THE UNIVERSITY
of EDINBURGH

Head Motion Synthesis: Evaluation and a Template Motion Approach

David Adam Braude

Doctor of Philosophy

Institute for Language, Cognition and Computation

School of Informatics

University of Edinburgh

2016

Abstract

The use of conversational agents has increased across the world. From providing automated support for companies to being virtual psychologists they have moved from an academic curiosity to an application with real world relevance. While many researchers have focused on the content of the dialogue and synthetic speech to give the agents a voice, more recently animating these characters has become a topic of interest. An additional use for character animation technology is in the film and video game industry where having characters animated without needing to pay for expensive labour would save tremendous costs.

When animating characters there are many aspects to consider, for example the way they walk. However, to truly assist with communication automated animation needs to duplicate the body language used when speaking. In particular conversational agents are often only an animation of the upper parts of the body, so head motion is one of the keys to a believable agent. While certain linguistic features are obvious, such as nodding to indicate agreement, research has shown that head motion also aids understanding of speech. Additionally head motion often contains emotional cues, prosodic information, and other paralinguistic information.

In this thesis we will present our research into synthesising head motion using only recorded speech as input. During this research we collected a large dataset of head motion synchronised with speech, examined evaluation methodology, and developed a synthesis system.

Our dataset is one of the larger ones available. From it we present some statistics about head motion in general. Including differences between read speech and story telling speech, and differences between speakers. From this we are able to draw some conclusions as to what type of source data will be the most interesting in head motion research, and if speaker-dependent models are needed for synthesis.

In our examination of head motion evaluation methodology we introduce Forced Canonical Correlation Analysis (FCCA). FCCA shows the difference between head motion shaped noise and motion capture better than standard methods for objective evaluation used in the literature. We have shown that for subjective testing it is best practice to use a variation of MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) based testing, adapted for head motion. Through experimentation we have developed guidelines for the implementation of the test, and the constraints on the length.

Finally we present a new system for head motion synthesis. We make use of simple templates of motion, automatically extracted from source data, that are warped to suit the speech features. Our system uses clustering to pick the small motion units, and a combined HMM and GMM based approach for determining the values of warping parameters at synthesis time. This results in highly natural looking motion that outperforms other state of the art systems. Our system requires minimal human intervention and produces believable motion. The key innovates were the new methods for segmenting head motion and creating a process similar to language modelling for synthesising head motion.

Acknowledgements

Firstly I wish to thank my supervisor Dr. Hiroshi Shimodaira for his advice, support, and guidance throughout my Ph.D. and particularly during the writing of this thesis. Second I wish to thank my annual review panel, Dr. Taku Komura and Prof. Simon King, who kept me on track and made sure my research was appropriate, relevant, and interesting. I also wish to thank all the members of The Centre for Speech Technology Research (CSTR) in the School of Informatics for their companionship and help.

Finally I wish to thank my father, Colin Braude, my sisters Talia Talmud, Avri Spilka-Drewnicki, and Ilana Spilka, my partner Claire Giry, and my friends Benjamin Rosman, Nicole Abvajee, Helle Hang, Philipp Petrenz, Peter Orchid, and Kat McNabb for their immense emotional (and at times financial) support and advice throughout my Ph.D.

Financial acknowledgements

This work was supported by EU FP7 SSPNet (grant agreement no. 231287)

The financial assistance of the South African National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*David Adam Braude*)

Dedicated to my family,

the memories of those who have gone,

and the memories that I make with those still with me.

Table of Contents

List of Figures

List of Tables

Supporting Publications

Chapter 3: Data Collection and Statistics of Head Motion

Braude D. A., Shimodaira H., Ben Youssef A., "The University of Edinburgh Head-Motion and Audio Storytelling (UoE-HAS) Dataset", *Intelligent Virtual Agents*, pp. 466–467, 2013

Chapter 4: Methodology for Head Motion Synthesis Evaluation

Ben Youssef A., Shimodaira H., Braude D. A., "Articulatory Features for Speech-Driven Head Motion Synthesis", *Interspeech*, pp. 2758–2762, 2013

Chapter 5: Temple - Warping Based Head Motion Synthesis

Braude D. A., Shimodaira H., Ben Youssef A., "Template-Warping Based Speech Driven Head Motion Synthesis", *Interspeech*, pp. 2763–2767, 2013

Ben Youssef A., Shimodaira H., Braude D. A., "Head Motion Analysis and Synthesis over Different Tasks", *Intelligent Virtual Agents*, pp. 285–294, 2013

Ben Youssef, A., Shimodaira, H. and Braude, D. A., "Speech Driven Talking Head from Estimated Articulatory Features", *in Proc. Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4573–4577, 2014

Chapter 1

Introduction

For millions of years, mankind lived just like the animals.
Then something happened which unleashed the power of our imagination.
We learned to talk and we learned to listen.

Stephen Hawking, 1993

Communication is the cornerstone of society. We not only communicate with each other but with everything we encounter, from dogs to computers we are constantly interacting and thus communicating with our environment. When discussing communication we mostly think of speech, but that is only one part of a multifaceted system. A far older method is the use of body language, and we still use it today to 'talk', especially if we cannot use speech. For instance we immediately know that a dog is curious when it tilts its head to the side. While not a word was spoken, an exchange of information took place. Non-verbal communication is an often neglected area of research in the greater field of human computer interaction.

Since the beginning of modern computing it has been a goal to be able to communicate with computers as naturally as talking to another human. The famous Turing test is perhaps the best known expression of this desire. This test for intelligence will be passed when a human is talking to two conversational agents and they cannot tell which

agent is the computer and which is another human. While it is obvious that the content of the dialogue is perhaps the most important aspect to this test, to truly have the same level of interaction, a computer must also be able to listen, speech, and move. This last aspect, body language, is the concern of this thesis. Though in this case we will limit ourselves to animated models.

When we talk there are many types of movement; arm gestures, posture, lip movement, and facial expression are all examples. What this thesis aims to present is a method for animating the movement of the speaker's head. Previously head motion has been shown to be an important communication channel that increases intelligibility (Munhall et al., 2004). Furthermore, even during human to human communication, the lack of natural visual queues, such as head motion during video-conferencing, severally reduces the perceived quality of communication (Suwita et al., 1997). This is due to the fact that head motion provides semantic meaning, important conversational clues, and expression (McClave, 2000). Examples of each would be shaking to indicate disagreement, turning to face the next speaker, and vigorous motion to show enthusiasm for the topic, respectively. Clearly heads are important.

While the first two types of motion above rely on higher-level cognitive processes, it is possible that the last, the expression of the speaker, could be predicted directly from other modalities of communication, for example the verbal parts of the speech. Apart from expression there are other influences on head motion such as the movements of the speaker's articulators. Articulators are the parts of the body that are responsible for speech production such as the lips, tongue, and jaw. Table 1.1 shows some additional examples of categories of head motion during speech and what would cause them. These are based on the findings by Hadar et al. (1983), McClave (2000), and Ishi et al. (2013), renamed to emphasise the origin of the cause of the production. There are of course other types of motion, for instance the head moves along with the body when the speaker adjusts their seating position or when they are laughing. The rows in

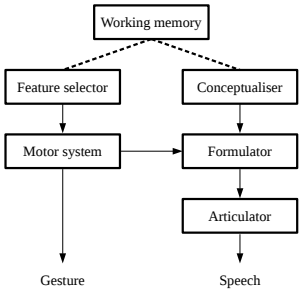
Table 1.1: Types of head motion as part of speech and their causes. Rows highlighted are the focus of the thesis

Type	**Examples**	**Cause**
Semantic	Nodding or shaking without speech	Direct desire to convey a particular meaning by substituting spoken words for gestures.
Dialogue cues	Turning to face the next speaker	Used to facilitate conversation.
Back-channel	Tilting	Serves the same function as spoken back-channel, mostly to indicate attention.
Cognitive	Upwards tilt during disfluency	Used to indicate cognitive processes such as recall or problem solving
Emphatic	Nodding while speaking	Movements used to emphasis the content of the speech.
Motor	Centring the head	Needed for comfort or to allow movement of the articulators such as the jaw

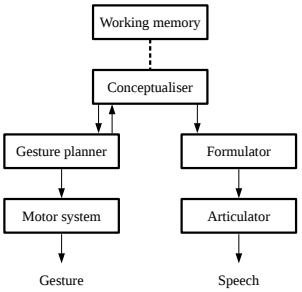
Table 1.1 that are highlighted are the types of motion that this thesis focuses on trying to synthesise.

The dependencies for gestures during speech has been studied previously. In the Wagner et al. (2014) survey they show several models that have been proposed by other authors, diagrammatically they are illustrated in Figure 1.1. In terms of this thesis the closest model to the one that describes the dependencies we have assumed is the one from de Ruiter (2000)). However, it still needs to be adapted to include some additional dependencies based on the emotions of the speaker, and specifically with head motion the dependencies between the articulators and the motor system need to be made more explicit to explain the origins of the motion we are trying to synthesise.

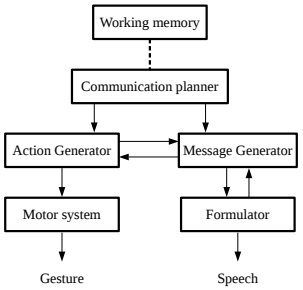
The primary aim of this thesis is to present a method for mapping speech to natural,



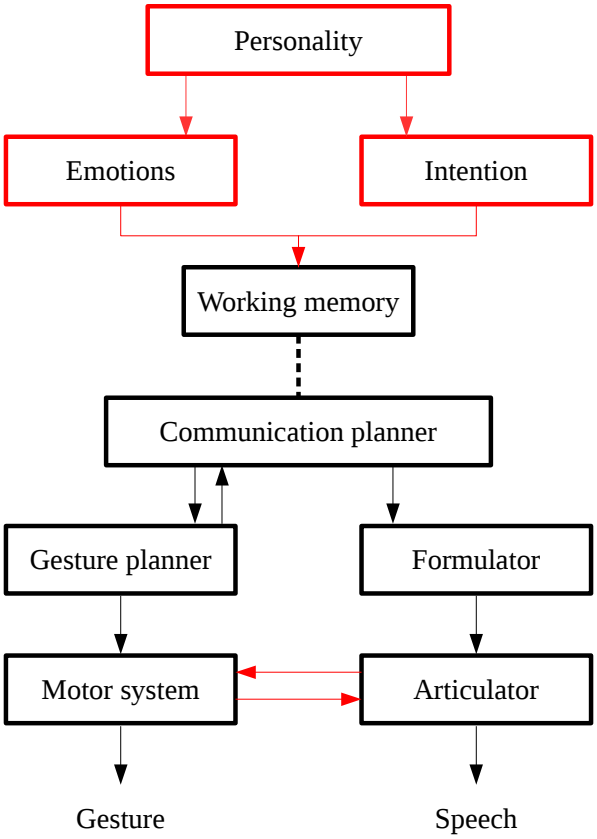
(a) Krauss and Hadar (1999)



(b) de Ruiter (2000)



(c) Kita and Özyürek (2003)



(d) Gesture production with additional dependencies

Figure 1.1: Different models of gesture and speech production; (a), (b), and (c) reproduced from (Wagner et al., 2014); (d) is proposed by this thesis to illustrate additional dependencies shown in red.

expressive head motion and show the mapping's validity and necessity, with the caveat that this does not include semantic gestures. In addition we will present what aspects of head motion need to be considered, what features of speech to use, how to evaluate

the synthesised motion, and speaker and task dependence. An important consideration is that the head motion must be synchronised with the speech, this goes to its believability. In Figure 1.1 what we are focusing on are the parts of the head motion that are generated by the processes and influences indicated by the red boxes and arrows. This is because they are reflected in the speech as well as the motion.

There are many applications for this research. Conversational agents are increasingly being used for a range of applications from interaction with corporate websites to virtual psychologists and currently embodying them (creating an avatar) is becoming more popular. This is because a body increases the levels of engagement with the agent (Cassell and Thorisson, 1999; Bickmore and Cassell, 2005). Additionally head motion synthesis also has applications in both computer games and films. By giving a rough draft as to how the head could move, the animation process becomes shorter. This is especially important in the independent sector of those industries, where budgets for animation are lower. In this case starting with something believable will improve the quality of animations while keeping costs low.

At a higher level our approach to synthesis is to treat the head as a Three Degree-Of-Freedom (3DOF) rigid object, with movement specified as rotations, this is a common approach as will be shown in Chapter 2. We then break the movement into patterns found in recordings of multiple speakers. Motion is then synthesised by estimating parameters that describe these patterns. This is done using a data driven model. Where this approach distinguishes itself from the existing body of research is in how the patterns are determined, and the model's structure.

As it is both common and intuitive we will be treating head motion synthesis as an optimisation problem that is probabilistic in nature. To model the probabilities involved we will utilise a structure that is commonly used in speech technologies which is the Hidden Markov Model (HMM). Speech technologies include Automatic Speech Recognition (ASR), speech synthesis, and speech to animation. Due to the HMM's ability

to take into account temporal and contextual information it is well suited to speech technology. A more detailed description of HMMs is presented in Section 2.2.2.

There are many modifications to the basic structure of HMMs that have been proposed. In this research we have made extensive use of a hierarchical model. One of the advantages of using a hierarchical model is that each level can be examined individually for speaker dependence and the level of that dependence. For example the patterns in the data might be common to all speakers, but how often each pattern appears is individual. Another advantage of a hierarchical model is that allows for the flexibility in the exact model used at each level, for instance in a particular layer a Gaussian Mixture Model (GMM) might be better suited than an HMM.

There are several problems with HMMs. The largest is that in their basic form they are too flexible. By this we mean that any trajectory is possible. However in head motion there are a limited amount of patterns that can be found in the data. In speech synthesis this is tackled by using the first and second derivatives but this approach is not well suited to head motion synthesis. The dependencies in head motion are much longer than in speech synthesis as the head moves much slower than articulators. This thesis presents an alternative method of using HMMs for higher level decisions and not for synthesising trajectories frame-by-frame. Instead the details are taken from existing patterns. One could view this as a combination of an HMM and unit selection approach. However, true unit selection would require that there are natural boundaries in head motion analogous to phonemes in speech. Unfortunately this is not the case.

For head motion there is no natural method for segmentation similar to phonemes for speech. Instead we propose a rule based segmentation that will mean that there is no need for detailed manual annotation. A fringe benefit of this approach is that the amount of optimisations is reduced which will speed up processing time.

In addition to speaker dependence we also look at task dependence. Different tasks include storytelling, reading, or dialogue. This thesis focuses on the first two which

reduces the complexity of the problem by not including the effects of another speaker. Collection of data for these tasks is also discussed as previously no appropriate large dataset existed. We will present how we collected this new data in terms of hardware, software, physical layout, and the method for eliciting speech from the participants.

While the design of a head motion synthesis system is important, it is not useful unless it is capable of good performance in both subjective and objective evaluation. Due to the nature of possible applications of this model it is obvious that no matter how well it theoretically performs, if people are not convinced or find it 'creepy', then the method is not a valid approach, thus subject testing is needed. However, when examining the literature we could not find a formal study on the subjective testing methodology. We present our research into how long the test should be, the type of animation used for the model, who could participate in the evaluation and if participants need to come to a laboratory or if they could conduct the evaluation over the internet.

Objective measures on the other hand can show the individual performance of different layers of the hierarchy and show if the approach is sound before subjective testing. Part of this thesis will discuss available objective measures and their suitability for this research. We will then present our own objective measure that better reflects the performance of a head motion synthesis system than the current practice.

Mapping from speech to head motion requires that a certain set of assumptions are true, and it is useful to take note of some of them. The most important is that head motion is at least partially predictable from speech. This does not necessarily mean that it is predictable at all times based purely on speech. There are many underling psychological and linguistic factors that are not available directly from speech features. This assumption is not completely unreasonable, for instance we are able to imagine how a person we are talking to on the telephone is moving. Also part of head motion is related to how our articulators move, for instance it unlikely the head is looking down while the jaw is wide open.

On the other hand psychological and linguistic features would definitely influence the motion. Their exclusion reduces the predictability of head motion and means that it would not be deterministic, as differences in the psycho-linguistic context may not reflect in the voice, but in the body language. Despite this, the fact that head motion and speech features are related has been examined before and while there is no linear mapping there does seem to be a link between the two (Yehia et al., 2002; Ishi et al., 2013). A second assumption coupled to the first is that given not all of the head motion will be predictable, sufficiently large parts will be so that the inclusion of this mapping would be useful. A moderate gain in objective and subjective measures compared to shaped noise would prove this assumption is correct and thus that this system would have practical applications.

This does raise the point that the type of speech features taken into account is important. With the wide variety available, some are undoubtedly better predictors of head motion than others. Part of this thesis will focus on the choice of speech features.

A potential problem is that there is a low direct correlation between speech features and head motion trajectories at a frame-wise level (Hofer, 2009; Ben Youssef et al., 2013a). On short time-scales there is a high correlation (Yehia et al., 2002; Busso et al., 2005). Thus, as Hofer (2009) showed with user defined modelling units, recognition of modelling units is possible over a short period, but a frame-wise would not have the same performance. Consequently it is important that the correct modelling unit is chosen.

Hofer (2009) used hand annotation to find human-meaningful motion segments in the head motion, for example nodding and shaking. This manual annotation is a time consuming process and is subject to error. This leads to the investigation of data driven modelling units. This thesis presents one such modelling unit and shows that it is capable of capturing the nuances of head motion and that it predictable from speech features.

In the following sections we will formalise the objectives of the research. Additionally we will show the scope of the research, briefly discuss some the challenges that needed to be overcome, and layout the structure of the thesis.

1.1 Objectives and Scope

One way of viewing a speech to head motion system is as a black-box. Speech goes into the system, and head motion is predicted. Internal to this black box, the speech is parametrised, mapped to head motion parameters, and then those parameters are converted into the animation. With that in mind the primary goal of this thesis is to:

> **Find a method to synthesise head motion from speech that will show emphasis and be consistent with physical considerations.**

This can be broken down into three distinct parts:

- Find a method to parametrise speech for head motion synthesis.
- Find a method to parametrise head motion suitable for analysis, synthesis, and animation.
- Find a method for predicting head motion parameters from the parametrised speech.

There are many standard methods for parametrising speech that are used in ASR and speech synthesis, such as extracting the Mel-frequency Cepstral Coefficients (MFCC). In addition there are other features that can be estimated from the speech features like the position of articulators. The various options are discussed in Section 2.2.4. Based on how the model is designed and the head motion parametrisation a secondary objective would be to find the best choice of speech features.

With regards to the parametrisation of the head motion there are many ways to describe rotation of three dimensional objects. We will discuss some of the more common

options in Section 2.2.3. There we will also explain why we chose to use rotation vectors to represent the head motion in our system instead of Euler angles or any other common system used in 3D graphics.

