the original frames and the squares and triangles correspond to the images in green boxes in (b) and (c), respectively. The solid and dashed lines show the synthesized trajectories projected on the eigenvectors.**

## 5.3 Evaluation of Synthesizing Videos for Novel Utterances

In the next experiment, we gave an example of synthesizing a video for a novel utterance. The utterance is kept short and simple, allowing us to show the synthesis results in a detailed way. Despite of the simplicity, the example illustrates the key ideas of video synthesis presented in this paper. We followed the way in [18] to define visemes except that the viseme of silence, /SIL/, was merged with the bilabial class (/M/, /P/ and /B/), resulting in 12 visemes in total. The novel utterance we chose to demonstrate our method was 'See me', which had a phoneme sequence $\mathcal{Q} =$ /S-IY-M-IY/. After comparing it with the existing utterances in the database, we obtained a sequence of phoneme combinations, $\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3$, that best approximated $\mathcal{Q}$. Here $\mathcal{Q}_1 =$ /S-IY/ from the utterance 'See you', $\mathcal{Q}_2 =$ /IY-SIL/ and $\mathcal{Q}_3 =$ /M-IY/. The latter two were both from the utterance 'Excuse me'.

Following the procedure described in Section 4, we first generated dense videos for $\mathcal{Q}_i$, $i = 1, 2, 3$ at the frame rate of one frame per millisecond and concatenated them. To show the discontinuities within the concatenated video, we projected the dense videos in the PCA domain. Figures 3(a)-

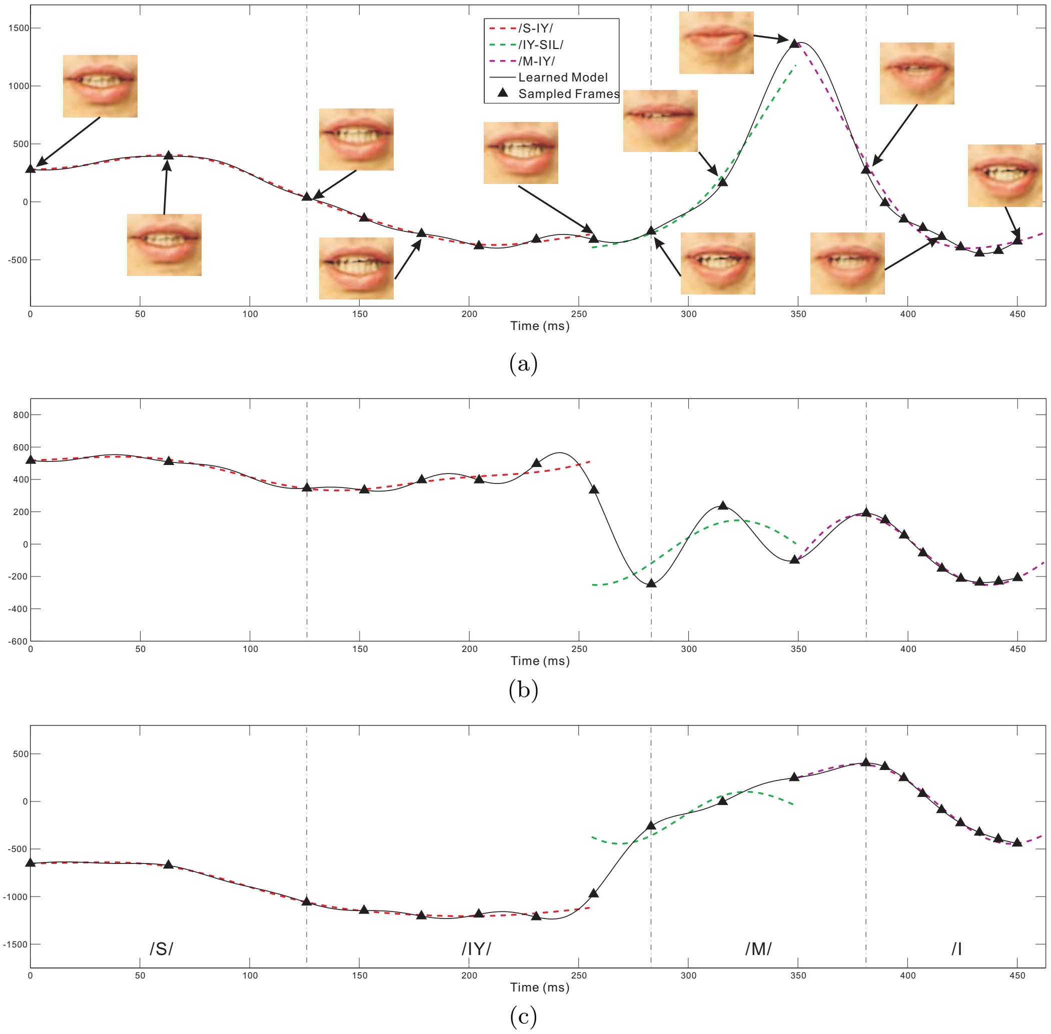


**Figure 3: Figures (a), (b) and (c) show the projected curve of the synthesized dense video for phoneme segments /S-IY/, /IY-SIL/ and /M-SIL/ along the three eigenvectors with the largest eigenvalues in the R channel by the red, green and magenta dashed lines, respectively. The vertical lines locate the boundaries of phonemes as identified in (c). The solid black line illustrates the curve synthesized by the model learned from the downsampled frames. The final synthesized video frames are sampled from the synthesized curve at the places marked by the black triangles.**

(c) show the projected curves along the three eigenvectors with the largest eigenvalues in the R channel. In the figures, the blue, red and green dashed lines are the curves projected from the dense videos synthesized for $\mathcal{Q}_1$, $\mathcal{Q}_2$ and $\mathcal{Q}_3$. The discontinuity can be easily found in all of the figures.

