MODELLING TALKING HUMAN FACES

Thesis submitted to Cardiff University in candidature for the degree of Doctor of Philosophy.

Samia Dawood Shakir

School of Engineering
Cardiff University
2019

DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed .....(candidate) Date .....

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of PhD.

Signed .....(candidate) Date .....

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff Universitys Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed .....(candidate) Date .....

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available online in the Universitys Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .....(candidate) Date .....

ABSTRACT

This thesis investigates a number of new approaches for visual speech synthesis using data-driven methods to implement a talking face.

The main contributions in this thesis are the following. The accuracy of shared Gaussian process latent variable model (SGPLVM) built using the active appearance model (AAM) and relative spectral transform-perceptual linear prediction (RASTAPLP) features is improved by employing a more accurate AAM. This is the first study to report that using a more accurate AAM improves the accuracy of SGPLVM. Objective evaluation via reconstruction error is performed to compare the proposed approach against previously existing methods. In addition, it is shown experimentally that the accuracy of AAM can be improved by using a larger number of landmarks and/or larger number of samples in the training data.

The second research contribution is a new method for visual speech synthesis utilising a fully Bayesian method namely the manifold relevance determination (MRD) for modelling dynamical systems through probabilistic non-linear dimensionality reduction. This is the first time MRD was used in the context of generating talking faces from the input speech signal. The expressive power of this model is in the ability to consider non-linear mappings between audio and visual features within a Bayesian approach. An efficient latent space has been learnt

using a fully Bayesian latent representation relying on conditional non-linear independence framework. In the SGPLVM the structure of the latent space cannot be automatically estimated because of using a maximum likelihood formulation. In contrast to SGPLVM the Bayesian approaches allow the automatic determination of the dimensionality of the latent spaces. The proposed method compares favourably against several other state-of-the-art methods for visual speech generation, which is shown in quantitative and qualitative evaluation on two different datasets.

Finally, the possibility of incremental learning of AAM for inclusion in the proposed MRD approach for visual speech generation is investigated. The quantitative results demonstrate that using MRD in conjunction with incremental AAMs produces only slightly less accurate results than using batch methods. These results support a way of training this kind of models on computers with limited resources, for example in mobile computing.

Overall, this thesis proposes several improvements to the current state-of-the-art in generating talking faces from speech signal leading to perceptually more convincing results.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my PhD supervisors, Dr. Yulia Hicks and Prof. David Marshall, for their advice, guidance, support and patience in helping me to complete this work.

I would also like to thank Prof. Neil Lawrence for the publicly available Gaussian process software and Prof. Barry-John Theobald for the publicly available audio-visual corpora used in this thesis.

Further gratitude and appreciation is expressed to my loved country, Iraq represented by Ministry of Higher Education and Scientific Research for funding my study and giving me this opportunity.

I would also like to thank my family, friends for their support and patience.

This PhD is dedicated to my father, mother and all other family members.

LIST OF ACRONYMS

3D	Three-dimensions
AAM	Active Appearance Model
ACC	Average Correlation Coefficient
AMSE	Average Mean Squared Error
ANN	Artificial Neural Network
ARD	Automatic Relevance Determination
ASM	Active Shape Model
BEEP	British English Example Pronunciation Dictionary
BLSTM	Bidirectional LSTM
CAS	Computer Assisted Craniofacial Surgery
CC	Correlation Coefficient
CHMMs	Coupled HMMs
CNN	Convolutional Neural Network
DBLSTM	Deep BLSTM
DFT	Discrete Fourier Transform
DNNs	Deep Neural Networks
DPDS	Deterministic Process Dynamical System
DTC	Deterministic Training Conditional
DTW	Dynamic Time Warp Cost
EM	Expectation-Maximisation

EMA	Electromagnetic Midsagittal Articulography
EVD	Eigenvalue Decomposition
FACS	Facial Action Coding System
FFNNs	Feed-Forward Neural Networks
FFT	Fast Fourier Transform
FITC	Fully Independent Training Conditional
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model
GP	Gaussian Processes
GPLVM	Gaussian Process Latent Variable Model
GPR	Gaussian Process Regression
HMM	Hidden Markov Model
ITU-T	International Telecommunication Union Telecommunication
KL	Kullback-Leibler divergence
LDS	Linear Dynamical System
LLD	Low Level Descriptors
LPC	Linear Predictive Coding
LSTM	Long Short-Term Memory
LSFs	Line Spectral Frequencies
LSPs	Line Spectral Pairs
MAE	Maximum Absolute Error
MAP	Maximum-a-Posteriori
MFCCs	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood
MLP	Multi-Layer Perception
MOS	Mean Opinion Scores

MRD	Manifold Relevance Determination
MSE	Mean Squared Error
PCA	Principal Component Analysis
PD	Probabilistic Descent
PFSA	Probabilistic Finite State Automation
PFT	Prediction Suffix Tree
PITC	Partially Independent Training Conditional
PLP	Perceptual Linear Predictive
PDM	Point Distribution Model
PPCA	Probabilistic Principal Component Analysis
PSL-UCSC	Perceptual Science Laboratory at University of California at Santa Cruz
RASTA-PLP	Relative Spectral Transform-Perceptual Linear Prediction
RAVDESS	Ryerson Auditory-Visual Dataset of Emotional Speech and Song
RBF	Radial Basis Function
RGB	Red, Green and Blue
RNN	Recurrent Neural Network
RMSE	Root Mean Squared Error
SAD	Sum of Absolute Differences
SGPDM	Shared Gaussian Process Dynamical Model
SGPLVM	Shared Gaussian Process Latent Variable Model
SLDS	Shared Linear Dynamical System
snakes	Active Contour Models
SNR	Signal to Noise Ratio
SSGPDM	Switching States Gaussian Process Dynamical Model
SVD	Singular-Value Decomposition

TTS	Text-to-Speech
VLMM	Variable Length Markov Model

LIST OF SYMBOLS

Facial Modelling Symbols

$\mathbf{x}$	Shape vector
$\mathbf{g}$	Texture vector
$\bar{\mathbf{x}}$	Mean of shape vector x
$\bar{\mathbf{g}}$	Mean of texture vector g
$\mathbf{P_x}$	Eigenvectors of shape vector x
$\mathbf{P_g}$	Eigenvectors of texture vector g
$\mathbf{Q}$	Eigenvectors of combined shape and texture
$\mathbf{b_x}$	Shape principal components
$\mathbf{b_g}$	Texture principal components
$\mathbf{b}$	Combined shape and texture principal components
$\mathbf{W_x}$	Diagonal matrix of weights
α	Scaling value
β	Offset value
U	Eigenvectors of $n \times n$ matrix
Λ	Eigenvalues of $n \times n$ diagonal matrix
$\mathbf{1}$	Vector of Ones
Ω	Eigenspace

Speech Processing Symbols

$s_t^{'}$	Hamming window
$X(F)$	Fourier transform of function x of t
$x(t)$	Inverse Fourier transform of function X of F
$X(k)$	Discrete Fourier transform
$x(n)$	Inverse Discrete Fourier transform
$E[z]$	z transform of e of n
$A[z]$	Transfer function of an an inverse filter
$s[n]$	Speech frame

Gaussian Processes Symbols

$k(\mathbf{x}, \mathbf{x}')$	Kernel function
Y	D-dimensional vector of data points
X	q-dimensional vector of latent points
θ	Gaussian process hyperparameters
$\mathbf{w}$	Automatic relevance determination weights
D	Dimensionality of data point
q	Dimensionality of latent point
α	Variance of the kernel
γ	Inverse width
δ_{ij}	Kroneckers delta function
β	Inverse variance
$\mathbf{W}$	Back-constraining mapping parameter set

Probability and Statistics, Information Theory Symbols

$P(\mathbf{y})$	Probability of y

$P(\mathbf{x}, \mathbf{y})$	Probability of x and y
$P(\mathbf{x}\|\mathbf{y})$	Probability of x given y
μ	Mean of univariate variable
$\boldsymbol{\mu}$	Mean of multivariate variable
σ	Variance
$\boldsymbol{\Sigma}$	Covariance matrix
$\mathbb{E}$	Expectation

Probabilistic Generative Models Symbols

$\mathbf{A}$	Transition probability matrix
$\mathbf{B}$	Emission probability distributions
$\mathbf{C}$	Observation matrix
$\mathbf{O}$	Observation matrix
$\mathbf{Q}$	Hidden state sequence
ϵ	Noise term

Variable Length Markov Model Symbols

Q	Finite set of states
Σ	Finite alphabet
τ	Transition function
γ	Output probability function
$\mathbf{s}$	Probability distribution over initial states
w	String prefix
σw	The suffix of word w

List of Figures

List of Tables

CONTENTS

Chapter 1

INTRODUCTION

Speech animation synthesis is the process of animating a face model to produce articulated movements that match accompanying vocal speech data. The issue of synthesising realistic talking faces is multifaceted, requiring the production of high-quality facial images, lip movements synchronised with the auditory input, and reasonable facial expressions. This is challenging because humans are expert at detecting subtle abnormalities in facial motion and auditory-visual synchronisation. Previous studies in visual speech animation can be categorised into two different classes: viseme-driven methods and data-driven methods [43]. Viseme-driven methods require key mouth shapes to be designed for phonemes to synthesise new speech animations, whereas data-driven methods need a pre-recorded facial motion dataset for synthesis purposes. Data-driven methods synthesise novel speech motions by combining prerecorded motion frame sequences (sample-based methods) or sampling from statistical models learned from the facial motion data (learning-based methods) [43].

Synthesising realistic visual speech animations has been a challenging task for decades, one of the major challenges being the phenomena of speech coarticulation, which complicates the mappings between speech signals and visual speech movements. Coarticulation refers to

changes in the articulation of a speech unit according to preceding or backward coarticulation and upcoming units or forward coarticulation. The animation parameters are typically synthesised utilising either a unit-driven approach [13, 57, 85] or a feature-driven [12, 48, 186]. Unit-driven methods use an indirect mapping of audio to visual speech, where trajectories of parameter values are formed from typically phoneme, diphone or triphone level representations of a sentence based on unit by unit. For unit-driven synthesis, longer-term coarticulation effects can be determined utilising phonetic context [172]. The disadvantages of these methods are that the dataset from which speech units are chosen do not include all possible phonetic contexts. The advantages of feature-driven methods which synthesise animation parameters as a direct mapping from parameterised auditory speech on a video frame by frame basis is that the articulators of speech are physically located to form the speech sound.

Auditory-visual mapping is highly non-linear because of ambiguities in both the auditory and visual domains [125]. High dimensional time series are endemic in applications of machine learning. Practical non-linear probabilistic methods for this data are required [33]. In this work, we focus on the auditory-visual mapping issue, aiming to model the non-linear relationship between auditory and visual features.

An overview of a general audio-driven visual speech synthesis is illustrated in Figure 1.1. An audio-visual corpus for visual speech synthesis is used. In the training stage, feature parameters of the auditory and visual signals are extracted then a mapping from auditory features to visual features is learned. In the synthesis stage, the trained model is utilised to estimate the visual parameters associated with an input

speech data.

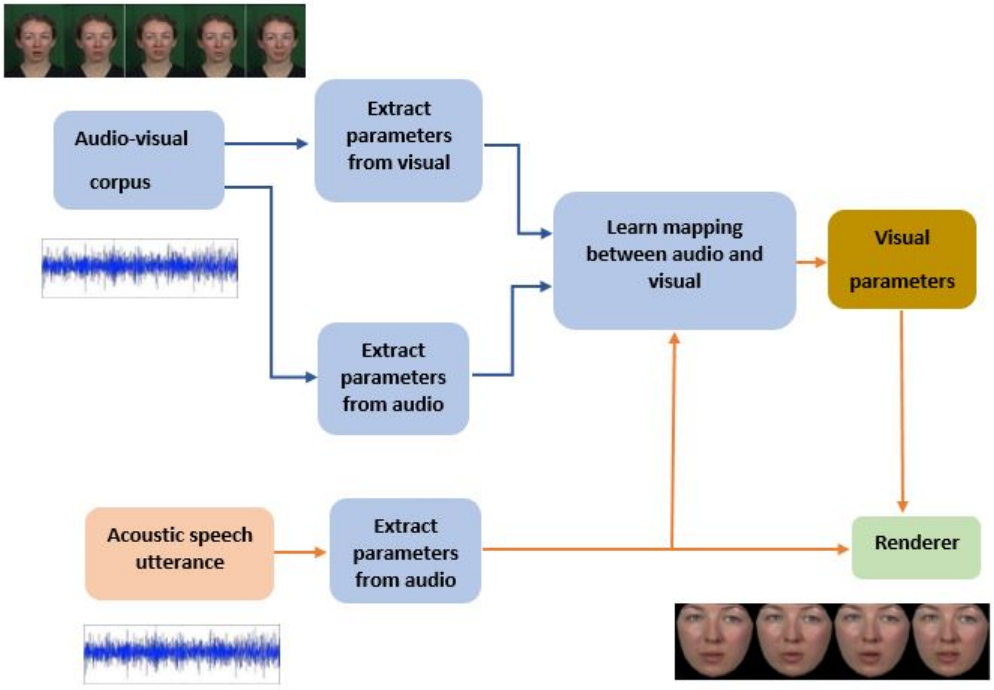


Figure 1.1. A general overview of the audio-driven visual speech synthesis. Training process is marked using the blue arrows, and synthesis process is marked by the orange arrows.

1.1 Applications

Over recent years, synthesis of realistically looking videos of moving human faces from speech found applications in such diverse areas as Massaro at the Perceptual Science Laboratory at University of California at Santa Cruz (PSL-UCSC) have been enhancing the precision of visible speech generated by Baldi, a synthesised talking tool [113]. Baldi has been utilised efficiently to offer curricular lessons, also to train vocabulary to profoundly deaf student at the TuckerMaxon Oral School in Portland Oregon [6, 118]. The PSL-UCSC coarticulation algorithm has been effectively utilised in American English and Mexican Span-

ish [113], and French [68]. Recently, Baldi presently talks Italian [28] and Arabic [132].

Some of applications include talking faces for teaching English as a second language such as Massaro's Baldi system [115]. Cohen and Massaro introduced a novel visual speech animation coarticulatory control strategy, utilising dominance and blending functions [117]. In addition, Baldi has control on the paralinguistic information therefore facial expressions and gestures can be displayed in the face, thus happiness, sadness and anger can be showed [119]. The system presents text-to-visible speech animation and alignment with normal speech. Baldi can be represented in different configurations, for instance, the skin can be made transparent thus the inside of the mouth (i.e tongue) can be seen, also the head can be turned around to be viewed from the side or back [28]. Massaro et al. [119] augmented Baldi with a body, to develop communication through gesture.

The recently released, publicly accessible resources provide articulation videos, which can be appropriate when learning/teaching the pronunciation of sounds absent in the learners local. There are some studies that have introduced improved pronunciation of non-native sounds through training with animated faces revealing the operations of internal speech articulators [44,114,191]. Computer assisted auditory-visual language learning enhances user engagement when compared to auditory alone. Karaoke, also called KTV, is a main pastime among Chinese nation, with various KTV clubs found in great cities in China. A karaoke-like feature has been added to Engkoo — a specialised search engine located at www.engkoo.cn — prepared for Chinese speakers learning English as a foreign language [155, 188], this allows English

learners to train the pronunciation online by mimicking a photo-realistic talking face lip-synchronously inside an inquiry and revelation environment. The KTVfunction is showed as videos produced from a large set of example sentences obtained from the web. Videos can easily be lunched by select the desired sentence. The videos show the sentence on the screen, while a model speaker utters it aloud, teaching the users how to pronounce the words, as appeared in Figure 1.2. Wang and Soong [189] motivation is on producing a photo-realistic, lip-synced talking face as a language assistant, web-based and low cost language learning. Their long-term target is to produce a technology that can assist users everywhere, and anytime from detailed pronunciation practicing to conversational training.

One of the motivating applications of facial animation in medicine is computer assisted craniofacial surgery (CAS) [67]. Patients with facial deformities or paralysis are restricted in their capability to communicate with other persons, so that, the re-formation of aesthetic appearance and natural facial expressions remains the main interest of corrective surgery. Such applications need a tissue and muscle model of the face to assist the surgeons planning their surgery. Animation approaches allow the construction of real dynamic faces helpful in the planning of reconstructive forensic medicine and facial surgery.

Talking faces are valuable in applications of human-machine interaction. Ostermann and Weissenfeld [130] have shown that confidence and reliance of humans toward machines growing by 30% when communicating with a talking head instead of text only. So that, visual speech can catch the interest of a user, making the human-machine interface more attractive. Avatars, with visual speech synthesis, are

increasingly being utilised to communicate with humans on a different of electronic devices, for example mobile phones, computers, kiosks, PDAs, and game consoles. Avatars can be found across many fields, such as technical support and customer service, also in entertainment. Following is some of the many uses of avatars [199]:

- Providing instructions and advice to guide users through Web sites;
- Reading news and other information to users;
- Displaying personalised messages on public Web sites;
- Practicing users to achieve complex tasks;
- Representing digital helpers and automated operators for help desks, contact centres, and self-service;
- Acting as character roles in games;
- Producing new branding opportunities for communities;
- Catching users attention in advertisements and announcements;
- Attracting users attention in advertisements;

Figure 1.3 shows applications of human-machine interaction with talking faces.

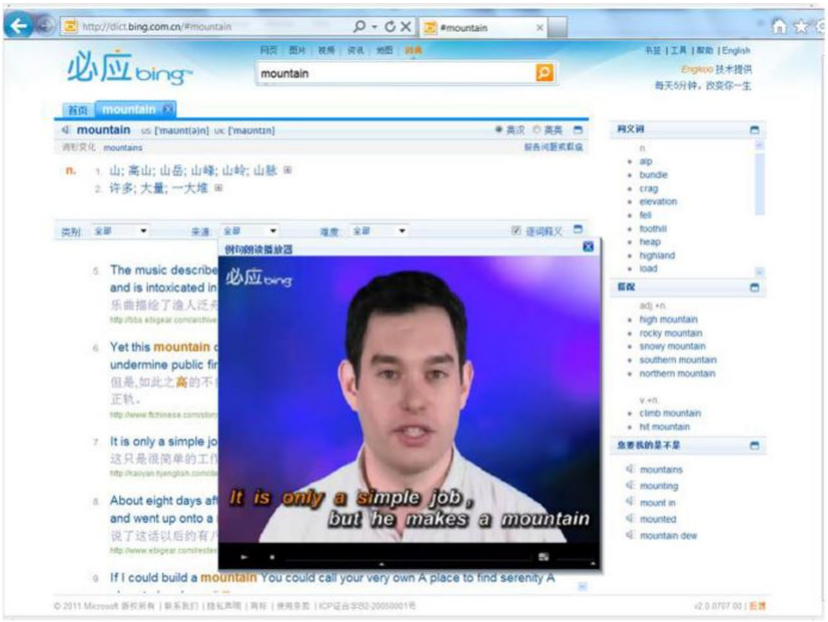


Figure 1.2. Screenshot of Karaoke-like talking face on Engkoo. This service is available at `http://www.engkoo.com` [189].

Figure 1.3. Applications of human-machine interaction with talking faces [185].

There is a large variety of multimedia on the Internet with the main

purpose of attracting human attention. MPEG-4 is an object-based multimedia compression standard, which permits to independently encode various visual objects in the scene. The visual objects might have a natural or synthetic content. The MPEG-4 standard expects that talking faces will play an important role in future customer service applications. MPEG-4 enables face animation over low bit rate communication channels. In particular, MPEG-4 facial animation parameters (FAPs) [91,99,203], where FAPs are popular for synthesising animation of human talking faces.

1.2 Research aims

Visual speech animation is an important aspect of synthesising a realistic talking head. Poor visual speech animation can be distracting, and confusing. It is known that there is a strong relationship between lip motions and speech. Mismatch between auditory and visual speech can change what the observer believes they heard [125]. Therefore, if the animation is not synchronised with the speech, the animations of a talking face cannot look realistic. One of the main challenge in visual speech synthesis is realism. The choice of approaches utilised for facial modelling and auditory-visual mapping greatly influences the amount of realism achieved.

This thesis focuses on generating realistic visual speech animation utilising a 2D shape and appearance models for facial modelling and a learning-based method to auditory-visual mapping. A 2D appearance facial modelling method [27] is used because it is a generative parametric model and commonly utilised to track and synthesise faces in video utterances. The model consists of two components: shape variation

model and appearance variation model. Therefore, using such models is attractive in speech animation because the geometry and the texture of the face are captured together.

The aim of this thesis is to improve visual realism of speech driven talking head by adopting a learning-based visual speech synthesis approach. Specifically the approach focusing on developing robust auditory-visual mapping. A novel principled method to learning a latent variable space of auditory and visual dynamic of speech is introduced. In contrast to previous methods the model is fully Bayesian, yielding the ability to estimate of the dimensionality and the structure of the latent space to be done automatically. The model can capture structure underlying high dimensional representations.

1.3 Main contributions

The main contributions in this thesis are the following:

- **A more accurate active appearance model (AAM) of talking faces:** We show that using larger dataset and more landmarks for building AAM produces more accurate model. Deena et al. [38] uses 56 markup points identified for each frame; 24 of them described the inner and outer mouth shape. In this thesis, we build a more accurate active appearance model, with 97 facial landmarks identified for each frame; 38 of them described the inner and outer mouth shape. Experiments are conducted to investigate the hypothesis of that increasing the number of landmark points can increase the accuracy. Quantitative evaluation demonstrates that using more landmark points around the mouth can give more accurate model and a smoother facial boundary can

be obtained using more landmark points for each frame, details are given in Chapter 4. In addition, experiments are conducted showing that building AAM on larger dataset improves the model accuracy, details are also given in Chapter 4.

- **Improving the accuracy of shared Gaussian process latent variable model:** A more accurate shared Gaussian process latent variable model (SGPLVM) is built on the AAM and relative spectral transform-perceptual linear prediction (RASTA-PLP) features which are extracted from video sequences. Previous method [38] used 184 images by selecting 4 random frames throughout the LIPS dataset, from each of the 45 sounds. An AAM was built using this number of images, then the remaining dataset was projected to AAM parameters. In this work, the AAM is built using larger dataset of around 6000 frames. In addition, the AAM is built with more facial landmarks identified for each frame. This is the first study to report that building an AAM using a larger dataset and a larger number of landmarks improves the accuracy of SGPLVM. Objective evaluation via reconstruction error is performed to compare our proposed approach against previously existing methods. The quantitative evaluation shows that the SGPLVM model using a larger dataset and more landmark points for each frame to build the AAM gives better results, with a full description is given in Chapter 4.

- **First application of manifold relevance determination model for visual speech synthesis:** A novel model for visual speech synthesis is introduced in order to produce more accurate

coarticulation, namely manifold relevance determination model (MRD), which explicitly models the non-linearities in auditory-visual mapping. MRD has not been used previously for generating videos of talking faces from audio features. In contrast to previous methods the model is fully Bayesian, allowing the automatic estimation of the dimensionality and the structure of the latent spaces. The model is able to capture the structure of data with extremely high dimensionality. Accurate visual features can be inferred directly from the trained model by sampling from the discovered latent points. The accuracy of generating videos of talking faces using MRD instead of SGPLVM has been improved. Statistical evaluation of inferred visual features against ground truth data is obtained and compared with the current state-of-the-art visual speech synthesis approach. The results show a noteworthy difference between the errors obtained from the MRD and SGPLVM methods, the analysis and learning of manifold relevance determination is described fully in Chapter 5.

- **Facilitating the performance of MRD by utilising incremental eigenmodels:** To produce realistic videos of talking faces, a generative model of the face that captures both the shape and texture variation is used for training MRD. Moreover, a multiple eigenvector adding algorithm [76, 77] is used, thus allowing for incremental updating of data models. This approach opens a way of training these kind of models on computers with limitted resources, for example mobile computing. Moreover, the incremental eigenmodels method is appropriate for real-world applications where the online learning from real-time dataset is needed.

The quantitative results demonstrate that MRD using incremental AAMs provides only slightly less accurate results than using batch methods. The proposed methods compare favourably against several other state-of-the-art methods for visual speech synthesis, which is shown in quantitative and qualitative evaluation. Full details are given in Chapter 6.

1.4 Thesis structure

Chapter 2 reviews background on the human speech production process, and a description of a phenomenon of coarticulation, followed by methods that have been used to generate talking head including Three-dimensions (3D) head model and image based facial animation. Then, visual speech animation methods and synthesiser's input requirements are presented. A review of deep neural networks approaches for visual animation and evaluation methods of synthesis methods for talking faces are described.

Chapter 3 reviews approaches for facial parameterisation and auditory speech feature extraction. Then, details of visual and audio processing are described. Finally, the synchronisation between audio and visual speech parameters is presented.

Chapter 4 different latent variable models are presented using the probabilistic graphical model structure. Moreover, shared linear dynamical system (SLDS) and a SGPLVM to model audio and visual features of a talking face are presented. The SGPLVM is then applied to predict visual from auditory features and a comprehensive evaluation of its performance is dealt with. A more accurate active appearance model was built using more landmark points for each frame and larger dataset. Experiments are conducted to compute the performance accuracy of increasing landmarks around the mouth shape. In addition, experiments are performed to investigate that building an AAM on larger dataset can improve the model accuracy.

Chapter 5 deals with MRD method which is used to represent

auditory and visual signal as a set of factorised latent spaces. Training and inference for MRD are described. Quantitative evaluation of the quality of visual speech of two different datasets are presented for MRD and SGPLVM models.

Chapter 6 a generative model of the face which captures the shape and texture variation and a method for adding eigenspaces is described. The MRD method using batch and incremental approaches for visual representation are compared. Experiments are conducted to initialise the latent space. Experiments are performed to compare two normalisation method utilising Mean-centering AAM parameters and a z-score normalisation. Then, quantitative and qualitative evaluation of the quality of visual speech of the talking head results are presented. MRD methods are evaluated against multiple methods of visual speech synthesis.

Chapter 7 concludes the work described in this thesis with a summary and indicates directions for future work.

1.5 Publications

Below is a publication based on the novel contribution related to Chapter 5

- S. Dawood, Y. Hicks, and D. Marshall, "Speech-Driven Facial Animation Using Manifold Relevance Determination." *In European Conference on Computer Vision.*, Springer International Publishing, 2016.

In addition a journal paper is in preparation

- S. Dawood, Y. Hicks, and D. Marshall, "The Manifold Relevance Determination for Audio Visual Mapping Based on Appearance Facial Model."

Chapter 2

LITERATURE REVIEW

This chapter provides background on human speech production process, followed by a review of approaches that have been utilised to synthesise visual speech including three-dimensions (3D) head model and image based facial animation, and a phenomenon of coarticulation. An input requirement to the synthesis system is then presented, which can be achieved by means of sound/viseme sequence or plain text (namely, text-driven systems), or by means of an audio speech signal (namely, speech-driven systems). In addition, different deep neural network (DNN) techniques in the area of facial animation have been reviewed. A review of different datasets for visual speech synthesis and evaluation approaches for talking head have also been presented.

2.1 Human speech generation and the multimodal nature of speech

Human speech is generated through the increase and decrease of air pressure over the larynx and vocal tract out of the mouth or oral cavity and nose or nasal cavity by the action of the diaphragm, as shown in Figure 2.1. The sounds are produced because of the interaction of the different cavities, associated with the vibration of the vocal cords.

To produce speech, the different fragments of the larynx and mouth should be in certain positions. Voiced sounds are produced when the vocal chords are tightened; the air flow becomes restricted causing them to vibrate. On the contrary, as the vocal chords are relaxed this results in voiceless sounds. As an example, the long sound of phoneme v in the word "vet" causes the vocal cords to vibrate, whereas the phoneme f in the word "fish" does not. The portions of the vocal tract that produce voices are called articulators [135].

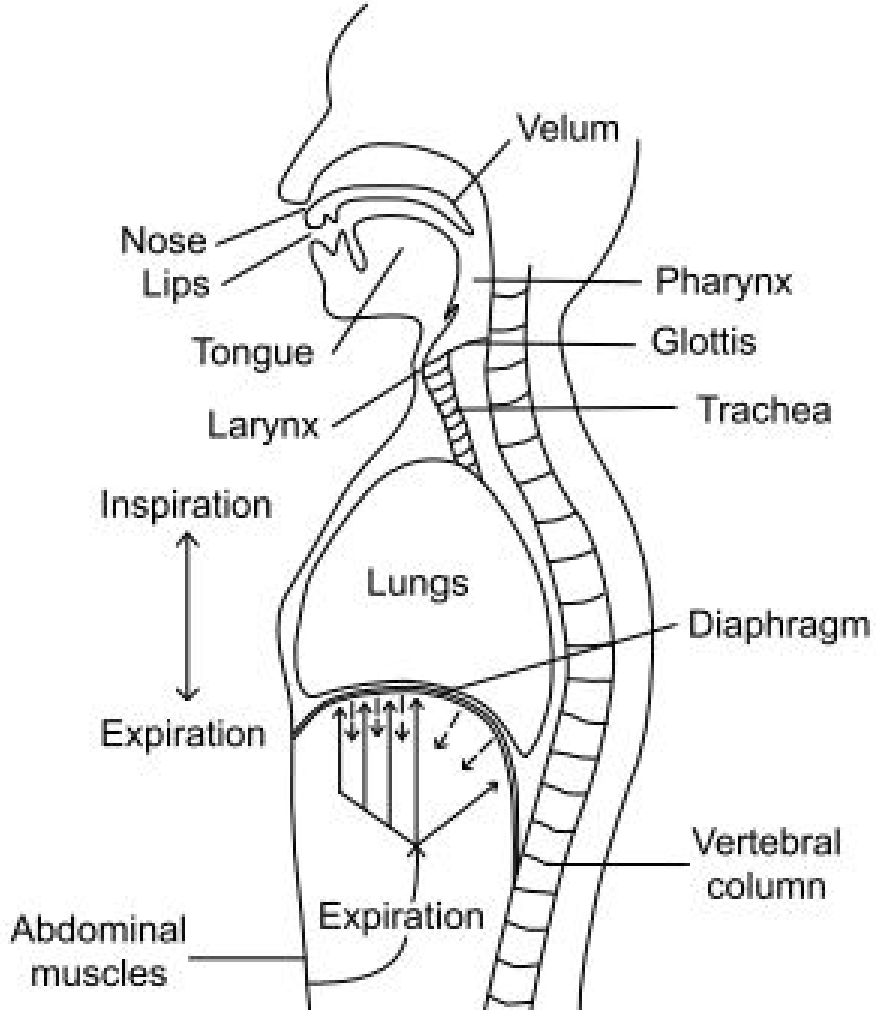


Figure 2.1. Sketch of the articulators utilised in human speech production Parke and Waters [135].

Some parts of the human speech generation system, for example the cheeks and the lips, are clearly visible when looking at a talker's face. Moreover, other articulators for example the teeth and the tongue are sometimes visible, depending on the speech sound that is being generated. The appearance of the visible articulators is highly correlated with the uttered sequence of speech sounds, because the auditory speech signal is the result of the movement of the articulators of the human speech production system. Thus, two distinct data streams are received when looking at the face of somebody who is speaking: visual speech data containing a variation of the talkers visible articulators and acoustic speech data containing a sequence of speech sounds. Auditory-visual speech data is the combination of these two data streams. Most developments in speech-based automatic recognition depend on auditory speech as the sole input data, ignoring its visual counterpart. However, the combination of audio and visual modalities have been shown to be suitable for improving recognition accuracy and robustness in both humans and machines than can be obtained with a single modality [21, 141, 142]. This is because the complementary nature of the auditory and visual modalities. For example, some sounds, such as "n" and "m" that are confusable by ear are simply distinguishable by eye.

Although the acoustic speech mode is often regarded as the primary communication channel, receiving also the visual signal helps to better understand the message [116, 179]. Erber [56], and Sumby and Pollack [161] have shown an effective improvement in signal to noise ratio (SNR) when the acoustic speech is corrupted by noise. Expressions of emotion and extra metacognitive information are also utilised by the speaker to emphasise the linguistic data transferred by speech, [15,162].

Visual prosody is utilised to assign stress or to add an emotion to the message [163]. The decoding of the audio data is influenced by the captured visual data and vice versa. This observation explains the existence of auditory visual speech perception effect called McGurk [125]. For example, this occurs when an auditory presenting syllable /*ba*/ is dubbed onto a video of a mouth /*da*/, the audio perception is altered by the visual information and the observers heard /*ga*/ [116]. Another effect, known as visual capture. This, happens when viewers who are perceiving unmatched audio-visual speech reported hearing the visually given syllable instead of the acoustic syllable that was given [1].

2.2 Phonemes and visemes

The phoneme is a basic unit of speech and viseme is a basic visual segment that corresponds to the phoneme. Correlating visemes to sounds is defined as visemetrics. Visemes are a unique posture of the mouth parts, Figure 2.2 shows some key viseme postures [194]. The combination of some postures can be derived from their audible equivalents. The transition between adjacent phonemes is called diphone, whereas triphones are a collection of three phonemes. They are utilised in speech analysis and synthesis in which acoustic models are necessary to be labelled to words in a language. Coarticulation is defined by such labelling which spans two sounds for diphones and three for triphones. If the number of sounds in a language is N, then the potential total number of diphones can be N^2 and N^3 for triphones [135]. The same concepts can be applied to visual labelling.