Additionally to how the rotation is represented, head motion can also be parametrised by dividing it into segments or modelling units. This would be analogous to phonemes in speech. A natural interpretation can be a nod or a shake, however, this is difficult to segment without manual annotation. Instead automatic segmentation is desired, this can be achieved with a simple rule based approach, providing that the rule is universal to all speakers and does not need to be adapted. A modelling unit is desirable because it would be difficult to have a universal frame-by-frame mapping. This frame-wise approach has been attempted before and will be discussed in Section 2.3. Doing this approach is very similar to unit selection if the segments are copied, with some transform to create the head motion. Due to the small number of patterns observed in the dataset this is the approach that this thesis follows. However, it must be proved that there are only a small amount of patterns. This should be achieved in a data driven approach. In other words the number of patterns should be determined from data. Additionally the shape of the pattern should be determined by the data.

The speech to head motion on the other hand has to be able to determine the correct modelling unit to use, based on the speech features. As was mentioned previously this thesis will make use of a hierarchical model, based around HMMs. But there are many aspects to how it would need to be structured, and no implementation of any machine learning model is transfer to implement in a new domain. So the challenges involved must be addressed. The design elements and general theory of HMMs are covered in Section 2.2.2.

It is important to use a machine learning approach rather than a rule based one, because it would mean that speaker individuality can be synthesised without much extra intervention. In a rule based paradigm for each new speaker a new set recordings would

be made and from that a new set of rules would need to be drawn up. With a machine learning approach the data would be captured and then the model retrained which is far less labour intensive. In addition a machine learning approach can combine different recordings from different people to create a new 'personality'. Alternatively one could still make believable head motion from less data by bootstrapping models from a larger dataset this is known as speaker adaptation in speech synthesis.

As the model's usability relies on its performance, objective evaluation methods are only valid if it indicates how users or subjective evaluators would rate the quality of the final synthesis. While there are standard approaches for evaluating head motion objectively in the literature there has previously been no testing on how well they predict the quality of synthesised head motion. With regards to the subjective evaluation there are many factors, such as the realism of the model for animation, and the format of the testing that need to be considered.

Thus to address the three parts of the objective stated at the start of the section the following points must be considered:

- Find a suitable model, whose parameters are learnt from data, that maps the speech to the head motion.
- Find a data driven head motion modelling unit that can be estimated from the head motion recording
- Determine the optimal choice of speech features.
- Prove the model's validity through both objective and subjective evaluation. This in turn requires:
  - Testing the appropriateness of objective measures.
  - Developing new measures as needed.
  - Designing and conducting subjective tests.

To address these issues there is an additional task which is to find a suitable dataset. While the need for data is obvious there are not many datasets that are available for this research and those that existed previously are often not very large. With a larger dataset it is possible to test how dependant head motion is on the speaker, and how dependant it is on the task. For instance one can expect that people move their head differently if they are reading or telling a story from memory. It is ideal if the model can replicate the differences between the speakers, however, this would only be possible if there are differences so this must be measured. Expanding on this dependence it has been shown that people are able to identify the gender of a speaker based on head motion alone (Hill and Johnston, 2001). We can examine the dataset to see if these differences can be detected in objective statistics and see how speakers relate to one another. Additionally if we collect data from multiple tasks we can see if there are dependencies based on those tasks too, and if speaker or task dependency is greater. The dataset that we collected contains two tasks and we will show our evidence that there is both speaker and task dependence and that the speaker dependency is greater than the task dependency.

It is also important to note what falls outside the scope of the research. Firstly semantic gestures will not be directly synthesised as this would require knowing the intention of the speaker which is not possible using the speech features alone. In addition linguistic information will not be used as input information as this would either require hand transcription which is time consuming and expensive, or ASR which is not yet capable of creating highly accurate transcription of new speakers without a large amount of training data. Thus by including linguistic information we would actually be limiting the applications in which this model could be used. A final limitation on this thesis is that it is only concerned with monologues. However, in principle the same types of models could be trained to synthesise head motion during a dialogue. This is primarily due to time constraints, as recording datasets is a long process. Additionally it is known that when humans speak their head motions influence each other, this would mean that

the head motion of the other participant would need to be taken into account, and that there are conversational cues that are indicated by head motion which are difficult to predict without psychological or linguist information.

To summarise the following are the contributions presented in this thesis:

- A novel approach to synthesis using motion templates, with amplitude and time warping which outperforms existing state of the art systems in a subjective test.
- Recommendations for both subjective and objective head motion synthesis evaluation, specifically that MUSHRA testing should be used and providing a framework to conduct subjective tests.
- Collection of a large dataset of head motion motion capture synchronised with speech.
- A demonstration that head motion is both speaker and task dependent and that it more speaker dependent than task dependent.

1.2 Thesis Structure

Chapter 2: Details the underling theoretical concepts and the prior research into speech to head motion mapping

Chapter 3: Contains the data collection method and basic statistics of the dataset. Additionally it discuses dependencies, and correlations which show that head motion synthesis is speaker dependant, and that the task for participants effects the head motion. This is the third and forth points listed above.

Chapter 4: Explains both the objective and the subjective evaluation methodology. We will present the results of some experiments that show that realistic models should be used during evaluation, that the length of a test should be approximately 30 samples long, that participants can evaluate motion from home, and that non-native speakers

can be used for subjective evaluations. Additionally we show a new method for objective evaluation that predicts the results of a subjective test better than Conical Correlation Analysis which is the current preferred method. This covers second point in the list above.

Chapter 5: Presents a novel approach to synthesis using motion templates. Additionally this chapter presents the evaluation and improvements made on the system. This covers the first point above. We show that the system gives results that compare favourably to motion capture and out performs other sate of the art systems.

Chapter 6: Provides a discussion of the results some concluding remarks and possible directions for future research.

The Appendix contains details of results that are presented elsewhere in the thesis.

Chapter 2

Theoretical Concepts and State of the Art Head Motion Synthesis

2.1 Links to Speech of Head Motion

Head motion synthesis forms part of the larger field of character animation. The need for good non-verbal communication channels, such as head motion, cannot be understated for use in embodied characters. With the correct body language an embodied conversational agent or any other animated character can appear, "credible, trustworthy, confident, and non-threatening" (André et al., 2011).

Body language does not only provide emotional content and a sense of the personality of the agent. Additionally understanding of spoken words is enhanced with the correct body language (Wagner et al., 2014). Specifically 'visual prosody' has been shown to increase understanding (Munhall et al., 2004) and visual prosody has in turn been shown to synchronise with spoken prosody (Wagner et al., 2014).

McClave (2000) showed that head motion in particular has linguistic content and that speakers are very sensitive to head motion that is out of alignment with spoken parts

of communication.

Graf et al. (2002) showed that head motion is related to pitch accents. They also demonstrated that the inclusion of realistic head motion improves engagement with 3D avatars. This provides a good direction for a head motion system, as pitch accents can be predicted from speech features (Ladd et al., 1999). This in turn means that it should be possible to predict head motion from speech.

More recently Ishi et al. (2013) showed explicitly that head motion that has linguistic context in dialogues are aligned with speech events. Thus combing the independent findings of McClave (2000) and Graf et al. (2002).

Head motion has also been shown to be strongly linked to speech features especially F0 (a measurable approximation of speech see Section 2.2.4) (Kuratate et al., 1999). Head motion from speech has been attempted before, however, we will cover the existing attempts at head motion synthesis in Section 2.3.

In the rest of this chapter we present some of the most important techniques and concepts at a theoretical level. First we will show the machine learning techniques that we used in the proposed synthesis system. We will then describe some details about how rotation can be represented and explain why we chose to use rotation vectors. Moving onto head motion synthesis in particular we will explain some of the speech features that are available, such as F0, that can be the inputs to the head motion synthesiser. Finally we will provide a review of some existing methods for head motion synthesis, focusing on HMM based systems and a particular system that we will use as a baseline.

2.2 Key Theoretical Concepts

2.2.1 Gaussian Mixture Models

Distributions of data are often approximated by a multivariate Gaussian distribution. However, it is not always the case that this is a good approximation. Much like any signal can be described as the addition of sinusoids of different frequencies and amplitudes by the Fourier transform, any distribution can be described as the addition of several multivariate Gaussian distributions with different means, covariance matrices, and mixing coefficients (weights). This combination is known as a Gaussian Mixture Model (GMM) or a mixture of Gaussians (Bishop, 2006, p. 110).

Similar to how a Riemann sum becomes a better approximation of the integral as the number of partitions increase, a GMM becomes a better approximation of the original distribution as the amount of components increases. However, along with this increase in accuracy, as the amount of components increases, so too does the number of training examples required and the length of time needed to train the model. Also the increase in accuracy becomes negligible beyond a certain number of components. One way to figure out the optimal amount of components is the Bayesian Information Criterion (BIC) (Bishop, 2006, p. 216-217). The BIC is defined as

$$BIC = -2\ln P(\boldsymbol{X} \mid \hat{\boldsymbol{\lambda}}_k) + k\ln(T), \tag{2.1}$$

where $\boldsymbol{X}$ is the observed data, T is the number of observations, k is the number of free parameters and $\hat{\boldsymbol{\lambda}}_k$ is the fitted model with k free parameters. The BIC is actually defined for any model with a finite number of parameters. In the case of a GMM each time a new component is added the additional free parameters are the mean vector, covariance matrix, and mixing weight.

The BIC helps find the optimal number of components by rewarding accuracy and penalising model complexity. Using this measure a model with a lower BIC is con-

sidered to be a better trade-off between accuracy and complexity than a similar model with a higher BIC score. This is because the model with the higher BIC score did not accurately describe the distribution of the date because it has too few components. Alternatively the model had more than enough components and some were not increasing the accuracy of the model. So given two models the appropriate model to use is the one that minimises the BIC.

Due to its popularity for determining the amount of mixing components that should be used, many implementations of GMMs have the option to compute the BIC. For instance the Python library Scikit-learn, (Pedregosa et al., 2011) and the Matlab Statistics Toolbox[1] both calculate the BIC.

GMMs are often trained using Maximum Likelihood Estimation (MLE). As there is no closed form solution it is solved iteratively (Bishop, 2006, p. 112-113). For an M component GMM each mixture component m is a combination of a mixing weight α_m, a mean vector $\boldsymbol{\mu}_m$, and a covariance matrix $\boldsymbol{\Sigma}_m$ and all the parameters for all the components together are denoted $\boldsymbol{\lambda}$, The maximisation problem is:

$$\boldsymbol{\lambda}^* = \arg\max_{\boldsymbol{\lambda}} \; P(\boldsymbol{\lambda} \,|\, X) \tag{2.2}$$

$$= \arg\max_{\boldsymbol{\lambda}} \; \sum_{t=1}^{T} \ln \left(\sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{X} | \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \right). \tag{2.3}$$

Where $X = \{\boldsymbol{x}_t\}, t \in [0, T]$ are the training observations that are available. It should be noted that a GMM is capable of modelling multivariate data because the mean is a vector, and the covariance is a matrix. Also note that there are no temporal dependencies. If one wanted to include a temporal dependence then either the derivatives or multiple observations must be appended to the base observations and then be treated as though they were a single observation.

A GMM can be used as a generative model. Considering input observation $\boldsymbol{x}_t$ an output $\boldsymbol{y}_t$ can be estimated in two ways. The deterministic output is obtained by the

[1] `http://www.mathworks.co.uk/help/stats/gmdistributionclass.html`

minimisation of error or the maximisation of likelihood. Alternatively a GMM can generate an output by sampling from the marginal distribution of the output variables, given the inputs.

Toda et al. (2007, 2008) demonstrations an example of the minimisation approach which we will explain here. In this case the mapping from input to output uses the Minimum Mean Square Error (MMSE) criterion. The mean and covariance can be decomposed into separate parts for the input and the output as follows:

$$\boldsymbol{\mu}_m = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix} \qquad \boldsymbol{\Sigma}_m = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}$$

Then the marginal likelihood is given by

$$P(\boldsymbol{y}_t \,|\, \boldsymbol{x}_t, \boldsymbol{\lambda}) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}) P(\boldsymbol{y}_t|m, \boldsymbol{x}_t, \boldsymbol{\lambda}), \tag{2.4}$$

$$P(m \,|\, \boldsymbol{x}_t, \boldsymbol{\lambda}) = \sum_{m=1}^{M} \frac{\alpha_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)})}, \tag{2.5}$$

$$P(\boldsymbol{y}_t \,|\, m, \boldsymbol{x}_t, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{m,t}^{(y|x)}, \boldsymbol{\Sigma}_{m,t}^{(yy|x)}). \tag{2.6}$$

Where $\boldsymbol{\mu}_{m,t}^{(y|x)}$ is the marginal mean, and $\boldsymbol{\Sigma}_{m,t}^{(yy|x)}$ is the marginal covariance which are given by

$$\boldsymbol{\mu}_{m,t}^{(y|x)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}), \tag{2.7}$$

$$\boldsymbol{\Sigma}_{m,t}^{(yy|x)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \boldsymbol{\Sigma}_m^{(xy)}. \tag{2.8}$$

As mentioned above at this point one could sample from the marginal distribution, however, Toda et al. (2007, 2008) continue and derive the optimal values for the MMSE criterion. Using (2.5) and (2.7) the optimal output is given by

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} P(m \,|\, \boldsymbol{x}_t, \boldsymbol{\lambda}) \boldsymbol{\mu}_{m,t}^{(y|x)}. \tag{2.9}$$

It should be noted that the predicted values do not the maximise the observation likelihood (Toda et al., 2007, 2008), they minimise the prediction error. For the purposes of

this research the final goal was to minimise the Mean Square Error (MSE) so optimising the MMSE criterion is suitable. This will be discussed further in Chapter 5.

2.2.2 Hidden Markov Models

A Hidden Markov Model (HMM) is so called because it describes a process by using the Markov assumption and hidden states. The Markov assumption is that the next observation and the next state will only depend on the current state. The fact that the state sequence cannot be directly determined from the observations is what makes it hidden.

A standard HMM is composed of a transition matrix, A, which defines the probability from moving from one state to another. If it allows any state to transition to any other state then it is known as an 'ergodic' HMM, if it only allows specific state sequences, by restricting A to being an upper triangle matrix, then it is called a 'left-to-right' HMM. The rows of A must total to 1, because they are probabilities. Each state s_i is tied to its own observation probabilities. If the observations are continuous then it can be a Probability Density Function (pdf) such as a GMM, if the observations are discrete then it would be a multinomial distribution. If the observations are multivariate then they can be split into streams. By streaming the data we compute the observation probability of subsets of the observations separately, then combine them proportionately to their stream weights. This is especially appropriate if there is a mixture of continuous and discrete variables.

An HMM can:

- Determine the observation likelihood given the model parameters:

$$P(X \mid \boldsymbol{\lambda}) \tag{2.10}$$

- Estimate the most likely state sequence, Q, and find the state sequence probabil-

ity, given an observation sequence, X, and the model parameters, $\boldsymbol{\lambda}$:

$$Q^* = \arg\max_{Q} P(Q|X,\boldsymbol{\lambda}) \tag{2.11}$$

- Train the model parameters given a set of observed sequences $\boldsymbol{X}$, without requiring a known state sequence:

$$\boldsymbol{\lambda}^* = \arg\max_{\boldsymbol{\lambda}} P(\boldsymbol{\lambda}|\boldsymbol{X}) \tag{2.12}$$

To estimate the observation likelihood or find the most likely state sequence by examining all possible state sequences would be too computationally expensive, $O(T^k)$ for a T length observation and a k state HMM. However, both of these measures can be solved by much less costly algorithms. Observation likelihood can be estimated with the forward algorithm, and the optimal state sequence (and its likelihood) can be determined via using the Viterbi algorithm. At a higher level both of these algorithms are iterative and rely on the Markov assumption. This allows them to only consider the best possible way to achieve the current state and not all possible ways. These algorithms and their implementation are both discussed in detail by Rabiner (1989).

There are two standard approaches to training, the first is based on the Viterbi algorithm and the second is based on the forward - backward algorithm. The second approach is slower but produces better trained models (Bishop, 2006, pp. 625 – 631).

Like a GMM one can use an HMM as a generative model. While there is no closed form solution for a specified duration, there are some approximations such as Trajectory HMMs (Tokuda et al., 2004; Zen et al., 2007). Additionally one could use some input features as part of the sampling which is known as an Input-Output HMM (IOHMM) (Bengio and Frasconi, 1995).

There are other modification for HMMs such as restructuring them as a hierarchical models (HHMM). In this case each state is a self contained HMM (Fine et al., 1998). Or using Infinite HMMs where there are a countably infinite set of states (Beal et al., 2001) and the model is described using hyper-parameters.

HMMs are used extensively in speech technology for both recognition and synthesis. They have also been used previously in head motion synthesis (see Section 2.3.1) and are one of the core techniques used in this research.

2.2.3 Representing Rotation

When describing rotation there are multiple possible representations. The most common ones are using three angles, quaternions, axis-angle or rotation vectors, and rotation matrices. The three angles used in that representation are called Euler Angles. These are usually denoted by:

- α or φ – rotation about the *y*-axis
- β or θ – rotation about the *x*-axis
- γ or ψ – rotation about the *z*-axis

There are, however, a number of problems with the use of Euler Angles. The first is that there are singularities where that actual rotation of the object is ambiguous, would not affect head motion synthesis as this condition only happens at the poles and the the normal range of human head motion is not that large. A far greater problem is that they are order dependant. By this we mean that applying the rotations in the order α β γ is not equivalent to the order γ α β. At first glance this may not seem to be an issue, but when reporting results in the literature the order used in the research is often not included. While working on one's own programs it is trivial to be consistent, when collaborating with other researchers, trying to reproduce results in the literature, or using commercial software the order may not be obvious. Another problem is that the axes of rotations are not fixed. It is possible to represent any 3D rotation by using any successively orthogonal axis, for instance rotation about the *y*-axis, then *x*, and then *y* again can represent any rotation. Another problem is that where the head is facing is also not known unless reported, but this ambiguity is common among rotation

representation methods. To address this a common convention in the literature is to use 'yaw', 'roll', and 'pitch', this still does not satisfy the order ambiguity.

The use of a rotation matrix does solve many of the issues of Euler angles. There is no order dependence because the rotation matrix rotates the object simultaneously about all axes. It is also possible to covert a rotation matrix to a given order of Euler angles, and if the object is not at a singularity it is also possible to covert from Euler angles to a rotation matrix. Though each of the possible orders has their own conversion.

However, it is difficult to interpret the meaning of individual elements of the rotation matrix. The rotation matrix consists of nine elements for three dimensional rotation. The redundancy makes it difficult to visualise the rotation represented by the matrix without actually using a computer. In other words there is no natural meaning to any one of the elements of the matrix. If we were to try and synthesise a rotation matrix directly it must satisfy these constraints:

- Orthogonal
- Determinant of 1
- Real entries

It should also be noted that the addition of multiple rotations is done through matrix multiplication, if the object is rotated by matrix R_1 then R_2 the total rotation R_T is given by

$$R_T = R_2 R_1. \tag{2.13}$$

Obviously this is not problematic, but it would mean that calculating the differentials of the rotation for the angular velocity and acceleration would be more difficult.

The key difference between a rotation matrix and Euler angles is the amount of parameters. The addition of six extra parameters is responsible for the lack of ambiguity, but there is redundancy in this information and hence the resulting co-dependence of the elements of a rotation matrix make it difficult to synthesise. On the other hand

Euler angles are independent in terms of rotation which makes synthesis easier. The independence we are referring to is mathematical independence, it is highly likely that when representing the movement of the head Euler angles would have some cross dependencies.