We downsampled the concatenated video frames at the rate of one frame per 25 milliseconds (or 40 fps) and learned a model for the utterance 'See me'. The black solid lines are the curves synthesized by the model to represent the whole utterance $\mathcal{Q}$. The model was then used to generate a 30 fps video according the lengths of the phonemes in the input phonetic transcript. Time-alignment was achieved by sampling image frames along the curves and the locations of the frames were marked by the black triangles in the figures. It can be seen that the learned generative model is capable of smoothing the discontinuities naturally, generating time-aligned videos easily and capturing the video dynamics for the utterance.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a video-realistic animation system for synthesizing a talking mouth. The system consists of two major components: 1) learning generative models from the recorded videos and 2) synthesizing new videos for novel utterances. A generative model is constructed on the assumption that an utterance can be modelled by function $\mathcal{F}^n(t)$ which is formed by a set of trigonometric functions embedded within a path graph. The transformation from $\mathcal{F}^n(t)$ is then found through graph embedding. To synthesize a video for a novel utterance, we compare it with the existing utterances and use the learned generative models to generate dense videos that best approximate the novel utterance. The videos are then concatenated and downsampled. A new generative model is constructed on the remaining image samples for diminishing discontinuity and easy time-alignment.

This work is part of our efforts towards a video-realistic animation system for synthesizing a talking head. As mentioned before, the visual realism of a synthesized talking head depends largely on how well we can synthesize a visually realistic talking mouth since the mouth region contains the most prominent dynamic changes on the face when we talk. The work of synthesizing such a mouth is already presented in this paper. In future, to complete the system of synthesizing a talking head, we plan to collect a new speech corpus for training, to investigate the methods of combining the synthesized mouth images with the background video frames, and to develop a fast searching algorithm for comparing novel utterances with the existing ones to find the best phoneme approximations.

## 7. ACKNOWLEDGEMENTS

This research was supported by the Academy of Finland.

## 8. REFERENCES

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

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of Advances in Neural Information Processing Systems*, volume 14, pages 585–591, Vancouver, Canada, 2001.

[3] C. Bishop. *Neural networks for pattern recognition.* Oxford University Press Inc., New York, NY, 1995.

[4] M. Brand. Voice pupperty. In *Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 21–28, Los Angeles, CA, 1999.

[5] C. Bregler, M. Covell, and M. Slaney. Video rewrite: driving visual speech with audio. In *Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 353–360, Los Angeles, CA, 1997.

[6] Y. Cao, W. Tien, P. Faloutsos, and F. Pighin. Expressive speech-driven facial animation. *ACM Transactions on Graphics*, 24(4):1283–1302, 2005.

[7] F. Chung. *Spectral graph theory (CBMS regional conference series in mathematics, No. 92).* American Mathematical Society, 1996.

[8] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

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

[10] D. Cosker, D. Marshall, P. Rosin, and Y. Hicks. Video realistic talking heads using hierarchical non-linear speech-appearance models. In *Proceedings of Mirage*, pages 2–7, Rocquencourt, France, 2003.

[11] Z. Deng, U. Neumann, J. Lewis, T.-Y. Kim, M. Bulut, and S. Narayanan. Expressive facial animation synthesis by learning speech coarticulation and expression spaces. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1523–1534, 2006.

[12] R. Diestel. *Graph Theory.* Springer-Verlag Heidelbery, New York, NY, 3rd edition, 2005.

[13] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. *ACM Transactions on Graphics*, 21(3):388–398, 2002.

[14] G. Golub and C. V. Loan. *Matrix Computations.* The John Hopkins University Press, Baltimore, MD, 3rd edition, 1996.

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

[16] X. He, D. Cai, S. Yan, and H. Zhang. Neighborhood preserving embedding. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, volume 2, pages 1208–1213, Beijing, China, 2005.

[17] I. Pandzic. Facial animation framework for the web and mobile platforms. In *Proceedings of the 7th International Conference on 3D Web Technology*, pages 27–34, Tempe, AZ, 2002.

[18] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[19] B. Theobald. Audio visual speech synthesis. In *Proceedings of International Congress on Phonetic Sciences*, pages 285–290, Saarbrücken, Germany, 2007.

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

[21] L. Xie and Z.-Q. Liu. A coupled HMM approach to video-realistic speech animation. *Pattern Recognition*, 40:2325–2340, 2007.

[22] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

[23] G. Zhao, M. Barnard, and M. Pietikäinen. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia*, 11(7):1254–1265, 2009.