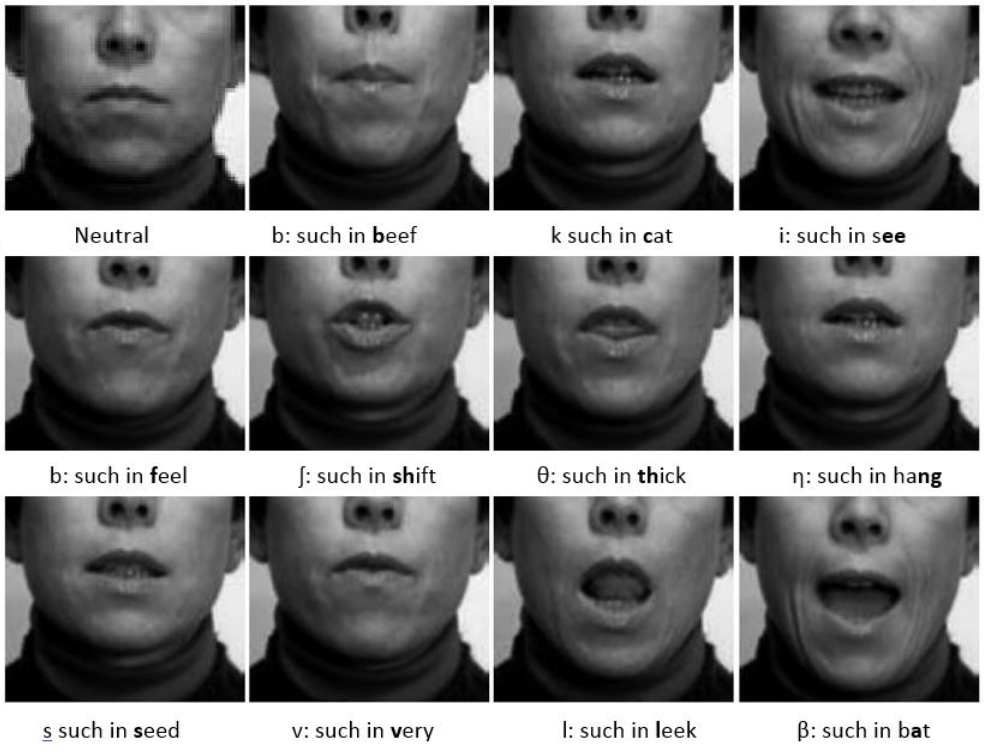


Figure 2.2. American English visemes Parke and Waters [135].

2.3 Coarticulation

A speech utterance can be transcribed into a group of functional segments known as phonemes. Phoneme is the basic segment of speech and viseme is the corresponding visual segment. There are 44 phonemes in the British English Example Pronunciation Dictionary (BEEP) phone set [169], which can be grouped into 14 visemes according to MPEG-4 standard [167]. The visual appearance of a phoneme depends on the phonemes which come before and after it, this phenomenon is known as *coarticulation*. Speech is a continuous process and in order to form the sounds of speech the articulators need to be physically moved to their

sequence of required positions. Coarticulation refers to changes in the articulation of a speech unit according to preceding or backward coarticulation and upcoming units or forward coarticulation. An example of preceding coarticulation is a difference in articulation of a final consonant in a word relying on the preceding vowel, e.g. boot vs beet. An example of upcoming coarticulation is the anticipatory lip rounding at the beginning of the word "stew" [22]. The speech articulators need to transition from the current positions to the following configuration, so that there is a blurring at the boundaries of the phonetic segments. For example, the utterance of the sound /ih/ in milk and sit. In the first case, the phoneme /m/ preceding the /ih/ would cause a lip-rounding through the utterance of /ih/, which in turn has to transition to the semi-vowel /l/ before reaching the sound /k/. The shape of the mouth through the utterance of /ih/ is more elongated vertically. In the case of the word sit, the sound /ih/ is encapsulated between /s/ and /t/, making the occurrence of /ih/ of more elongated horizontally and shorter duration. In both cases, the visual appearance would change, because of the preceding and upcoming sounds that occur, so making speech production a highly context-bound process [39].

2.4 Facial animation

Facial animation involves controlling a face model utilising geometric manipulations or image manipulations. A review of facial modelling and animation techniques is presented in this section.

2.4.1 3D Head model and image based facial animation

Traditional facial animation methods are graphics-based, where points on the face are represented as vertices in three-dimensions. To form a connected mesh, the skin is approximated by connecting the vertices. Using time-varying parameters, these mesh vertices are manipulated which influence the mesh geometry either directly, or utilising a physically-based method [135]. The pioneering work of Parke [136] was the first to build a three-dimensional geometric model of a human face using a polygon mesh, by painting the polygon topology onto a human's face, then 3D coordinates of the vertices were reconstructed by measuring their distances in multiple photographs. Using key shape interpolation, these head models were then animated, afterward the face was manipulated through the use of parameters that controlled interpolation, translation, rotation and scaling of different facial features [134, 136, 137]. Figure 2.3 shows the polygon topology and two type of rendering faces for Parke's [134] model which is an arbitrary network instead of a regular grid. The polygons are sized and positioned to correspond to the features of the face.

Cohen and Massaro's Baldi [22, 23, 113] is a descendant of Parke's software and his specific 3-D talking face [133]. The resolution of the model was increased, modified and additional control parameters were added, asymmetric facial movements were allowed, a complex tongue was trained, a coarticulation algorithm was implemented, and controls for paralinguistic data were added and have an effect in the face. Baldi [131] can either be controlled by text-to-speech (TTS) synthesis or aligned with natural speech to predict bimodal (audio/visual) speech. The control parameters move vertices (and from these vertices,

the neighbouring polygons were formed) on the face using geometric functions such as rotation (e.g. jaw rotation which determines the mouth opening) or translation (e.g., mouth widening, upper and lower lip height). Other parameters controlled scaling and interpolating various face sub-areas. Interpolation was used in the face shape parameters for example cheek, forehead shape, neck, and smiling.

Another approach for 3D animation utilises motion capture to map recorded movement onto a character [204]. By utilising reflectance markers located on the actor, which are tracked by cameras, feature points on the face are recorded. Ma et al. [112] achieve a small set of facial expressions utilising a real-time 3D scanning technique to record a high-resolution appearance and geometry of an actor. A group of motion capture markers was placed on the face to track large scale deformations. These large scale deformations were mapped to the deformations at finer scales. This relation was represented in the form of deformation-driven polynomial displacement maps, encoding variations in medium-scale and fine-scale displacements.

An anatomy-based method mentioned in a recent review paper discusses a strategy for defining the 3D polygon mesh [124]. The authors mention that in this approach the deformations of the polygon mesh cannot be directly parameterised. Alternatively, the facial gestures are simulated by modelling the anatomy of the face: muscles, bones and skin. Sifakis et al. [156] presented a physics-based method for generating animations of words and phrases from text and auditory input based on the analysis of motion captured speech samples. A high resolution, anatomically accurate flesh and muscle model was built for a specific subject. After that a motion captured training set of

speech samples was translated into muscle activation signals, and segment those into time segments corresponding to individual phonemes. Then, novel words and phrases were synthesised using these samples. Physics based approaches model face movement depending on simulating the effects of muscle interaction, allowing anatomically plausible motion. However constructing such models needs noteworthy effort. Waters [193] proposed a physical model which can predict a persons facial expressions given a neutral 3D range scan. Human emotions such as fear, anger, surprise, disgust, happiness, and joy were animated utilising vector based linear and orbicularis oris muscles using the facial action coding system (FACS) [53].

Image-based methods have been developed to generate a photorealistic human face model and can produce animated sequences with a high degree of both static and dynamic realism. Photorealism denotes that the novel synthesised images show the accurate dynamics, motion, and coarticulation effects. In Theobald [169], image-based animation approaches are classified into two major groups, namely morphing approach and a concatenative approach. The morphing method is equivalent to traditional key-frame animation, in which images are selected to represent mouth shapes or specific expressions and in-between frames created utilising the morph. In a concatenative approach, the face in each frame of the dataset is labeled with a set of parameters, then these parameters are utilised to select frames from the dataset closest to the required frame. The pioneering work of Bregler et al. [13] is the Video Rewrite system where key-frame-based interpolation techniques based on morphing between 2-D key-frame images were developed. The most frequently utilised key-frame set is visemes, which form a set of images

spanning many of mouth shapes. The transitions from one viseme to other viseme can be computed and interpolated automatically, utilising morphing approaches. The speech and video were modelled together by breaking down the recorded video corpus into a group of smaller audio-visual basis units, where each of the units is a triphone segment. Photorealism in Video Rewrite is achieved by only utilising recorded sequences to synthesis the new video. Videorealism is realised by utilising triphone contexts to model coarticulation. To deal with all the possible triphone contexts, the system needs a library with tens and perhaps thousands of subsequences, which appears to be too redundant. To decrease the number of frames to be stored, other approaches morph between keyframes [58] of visemes, which are the visual analogue of sounds. A developed statistical analysis of video footage has yielded other essential mouth shapes that can be encoded as a vector space of warp-fields and textures [57]. Cosker et al. [30] proposed a hierarchical image based facial model capable of producing coarticulted mouth animation given speech input. A novel modelling and synthesis algorithm is incorporated for learning and producing coarticulated mouth animation.

Producing synthesised talking heads that look like real humans is challenging. The existing methods to talking heads utilised image-based models [106, 123, 186, 190]. Wang and Soong proposed a system for producing photo-realistic talking head from video clip, and focus on the articulator movements rendering such as lips, teeth, and tongue. A hidden Markov model (HMM) trajectory-guided, real image sample concatenation method to photo-realistic talking head animation is introduced by Wang and Soong [189]. They suggest to model and

synthesise the lip movement trajectory using statistical HMM model, the model is initialised with maximum likelihood training and refined under minimum generation error principle. Then, they utilised the trajectory-guided sample selection approach, in which the HMM synthesised visual trajectory is utilised as guidance to select the natural mouth images from the recorded collection for a photo-realistic head rendering as illustrated in Figure 2.4. This system won the first place in the auditory-visual contest in the LIPS2009 Challenge having been perceptually evaluated by human subjects.

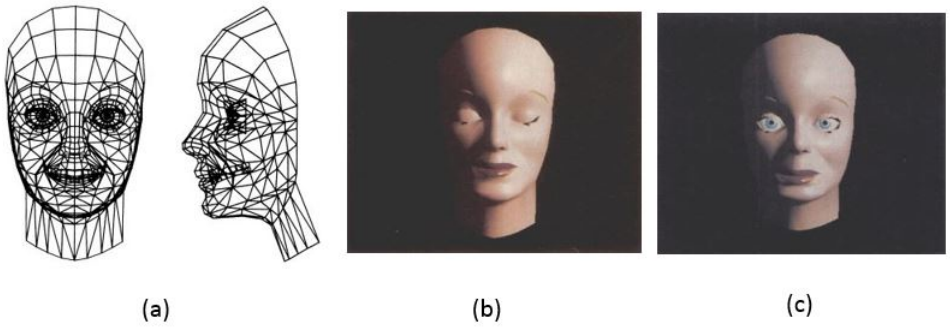


Figure 2.3. The topology and the rendering images (a)Frontal and side views of the topology, (b) The generated face with the eyelids closed and neutral expression, (c) A smaller chin face Parke [134].

Figure 2.4. HMM trajectory-guided sample selection method. The top-line images are the HMM generated visual trajectories. The bottom colored images are real sample image candidates in which the optimal lips sequence (red arrow path) is selected using Viterbi decoding (Wang and Soong) [189].

2.4.2 Visual speech animation

Visual speech animation can be considered as visual motions of the face (specifically the mouth part) when humans are speaking. Synthesising realistic visual speech animations corresponding to novel text or audio speech input has been a hard task for decades, because the large number of phonemes in the human languages, such as English, also the phenomena of coarticulation in speech that complicates the mappings between phonemes (or auditory speech signals) and visual speech motions. Previous studies in visual speech animation synthesis can be roughly categorised into two classes: viseme-driven methods and data-driven methods. Viseme-driven methods need animators to design key mouth shapes for phonemes (visemes) in order to synthesis novel visual speech. In contrast, data-driven methods require a pre-recorded facial motion dataset for synthesis purposes.

Viseme-driven methods

Viseme-driven methods typically require the creation of key mouth shapes (visemes) for each phonetic realisation, and then smoothing functions or coarticulation rules are utilised to infer new speech animations. There is no agreement as to which phonemes are grouped to form each viseme, and how many visemes are optimum. Many researchers give different phoneme to viseme groupings [20, 126], the main cause is the lack of standardisation for visemes, and also because there are different phonetic alphabets for different languages. Some attempts have been made to utilise machine learning methods to identify visemes objectively [83]. However, these have yet to produce a generic

set of visemes. Most viseme-based methods assume a many-to-one mapping between phonemes and visemes, and utilise an approximate set of mouth shapes; as an example, Tekalp and Ostermann [167] utilised 14 visemes. Cohen and Massaro [22] introduce the coarticulation model based on Lofqvist's gestural theory of speech production [110] for synthesis speech animations. In their method, a viseme shape is defined utilising dominance and blending functions that are defined in terms of each facial measurement. The final mouth shapes are then determined using a weighted sum of dominance values. Dey et al. [45] implement talking head utilising viseme-driven speech animation. In their work, 15 visemes were utilised, and meshes were built utilising Facegen modelling software [88]. Tongue positions visemes were initially adapted using Lazalde's tongue models [103]. For a speech tutoring application, the head was integrated into a GUI. The talking head explains how to vocalise sounds, and displaying the proper mouth movements. Acoustic speech and phonetic labels with durations were generated from text. An animation sequence was generated by mapping each phoneme label to the corresponding viseme. Coarticulation was performed based on Cohen and Massaro's model, utilising a dominance function to represent the impact over time that a viseme has on an utterance [22].

Data-driven methods

Data-driven methods infer new talking heads by concatenating prerecorded facial motion database or building a statistical model that captures the mapping between auditory and visual parameters. There are two ways to deal with the constructed facial motion data, either by utilising learning-based methods in which statistical models for

facial motion control are trained from the data, or facial motion data is further organised and processed utilising sample-based methods. Lastly, given novel acoustic speech or text input, corresponding visual speech animations are synthesised by sampling from the trained statistical models, or recombining facial motion frames optimally chosen from the facial motion dataset.

Data-driven methods can be further classified in terms of the input utilised to produce facial animation as well as on the method utilised to achieve auditory-visual mapping. The input can be either text or speech features. Text input in a given language is driving the facial animation and in the synthesis stage, text-to-speech synthesis is involved, followed by a mapping between underlying phonemes and visual speech [91, 121, 187]. Speech-driven facial animation involves mapping the discrete phonemes or continuous speech parameters onto the face. Audio information can be represented utilising phonemes which are obtained by phonetically aligning auditory data to phonemes [87]. Speech-driven facial animation in the work of Deena et al. [38] refers both to phoneme-driven and continuous speech-driven facial animation.

Sample based approaches

Sample based methods concatenate visual speech units in a dataset, where the units may be variable length [111, 166] or fixed-length (e.g. phonemes, visemes, or words) [85, 122, 172]. To form the animation, a set of units are found by minimising a cost function, based on phonetic context and smoothness of concatenation. Bregler et al. [13] proposed the video rewrite approach for inferring 2D talking heads given novel

voice input, based on the collected triphone video units. This method models the coarticulation effect with triphone video units. Kshirsagar and Thalmann [97] present a syllable motion based method to infer novel speech animations. In their method, captured facial motions are segmented into syllable motions, and then new speech animations are achieved by concatenating syllable motion segments optimally. To allow natural smoothing between concatenated units and avoids discontinuities, it is important to have a sufficiently large database. To overcome these problems, statistical approaches are used to learn and then predict visual speech parameters from speech or phonetic context.

Learning-based approaches

Learning-based methods model speech coarticulations as implicit functions in statistical models. Xie and Liu [197] used a coupled HMM to model the auditory and visual speech separately with coupled states having a mapping on both data streams. Visual parameters are predicted from auditory parameters using a Baum-Welch auditory-visual inversion algorithm. Brand [12] utilised an entropy minimisation learning algorithm to learn a HMM-based facial control model from audio and visual training data, then full facial motions were predicted from novel audio speech. To synthesis facial configuration sequences a Viterbi algorithm is used through vocal HMMs to search for most likely facial state sequence. Lehn-Schioler et al. [105] utilised a linear dynamical system (LDS) to jointly model voice and video parameters. A Kalman filter was used during inference to synthesise the underlying states from voice data, and a linear mapping was then utilised to infer the visual parameters from the synthesised states. Variant of switching

linear dynamical systems (SLDS) was used in the work of [55], named deterministic process dynamical system (DPDS) to model video data while voice data was modelled utilising a HMM. The two models were joined by the phonemes, which represent the states of the HMM and the switching states of the switching linear dynamical systems. During inference, synthesised phonemes from the HMM were utilised to find the most likely video parameters utilising the DPDS. Lehn-Schioler et al. [105] and Englebienne et al. [55] methods only model carry-over coarticulation through use of autoregressive linear dynamics models because the state vector for the current frame is synthesised from that of the previous frame in the inference stage. In Deena et al's. method [36] shared Gaussian process latent variable model (SGPLVM) is proposed to model a mapping between facial motion and speech data. A shared latent space is computed by maximising the joint likelihood function of the auditory and visual features, utilising Gaussian kernels. During the inference stage, intermediate latent points are obtained from the auditory data, and then utilised to synthesise the corresponding visual data using the Gaussian process mapping, visual data is represented using active appearance models (AAMs) [25]. Chen et al [19] introduced a nonparametric switching state-space model, to account for multiple types of dynamics. This approach is an extension of the shared GPDM [36], where multiple shared GPDMs are indexed by switching states. Later, Deena et al [38] utilised switching shared Gaussian process dynamical model (SSGPDM) with a variable-order Markov model on phonemes for talking head synthesis. We adopt such a learning-based method in this thesis because it allows for a compact representation of facial data. More details on specific learning-based approaches

applied to visual speech animation and related to this thesis are presented in Chapter 4.

2.4.3 Input text and speech driven systems

To synthesise the target audio visual speech data, the sequence of phonemes that must be uttered by the computer system is required. The input data can be plain text, which is called text-to-speech synthesis systems. A sequence of visemes can describe speech instead of phonemes, because visemes are more suitable to describe visual speech data [62]. A many-to-one mapping from phonemes to visemes is the standard method that utilised, in this approach a multiple visually similar phonemes are mapped on the same viseme. However, a many-to-one phoneme-to-viseme mapping does not take visual coarticulation effects into account. Mattheyses et al. [122] introduced a novel approach to define a many-to-many phoneme-to-viseme mapping, and showed that a many-to-many relationship more accurately describe the visual speech data as compared to many-to-one viseme-based and phoneme-based speech labels. Some visual speech animation synthesis a novel visual speech data based on an audio speech data that is given as input to the system. These systems estimate the target facial data based on features extracted from the audio input data. To obtain appropriate visual features, a training set is utilised to train a statistical model on the auditory speech features and their corresponding visual features. In the synthesis stage, this model is utilised to infer the target visual features giving a novel audio speech data. The synthesised visual features can then be utilised to drive the visual speech. Different methods have been utilised to represent visual speech and audio features for vi-

sual speech animation driven by speech. In addition, diverse techniques have been suggested to learn the mapping between the audio and visual features, such as Gaussian mixture models [206], hidden Markov model [11,12,30], regression techniques [84], an artificial neural network (ANN) [2], DPDS [54] and SGPLVMs [37].

More relevant literature relating to visual speech synthesis are given in the following Chapters.

2.5 Deep Neural Networks

Recently, deep neural networks have been successfully applied to facial animation. The long short-term memory (LSTM) is an extension of the recurrent neural network (RNN) architecture. Conventional RNNs are only able to utilise previous context data. However, modelling speech is highly related with preceding and succeeding speech contexts. Such that, bidirectional RNNs can access both the past and future contexts with two separate hidden layers. Pham et al. [140] introduced a regression framework based on LSTM RNN to determine rotation and activation parameters of a 3D blendshape face model [17] from a collection of audio features, for real-time facial animation. A set of acoustic features to capture contextual progression and emotional intensity of the speech were extracted from input audio. The blendshape model of [17] was used for animation. It can represent different emotional states, such as happiness, sadness, etc., without explicitly specifying them. Fan et al. [59] proposed a deep BLSTM (DBLSTM) method in modelling nonlinear mapping between auditory and visual data streams for photo-realistic talking face, which showed some enhancements over HMMs utilising a small database. A deep BLSTM RNN can be con-

structed by stacking multiple BLSTM RNN hidden layers. The output sequence of one layer is utilised as the input sequence of the following layer. In their work, the auditory and visual data were converted into utterances of contextual labels and visual parameter, respectively. The deep BLSTM network was trained to learn the regression model between the auditory and visual utterances by minimising the generation errors. The input layer of the network was the labels sequence and the output layer was the AAM visual features sequence. In following work, Fan et al. [60] introduced a 2D image-based video-real speech animation. Using AAM learned from a group of facial images, the lower face area of the speaker was modelled. A deep neural network model was trained to learn an auditory to visual mapping. To enhance the realism of the presented talking face, the trajectory tiling approach was adopted to utilise the generated AAM trajectory as a guide to select a smooth and natural video sequence from the recorded audio-visual database. Inspired by the encouraging performance of low level descriptors (LLD) in speech emotion recognition, Lan et al. [99] investigated LLD based DBLSTM bottleneck feature for speech driven talking avatar that considers the contextual auditory feature correlations and the textual information. The proposed method demonstrated some improvements over the conventional spectrum related features.

Recent studies have shown that utilising deep neural networks results in improved synthesis of head motion, especially when using BLSTM. Ding et al. [47] proposed a neural network method for speech-driven head motion synthesis to model a non-linear mapping from auditory speech to head motion. They showed that using a one-hidden-layer multi-layer perceptron (MLP) with Mel-frequency cepstral co-

efficient (MFCC) feature input enhanced the head motion synthesis accuracy significantly. Their study showed that feed-forward neural networks (FFNNs) have dramatically outperformed a HMM in head motion prediction. A bidirectional LSTM (BLSTM) was used in their recent work [46] and showed that the BLSTM networks outperform the FFNNs because of their capability of learning long-range speech dynamics. Haag and Shimodaira [75] presented a novel method which combines stacked bottleneck features and a BLSTM network, to model context and expressive variability for the task of expressive speech-driven head motion synthesis. The proposed method outperforms the conventional feed-forward DNNs. Sadoughi and Busso [151] built a model to learn the distribution of head motions conditioned on speech prosodic features employing a conditional generative adversarial network (GAN) with BLSTM. The conditional GAN model showed improvement performance over some baseline systems.

Karras et al. [93] introduced a deep convolutional neural network (CNN) that learns a mapping from input auditory coefficients to the 3D vertex coordinates of a face model to generate an expressive 3D facial animation. Taylor et al. [165] introduced a framework utilising deep neural network to estimate AAM coefficients from input phonemes, which can be generalised to any input speaker and languages. Their method was a continuous deep learning sliding window predictor which means that the predictor can represent a complex non-linear regression between the input phonetic context and output visual representation of continuous speech that includes coarticulation effects. Their sliding window predictor can be viewed as a variant of a convolutional deep learning architecture. Song et al. [158] presented a synchronised audi-

tory to talking video generation method using the recurrent adversarial network to model the temporal correlation between auditory features and face image features at the same time. In addition, a sample selection approach which decreases the training size by eliminating highly repeated samples without affecting performance. The limitation of the end-to-end training performance is that the auditory encoder network needs the off-the-shelf auditory extractor.

The main limitation for using a deep neural network for auditory-visual mapping is a lack of publicly available auditory-visual speech corpus which is either of restricted vocabulary or size [24,139]. Employing modern machine learning techniques such as deep learning required a sufficiently comprehensive dataset, because such approaches are generally highly under constrained. In addition, training deep learning is computationally very expensive and needs a GPU with high performance.

2.6 Datasets

For visual-speech synthesis purposes, no standardised databases are available. Therefore, each visual-speech synthesis method is developed and assessed utilising its own visual-speech database. Recently, Taylor et al. [164] studied the problem of mapping from audio to visual speech to generate speech animation automatically from an auditory speech data. A sliding window DNN that learns a mapping from a window of audio parameters to a window of visual parameters from a large auditory-visual speech corpus is presented. The KB-2k corpus is used from [166] which is expected to be set for future release. It is a large auditory-visual speech corpus including a male actor speaking

about 2500 phonetically balanced TIMIT [65] utterances in a neutral style. The actor was requested to speak in a neutral talking style and maintain, with a fixed pose.

Ding et al. [46] used the MNGU0 electromagnetic midsagittal articulography (EMA) dataset in their experiments for training of BLSTM-RNNs. The corpus consists of 1263 sequences recorded from a single talker. Pham et al. [140] utilise the ryerson auditory-visual dataset of emotional speech and song (RAVDESS) [109] for training and evaluation DNN. The database includes 24 professional actors talking and singing with several emotions such as neutral, happy, calm, angry, sad, fearful, surprised and disgust. Video utterances of the 20 actors are used for training, with about 250,000 frames, and the model is evaluated on the data of four actors.

Vougioukas et al. [181] suggested a temporal GAN, capable of synthesising a video of a talking face from an auditory data and a single still image. Evaluation was performed on the GRID [24] and TCD TIMIT [79] datasets. GRID an auditory visual dataset consists of 1000 sentences per talker such as "place blue at F 9 now" spoken by each of 34 speakers giving a total collection size of 34,000. The sentences structure is drawn from the following grammar "command:4, color:4, preposition:4, letter:25, digit:10, adverb:4". GRID had been collected to support the utilise of common material in speech perception and automatic speech recognition research's. TCD-TIMIT consists of auditory and video footage of 62 talkers reading 6913 phonetically rich sentences from the TIMIT corpus.

Kuhnke [98] demonstrated a structure to construct a visual speech synthesis method from 3D performance capture data utilising a pub-

licly available 3D database [61]. BIWI 3D audio-visual dataset [61] of Affective Communication includes 14 different subjects, citing 40 utterences once neutral and once emotional. The 3D mesh utterances are registered at 25 fps and phoneme labels of the auditory utterances provided with the corpus.

In 2008, the LIPS visual speech synthesis challenge was organised to evaluate and compare different auditory-visual speech synthesis approaches utilising the same original speech data [171]. An English visual speech dataset appropriate for auditory-visual speech synthesis was released. A great part of the work described in this thesis employed the LIPS 2008 dataset [171]. This public database has 278 video files with corresponding audio data, each being one English utterance from the phonetically-balanced Messiah corpus [169] spoken by a native English female speaker with neutral emotion. Another dataset closer to natural speech is used in this work namely DemocracyNow! corpus (DEMNOW dataset) [54] consists of 803 utterances for a total duration of 1h 7m 29.20s featuring a female American anchor reading broadcast news items. A more detailed description of the datasets used in this work is described in Chapter 3.

2.7 Evaluation methods of synthesis methods for talking faces

The quality of a visual speech synthesiser can be determined utilising various methods, which can be categorised as either objective and subjective evaluation methods. Objective methods include determining the error or correlation between real and synthetic visual features [29, 39, 169] in addition to comparing the evolution of their trajectories over time. Error measures such as the sum of absolute dif-

ferences (SAD), the root mean squared error (RMSE) or mean squared error (MSE) (the difference between the two-error measure is in units), the maximum absolute error (MAE) and average mean squared error (AMSE), in addition to the correlation metrics such as average correlation coefficient (ACC), give a measure of the static comparison between real and synthetic visual parameters. Whereas comparison between real and synthetic visual feature trajectories provide the dynamic correlation of the two features. The RMSE can be utilised as a performance evaluation, either in terms of the pixel values in the synthesised shape-free texture images [48] or the synthesised landmarks [3].

Objective quality evaluations are easy to achieve and are less time-demanding in comparison to other evaluation methods because the numeric quality metric can be determined directly from the visual speech data. The main disadvantage of utilising purely objective measures is that it is difficult to compute the overall realism, naturalness and intelligibility of the synthetic output in comparing features of original and synthetic visual speech. These measures cannot consider cognitive issues with respect to how the brain perceives a speaking head.

In Englebienne [54], it was shown that objective results do not necessarily correlate with subjective results since a synthesiser, which provided less appropriate objective results was found to be better based on perceptual judgments.

A subjective evaluation is the most frequently utilised method for assessing the naturalness, and the degree of viewer agreement of synthesised visual speech signals. A subjective evaluation of the visual speech quality includes a set of test subjects contributing in an evaluation experiment. In some cases of subjective evaluation the quality measure is

computed by collecting the test participants opinion utilising a Likert scale [45, 122]. To compare different synthesis methods, it is useful to use comparative mean opinion scores [38, 55, 120].

RMSE shows the parameter prediction errors and gives a more average-case statistical comparison. Therefore, in this research we utilise the AMSE or RMSE as the error measure between the true and estimated visual parameters as has been used by researchers such as Deena et al. [38], Gutierrez-Osuna et al. [73], Xie and Liu [197], and Terissi et al. [168]. MSE is used by the researchers such as Xie et al. [198], Wang et al. [183], and Taylor et al. [165]. We also use ACC in our experiments. ACC describes how the synthesised trajectory is correlated with the ground truth and it is used by the researchers such as Deena et al. [38], Gutierrez-Osuna et al. [73], Xie and Liu [197], and Terissi et al. [168]. Note that lower AMSE and higher ACC correspond to better performance.

The advantage of objective evaluations is that they can be calculated automatically, much less time consuming, and the tests are inexpensive to perform than subjective evaluations. However, humans certainly not repeat the same words exactly in the same way, therefore there might be differences in the parameters performing repetitions of the same words, so that a synthesiser will not be able to exactly resynthesise an occurrence of the words. The differences between the reference and generated parameters could appear in the difference viewed in normal speech production, which are not perceived by observers. On the contrary, the variation might be because of errors in the generated visual speech.

Therefore, the objective evaluations might be not enough to evaluate

our work. Thus we also use another useful evaluation method such as subjective evaluation.

2.8 Summary

A description of the human speech generation and the multimodal nature of speech and a review of approaches to synthesise visual speech including 3D head model and image based facial animation was presented. In addition, classes of visual speech animation including viseme-driven and data-driven approaches were described. The phenomenon of coarticulation has also been addressed, followed by a description of the input requirement to the synthesis system. Moreover, various DNN methods in the domain of visual speech synthesis were reviewed. A review of different databases for visual speech synthesis and evaluation methods of synthesis approaches for talking face was then presented.

The choice of approaches utilised for face modelling and auditory-visual mapping importantly effects the level of realism achieved. The aim of this work is to improve the state-of-the-art in the area of learning-based speech-driven talking face. Our focus is on the auditory-visual mapping, aiming to automatically model the relationship between auditory and visual features by jointly modelling auditory and visual speech utilising non-linear mappings within a Bayesian framework. In addition, 2D appearance-based face modelling and a construction eigenspace models are used in this work.

Chapter 3

DATA PRE-PROCESSING AND MODELLING

A dataset of auditory-visual recordings of a talking head that captures the different phonetic combinations in the language is required for visual speech synthesis. In this chapter, we begin by reviewing methods for facial parameterisation and audio speech feature extraction. Afterwards, details of visual and auditory processing are presented. The synchronisation between auditory and visual speech parameters is discussed as there is typically a mismatch between the auditory and visual frequency because of the requirement for an auditory window in which the audio speech data is stationary.

In modelling a talking face, we require some way of representing the visual speech, either in two or three dimensions. In this work, we focus on 2D appearance models, which can deliver high levels of realism [106, 185].

3.1 Data

Two datasets are utilised in this work: the first one is LIPS dataset [171] and the second is DEMNOW dataset [55], because they are most popu-

lar, large datasets and also we want to compare our proposed approach using different training databases. The LIPS corpora is phonetically balanced, while DEMNOW is closer to natural speech.

The LIPS dataset has 278 video files and all together 61,028 face images with corresponding auditory track, each being one English sentence from the Messiah corpus [169] spoken by a single British female with neutral emotion. The duration of each sentence is approximately 3-6 seconds. This dataset was made available for the LIPS2008 Visual Speech Synthesis Challenge. The auditory signal is in the form of WAV files with 16-bits/sample and a sampling frequency of 44.1 KHz. The phonetic annotation for each frame has also been made available for the corpora. The acoustic speech for each utterance has been aligned to the corresponding video. Utilising the British English Example Pronunciation (BEEP) phonetic dictionary, the corpus was phonetically aligned with the HTK audio speech recognition software [202], and the full contextual labels are generated with a phoneme dictionary which has 50 phonemes. The phonetic labels are specified in terms of the timings and this is processed to align the phonemes with each frame of the utterance. The dataset consists of video stream of size 576×720 sampled at a rate of 50 frames per second according to the PAL standard [90]. The corpus is available for download at the website: `http://www.lips2008.org/`.

The DEMNOW datasets (originally called the Democracy Now! news broadcast) [54] consists of 803 utterances featuring a female American news reader, which are available for download at the website `http://gwenn.dk/demnow/`. The sequences were extracted from broadcasts totalling one hour and seven minutes of video. Englebienne [54] manually extracted short video sequences of about $3 - 10$

seconds (eliminating inserts, telephone talks, etc.). According to the American NTSC standard [129] utilised for the recordings, a frame rate of 29.97 Hz was extracted. The DEMNOW dataset was phonetically aligned utilising HTK [202] with the CMU phonetic dictionary [195], because Amy Goodman, the news reader, is American.

3.2 Statistical models of shape and texture as a representation of the face

Statistical models of shape and texture have been widely utilised for recognition, tracking, and synthesis. These models have been used to recognise and track objects, including in video sequences [27]. The active appearance model (AAM) approach seeks to find the model parameters which produce a synthetic image as close as possible to the target one. To allow an easy mapping from audio features to visual features, some speech-driven 2D talking head synthesisers utilise a mathematical model to parameterise the visual speech data. For example, AAMs are ustilised by Cosker et al., Englebienne et al. and Deena et al. [31,36,55]. Our work will focus on a statistical method of AAM to obtain appropriate visual features. Kass et al [94] introduced Active Contour Models (snakes) which can be utilised to "snap" onto nearby edges through an energy-minimising spline function. Sirovich and Kirby [95,157] first applied principal component analysis (PCA) to efficiently find the lower dimensional model of human faces. It was argued that any face picture could be resynthesised as a weighted sum of a small collection of pictures which define the eigenfaces, and a mean image of the face. Turk and Pentland [178] utilised PCA to represent the intensity patterns in the images in terms of a set of eigenfaces. A set of basis vectors were com-

puted from a training data of faces and every training face decomposed into its principal components, which was applied to facial classificiation. PCA was used by Cootes et al. [26] to model the shape variation of the face example, then they used active shape model (ASM) search approach (also known as smart snakes) to fit a shape model to images. Craw and Cameron [32] used manually selected points to warp the input images to the mean shape, yielding shape-free images. Moghaddam and Pentland [127] used view-based eigenface models to describe different viewpoints, for the shape-free representations. Active appearance model (AAM) were developed by Cootes et al. [25], they combined the modelling of both shape and texture using PCA. The AAM method offers a convenient hybrid of model-based and image-based approaches in the form of a compact model that describes the variations of the shape and appearance of the face.