Quaternions have four parameters and like a rotation matrix describe rotation unambiguously. Thus we can conclude that this is probably the ideal number of parameters. There is, however, still an issue with interpretation. Euler's rotation theorem states that any rotation in 3D can be represented by an axis about which the object will rotate and an angle which is the magnitude of the rotation. The four elements of a quaternion that represents rotation are by convention called: w, x, y, and z.

This may lead one to believe that x, y, and z are the axis of rotation, and w is the angle. However, this is not the case. Quaternions have been formulated in such a way that applying multiple rotations is simple, but this means that the elements still have no easy interpretation.

The axis-angle representation of rotation is a far more direct interpretation of Euler's rotation theorem. As the name implies the rotation is specified by a 3D vector which is the axis of rotation and an angle which is the magnitude of the the rotation. An extension of this representation is a rotation vector. A rotation vector is a conversion of the 4D axis-angle down to a 3D rotation vector. This is achieved by normalising the magnitude of the vector representation axis of rotation, and then setting the magnitude of the rotation vector to be the angle of rotation. This replaces the forth parameter that we need to uniquely describe rotation with prior knowledge. Mathematically if α is the angle and $\mathbf{v}$ is the axis of rotation then the rotation vector $\mathbf{r}$ is given by

$$\mathbf{r} = \alpha \frac{\mathbf{v}}{||\mathbf{v}||} \tag{2.14}$$

The rotation vector representation has many advantages. First, it is unambiguous with regards to order, as one can think of it as simultaneous rotation about all three axes.

Second, the components are measured in radians so they are easy to interpret. Third, there is no singularities when there is a rotation, and no rotation is given by the zero vector which is intuitive. The major disadvantage of rotation vectors is that they are not easy to add together. Normally addition of rotations is done by converting to either quaternions or rotation matrices first (Diebel, 2006). What could also be a theoretical problem with this formulation is that there is a discontinuity at π radians, however, as the human head cannot turn this far this problem can be disregarded for this application.

For the purposes of this research we have chosen to use rotation vectors, for which the components will be donated as r_x, r_y and r_z.

The issue of addition is not problematic, because we only specify absolute angles of rotation. And when applying the rotation we will convert to a rotation matrix with

$$\alpha = ||\mathbf{r}|| \qquad s = \sin\left(\frac{\alpha}{2}\right) \qquad c = \cos\left(\frac{\alpha}{2}\right)$$

$$R = [\mathbf{v}_1(\mathbf{r}) \quad \mathbf{v}_2(\mathbf{r}) \quad \mathbf{v}_3(\mathbf{r})] \tag{2.15}$$

$$\mathbf{v}_1(\mathbf{r}) = \frac{1}{\alpha^2}\begin{bmatrix} \left(r_x^2 - r_y^2 - r_z^2\right)s^2 + \alpha^2 c^2 \\ 2s\left(r_x r_y s - \alpha r_z c\right) \\ 2s\left(r_x r_z s + \alpha r_y c\right) \end{bmatrix} \tag{2.16}$$

$$\mathbf{v}_2(\mathbf{r}) = \frac{1}{\alpha^2}\begin{bmatrix} 2s(r_x r_y s + \alpha r_z c) \\ \left(r_y^2 - r_z^2 - r_x^2\right)s^2 + \alpha c^2 \\ 2s(r_y r_z s + \alpha r_x c) \end{bmatrix} \tag{2.17}$$

$$\mathbf{v}_3(\mathbf{r}) = \frac{1}{\alpha^2}\begin{bmatrix} 2s\left(r_x r_z s - \alpha r_y c\right) \\ 2s\left(r_y r_z s + \alpha r_x c\right) \\ (r_z^2 - r_x^2 - r_y^2)s^2 + \alpha^2 c^2 \end{bmatrix} \tag{2.18}$$

Where R is the rotation matrix.

While calculating the true angular velocity would be difficult for our purposes the difference between the two vectors is a sufficient approximation. Because, as will be

explained in Chapter 5, for the proposed method of head motion synthesis the only aspect of the velocity that is important is when the head changes direction. Thus the rotation vector is the representation used in the majority of this research. And unless it is specified otherwise the reader can assume we used rotation vectors throughout this thesis.

2.2.4 Speech Features

In Automatic Speech Recognition (ASR) and statistical Text To Speech (TTS) systems there are several methods that are widely used for encoding speech. This is to make it easier to build statistical models compared to modelling the original waveform. Some examples of encoding systems are Mel-Frequency Cepstral Coefficients (MFCC), the related Mel-Cepstral Coefficients (MCEP), and their generalised form Mel-Generalized Cepstral Coefficients (MGCEP). These are all based on representing the cepstrum: The Fourier Transform of energy of the signal. Normally for speech the cepstrum is determined over a window in time that moves at a fixed rate through the signal. For example in TTS applications a common choice of parameters is 5 to 10 ms frame shift with a 25 ms window over which the analysis is performed to extract the speech features.

A simple explanation of how waveforms are converted into the various coefficients is that a bank of lifters (cepstral domain filters) are applied to the cepstrum of the waveform and the energy of the signal in each of the lifters forms the coefficients. The various types of Mel-coefficients are different ways of constructing the lifters and are designed to approximate the parts of the signal that humans can hear the clearest. It is thought that these bands are the most important for comprehension of speech, both for understanding what was said by a human and for making synthesised speech

The author contributed to research in the use of EMA features for head motion synthesis by preparing the data and helping design and carry out the evaluations presented in Ben Youssef et al. (2013a,b, 2014)

understood.

A related set of features to the MFCC, MCEP, and MGCEPs are Linear Predictive Coefficients (LPC) and Linear Spectral Pairs (LSP). With the correct parameters the MGCEPs are equal to LPCs and there are algorithmic conversions between LSPs and LPCs. LPCs and LSPs are another way of dividing up the cepstrum but focus on the ease of signal processing and interpolation, and not human hearing characteristics.

Another category of speech features are often called prosodic features. It is important here to differentiate between perceptual features and measurable features. As the name implies perceptual features are what a listener would be able to perceive. Examples of perceptual features are pitch and loudness. On the other hand measurable features can either be approximated or directly measured from a speech signal. While pitch is perceptual the fundamental frequency of the glottal folds (sometimes called vocal chords) can be approximated, this is called F0, and while loudness is perceptual we can measure the energy of the signal over the window. F0 and energy can function as stand-ins for pitch and loudness.

A large problem with F0 is that it is not continuous in time. The glottal folds do not always vibrate during speech, for instance when pronouncing the English letter 's' there is no movement in the glottis, instead the sound is created by the shape of the lips, tongue position, and air from the lungs. The regions of the signal where F0 exists are called the voiced region and where it does not exist is called the unvoiced region. There are two common methods for dealing with this problem. Either when building the models the F0 is handled in a different manner to the other features which takes into account the fact that it is not always measurable. Alternatively F0 can be interpolated in the unvoiced regions, though this would be an approximation.

There are many software tools available for extracting all of the above speech features from speech signals. Some common ones are SPTK[2], openSmile (Eyben et al., 2010),

[2] `http://sp-tk.sourceforge.net/`

and STRAIGHT (Kawahara et al., 1999). Some of these tools also include methods for estimating the perceptual features from the speech features. For instance openSmile outputs a pitch feature.

Another method for representing speech is based around how speech is produced in humans. In humans speech is produced by the movement of air over the lips, tongue, teeth, and glottis (also known as the voice box). These are collectively known as the articulators. Studies in speech production in humans have measured how the articulators move during speech, initially using x-rays, and later (once the dangers of x-rays became known) a device known as the Electro-Magnetic Articulograph (EMA). An EMA machine has the disadvantage of not being able to measure the glottal movements directly, but compared to x-rays has the advantage of not giving study participants cancer. An EMA works by attaching magnetic coils to the articulators of the participant other than the glottis, and determining the coils' movement by measuring changes in the magnetic field. Often only two dimensions are considered. This is because in most languages, during non-impaired speech, the articulator movement is symmetrical about the left-right axis when facing the speaker.

Through a process known as speech inversion EMA measurements, i.e. the movement of articulators, can be estimated using speech features. Recent research has examined the use of EMA features estimated from speech for head motion synthesis (Ben Youssef et al., 2013a,b, 2014). It was found that predicted EMA features are more highly correlated with head motion than the standard array of speech features used in ASR and TTS.

2.3 Existing Head Motion Synthesis Methods

In this section we will present a discussion on a selection of existing head motion synthesis systems. Part of this will entail a review of prior work analysing the relationship

between speech features and head motion trajectories. Additionally we will look at how head motion has been encoded in the existing literature. As the method proposed in Chapter 5 is an HMM based system, the literature on HMM head motion synthesis will be presented separately.

An early attempt at synthesising head motion from speech features was carried out by Yehia et al. (2002). They used a linear estimator to predict head motion trajectories from F0. What they found that a linear estimator trained on the entire dataset found no correlation between speech and head motion. However, a linear predictor worked well when only using small samples of their data. Their conclusion was that the dependency changes from utterance to utterance. We will discuss this problem and our attempts to mitigate it in Section 4.3.

Even if it is the case that the dependencies are only found locally, it still means that a simple linear estimator is not suitable for head motion synthesis. This is because even if a linear estimator is trained for every sample, it would not be possible to determine which estimator is the appropriate choice without some additional modelling.

On the other hand, the work by Yehia et al. (2002) shows that it should be possible to predict head motion from speech features, if a more sophisticated modelling technique is utilised. One problem with their research is that they did not provide any details of subjective testing which means that it would be difficult to compare results to this system.

More recently Le et al. (2012) proposed another system that relies on a simple predictor. In their case their system predicts Euler angles, and each angle's trajectory is treated separately. To do this they take into account speech features, the current value of the Euler angle trajectory, and the total angular velocity and acceleration of all three angles. They then use a probabilistic modelling technique, specifically GMMs, to estimate head motion. As they found their system performs well on subjective tests compared to other synthesis systems previously proposed by Busso et al. (2005), Chuang

and Bregler (2005), and Levine et al. (2009). As this includes some of the HMM based systems described below, it is an excellent choice for comparing against state of the art systems for both GMM and HMM based systems. As such we will be using it as a baseline state of the art system and so will examine their approach in more detail.

Le et al. (2012) used 40 minutes of data from one speaker, though they did not mention how much was used for training and how much was held out for testing. Head motion was recorded using a motion capture system. The head motion trajectory was parametrised for time, t, into three Euler angles: yaw (α_t), pitch (β_t), and roll (γ_t). Though as is so often the case they did not report the order of operation. Additionally for training they calculated angular velocity, v_t and angular acceleration, a_t, using respectively

$$v_t = \left\| \begin{bmatrix} \alpha_t \\ \beta_t \\ \gamma_t \end{bmatrix} - \begin{bmatrix} \alpha_{t-1} \\ \beta_{t-1} \\ \gamma_{t-1} \end{bmatrix} \right\|, \tag{2.19}$$

and

$$a_t = \left\| \begin{bmatrix} \alpha_t \\ \beta_t \\ \gamma_t \end{bmatrix} - 2 \begin{bmatrix} \alpha_{t-1} \\ \beta_{t-1} \\ \gamma_{t-1} \end{bmatrix} + \begin{bmatrix} \alpha_{t-2} \\ \beta_{t-2} \\ \gamma_{t-2} \end{bmatrix} \right\|. \tag{2.20}$$

For speech features they used the OpenSmile (Eyben et al., 2010) approximations of pitch, p_t and loudness, l_t, for time t.

For synthesis they would find the estimated trajectory of the $(\alpha_t^*, \beta_t^*, \gamma_t^*)$ tuples by optimising

$$\begin{aligned}(\alpha_t^*, \beta_t^*, \gamma_t^*) = \arg\max_{\alpha,\beta,\gamma} & \prod_{\kappa \in \{\alpha,\beta,\gamma\}} P(\kappa_t, p_t, l_t) \\ & \times P\left(\left\| \begin{bmatrix} \alpha_t \\ \beta_t \\ \gamma_t \end{bmatrix} - \begin{bmatrix} \alpha_{t-1}^* \\ \beta_{t-1}^* \\ \gamma_{t-1}^* \end{bmatrix} \right\|, p_t, l_t \right) \\ & \times P\left(\left\| \begin{bmatrix} \alpha_t \\ \beta_t \\ \gamma_t \end{bmatrix} - 2\begin{bmatrix} \alpha_{t-1}^* \\ \beta_{t-1}^* \\ \gamma_{t-1}^* \end{bmatrix} + \begin{bmatrix} \alpha_{t-2}^* \\ \beta_{t-2}^* \\ \gamma_{t-2}^* \end{bmatrix} \right\|, p_t, l_t \right),\end{aligned} \tag{2.21}$$

Under the assumption that the system is causal, ie there is no backwards dependence in time. To solve the optimisation they used the gradient decent algorithm. They estimated P by using five 10 component GMM, where each GMM found the joint probability of:

- The angles and the speech features, each angle trained separately
- The angular velocity
- The angular acceleration

The GMMs were all trained using expectation maximisation.

For a subjective evaluation they used an A/B test (see Section 4.2.1 for details on different types of subjective evaluation). They compared their system to the ones stated above (Busso et al., 2005; Chuang and Bregler, 2005; Levine et al., 2009) and to motion capture. What they found was their system outperformed the other systems the majority of the time. It also was sometimes preferred to motion capture, but in this case it was not the majority of the time.

Brkic et al. (2008) proposed a similar encoding to the one proposed in this thesis, as

such we will mention how they differ to the template based encoding presented in Chapter 5 here, despite not having described how it works yet. It should also be noted that in their paper they did not attempt to perform synthesis only encoding. In their system they define two types of gestures:

- **Nods**: "an abrupt swing of the head with a similarly abrupt motion back"
- **Swings**: "an abrupt swing of the head without the back motion"

Using those two categories, they annotated their dataset with gesture type, direction, amplitude, and duration. Then using this annotation they encoded the trajectories using simple mathematical functions that described the same motions as what was annotated. They then conducted a subjective test of the encoded trajectories and found that subjects rated the encoded motion very highly.

Our system does not work on the head motion trajectory as a whole, instead we work with each angle individually. Additionally our system requires no manual annotation, and relies entirely on automatically segmented data. The main similarity is that we encoded the motion using simple mathematical functions.

While Choi et al. (2001) describes a synthesis system based on HMMs it will be discussed here as there is a substantial difference between it and the HMM systems discussed in the next section. This paper illustrates a completely different method of synthesis that occasionally is used in the literature. The Choi et al. (2001) method does not estimate the trajectory of the head directly. Rather head motion is generated indirectly. The author's method morphs a 3D facial model and this necessitates motion in the rest of the 3D model of the head. The HMM in their paper is used to optimise the synthesised trajectory of the morphing parameters given the speech features.

Our analysis into the use of EMA features follows a similar logic, where the head motion is guided by the articulators which were estimated from speech. However, by estimating head motion trajectories directly instead of relying on morphing the rest of

a 3D head model we allow for a wider variety of head motion such as the prosodic visual gestures discussed in the introduction.

2.3.1 HMM Based Methods

The systems presented in this section are all HMM based. They are separately discussed because the system presented in this thesis is also HMM based. However, it is not needed to directly compare to HMM systems in subjective evaluation because they have already been compared to the system proposed by Le et al. (2012), and the Le et al. (2012) system was already shown to outperform all of the systems it was compared to in subjective testing. As the system we present has outperform Le et al. (2012) (see Chapter 5) there was no need to go back to comparing to these earlier systems.

One of the first HMM based systems was proposed by Busso et al. (2005) and updated in Busso et al. (2007). In the more recent approach they train HMMs on clusters of head motion built using Linde - Buzo - Gray vector quantization (Linde et al., 1980). These are meant to represent typical head motion poses. They then pick the most likely head motion sequence based on the acoustic features. This gives the target poses which they then interpolate between they then add noise to create an interesting trajectory. Both their subjective and objective results have proved promising. Furthermore using different training data they were able to simulate different emotions. However, apart from using an HMM for modelling probabilities our proposed system takes a different approach.

A similar approach was developed by Sargin et al. (2006) and expanded in Sargin et al. (2007, 2008). Like Busso et al. (2005, 2007) they used automatic segmentation. This segmentation gave them modelling units. They then used a multi-stream HMM to recognise the modelling units and simultaneously generate a trajectory. Because of the different modelling units they employed a parallel HMM structure with the HMMs in each branch trained on different units.

Hofer and Shimodaira (2007); Hofer (2009) used a different approach to the segmentation. Unlike the other two HMM based systems by Busso et al. (2005, 2007) and Sargin et al. (2007, 2008) used manual segmentation. The segmentation was done using linguistically meaningful units, such as nodding or shaking. Their dataset was then manually annotated with these modelling units. At this point they trained a bank of HMMs to recognise these units in a manner analogous to whole word ASR. The HMMs that were trained for recognition also included a stream for the head motion trajectories. When performing the recognition these streams were adjusted so that they did not affect the observation probabilities. Once the model sequence was determined the HMMs were then used for synthesising the head motion. This was achieved by using a technique of adapting HMMs to predict fixed length segments called Trajectory HMMs (Sako et al., 2000).

The other system that was compared to Le et al. (2012) was proposed by Levine et al. (2009). In this system Levine et al. (2009) still use HMMs to recognise the type of motion to use, however, instead of using motion generated by an HMM they use units selected from the original database. Like Busso et al. (2005, 2007) and Sargin et al. (2007, 2008) they used automatically segmented data. However, their system was actually designed for whole body gesture generation, and for dialogue systems, not monologues. Due to all these differences, the Levine et al. (2009) system is not a suitable comparison for the system presented here.

2.4 Section Summery

In this section we have presented the key background needed for understanding this thesis. In particular we discussed the machine learning techniques that are going to be used in the presented system, namely Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM). We also covered different methods of representing rota-

tion and gave reasons for our chose in using rotation vectors. Finally presented some background information on different types of speech features, including EMA features which are estimated from lower level speech features.

Additionally we have discussed some other approaches to head motion synthesis. We focused on similar methods to the one proposed in Chapter 5, particularly the method by Le et al. (2012) which we will be using for comparisons as it outperformed all of the existing systems they compared themselves to.

Chapter 3

Data Collection and Statistics of Head Motion

3.1 Dataset Considerations

With any machine learning based approach training data is needed. Mostly this is taken from datasets, though some algorithms will learn from data generated from its own 'experience'. Head motion synthesis systems, as they stand, use machine learning techniques that are dataset based to build their mapping from speech to head motion. The model presented in this thesis are no different. This section describes the data that was used for training and evaluation. First we will discuss the considerations in recording head motion and the available datasets and then the dataset that was collected.

To begin with there are several ways with which head motion and speech can be simultaneously recorded. For instance, Hadar et al. (1983, 1985) used a polarised-light

This work was submitted to Intelligent Virtual Agents 2013. It was edited to present as a poster (Braude et al., 2013b)

goniometer to record the head movements. With improvements in computer vision technologies it is now possible to use videos, though this is not common in the head motion synthesis literature. The most prevalent type of data is motion capture, which is what was used for in this research. Another less common device that could be used is an Electro-Magnetic Articulograph (EMA). Some examples of EMA datasets are the ones collected by Turk et al. (2010) and Richmond et al. (2011).

While EMA data is highly accurate, it does have some potential problems. The subject has wires glued to their tongue and face for the sensors to detect. Additionally the subjects are restricted in their range of head motion by an enclosure which houses the magnetic sensors. This could lead to movement that is less natural. Motion capture is another approach, but because there is no physical anchors such as the wires or the enclosure from the EMA the movement, the motion captured data should be more natural. On the other hand the subjects of the recording usually have some motion capture markers, which can also reduce the naturalness though probably not as much as an EMA. Video is the recording method that should give the most natural motion as there nothing other the fact they are being recorded influencing the subjects. But computer vision does not yet have the same level of accuracy as the other two methods. For this research the motion capture was thought to be the best trade-off between accuracy and naturalness.

Another consideration is the scenario in which the data is recorded. Some researchers such as Busso et al. (2007) have used short utterances, others have used longer story-telling monologues such as Sargin et al. (2008). Some researchers have also attempted to use dialogues for example Le et al. (2012). For the bulk of the research story telling monologues were the chosen scenario. Monologues are easier to capture because there is only one person that needs to be recorded. And secondly there are no additional influences from other speakers on the motion of the subject of interest.

One must also decide on how the speaker will know what to say. One can either give

the subjects something to read, or give them the ability to say what they like, perhaps with some prompting, which we will call 'free' speech as opposed to 'read' speech. Both approaches have been used previously in the literature, for example Busso et al. (2005) used read speech and Hofer and Shimodaira (2007) used free speech. While free speech is probably more natural we were able to capture both so there was no need to compromise on this point. This allows us to compare the different types of head motion.

When collecting a dataset one must take into account the number of speakers and how much data we have for each of them. Typically in available datasets there are less than four speakers with very little training data each. We will seek to gather data from a much larger number of speakers with over 15 minutes of data each.

3.2 Existing Datasets

The noted exception to the short dataset generalisation is the IEMOCAP dataset (Busso et al., 2008) for which there are ten actors. However, it was unsuitable for our needs as it contained only dialogues which as previously mentioned would mean that there are two parties to take into account when trying to estimate head motion.

Some other datasets that are publicly available and might be considered but are unsuitable are SEMAINE (McKeown et al., 2012) and CID (Ferré et al., 2007) which do not have motion capture, while IDIAP Tosato et al. (2012), CMU (Carnegie Mellon University, 2013), and HID (Rett et al., 2007) have motion capture but not audio. Due to the missing data none of these datasets could be used. Table 3.1 summarises the differences between the different datasets and shows why a new one was necessary.

While it was not publicly released as a dataset, the data used by Hofer (2009) was also available to us. This was a single speaker telling personal anecdotes and jokes without a script. The speaker was a semi-professional actor. For this research we

Table 3.1: Comparison of data available in existing candidate datasets

Name	Body	Clean audio	Transcription timing
IDIAP	Yes	No	None
CMU	Yes	No	None
HID	Yes	No	None
mngu0	No	Yes	Phone (automatic)
ESPF	No	Yes	Sentence (manual)
IEMOCAP	Yes	Yes	None

needed multiple speakers. Additionally each speaker should have different methods for eliciting the data such as read and free speech, which we will call tasks. This would mean our results are comparable to other researchers who use read speech only, while still providing data that one would expect to be more natural, namely free speech.

The data that was recorded was released under the name of the University of Edinburgh Head motion and Audio Storytelling (UoE – HAS) dataset (Braude et al., 2013b).

3.3 Recording Scenario and Participants

For our recordings the subjects recruited were all native speakers of English who had been raised in the United Kingdom. Most of the subjects were undergraduate university students, with the exception of one who reported her age as greater than 50, who we refer to as Irene. None of the subjects were professional actors.

We recorded 16 speakers each telling classical fairy-tales which the subjects were familiar with either from childhood or from watching the Disney films. Copies of these stories from Project Gutenberg[1] were provided a week in advance for the subjects to read to refresh their memories. The stories chosen were:

[1] https://www.gutenberg.org/

- Red Riding Hood
- Rumplestiltskin
- Repunzel
- The Emperor's New Clothes
- Sleeping Beauty

During the recording session the subjects would first read the story off a Teleprompter, and then retell the same story in their own words. They were allowed to either reuse the story as they had read it, or tell another version of the story that they were more familiar with. This provided both read and free speech. Preliminary testing had showed that by priming the speakers with a story we were able elicit stories that were at least five minutes long. When the subjects were left to choose their own story the speakers normally told stories that were approximately two minutes long.

During the recording session the speakers were seated and facing the recorder so they would be able to focus on a person when telling the free speech. The recorder was a native English speaker. During the read speech they focused on the Teleprompter. They were instructed to tell the story as if they were speaking to an adult native English speaker who did not know the story.

3.4 Physical Layout, Hardware, and Software

It is important that both the head and body motion are tracked. This is so that it is possible to remove the movement of the body from the head motion trajectory. As previously mentioned to record the movement a motion capture system was used. The system that was available to us was the Optitrack system from Natural Point[2]. Our system consisted of seven V100:R2 cameras arranged on four tripods facing the speaker.

[2] http://www.naturalpoint.com/optitrack/

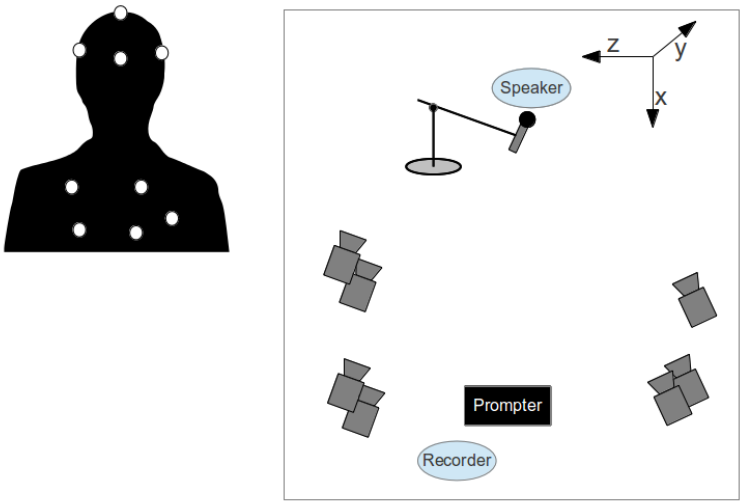


Figure 3.1: Placement of motion markers on the participant and layout of recording studio

The speaker wore a peak-cap with four markers and a jacket with five markers. In addition to head motion, body movements also needed to be captured so that it would be possible to remove it from the head movements. The position of the markers on the participant and the layout of the recording studio is shown in Figure 3.1. Also included in Figure 3.1 is the axes we used through the research. Note that this is a right handed coordinate system.

The motion capture system software that we used was initially Natural Point Tracking Tools, however, Tracking Tools was discontinued and replaced by Motive in 2012. The final dataset only includes data that was recorded with Motive. The system was set to record at 100 Hz, and give a synchronisation signal at the start and end of the recording. The software exports the tracking data into a .CSV file which contains the marker coordinates in 3D space. Additionally one can predefine 6DOF objects and Motive will estimate their position and rotation in real-time and include this information in the .CSV file.

For our dataset we tried to define the head and body as rigid objects. However, the original rigid body tracking was fairly poor due to the limitations of real-time tracking.

Thus the motion of the head and body was re-estimated using Singular Value Decomposition (Söderkvist and Wedin, 1993). The mean position of all the participants was reset to the zero vector to account for differences in height and sitting position of the participants.

Audio was captured using a free-standing directional microphone and a MOTU-8pre mixer[3]. The synchronisation signal was recorded on a second channel in the mixer. The audio was captured at 44100 Hz with 32-bit depth. Audacity[4] was used to trim the audio files down to exactly match the start and length of the motion capture samples, as part of this process Audacity down-samples the bit depth to 16 bits. For the extraction of the acoustic features the sampling frequency was reduced to 16 kHz using SOX[5].

3.5 Dataset Statistics

The rest of this chapter will focus on some statistics about the dataset, including an analysis of the speaker and task dependency.

The lengths of the recordings are given in Table 3.2. It should be noted that the only speaker who was not a university student is Irene who is also the subject that reported her age as greater than 50. Also to note that these are all pseudonyms and gender is self reported. Due to issues with the tracking the data from Simon was discarded, though we have still made it available with this fact noted.

To estimate the average speaking rate of each of the subjects we employed the script made available by de Jong and Wempe (2009). Rather than using an Automatic Speech Recognition (ASR) system, they use signal processing techniques to estimate the number of syllables in each utterance. This is both far faster and has the advantage of not needing to train recognition models. Additionally it is speaker independent.

[3] http://www.motu.com/
[4] http://audacity.sourceforge.net/
[5] http://sox.sourceforge.net/

Table 3.2: Lengths of recordings (min:sec)

Gender	Name	Read	Free	Total
Female	Ally	26:37	16:33	43:10
	Carla	10:58	14:49	25:47
	Irene	26:48	28:23	55:11
	Jane	26:12	23:50	50:02
	Nadine	26:45	26:07	52:52
	Natalie	25:25	17:48	43:13
	Nicole	21:05	21:00	42:05
	Rebecca	16:27	17:24	33:51
	Robin	22:13	21:36	43:49
Male	Desmond	26:41	24:43	51:24
	Gary	25:35	17:46	43:21
	Mark	23:32	17:35	41:07
	Marvin	26:00	26:33	52:33
	Ray	25:21	14:12	39:33
	Sam	14:51	14:58	29:49
	Simon	27:05	20:05	47:10
Total		371:14	323:22	694:36

The mean estimated speaking rates of each of the participants is provided in Table 3.3. Below we will see if differences in speaking rate correlates with differences in head motion, which will be examined first.

3.5.1 Speaker and Task Dependency

In order to examine the differences in head motion between speakers and speaking tasks, a distance measure is needed. In this case the cross entropy distance is appropriate as it gives an indication as to the likelihood of two trajectories being generated by the same source and has been shown to be more reliable than KL-divergence (Helén and Virtanen, 2010). Formally cross entropy distance is defined a s

$$E(\mathbf{A},\mathbf{B}) = \frac{1}{T_A}\log\left(\prod_{t=1}^{T_A}\frac{p_A(a_t)}{p_B(a_t)}\right) + \frac{1}{T_B}\log\left(\prod_{t=1}^{T_B}\frac{p_B(b_t)}{p_A(b_t)}\right), \tag{3.1}$$

where $\mathbf{A} = (a_1, \ldots, a_{T_A})$ and $\mathbf{B} = (b_1, \ldots, b_{T_B})$ are random vectors of data from two observations, T_A and T_B are the lengths of those observations, and $p_A(\cdot)$ and $p_B(\cdot)$ are the Probability Density Functions (PDF) estimated from $\mathbf{A}$ and $\mathbf{B}$ respectively.

While cross entropy distance is usually only used for single-variate data, it is trivial to extend it to the multivariate case. The only change is that $p_A(\cdot)$ and $p_B(\cdot)$ are multivariate probability density functions. By making an assumption that the variables are independent it will be possible to check that the goodness of fit for each variable separately.

In this dataset we are looking at the difference between the trajectory of each utterance. In this instance $\mathbf{A}$ is the trajectory for one sample, and $\mathbf{a}_t$ are the components of the rotation vector. However, the static position does not give any indication as to the way the head moves, so $\mathbf{a}_t$ also contains the first and second derivatives. $\mathbf{B}$ and $\mathbf{b}_t$ are similarly defined for another sample. If $\mathbf{A}$ and $\mathbf{B}$ are the same sample then the distance will be zero.

To determine the appropriate PDF it is useful to examine the distribution of the data.

Table 3.3: Mean speaking rate (syllables / sec) for each speaker for different tasks

	Free		Read		Overall	
Speaker	Mean	Std	Mean	Std	Mean	Std
Ally	2.73	0.10	3.47	0.13	3.10	0.39
Carla	3.41	0.13	3.45	0.10	3.43	0.12
Irene	3.97	0.12	3.94	0.14	3.96	0.13
Jane	3.50	0.11	4.17	0.19	3.83	0.37
Nadine	3.76	0.17	4.21	0.09	3.99	0.26
Natalie	3.64	0.13	3.86	0.20	3.75	0.20
Nicole	2.74	0.12	3.35	0.03	3.04	0.32
Rebecca	3.74	0.09	4.17	0.03	3.95	0.22
Robin	3.71	0.14	4.18	0.14	3.94	0.28
Desmond	3.49	0.14	3.98	0.10	3.73	0.28
Gary	2.84	0.10	3.37	0.10	3.10	0.29
Mark	3.21	0.12	3.45	0.14	3.33	0.18
Marvin	3.35	0.13	3.64	0.18	3.49	0.21
Ray	2.75	0.15	3.73	0.14	3.29	0.51
Sam	2.99	0.12	3.83	0.07	3.41	0.43

Table 3.4: Skew, Kurtosis, and Kolmogorov-Smirnov statistic, assuming normal distrubtion of all the rotation vectors components, measured in radians

	r_x	r_y	r_z	Δr_x	Δr_y	Δr_z	$\Delta^2 r_x$	$\Delta^2 r_y$	$\Delta^2 r_z$
Skew	-0.08	-0.16	0.55	0.03	-0.03	-0.10	-0.01	0.01	-0.15
Kurtosis	4.78	5.12	6.21	22.2	18.8	10.6	303	59.4	108
KS - statistic	0.42	0.41	0.42	0.50	0.49	0.50	0.50	0.50	0.50

This is shown in Figure 3.2. Based on these diagrams the data may be normally distributed, however, to determine if this is true we examined the Skew, Kurtosis, and Kolmogorov-Smirnov statistics of the rotation vector components. The results of this analysis are presented in Table 3.4. While the Skew is quite low, the Kurtosis is very high, especially for the second derivatives. Furthermore, the Kolmogorov-Smirnov statistic also indicates that the data is probably not normal. As such a normal distribution is not appropriate.

Because the data is not normally distributed we instead chose to use histograms to calculate the likelihoods. The bin edges were evenly spaced on the range calculated from all available samples so that the bin edges would be the same. 50 bins were used for this experiment. One sample was added to all the bins to ensure that there would be no zero probabilities.

To visually inspect the data we employed a technique known as MultiDimensional Scaling (MDS). This is a process by which we bring the distances between samples to a lower number of dimensions while attempting to preserve the distances between the points. The amount by which the distances do change is known as the stress. There are several methods for calculating stress but the most common, and the one used by default in libraries such as MATLAB[6] and Scikit - Learn in Python[7], is the Kruskal's normalized stress 1 criterion method (Kruskal, 1964) so this was used for this research.

[6] `http://uk.mathworks.com/products/matlab/`

[7] (Pedregosa et al., 2011) `http://scikit-learn.org/`

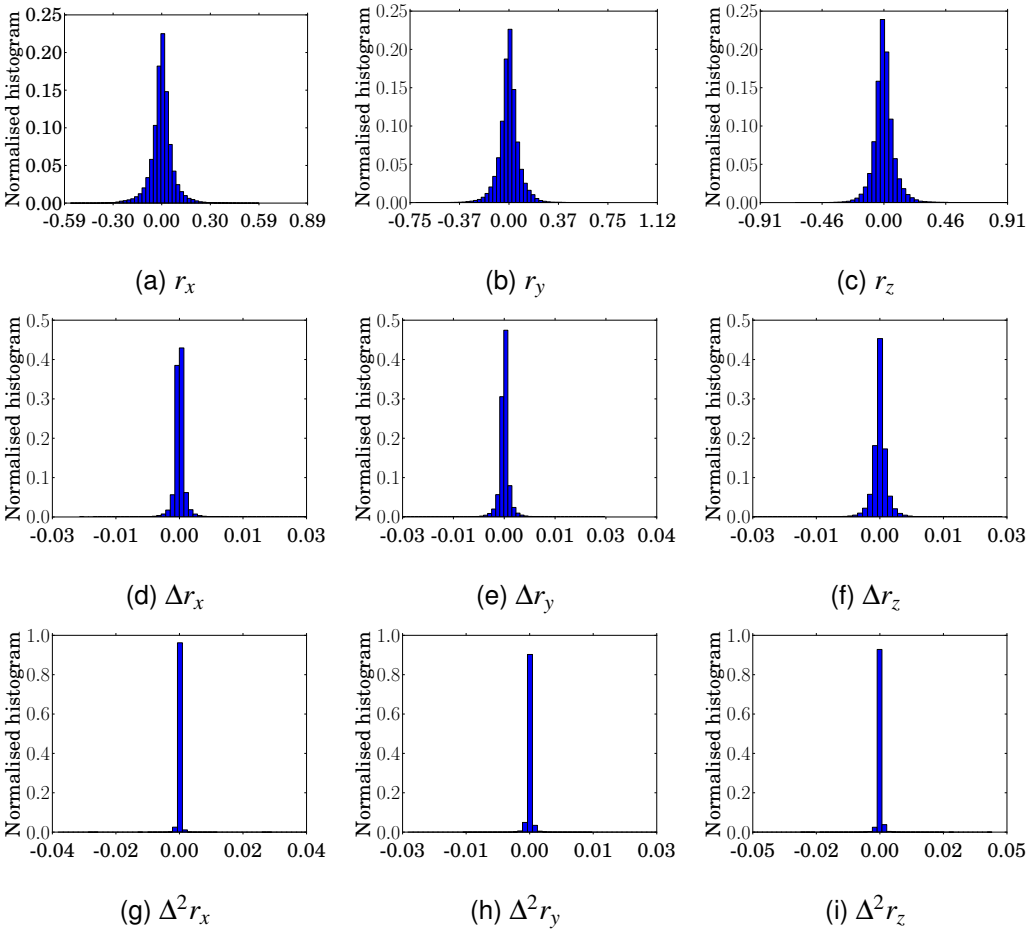


Figure 3.2: Distribution of rotation vector components and the first and second derivatives in the dataset, using 50 bins. Rotation vector components are measured in radians

Kruskal's normalized stress 1 criterion is defined as

$$S = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left(d_{ij}-\hat{d}_{ij}\right)^2}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(d_{ij})^2}, \tag{3.2}$$

where d_{ij} is the original distance between samples i and j, $\hat{d}_{ij}$ is the scaled distance, and there are N samples. In other words it the sum of the square error normalised by the original distances. If the stress is zero it would be a perfect scaling, and there is no theoretical upper limit.