3.3 Appearance models

Appearance models [25, 27] are generated by combining a model of shape variations with a model of the appearance variations. In our work, we used a statistical approach of appearance models to build a face model. A training set of labeled images is required, where landmark points are marked on each face. The appearance models allows any face to be represented using a compact set of parameters, which can be used to synthesise the original face.

3.3.1 Statistical models of the shape

In order to begin building an appearance model, the construction of a point distribution model (PDM) is required [26]. Statistical models

of shape are used to represent objects in images. In a 2D image n landmark points, $\{(x_i, y_i)\}$ can be represented for a single shape as the $2n$ element vector,

$$\mathbf{x} = (x_1, ..., x_n, y_1, ..., y_n)^T \tag{3.1}$$

where x and y coordinates represent image landmarks, n is the number of landmarks in an image, s such vectors $\mathbf{x}_s$ can be generated for s training examples. Figure 3.1 shows a face image from LIPS dataset marked with points defining the main features. A PDM captures shape variation, for instance, in the face example, we would be interested in capturing variation which occurs in regions such as the mouth, for example, when it opens or closes. However, the variation which might occur due to pose changes in the face, for example, when a person moves their face out-of-plane is not appropriate. Therefore, before statistical analysis can be applied on these shapes, it is necessary that the shapes are in the common co-ordinate frame, the shape training set is firstly normalised with respect to pose using an alignment algorithm. The most popular method of aligning shapes in a same co-ordinate frame is Procrustes analysis [71]. In this method each shape is aligned such that the sum of distances of each shape to the mean ($F = \sum |\mathbf{x}_i - \bar{\mathbf{x}}|^2$) is minimised, where $\mathbf{x}_i$ is a shape vector, $\bar{\mathbf{x}}$ is the mean shape.

A simple iterative approach for minimising F, can be described as follows:

1. Each shape is translated so that its centre of gravity is at the origin.

2. One shape vector is chosen as an initial estimate of the mean and

scale so that $|\mathbf{x}| = 1$.

3. The first estimate is recorded as $\bar{\mathbf{x}}_0$.

4. Each vector is aligned with the current estimate $\bar{\mathbf{x}}_0$.

5. The mean is re-estimated from the aligned vectors.

6. If mean estimation has not converged then return to step 4. Convergence is declared if the mean does not change significantly between iterations.

By applying PCA on the shape vector set, variation can be approximated in terms of principle axis. The principle components are computed as follows

1. The mean of the distribution is calculated using

$$\bar{\mathbf{x}} = \frac{1}{s}\sum_{i=1}^{s}\mathbf{x}_i \tag{3.2}$$

2. The covariance of the data, is calculated using

$$\mathbf{S} = \frac{1}{s-1}\sum_{i=1}^{s}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{3.3}$$

3. The eigenvectors ϕ_i and eigenvalues λ_i of $\mathbf{S}$, ordered such that $\lambda_i \geq \lambda_{i+1}$, i.e. in descending order of energy are calculated. By performing PCA on the matrix $\mathbf{S}$, the data can be represented as a linear model

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_x\mathbf{b}_x \tag{3.4}$$

where $\mathbf{P}_x = (P_1, P_2, ..., P_t)$ is the matrix of the first t eigenvectors corresponding to the largest eigenvalues, and $\mathbf{b}_x = (b_1, b_2, ..., b_t)$ is a shape parameter represented as

$$\mathbf{b}_x = \mathbf{P}_x^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{3.5}$$

By varying the elements of $\mathbf{b}$, new shapes are allowed to be defined. Suitable limits of $\mp 3\sqrt{\lambda_i}$ (since most of population lies within three standard deviation) might be applied to the parameter b_i of $\mathbf{b}_x$ to ensure that synthesised shapes are similar to those present in the original training set.

The choice of t determines the number of principle components (or the proportion (e.g. 98%) of total energy retained in the model). To compute the number of eigenvectors required to retain a certain proportion of shape variation, t might be chosen such that

$$\sum_{i=1}^{t} \lambda_i \geq p_v \sum \lambda_i \tag{3.6}$$

where p_v represents the percentage of the total variation.

Figure 3.1. A labeled training image gives a shape free patch and a set of points.

3.3.2 Statistical models of texture

After PDM construction, the next stage in building an appearance model is to construct a statistical model of texture, which is either gray-scale pixel intensities or Red, Green and Blue (RGB) colour values. To achieve this numerous image warping methods exist, including Bookstein thin-plate spline warping [10] and piecewise affine warping [159]. To obtain shape-free patches (also called shape normalised textures), we use a triangulation piece-wise affine warping to warp each image texture from the landmark coordinates to the mean landmark shape, because this approach is less expensive to calculate. The procedure includes implementing Delaunay triangulation on image

landmark data in the original and target images, then affine warping corresponding triangles. In order to get a texture vector, $\mathbf{g}_i$, the intensity information from the shape-normalised image over the region covered by the mean shape needs to be sampled. Figure 3.2 summaries the building of a shape-free patch using piece-wise affine warping.

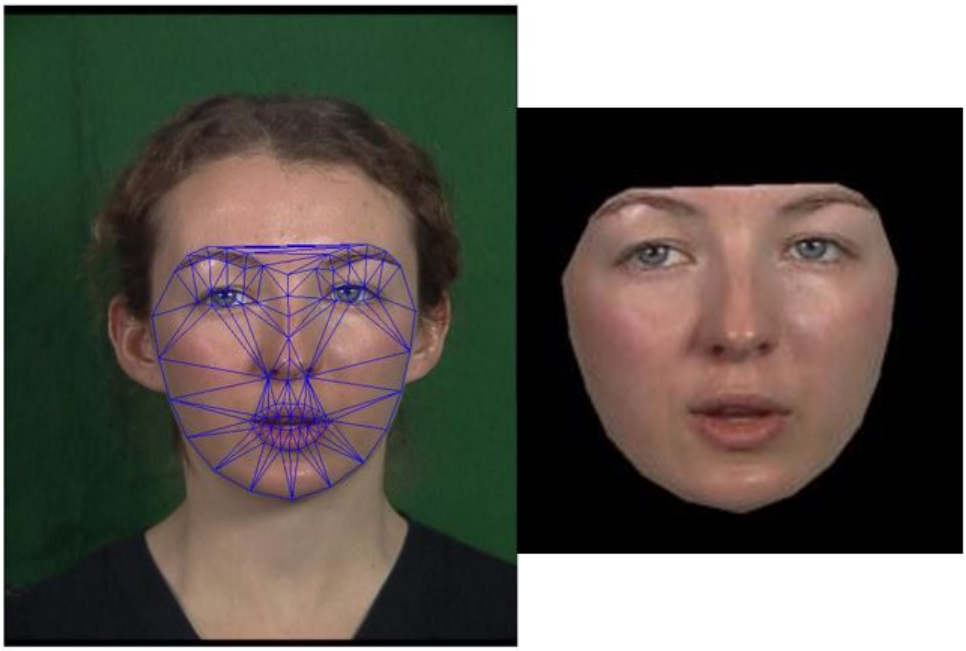


Figure 3.2. Building a shape-free patch. The texture in each triangle in the original shape (left) is warped to the corresponding position in the mean shape. Reiterating this for each triangle results in a shape-free patch of the training face (right).

The texture needs to be normalised to minimise the effect of global lighting variation, following the method described in [27]:

1. Select a texture from the texture collection as an initial estimate

of the mean texture $\bar{\mathbf{g}}$.

2. A scaling value $\alpha = \mathbf{g}_i \cdot \bar{\mathbf{g}}$ and an offset value $\beta = (\mathbf{g}_i \cdot 1)/n$ can be determined for each texture $\mathbf{g}_i$, where $i = 1, \cdots, N$ and n is the number of elements in $\mathbf{g}_i$, then $\mathbf{g}_i = (\mathbf{g}_i - \beta \cdot \mathbf{1})/\alpha$ can be recalculated.

3. Determine a new estimate for the mean $\bar{\mathbf{g}}$.

4. The steps 2 and 3 can be repeated until $\bar{\mathbf{g}}$ converges.

PCA is then applied to the data giving the linear model

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{3.7}$$

Where $\mathbf{P}_g$ is a set of orthogonal modes of variation, $\bar{\mathbf{g}}$ is the mean vector, and $\mathbf{b}_g$ is a texture parameter vector. Then, the shape and texture are separately projected to PCA parameters as follows

$$\mathbf{b}_x = \mathbf{P}_x^T(\mathbf{x} - \bar{\mathbf{x}}) \tag{3.8}$$

$$\mathbf{b}_g = \mathbf{P}_g^T(\mathbf{g} - \bar{\mathbf{g}}) \tag{3.9}$$

3.3.3 Combined shape and texture models

The shape and texture of any image can then be represented using the parameter vectors $\mathbf{b}_x$ and $\mathbf{b}_g$. However, the correlations might be exist between the shape and texture variations, so that, further PCA can be applied to the combined parameters. For each example in the training set, the concatenated vector $\mathbf{b}$ is generated as follows

$$\mathbf{b} = \begin{pmatrix} \mathbf{W}_x \mathbf{b}_x \\ \mathbf{b}_g \end{pmatrix} = \begin{pmatrix} \mathbf{W}_x \mathbf{P}_x^T (\mathbf{x} - \bar{\mathbf{x}}) \\ \mathbf{P}_g^T (\mathbf{g} - \bar{\mathbf{g}}) \end{pmatrix} \tag{3.10}$$

where the matrix $\mathbf{W}_x$ is a diagonal matrix of weights for each shape parameter, witch allows for the difference in units between the shape and texture models. PCA is then applied on these vectors to obtain a further model $\mathbf{b}$,

$$\mathbf{b} = \mathbf{Q}\mathbf{c} \tag{3.11}$$

where $\mathbf{Q}$ contains the first t eigenvectors, and $\mathbf{c}$ is an appearance parameters controlling both the shape and texture of the model. The matrix $\mathbf{Q}$ contain both shape and texture related elements, which is given by

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_x \\ \mathbf{Q}_g \end{pmatrix} \tag{3.12}$$

where the dimensions of $\mathbf{Q}_x, \mathbf{Q}_g$ are $(n, t), (m, t)$ respectively, t is the number of eigenvectors in $\mathbf{Q}$, n is the number of eigevectors in $\mathbf{P}_x$ and m is the number of eigenvectors in $\mathbf{P}_g$. Example of appearance model are shown in Figure 3.3, where appearance model trained on 50 sequences (totalling 5332 images) of LIPS dataset using adding eigenspace as described in Chapter 6. The first 5 mode are shown at ∓ 2 standard deviations from the mean.

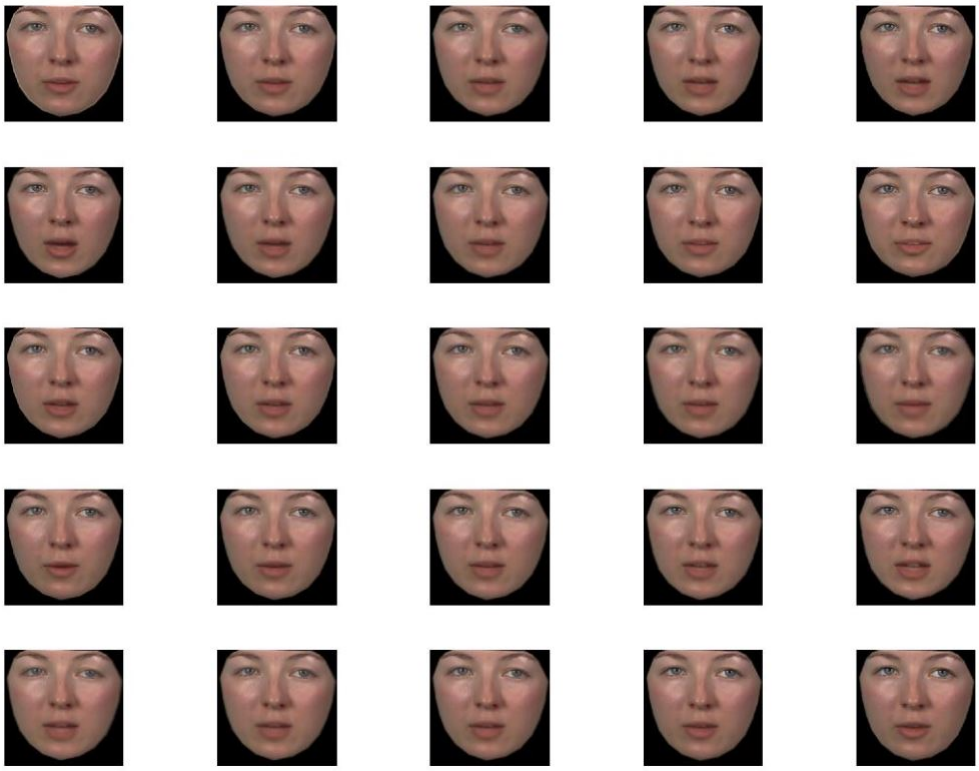


Figure 3.3. Effect of varying each of the first five parameters ($-2\sqrt{\lambda_i} \leq b_i \leq 2\sqrt{\lambda_i}$) of the appearance model.

3.3.4 Synthesis of an AAM

To synthesise an example image, the shape $\mathbf{x}$ and texture $\mathbf{g}$ are reconstructed as a function of AAM parameters $\mathbf{c}$ as follows:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_x \mathbf{W}_x^{-1} \mathbf{Q}_x \mathbf{c} \tag{3.13}$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{Q}_g \mathbf{c} \tag{3.14}$$

An output image can be synthesised for a given $\mathbf{c}$ by warping the texture $\mathbf{g}$ from the mean shape $\bar{\mathbf{x}}$ to the new shape $\mathbf{x}$.

3.4 Mean-centering AAM parameters and z-score normalisation

Pose differences between real and synthetic videos may lead to quantitative results that are not meaningful. So that the pose normalisation procedure is required to make the quantitative results more reliable since they correspond to visual features associated to auditory speech motions in a normalised coordinate space. The LIPS corpus contains many variabilities across utterances because of discrepancies in the orientation of the speaker with respect to the camera, so that visual normalisation needs to be performed to remove pose variations in the visual data. This is done by calculating the mean of the parameters for each mode of variation for a given collection, then subtracting that mean from the corresponding parameters, as demonstrated in equation 3.15,

$$\mathbf{c}_j = \mathbf{c}_j - \bar{\mathbf{c}_j} \tag{3.15}$$

where j is a given mode of variation and $\bar{\mathbf{c}_j}$ is the mean of the jth mode of variation across all frames in the utterance. The above is the mean-centering parameters, we also used in this thesis z-score normalisation as represented in equation 3.16,

$$\mathbf{c}_j = \frac{\mathbf{c}_j - \bar{\mathbf{c}_j}}{\sigma_j} \tag{3.16}$$

where σ_j refer to standard deviation of the jth mode of variation across all frames in the utterance.

3.5 Speech pre-processing

To obtain robust and uncorrelated continuous speech features, the continuous speech signal is first processed. Several approaches exist which allow this, including relative spectral transform-perceptual linear prediction (RASTA-PLP) [82], linear predictive coding (LPC) [42] and Mel-Cepstral analysis [40]. These approaches are utilised frequently in speech recognition [40]. Deena et al. [39] found that using RASTA-PLP for speech animation gives better results for LIPS dataset, therefore in this thesis RASTA-PLP analysis is utilised. A description of Fourier transform, RASTA-PLP, LPC, line spectral pairs/frequencies (LSPs/LSFs), and Mel-frequency cepstral coefficients (MFCCs) which is utilised in many of the audio speech processing approaches are given in this section. This is followed by the details of auditory processing and synchronisation with video in the next sections.

3.5.1 Windowing

For audio speech parameterisation, a sliding window is utilised to represent the part of the signal that is considered for analysis at each time point. Features are extracted from the speech signal within the rectangular time window of t_w which is known as a Dirichlet window [144] and the information outside of this window are discarded. However, spurious high-frequency components at the edges of the window is introduced. In order to avoid this problem, soft window boundaries can be utilised, such as the Hamming window [78] which is defined as:

$$s'_t = \left[0.54 - 0.46cos(\frac{2\pi(t-1)}{N-1})\right] s_t \tag{3.17}$$

The discontinuities at the edges are attenuated, because the Hamming window is based on the cosine function. Moreover, the window utilised is typically overlapping to capture dynamical properties of audio speech. The Hamming window is widely utilised in speech recognition, despite of the fact that the Hamming window has small discontinuities at the edges. Artifacts in the extracted features are introduced, however those artifacts have quite restricted impact. The size of the audio window utilised at any time point is denoted as window size while the length of the overlap that moving from one time point to the next is the hop size. Typically, a window size of $10 - 25$ ms is required by feature extraction approaches, where the speech data remains relatively stable. So that, audio speech is typically parameterised with a window of 25 ms-length frames with 10 ms overlapping between the windows, resulting in an audio speech processing frequency of 100 Hz. We utilised RASTA-PLP for the features extracted in this work and utilised a Hamming window of 50 ms, and sliding the window by 40 ms for every feature vector to compute 20 RASTA-PLP features.

3.5.2 Fourier transform

The Fourier transform [63] is described here because it is utilised in the perceptually-motivated approaches of speech parameterisation. The continuous Fourier transform of a function $x(t)$ and its inverse will be given here as

$$X(F) = \int_{-\infty}^{\infty} e^{-2\pi iFt} x(t) dt, \tag{3.18}$$

$$x(t) = \int_{-\infty}^{\infty} e^{2\pi iFt} X(F) dF, \tag{3.19}$$

where F represents the frequency components.

The discrete Fourier transform (DFT) [64] can be defined as a set of N samples $\{X(k)\}$ of the Fourier transform $X(w)$ for a finite duration sequence $\{x(n)\}$ of length $L \leq N$. The sampling of $X(w)$ occurs at the N uniformly spaced frequencies $w_k = 2\pi k/N$, where $k = 0, 1, 2, \cdots, N-1$, and $X(w)$ is defined as

$$X(w) \equiv F\{x(n)\} = \sum_{n=-\infty}^{\infty} x(n)e^{2\pi iwn} \tag{3.20}$$

The definition of the DFT pair is as follows:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{2\pi ikn/N} \quad k = 0, 1, \cdots, N-1 \tag{3.21}$$

$$x(n) = (1/N)\sum_{k=0}^{N-1} X(k)e^{-2\pi ikn/N} \quad n = 0, 1, \cdots, N-1 \tag{3.22}$$

The fast Fourier transform (FFT) [14] is an efficient algorithm to determine the DFT and its inverse.

3.5.3 RASTA-PLP

The perceptual linear predictive (PLP) technique presented by Hermansky [81] to study auditory like spectral modifications. PLP analysis yields a low-dimensional representation of speech. This method utilises three ideas from the psychophysics of hearing to extract an estimate of the audio spectrum: the critical-band spectral resolution, the equal-loudness curve, and the intensity-loudness power law. The audio spectrum is then approximated using an autoregressive all-pole model. A block diagram of the PLP approach is shown in Figure 3.4. The

speech fragment is weighted by the Hamming window

$$W(n) = 0.54q - 0.46\cos[2\pi/(N-1)] \tag{3.23}$$

where N is the length of the window. The typical length of the window is around 20 ms. The windowed speech fragment is transformed into the frequency domain using DFT. In comparison with PLP analysis, RASTA-PLP analysis proposed by Hermansky et al. [82], is more robust to linear spectral distortions than PLP, since each frequency channel of PLP is band-pass filtered [78]. A block diagram of the RASTA-PLP method is shown in Figure 3.5.

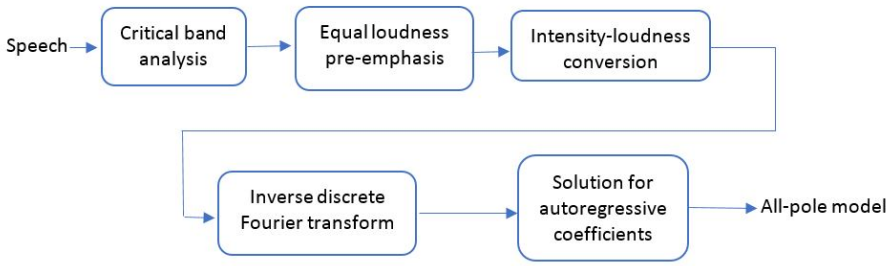


Figure 3.4. A block diagram of the PLP approach.

The steps of RASTA-PLP method are as follows [82].
For each analysis frame,

- The critical-band spectrum is computed (as in the PLP) and its logarithm is taken.
- The temporal derivative of the log critical-band spectrum is estimated utilising regression line through five sequential spectral values.

- Non-linear processing (i.e. applying median filtering or threshold) can be achieved.

- The log critical-band temporal derivative is re-integrated utilising a first order Infinite Impulse Response (IIR) system. The effective window size can be determined by adjusting the pole position of this system. This value can be set to 0.98, producing an exponential integration window with a 3-dB point after 34 frames.

- In accord with the PLP, the equal loudness curve is added and multiplied by 0.33 to simulate the power law of hearing.

- The inverse logarithm of the relative log spectrum is taken, producing a relative audio spectrum.

- An all-pole model of the spectrum is computed, following the conventional PLP method.

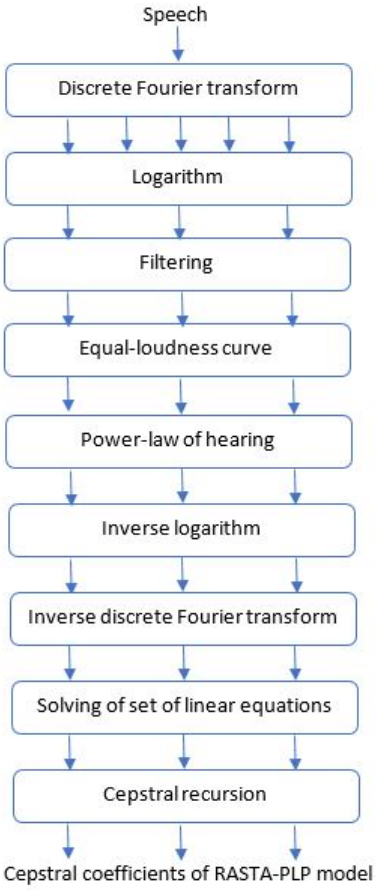


Figure 3.5. A block diagram of the RASTA-PLP approach [81].

3.5.4 Linear Predictive Coding

Linear Predictive Coding [42] models the sound signal as being the result of filtering by an all-pole filter. The linear prediction also called an autoregressive refers to the mechanism of utilising a linear combination of the previous speech frames, $s[n-1], s[n-2], \cdots, s[n-M]$, to predict the frame $s[n]$:

$$s[n] \approx \hat{s}[n] = -\sum_{i=1}^{M} a_i s[n-i] \tag{3.24}$$

where $\hat{s}[n]$ represent the predicted frame, and $a_i, i = 1, 2, \cdots, M$ are the LPC coefficients

The sampled error for the prediction can be defined as

$$e[n] = s[n] - \hat{s}[n] = s[n] + \sum_{i=1}^{M} a_i s[n-i] = \sum_{i=0}^{M} a_i s[n-i] \tag{3.25}$$

where $a_0 = 1$. The z transform of equation 3.25 can be obtained as,

$$E[z] = S[z] + \sum_{i=1}^{M} a_i S(z) z^{-i} = S(z)[1 + \sum_{i=0}^{M} a_i z^{-i}] = S(z)A(z) \tag{3.26}$$

equation 3.26 can be written as,

$$S[z] = \frac{E(z)}{A(z)} \tag{3.27}$$

So that, the sound signal can be represented as an output of a transfer function of an all-pole digital filter, where the input to the filter is the LPC error data $e[n]$, and the transfer function is $\frac{1}{A(z)}$. Consequently, equation 3.26 can be interpreted as an inverse filter whose transfer function is $A(z)$. The mechanism is as follows, if the sound signal $s[n]$ passed into an inverse filter, then the output will be the error data $e[n]$.

3.5.5 Line Spectral Pairs/Frequencies

Line spectral frequencies (LSF) [86, 89] parameters are one of the most efficient choices of transmission parameters for the LPC coefficients. They are particularly appropriate for transmission over a channel, as in a communication system. This is due to quantisation required to be performed for transmission vector and LPC is not very strong to

quantisation noise. The idea of LSP decomposition is to decompose the LP polynomial $A(z)$ into a symmetrical and antisymmetrical part represented by the $P(z)$ and $Q(z)$ respectively:

$$P(z) = A(z) + z^{-(m+1)}A(z^{-1}) \tag{3.28}$$

$$Q(z) = A(z) - z^{-(m+1)}A(z^{-1}) \tag{3.29}$$

The linear predictor $A(z)$ can be defined in terms of $P(z)$ and $Q(z)$ as follows:

$$A(z) = \frac{1}{2}(P(z) + Q(z)) \tag{3.30}$$

The process is shown in Figure 3.6. The LSP parameters are represented as the roots or (zeros) of $P(z)$ and $Q(z)$. Since all zeroes are placed on the unit circle, it is essential to specify the angle w to express the LSP. If LSP is represented in terms of the angular frequency, then the solutions are called line spectrum frequencies (LSF). The LSFs coefficients are generally the preferred feature vectors utilised in vector quantisation.

Figure 3.6. The decomposition of the linear predictor A(z).

3.5.6 Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients are found by warping the frequencies of auditory signal onto the mel scale [160]. An approximation

of the non-linear frequency response of the human ear is namely mel scale. The mel coefficients are distributed approximately logarithmically above 1 kHz and linearly up to 1 kHz , and may be computed in the following manner [16, 29]:

- Divide the auditory signal into windows.
- Take the DFT of the signal in each window.
- Take the spectral magnitude and logarithm of each frame.
- Warp the frequencies according to the mel scale.
- Take the inverse DFT Fourier transform.

3.6 Audio processing and synchronisation with video

A good visual speech synthesis system, which produces facial animation given audio speech input, does not only require generating both kinds of data in high quality; it also requires achieving good temporal alignment between the two, for example, synchrony between audio and video, to deliver a visual signal that is consistent with those delivered by a real talking person.

The method used in this work for features extraction was RASTA-PLP features, which are perceptually-motivated. To satisfy the requirement of having a window where the acoustic data is stable, a time window size of 25 ms with a hop size of 10ms is typically utilised resulting in a frequency of 100 Hz. This results in a mismatch with the video frame rates of 25 fps. So that, the audio speech parameters need to be downsampled to match the visual processing rate. Deena [39] conducted experiments to investigate three methods to matching the auditory features to visual

features. The first method consists of utilising an audio speech window of 50 ms and a hop window of 40 ms. The second and third methods consist of utilising an auditory window of 25 ms and a hop window of 10 ms to obtain acaustic features at 100 Hz, then downsampling to match the visual features utilising polyphase quadrature filtering [150] and median filtering [4]. It was found that features computed from the first approach were smoother but some salient features may be potentially discarded. Theobald and Wilkinson [173] conducted experiments to determine the effect of the audio window size on the auditory-visual correlation. The window size is assumed to be 40 ms duration so the audio features were in direct correspondence with 25 Hz visual features. It was found that smoother acoustic properties can be obtained utilising auditory features over a longer period that have higher linear correlation with visual features, as compared to the correlation between upsampled video to match the acoustic parameterised at 100 Hz and the audio speech parameters. In this work, we downsampled the visual data to 25 fps in order to obtain a reasonable corpus size, and processed the speech signal in frames of 50 ms with a 40 ms overlapping (at the visual frame rate 25 Hz). Figure 3.7 shows the mean trajectories of the RASTA-PLP parameters for a sequence of the LIPS corpora.

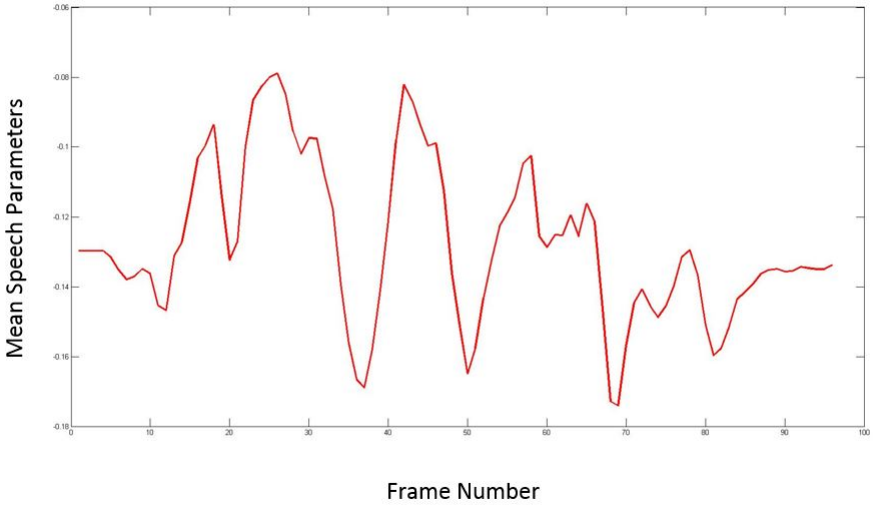


Figure 3.7. Mean RASTA-PLP trajectories for a given LIPS utterance.

3.7 Summary

A review of shape and appearance models has been presented in this chapter. This requires hand-annotated images for training the point distribution model and can handle variations in pose. The appearance is described by the shape-free texture model, built by warping each labelled image from the landmarks to the mean shape and performing a PCA on the resultant images. A combined model is constructed by calculating and concatenating the shape and shape-free texture parameters for the labelled images and processing a third PCA on the resulting parameters. The AAM algorithm seeks to find the model parameters which generate a synthetic image as close as possible to the target image. It has been used in this work to automatically project the face of a talker in each frame of a video sequence onto the principal components. We have also described several techniques to representing

speech signal that are commonly used in speech feature extraction including RASTA-PLP. These approaches are used frequently in speech recognition. In this thesis RASTA-PLP analysis is used to provide robust speech features. We have also presented the data corpora used in our work as well as details of auditory and visual parameterisations on the data corpora.

Chapter 4

IMPROVING THE ACCURACY OF AUDITORY-VISUAL MAPPING USING STATE-SPACE MODEL

One of the challenges in visual animation is producing accurate videos of faces from the audio signal alone. Our hypothesis is that using a more accurate AAM in the shared Gaussian process latent variable model (SGPLVM) will improve the accuracy of the produced videos of the talking faces.

In this chapter, we improve the performance of the SGPLVM [38]. As stated above, it is required to build a more accurate AAM to develop the SGPLVM model of the talking faces. An AAM is built on a larger number of sequences with more landmark points for each frame than in Deena et al. [38]. Experiments are conducted to investigate the hypothesis of increasing the number of landmark points can improve the accuracy of AAM. In addition, experiments are performed to

investigate that constructing an AAM on larger dataset can improve its accuracy. From the quantitative evaluation, we found that our improved SGPLVM model using a more accurate AAM outperforms the existing SGPLVM model [38].

Initially we describe a non-linear state-space model that can be utilised to jointly model the auditory and visual features of a visual speech synthesis. Different latent variable models are described utilising the probabilistic graphical model structure, emphasising their previous use for visual speech animation. In addition, a shared linear dynamical system (SLDS) to model auditory and visual modalities of a talking face and a shared latent variable model utilising Gaussian processes are presented.

4.1 Graphical models

Before we present the specific models in which we are interested, it is beneficial to have a general overview of the graphical models. In a graphical model, a node denotes the variables and the arcs denote probabilistic dependencies between the variables. There are two types of graphical models: undirected graphical models, for example Markov random fields, and directed graphical models, for example Bayesian networks [9]. Markov random frameworks are appropriate for the representation of soft constraints between points. Bayesian frameworks are suitable for modelling the relationships between random variables and can be utilised to model generative processes. In Bayesian networks, the existence of a link between variable X and Y indicates a conditional dependency between Y and X. The lack of links is appropriate because it decodes to conditional independence properties. In our

work, we deal with the Bayesian networks. Utilising the product and sum rules of probability, graphical models permit us to do inference of the hidden variables given several observations. Inference typically involves marginalising over several variables, which involves integrations in the continuous case and summations in the discrete case. The generalised approach for achieving inference in graphical models is named the sum-product algorithm [96], which depend on converting the graphical model to a factor graph. The more general process for graphical model including those with loops is Belief propagation [201].

4.2 Probabilistic principal component analysis

Principal component analysis (PCA) [92] is a well-established approach for dimensionality reduction. Tipping [174] proposed a probabilistic formulation of PCA from a Gaussian latent variable model. In linear subspace models, a D-dimensional observation data Y is related to a corresponding q-dimensional latent variables X, as follows:

$$Y = \mathbf{W}X + \boldsymbol{\mu} + \epsilon \tag{4.1}$$

The matrix $\mathbf{W}$ relates the two sets of variables, while $\boldsymbol{\mu}$ permits having non-zero means in the model and ϵ refer to the noise term. With $q < D$, the latent variables could give a more parsimonious explanation of the dependencies between the observations. A reformulation of PCA as a latent variable model is given by the model equation 4.1 which is named as probabilistic principal component analysis (PPCA) in the case where the noise ϵ follows an isotropic Gaussian distribution. In this work, the probabilistic principal component analysis is one of the

approaches utilised to initialise latent spaces of the proposed manifold relevance determination methods.