Borg and Groene (2005) provides a detailed description of MDS theory. Here we present a brief overview of the specific MDS algorithm that was used. The method is one of the family of none linear methods were the stress is optimised iteratively. First it is initialised with what is referred to as classical MDS scaling:

The distances between all n utterances are arranged into an $n \times n$ square distance matrix D. From here a centring matrix C is calculated as

$$C = I(n) - \frac{1}{n}O(n), \tag{3.3}$$

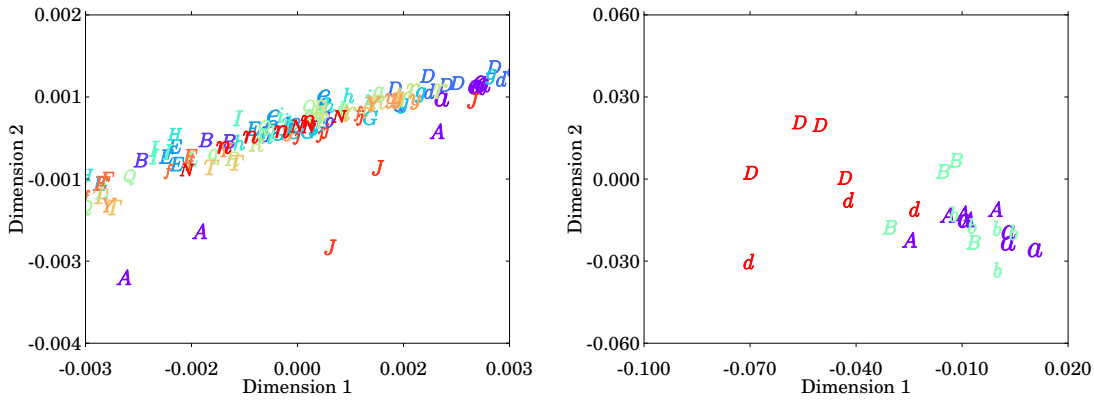
where $I(n)$ and $O(n)$ are the $n \times n$ identity matrix and one filled matrix respectively. The centred matrix D' is then given by

$$D' = \frac{1}{2}CD^2C. \tag{3.4}$$

We then calculate the eigenvectors and their respective eigenvalues of D'. Any negative eigenvalues are discarded and the remaining eigenvalues are arranged into descending order in a diagonal matrix λ, and the respective eigenvectors are placed into a matrix X. The scaled matrix S is then found by

$$S = \lambda X. \tag{3.5}$$

The first p columns are kept to reduce to p dimensions, and then the stress can be calculated. If there are less than p columns in S then extra columns can be added with zero padding.



(a) All speakers (Stress: 0.78)

(b) Three speakers (Stress: 0.35)

Figure 3.3: Relative cross entropy distances between samples reduced to two dimensions using MDS, lower case letters represent read samples and capital letters free speech. Speaker identity is given by letter and colour

At this point S is optimised iteratively with respect to the stress criterion used. A detailed description of the process is available in (Coxon, 1982, pp 60 – 84), however, in essence it is a variation of Hill Climbing Optimisation.

Figure 3.3 shows the cross entropy distances between the samples rescaled with MDS down to two dimensions. In the diagram the read speech clusters are represented with lower case letters and free speech with capital letters. Each speaker has its own letter and colour. Trying to view all the speakers together does not reveal any obvious patterns and is difficult to understand. By inspecting three speakers only we can see that there does seem to be some clusters within the data, though they are not very distinct.

The mean and variance of the distances between two samples under different conditions, namely different tasks and speakers, are given in Table 3.5. The mean distance between tasks was calculated so that the only the distances between samples of the same speaker were considered. As can be seen in Table 3.5 the difference between both conditions is greater than distances within the condition for both speaker and task.

In Appendix A, Tables A.1, A.3, and A.2 show the mean distances between each speaker, overall, for free speech, and for read speech respectively. As in this representation it is difficult to see any patterns the results are also shown in Figures 3.4, 3.6, and 3.5 as "heat maps". Additionally the mean results are given in Table 3.5. Here we see that there is a strong diagonal as would be expected if head motion is speaker dependant. We can also see that for read speech speakers are more distinct from one another than for free speech.

Table 3.5 shows that the mean distance between free speech samples is lower than the distance between tasks, while the mean distance between read speech samples is greater than the distance between tasks. This might imply a greater variety in read speech samples than free speech. Though the difference is very small so the effect is probably slight. Interestingly both free and read speech have lower average distance than the mean distance between read and free samples. This does imply there is some task dependency as one would expect.

In Table 3.5 also shows the mean distance between speakers is greater than the distance of samples from one speaker. This implies that different speakers have very different head motion. Additionally the heat maps confirm this through the strong diagonal. It is important to note that the mean distance between speakers is greater than the distance between tasks. The implication is that head motion is more speaker dependent than task dependent.

Overall the data implies that counter-intuitively it is perhaps better to use read speech as it will have more variety, and hence be more interesting. This is despite the fact that one would think that the free speech would give more natural head motion. Alternatively the method for eliciting free speech might not be appropriate for this type of participant. By this we mean that either other methods of eliciting free speech, or a different type of participant such as a professional actor would produce different results.

Table 3.5: Overall cross entropy distance statistics between speakers and free or read speech

	Mean	Variance
Read	8.43e-01	1.12e-02
Free	8.28e-01	1.12e-02
Between tasks	8.49e-01	5.31e-04
Within speaker	7.54e-01	7.17e-02
Between speakers	8.48e-01	5.55e-04

Table 3.6: Pearson correlation coefficient between difference in speaking rate and motion distance

Task	Correlation
All	0.273
Free	0.191
Read	0.058

The second implication is that speaker independent models are not appropriate as there seems to be a large dependency on the participant.

As mentioned earlier it is interesting to see if there is a correlation between the cross entropy distance of the head motion and the speaking rate of the samples. This correlation is given in Table 3.6.

Considering correlations greater than 0.1 are considered significant it is interesting to note that the correlation between difference in speaking rate and difference in head motion is only present between free speech samples, not between read samples. A possible cause is that people who speak faster during free speech have similar speaking style and head motion styles, while the differences in read speech are dominated by the speaker, and so the difference in speaking rate is of secondary importance.

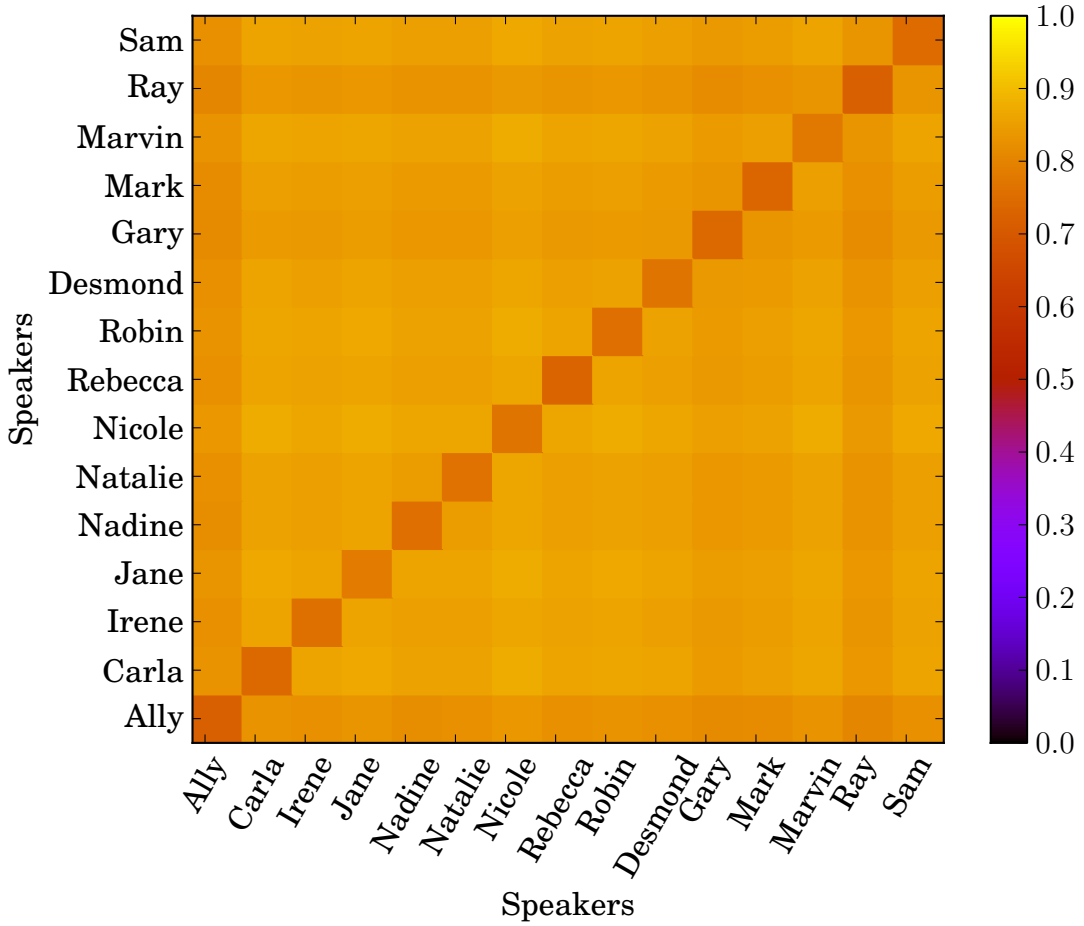


Figure 3.4: Heat map of the mean distance between all samples of one speaker to another

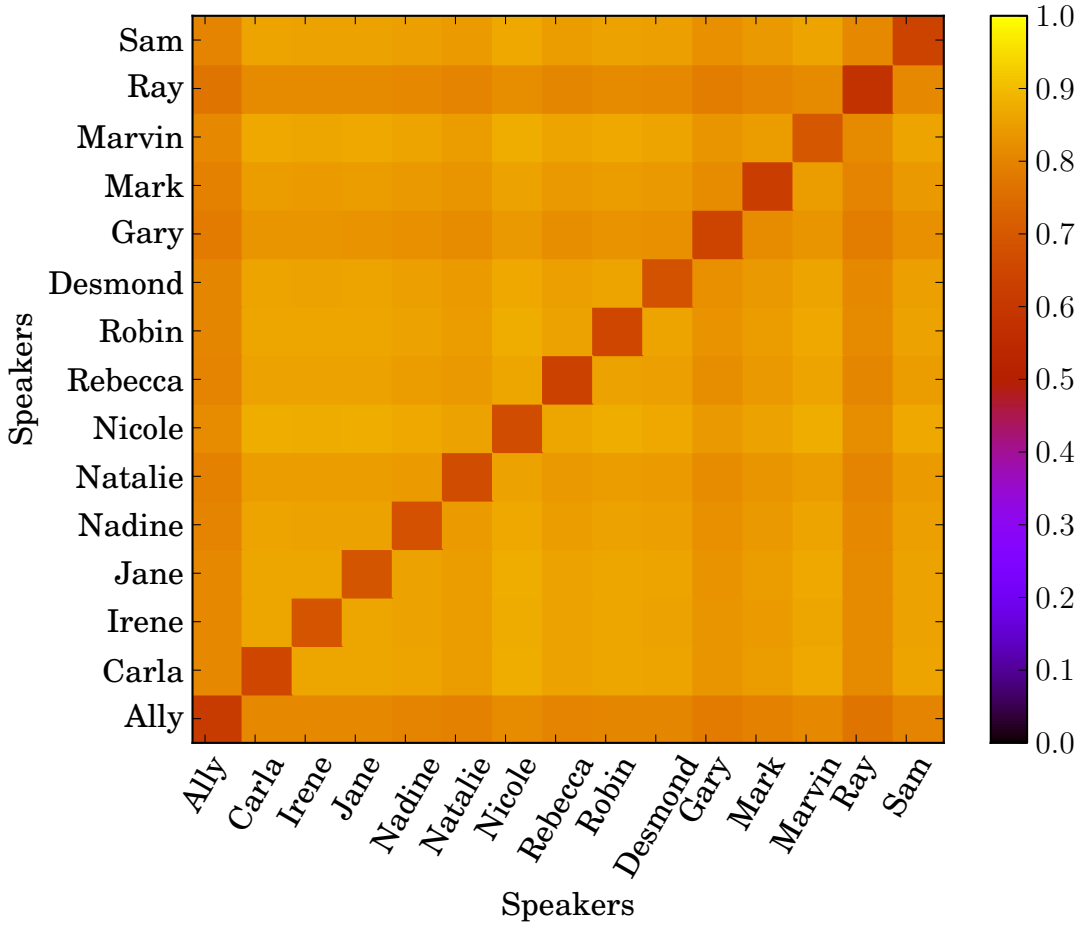


Figure 3.5: Heat map of the mean distance between samples from one speaker to another for free speech only

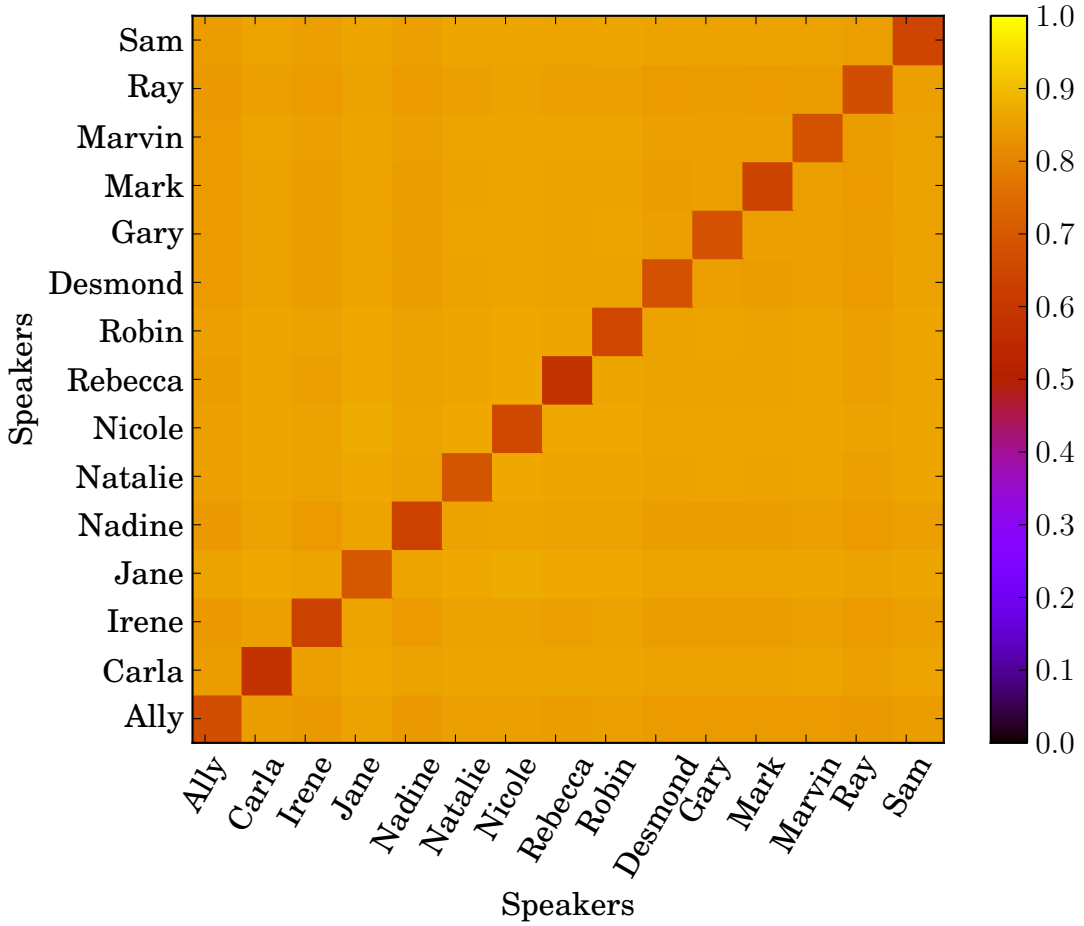


Figure 3.6: Heat map of the mean distance between samples from one speaker to another for read speech only

Chapter 4

Methodology for Head Motion Synthesis Evaluation

4.1 Subjective and Objective Evaluation Background

There are two types of head motion synthesis evaluation. Objective evaluation measures show how similar head motion is to the original. This gives an indication if the current synthesis method preforms better than either existing systems or earlier versions of new methods as they are being developed. This is because one would assume that motion capture is the gold standard.

Subjective testing on the other hand is arguably more important as this reflects the usability of the developed synthesis systems. However, subjective testing is more time consuming and expensive. Thus when building a system one would generally first use objective evaluation during prototyping and then subjective testing for the final analysis.

As far as can be ascertained there has been no formal study into the evaluation of head motion synthesis in the literature. This is not to say that reasonable methods have not

been used in the literature, but which of these methods is the most appropriate has not been tested. As the goal of the objective evaluation is to predict the results of subjective testing, first the method and typical results of a subjective test must be found. Once this is done then an objective test that gives similar these should be developed.

This chapter presents both the subjective and objective testing methodology that we developed and tested. We will first show the theoretical basis of the methods we chose for comparison and then the results of experiments that confirm the appropriateness of the methods. The end of the chapter presents an analysis and discussion of the experiments in head motion evaluation. From this discussion we are able to make recommendations for subjective evaluation and show an objective evaluation method that is more informative than those currently being used.

4.2 Subjective Evaluation

4.2.1 Types of Subjective Evaluation and Their Prior Use

In Section 2.3 we discussed existing methods for synthesising head motion. Here we will briefly present how the researchers performed their subjective evaluation on their systems. As was mentioned in Section 2.3 head motion synthesis is either done with 3D models or by joining parts of videos together. As this research utilises 3D models only methods for that used this type of animation will be discussed and examined. In terms of rendering style and model shape there are three methods that are appropriate for 3D models that have been previously used in the literature. The first, and most common, is to use 3D models that are textured to have a human-like skin, eye colour, mouth, and hair. Choi et al. (2001), Busso et al. (2005) , Sargin et al. (2008), and Le et al. (2012) are all examples of prior research that have used this method. Less common is to use 3D models that are still realistic in structure but have not been textured i.e. smooth

shaded. Massaro et al. (1998), Munhall et al. (2004), and Hofer (2009) are examples of this approach. While not used for head motion synthesis so far, some early work by Hadar et al. (1983) used the information available from silhouettes for analysis. The results from the Hadar et al. (1983) study are widely used in head motion research, thus it is appropriate to evaluate if silhouettes would be useful for subjective evaluation.

There are two common ways of presenting different types of synthesis used in speech technologies like TTS and head motion synthesis. The first is an A/B test, the second is a Mean Opinion Score (MOS) test. In an A/B test participants state a preference between two samples. Depending on the implementation they may be able to select no preference. Alternatively the participants have to pick one of the samples, in which case it is called an A/B Forced choice.

A MOS test has the participants rate the samples on a fixed scale, usually between 5 and 10 points. Where the scale represents quality from bad at zero to excellent at the other extreme. In this test each sample is presented individually.

For evaluating similar quality signals the International Telecommunication Union (ITU) recommends an variation of a MOS test called a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test[1]. In this test subjects rate their preference with a continuous slider. On the slider certain regions have been marked as: Bad, Poor, Fair, Good, and Excellent. In a MUSHRA test all of the different types of stimuli are shown at once. The participants are able to repeatedly listen to the samples until they have decided on all of the ratings. This is as opposed to a MOS test where only one type of sample is shown at a time.