4.3 Gaussian Mixture Models

A Gaussian mixture model (GMM) is a non-linear modelling approach. It is a mixture of some Gaussian distributions and different subclasses can be represented inside one class. A GMM is described by

$$P(y) = \sum_{i=1}^{M} \alpha_i g(\mathbf{y}, \mu_i, \mathbf{\Sigma}_i) \tag{4.2}$$

where α_i are the prior probability of each Gaussian, μ_i are the means of the Gaussians, $\mathbf{\Sigma}_i$ are the covariance matrices and M represents the number of Gaussians [49]. In a GMM, the latent states $\pi \in 1, ...M$ are discrete. GMMs can be utilised to segment information into clusters of Gaussian distributions.

Expectation-Maximisation (EM) [41] is a widely utilised approach for estimating the parameters of the GMM. It is an iterative method. To proceed EM iteratively, there are two steps, the expectation and the maximisation. In the expectation step, the expectation of the log-likelihood over all possible assignments of data points to sources are calculated. In the maximisation step, the expectation by differentiating written current parameters are maximised.

4.4 Hidden Markov Models

Hidden Markov models (HMMs) have been utilised since 1970s in speech recognition field [40]. At the beginning, they were utilised to model auditory features but more recently they have been utilised to

model auditory-visual speech features [12, 30, 189]. A hidden Markov model is a probabilistic model that allows the temporal dependencies of information to be modelled utilising a transition matrix, where each element of the matrix represents the conditional transition probability from one state to another. The state inferred from the observation. It is hidden because the observation is a probabilistic function of the state. The basic elements of an HMM are as follows:

- k is the number of states in the model, the state at discrete time t is given by $q_k(t)$
- $\mathbf{A} = \{a_{ij}\}$ describes the transition probability matrix, where

$$a_{ij} = P(q_j(t+1)|q_i(t)), \quad 1 \leq i, j \leq k \tag{4.3}$$

  that is, a_{ij} is the probability of making a transition from state i to state j, $\mathbf{B} = \{b_j(\mathbf{O})\}$ is the set of emission probability distributions, where $b_j(\mathbf{O})$ is the probability distribution for state j.

- The initial probabilities of being in state i at $t = 1$

$$\pi = \pi_i \tag{4.4}$$

The model parameters of an HMM are defined as:

$$\lambda = (\mathbf{A}, \mathbf{B}, \pi) \tag{4.5}$$

The three basic problems with HMMs identified by [143] are:

- Given the observation sequence $\mathbf{O} = (O_1, O_2, ..., O_N)$ and a model

parameters $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$, compute the probability of the observation sequence $P(\mathbf{O}|\lambda)$.

- Given the observation sequence $\mathbf{O} = (O_1, O_2, ..., O_N)$ and a model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$, compute a corresponding hidden state sequence $\mathbf{Q} = (q_1, q_2, ..., q_N)$, which best explains the observation.

- Adjust the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$ to maximise $P(\mathbf{O}|\lambda)$.

Brand [12] applied HMMs to visual speech synthesis. HMM-based to map from auditory parameters (RASTA-PLP/LPC) to marker data is built. Facial landmarks tracked on the face of the talker to represent facial features. The HMM is trained to recognise the auditory signal, then for animation a Viterbi algorithm is utilised through vocal HMMs to search for most likely hidden state sequence. Wang et al. [186] utilised the maximum likelihood (ML) based estimation for the auditory-visual joint HMM training. The ML-based training does not explicitly optimise the quality of generated trajectory. To address this issue, Wang et al. [190] propose to utilise the minimum generated visual trajectory error method to enhance speech animation. The model parameters were improved by minimizing the mean square errors between the synthesised visual trajectories and the real ones utilising a probabilistic descent (PD) algorithm. The MGE training was incorporated into HMM trajectory-guided talking face rendering system. The HMM-based method can produce an efficient estimate of the visual speech given any new speech input, the animation synthesised is synchronised with speech, but lacks photo realism. Therefore, Wang and Soong [189] combined both sample-based concatenation and the HMM-based modelling methods to get both lip in synced with speech and

photo-realistic. In this work, we limit ourselves to learning-based approaches when answering research problems related to auditory-visual mapping.

4.5 Linear Dynamical System

The linear dynamical system (LDS) is a continuous state space model, and it is more suitable for synthesis because there is a dynamical mapping from the preceding state to the next. The graphical model of the LDS is similar to the HMM with the difference that the latent points are continuous instead of discrete. The LDS is a generative model in which the observations can be generated from the states. Saisan [153] utilised LDS for the synthesis of lip articulation with speech as the driving input. Lehn-Schiler et al. [105] applied SLDS to model auditory and visual modalities of a talking head. The state-space equations of the model is as follows:

$$x_t = \mathbf{A}x_{t-1} + n_t^x \tag{4.6}$$

$$y_t = \mathbf{B}x_t + n_t^y \tag{4.7}$$

$$z_t = \mathbf{C}x_t + n_t^z \tag{4.8}$$

where x_t is a hidden vairiable, y_t is the audio features, and z_t is the visual features, n is Gaussian noise parameter that add to each equation, $\mathbf{A}$ is the prediction matrix that maps previous states to next states, $\mathbf{B}$ and $\mathbf{C}$ are the observation matrices which transforms latent states to observations. Through training the audio and visual features are

known, and both the observation equations can be represented in one.

$$\begin{pmatrix} y_t \\ z_t \end{pmatrix} = \begin{pmatrix} \mathbf{B} \\ \mathbf{C} \end{pmatrix} x_t + \begin{pmatrix} n_t^y \\ n_t^z \end{pmatrix} \tag{4.9}$$

The parameters $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \Sigma^x, \Sigma^y, \Sigma^z\}$ can be found utilising the EM algorithm [41] on the training data, where $\Sigma's$ are the diagonal covariance matrices of the noise parameters. Given a new audio sequence, the corresponding visual features can be obtained, $y_t = \mathbf{C}x_t$.

Englebienne [54] used a more robust model called as switching linear dynamical system (SLDS) by augmenting the linear dynamical system with switching states, where phonemes were utilised as the switching states and visual speech data were the observations. However, Englebienne [54] displayed that the parametric assumptions in the SLDS are not appropriate for speech animation and simplified the model to acquire a model called the deterministic process dynamical system (DPDS).

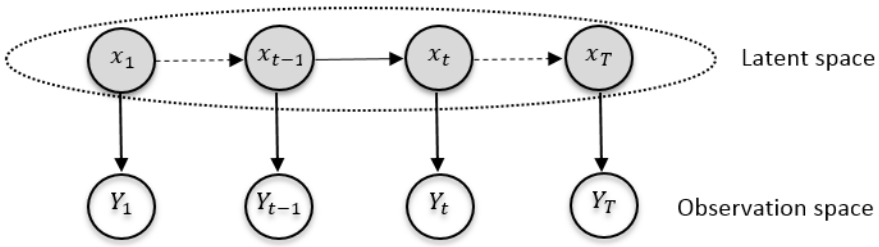


Figure 4.1. Graphical model for LDS.

4.6 Gaussian Processes

Gaussian processes (GP) are generalisations of Gaussian distributions specified over infinite index sets [147]. A GP is defined as a collection

of random variables, any finite number of which have joint Gaussian distributions. It can be utilised to define distribution over functions. A GP is completely determined by its mean function μ, which is often taken to be zero and its covariance function k defined over infinite index sets, $\mathbf{x}$ [145, 146].

$$f \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{4.10}$$

where

$$\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{4.11}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))] \tag{4.12}$$

The radial basis function (RBF) kernel is widely utilised in the GP. The covariance function for the RBF kernel is represented by:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \alpha \, exp(-\frac{\gamma}{2}(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)) \tag{4.13}$$

α is the variance of the kernel, γ is its inverse width. An elegant structure of GPs for regression is provided. Given a set consisting of N input points $X = \{\mathbf{x}_n\}_{n=1}^N$ of dimension q and a set of corresponding output points targets $\mathbf{y} = \{y_n\}_{n=1}^N$. A function f is fitted to the data such that:

$$y_n = f(\mathbf{x}_n) + \epsilon \quad \epsilon \sim \mathcal{N}(0, \beta^{-1}) \tag{4.14}$$

where β is inverse variance. One widely utilised covariance function which combines an RBF function, and a noise term is

$$k(\mathbf{x}_i, \mathbf{x}_j) = \alpha exp(-\frac{\gamma}{2} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2) + \beta^{-1}\delta_{ij} \tag{4.15}$$

δ_{ij} is Kroneckers delta function. By integrating over f, a marginal likelihood is obtained:

$$P(Y|X,\theta) = \int P(\mathbf{y}|f)P(f|X,\theta)df \tag{4.16}$$

The parameters θ of the covariance function k are denoted as the hyper-parameters of the GP given by $\theta = \{\alpha, \gamma, \beta\}$. The GP parameters are obtained utilising maximum likelihood, typically utilising gradient-based optimisation approaches such as conjugate gradient optimisation [69].

$$\theta = argmax_\theta P(Y|X,\theta) \tag{4.17}$$

Then the learnt regression model can be utilised to predict function values y_* at previous unseen input points x_*. The Gaussian distribution of the predictive distribution is:

$$P(y_*|\mathbf{x}_*, X, \mathbf{y}) \sim \mathcal{N}(\mu_*, \sigma_*^2) \tag{4.18}$$

The joint distribution between $\mathbf{y}$ and y_* given by

$$\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} \mathbf{K} + \beta^{-1}\mathbf{I} & \mathbf{k} \\ \mathbf{k}^T & k(\mathbf{x}_*, \mathbf{x}_*) + \beta^{-1} \end{bmatrix}\right) \tag{4.19}$$

where $\mathbf{K} = k(X, X)$ and $\mathbf{k} = k(X, \mathbf{x}_*)$

4.6.1 The Gaussian Process Latent Variable Model

The Gaussian process latent variable model (GPLVM) [101] is an algorithm for dimensionality reduction using GPs. It is a generative model,

where observation space $y_n \in \mathbb{R}^D$ is assumed to be generated from a latent space $x_n \in \mathbb{R}^q$ through a mapping f that is corrupted by noise. The relationship between the data points and the latent points is the same as for GP regression as given by equation 4.14. By placing a zero mean GP prior on the mapping f and marginalising it, the likelihood $P(Y|X,\theta)$ is obtained, which is a product of D GPs, while θ represents the hyper parameters of the covariance function

$$P\left(Y|X,\theta\right) = \prod_{i=1}^{D} \frac{1}{(2\pi)^{\frac{N}{2}}|\boldsymbol{K}|^{\frac{1}{2}}} exp(-\frac{1}{2} y_{:,i}^{T} \boldsymbol{K}^{-1} y_{:,i}) \tag{4.20}$$

where $y_{:,i}$ is the ith column from the data matrix, Y.

Maximising the marginal likelihood in equation 4.20 with respect to both the latent points X and the hyper-parameters θ of the covariance function results in the latent space representation of the GPLVM:

$$\left\{\widehat{X}, \widehat{\theta}\right\} = argmax_{X,\theta} P\left(Y|X,\theta\right) \tag{4.21}$$

The back-constrained GPLVM.

A smooth mapping from the latent space X to the data space Y is specified using a smooth covariance function, which means that points close in the latent space will be close in the data space. However, it does not ensure the opposite case. Therefore, an extension to the GPLVM is proposed by using an inverse parametric mapping that maps points from the observation space to the latent space. This constrains points that are close in the data space to be close in the latent space [52]:

$$x_i = g\left(y_i, \boldsymbol{W}\right) \tag{4.22}$$

where $\boldsymbol{W}$ is the mapping parameter set of the back-constraint kernel function. This is typically computed with a RBF network or multi-layer perception (MLP). The maximisation in equation 4.21 is then changed from optimisation with respect to the latent points X to optimisation the parameters of the back-constraining mapping $\boldsymbol{W}$:

$$\left\{\widehat{\boldsymbol{W}}, \widehat{\theta}\right\} = argmax_{\boldsymbol{W},\theta} P\left(Y|\boldsymbol{W}, \theta\right) \tag{4.23}$$

Dynamics.

An extension of the GPLVM was proposed by Wang [184]; this produces a latent space that preserves sequential relationships between points on the data variables, as well as on the latent variables. This is done by specifying a predictive function over the sequence in latent space, x_t:

$$x_t = h(x_{t-1}) + \epsilon_{dyn} \tag{4.24}$$

where $\epsilon_{dyn} \sim \mathcal{N}(0, \sigma_{dyn}^{-1} I)$. A GP prior can then be placed over the function $h(x)$, and marginalising this mapping results in a new objective function. By optimising this objective function, the latent points that preserve temporal relationships in the data are obtained. The new objective function is given by:

$$P\left\{\widehat{X}, \widehat{\theta}_Y, \widehat{\theta}_{dyn}\right\} = argmax_{X,\theta_Y,\theta_{dyn}} P\left(Y|X, \theta_Y\right) P\left(X|\theta_{dyn}\right) \tag{4.25}$$

where θ_{dyn} being the hyper-parameters of the dynamics kernel.

To construct a shared latent structure between two views, $Y \in \mathbb{R}^{N \times D_Y}$ and $Z \in \mathbb{R}^{N \times D_Z}$ with a shared latent space $X \in \mathbb{R}^{N \times Q}$, the GPLVM is modified to learn separate sets of GPs for each of the differ-

ent observation spaces from a shared latent space. The latent space is given by maximising the joint likelihood of the two observation spaces:

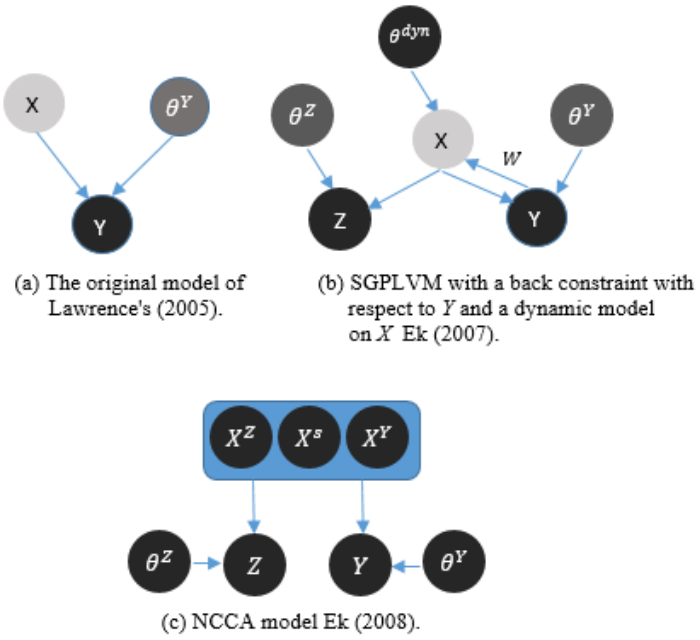
$$P\left(Y, Z|X, \theta_s\right) = P\left(Y|X, \theta_Y\right) P\left(Z|X, \theta_Z\right) \tag{4.26}$$

where $\theta_s = \{\theta_Y, \theta_Z\}$ is two different sets of hyper-parameters.

In Ek et al. [52], the SGPLVM is used to learn a mapping between silhouette and pose features; the pose ambiguities from the silhouette observations are resolved by considering sequential data. This is done by learning a dynamical model over the latent space to disambiguate ambiguous silhouettes. Deena et al. [36] used the same SGPLVM approach to model coarticulation. First, placing a back constraint with respect to auditory features ensures a smooth mapping from the latent space to the observation space. Second, a dynamical model is placed on the latent space to respect the data's dynamics in the training and inference phases. Canonical correlation analysis (CCA) coupled with linear regression was used in Theobald and Wilkinson [173] to model the relationship between auditory and visual features; it was also used to predict visual features from the auditory features.

An extension of CCA proposed in Ek et al. [51], called the non-consolidating components analysis (NCCA) model, is used to address the ambiguities in a human motion dataset by decomposing the latent space into subspaces whereby a private latent space for each of the observation spaces is learned in addition to the shared latent space. The NCCA model encodes the variance in the data separately, so that it does not influence the inference procedure; this represents the advantage of using this model compared to other conditional models. An NCCA

model is also used for modelling and mapping human facial expression space, represented by facial landmarks, to a robot actuator space [50]. The ambiguity in this case relates to robot poses, with multiple robot poses that are most likely to be the solution to a facial expression in the facial expression space. Figure 4.2 shows various types of graphical GPLVMs. Recently, Damianou et al. [34] used a manifold relevance determination (MRD) framework to predict a 3D human pose from a silhouette in an ambiguous setting. To perform disambiguation, they include latent space priors that incorporate the dynamic nature of the data.



(a) The original model of Lawrence's (2005).

(b) SGPLVM with a back constraint with respect to Y and a dynamic model on X Ek (2007).

(c) NCCA model Ek (2008).

Figure 4.2. The structure of different GPLVM models. (a) In Lawrence's (2005) original model the observed data Y is represented using a single latent variable X. (b) The SGPLVM with a dynamic model on the latent points and with a back constraint with respect to the observation Y proposed by [52]. (c) Private latent spaces introduced by [51] to explain variance specific to one of the observations.

Sparse Approximations

The computational complexity of GP can be reduced from $O(N^3)$ to a more tractable $O(k^2N)$, where k refer to the number of points retained in the sparse representation. The sparse approximations require augmenting the function values at the training points, $\mathbf{F} \in (\mathbb{R})^{N\times D}$, and the function values at the test points, $\mathbf{F}_* \in (\mathbb{R})^{\infty\times D}$, by an additional collection of variables called the "active points", "support points" or "inducing variables", $\mathbf{U} \in (\mathbb{R})^{k\times D}$. There are different approximations considered in the context of the GPLVM in order to make training and inference tractable [102]. The techniques in sparse Gaussian process regression (GPR) that applied to the GPLVM were deterministic training conditional (DTC) approximation, fully independent training conditional (FITC) approximation and partially independent training conditional (PITC) approximation. Deena [39] found that training shared Gaussian process dynamical model (SGPDM) with sparse approximation FITC using $k = 100$ as proposed by Lawrence [102] give better performance than PITC and DTC for LIPS dataset.

4.7 Switching Shared Gaussian Process Dynamical Model

The above SGPDMs assume a single dynamics in the latent space. Chen et al. [19] proposed the switching shared Gaussian process dynamical model (SSGPDM) to deal with composite types of dynamics when jointly modelling silhouettes and 3D pose data, this model is a non-parametric switching state-space model which multiple SGPDMs are indexed by switching states $\boldsymbol{\pi} = [\pi_1, \cdots\cdots, \pi_N]^T$. Deena et al. [38] suggested the SSGPDM approach to learn the mapping from auditory to visual features. In their method auditory and visual signal from a

talking head corpus are jointly modelled, then variable length Markov models (VLMMs) trained on labelled phonetic data in order to get the switching states. The graphical model of the SSGPDM is shown in Figure 4.3, where the three layers of an SSGPDM are represented

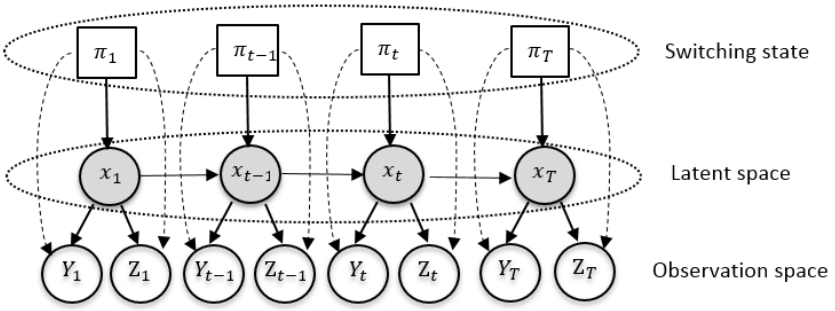


Figure 4.3. Graphical model for SSGPDM by Deena et al. [38].

4.7.1 Variable Length Markov Model

Training an nth-order Markov model is infeasible as the size of Markov chain grows exponentially, and needs large number of data to estimate their parameters. Ron et al. [148] suggested a robust extension of nth-order Markov models which allow the memory length to vary locally based on the specific realisation of backward states. A VLMM was formulated as a probabilistic finite state automation (PFSA). The PFSA is a 5-tuple (Q, Σ, τ, γ, s), where Q is a finite set of states, Σ is a finite alphabet. Each VLMM state corresponds to a string of tokens of at most length N, representing the memory in the conditional transition of the VLMM. $\tau : Q \times \Sigma \rightarrow Q$ is the transition function, $\gamma : Q \times \Sigma \rightarrow [0, 1]$ is the output probability function, and $s : Q \rightarrow [0, 1]$ is the probability distribution over the initial states, s. A VLMM of order N can be trained on a stream of discrete collections of symbols

from Σ, resulting in a predictive model that can predict a symbol σ using a previous string of symbols or context w of maximum length N. A predictive model can be obtained by training a VLMM of order N on a stream of discrete collections of symbols from Σ which can predict a symbol σ utilising a previous context w of maximum length N. Training a VLMM requires scanning through the training collections and constructing a prediction suffix tree (PFT) in which each node represents a string prefix (i.e., context) w of at most length $N-1$. To train a VLMM of maximum order N, w can be considered a prefix of length $N-1$ which can be utilised to predict the next character σ' according to an estimate $\hat{P}(\sigma'|w)$ of $P(\sigma'|w)$. Given a context σw and its parent w, the amount of information gained is then measured utilising weighted Kullback-Leibler divergence (KL):

$$\Delta H(\sigma w, w) = \hat{P}(\sigma w)\Sigma_{\sigma'}\hat{P}(\sigma'|\sigma w)log\frac{\hat{P}(\sigma'|\sigma w)}{\hat{P}(\sigma'|w)} \tag{4.27}$$

The longer memory σw is retained, when $\Delta H(\sigma w, w)$ exceeds a given threshold ϵ, otherwise the shorter memory w is sufficient. Converting the suffix tree to a PFSA which representing the trained VLMM is the final stage of training. For more details refer to [74, 148].

4.8 Inference utilising the SGPDM

$\hat{Z}$ can be inferred from $\hat{Y}$, by first obtaining the corresponding latent points, $\hat{X}$. The optimisation of latent points needs to be performed with respect to both the GP mapping from X to Y and the dynamical GP, this is done by formulating a joint likelihood as given in equation 4.28. Using conjugate gradient optimisation, the likelihood is then optimised

to find the most likely latent coordinates for a sequence of auditory features.

$$\hat{X} = argmax_{x_*} P(\hat{Y}, X_*|Y, X, \theta_Y, \theta_{dyn}) \tag{4.28}$$

Where X_* refer to the initialisation of the latent points. Then the observation space $\hat{Z}$ can be found utilising the mean prediction of the GP from the latent space X to the visual space Z:

$$\hat{Z} = k(\hat{X}, X)^T \mathbf{K}^{-1} Z \tag{4.29}$$

4.9 Auditory-visual mapping using SGPLVM

The SGPLVM [39] can be utilised to combine auditory and visual information through a shared latent space. The advantage of utilising the SGPLVM over HMMs are that the state-space is continuous and offers a richer representation, thus bypasses the need to interpolate between discrete states to synthesis image signal. Moreover, the generative process of speech is represented by utilising a shared latent space that maps to both the auditory and image modalities, which is more appropriate than training a HMM on image signal then remapping it to auditory signal as done by Brand [12]. The better performance of SGPLVM as compared to shared LDS can be explained by the fact that the observation and dynamical mappings are non-linear GPs. The dynamics of speech are highly non-linear and a shared LDS only provides a linear approximation to the dynamics. Moreover, the SGPLVM can be utilised to satisfy the many-to-one mapping between phonemes and visemes.

4.10 Experiments

Our motivation is to improve the performance of SGPLVM [38], by building an AAM on a larger dataset with more facial landmarks identified for each frame to represent different visual realisations of sounds. Experiments described in Subsection 4.10.1 are performed to compare our SGPLVM method with Deena's method [38] using the same dataset for training and synthesis. The dimensionality of the latent space is fixed to be 5, as Deena [39] found that the optimal latent space is 5 for LIPS dataset. The training time would be affected negatively when using higher latent spaces especially in the case of introducing dynamics and back-constraints. The SGPLVMs are trained on the training set, then the inference is performed utilising only auditory signal from the test collection with the sequence optimisation approach described in Section 4.8. In addition experiments described in Subsection 4.10.2 are performed to investigate the effect of increasing the landmark points around the mouth to have a smoother lip boundary. Experiments described in Subsection 4.10.3 are conducted to investigate that constructing AAM on a larger dataset can improves the performance accuracy.

4.10.1 Experiment 1: Objective evaluation for the SGPLVM

In this work we want to improve the accuracy of SGPLVM [38] by training AAM on a large number of sequences with more landmarks around the mouth and compare our approach with Deena's method [39]. The maximum number of images used is 6000 due to the $O(N^3)$ complexity of SGPLVM training. Deena [39] conducted experiments and compared two back-constraint approaches, called KBR [9] and MLP [8] and examined the effect of varying their parameters. They showed

that for LIPS dataset, utilising a KBR back-constraint produces better results than not utilising it. In addition, the results showed that the MLP back-constraints lead to worse results than not utilising back-constraints at all. So that having a proper back-constraint allows them to model the many-to-one relationship from phonemes to visemes. In addition, in their experiments they found that 5 is the optimal latent space, and the performance of the sparse approximation FITC was better than PITC and DTC utilising $k = 100$ support points, as proposed by Lawrence [102]. Such that, in our experiments we used KBR back-constraints with respect to auditory signal and an autoregressive dynamical model on the latent space, and fix the dimensionality of latent space and the sparse approximations accordingly. Moreover, we used both ASM (refer to Chapter 5) and AAM features for visual representation and RASTA-PLP for audio representation. Deena [39] train the AAM choosing 184 prototype images, which was done by selecting 4 random frames throughout the dataset, from each of the 45 sounds plus breathing and silence. Afterwards, 56 markup points were located around the lips, face and nose in each of the example images. An AAM was built on the shapes and images, then the remaining images was projected to AAM parameters.

In our experiments, we have used the same number of sequences for training and testing as by Deena [39], such that the training and testing collections do not overlap and trained AAM on a large number of sequences with more landmarks around the mouth. Due to the complexity of GPLVM training, optimisation of a GPLVM likelihood becomes intractable when the amount of training frames exceeds a few thousand frames. Therefore a repeated random subsampling method

for choosing about 10 sets, each set of approximately 55 sequences from the 236 auditory-visual collection pairs is utilised for training ten SGPLVMs, giving an average of 6000 frames for each of ten training sets. Two sets of about 21 utterances for each set from the remaining of the auditory-visual collection pairs is utilised for testing of all ten SGPLVMs, such that the training and testing sets do not overlap.

Moreover, Deena [39] used mean-centering normalisation for LIPS corpus to obtain normalised data varying around the zero baseline. Hence, mean-centering AAM parameters which is discussed in more detail in Section 3.15 are used in the following experiments. AAMs are trained on each of the 55 sequences from the 236 auditory-visual sequence pairs, with 97 facial landmarks identified for each frame; 38 of them described the inner and outer mouth shape.

In assessing the results of this method, we used the average mean squared error (AMSE) between test feature vectors and ground truth, as this is the most commonly used error for multivariate data [39]: this is shown in equation 4.30,

$$AMSE = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{I}(z_{k,i} - \widehat{z}_{k,i})^2 \tag{4.30}$$

Table 4.1 shows the AMSE between the ground truth and synthesised AAM features, obtained from our method and Deena's et al. [38] approach. The results show that the AMSE error obtained from our method is lower than Deena's approach, due to modelling AAM using larger visual dataset and using more landmarks around the mouth.

Figures 4.4 and 4.5 show typical examples of the mouth landmark parameter reconstructed from the AAM parameters against ground truth and synthesised output shape parameters against ground truth

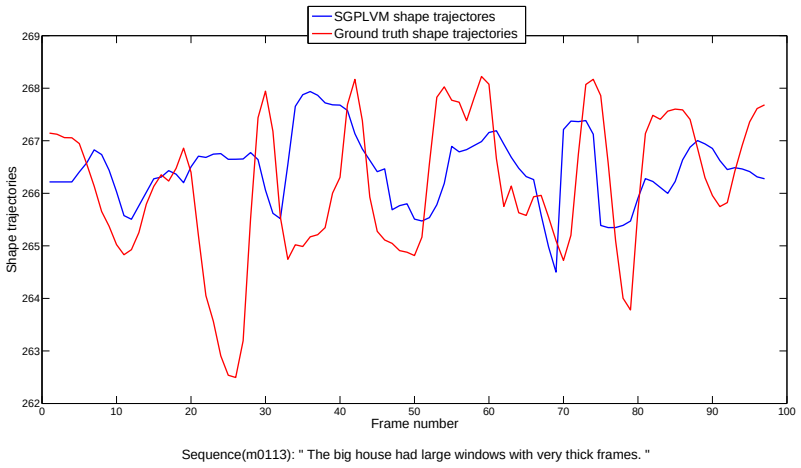
parameters respectively. The synthetic trajectories are calculated using two test sequences of LIPS dataset. The comparisons of synthesised trajectories to ground truths gives a good overall representation of a talking faces lip-synch ability. However, the results show that the trajectories are smoothed out as compared to the ground truth. This may be due to the absence of a Bayesian formulation so that the dimensionality of the latent spaces is not allowed to be estimated automatically. As a result, model selection experiments were performed by Deena [39] to determine the optimal latent dimensionality and other free parameters in the model.

The goal is to produce synthetic video indistinguishable from real video. So that to learn an efficient latent space, in the next chapter a novel method to visual speech synthesis utilising a joint probabilistic model is introduced, namely the Gaussian process latent variable model trimmed with manifold relevance determination model. This is a fully Bayesian latent variable model that uses conditional non-linear independence structures.

More results and comparison to manifold relevance determination are presented in Chapter 6.

Table 4.1. Quantitative evaluation of our SGPLVM using a more accurate AAM vs. Deena's methods.

Method	Audio representation	Visual representation	AMSE
Our method (SGPLVM)	Continuous	AAM	0.01826±0.0053
SSGPLVM [38]	Continuous	AAM	0.0413±0.0063
SGPLVM [38]	Continuous	AAM	0.0535±0.0090



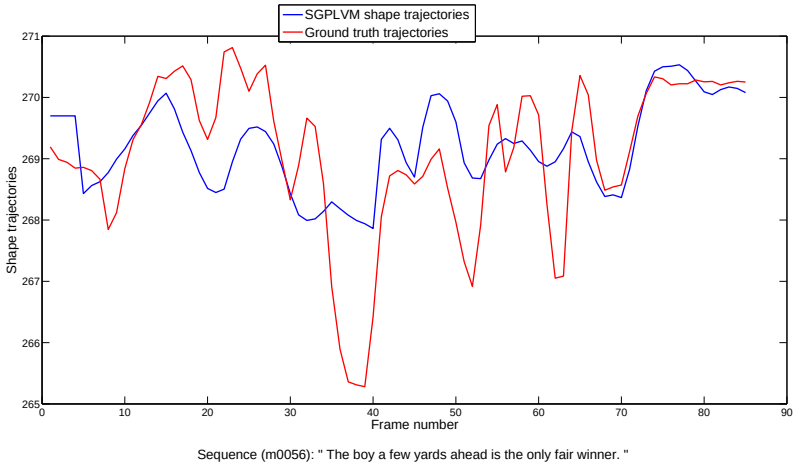


Figure 4.4. Sample shape trajectories obtained from SGPLVM and the corresponding ground truth trajectories.

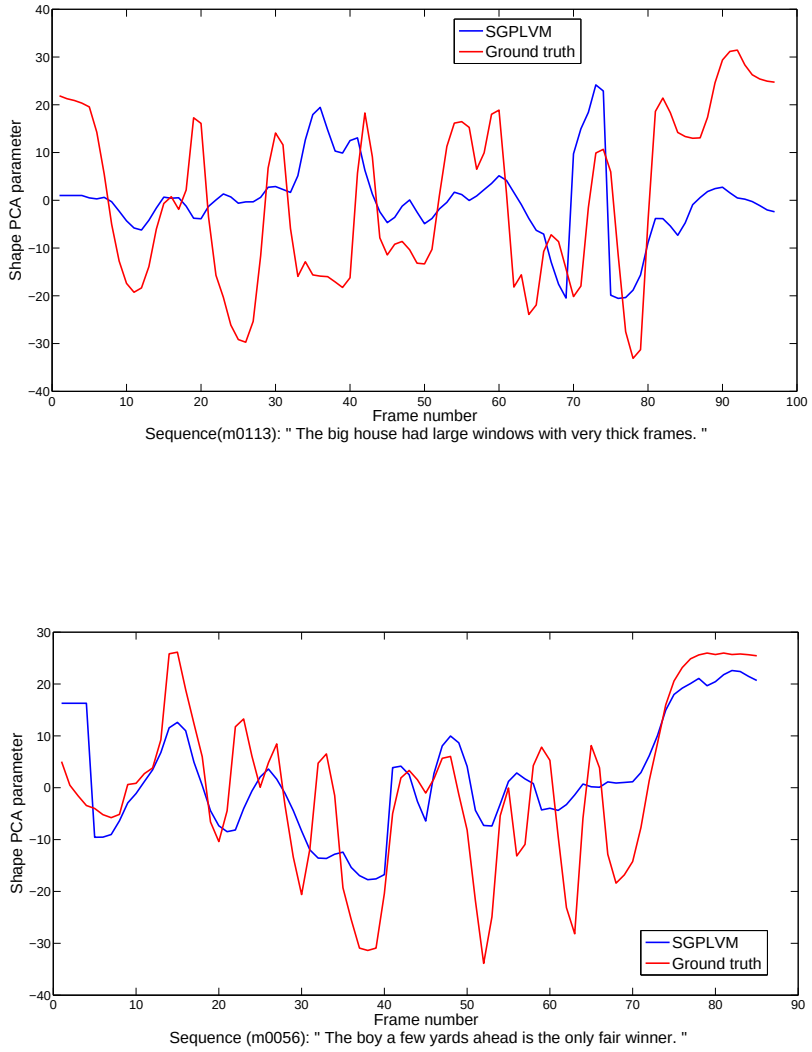


Figure 4.5. Synthetic sample against ground truth shape PCA parameter trajectories.

4.10.2 Experiment 2: Increasing the number of landmark points

In this experiment, the effects of the number of landmarks on the AAM performance are considered. We want to investigate that increasing the number of landmark points can improve the accuracy of AAM. For visual processing of the LIPS dataset Deena et al. [38] placed 56 landmark points around the face, lips and nose in each

of the 184 images. Then an AAM was built using this number of prototype images. A more accurate active appearance model, with 97 facial landmarks identified for each frame is built to improve the performance of SGPLVM [38]: 38 of them described the inner and outer mouth shape. In this section, experiments are conducted to compute the performance accuracy of increasing landmarks around the mouth shape. We use only mouth shapes because most of the important visual speech information is contained in the mouth shapes. Figure 4.6 shows the bar chart of root mean squared error (RMSE) in shape normalised images compared to the ground truth lip images for the 85 frames of the sequence (m0056): "The boy a few yards ahead is the only fair winner". It can be seen from the figure that the RMSE value decreases with the increasing of the number of landmarks. This means that increasing landmarks around the mouth can give better results and improve the performance accuracy. Figure 4.7 shows the ground truth mouth and several shape normalised mouth images obtained using different number of landmarks. It can be shown that a smoother lip boundary can obtained using more landmark points.

In these experiments, the eigenspaces are computed using the eigenvalue decomposition (EVD) of the covariance matrix of the data, which is a standard approach. Suppose there is a set of N data samples, each n dimensional where $n/2$ is number of the landmark points in an image. The samples coordinates are combined into a matrix of size $N \times n$. The EVD of the covariance of the data can be defined by

$$(X - \mu\mathbf{1})(X - \mu\mathbf{1})^T = (1/N)U\Lambda U^T \tag{4.31}$$

μ is the observations mean, $\mathbf{1}$ is a row N $\mathbf{1}'s$, U is composed of eigenvectors of size $n \times n$, and Λ is an $n \times n$ diagonal matrix of eigenvalues.

The covariance matrix of the data samples $N \times n$ is of size $n \times n$, it is often supposed that only those eigenvectors that correspond to p largest eigenvalues are of interest, the others are discarded by deleting columns from the eigenvector matrix.

In these experiments, 38 landmark points around the mouth for each image were used giving $n = 76$. A set of 5982 data samples was used giving $N = 5982$. The covariance matrix will be of size 76×76, while it will become 38×38 when we use 19 landmark points. In these experiments 6 eigenvectors were kept to retain a 99% of energy. The resulting eigenvectors matrices will be of size 76×6 and 38×6 respectively. It can be found that increasing the size of the landmark points to 38 have little effect on computer memory. Therefore, we conclude that we can gain considerable accuracy when increasing the number of landmark points around the mouth with trivial memory problems. Refer to Chapter 6 for more details on the EVD method.

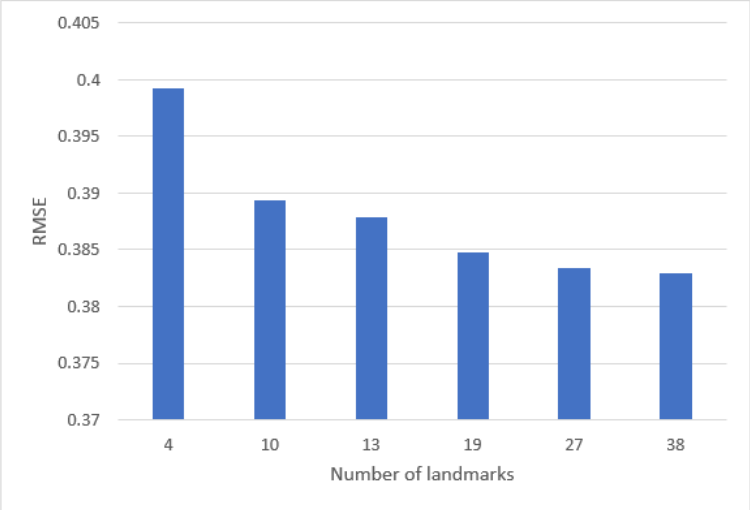


Figure 4.6. RMSE in shape normalised images compared to the ground truth lip images against different number of landmarks.

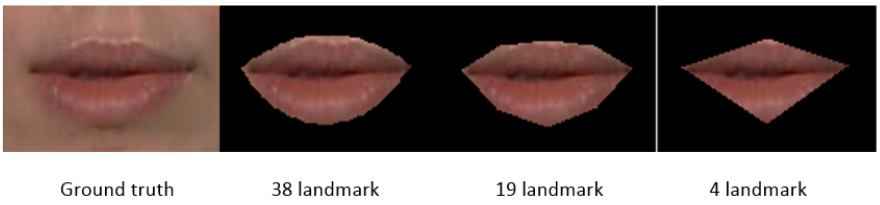


Figure 4.7. Ground truth mouth and several shape normalised mouth images obtained using different number of landmarks.

4.10.3 Experiment 3: Building AAM on different number of images

The aim of the experiments is to investigate that building active appearance model (AAM) on larger dataset can improve its accuracy. In our experiments we built AAMs using 1000, 2000, 3000, 4000 and 5000 prototype images. A non-overlapping subset of 20 sequences totalling 2756 prototype images were chosen and then projected to AAMs

parameters. Figures 4.8 and 4.9 show the bar charts of RMSE and correlation coefficient (CC) between ground truth and synthesised AAM features. These figures show that building AAMs on larger dataset gives better results.

In our experiments low-dimensional approaches are utilised to determine eigenspace models and are necessary when the dimensionality of the dataset is very large compared to their number. Therefore, they can be utilised to determine eigenspace models that would otherwise be inappropriate. Determining an eigenspace model needs that we built an $n \times n$ matrix, where n is the dimension of each sample in the dataset. In practice the model could be determined by utilising an $N \times N$ matrix, where N is the number of samples. This method is efficient in applications such as, image processing where the number of samples is less than the number of dimensions in each image $N << n$. The covariance matrix of the 5000 texture vectors of dimension 318753 is of size 318753×318753. Since $N << n$ we use the inner product approach [157]. Typically, only p of the n eigenvectors required to be kept where p eigenvalues are significant. To keep the p largest eigenvectors we set out $p = 100$ as a specified integer and thus retain the 100 largest eigenvectors. Constructing eigenspace for texture dataset 5000×318753 or more is difficult because storing such matrix of pixels in memory is intractable and need a machine with high RAM capacity. Therefore, we investigate incremental approaches in Chapter 6 which not require all observations at once thus decreasing storage requirements and making large problems computationally simpler. Refer to Chapter 6 for more details on the EVD method.

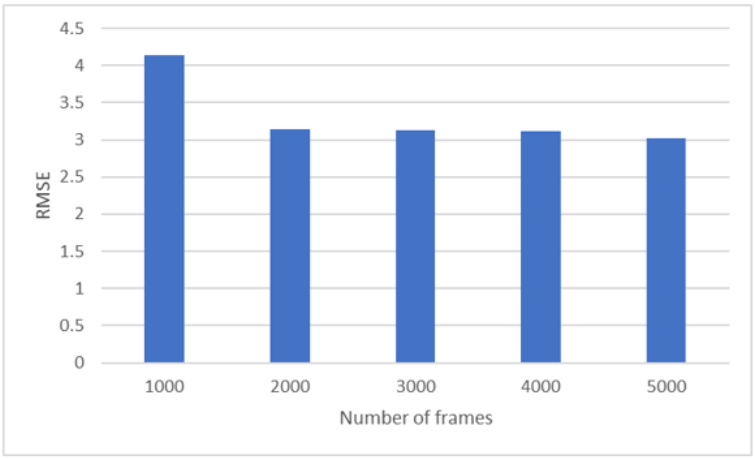


Figure 4.8. RMSE between ground truth and synthesised AAM features against different number of prototype images.

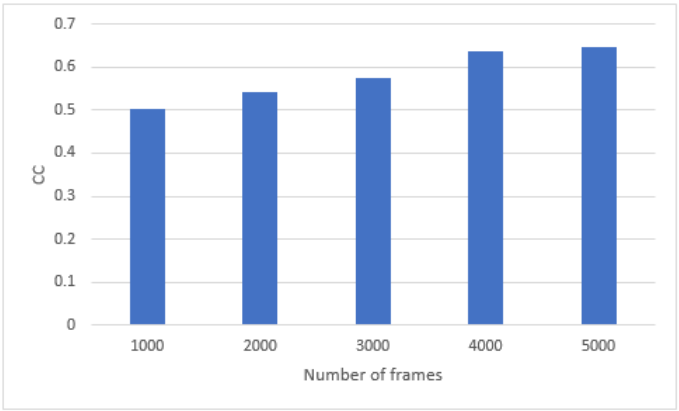


Figure 4.9. CC between ground truth and synthesised AAM features against different number of prototype images.

4.11 Limitations of the SGPLVM method

The SGPLVM is intractable for large training sets. For a training set of N frames, the SGPLVM has $O(N^3)$ space complexity. Without using sparse approximations, the time complexity for each iteration of the training and inference algorithm is $O(N^3)$, and it becomes $O(k^2N)$

when using sparse approximations, where k is a support points. Although using sparse approximations, the SGPLVM comes with a high computational complexity. In addition, the SGPLVM has a large number of free parameters that need to be adjusted to obtain the optimal parameters.

4.12 Summary

This chapter has presented a review of various probabilistic models utilised in speech animation, using the structure of graphical models. The SGPLVM was described, which allows the coupling of two data spaces. In addition, experiments were performed to examine the hypothesis of increasing the number of landmark points can increase the accuracy of the AAM. Moreover, experiments were conducted to examine that building an AAM on larger dataset can improve the accuracy. The quantitative results show that modelling an AAM on large number of visual data and using more landmarks around the mouth give the best results and higher correlation with ground truth.

In this work, the performance of the SGPLVM that is used to jointly model the auditory and visual features was improved using a more accurate AAM. An AAM was constructed on a large number of video frames with more landmark points for each frame. Objective results were introduced for the SGPLVM. Our objective results revealed that our model performs better than comparable method of visual speech synthesis.

Chapter 5

MANIFOLD RELEVANCE DETERMINATION FOR AUDIO VISUAL MAPPING BASED ON ACTIVE SHAPE MODEL

To abtain a successful speech driven facial animation system, the synthesised visual speech has to be smooth and stay within the limits allowed by the facial articulators. In this and the next chapter we continue addressing the problem of producing accurate and realistic videos of talking faces using audio signal as input. Here we hypothesise that we can improve the accuracy of produced videos of talking faces if we use manifold relevance determination for audio visual mapping instead of shared Gaussian process latent variable model (SGPLVM). Manifold relevance determination (MRD) has not been used previously for generating videos of talking faces. Here, MRD is used to represent audio and visual data as a set of factorised latent spaces. Objective evaluation is presented for MRD and compared results by SGPLVM [38]. In

this chapter, we shall refer to ground truth as the features or videos corresponding to real visual speech sequences parameterised with the active shape model (ASM). Some of the work in this chapter appeared in Dawood et al. [35]. An overview of the proposed approach for visual speech synthesis is illustrated in Figure 5.1.

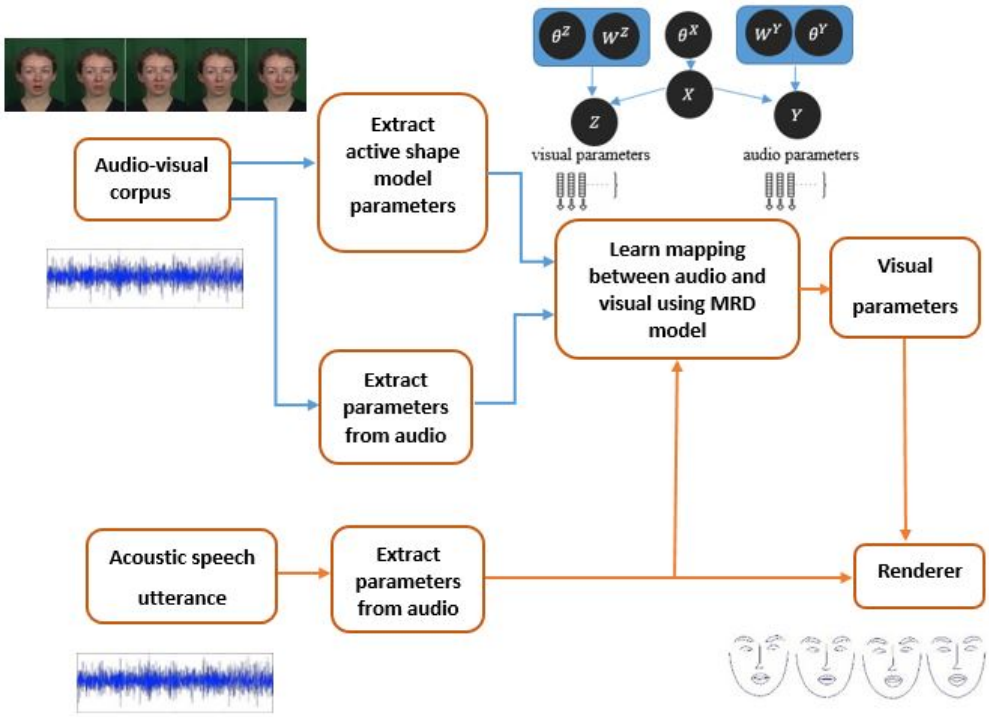


Figure 5.1. An overview of the proposed approach for visual speech synthesis. Training process is marked using the blue arrows, and synthesis process is marked by the orange arrows.

5.1 Our proposed model

Deena et al. [38] introduced a framework that jointly models auditory and visual features using a shared latent points. To cater for the various dynamics involved in speech, they augmented the model with switching states by training a variable-length Markov model [149] on

phonetic labels. In this work, we present a new generative method for speech-driven facial animation using a joint probabilistic model of audio and visual features, which explicitly models coarticulation. The proposed framework to jointly model speech and visual features is based on Bayesian techniques. The latent variable is factorised to represent private and shared information from audio and visual features. To obtain a smooth, continuous representation, a relaxation of the structural factorisation of the model is introduced, where a latent variable might be more important to the shared space than the private space. In contrast to previous methods, using this model allows the dimensionality of the latent space to be estimated automatically. MRD is a powerful and flexible approach to capture structure within very high dimensional spaces [34], it models raw images with many thousands of pixels. In addition, this method has been applied successfully in several multiple-views tasks, such as human pose prediction in an ambiguous setting. The disambiguation is performed by including latent point priors, which combine the dynamic nature of the data. In this section, we describe the model and the variational approximation.

Two observation spaces of a dataset, $Y \in \mathbb{R}^{N \times D_Y}$ and $Z \in \mathbb{R}^{N \times D_Z}$, are assumed to be generated from a single latent point $X \in \mathbb{R}^{N \times Q}$ through the non-linear mappings $(f_1^Y, ..., f_{D_Y}^Y)$ and $(f_1^Z, ..., f_{D_Z}^Z)$ $(Q < D)$, giving a low-dimensional representation of the data. The assumption is that the observation is generated from a low-dimensional manifold and corrupted with Gaussian distributed observation noise $\epsilon^{\{Y,Z\}} \sim \mathcal{N}(0, \sigma_\epsilon^{\{Y,Z\}} I)$:

$$y_{nd} = f_d^Y(\mathbf{x}_n) + \epsilon_{nd}^Y \tag{5.1}$$

$$z_{nd} = f_d^Z(\mathbf{x}_n) + \epsilon_{nd}^Z \tag{5.2}$$

where nd represents the dimension d of point n. This leads to the joint likelihood under the model, $P(Y, Z|X, \theta)$, where $\theta = \{\theta^Y, \theta^Z\}$, representing two different sets of hyper-parameters of the mapping functions and the noise variances σ_ϵ^Y and σ_ϵ^Z. A GP prior distribution is proposed to place over the mappings [100]; the resulting models are known as Gaussian process latent variable models (GPLVMs). In the GPLVM approach, each generative mapping is modeled as a product of D separate GPs parameterised by a covariance function $k^{\{Y,Z\}}$ evaluated over the latent points X:

$$p\left(F^Y|X, \theta^Y\right) = \prod_{d=1}^{D_Y} \mathcal{N}\left(\mathbf{f}_d^Y|0, K^Y\right) \tag{5.3}$$

where $F^Y = (f_1^Y, ..., f_{D_Y}^Y)$ with $f_{nd}^Y = f_d^Y(\mathbf{x}_n)$, and the same definitions for F^Z. The non-linear mapping can be marginalised out analytically, obtaining a joint likelihood:

$$P\left(Y, Z|X, \theta\right) = \prod_{\mathcal{K}=\{Y,Z\}} \int p\left(\mathcal{K}|F^{\mathcal{K}}\right) p\left(F^{\mathcal{K}}|X, \theta^{\mathcal{K}}\right) dF^{\mathcal{K}} \tag{5.4}$$

To obtain a fully Bayesian treatment, integration over the latent representation X is required. This is intractable, because X appears non-linearly in the inverse of the covariance matrices $\{K^Y, K^Z\}$ of the GP priors over the mapping $\{f^Y, f^Z\}$. By variationally marginalising out X, an approximated Bayesian training and synthesis procedure can be obtained. The automatic relevance determination (ARD) priors can

then be introduced, such that each observation is allowed to estimate a separate vector of ARD parameters. In this case, the observations are allowed to set the private and shared latent subspaces relevant to them.

In the following, we can summarise how MRD has its origins in principal component analysis (PCA),

- GPLVMs can be acquired after generalising PCA to be nonlinear and probabilistic.
- A SGPLVMs can be acquired when GPLVMs are generalised to the case of multiple views of the information.
- When private subspaces to shared GPLVMs are found, then a factorisation of the latent representation that encodes variance specific to each view is recovered.
- Lastly, the dimensionality and factorisation of the latent variable can be automatically determined when the latent space is marginalised instead of optimised.

Automatic determination of the dimensionality and structure of the nonlinear latent variable from multiple views can be known as manifold relevance determination.

5.1.1 Manifold Relevance Determination

Damianou et al. [34] tried to improve factorised latent spaces so that the variance shared (i.e. correlated) between different data spaces can be aligned and disjointed from variance that is private (i.e. independent). In this model, the variance contained in the data space does not need

to be governed by geometrically orthogonal subspace, as supposed in [154]. The manifold model has the ability to treat non-linear mappings within a Bayesian approach. In particular, the latent functions f_d^y that are selected to be separate draws with a zero-mean GP, and ARD covariance function, which is given by:

$$k^Y(\mathbf{x}_i, \mathbf{x}_j) = \left(\sigma_{ard}^Y\right)^2 e^{-\frac{1}{2}\sum_{q=1}^{Q} \mathbf{w}_q^Y (x_{i,q} - x_{j,q})^2} \tag{5.5}$$

and analogously for f^Z. A common latent space can be learned; however, the two groups of ARD weights $\mathbf{w}^Y = \left\{w_q^Y\right\}_{q=1}^Q$ and $\mathbf{w}^Z = \left\{w_q^Z\right\}_{q=1}^Q$ are allowed to automatically infer the responsibility of every latent dimension to produce points in the Y and Z spaces, respectively. After that, the segmentation of the latent points $X = \left\{X^Y X^s X^Z\right\}$ can be recovered, where X^Y and X^Z are private spaces, $X^s \in \mathbb{R}^{N \times Q_s}$ is a shared space defined by a group of dimensions $q \in [1, ..., Q]$ and $w_q^Y, w_q^Z > \delta$ with δ is a number near to zero, $Q_s \leq Q$. This allows for softly shared latent points, if the two sets of weights together are greater than δ and they are different. The two subspaces X^Y and X^Z are inferred automatically:

$$X^Y = \{\mathbf{x}_q\}_{q=1}^{Q_Y} \tag{5.6}$$

$$X^Z = \{\mathbf{x}_q\}_{q=1}^{Q_Z} \tag{5.7}$$

where Q_Y and Q_Z are the dimensionality of X^Y and X^Z respectively, $\mathbf{x}_q \in X, w_q^Y > \delta, w_q^Z < \delta$. Figure 5.2 shows the graphical model of MRD. In this figure, the ARD weights $\mathbf{w}^{\{Y,Z\}}$ are separated from the full set of model hyper-parameters $\theta^{\{Y,Z\}} = \left\{\sigma_{\in}^{\{Y,Z\}}, \sigma_{ard}^{\{Y,Z\}}, \mathbf{w}^{\{Y,Z\}}\right\}$ to describe

the utilisation of ARD covariance functions. The latent variable X is marginalised out and a distribution of latent points is learned for which additional hyperparamters encode the relevance of each dimension independently for the observation spaces, then, a factorisation of the data is automatically defined. The distribution $p(X) = p\left(X|\theta^X\right)$ placed on the latent variable allows the incorporation of prior knowledge about its structure. Despite the similarity with the graphical model of a fully shared latent space, in this figure the role of X is totally different. The latent space X is marginalised out and, both with the additional weight parameters, runs in a Bayesian factorised model.

Bayesian training

The fully Bayesian training technique needs maximisation of the logarithm of the joint marginal likelihood

$$p(Y, Z|\theta) = \int p(Y, Z|X, \theta)\, p(X) dX \tag{5.8}$$

where a prior distribution is located on X. This prior might be a standard normal distribution or might rely on a group of parameters θ^X. From equation 5.4 it can be seen that the integral is intractable because of the nonlinear way in which X represents in $p\left(F^{\{Y,Z\}}|X, \theta^{\{Y,Z\}}\right)$. In this situation standard variational approximations are intractable as well. A non-standard approach will be reported here to obtain an analytic solution.

Damianou et al. [34] tried to maximise a variational lower bound $F_v(q, \theta)$ on the logarithm of the true marginal likelihood depending on a variational distribution which factorises as $q(\Theta)q(X)$, where $q(X) \sim \mathcal{N}(\mu, S)$ can be assumed. In this method $q(\Theta)$ is a distribution which

relies on additional variational parameters $\Theta = \left\{\Theta^Y, \Theta^Z\right\}$ such that $q(\Theta) = q(\Theta^Y)q(\Theta^Z)$. These additional parameters Θ and the exact form of $q(\Theta)$ form the most crucial ingredient of non-standard variational method. For simplicity, hyperparameters θ is dropped from the expressions and Jensens inequality is used to obtain a variational bound $F_v(q) \leq \log p(Y, Z)$:

$$\begin{aligned} F_v(q) &= \int q(\Theta)q(X)log\left(\frac{p(Y|X)p(Z|X)p(X)}{q(\Theta)q(X)}\right)dX \\ &= \mathcal{L}_Y + \mathcal{L}_Z - \mathbf{KL}\left[q(X)||p(X)\right] \end{aligned} \tag{5.9}$$

where $\mathcal{L}_Y = \int q(\Theta^Y q(X)log\frac{p(Y|X)}{q(\Theta^Y)}dX$ and analogously for $\mathcal{L}_Z$. "data augmentation" principle is applied to expand the joint probability space with M extra samples U^Y and U^Z of the latent functions f^Y and f^Z estimated at a group of pseudo-inputs namely "inducing points" $\bar{X}^Y$ and $\bar{X}^Z$ respectively. $U^Y \in \mathbb{R}^{M_Y \times D_Y}$, $U^Z \in \mathbb{R}^{M_Z \times D_Z}$, $\bar{X}^Y \in \mathbb{R}^{M_Y \times Q}$, $\bar{X}^Z \in \mathbb{R}^{M_Z \times Q}$ and $M = M_Y + M_Z$. The expression of the joint probability becomes:

$$p(Y|X, \bar{X}^Y) = \int p(Y|F^Y)p(F^Y|U^Y, X, \bar{X}^Y)p(U^Y|\bar{X}^Y)dF^Y dU^Y \tag{5.10}$$

and analogously for $p(Z|X)$. It can be seen that the inducing points are variational rather than model parameters. Now, $q(\Theta) = q(\Theta^Y)q(\Theta^Z)$ can be represented as

$$q(\Theta) = \prod_{\mathcal{K}=\{Y,Z\}} q(U^{\mathcal{K}})p(F^{\mathcal{K}}|U^{\mathcal{K}}, X, \bar{X}^{\mathcal{K}}) \tag{5.11}$$

where $q(U^{\mathcal{K}})$ are free from distributions. The final objective function can be obtained by replacing equations 5.11 and 5.10 back to 5.9. This

function is jointly maximised with respect to the model parameters, including the latent space weights $\mathbf{w}^Y$ and $\mathbf{w}^Z$ and the variational parameters $\left\{\mu, S, \bar{X}\right\}$. This incorporates additional strength to the model, since previous methods depend on maximum-a-posteriori (MAP) approximations for the latent points.

Dynamics

For the dynamic version, the model can represent the correlations between datapoints of the same output space, for instance when Y and Z are multivariate time-series. In this case the prior on the latent representation is chosen to depend on the observation times $t \in \mathbb{R}^N$, for example a GP with a covariance function $k = k(t, t^{'})$. As in standard variational inference, this optimisation gives an approximation of $p(X|Y, Z)$ by $q(X)$, in other words, a distribution over the latent space is obtained.

Damianou et al. [34] consider a collection of human poses and associated silhouettes. They utilised the MRD model to synthesis the poses corresponding to the test silhouette features. This is a challenging because the data are multi-modal. The silhouette features might be created from more than one pose. Damianou et al. compare the method with the SGPLVM [50, 51] which optimises the latent space utilising MAP. They showed that the MRD performs better than the SGPLVM method in the task of predicting human pose in an ambiguous setting.

Our novel method for audio-visual mapping is based on the MRD framework. Utilising a softly shared latent space, the non-linear relationship between auditory and visual dynamics during speech can be

automatically modelled. MAP estimates utilised in SGPLVM model mean that the structure of the latent space cannot be automatically calculated. For the MRD model, ARD priors are introduced to calculate which of the emerging private latent space and shared latent space are relevant to the views. The model learns a distribution over the latent space variationally which allows the dimensionality of the latent representation automatically and incorporate prior knowledge about the structure to be determined. The model forces the whole collection of training and test inputs to generate smooth paths in the latent space.

5.1.2 Training

MRD is learned between Y, represented by the relative spectral transform-perceptual linear prediction (RASTA-PLP) feature vector, and Z, represented by the ASM feature vector. The obtained latent space is a non-linear embedding of both audio and visual features that can generate the two spaces Y and Z. Probabilistic principal component analysis (PPCA) is used as an initialisation of the latent space variational means; this is done by performing PPCA on each dataset separately and then concatenating the two low dimensional representations to initialise X. We perform experiments to compare this method with Deena's SGPLVM approach in Section 5.4.

5.1.3 Inference

Given a trained model that jointly represents the audio features Y and the ASM parameters Z with a single but factorised input space X, we wish to infer a new set of output points $Z^* \in \mathbb{R}^{N^* \times D_Z}$ given a set of test points $Y^* \in \mathbb{R}^{N^* \times D_Y}$. The inference procedure is done in

three steps; first, the sequence of latent points $X^* \in \mathbb{R}^{N^* \times Q}$ that is most likely to have inferred Y^* is predicted. An approximation to the posterior $p(X^*|Y^*, Y)$, which has the same form as for the standard Bayesian GPLVM is used [176] and given by a variational distribution $q(X, X^*)$. In order to find $q(X, X^*)$, the variational lower bound on the marginal likelihood $p(Y, Y^*)$ which has analogous form with the function 5.9 is optimised. Precisely, Z is ignored and Y is replaced with (Y, Y^*) and X with (X, X^*). Second, the training latent points X_{NN} that are nearest to X^* in the shared latent representation are found. Finally, the output sequence Z from the likelihood $p(Z|X_{NN})$ is determined. To infer novel outputs, the recovered information has to propagate. Algorithm 1 summarise the MRD inference procedures.

Algorithm 1 MRD Inference Algorithm, assuming two views audio (Y) and visual (Z)

1. Given a trained MRD model on views (Y, Z) representing audio and visual features respectively to obtain factorised latent space $X = \left(X^Y, X^{YZ}, X^Z\right)$.
2. Given test point from audio features y^*.
3. Do optimisation $q(X, x^*) \approx p(X, x^*|Y, y^*)$.
4. The marginal $q(x^*)$ with mean $x^* = (x^{*Y}, x^{*YZ}, x^{*Z})$ is obtained.
5. K points $x_{NN(k)}$ from a $K-$nearest neigbour search between x^{*YZ} and X^{YZ} is found.
6. **for** $k = 1,\ldots,K$ **do**
7. $x^*_{NN(k)} = (x^{*Y}, x^{*YZ}, x^Z_{NN(k)})$ is generated by joining $x_{NN(k)}$ and x^*.
8. $z^*_{(k)}$ is generated from the likelihood $p(z|x^*_{NN(k)})$.
9. All test audio points Y^* are treated together using dynamical MRD version, such that the variational distribution $q(X, X^*)$ will form a timeseries.

5.2 Applications of MRD

Following is a potential application suggested by other researchers for MRD method:

The work of Bekiroglu et al. [7] applied MRD to the problem of transferring between stable and unstable robot grasps. Bekiroglu et al. used

MRD for correcting unstable robot grasps to stable robot grasps. The key characteristic of this method is the utilisation of a factorised latent space which individually models the data that is shared between the views. The factorisation permits to achieve efficient inference in ambiguous setting where the observations are not sufficient to select the required output.

Damianou et al. [34] applied MRD method to the Yale dataset [66, 104] which contains images of human faces under various poses and 64 illumination conditions to model very high-dimensional spaces. A single pose for each subject was considered so that the only variations were the location of the light source and the subjects appearance. The model was directly applied to the raw pixel values such that image feature extraction to process the data was not needed, and novel outputs can be directly sampled. The information about the position of the light source and not the face characteristics was encoded successfully by the shared space.

The work of Trautman [177] presents possible applications of MRD in sensor fusion, multi-agent SLAM, and "human-appropriate" robot movement. In his work Trautman show how MRD can be utilised to construct the underlying models in a data driven manner, instead of directly using first principles theories for example physics and psychology.

5.3 Data and pre-processing

We use a phonetically balanced LIPS corpus [171] consisting of 278 high-quality sequences featuring a female British subject speaking sentences from the Messiah corpus [169]. The sentences were spoken in a

neutral speaking style (no expression). The original LIPS corpus consists of images of 50 Hz video stream with 576x720 pixels. Figure 5.3 shows video frames from the LIPS corpus. In addition, high-quality audio in the form of WAV files and the phonetic annotation for each frame have been made available. Using the British English Example Pronunciation Dictionary (BEEP), the LIPS dataset has been phonetically aligned. The DEMNOW dataset [55] has also been used which is closer to natural speech, as this dataset has been acquired in a real world setting of a newsroom.

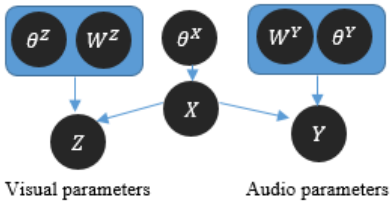


Figure 5.2. Graphical model of the MRD method.

Figure 5.3. Video frames from the LIPS dataset.

5.3.1 Audio proccessing

Deena et al. [39] performed experiments to determine which speech parameterisation technique out of linear predictive coding (LPC), line spectral frequencies (LSF), Mel-frequency cepstral coefficient (MFCC) and RASTA-PLP is better for predicting visual features. They found

that RASTA-PLP processed at 25 Hz is a better predictor of visual features for LIPS, whilst MFCC coefficient downsampled using polyphase quadrature filtering gives the best results for the DEMNOW dataset [54]. In this work, we used RASTA-PLP features to parameterise speech to represent the acoustic variability within and between the different phonemes. To satisfy the requirement of having a window where the speech signal is stationary, a window size of 25 ms and a hop size of 10 ms is typically used, resulting in an audio processing frequency of 100 Hz. The speech parameters need to be downsampled to match the visual processing rate of 25 fps used for LIPS video sequences, so we use an auditory window of 50 ms and a hop window of 40 ms to obtain speech features at 25 Hz. In addition, we use 20 parameters to represent the RASTA-PLP features. The speech features required to be downsampled to match the visual processing rate 29.97 fps utilised for the DEMNOW dataset. Therefore, we use an auditory window of 50 ms and a hop window of 33 ms to get the speech features at 29.97 Hz.

5.3.2 Visual processing

We use an ASM [25] for visual parameterisation, because such models capture the statistical variation in shape and build a generative model to obtain novel shapes. A training set of annotated prototype face images is required. For the LIPS dataset we use 97 landmark points around the face, eyebrows, lips and nose in each of the prototype images as shown in Figuer 3.1.

For the DEMNOW dataset, 12922 images corresponding to 70 sequences were annotated automatically with 68 facial landmarks iden-

tified for each frame; 20 of them described the inner and outer mouth shape using open source tool (OpenFace). Figure 5.4 shows an image from the DEMNOW dataset with landmark points.

An ASM has been built on the shapes in several steps. First, the shape vectors have been normalised by removing rotations and translations, and then aligned with respect to the mean shape using Procrustes analysis. Following this, PCA has been applied to the normalised shape vectors. After training the PCA model and retaining 95% of the variance of the shape, ASM parameters can be obtained from novel shapes by projecting the shape vectors to the corresponding retained eigenvectors.

Figure 5.4. A labeled training image from the DEMNOW dataset.

5.3.3 Computational complexity

GP models are intractable for large datasets and have a time complexity scales and storage of $O(N^3)$ and $O(N^2)$ respectively, where N is the number of training examples. To overcome this limitation, several approximation approaches have been described in the literature to construct a sparsification dependent on a small set of M inducing points to reduce the typical time complexity from $O(N^3)$ to $O(NM^2)$ [175],

where N and M are the total numbers of training and inducing variables, respectively. In the MRD model [34], the datasets with very large numbers of features can be modelled because the objective function involves the matrices Y and Z in expressions of the form YY^T and ZZ^T, which illustrates that the model does not rely on the number of features $\{D_Y, D_Z\}$ in the datasets.