In a MUSHRA test a labelled reference is given against which the other samples are compared. Additionally among the samples that are being evaluated there is a ‘hidden

[1] The ITU standard is for MUSHRA implementation is: *Method for the Subjective Assessment of Intermediate Quality Level of Coding System*, International Telecommunications Union Radiocommunications Assembly, Standard number: ITU-R Rec. BS.1534-1, (2003), Available at `http://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-0-200106-S!!PDF-E.pdf`

reference' and a 'hidden anchor'. What this means is that among the samples the reference signal is given again. Also a sample that has been deliberately degraded to a worst case scenario which is called the anchor. These give the upper and lower bounds against which the other samples are rated. In the case of head motion synthesis we have a convenient reference: the original motion capture. The ITU recommends using a low pass filtered version of the reference as the anchor.

The original ITU recommendations are based on testing quality of audio modification an example application would be testing compression algorithms. The standard was not specifically intended for speech technology, so not necessarily all of the recommendations are relevant to this application.

The problem with MUSHRA and MOS testing is that they are prone to bias which can originate from many different sources (Zielinski et al., 2007). For the most part these can be addressed by designing the interface and experimental conditions according to recommended practice. Additionally a researcher should obtain obtaining a sufficient number of samples from all of the subjects for statistical confidence. However, we wanted to examine which of these biases would be important in head motion synthesis evaluation.

It is important to remember that what we seek in an objective evaluation is the largest difference in the scores between good and bad synthesis, not the highest possible score. For example if the participant was shown no movement and motion capture and they rated both very highly it will not be a useful test.

4.2.2 Considerations for Subjective Evaluation

While there are potentially hundreds of factors that can influence the outcome of evaluations we limited ourselves to what we thought would be the most important factors that could cause bias and can be mitigated by experimental design. These factors are:

- What type of animation should be used?
- How long does the training phase need to be?
- How long can the test last before the participants become bored and stop paying attention?
- Can only native speakers of the language be used as participants?
- Does the listening environment have a significant impact on the results?

Going through these factors in more detail. The famous uncanny valley effect (Seyama and Nagayama, 2007) could mean that too realistic animation might all be considered 'creepy'. While with a less realistic animation the participants can focus on the movement. However, if the animation is too unrealistic it may not be possible for participants to tell the difference between good and bad animation.

In fact the type of animation has already been shown to have a large impact on a subjective evaluation (Hofer, 2009). However, (Hofer, 2009) did not compare his results to head motion that is guaranteed to look natural but is definitely not synchronised. He compared the subjective evaluation results of different types of animation using motion capture, synthesised motion, and randomly generated head motion gestures. This random case may create results that look very unnatural, and one cannot be sure of the quality of synthesised motion.

It should be noted that eye and lip movement have a large impact on the perception of naturalness (Massaro et al., 1998; Lee et al., 2002). Thus they need to be excluded or participants will also focus on their quality. Even if the eye and lip motion are the same for all types of head motion, the participants will rate both good and bad head motion as being unnatural if the lip and eye motion is poor. In the (Hofer, 2009) study lip and eye motion were also included which may have influenced the results.

Most people are not used to evaluating head motion. Thus they are not capable of immediately differentiating between good and bad head motion. Participants also need

to get used to the interface used in the experiment. In speech synthesis research this is addressed by including a training phase (Benoît et al., 1996). For head motion it is not known how long the training phase would need to be. A training phase is also needed in other speech technology applications, and the ITU MUSHRA standard (Footnote 1) includes a training phase. These points strengthens the case for the use of a training phase in head motion synthesis testing.

Mental fatigue during difficult tasks is a well studied field. In particular Orden et al. (2000) showed that performance in visual tasks would degrade over a sustained period. As head motion comparisons are a visual task this would mean that the participants' ability to differentiate head motion quality would get worse over time. Additionally Persson et al. (2007) found that fatigue impacts on 'interference tasks'. Simply put this is the process of filtering out relevant data, which would be part of making comparisons between stimuli, further affecting the results. This effect is not only psychological, by using an EEG Boksem et al. (2005) were able to show that as fatigue increased so attention decreased indicating that there would not be method to prevent this effect with the correct instructions, and that all participants would probably experience this problem. All told this means that subjective evaluation cannot last for too long. As fatigue during head motion evaluation has not been directly studied, the maximum length of the evaluation will also need to be determined.

Being able to conduct head motion synthesis evaluation over the internet would be far more convenient compared to bringing participants into a laboratory. However, Reips (2002) warns that it is possible that other factors in the environment can skew results. Further (Kittur et al., 2008) warns that not all experiments can be done online. Nevertheless, Buchholz and Latorre (2011) showed that online evaluation is suitable for preference testing. Furthermore Wolters et al. (2010) found that online evaluation can be used for speech quality testing, despite the fact that audio quality is known to change between headphones and speakers. So while there is little doubt that the

environment in which one performs the evaluation will have an impact on the results, the effect may not be large. Alternatively if the results from online testing have a similar distribution to those from an evaluation performed in a laboratory, possibly after some reasonable transform, then the different conditions can be controlled for.

There is little doubt that culture has an impact on which gestures you produce when you speak, and this has long been studied (Graham and Argyle, 1975). In particular native language has a large impact on the types of gesture you would make Kita (2009). However, despite producing different gestures, what we perceive as natural may not be dependent on culture. In effect we may be taking culture into account when evaluating head motion. A broad test for this would be to determine if native English and non-native English speakers give the same, or similar results. If this is the case then both can be used for subjective evaluation, if not, then it would be important to determine if other native speakers of English, but from a different culture give the same results, for example US native English speakers, recalling that our dataset consists only of UK native English speakers. Otherwise if culture has a large impact on the evaluation of head motion synthesis it is important that it is taken into account when conducting evaluation.

To find the impact of the above mentioned factors two experiments were conducted. The first will examine:

- The length of the training phase
- The length of the test
- The type of rendering

The second will examine:

- The types of participant
  - Native or Non-native English Speakers

  - Male or Female

- The listening environments
  - Laboratory or over internet
  - Headphones or speakers
- The expected results for different types of head motion trajectories

To meet these aims the first experiment will be very long and show multiple types of rendering, be conducted in a laboratory, and only use native English speakers. While the second experiment will be conducted over the internet, will use all types of participants, and both headphones and speakers. In this experiment only the type of rendering that gave the best results in the first experiment will be used, and the length of the experiment will be limited to what was found to be appropriate in the first experiment.

To determine the length of training phase the results between the participants will be compared and when they are similar then the training phase is complete and the results before that point should be discarded. The length of the test, or point of boredom, will either show up as the scores all tending to same regardless of the rendering type, becoming random regardless of the rendering type, or the results between participants no longer being consistent.

The second experiment uses the same interface as the first, except only using the correct rendering type and the correct length for the experiment. For the second experiment the difference between native and non-native speakers and between headphones and speakers can all be determined using statistical significance testing. The necessity of using a laboratory can be evaluated by comparing the results to the first experiment again using statistical significance.

The other goal of the second experiment is to determine the expected results with an anchor, in this case the anchor will be head motion shaped noise, the generation of which is described below.

The next two sections will give details of each experiment, the results, and some preliminary analysis. The full analysis and discussion will is presented in Section 4.4.

4.2.3 Experiment 1: Length of Test and Animation Style

4.2.3.1 Experimental setup

The aims for this experiment were to determine which type of rendering should be used and how long the test should run for. In Figures 4.1 and 4.2 stills of the videos used for the different types of rendering are shown. While Figure 4.3 shows a screen grab of the web interface used for the test. The users are able to repeat each video as many times as they wish, in any order they like. As both data from Male and Female speakers was used, two models were needed. All speakers on one gender used the model which was appropriate. These 3D models are bundled with Poser [2] which was also used to create the animations from the trajectories. The silhouette was generated from the same the 3D models used for the other types of rendering.

As was previously mentioned it is important to remove the impact of eye and lip motion. Eye and lip motion are linked to head motion, so reusing them will not be appropriate (Lee et al., 2002). This means that the eyes and lips must be obscured. We showed three volunteers some samples and asked them to pick which looked the best from:

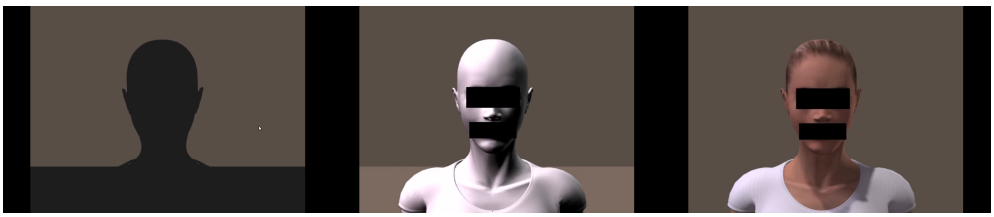
- A black block over the eyes and mouth
- A head with no eyes or mouth
- A "news blurring" effect

The feedback was that while the news blurring and black box seemed about the same, removing the nose and mouth was very creepy. However, news blurring made it more

[2] http://my.smithmicro.com/poser-3d-animation-software.html

(a) Silhouette (b) Smooth shaded (c) Textured

Figure 4.1: Animation stills for male speakers for experiment to determine the type of rendering that is appropriate for subjective evaluation



(d) Silhouette (e) Smooth shaded (f) Textured

Figure 4.2: Animation stills for female speakers for experiment to determine the type of rendering that is appropriate for subjective evaluation

difficult to see the head movement, so the black box option was chosen.

As the audio was always recorded the only differences between samples were the type of rendering and whether the type of head motion trajectory

Four speakers, two male and two female, were chosen at random from the dataset for the evaluation. Two types of motion were included in the evaluation. The first was the original motion capture and the second was also motion capture but from a different sample of the same speaker. The former was the gold standard, and the second was used to be analogous to natural looking but desynchronised speech synthesis. This would be more difficult to differentiate from the original than completely random motion or 'head motion shaped' noise. While random motion is a worst case scenario, the

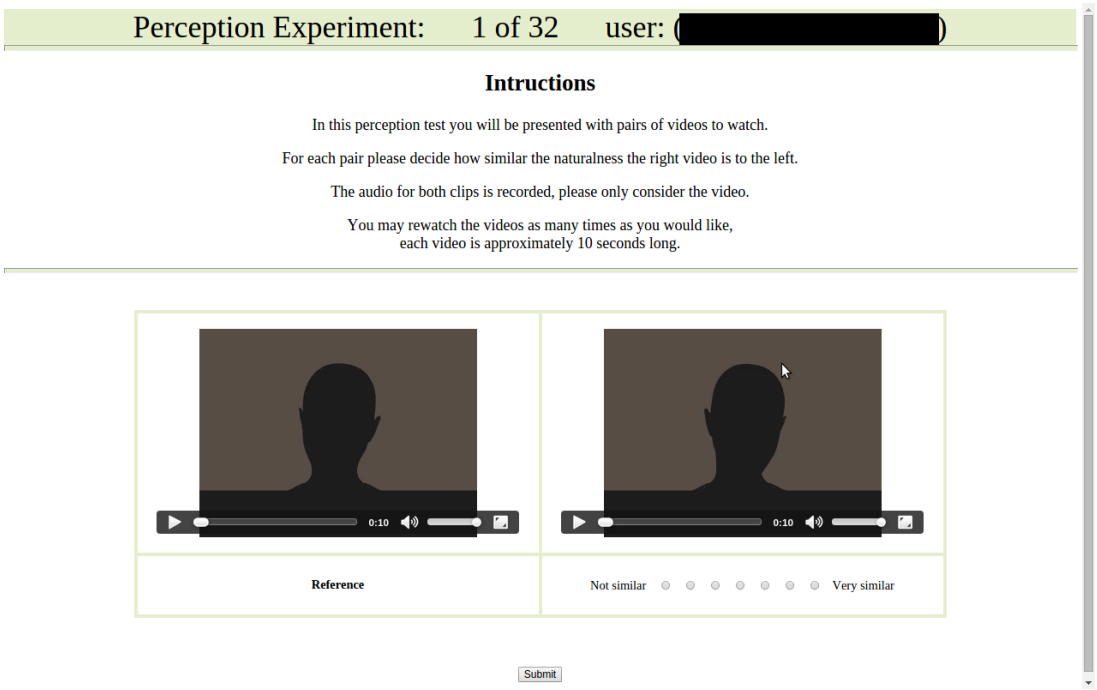


Figure 4.3: Web interface used for all subjective evaluations, showing layout and instructions

aim of this experiment is not to find the full scale. Instead we aimed to find which rendering with which the participants were best able to distinguish between high quality synthesis systems and motion capture. Including a third animation type like random motion would greatly increase the length test and was not needed for this goal so it was excluded. The same amount of each animation type, speaker, and motion type were shown to each participant.

The experiment consisted of 72 pairs of videos. This would take approximately an hour for a participant to complete. Each video was approximately 10 s long, though they were cut so that they did not end mid-sentence. More videos were generated than any one participant would see. The videos were shown in a random order of animation type, speaker, and sample from the speaker, which was different for each participant. This should remove any order bias in the average scores. The reference video was always on the left and was always motion capture. The participants were told that it was the reference. This meant that the higher the score the better the animation,

making the results easier to interpret. If it had not always been motion capture then interpretation of the results becomes difficult. The reference videos were chosen from different samples than the evaluation videos. The reference and test videos were the same animation type. For the all the animations the head motion included rotation and translation. While it would be ideal to include an anchor in the test, the amount of stimuli would mean that adding a third type of animation would mean a much longer test, and require many more participants, thus it was left for the next experiment to determine the range of expected results by including the anchor.

For this test the participants were seated in a sound booth and presented with the stimuli on a computer, using a web browser. The participants had headphones and the sound booths are a low noise environment. This should remove environmental factors in the evaluation. By using the same model of headphones for all participants the listening environment should be the same and so should also not be a factor in this experiment. Adjacent to the sound booths is a control room where the experimenter was seated.

The participants were asked to rate on a seven point scale "How similar is the video on the right to video on the left". This is a combination of MOS and MUSHRA methods. Due to screen size constraints only one stimulus is shown at a time like a MOS test, but a reference is provided so that the results have an upper anchor. More than two videos on the screen would mean that they become too small and hard to judge. As mentioned previously time constraints mean that a lower anchor could not be included, which meant a proper MUSHRA test would not be possible. A small pilot version of this experiment with only three participants showed that a reference was needed, otherwise pure MOS methodology would have been used at this stage. As a MUSHA test has 6 boundaries between 5 categories and we wanted to include a 'no preference' option, a seven point scale was used.

The participants were specifically told that the audio was recorded and not to base their evaluation on it. The participants then had the opportunity to ask questions from

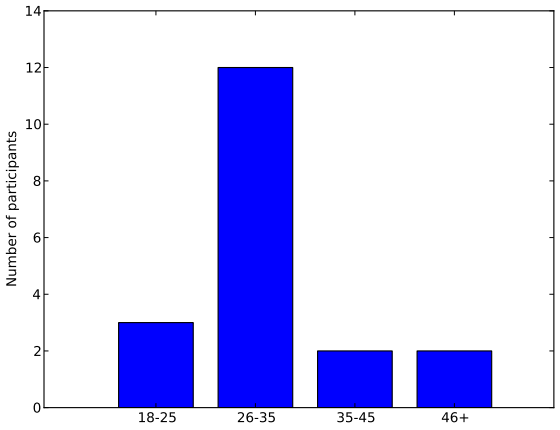


Figure 4.4: Distribution of age for participants in Experiment 1

the experimenter before they began. The experimenter remained in a control room during the evaluation in case the participants had further questions once they started. The participants were not allowed to re-evaluate videos once they made their decision. The participants were allowed to view both videos as many times as they liked before making their decision. The experimenter did not know the order of presentation for any of the speakers so as to maintain a double blind standard.

4.2.3.2 Results

Twenty participants were recruited, however, the results of one were excluded because they rated all videos the same. All the participants were native English speakers of which 11 were native to the UK. There were 10 male and 9 female participants and the age distribution is shown in Figure 4.4.

The most important output of this experiment is the difference between the scores from each participant between motion capture and desynchronised motion capture. It is expected that initially participants will need to become accustomed to the test and so they first will have a training phase, also after time they will become bored and their answers will be meaningless.

Different participants will perceive the differences to greater or lesser extents. So all the scores are scaled to be on the unit interval for each participant individually. Additionally each participant will be able to evaluate some speakers better than others so the results from each speaker also have to be scaled separately. It is important to note that this type of test will have each participant evaluate multiple pairs of videos sequentially. Thus rather than examining the difference in scores for each pair individually, it is more useful to look at a window of pairs. This window represents the expected results for a test where everything before is regarded as training examples and everything after is meaningless due to participants loosing interest. The effect of this process is similar to smoothing.

In Figure 4.5 the mean difference in score over the window between the original motion capture and the desynchronised motion capture. This is graphed separately for each animation type. From this graph it seems that the smooth shaded model is the least suitable as it does not have discriminatory power of either silhouettes or fully textured videos.

In Figure 4.6 we have replotted the same data as Figure 4.5 but now we have included the standard deviation for both of the rendering types that might be suitable. From the figure it is clear that the standard deviation of the silhouette animation is lower than that of the textured ones. However, the textured animation is consistently above zero, which means participants consistently rate the synchronised higher than the desynchronised samples. Also the peak difference is higher for textured models. This implies that it is better to use realistic models, as would be expected.

It is important to note that this test was meant to be a worst case scenario, normally a test would also have a condition such as noise, to anchor the lower end of the test. In the second experiment we will include head motion shaped noise so that we would also have an idea of the range of results one could expect.

With regards to the length of the test, Figure 4.6 shows that the due to the high variance

at least the first 10 samples should be considered training data at a minimum. The useful period seems to be from approximately sample 10 to 40. After this period the variance begins to increase again which would indicate that the participants have lost focus.

A total of 40 samples equates to a test of up to approximately half an hour. Table 4.1 shows some key statistics about the results from the period. First it should be noted that the variance of smooth shaded rendering is higher than both silhouettes and textured rendering. The variance of silhouettes and textured rendering are very similar. Also the mean difference in score the silhouette is slightly higher than the textured rendering but both are much great than smooth shaded. It is interesting to note that there was no significant difference in the amount of time it takes participants to make their decision.

Table 4.1: Key statistics of the difference between the rescaled scores of motion capture and desynchronised motion for pairs 10 to 40

	Silhouette	Smooth shaded	Textured
Mean difference	0.20	0.04	0.14
Variance	0.06	0.10	0.08
Mode difference	0.20	0.04	0.14
Mean time for decision	32.3 s	31.7 s	32.1 s
Variance	16.7	18.0	14.8

Figure 4.7 shows the distribution of scores after they been normalised as above. This figure shows that for textured samples more synchronised samples are rated higher and desynchronised samples are rated lower than samples with the other types of rendering. Though the difference is not very large.

What is somewhat surprising is that the smooth shaded model performed so poorly. While it is impossible to state for certain what was the cause, based on informally speaking to some of the participants after the experiment, it is probably the case that

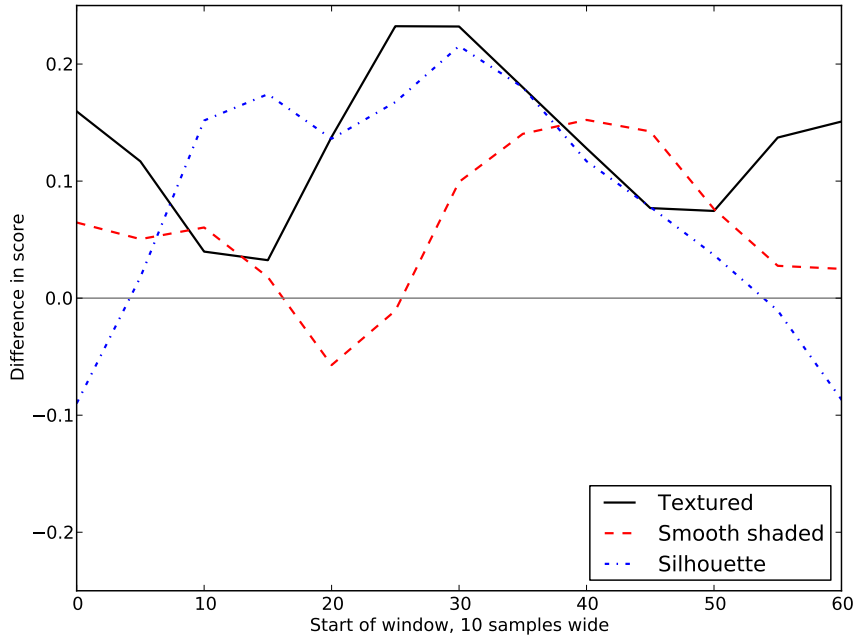
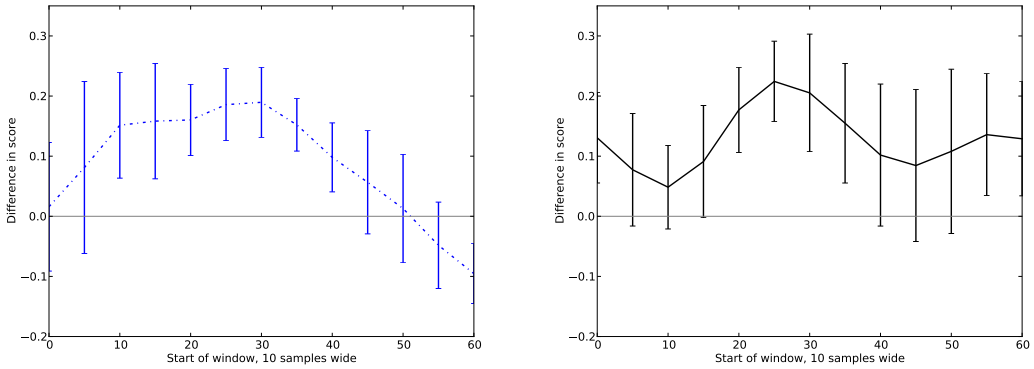
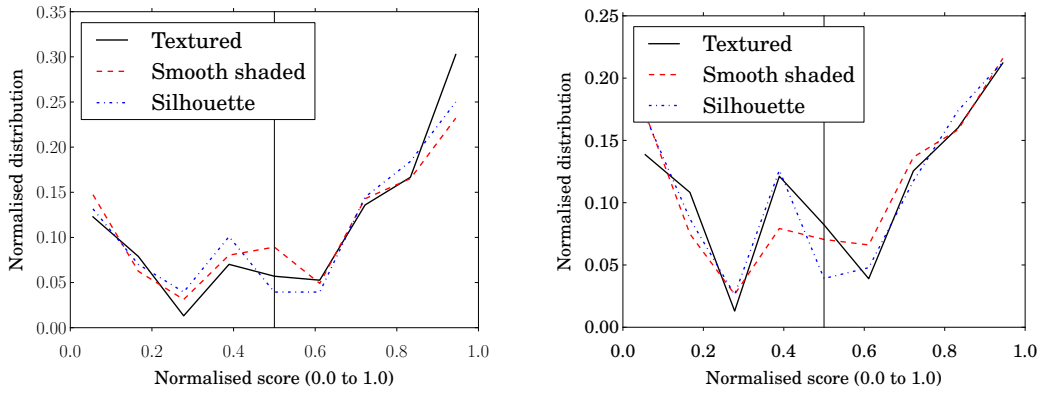


Figure 4.5: Difference between a 10 sample wide window of scores normalised by participant and speaker of motion capture and desynchronised motion capture for different types of rendering as the number of samples already viewed increases



(a) Silhouette (b) Textured

Figure 4.6: Differences of score showing standard deviation



(a) Motion capture

(b) Desynchronised

Figure 4.7: Histogram of the scores for synchronised and desynchronised speech. Scores are normalised by speaker and participant, the histogram is normalised to have an area totalling 1.0 so that it can be used as a pdf.

this unusual presentation was too distracting. One participant in particular stated that to him, the silhouette was like listening to a speaker and only being able to see their shadow. The implication is that while silhouettes are less human-like, it is not implausible to listen to someone and only see a silhouette. On the other hand the textured model is as closest to looking like an actual person in this experiment. This leaves the smooth shaded model as neither the best at looking human, nor a proxy for a reasonable scenario, which may explain why it performed worse than the other options.

Because the difference in mean score is not very large between textured and silhouette animation, the deciding factor as to which is best to use falls to consistency. In this case, when participants have textured animation they are always able to rate the synchronised motion higher than the desynchronised motion, while with silhouettes there are times when they rate desynchronised motion higher. This in turn means that we should be using textured head motion in subjective evaluations.

Table 4.2: Make-up of participants for Experiment 2

Male	10	Female	13
Native	15	Non-native	8
Speakers	15	Headphones	8

4.2.4 Experiment 2: Participant and Environment Suitability

4.2.4.1 Experimental setup

The second experiment had a similar experimental setup to the first. In this experiment the length of the test was 36 pairs of videos as this was the within the bounds found to be appropriate by the first experiment. Similarly only textured models were used as they were found to be the type of rendering with which participants were best able to differentiate between different quality head motion trajectories.

The aim of this experiment was to determine if participants needed to be in a sound booth or would over the internet experiments be suitable. Related to this is whether the participants needed to use headphones or would speakers suffice. We also wished to establish who would be suitable participants. An additional aim was to determine the range of MOS scores one could expect from as subjective test with a hidden reference and hidden anchor as used in MUSHRA testing.

Participants were thus recruited to perform the experiment online. They were allowed to use headphones or speakers which accounts for both possible methods which participants would view the videos. Additionally not all the participants were native English speakers. At the start of the experiment these demographics were collected. The demographics of the participants that took part in this experiment are given in Table 4.2. Note that the participants from the first experiment were not brought back. This is due to the fact that they already have experience and this could skew the results.

The types of rendering shown were motion capture, desynchronised motion capture, and the anchor was random smooth 'head motion shaped noise'. This was generated by first creating white - noise then taking the Fourier Transform of the head motion trajectories and applying it as a filter in the frequency domain to the white - noise. Then we invert the noise back into the time domain and apply five point Gaussian smoothing.

The instructions were the same as the first experiment and the scale was still seven points. Other than the inclusion of a demographics questionnaire at the start of the experiment the interface was the same as the first experiment.

We wished to either confirm or reject the Null Hypothesis that two groups gave different results, thus the Student's T-test was the appropriate measure (Field, 2013, pp. 303 – 304). This tests the similarity of two distributions, if they are too similar then the Null Hypothesis cannot be rejected, implying that the results of the two groups would be the same. This is the case when the T-test's score, p, is high. So a high p value would indicate that the two different groups being tested would give the same ratings to each type of head motion trajectory (synchronised, desynchronised, or noise).

4.2.4.2 Results

The Student's T-test was carried out after each participant's results were scaled to the unit range as per Experiment 1. The results are given in Table 4.3. Bearing in mind that p values of greater than 0.05 are considered high (Field, 2013, pp. 303 – 304), the p values are all beyond the range where we can exclude the null hypothesis. Thus we can conclude that the different the groups of participants give the same distribution of normalised results. That being said, some pairs of groups are more different than others, but none of the differences are statistically significant.

The second aim of the experiment was to get a sense of the values one might expect

Table 4.3: Student's T-test results (p values) for the distribution of opinion scores, normalised by participant and speaker, between pairs of groups of different participants

Participants	Original	Desynchronised	Head Motion Shaped Noise
Male and Female	0.67	0.34	0.51
Native and nonnative	0.66	0.77	0.69
Headphone and Speakers	0.98	0.12	0.22

Table 4.4: Mean opinion scores from Experiment 2, normalised results are obtained by rescaling results to the unit range by participant and speaker, then calculating the mean

Participants	Original	Desynchronised	Head Motion Shaped Noise
Male	4.51	4.33	3.16
Female	4.90	4.62	3.49
Native	4.77	4.56	3.45
Nonnative	4.67	4.38	3.13
Headphone	4.71	4.57	3.51
Speakers	4.74	4.46	3.04
Mean results	4.73	4.51	3.34
Mean normalised results	0.57	0.52	0.41

from a seven point MOS test. As stated above the conditions we tested were original, desynchronised, and head motion shaped noise. The MOS scores broken down by group are given in Table 4.4. We have given the unnormalised form so that the absolute values can be seen. We can see that certain groups have higher ratings than others, but the test before shows that this is only the absolute rating. The ratings of the types of trajectories relative to other types are the same across all groups of participants.

4.3 Objective Measures

4.3.1 Canonical Correlation Analysis and Head Motion Synthesis

One of the most commonly used correlation tests for two streams of multivariate data, such as head motion trajectories and speech features, is Canonical Correlation Analysis (CCA) (Alpert and Peterson, 1972; Lambert and Durand, 1975). It was formally introduced by Hotelling (1936) and is analogous to a multidimensional extension to Pearson's correlation. The idea is to map two streams of data, which may not be the same width, onto a common hyperplane and then find the Pearson's correlation between vectors in that plane. For two streams of multivariate data arranged into a matrix, where each of the rows corresponds to one observation, $\mathbf{X} \in \mathbb{R}^{n \times T}$ and $\mathbf{Y} \in \mathbb{R}^{m \times T}$, and cor is the Pearson's correlation function, the canonical correlation score $\rho^{(c)}$ is defined to be

$$\rho^{(c)} = \text{cor}(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y}), \tag{4.1}$$

where $\mathbf{a}$ is the $n \times 1$ vector and $\mathbf{b}$ is the $m \times 1$ vector that satisfying

$$\mathbf{a}, \mathbf{b} = \arg\max_{a,b} \text{cor}(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y}). \tag{4.2}$$

It is also useful to define

$$U_1 = \mathbf{a}'\mathbf{X} \tag{4.3}$$

$$V_1 = \mathbf{b}'\mathbf{Y} \tag{4.4}$$

This process is then repeated with the added constraint that new **a** and **b** are uncorrelated with the first and so on. The complete set of these vectors is **A** and **B**. The complete sets of **U** and **V** are known as canonical variables or scores. These vectors lie on the hyperplane mentioned above. **X** is mapped onto the hyperplane by **A** and **Y** is mapped onto the hyperplane by **B**.

There are several methods for solving for **A** and **B**. For instance the one implemented in Matlab is based on singular value decomposition [3].

Canonical correlation scores are always positive and lie on the range $[0,1]$. Because it is maximised, negative correlation would be converted to a positive $\rho^{(c)}$ value by the mapping vectors **a** or **b** changing direction. Like Pearson's correlation a result of zero indicates no correlation, while a result of one would indicate a perfect correlation. Traditionally any correlation above 0.1 is considered significant (Alpert and Peterson, 1972). CCA values are traditionally denoted ρ.

There are is an important limitation in that CCA operates on each row individually, i.e. it is a frame-wise function. This means that CCA does not take into account the temporal ordering of the streams (Alpert and Peterson, 1972). One can mitigate this effect by adding derivatives to the data streams. It is also known that if there are too few data points CCA can identify spurious correlations and thus will show a high correlation when there is none (Lambert and Durand, 1975). Despite these limitations many researchers, such as Busso et al. (2007) and Hofer (2009), use CCA as part of the object evaluation of head motion synthesis.

Due to these limitations we tested CCA to determine if it was suitable for head motion

[3] http://www.mathworks.co.uk/help/stats/canoncorr.html

synthesis evaluation. The key benchmark is whether or not it is able to show the difference between trajectories synchronised with audio and those that may follow natural head motion trajectories but are unsynchronised with the audio. Beyond that it would be useful if the results of the objective test predict the results of a subjective evaluation. In the rest of this section we will show that CCA in its standard form is not suitable for head motion synthesis evaluation based on these criteria. We will then show in the next section a modification to the standard method of applying CCA that provides more useful results.

It is common in the literature to use short sentences for analysis of the relationship between acoustic and head motion. We will show that this is not a good choice for analysing this relationship. This is due to the fact that it would only show correlations that are short term which cannot be used for head motion synthesis evaluation. By this we mean that the correlations that are found on short utterances only exist for that sample.

Concatenating samples is the correct way to find the correlation between two streams of data. This is because taking the mean of multiple samples shows only the short term correlations, while concatenating shows the overall correlation. Because CCA does not take ordering of samples into account the concatenation boundaries do not have to be treated specially. We will term the mean CCA score of all the short utterances as local correlation, ρ_l, and CCA of all the samples concatenated as global correlation, ρ_g. If the short utterances are concatenated then these correlations almost disappear. In other words despite ρ_l being high, when calculating ρ_g the correlation is shown to be much lower.

To examine the difference between local and global CCA we took the short utterances recorded by Hofer (2009) and concatenated them. This gave approximately 10 minutes of data. From here we calculated the global CCA between MFCCs, F0, and Energy, and the head motion trajectory (these were the features used by Hofer (2009)). We

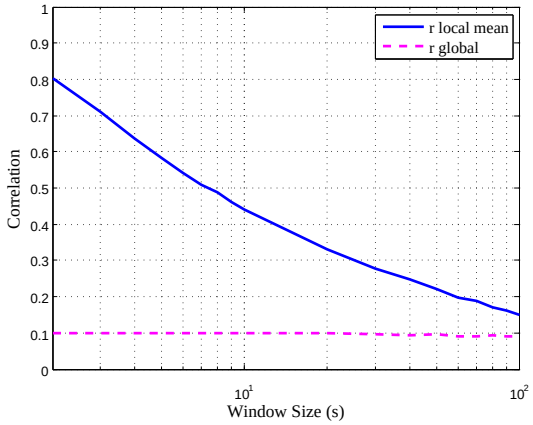


Figure 4.8: Global and local CCA between speech features and head motion. CCA was calculated with different sized windows, for global the windows were concatenated, for local the mean correlation of the windows was found.

then divided the data into various sized windows to simulate different lengths of utterances. We started at 2 s and went to 100 s (compared to the 600s available). The results are given in Figure 4.8. This figure shows that local CCA quickly trends towards global CCA as the size of utterance increases. So CCA should be computed on larger segments of data.

To test if CCA is suitable under other conditions we computed several correlations between different types of head motion trajectory and speech features. The first was the original head motion to provide a baseline. The second was to head motion from the same speaker but a different recording, in other words the desynchronised condition from the subjective evaluation testing. We also calculated the CCA scores between speech features as head motion from a different speaker, and the last test was the correlation of speech features and shaped noise generated the same way as in the subjective evaluations. In all cases we included the first derivative of the trajectory.

Instead of only using the speech features we used above, we tested a wide variety of options, with and without the dynamic features (first and second derivatives) included. The legend for all the results is provided in Figure 4.9, where E is short for energy and

D indicates that the dynamic features were included, we have also included a sample correlation plot.

The correlations between head motion trajectories and speech features are shown in Figure 4.10. In this graph the samples marked long are all from the UoE-HAS dataset, while the samples marked read and free are short utterances that were recorded with the same recording conditions as the UoE-HAS database, except that these were short utterances. The free speech was elicited by getting the participant to say a quote from a film they knew and the read speech was also film quotes, but read off a Teleprompter. Due to problems with the tracking short free speech was only available from one of the two speakers who recorded short utterances. Thus it was not possible to calculate the CCA for motion from a different speaker who recorded under the same conditions. As we will discuss below, short recordings are not suitable for training so there was no need to recapture the data.

Figure 4.10 shows some correlation. The problem is that for long utterances the correlation does not change under different conditions significantly. For the short utterances there were some differences. This shows that the behaviour of short and long utterances are different. A real system would rely on longer utterances as they are the more interesting and the avatars would need to be able to speak for long periods. The fact that shorter utterances have different behaviour means that they are less suitable to be used for training data compared to long utterances. The long utterances, even though they are from different speakers, all have similar behaviour. Regardless, the fact that longer utterances do not have different CCA scores despite the different conditions shows that CCA is not suitable for objectively measuring the quality of head motion synthesis.

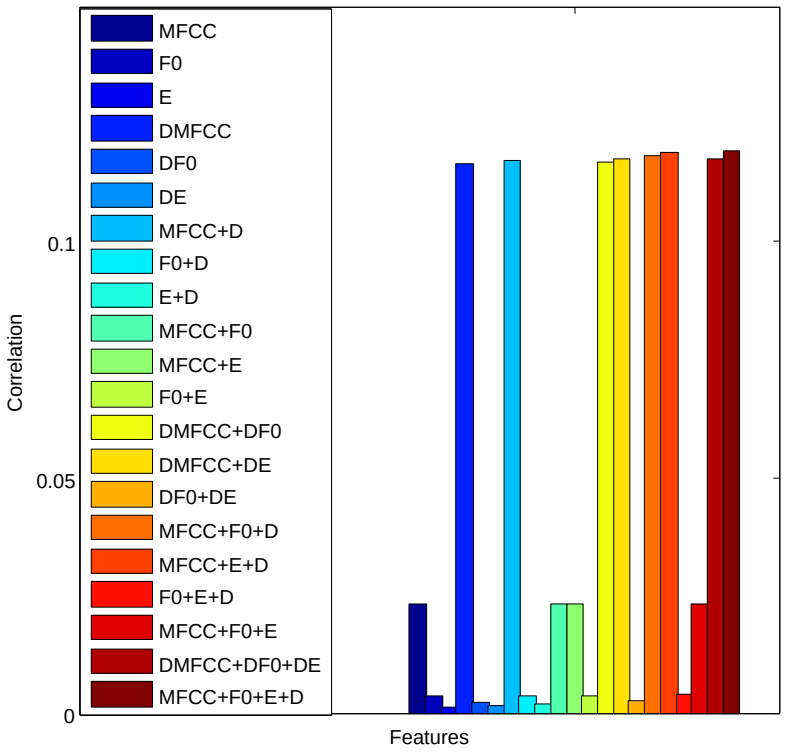


Figure 4.9: Legend of speech features used for testing objective measures for head motion synthesis in Section4.3. E is short for energy, and D indicates that the first and second derivative was included. Also included is a smaple plot of correlation between speech features and head motion trajectories. Note that the order of the legend is the same as in the plots.