5.4 Experiments on auditory and visual signal

MRD and SGPLVM models are trained on ASM and RASTA-PLP data for both LIPS and DEMNOW datasets because we want to assess the approaches using two different datasets. The LIPS dataset is phonetically balanced whereas the DEMNOW dataset is closer to natural speech, as this dataset has been obtained from a real world setting of a newsroom. In Subsection 5.4.1 experiments are presented for training MRD and SGPLVM methods using LIPS shapes for visual representation and RASTA-PLP features for audio representation. In Subsection 5.4.2 experiments are performed using the same methods for audio-visual mapping with DEMNOW shapes for visual and RASTA-PLP features for audio representations respectively. The results of MRD are compared with these of the SGPLVM.

5.4.1 Experiment 1: Quantitative evaluation for LIPS dataset

Due to the complexity of MRD training, we trained an MRD model on 50 training sequences from LIPS dataset, totaling 5332 frames, by taking Y as the RASTA-PLP features and Z as the mean-centering ASM features; the obtained latent space was represented by six dimensions. In our experiment, we set the inducing points to 100. We then

used a validation set of 10 sequences totalling 1234 frames to predict visual parameters from audio parameters. As illustrated in the inference subsection 5.1.3, given test point y^*, one of the N^* audio tests, the model optimised a variational distribution and found a sequence of K candidate initial training data $\left(x_{NN}^{(1)}, ..., x_{NN}^{(K)}\right)$; these were ordered according to their similarity to x^*, and only the shared dimensions were taken into account. Based on these initial latent points, a sorted series of K novel visual features $(z^1, ..., z^K)$ were found. In our experiments, we performed PPCA on each dataset separately and then concatenated the two low-dimensional representations to initialise X. The results of the MRD were compared against the SGPLVM method [36]. Table 5.1 shows the average mean squared error (AMSE) and average correlation coefficient (ACC) for our approach using MRD and Deena et al's method [36]. The table demonstrates that there was a distinction between the errors obtained using MRD and those using Deena's approach (SGPLVM). In addition, Figure 5.5 shows the shape frames obtained from the ground truth, MRD and SGPLVM. The corresponding audio contained a sentence from the LIPS dataset. We found that the shape uttering /b/ from the MRD method showed proper lip synchronisation with the audio and appeared to be the best, whilst the SGPLVM gave lip synchronisation with a few jerks in the animation. It can be seen that the difference between the quality of mouth articulation between real and MRD synthetic videos was non-significant. In addition, we observe smooth lip movements compared with the ground truth and the SGPLVM methods. Figure 5.6 shows the shape frames obtained from the ground truth, MRD and the SGPLVM. The phonemes correspond to six different visemes of the words ("house", "had") from

the test audio sentence (The big house had large windows with very thick frames). It can be found that the shapes uttering /aw/ from MRD and the SGPLVM methods showed proper lip synchronisation with the auditory and the differences of the quality of mouth articulation among real and both of MRD, the SGPLVM synthetic videos were non-significant. Also, it can be observed smooth lip movements of MRD synthetic videos compared with the ground truth and SGPLVM approaches. Figure 5.7 shows the results normalised between 0 and 1 obtained across the ten runs of the experiment. The results show a noteworthy difference between the errors obtained from MRD and Deena's method. Generally, the AMSE errors for MRD are distinctly lower than those for the SGPLVM, mostly due to a softly shared latent space. In addition, the results are in higher correlation with the ground truth as compared to other method.

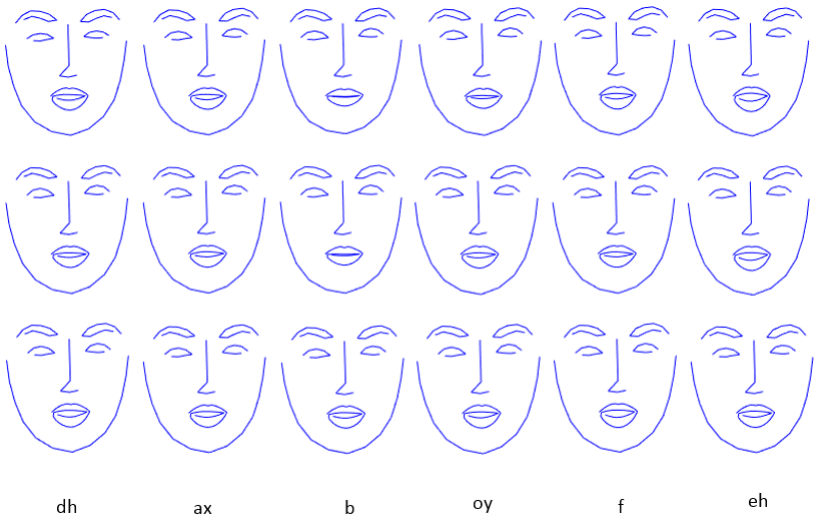


Figure 5.5. Example frames of the shapes obtained from the LIPS synthesis results using ground truth (top row), MRD (middle row) and SGPLVM (bottom row). The phonemes correspond to six different visemes of the words ("the", "boy", "fair") from the test audio sentence (The boy a few yards ahead is the only fair winner).

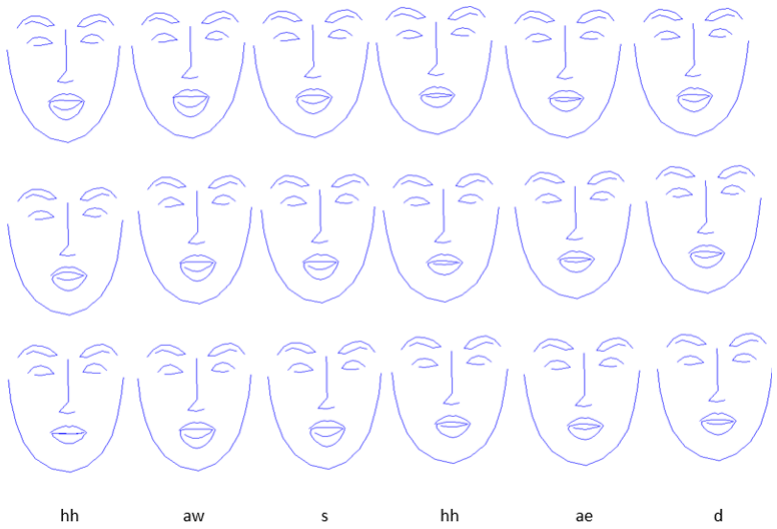


Figure 5.6. Example frames of the shapes obtained from the LIPS synthesis results using ground truth (top row), MRD (middle row) and SGPLVM (bottom row). The phonemes correspond to six different visemes of the words ("house", "had") from the test audio sentence (The big house had large windows with very thick frames).

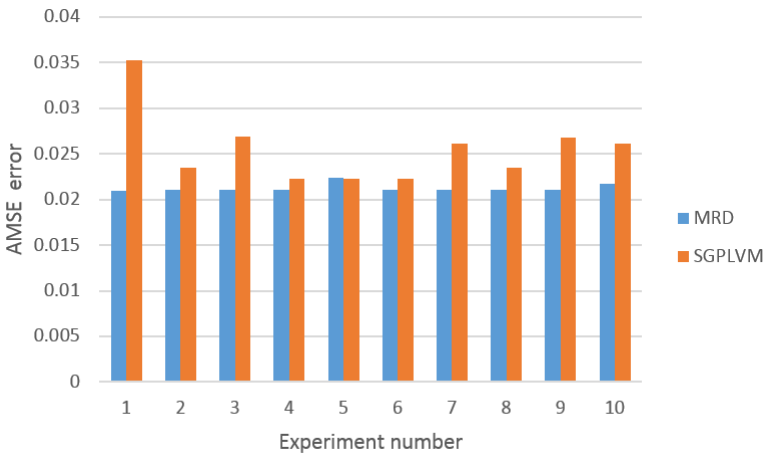


Figure 5.7. AMSE errors obtained between ground truth ASM feature vectors and 1- MRD 2- SGPLVM.

Table 5.1. Objective measure computed between original ASM parameters and the corresponding synthesised parameters using LIPS dataset.

Method	Audio representation	Visual representation	AMSE	ACC
MRD	Continuous	ASM	54.8216	0.7095
SGPLVM	Continuous	ASM	78.7457	0.7071

5.4.2 Experiment 2: Quantitative evaluation for DEMNOW dataset

In this experiment we asses the proposed approaches on another dataset which is closer to natural speech, as this dataset has been acquired in a real world setting of a newsroom.

For the DEMNOW dataset ASMs are built on 50 training sequences, totaling 9199 frames. By retaining 99% of the variance of the shape a 6 dimensional vector of ASM parameters is obtained. The speech features need to be downsampled to match the visual processing rate 29.97 fps utilised for the DEMNOW dataset. Therefore, we utilise an auditory window of 50 ms and a hop window of 33 ms to get the speech features at 29.97 Hz.

For the DEMNOW dataset MRD and SGPLVM models are trained on 50 training sequences, totaling 9199 frames. Then, a validation set of 20 sequences totaling 3723 frames is used to predict visual parameters from auditory. The models are learned between RASTA-PLP feature vector and the ASM feature vector. Table 5.2 shows the AMSE and ACC for our approach and Deena's method using the DEMNOW dataset. Again, as for the LIPS dataset, the results show that the MRD model gives better results. The trajectories of the first mouth landmark (y-coordinate) feature reconstructed from the active appearance model (AAM) features of the two methods against ground truth

are also compared as shown in Figure 5.8. The synthetic trajectories are calculated utilising three test utterances of DEMNOW dataset. The results for MRD approach show that the trajectories are smoothed out as compared to the SGPLVM method.

The DEMNOW corpus contains more variability than the LIPS dataset. These dataset contains large variations in pose and expression within a particular sequence as well as the variability across sequences. Moreover, the DEMNOW dataset includes an American speaker presenting news with a fast speaking rate. In our experiments, z-score AAM parameters normalisation and RASTA-PLP parameters method are used for DEMNOW dataset. Other normalisation and speech parameterisation methods could be used to address the issues of large variability and hyper-articulated that is happened because of the fast speaking rate.

Table 5.2. Objective measure computed between original ASM parameters and the corresponding synthesised parameters using the DEMNOW dataset.

Method	Audio representation	Visual representation	AMSE	ACC
MRD	Continuous	ASM	22.5922	0.6701
SGPLVM	Continuous	ASM	37.2289	0.6398

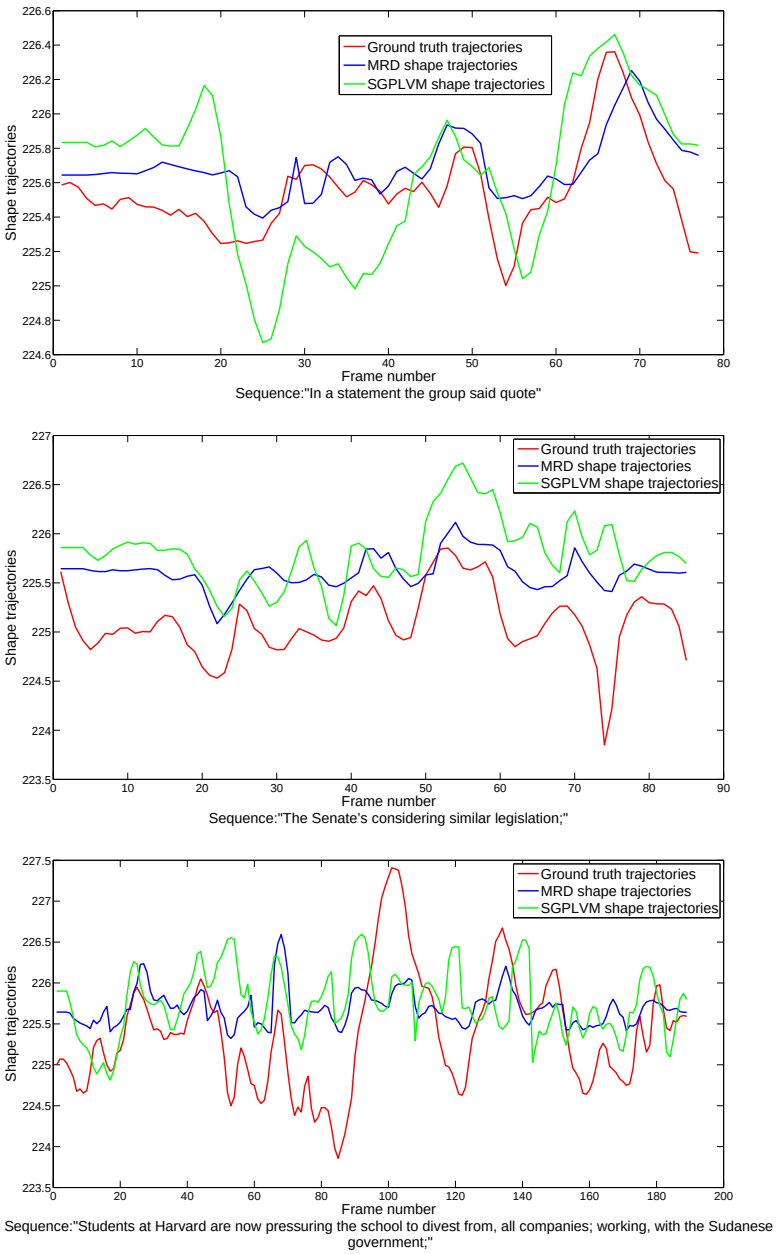


Figure 5.8. Sample shape trajectories obtained from MRD, SGPLVM and the corresponding ground truth trajectories.

More experiments and analysis on MRD are presented in Chapter 6.

5.5 Summary

This chapter presented a new factorised latent variable model for synthesising visual speech from auditory signals. We showed how MRD can be utilised for auditory-visual mapping. A compact and intuitive representation of audio-visual data was learned, represented by the synthesis of novel shapes by sampling from the latent space in a structured manner. Objective results were presented for both MRD and the SGPLVM. We then presented training and synthesis approach for MRD. Experiments were finally dealt with. Two datasets were utilised in our experiments the first one was LIPS dataset [171] and the second was DEMNOW dataset [55]. It can be found that using MRD method to modelling auditory and visual features decreased the AMSE error of the resulting animation compared to the SGPLVM approach for both the LIPS and DEMNOW datasets. The results were in higher correlation with the ground truth as compared to other method. In addition, synthesis of facial animation using MRD produced visuals with the correct facial dynamics and proper synchronisation with the audio signal.

Chapter 6

THE MANIFOLD RELEVANCE DETERMINATION FOR AUDIO VISUAL MAPPING BASED ON APPEARANCE FACIAL MODEL

We have seen in the previous chapter how flexible the manifold relevance determination (MRD) approach is at using a shape model as a representation of the face. In this chapter, we continue addressing the problem of producing accurate and realistic videos of talking faces from audio signal. A generative model of the face that captures the shape and texture variation is used for training the MRD model. In Chapter 4 we showed that building active appearance models (AAMs) using larger dataset and more facial landmarks leads to more precise models. However, batch learning includes processing of the whole dataset which is restricted to its applications as batch learning methods are more time consuming and needs the whole dataset prior to training. This means that the computers with low resources cannot be used for building such

models from all data simultaneously.

Our motivation for this chapter arose in the context of constructing eigenspace models for many images. Nonetheless, method for incremental learning of models exists [76,77]. Furthermore, in online learning algorithm, AAM parameters can be updated when a new dataset arrives. It can be shown in this chapter that such method produces models almost as good as the ones trained on all the data simultaneously.

An eigenspace model includes the number of observations, their mean, the support vectors (eigenvectors) over the observations, and a measure of the spread of the observations (eigenvalues) through each support vector. Eigenspace models can be computed utilising either eigenvalue decomposition (EVD) of the covariance matrix of the data (also named as principal component analysis) or singular-value decomposition (SVD) [76, 77]. In an incremental computation, an eigenspace model can be updated utilising new observations. Incremental approaches do not require all observations at once so, they decrease storage requirements and making large problems computationally tractable.

An overview of the basic method that we utilise is as follows. The first stage is to build a 2D appearance-based model of the face using a larger dataset which allows different facial poses to be represented utilising a small number of parameters. The auditory data is paramterised utilising relative spectral transform-perceptual linear prediction (RASTA-PLP). MRD is then learnt on the audio and visual data. MRD using batch and incremental approaches for visual representation are compared. Moreover, two approaches are investigated for visual normalisation, namely mean-centering of AAM parameters and z-score normalisation. Quantitative evaluation including the analysis of the

error and correlation measures between ground truth and synthetic features is performed to compare our proposed approach against other related state-of-the-art approaches. Qualitative evaluation with human volunteers has been performed to evaluate the perceptual characteristics of the synthesised animations. The proposed approach using MRD is compared with other related approaches such as the shared Gaussian process latent variable model (SGPLVM) [38], bidirectional LSTMs (BLSTMs) [60], and hidden Markov models (HMMs) [80]. In addition, experiments to determine the optimal latent space initialisation are performed.

6.1 Memory problems

Given a set of facial images, an appearance model needs to be built for these images. In this work, we try to construct a more accurate active appearance model using a larger dataset. But several problems arise when constructing this kind of models because of the size of training dataset and the subsequent effect on computer memory, and also within online learning applications. For instance, the training set in our experiments consists of 5982 texture vectors of dimension 106251 by 1. Implementing principal component analysis (PCA) on this collection is very costly in terms of memory. Moreover, the addition of colour data increase the dimension to 318753. Colour is incorporated into appearance model by determining texture vectors as concatenated Red, Green and Blue (RGB) data vectors [170]. For example, the texture of the $j-th$ face image, t_j, can be defined by a vector concatenating the RGB

value of every pixel that lies inside the mean shape:

$$t_j = (R_{j1}, R_{j2}, \cdots, R_{jU}, G_{j1}, G_{j2}, \cdots, G_{jU}, B_{j1}, B_{j2}, \cdots, B_{jU}) \quad (6.1)$$

where $j = 1, 2, \cdots\cdots, J$ and U is the total number of pixels in the face region. In our work we utilise eigenspace addition algorithm proposed by Hall et al. [77] to solve the computer memory problem and allow online learning.

6.2 Adding eigenspaces

The principle behind this approach is that eigenmodels may be added together. So, given memory constraints in a system, building and following that addition of several smaller PCA models is more efficient than building of one large model. Let $X = [\mathbf{x}_1, ..., \mathbf{x}_N]$ be a collection of N observations, each n dimensional. The EVD of the covariance of the observations is determined by

$$(X - \mu\mathbf{1})(X - \mu\mathbf{1})^T = (1/N)U\Lambda U^T \quad (6.2)$$

μ is the observations mean, $\mathbf{1}$ is a row N $\mathbf{1}'s$, U is an eigenvectors of $n \times n$ matrix, and Λ is an eigenvalues of $n \times n$ diagonal matrix.
Only the eigenvectors that correspond to large eigenvalues are of interest, the others are ignored. Typically, only $p(n, N)$ of the support vectors have considerable spread values and, consequently, only p of the n support vectors need to be retained. This reduction leaves p support vectors in a $n \times p$ matrix U_{np}, and p spread values in a diagonal matrix Λ_{pp}. After the reduction, p eigenvectors in a $n \times p$ matrix U_{np} and p

eigenvalues in a diagonal matrix Λ_{pp} can be obtained; p can be considered as the dimension of the eigenspace. Because of the reduction, we can have

$$(X-\mu)(X-\mu)^T \approx U_{np}\Lambda_{pp}U_{np}^T \tag{6.3}$$

$$U_{np}^T U_{np} = I \tag{6.4}$$

$$U_{np}U_{np}^T \neq I \tag{6.5}$$

An eigenspace model Ω can be defined as the mean, a reduced collection of eigenvectors, their eigenvalues, and the number of observations which is specified as:

$$\Omega(X) = (\mu(X), U(X)_{np}, \Lambda(X)_p, N(X)) \tag{6.6}$$

where $\mu(X)$ is the mean of data points; $U(X)_{np}$ is a p column eigenvectors; $\Lambda(X)_p$ is a vector of p eigenvalues, and N is the number of observations.

If there is another set of data points $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_M]$, the EVD eigenspace can be defined as

$$\Omega(Y) = (\mu(Y), U(Y)_{nq}, \Lambda(Y)_q, N(Y)) \tag{6.7}$$

This set generally differs from X, however this difference is not a requirement. In general $q \neq p$ even when $Y = X$, because PCA reduction may happen in different ways.

For addition utilising EVD, the eigenspace for the concatenated pair $Z = [X, Y]$ is determined

$$\Omega(Z) = (\mu(Z), U(Z)_{nr}, \Lambda(Z)_r, N(Z)) = \Omega(X) \oplus \Omega(Y) \tag{6.8}$$

Generally, the number of eigenvectors and eigenvalues r differs from both p and q. Incremental computation of $N(Z)$ and $\mu(Z)$ can be defined as follows

$$N(Z) = N(X) + N(Y) \tag{6.9}$$

$$\mu(Z) = (N(X)\mu(X) + N(Y)\mu(Y))/N(Z) \tag{6.10}$$

Eigenvectors $U(Z)$ should support all data in both sets X and Y, both $U(X)$ and $U(Y)$ should be subspaces of $U(Z)$. Generally, these subspaces might be expected "intersect" in the sense that $U(X)^T U(Y) \neq 0$. The null space of each of $U(X)$ and $U(Y)$ might contain some component of the other, $H = U(Y) - U(X)(U(X)^T U(Y)) \neq 0$. $U(Z)$ can still be of larger dimension, even if $U(X)$ and $U(Y)$ support the same subspace. This is since some component h of the vector joining the means $\mu(X) - \mu(Y)$ might be in the null space of both subspaces, simultaneously. Putting issues relating to changes in dimension, adding data acts to rotate the eigenvectors and scale the values relating to spread of the data. The new eigenvectors should be linear combination of the old. Hall et al. [77] deal with a change in dimension by building a basis sufficient span $U(Z)$, for which they used $U(X)$ augmented by v, v spans $[H, h]$, which is in the null space of $U(X)$. Note that v spans a t-dimensional subspace, $t \leq q + 1$. Then the new eigenvectors can be obtained

$$U(Z) = [U(X), v]R \tag{6.11}$$

where R is an orthonormal matrix. Addition for eigenspaces diverge

only in the manner in which R is determined.

A more detailed description of the algorithms for adding eigenspaces is given by Hall et al. [76, 77]. Algorithm 2 and Figure 6.2 summarise adding eigenspaces with eigenvalue decomposition procedures.

Algorithm 2 A sequence of instructions for adding eigenspaces.

1. Given two sets of observations,

   $X = [\mathbf{x}_1, ..., \mathbf{x}_N]$, and $Y = [\mathbf{y}_1, ..., \mathbf{y}_M]$

2. Calculate an eigenspace for X,

   $\Omega(X) = (\mu(X), U(X)_{np}, \Lambda(X)_p, N(X))$

3. Calculate an eigenspace for Y,

   $\Omega(Y) = (\mu(Y), U(Y)_{nq}, \Lambda(Y)_q, N(Y))$

4. Calculate an eigenspace for the concatenated pair, $Z = [X, Y]$,

   $\Omega(Z) = (\mu(Z), U(Z)_{nr}, \Lambda(Z)_r, N(Z)) = \Omega(X) \oplus \Omega(Y)$

   steps 5 to 7 represent the process.

5. Compute the number of the data points in eigenspace $\Omega(Z)$

   $N(Z) = N(X) + N(Y)$

6. Compute the data mean in eigenspace $\Omega(Z)$

   $\mu(Z) = (N(X)\mu(X) + N(Y)\mu(Y))/N(Z)$

7. Compute eigenvectors $U(Z)$ which should support all data in both sets X and Y, both $U(X)$ and $U(Y)$ should be subspaces of $U(Z)$.

   $U(Z) = [U(X), v]R$

   where R is an orthonormal matrix, v spans a t-dimensional subspace; $t \leq q + 1$.

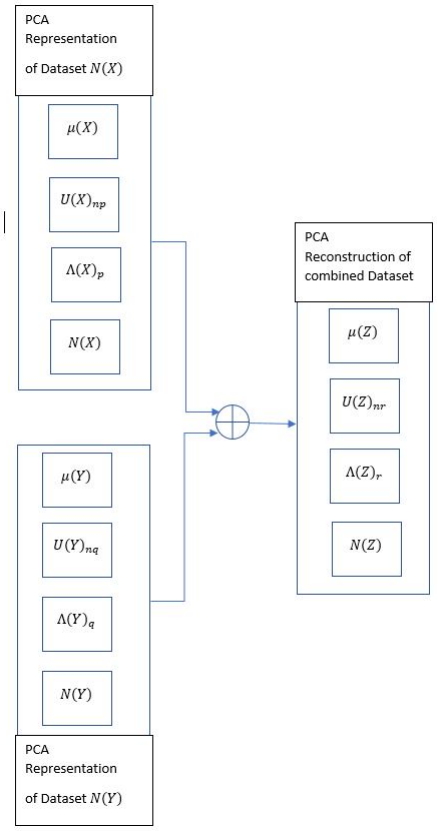


Figure 6.1. Adding eigenspaces with eigenvalue decomposition.

6.3 Visual data pre-processing

In our method, we utilise the AAM [25] for visual parameterisation. An AAM is constructed on the shapes and images, by first aligning the shapes then computing a mean shape. Using a piecewise affine warp algorithm, the texture sampled from within the convex hull of the shape for each image is warped to the mean shape. The warping is performed using Delaunay triangulation on image landmark data in both original and target prototypes, and affine warping corresponding triangles. Principal component analysis is applied to the shape and texture separately and then to the concatenation of the PCA parameters for shape and texture.

As mentioned in Chapter 4 that Deena et al. [38] chose 184 images by

selecting 4 random frames throughout the dataset, from each of the 45 sounds plus silence and breath for training the AAM. 56 landmark points were used in each of the example images, then AAM was constructed on the shapes and images. After that, they projected the remaining corpus to AAM parameters. Our motivation is to build AAM on a larger dataset in order to find better visual realisations of phonemes.

6.4 Experiments

The experiments which are conducted in this section are organised as follows. Subsection 6.4.1 compares the performance of two methods for building eigen models using adding eigenspaces and not using it. The purpose of this experiment is to use an online learning algorithm to add AAMs as they do not need retraining whenever new training data arrives and asses our MRD approach using these models. Subsection 6.4.2 investigates two initialisation approaches to initialise the latent space, namely: PCA and probabilistic principal component analysis (PPCA), because the training process has to proceed with appropriate initialisation of the latent space. Subsection 6.4.3 performs experiments that examine feature improvement approaches to reduce speaker variability and compare two normalisation technique using mean-centering AAM parameters and a z-score normalisation.

6.4.1 Experiment 1: Adding eigenspaces

We have chosen 55 sequences from the LIPS dataset giving 5982 images, and more markup points are utilised with 97 facial landmarks identified for each frame; 38 of them described the inner and outer mouth shape

in order to have a smoother facial boundary for training the AAM. PCA cannot be applied on the texture, because there are too many images to store into memory at once, so incremental approaches are a prerequisite to our method. Such that, this dataset is partitioned into two collections. PCA is applied to the texture for each set, then the two constructed models are merged. The number of eigenvectors retained in any model, including a merged model, is set to be 100 to obtain a 90% of the variance of the texture because we find that utilising more principal components will not lead to further performance improvement. Subsequently, the PCA is applied to the concatenated shape and texture merged PCA parameters. By retaining 95% of the variance of the combined parameters, a 95−dimensional vector of normalised AAM parameters using z-score normalisation, and 34−dimentional vector of mean-centering of AAM parameters are obtained. We decided to retain this number of parameters in order to obtain the visual and auditory spaces of comparable dimension as the audio dimension is 20.

Figure 6.2 displays AAM features synthesised by our proposed approaches using both adding eigenspace models and one eigenspcae model for a test sequence "The boy a few yards ahead is the only fair winner" and those obtained from ground truth. The frames correspond to three words consisting of eight different visemes from the test auditory sentence. The corresponding phonetic labels are shown below each frame.

Table 6.1 shows the average mean squared error (AMSE) and average correlation coefficient (ACC) between the ground truth and the synthesised mean-centering AAM features, obtained from MRD using adding eigenspaces and without using it approaches across the 20 test

utterances, such that the training and testing sets do not overlap. We will call the approach based on the adding eigenspaces MRD-ADD and the approach without using adding eigenspaces MRD. The results show that the AMSE errors for MRD-ADD are slightly higher than those for MRD. This means that the addition of eigenspaces provides only slightly less accurate results than batch approaches.

Figure 6.3 shows trajectory comparisons among MRD, MRD-ADD, and ground truth for the sentence "The boy a few yards ahead is the only fair winner". The Figure shows how the synthetic trajectories for MRD and MRD-ADD follow the same general pattern as the ground truth trajectories. Figure 6.4 highlights selected frames from the animation. The trajectories at frames 14 to 16 associated with the articulation of "b", the ground truth mouth and the MRD-ADD synthetic mouth are both closed, while the MRD synthetic mouth is slightly opened. The differences which do happen between the MRD synthetic and ground truth signals in this period appear in terms of signal amplitude. Frames 18 and 20 refer to the articulations "oy" from the word boy and "a" respectively. It can be shown that the trajectories at point 20 are very close and both the synthetic and ground truth mouths are opened. The results at the period 14 to 20 show that the MRD-ADD model gives the better results, and follow approximately the same general pattern as the ground truth trajectories. The trajectories at frames 36 to 39 associated with the articulation of "eh" from the word "head", the ground truth mouth and both of the MRD, MRD-ADD synthetic mouths are opened, which can be shown in Figure 6.2. It can be shown that the trajectories for MRD and MRD-ADD at this period follow the same pattern as the ground truth trajectories.

It can be seen from Figure 6.4 that the difference between the quality of mouth articulation between real and MRD-ADD synthetic videos in frames 14, 18, 19 is non significant. The lip uttering /b/ from the MRD-ADD method showed proper lip synchronisation with the audio and appeared to be the best, whilst the MRD gave lip synchronisation with a few jerks in the animation. We observe that the MRD-ADD could estimate the form of lip movements reasonably. The trajectories show that the performance of MRD-ADD approximately the same of MRD.

PCA building steps

The process for building PCA models is as follows

1. Given a training set of 2857 frames, compute the eigenvectors and eigenvalues using PCA. We have used this number of frames due to the memory constraint.

2. Define this model as $\Omega(X)$.

3. Given another collection of 3125 frames, perform PCA on the dataset.

4. Build $\Omega(Z) = \Omega(X) \oplus \Omega(Y)$

After increasing the RAM capacity in the machine to 32 GB, active appearance model has been built using about 6000 images, so that we did experiments to compare the two approaches. As shown in Figure 6.2, there is very little difference in the accuracy of the synthesised images using adding eigenspace models and those without using this procedure.

The outcomes of MRD-ADD results are almost as good as when

training MRD on all data together. However, our approach can deal with patch input perhaps over several months/years without the need to rebuilt model from scratch it is tractable when machine memory limited.

Suppose there is a collection of N data samples, each n dimensional ($N \times n$), where n is the number of pixels in an image and N is the number of images. The covariance matrix is of size $n \times n$. In practice the eigenspace model could be computed by utilising an $N \times N$ matrix, where N is the number of samples in the dataset. This method is efficient in image processing where the number of samples is less than the number of dimensions in each image $N << n$. Since $N << n$ we use in our experiments the inner product method [157]. It is often supposed that only those eigenvectors that correspond to p largest eigenvalues are of interest, the others are discarded by deleting columns from the eigenvector matrix. Different standards for discarding eigenvectors and eigenvalues available and these fit different applications and different approaches of computation. Three popular approaches are:

1. Set out p as a specified integer and thus retain the p largest eigenvectors [128].

2. Retain the p eigenvectors whose size is larger than an absolute threshold [18].

3. Retain the p eigenvectors such that a fixed portion of energy is kept.

In our experiments for batch approach the training set consists of 5982

texture vectors and for incremental approach we used two batches of 2857 and 3125 texture vectors of dimension 318753. To keep the p largest eigenvectors we used the first approach and set $p = 100$. Conventional batch approaches cannot be utilised to build an eigenspace because there are too many images to store in memory at once, so incremental approaches are essential methods. Constructing eigenspace for texture dataset 5982×318753 is difficult because storing such matrix of pixels in memory is intractable and need a machine with high RAM capacity. In contrast incremental approaches do not require all observations at once thus decreasing storage requirements and making large issues computationally appropriate.

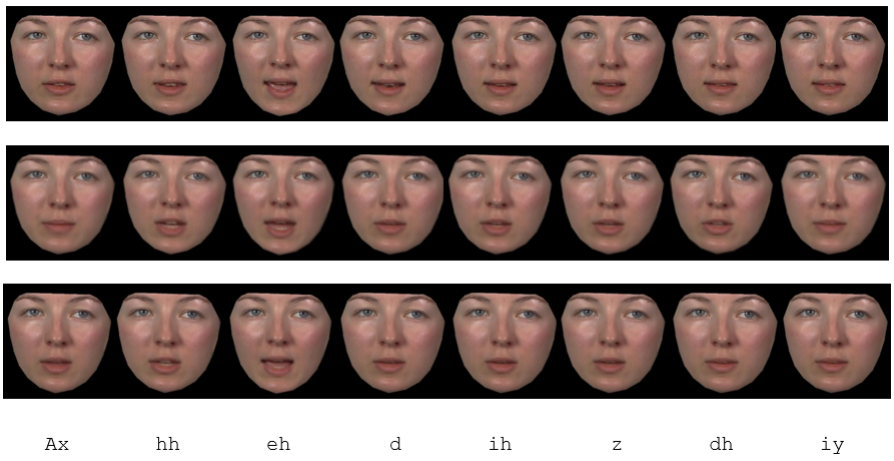


Figure 6.2. Example frames of the AAM features obtained from the Lips synthesis results using ground truth (top row), MRD approach using adding PCA models (middle row), MRD approach without using adding PCA models (bottom row). The phonemes correspond to the words (ahead, is, the) from the test audio sentence (The boy a few yards ahead is the only fair winner).

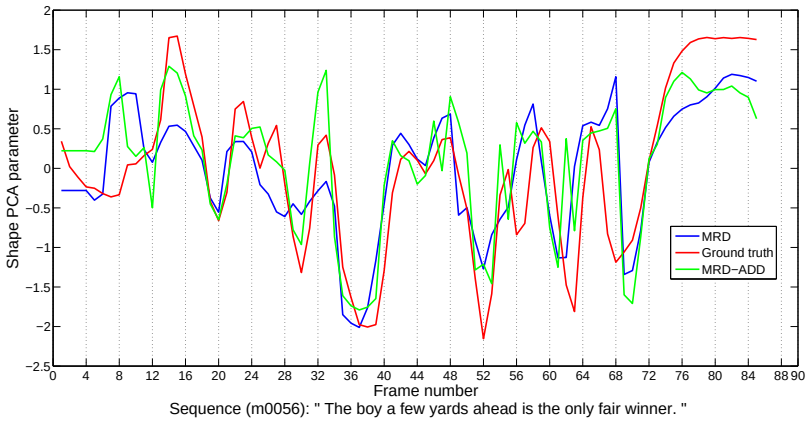


Figure 6.3. Shape trajectories obtained from MRD, MRD-ADD and the corresponding ground truth trajectories.

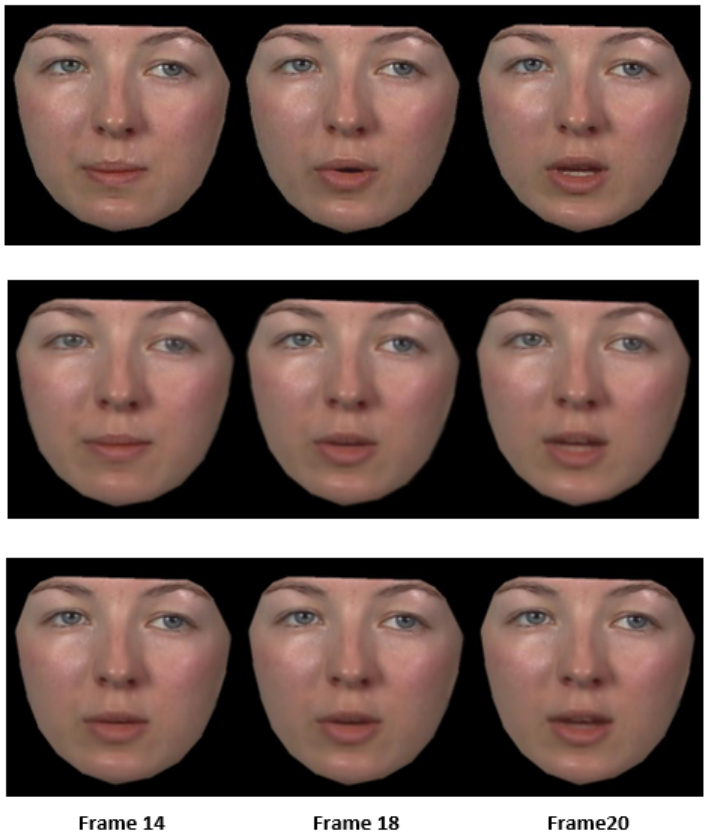


Figure 6.4. Selected frames of the AAM features obtained from the Lips synthesis results using ground truth (top row), MRD approach using adding PCA models (middle row), MRD approach without using adding PCA models (bottom row). The phonemes correspond to the words (boy, a) from the test audio sentence (The boy a few yards ahead is the only fair winner.)

6.4.2 Experiment 2: Latent space initialisation

The quality of the optimisation is relying upon many things, especially initialisation, numerical errors and the optimiser used because the optimisation procedure is gradient based, which means that there is no analytic form to a unique solution. Random number initialisation and initialised to zero cannot be used. Random number initialisation and initialised to zero might lead to the MRD training algorithm becoming stuck in local minima or longer convergence that does not recover the true embedded space. Damianou [34] initialised the latent space by concatenating two datasets, each consisting of images corresponding to a set of three different faces, under 64 different illumination conditions and performing PCA. An alternative method was to perform PCA on each dataset individually and then concatenated the two low dimensional subspaces to initialise the latent space X. They found that both initialisations obtained similar results.

In order to avoid bad local minimum, we try to control the initialisation by using appropriate initial variational distribution and signal to noise ratio (SNR). In this experiment, we set the iterations for initialising the variational distribution to 300 and the initial SNR to 150 in order to obtain high SNR of the optimised model. We also investigate two initialisations approaches to initialise the latent space variational means, namely: PCA and PPCA. PCA is performed on each dataset (audio and visual data) separately and then concatenated the two low-dimensional representations to initialise latent space X. This is also done with PPCA. The results are shown in Figure 6.5, it can be seen that PPCA initialisation gives better results than PCA. This is because that PPCA finds a latent space X which maximises the correlation be-

tween the two features and therefore better mappings are learnt from the latent space to each representation. In the next experiments, the initialisation approach is fixed to PPCA. More detail is described in section 6.5.

6.4.3 Experiment 3: Normalisation procedure

In this thesis, we attempt to minimise the effects of the face pose variability in the visual features and thus enhance the discriminative power of the visual features. Feature enhancement approaches considered include z-score normalisation and mean-centering AAM Parameters discussed in more detail in Section 3.15. We compare the trajectories of shape parameters for ground truth against synthesised output shape parameters obtained utilising the MRD and that gave best quantitative results than SGPLVM [38]. The plots for three test sequences of LIPS using mean-centering AAM parameters and z-score normalisation as a normalisation of visual features are shown in Figures 6.6 and 6.7 respectively. The Figures clearly show the high correlation between shape parameters trajectories obtained from MRD and ground truth as compared to another method, which is supported by the quantitative results. However, the MRD trajectory in Figure 6.6 shows some differences which do happen between the two signals (MRD and ground truth) appear in terms of signal amplitude. It can be seen that the differences in amplitude tended not to affect the perceived accuracy of the lip-synch.

In Figure 6.7 we found that there are no differences in signal amplitude between the MRD synthetic trajectories and the ground truth trajectories and it follows the same general pattern as the ground truth

trajectories. This is may be because all parameter values lie within a stable bound of ±2 standard deviations from the mean. The differences that happen between the two signals appear in terms of noise. This might occur because we have restricted ourselves to the limited amount of training data since the MRD is a non-parametric model using Gaussian processes, the size of the model grows with the data. Such that there is no use of the full variance of the visual data. In addition, there is a wide range of highly non-linear dynamics because of the phenomenon of coarticulation which when modelled utilising a single model produces an over-generalised predictive model.

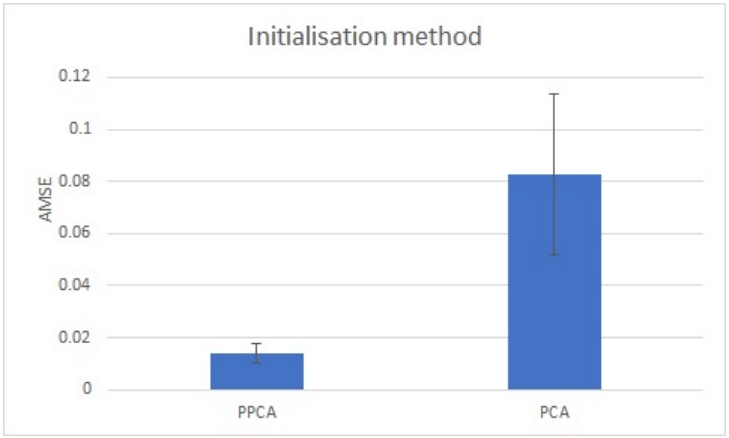


Figure 6.5. Varying latent space initialisation approaches.

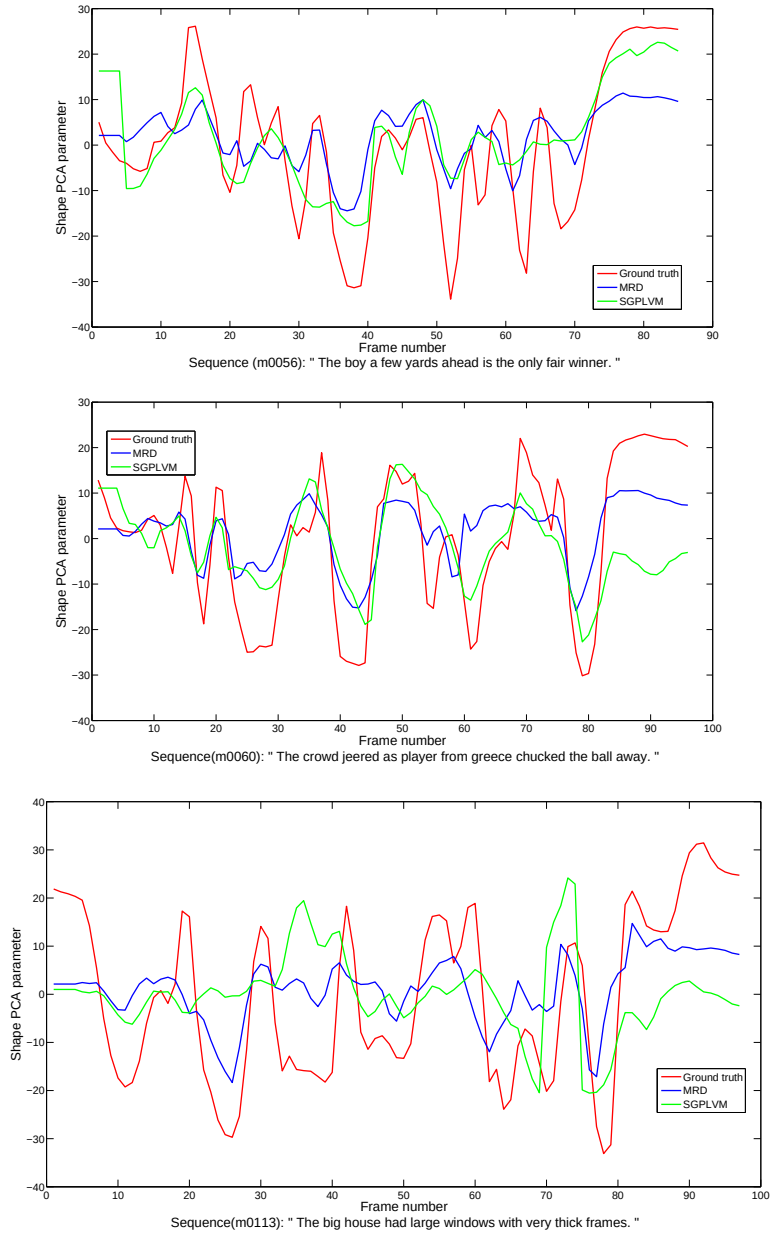


Figure 6.6. Shape trajectories obtained from MRD, SGPLVM, and the corresponding ground truth trajectories using mean-centering AAM parameter.

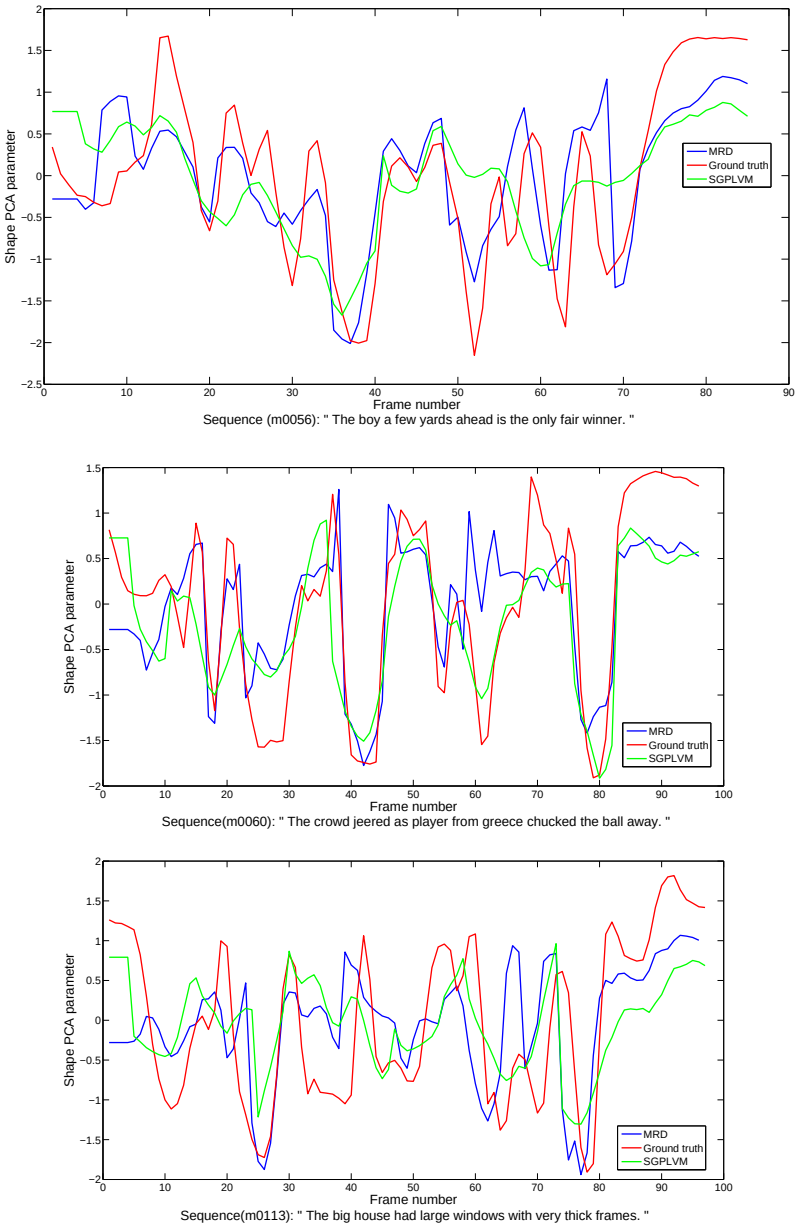


Figure 6.7. Shape trajectories obtained from MRD, SGPLVM and the corresponding ground truth trajectories using z-score.

6.5 Objective evaluation for the MRD

In this Section quantitative evaluation is performed using AMSE and ACC measures with the standard deviation errors. We evaluate our MRD approach against multiple approaches of visual speech synthesis. The visual features are extracted using an AAM.

Due to the complexity of MRD training, optimisation of MRD likelihood becomes intractable when the number of data points exceeds a few thousand. In our experiments, a repeated random subsampling approach for choosing about 10 sets, each set of approximately 55 sequences from the 236 auditory-visual collection pairs is utilised for training, giving an average of 6000 frames for each set. Two sets of about 21 utterances from the remaining of the auditory-visual collection pairs is utilised for testing, such that the training and testing sets do not overlap. We have restricted ourselves for choosing these number of sequences from the 236 auditory-visual sequence pairs due to the $O(NM^2)$ complexity of MRD training, where N is the number of data points and M is the inducing points in the model.

In their experiments Damianou et al. [34] set inducing points to 100, so that we set M = 100 in our experiments to reduce the complexity of the MRD training. The visual features used are the normalised AAM parameters and the auditory features are RASTA-PLP processed at $25Hz$. It is important when learning our fully Bayesian latent variable model of auditory and visual features to have the auditory dimension as being comparable to the visual dimension. As illustrated in the inference Subsection 5.1.3, given $\mathbf{y}^*$, one of the N^* test speech features, a test latent point $\mathbf{x}^*$ is optimised and a series of K candidate initial training inputs is found $\mathbf{x}_{NN}^1, \ldots, \mathbf{x}_{NN}^K$, ordered according to their sim-

ilarity to x^*, and only the shared dimensions is considered. A sorted series of K novel visual speech $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(K)}$ is then generated depending on those initial latent points. We have used Deena et al′s work [36] as a benchmark.

Objective methods include computing the error or correlation between ground truth and synthetic visual features in addition to comparing the evolution of their trajectories. Error metrics such as AMSE, root mean squared error (RMSE) and ACC, provide a measure of the static evaluation between ground truth and synthetic visual parameters.

We show quantitative results from our experiments. AMSE and ACC between synthesised AAM features vectors for the test sequences and ground truth is computed with the standard deviation of the errors. The ACC is given by

$$ACC = \frac{1}{K * I} \sum_{k=1}^{K} \sum_{i=1}^{I} \frac{(z_{k,i} - \mu_i)(\hat{z}_{k,i} - \hat{\mu}_i)}{\sigma_i \hat{\sigma}_i} \tag{6.12}$$

where μ_i is the mean of the ith dimension of z across the frames from 1 to K and σ_i is the corresponding variance and $\hat{\mu}_i$ is the mean of the ith dimension of $\hat{z}$ across frames from 1 to K and $\hat{\sigma}_i$ is the corresponding variance.

The results are compared against the SGPLVM method [36] utilising the same training and test collections. The AMSE error for MRD is lower than those for SGPLVM, mostly due to a smoother shared latent space obtained from fully Bayesian model, allowing estimation of both the dimensionality and the structure of the latent spaces to be achieved automatically. Table 6.1 presents the AMSE and ACC between the ground truth and synthesised AAM features, obtained from MRD and

SGPLVM approaches across the test utterances.

To test the statistical significance of the results we performed paired-sample t-test on the AMSE error values for the results of the methods tested [39]. The paired-sample t-test assumes that the two samples are dependent. Because the two samples of the error values obtained from the two methods MRD and the SGPLVM are dependent, the paired-sample t-test is the appropriate test to be utilised. This test gave a significance value of 0.00048 for the results which is lower than 0.05 value commonly utilised to determine statistical significance and consequently proves its statistical significance.

Our MRD approach also compared to the method followed by Havell et al. [80]. In their work, hidden Markov models are utilised to find the most likely sequence of appearance states given the auditory and visual training data. Using cluster analysis, a set of possible states was found. A structure based on Gaussian mixture models (GMMs) was proposed to model each phoneme separately to solve the problem of multiple possible visemes representing each phoneme and improved the quality of the HMM produced. Another method used by Havell et al. [80] to generate facial animations is based on coupled HMMs (CHMMs). A CHMM is a combination of multiple HMM chains coupled through cross-time and cross-chain conditional probabilities. In [196], a CHMM consist of two HMM chains describing the audio and video respectively and permitting for asynchronous progression of the chains, which is required in auditory-visual speech modelling. A number of CHMMs with different numbers of clusters were trained and a model with 10 states were build. The root mean squared error in shape normalised pixel values was compared to the ground truth images for the 750 frames

of 5 utterances. The RMSE is given by:

$$pixelerror = \sqrt{\frac{\sum_{i=1}^{N}(\mathbf{x}_{1,i} - \mathbf{x}_{2,i})^2}{N}} \tag{6.13}$$

where $\mathbf{x}_1$ refer to the ground truth and $\mathbf{x}_2$ refer to the reconstructed pixels for N pixels. The results are summarised in Table 6.1. The results show that the MRD model gives the best result. Only some implementations for the methods we compare to in Table 6.1, are available. The stated results of GMMs and CHMM have been obtained from the respective publication. This explains the omissions in the table.

In addition, our MRD method is compared to the methods followed by Fan et al. [60]. Two network topology using LSTMs are used in our experiments, a BLSTM and a deep BLSTM network. LSTMs network are trained utilising the same training and test sets. Table 6.1 presents the AMSE and ACC between the ground truth and synthesised AAM features, obtained from MRD, BLSTM, and DBLSTM across the test utterances. It can be seen that MRD performs better than LSTMs network, this is because the LIPS dataset is small and deep learning required a sufficiently dataset. Fan et al. [60] shows that for speech-driven talking head, the best network topology is with 64 nodes per layer. In our experiments we also used 64 nodes per layer. Figure 6.8 shows trajectory comparisons among MRD, BLSTM, and ground truth. The synthetic trajectories are calculated utilising three test utterances of LIPS dataset. The resultant synthesised parameter trajectories using MRD and the BLSTM follows the same general pattern as the ground truth trajectories. It is clear that the synthesised parameter trajecto-

ries using MRD slightly more closely follows the original ground-truth parameters than the BLSTM.

It should be noted from the results that the objective measures used in this work confirmed by qualitative evaluation obtained in Section 6.6, because a synthesiser that give more favourable objective results is found to be perceptually better based on subjective tests.

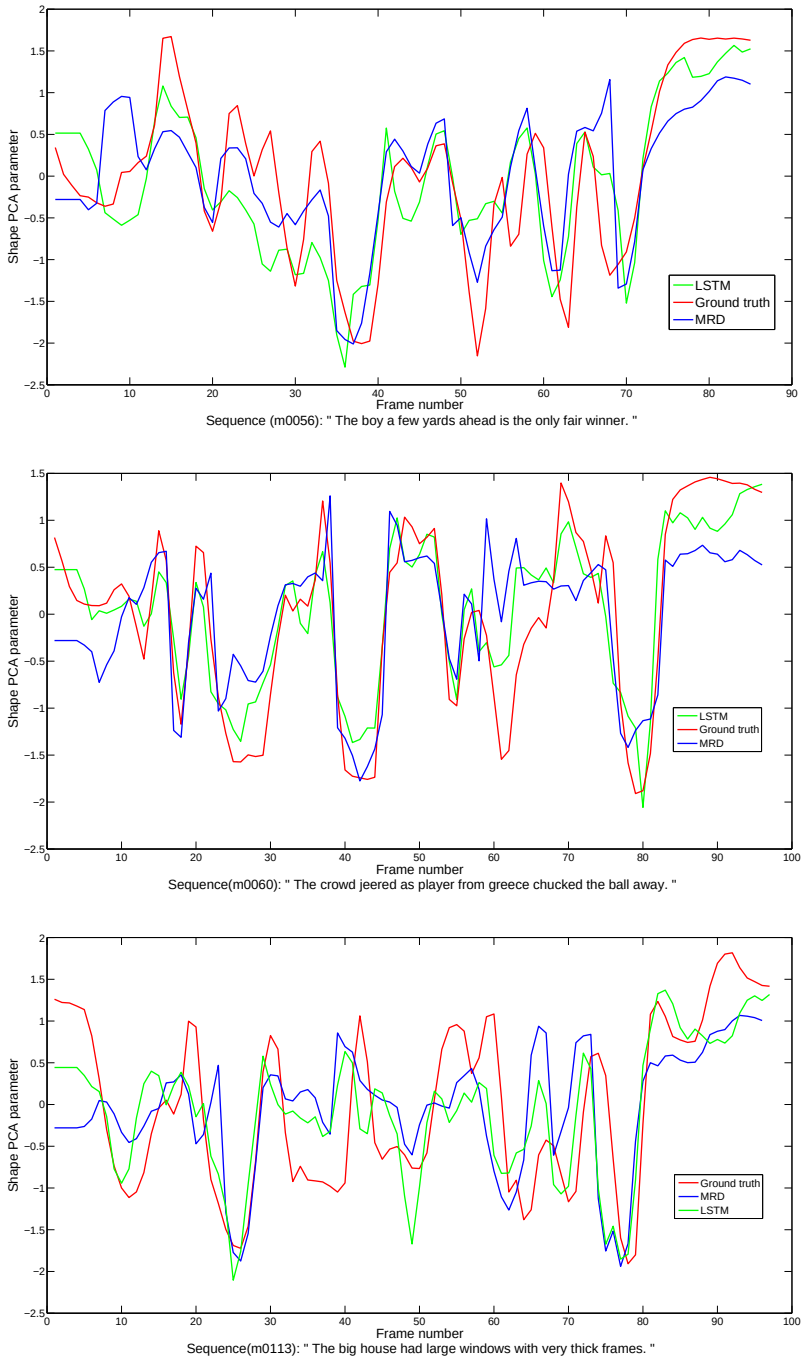


Figure 6.8. Shape trajectories obtained from MRD, BLSTM and the corresponding ground truth trajectories.

Table 6.1. Mean ± standard deviation objective measures of different approaches.

Method	AMSE	ACC	Average Pixel error
MRD	20.8847±4.5896	0.7623±0.0253	2.05
MRD-ADD	21.9897±4.8673	0.7576±0.0398	2.09
BLSTM	24.6471±7.1507	0.7517±0.024	2.21
SGPLVM	25.8953±5.7799	0.7259±0.0439	2.26
DBLSTM	27.4647±7.9427	0.7143±0.0224	2.34
GMMs	-	-	6.41
CHMM	-	-	6.62

6.6 Qualitative evaluation

In terms of validating this work, the objective comparisons might be not enough; performing auditory-visual perceptual experiments is another useful way to evaluate our work.

Subjective tests with human volunteers are performed to evaluate the facial animation. Image frames are generated from the synthesised AAM parameters, then the frames are encoded to video at the suitable frame rate and mixed with the test auditory file to produce a speech-synchronised talking head video. The perceptual test involved showing videos and asking volunteers to provide a score from 1 (poor) to 5 (excellent) on the quality of mouth articulation. The video utterances were selected randomly and subsequently fixed for all volunteers to calculate mean opinion scores (MOS) on the same utterances. MOS are one of the procedures that recommended by International Telecommunication Union Telecommunication Standardization Sector (ITU-T) for conducting subjective evaluations of transmission quality, with a scale from 1 to 5.

Many researchers have adopted MOS for evaluating the perceptual

aspects of facial animation [38, 91, 107, 169, 186]. When the videos are shown to the volunteers, comparative mean opinion scores can be collected which is suitable to compare different synthesis methods.

The volunteers were asked to assess 18 video sequences with 6 in each of the following classes: LIPS synthetic using MRD, the BLSTM and the SGPLVM. The sequences were selected randomly, they correspond to the following sentences:

1. No matter how overdone her praise where Amy was concerned I did not dare thank her.
2. The boy a few yards ahead is the only fair winner.
3. The crowd jeered as player from Greece chucked the ball away.
4. A burst pipe can cause damp carpets.
5. With artists trying to merge the courses prefer a job with Oxford Press.
6. The boy oyster came here to find a far airier atmosphere.

There was no time limit and the volunteers could play and replay each video any time they required. Some participants were native and most of them were non-native English speakers, but it was ensured that all volunteers had good command of the English language. The volunteers were asked to report their opinion about how well the lips synchronised with the audio and picked on a scale of 1 to 5 the quality of mouth articulation.

After that, MOS is calculated for each video class per participant. Table 6.2 gives MOS and standard deviations calculated over the 20 volunteers, mostly research students from the Schools of Engineering

and Computer Science at the University of Cardiff. The results show that MRD videos perform better than BLSTM and SGPLVM videos. Moreover, the subjective tests for lip synchronisation have proved that the objective evaluation are consistent with the subjective evaluation.

Table 6.2. MOS scores for perceptual test.

Class	Mouth articulation
MRD-LIPS synthesis	3.795±0.48
BLSTM-LIPS synthesis	3.235±0.49
SGPLVM-LIPS synthesis	2.95±0.3

6.7 Discussion

Adding eigenspaces is a powerful method that can be used to add several smaller eigenmodels. Because of the memory constraints in a system, it may be possible to use this method when it is not possible to use the whole training set at once. Another advantage of this method is that the eigenspace model can be updated with the new data, it is therefore appropriate for real-world applications where the online learning from real-time dataset is needed.

The improved results over the SGPLVM-based speech synthesis approach [36] can be accounted by the fact that MRD uses of automatic relevance determination (ARD) covariance functions for the mapping from the latent to the observation space which allows for automatic dimensionality detection. In addition, a different Gaussian process (GP) mapping is utilised per output modality, each with a different set of ARD hyperparameters. So that efficient initialisation and training of such a model is allowed, and a soft segmentation for the latent space is defined after optimisation the different sets of ARD hyperparamters. As opposed to SGPLVM model that has a large number of free parame-

ters which need to be modified to obtain the optimal model, because of the absence of a Bayesian formulation. The quantitative results show that using MRD to jointly model auditory and visual features results in smaller error in comparison to ground truth as compared to another method.

Since we use a joint probabilistic model of auditory and visual features, then spurious pose variations in synthesis are produced because of the correlations between the pose and the auditory. Therefore the normalisation methods introduced in Section 3.15 had to be adopted. Moreover, a standardised frame of reference to compute the errors between synthesised AAM parameters and ground truth can be obtained using the normalisation procedure.

6.8 Limitations of the MRD method

Although the MRD method is elegant, it is intractable for large training sets. Space and time complexity of training the model is increased. Training such a model with a high number of frames takes more time and requires larger storage to store the model. In our experiments a repeated random subsampling approach for choosing about 10 sets, each set of approximately 55 utterances from the 236 auditory-visual collection pairs is utilised for training, giving an average of 6000 frames for each set and two sets of about 21 utterances from the remaining of the auditory-visual collection pairs is utilised for testing, such that the training and testing sets do not overlap. Training such models takes a long time. Thus, the MRD comes with a high computational complexity. This could be an issue for very resource-limited devices, but does not compromise the possibility of online learning.

6.9 Summary

The main aim of this chapter was to construct a dynamic 2D appearance-based model of the face utilising a large dataset which allows different visual realisations of sounds to be represented using a small number of parameters. A description of a method to add eigenspaces, and visual data processing methods were presented. Because of the memory constraints in a system, it might be possible to utilise the incremental approach when it is not possible to use the whole training set at once. In addition, online learning methods are required if not all training data are available all the time and preferred over batch learning methods as they do not need retraining whenever a new training data is received.

Two eigenspaces were built: one utilising batch approach and another utilising incremental approach. The performance of our MRD approach using those two models for visual representation were then compared. The results show that there was very little difference in the accuracy of the synthesised images and the difference between the errors obtained using the two methods is quite small, which mean that MRD using the addition of eigenspaces provides only slightly less accurate results than using batch methods.

Moreover, two methods were presented for visual normalisation, namely mean-centering of AAM parameters and z-score normalisation to minimise the effects of the face pose variability in the visual parameters. Quantitative evaluation was used to compare our proposed method using MRD with the current state-of-the-art methods. The results reveal that the joint models of auditory and visual using MRD perform better than the comparable methods of visual speech synthesis.

Qualitative analysis with human participants has also been performed to evaluate the perceptual characteristics of the synthesised facial motions. Qualitative analysis demonstrating the improved performance of our MRD method. Furthermore, experiments to determine the optimal latent space initialisation were dealt with to avoid bad local minimum. Our MRD approach is an efficient model for modelling two views of the same process, which are in this work the auditory and visual components of speech. However, because our MRD model is a non-parametric, the size of the model grows with the data. Training such a model becomes intractable when the number of frames exceeds a few thousand. In addition, when a wide range of highly non-linear dynamics modelled utilising a single dynamical model, an over-generalised predictive model will be produced.

Chapter 7

CONCLUSION AND FUTURE WORK

This thesis presented a number of improvements for generating realistic speech animation of the human face using joint probabilistic models of speech and face appearance. The auditory and visual signals were represented as a collection of factorised latent spaces. To train the models two auditory-visual corpora were used, namely the LIPS [171] and DEMNOW [54] datasets. The LIPS corpus features a female British talker reading sentences from the Messiah corpus, while DEMNOW features a female American anchor giving news presentations. The facial features were extracted utilising active appearance model (AAM). Relative spectral transform-perceptual linear prediction (RASTA-PLP) was used as speech parameterisation to obtain speech features matching the visual frame rate. Quantitative evaluation of the proposed approaches was presented and compared with the current state-of-the-art methods. Furthermore, qualitative analysis with human volunteers was performed to evaluate the synthesised animations.

Our motivation was to explicitly model the non-linearities in audio-visual mapping utilising non-parametric, fully Bayesian latent variable model which utilises conditional non-linear independence structures to

learn an efficient latent space. In contrast to the shared Gaussian process latent variable model (SGPLVM) [38] approach, which presumes that a single latent variable is able of representing each modality, suggesting that the modalities can be fully aligned. In this work, a smooth continuous latent space, where a latent variable may be more important to the shared representation than the private representation, namely manifold relevance determination (MRD) was introduced. The model was exploited to combine the audio and visual features using a softly shared latent space. Our Bayesian approach allow us to automatically estimate of both the dimensionality and the structure of the latent space. In contrast to Deep Learning methods for auditory-visual mapping which required a sufficiently comprehensive dataset, because such methods are generally highly under constrained, our MRD method can be undertaken in smaller datasets.

A more accurate AAM, with more facial landmarks identified for each frame and using a larger dataset was created. Quantitative evaluation revealed that utilising more landmark points around the mouth and building an AAM using larger dataset can give better results and a smoother facial boundary can be obtained. It was also shown that SGPLVM produces more accurate results when using a more accurate AAM.

For a visual representation, the active shape and active appearance models [27] were utilised to extract visual parameters from images. To cater for all possible speech-related facial expressions, the AAM was trained using a large number of sequences. To overcome the limitations of the batch learning methods, incremental approach was used. This method is appropriate for real-world applications where the online

learning from real-time dataset is needed. Two eigenspaces were constructed for visual representation: one utilising a batch approach and another utilising an incremental approach [76, 77]; experiments were conducted to compare the two approaches. The incremental approach was performed very efficiently and provided only slightly less accurate results than the batch approach.

Quantitative evaluation results of the proposed approaches using MRD were compared with other related methods such as the SG-PLVM [38], bidirectional LSTMs (BLSTMs) [60], and hidden Markov models (HMMs) [80]. Our MRD approach was found to give the better quantitative results. Qualitative evaluation was also included demonstrating the improved performance of our MRD method in comparison to alternative methods.

7.1 Contributions

The contributions of this thesis are summarised below;

- **A more accurate AAM of talking faces has been built:** We showed that using larger dataset and/or more landmarks for building an AAM can produce a more accurate model. A more accurate AAM, with 97 facial landmarks identified for each frame was constructed; 38 of these landmarks described the inner and outer mouth shape, which contains most variation in talk. Experiments were performed to investigate the hypothesis that increasing the number of facial landmarks can increase the accuracy of AAM. Quantitative evaluation revealed that utilising more landmark points around the mouth can give more accurate model and a smoother facial boundary can be obtained utilising more

landmark points for each frame, details were given in Chapter 4. Furthermore, experiments were performed to investigate that constructing an AAM on larger dataset can also improve the accuracy, details were given in Chapter 4.

- **The accuracy of SGPLVM was increased by utilising a more accurate AAM within the algorithm:** This is the first study to report that using a more accurate AAM within SGPLVM improves the accuracy of SGPLVM. Objective evaluation via reconstruction error was performed to compare the proposed approach against the previously existing methods. The quantitative evaluation confirmed our hypothesis, with a full description given in Chapter 4.

- **First application of MRD model for visual speech synthesis:** A new model for visual speech synthesis was presented, namely manifold relevance determination model, which explicitly models the non-linearities in audio-visual mapping. The accuracy of generating videos of talking faces using MRD instead of the SGPLVM was improved. Statistical evaluation of synthesised visual features against ground truth data was obtained and compared with the current state-of-the-art visual speech synthesis approach, with the analysis described fully in Chapter 5. This was also presented at ECCV'2016 conference.

- **Facilitating the performance of MRD by utilising incremental eigenmodels:** Incremental procedure for learning eigenmodels was utilised, thus facilitating incremental updating of the visual speech models in real world applications where online learn-

ing is needed. The MRD methods utilising batch and incremental AAMs were compared and demonstrated very similar accuracy. Full details were given in Chapter 6.

7.2 Future work

Future work will be covered in this section, and focuses on areas in which our approach could be further developed:

- This work utilised two datasets, the first one (LIPS corpus) containing read sentences from the Messiah corpus by a single person and the second one (DEMNOW corpus) featuring a female American anchor presenting news. Future work should include applying the approaches developed here to different groups of people based on their ethnicity, age, gender, face shape dynamic, etc. Moreover, applying these approaches to different contexts such as conversational speech would produce different results. Emotional data should be considered for conversational agents. Possible methods for combining emotion to the visual speech need investigation.

- The results of this work could certainly be improved with a latent model of phonetic context. The MRD method can be extended by augmenting the model with switching states represented by the phonetic context to model backward and forward coarticulation. The switching states can be found utilising a variable length Markov model trained on a phonetic data. The auditory and visual features corresponding to those switching states can then be extracted and modelled utilising MRD.

- Deep Learning is a recent direction in artificial intelligence and machine learning research. Recently, new deep learning approaches are being born, outperforming state-of-the-art machine learning and existing deep learning techniques. In terms of further work, it would be useful to explore the use of deep learning architectures, such as have found modern success in text to speech synthesis [192]. The current overview of generative adversarial network GANs [70] has shifted the motivation of the machine learning group to generative modelling. GANs contain two challenging networks: generative and discriminative networks. The generator's target is to generate realistic samples while the discriminator's target is to discriminate between the real and produced samples. This competition leads the generator to produce robustly realistic samples. Vougioukas [180,182] proposed an end-to-end model using temporal GANs for speech-driven facial animation, capable of generating a video of a talking head from an audio signal. In future work, we would like to extend the network architectures of [72, 182, 205] to generate high quality video using GAN and compare these outputs to those describe in this thesis.

- The current model was built utilising 2D video image data. The next stage is to extend the 2D photo-realistic talking face to 3D, using a 2D-to-3D reconstruction methods [5, 200]. It may be promising to use the landmark updating optimization strategy [108] which can give high-quality 3D face models.

- Since the LIPS dataset includes expression-free visual speech recordings displaying a neutral prosody, the synthesised visual

speech shows neither expressions nor emotions. Adding expressiveness to the face could increase the naturalness of the synthetic visual speech, but it will probably affect the visual speech quality: when the expressions may not be perceived as natural or may not be so well correlated with the speech. However, it should be noted that expression is an important component of visual speech and therefore needs incorporating for more realistic facial animation models. In the past few years researchers such as [93, 138, 152] tended to use deep neural networks(DNNs) in the field of expressive talking head. Human speech-based communication does not only include a collection of gestures and speech sounds corresponding to the production of sentences. The realism of the communication is enhanced by adding expressions to the utterance. The BIWI audio-visual corpus [61] of effective speech and corresponding dense dynamic 3-D face geometries can be utilised for an expressive talking head. Each emotional sentence in the BIWI corpus was enriched by the states such as negative, anger, sadness, stress, contempt, fear, surprise, excitement, confidence, happiness and positive.

- The work presented has focused on synthesising neutral speech and it would also be interesting to extend the proposed methods to produce, in addition to more expressive speech, realistic head and eye movements. A possible plan of future work is to add additional modes to the MRD in order to combine prosodic information such as eye blinks and head movements. Prosodic information in both auditory and visual modalities would be correlated with the speech content.

BIBLIOGRAPHY

[1] Tobias S Andersen. The Mcgurk illusion in the oddity task. In *International Conference on Auditory-Visual Speech Processing (AVSP)*, pages paper S2–3, 2010.

[2] James A Anderson. *An introduction to neural networks*. MIT Press, Cambridge, MA, 1995.

[3] Harold A Arb. *Hidden markov models for visual speech synthesis in limited data environments*. PhD thesis, Air Force Institute of Technology, Wright- Patterson Air Force Base, Ohio, 2001.

[4] Gonzalo R Arce. *Nonlinear signal processing: a statistical approach*. New York, NY, USA:Wiley Interscience, 2005.

[5] Abdullah Taha Arslan and Erol Seke. Face depth estimation with conditional generative adversarial networks. *IEEE Access*, 7:23222–23231, 2019.

[6] Lecia J Barker. Computer-assisted vocabulary acquisition: The CSLU vocabulary tutor in oral-deaf education. *Journal of Deaf Studies and Deaf Education*, 8(2):187–198, 2003.

[7] Yasemin Bekiroglu, Andreas Damianou, Renaud Detry, Johannes A Stork, Danica Kragic, and Carl Henrik Ek. Probabilistic consolida-

tion of grasp experience. In *International Conference on Robotics and Automation (ICRA)*, pages 193–200. IEEE, 2016.

[8] Christopher M Bishop. *Neural networks for pattern recognition.* Clarendon Press, Oxford, 1995.

[9] Christopher M Bishop. *Pattern recognition and machine learning.* Springer, New York, 2006.

[10] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[11] Elif Bozkurt, Cigdem Eroglu Erdem, Engin Erzin, Tanju Erdem, and Mehmet Ozkan. Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation. In *The True Vision-Capture, Transmission and Display of 3D Video (3DTV)*, pages 1–4. IEEE, 2007.

[12] Matthew Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28. ACM Press, 1999.

[13] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360. ACM Press, 1997.

[14] E.Oran Brigham. *The fast Fourier transform: an introduction to its theory and application.* Prentice Hall, New Jersey, 1973.

[15] Peter Bull and Gerry Connelly. Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, 9(3):169–187, 1985.

[16] Joseph P Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.

[17] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.

[18] Shivkumar Chandrasekaran, Bangalore S Manjunath, Yuan-Fang Wang, Jay Winkeler, and Henry Zhang. An eigenspace update algorithm for image analysis. *Graphical models and image processing*, 59(5):321–332, 1997.

[19] Jixu Chen, Minyoung Kim, Yu Wang, and Qiang Ji. Switching Gaussian process dynamic models for simultaneous composite motion tracking and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2655–2662. IEEE, 2009.

[20] Tsuhan Chen. Audiovisual speech processing. *Signal Processing Magazine*, 18(1):9–21, 2001.

[21] Claude C Chibelushi, Farzin Deravi, and John SD Mason. A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, 4(1):23–37, 2002.

[22] Michael M Cohen, Dominic W Massaro, et al. Modeling coarticulation in synthetic visual speech. *Models and techniques in computer animation*, 92:139–156, 1993.

[23] Michael M Cohen, Rachel L Walker, and Dominic W Massaro. Perception of synthetic visual speech. In *Speechreading by humans and machines*, pages 153–168. Springer, 1996.

[24] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[25] T. F. Cootes, G. J. Edwards, and C. J. Taylor. *Active appearance models*, pages 484–498. Springer Berlin Heidelberg, 1998.

[26] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Training models of shape from sets of examples. In *BMVC*, pages 9–18. Springer, 1992.

[27] Timothy F Cootes, Cristopher J Taylor, et al. Statistical models of appearance for computer vision. Technical report, University of Manchester, 2004.

[28] Piero Cosi, Michael M Cohen, and Dominic W Massaro. Baldini: baldi speaks italian. In *7th Interna- tional Conference on Spoken Language Processing*, pages 2349–2352, 2002.

[29] Darren Cosker. *Animation of a Hierarchical Appearance Based Facial Model and Perceptual Analysis of Visual Speech.* PhD thesis, Cardiff University, 2006.

[30] Darren Cosker, Dave Marshall, Paul L Rosin, and Yulia Hicks. Speech driven facial animation using a hidden markov coarticulation model. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 1, pages 128–131. IEEE, 2004.

[31] Darren Cosker, David Marshall, Paul Rosin, and Yulia Hicks. Video realistic talking heads using hierarchical non-linear speech-appearance models. *Mirage, France*, 147, 2003.

[32] Ian Craw and Peter Cameron. Face recognition by computer. In *BMVC*, pages 498–507. Springer, 1992.

[33] Andreas Damianou, Michalis K Titsias, and Neil D Lawrence. Variational Gaussian process dynamical systems. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2011.

[34] Andreas C. Damianou, Carl Henrik Ek, Michalis K. Titsias, and Neil D. Lawrence. Manifold Relevance Determination. In *Proceedings of the 29th International Conference on Machine Learning*, pages 531–538. Omnipress, 2012.

[35] Samia Dawood, Yulia Hicks, and David Marshall. Speech-driven facial animation using manifold relevance determination. In *European Conference on Computer Vision*, pages 869–882. Springer, 2016.

[36] Salil Deena and Aphrodite Galata. Speech-driven facial animation using a shared Gaussian process latent variable model. In *Advances in Visual Computing*, pages 89–100. Springer, 2009.

[37] Salil Deena, Shaobo Hou, and Aphrodite Galata. Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–8. ACM, 2010.

[38] Salil Deena, Shaobo Hou, and Aphrodite Galata. Visual speech synthesis using a variable-order switching shared Gaussian process dy-

namical model. *IEEE Transactions on Multimedia*, 15(8):1755–1768, 2013.

[39] Salil Prashant Deena. *Visual Speech Synthesis by Learning Joint Probabilistic Models of Audio and Video*. PhD thesis, The University of Manchester, 2012.

[40] John R Deller Jr, John G Proakis, and John H Hansen. *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[41] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society*, 39(1):1–38, 1977.

[42] Li Deng and Douglas O'Shaughnessy. *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2003.

[43] Zhigang Deng and Junyong Noh. Computer facial animation: A survey. In *Data-driven 3D facial animation*, pages 1–28. Springer, 2008.

[44] Priya Dey. *Visual Speech in Technology-Enhanced Learning*. PhD thesis, University of Sheffield, 2012.

[45] Priya Dey, Steve C Maddock, and Rod Nicolson. Evaluation of A Viseme-Driven Talking Head. In *TPCG*, pages 139–142, 2010.

[46] Chuang Ding, Pengcheng Zhu, and Lei Xie. Blstm neural networks for speech driven head motion synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[47] Chuang Ding, Pengcheng Zhu, Lei Xie, Dongmei Jiang, and Zhong-Hua Fu. Speech-driven head motion synthesis using neural networks.

In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[48] Yangzhou Du and Xueyin Lin. Realistic mouth synthesis based on shape appearance dependence mapping. *Pattern Recognition Letters*, 23(14):1875–1885, 2002.

[49] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. Wiley-Interscience, New York, NY, USA, 2000.

[50] Carl Henrik Ek, Peter Jaeckel, Neill Campbell, Neil D Lawrence, and Chris Melhuish. Shared Gaussian process latent variable models for handling ambiguous facial expressions. In *Mediterranean Conference on Intelligent Systems and Automation*, volume 1107, pages 147–153, 2009.

[51] Carl Henrik Ek, Jon Rihan, Philip HS Torr, Grégory Rogez, and Neil D Lawrence. Ambiguity modeling in latent spaces. In *Machine learning for multimodal interaction*, pages 62–73. Springer, 2008.

[52] Carl Henrik Ek, Philip HS Torr, and Neil D Lawrence. Gaussian process latent variable models for human pose estimation. In *Machine learning for multimodal interaction*, pages 132–143. Springer, 2007.

[53] Paul Ekman and Wallace V Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[54] Gwenn Englebienne. *Animating faces from speech*. PhD thesis, University of Manchester, 2008.

[55] Gwenn Englebienne, Timothy F Cootes, and Magnus Rattray. A

probabilistic model for generating realistic lip movements from speech. In *NIPS*, volume 8, pages 401–408, 2007.

[56] Norman P Erber. Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, 40(4):481–492, 1975.

[57] Tony Ezzat, Gadi Geiger, and Tomaso Poggio. *Trainable videorealistic speech animation*. ACM, 2002.

[58] Tony Ezzat and Tomaso Poggio. Visual speech synthesis by morphing visemes. *International Journal of Computer Vision*, 38(1):45–57, 2000.

[59] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie. Photo-real talking head with deep bidirectional LSTM. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888. IEEE, 2015.

[60] Bo Fan, Lei Xie, Shan Yang, Lijuan Wang, and Frank K Soong. A deep bidirectional LSTM approach for video-realistic talking head. *Multimedia Tools and Applications*, 75(9):5287–5309, 2016.

[61] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010.

[62] Cletus G Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11(4):796–804, 1968.

[63] Joseph Fourier. *Theorie analytique de la chaleur, par M. Fourier*. Didot, 1822.

[64] John G. Proakis and Dimitris G. Manolakis. *Digital signal processing (3rd ed.): principles, algorithms, and applications*. Prentice Hall, 1996.

[65] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

[66] Athinodoros S Georghiades, Peter N Belhumeur, and David J Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (6):643–660, 2001.

[67] Evgeny Gladilin, Stefan Zachow, Peter Deuflhard, and H C Hege. Anatomy-and physics-based facial animation for craniofacial surgery simulations. *Medical and Biological Engineering and Computing*, 42(2):167–170, 2004.

[68] Bertrand Le Goff and Christian Benoît. A french-speaking synthetic head. In *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*, 1997.

[69] Gene H Golub and Charles F Van Loan. *matrix computations, 3rd*. Johns Hopkins Univ Press, 1996.

[70] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[71] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[72] Haohan Guo, Frank K Soong, Lei He, and Lei Xie. A New GAN-based End-to-End TTS Training Algorithm. *arXiv preprint arXiv:1904.04775*, 2019.

[73] Ricardo Gutierrez-Osuna, Praveen K Kakumanu, Anna Esposito, Oscar N Garcia, Adriana Bojórquez, José Luis Castillo, and Isaac Rudomín. Speech-driven facial animation with realistic dynamics. *IEEE Transactions on Multimedia*, 7(1):33–42, 2005.

[74] Isabelle Guyon and Fernando Pereira. Design of a linguistic postprocessor using variable memory length Markov models. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, volume 1, pages 454–457. IEEE, 1995.

[75] Kathrin Haag and Hiroshi Shimodaira. Bidirectional LSTM networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In *International Conference on Intelligent Virtual Agents*, pages 198–207. Springer, 2016.

[76] Peter Hall, David Marshall, and Ralph Martin. Merging and splitting eigenspace models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):1042–1049, 2000.

[77] Peter Hall, David Marshall, and Ralph Martin. Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition. *Image and Vision Computing*, 20(13):1009–1016, 2002.

[78] Richard Wesley Hamming. *Digital filters*. Courier Corporation, 1989.

[79] Naomi Harte and Eoin Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015.

[80] Benjamin Havell, Paul L Rosin, Saeid Sanei, Andrew Aubrey, David Marshall, and Yulia Hicks. Hybrid phoneme based clustering approach for audio driven facial animation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2261–2264. IEEE, 2012.

[81] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[82] Hynek Hermansky, Nelson Morgan, Aruna Bayya, and Phil Kohn. RASTA-PLP speech analysis technique. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 121–124. IEEE, 1992.

[83] Sarah Hilder, Barry-John Theobald, and Richard W Harvey. In pursuit of visemes. In *International Conference on Auditory-Visual Speech Processing(AVSP)*, pages 8–2, 2010.

[84] Chao-Kuei Hsieh and Yung-Chang Chen. Partial linear regression for speech-driven talking head application. *Signal Processing: Image Communication*, 21(1):1–12, 2006.

[85] Fu Jie Huang, Eric Cosatto, and Hans Peter Graf. Triphone based unit selection for concatenative visual speech synthesis. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 2037–2040. IEEE, 2002.

[86] Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Foreword By-Reddy. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[87] Xuedong Huang, Fileno Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, Kai-Fu Lee, and Ronald Rosenfeld. The SPHINX-II speech recognition system: an overview. *Computer Speech & Language*, 7(2):137–148, 1993.

[88] Singular Inversions. Facegen, 2008. http://www.facegen.com/.

[89] Fumitada Itakura. Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, 57, 1975.

[90] ITU-R. ITU-R Recommendation BT.470-6 Conventional Television Systems. 1998.

[91] Jia Jia, Shen Zhang, Fanbo Meng, Yongxin Wang, and Lianhong Cai. Emotional audio-visual speech synthesis based on PAD. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):570–582, 2011.

[92] Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

[93] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94, 2017.

[94] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[95] Michael Kirby and Lawrence Sirovich. Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[96] Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[97] Sumedha Kshirsagar and Nadia Magnenat-Thalmann. Visyllable based speech animation. In *Computer Graphics Forum*, volume 22, pages 631–639. Wiley Online Library, 2003.

[98] Felix Kuhnke and Jorn Ostermann. Visual speech synthesis from 3D mesh sequences driven by combined speech features. In *International Conference on Multimedia and Expo (ICME)*, pages 1075–1080. IEEE, 2017.

[99] Xinyu Lan, Xu Li, Yishuang Ning, Zhiyong Wu, Helen Meng, Jia Jia, and Lianhong Cai. Low level descriptors based DBLSTM bottleneck feature for speech driven talking avatar. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5550–5554. IEEE, 2016.

[100] Neil Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 2005.

[101] Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in Neural Information Processing Systems*, 16(3):329–336, 2004.

[102] Neil D Lawrence. Learning for larger datasets with the Gaussian process latent variable model. In *International Conference on Artificial Intelligence and Statistics*, pages 243–250, 2007.

[103] Oscar Martinez Lazalde, Steve Maddock, and Michael Meredith. A constraint-based approach to visual speech for a Mexican-Spanish talking head. *International Journal of Computer Games Technology*, 2008(3):1–7, 2008.

[104] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):684–698, 2005.

[105] Tue Lehn-Schiøler, Lars Kai Hansen, and Jan Larsen. Mapping from speech to images using continuous state space models. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 136–145. Springer, 2004.

[106] Kang Liu and Joern Ostermann. Realistic facial animation system for interactive services. In *International Speech Communication Association*, pages 2330–2333, 2008.

[107] Kang Liu and Joern Ostermann. Evaluation of an image-based talking head with realistic facial expression and head motion. *Journal on Multimodal User Interfaces*, 5(1-2):37–44, 2012.

[108] Peng Liu, Yao Yu, Yu Zhou, and Sidan Du. Single view 3d face reconstruction with landmark updating. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 403–408. IEEE, 2019.

[109] Steven R Livingstone, Katlyn Peck, and Frank A Russo. Ravdess: The ryerson audio-visual database of emotional speech and song. In *Annual meeting of the canadian society for brain, behaviour and cognitive science*, pages 205–211, 2012.

[110] Anders Löfqvist. Speech as audible gestures. In *Speech production and speech modelling*, pages 289–322. Springer, 1990.

[111] Jiyong Ma, Ron Cole, Bryan Pellom, Wayne Ward, and Barbara Wise. Accurate visible speech synthesis based on concatenating variable length motion capture data. *IEEE Transactions on Visualization and Computer Graphics*, 12(2):266–276, 2006.

[112] Wan-Chun Ma, Andrew Jones, Jen-Yuan Chiang, Tim Hawkins, Sune Frederiksen, Pieter Peers, Marko Vukovic, Ming Ouhyoung, and Paul Debevec. Facial performance synthesis using deformation-driven polynomial displacement maps. In *ACM Transactions on Graphics (TOG)*, volume 27, page 121, 2008.

[113] Dominic W Massaro. *Perceiving talking faces: From speech perception to a behavioral principle*, volume 1. MIT Press, 1998.

[114] Dominic W Massaro. A computer-animated tutor for spoken and written language learning. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 172–175. ACM, 2003.

[115] Dominic W Massaro. Symbiotic value of an embodied agent in language learning. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences*. IEEE, 2004.

[116] Dominic W Massaro and Michael M Cohen. Perception of synthesized audible and visible speech. *Psychological Science*, 1(1):55–63, 1990.

[117] Dominic W Massaro and Michael M Cohen. Perceiving talking faces. *Current Directions in Psychological Science*, 4(4):104–109, 1995.

[118] Dominic W. Massaro, Michael M. Cohen, Jonas Beskow, and Ronald A. Cole. Embodied conversational agents. chapter Developing and Evaluating Conversational Agents, pages 287–318. MIT Press, 2000.

[119] Dominic W Massaro, Slim Ouni, Michael M Cohen, and Rashid Clark. A multilingual embodied conversational agent. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 296b–296b. IEEE, 2005.

[120] Wesley Mattheyses, Lukas Latacz, and Werner Verhelst. On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009(1):169819, 2009.

[121] Wesley Mattheyses, Lukas Latacz, and Werner Verhelst. Optimized photorealistic audiovisual speech synthesis using active appearance modeling. In *International Conference on Auditory-Visual Speech Processing(AVSP)*, volume 10, pages 148–153, 2010.

[122] Wesley Mattheyses, Lukas Latacz, and Werner Verhelst. Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis. *Speech Communication*, 55(7):857–876, 2013.

[123] Wesley Mattheyses, Lukas Latacz, Werner Verhelst, and Hichem Sahli. Multimodal unit selection for 2D audiovisual text-to-speech synthesis. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 125–136. Springer, 2008.

[124] Wesley Mattheyses and Werner Verhelst. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217, 2015.

[125] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[126] Javier Melenchón, Elisa Martínez, Fernando De La Torre, and José A Montero. Emphatic visual speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):459–468, 2009.

[127] Baback Moghaddam and Alex Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[128] Hiroyasu Murakami and BVK Vijaya Kumar. Efficient calculation of primary images from a set of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):511–515, 1982.

[129] NTSC. National Television Systems Committee (NTSC) standard, 1953.

[130] Joern Ostermann and Axel Weissenfeld. Talking faces-technologies and applications. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 826–833. IEEE, 2004.

[131] Slim Ouni, Michael M Cohen, and Dominic W Massaro. Training Baldi to be multilingual: A case study for an Arabic Badr. *Speech Communication*, 45(2):115–137, 2005.

[132] Slim Ouni, Dominic W Massaro, Michael M Cohen, Karl Young, and Alexandra Jesse. Internationalization of a talking head. In *Proc. of 15th International Congress of Phonetic Sciences, Barcelona, Spain*, pages 286–318, 2003.

[133] Frederic I Parke. A model for human faces that allows speech synchronized animation. *Computers Graphics*, 1(1):3–4, 1975.

[134] Frederic I Parke. Parameterized models for facial animation. *IEEE computer graphics and applications*, 2(9):61–68, 1982.

[135] Frederic I Parke and Keith Waters. *Computer facial animation*. CRC Press, 2008.

[136] Frederick I Parke. Computer generated animation of faces. In *Proceedings of the ACM annual conference*, pages 451–457. ACM, 1972.

[137] Frederick Ira Parke. A parametric model for human faces. Technical report, DTIC Document, 1974.

[138] Jonathan Parker, Ranniery Maia, Yannis Stylianou, and Roberto Cipolla. Expressive visual text to speech and expression adaptation using deep neural networks. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4920–4924. IEEE, 2017.

[139] Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N Gowdy. CUAVE: A new audio-visual database for multimodal human-computer interface research. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–2017. IEEE, 2002.

[140] Hai X Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: a deep learning approach. In *The 1st DALCOM workshop, CVPR*, 2017.

[141] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, 2003.

[142] Gerasimos Potamianos, Chalapathy Neti, Giridharan Iyengar, and Eric Helmuth. Large-vocabulary audio-visual speech recognition by machines and humans. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[143] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[144] Lawrence R Rabiner and Bernard Gold. *Theory and application of digital signal processing*. 1975.

[145] C Rasmussen and C Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. 2005.

[146] Carl Edward Rasmussen and Hannes Nickisch. Gaussian processes

for machine learning (GPML) toolbox. *Journal of Machine Learning Research*, 11:3011–3015, 2010.

[147] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT Press Cambridge, 2006.

[148] Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia. In *Advances in neural information processing systems*, pages 176–183, 1994.

[149] Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 25(2-3):117–149, 1996.

[150] Joseph Rothweiler. Polyphase quadrature filters–a new subband coding. In *IEEE International Conference on Speech and Signal Processing (ICASSP)*, volume 8, pages 1280–1283, 1983.

[151] Najmeh Sadoughi and Carlos Busso. Novel realizations of speech-driven head movements with generative adversarial networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada*, pages 6169–6173, 2018.

[152] Najmeh Sadoughi and Carlos Busso. Speech-driven expressive talking lips with conditional sequential generative adversarial networks. *IEEE Transactions on Affective Computing*, 2019.

[153] Payam Saisan, Alessandro Bissacco, Alessandro Chiuso, and Stefano Soatto. Modeling and synthesis of facial motion driven by speech. *European Conference on Computer Vision (ECCV)*, pages 456–467, 2004.

[154] Mathieu Salzmann, Carl H Ek, Raquel Urtasun, and Trevor Darrell. Factorized orthogonal latent spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 701–708, 2010.

[155] Matthew R Scott, Xiaohua Liu, and Ming Zhou. Towards a specialized search engine for language learners. *Proceedings of the IEEE*, 99(9):1462–1465, 2011.

[156] Eftychios Sifakis, Andrew Selle, Avram Robinson-Mosher, and Ronald Fedkiw. Simulating speech with a physics-based facial muscle model. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 261–270. Eurographics Association, 2006.

[157] Lawrence Sirovich and Michael Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America. A (JOSA A)*, 4(3):519–524, 1987.

[158] Yang Song, Jingwen Zhu, Xiaolong Wang, and Hairong Qi. Talking Face Generation by Conditional Recurrent Adversarial Network. *arXiv preprint arXiv:1804.04786*, 2018.

[159] M Bille Stegmann. Active appearance models: Theory, extensions and cases. Master's thesis, Technical University of Denmark, 2000.

[160] Stanley Smith Stevens, John Volkmann, and Edwin B Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[161] William H Sumby and Irwin Pollack. Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215, 1954.

[162] Marc Swerts and Emiel Krahmer. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1):81–94, 2005.

[163] Marc Swerts and Emiel Krahmer. The importance of different facial areas for signalling visual prominence. In *International Conference on Spoken Language (ICSLP)*, pages 1280–1283. Pittsburgh, 2006.

[164] Sarah Taylor, Akihiro Kato, Ben Milner, and Iain Matthews. Audio-to-visual speech conversion using deep neural networks. pages 1482–1486, 2016.

[165] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):93, 2017.

[166] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284. Eurographics Association, 2012.

[167] A Murat Tekalp and Joern Ostermann. Face and 2-D mesh animation in MPEG-4. *Signal Processing: Image Communication*, 15(4):387–421, 2000.

[168] Lucas Terissi, Mauricio Cerda, Juan C Gomez, Nancy Hitschfeld-Kahler, Bernard Girau, and Renato Valenzuela. Animation of generic 3D head models driven by speech. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2011.

[169] Barry-John Theobald. *Visual speech synthesis using shape and appearance models*. PhD thesis, University of East Anglia, 2003.

[170] Barry-John Theobald, J Andrew Bangham, Iain A Matthews, John RW Glauert, and Gavin C Cawley. 2.5 D Visual Speech Synthesis Using Appearance Models. In *BMVC*, pages 1–10, 2003.

[171] Barry-John Theobald, Sascha Fagel, Gérard Bailly, and Frédéric Elisei. LIPS2008: Visual speech synthesis challenge. In *Interspeech*, pages 2310–2313, 2008.

[172] Barry-John Theobald and Iain Matthews. Relating objective and subjective performance measures for aam-based visual speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2378–2387, 2012.

[173] Barry-John Theobald and Nicholas Wilkinson. A real-time speech-driven talking head using active appearance models. In *International Conference on Auditory-Visual Speech Processing(AVSP)*, pages 264–269, 2007.

[174] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[175] Michalis K Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.

[176] Michalis K Titsias and Neil D Lawrence. Bayesian Gaussian process latent variable model. In *International Conference on Artificial Intelligence and Statistics*, pages 844–851, 2010.

[177] Pete Trautman. Manifold relevance determination: Learning the latent space of robotics. *arXiv preprint arXiv:1705.03158*, 2017.

[178] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[179] Virginie Van Wassenhove, Ken W Grant, and David Poeppel. Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4):1181–1186, 2005.

[180] Konstantinos Vougioukas, Samsung AI Center, Stavros Petridis, and Maja Pantic. End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 37–40, 2019.

[181] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal gans. *arXiv preprint arXiv:1805.09313*, pages 1–14, 2018.

[182] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *arXiv preprint arXiv:1906.06337*, 2019.

[183] Guang-Yi Wang, Mau-Tsuen Yang, Cheng-Chin Chiang, and Wen-Kai Tai. A talking face driven by voice using hidden markov model. *Journal of information science and engineering*, 22(5):1059–1075, 2006.

[184] Jack Wang, Aaron Hertzmann, and David M Blei. Gaussian process dynamical models. In *Advances in neural information processing systems*, pages 1441–1448, 2005.

[185] Lijuan Wang, Xiaojun Qian, Wei Han, and Frank K Soong. Photo-

real lips synthesis with trajectory-guided sample selection. In *Speech Synthesis Workshop(SSW)*, pages 217–222, 2010.

[186] Lijuan Wang, Xiaojun Qian, Wei Han, and Frank K Soong. Synthesizing photo-real talking head via trajectory-guided sample selection. In *Interspeech*, volume 10, pages 446–449, 2010.

[187] Lijuan Wang, Xiaojun Qian, Lei Ma, Yao Qian, Yining Chen, and Frank K Soong. A real-time text to audio-visual speech synthesis system. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[188] Lijuan Wang, Yao Qian, Matthew Scott, Gang Chen, and Frank Soong. Computer-assisted audiovisual language learning. *Computer*, 45(6):38–47, 2012.

[189] Lijuan Wang and Frank K Soong. HMM trajectory-guided sample selection for photo-realistic talking head. *Multimedia Tools and Applications*, 74(22):9849–9869, 2015.

[190] Lijuan Wang, Yi-Jian Wu, Xiaodan Zhuang, and Frank K Soong. Synthesizing visual speech trajectory with minimum generation error. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4580–4583. IEEE, 2011.

[191] Xiaoou Wang, Thomas Hueber, and Pierre Badin. On the use of an articulatory talking head for second language pronunciation training: the case of Chinese learners of French. In *10th International Seminar on Speech Production (ISSP)*, pages 449–452, 2014.

[192] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen,

Samy Bengio, et al. Tacotron: A fully end-to-end text-to-speech synthesis model. *arXiv preprint arXiv:1703.10135*, 2017.

[193] Keith Waters. A muscle model for animation three-dimensional facial expression. *Acm siggraph computer graphics*, 21(4):17–24, 1987.

[194] Keith Waters and Tom Levergood. An automatic lip-synchronization algorithm for synthetic faces. In *Proceedings of The second ACM international conference on Multimedia*, pages 149–156, 1994.

[195] R Weide. The CMU pronunciation dictionary, release 0.6. *Carnegie Mellon University*, 1998.

[196] Lei Xie and Zhi-Qiang Liu. Speech animation using coupled hidden markov models. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1128–1131. IEEE, 2006.

[197] Lei Xie and Zhi-Qiang Liu. A coupled HMM approach to video-realistic speech animation. *Pattern Recognition*, 40(8):2325–2340, 2007.

[198] Lei Xie, Naicai Sun, and Bo Fan. A statistical parametric approach to video-realistic text-driven talking avatar. *Multimedia tools and applications*, 73(1):377–396, 2014.

[199] Lei Xie, Lijuan Wang, and Shan Yang. Visual speech animation. 2016.

[200] Yifan Xing, Rahul Tewari, and Paulo Mendonca. A self-supervised bootstrap method for single-image 3d face reconstruction. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1014–1023. IEEE, 2019.

[201] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. chapter Understanding Belief Propagation and Its Generalizations, pages 239–269. 2003.

[202] Steve Young, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, and Dan Povey. *The HTK book Version 3.4*. Cambridge University Engineering Department, Cambridge, 2006.

[203] Shen Zhang, Zhiyong Wu, Helen M Meng, and Lianhong Cai. Facial expression synthesis using PAD emotional parameters for a Chinese expressive avatar. In *International Conference on Affective Computing and Intelligent Interaction*, pages 24–35. Springer, 2007.

[204] Jian Zhao, Wang Lirong, Zhang Chao, Shi Lijuan, and Yin Jia. Pronouncing rehabilitation of hearing-impaired children based on Chinese 3D visual-speech database. In *International Conference on Frontier of Computer Science and Technology (FCST)*, pages 625–630. IEEE, 2010.

[205] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. *arXiv preprint arXiv:1807.07860*, 2018.

[206] Xiaodan Zhuang, Lijuan Wang, Frank K Soong, and Mark Hasegawa-Johnson. A minimum converted trajectory error (MCTE) approach to high quality speech-to-lips conversion. In *Interspeech*, pages 1736–1739, 2010.