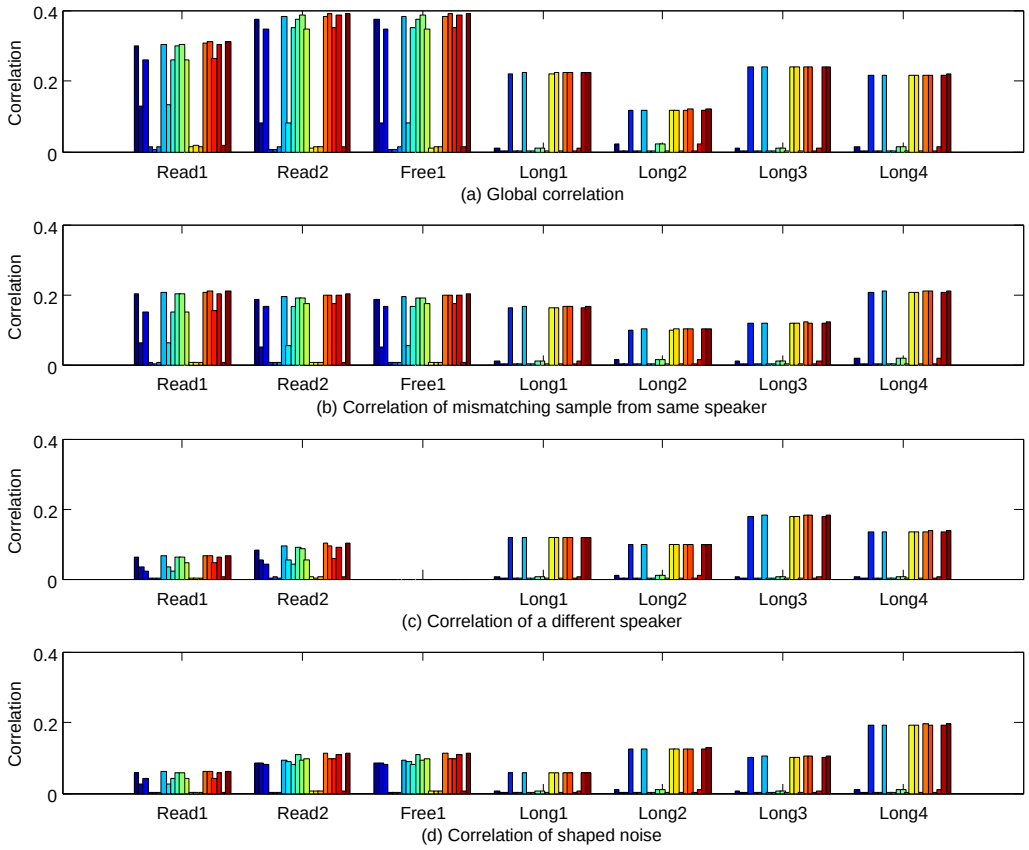


Figure 4.10: Global correlations between speech features and head motion trajectories for different types of trajectory. Read and Free are short utterances, Long samples are from UoE-HAS. Each column represents one sample from a different speaker.

4.3.2 Forced Canonical Correlation Analysis

While CCA directly does not give informative results, we can still exploit the relationship it finds. Considering that head motion is probably non-deterministic given only speech features and not any psycho-linguistic information, using RMS error would also not be a suitable measure. Instead we can look at how well the new head motion duplicates the relationship to the original head motion. As this relationship is already described by the CCA between the original head motion and the speech features we can reuse the mapping matrices **A** and **B** and then calculate the correlation with the synthesised head motion. Instead of seeing how well the new motion duplicates the the original motion this will show how well the new motion replicates the relationship between the original to the speech features.

We will term this new measure Forced CCA (FCCA) with the symbol ρ_f. The name is chosen to reflect that the scores are determined by forcing the reuse of the mapping optimised for the original head motion. Mathematically we define it as

$$\rho_f = \text{abs}(\text{cor}(\mathbf{a}'\mathbf{X}^*, \mathbf{b}'\mathbf{Y})) \tag{4.5}$$

where

$$\mathbf{a}, \mathbf{b} = \arg\max_{\mathbf{a},\mathbf{b}} \text{cor}(\mathbf{a}'\mathbf{X}, \mathbf{b}'\mathbf{Y}) \tag{4.6}$$

and $\mathbf{X}$ and $\mathbf{X}^*$ are the original and synthesised head motion trajectories respectively, and $\mathbf{Y}$ is the acoustic features trajectory. Or to put it into words, ρ_f is the correlation found by reusing the **A** and **B** matrices that were determined by calculating performing CCA between the original trajectory and the acoustic features.

The results from performing FCCA are shown in Figure 4.11. The results for FCCA with the original is obviously the same as the original CCA. What is clear is that there is a definite decrease in the correlation scores when using FCCA with conditions other than reusing the original trajectory. Also, in general, the lowest value was for

noise, which is ideal. The exceptions in Figure 4.11 show very low correlations for all conditions.

While some preliminary work is shown here about the use of speech features they will be chosen to suit the synthesis system. Though it is already clear that some speech features will be better suited for use with head motion synthesis than others.

4.4 Analysis and Discussion

This chapter presented our finding about subjective and objective evaluation methods.

In terms of subjective evaluation, by utilising a modified version of MUSHRA testing, we found that the current practice of using realistic avatars was the best approach for evaluating head motion. This is because during comparative testing participants are reliably able to tell when head motion is synchronised.

Of particular importance was the guideline for the length of the test that was developed. In general, the first 10 samples should be discarded as this would be a learning phase, and the test should last approximately 30 minutes. This equates to a total of about 40 samples that are approximately 10 seconds long. While we used 10 second clips the ideal length of samples that ensure the results are consistent should still be found.

While participants would give different absolute scores if one normalises the results from each participant to be on the unit scale then the results are the same no matter the demographics we tested nor whether was it important for them to come into a laboratory.

With regards to objective evaluation, we found that the current common practice of using CCA is not suited for this application. Firstly calculating the mean of the CCA with small sentences does not give a meaningful result, secondly CCA by itself cannot distinguish good and bad head motion. In the process we also showed that short and

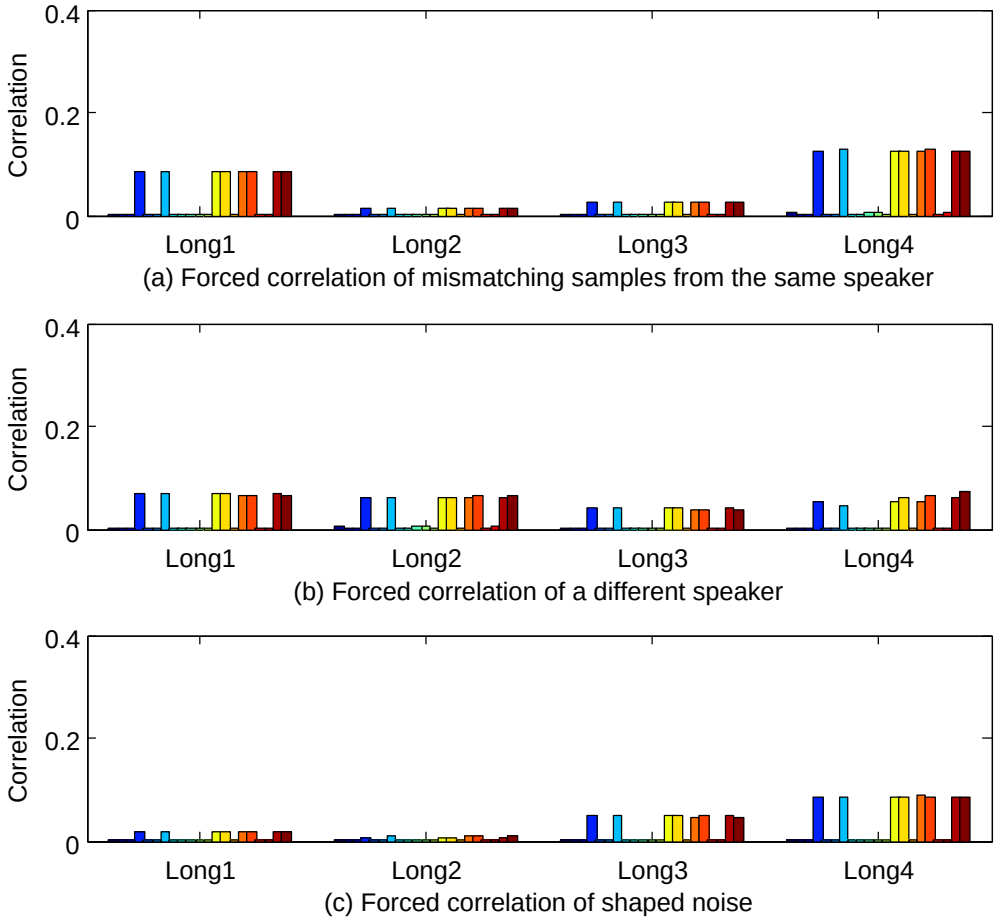


Figure 4.11: Forced CCA correlations between head motion and speech features for the same samples of long speech from UoE-HAS that were used for Figure 4.10.

long utterances have different relationships to speech features. This means that short utterances should not be used for training data.

Instead of using standard CCA we introduced Forced CCA. This objective measure shows the decrease in the quality of the head motion as the type of head motion decreases in appropriateness from original motion capture to head motion shaped noise.

Chapter 5

Template – Warping Based Head Motion Synthesis

5.1 High Level Description of Template - Warping Synthesis

In the majority of existing research rigid head motion is often expressed as Euler angles. In this representation, as the head cannot rotate past a certain point there are maximums in these angles. This is also true in rotation vectors. This in turn leads the angles to have a wave-like motion when plotted individually over time, with the trajectory showing clear peaks and valleys. An example of a head motion trajectory is given in Figure 5.1. This trajectory segment is completely unprocessed data from the UoE-HAS dataset. Normally one would apply some smoothing before any further processing is done. This is only true if there the angles are treated separately. However,

An earlier version of this work was presented at Interspeech 2013 (Braude et al., 2013a)

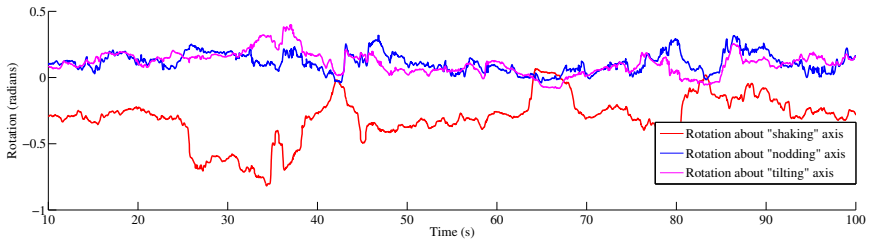


Figure 5.1: An example of Euler Angle trajectories taken from motion capture before any processing.

seldom in synthesis is this prior knowledge about the wave like nature of head motion exploited.

HMMs are capable of recovering highly complex patterns in data. This means that in standard HMM based synthesis methods dynamic constraints must be described explicitly or the movement may have discontinuities. If simple patterns are extracted then this would mean that the dynamic constraints become implicit. This in turn means that the synthesised head motion is guaranteed to be continuous and reasonably smooth. In the case of head motion a simple method of segmenting the data is to define the borders at the peaks and valleys of each of the angle's trajectories.

This approach is similar to unit selection in speech synthesis. The primary difference is that the nature of patterns allow for a much easier concatenation. This is due to the fact that if the correct segmentation rule is used, any pattern can be followed by any other pattern and it would still be continuous.

The number of patterns can be reduced by normalising the segments in both amplitude and duration and by disregarding the initial position or offset of the segment. We can ignore the initial position because the segments will be joined so that they are contiguous and hence the offset is predetermined. These patterns can then be modified to represent any particular segment by reversing the normalisation as appropriate. As these patterns form the basis of the trajectory they can be seen as templates of motion and so we refer to them as 'templates', and modifying the duration, amplitude and

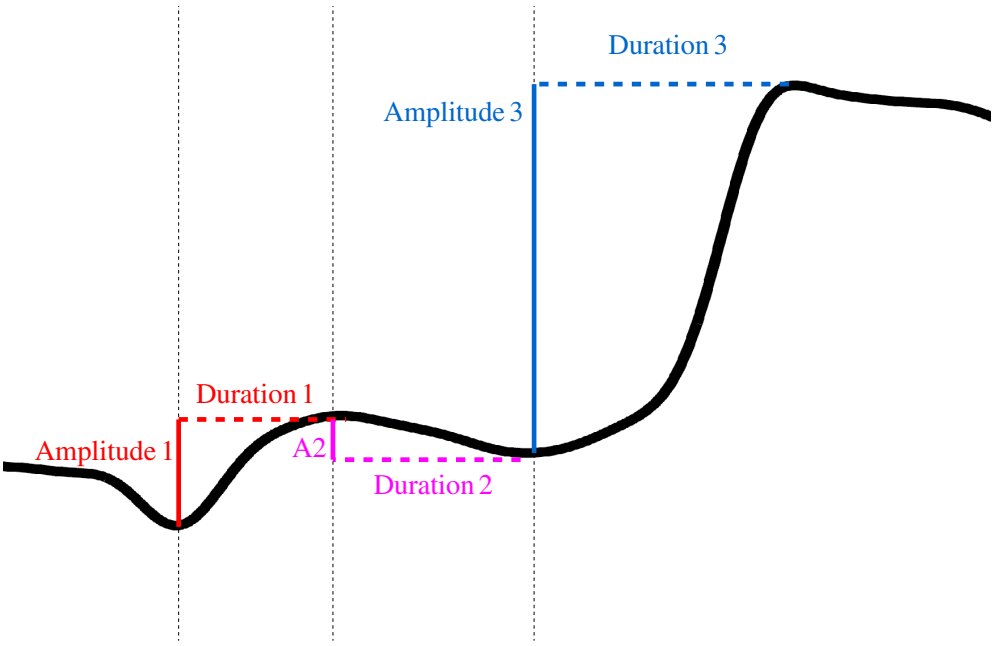


Figure 5.2: A head motion trajectory of one rotation vector component annotated with the segment boundaries (fine dashed lines) and the warping parameters for template - warping based synthesis.

offset will be referred to as 'warping'. Thus this method of creating the head motion trajectories will be called 'template - warping' based synthesis.

Template - warping based synthesis requires the estimation of amplitude, duration, and the offset which we will call collectively the warping parameters and are shown on a sample segment of a head motion trajectory in Figure 5.2. Warping parameters can be estimated from a statistical model.

Formally we will define a template as a function of time, duration, amplitude, and constant offset as follows:

Definition 1. A single head motion template index i for angle κ is expressed as

$$\kappa_t = \mathbf{g}_i^{(\kappa)}(t, \lambda_t^{(\kappa)}),$$
$$\lambda_t^{(\kappa)} = (d_t, a_t, c_t),$$

with duration d, amplitude a, and constant offset c, and time t.

Definition 2. $G^{(\kappa)}$ is the set of all $\mathbf{g}_i$ for $i \in [1 \dots I]$ and for angle κ.

In this chapter we will show that through the use of this method one can create head motion that is very similar to that found in data and find the appropriate amount of templates for head motion. In addition we will present the details of how the segmentation was performed to find the patterns and the choice of templates that were used in the remainder of this research.

We will then show a GMM based approach to synthesis head motion from speech features by predicting warping parameters using GMM regression, then performing a time and amplitude warp on the templates. Next we will show a more sophisticated approach than GMMs. In this modification we use machine learning to find clusters of warping parameters joined to speech features. We then use a two level HMM to recognise these clusters from speech features at synthesis time. Once the clusters have been recognised GMM based regression is used to estimate the warping parameters.

This chapter will also present the results of subjective and objective evaluation which was performed using the recommendations we developed in Chapter 4. These show that template - warping based synthesis is able to outperform other state of the art systems and is comparable to motion capture.

5.2 Segmentation

We will describe the rotation of the head in terms of rotation vectors (see Section 2.2.3), and the components of the vector as r_x, r_y, and r_z. Each component represents rotation about a different orthogonal axis which we have shown in Figure 5.3. As has been mentioned before we will be treating each angle separately. There are a few reasons for this. The first is that creating a rule would be difficult for all three angles simultaneously. Secondly, in the previous section it was mentioned that the head motion in each angle individually will exhibit a wave-like motion, this lends itself to the segmentation

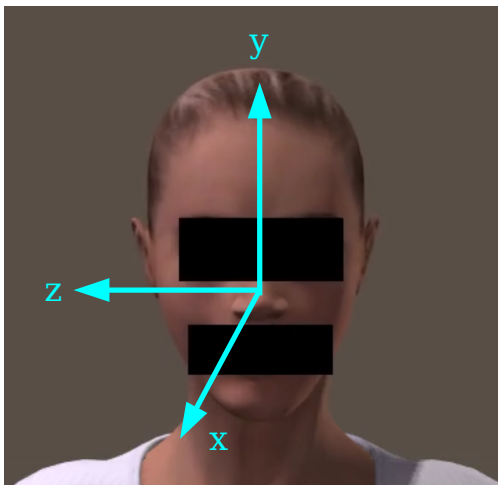


Figure 5.3: Coordinate system about which the head is rotating

rule described below. Additionally this allows for a reduction in the amount of clusters as will be described in Section 5.6.1, but essentially it is because with all three angles being segmented simultaneously one would need to account for all possible combinations of the segmentation that would be found by treating the angles individually. This would necessitate an exponential growth in the amount of training data needed as the amount of templates, clusters, and/or HMM states increase i.e. the famous curse of dimensionality. Finally, it lends itself to generalisation to other applications with a different amount of variables that need to be predicted.

The segmentation rule is defined to be:

Definition 3. Segment boundaries for angle κ occur $\forall t$ which satisfies

$$\frac{\mathrm{d}\kappa}{\mathrm{d}t} = 0.$$

In practice this means that the segments occur at the peaks and troughs of the wave like motion as shown by the fine dashed lines in Figure 5.2. However, estimation of the first derivative is prone to noise. To deal with this we first smoothed the trajectory with a Gaussian filter that was set to be 5 frames wide. Additionally the minimum segment was set to be 6 frames long. This is because Hadar et al. (1983) showed that the fastest head motion is below 7 Hz. Considering that the sampling period of the

motion capture system is 10 ms, would mean that 7 Hz is 14 frames and so 6 frames is below the Nyquest sampling criterion (Phillips et al., 2008, p. 240) and will thus be able to capture all the details of the head motion.

5.3 Choice of Templates

To determine how many templates should be used an unsupervised clustering approach was used. In brief the process in was to train an HMM on each segment then use cross-entropy distance for furthest neighbour clustering. This is based on the work by Smyth (1997).

For all clustering algorithms one needs some measure of the similarity or dissimilarity between samples. A popular choice is Euclidean distance. However, in this case it is not appropriate because the samples may not be the same length. Instead we chose to use cross entropy distance (see Section 3.5.1). As cross entropy distance requires a probabilistic model of each sample so we used left - to - right HMMs. Note that throughout this process the angles were treated separately and we are reporting the mean results of the three angles.

The shortest duration of a segment was 6 frames so it was not meaningful to use more than five states, as this would mean that not all the states would be occupied in all samples. As such 3, 4, and 5 state HMMs were tried but the amount of states did not have a significant impact on the results. This is not unexpected as the patterns of motion are fairly simple. As the warping will adjust the amplitude and duration, these are not needed during clustering and so all the segments were normalised so the amplitude was 1 radian and the duration was 1 second using linear warping. The normalisation took place before the clustering.

The next step was to estimate the cross-entropy distance between all pairs of segments. This is an $O(n^2)$ operation and it is not trivial amount of time to calculate the distances.

To make it computationally feasible only a subset of segments were used. We will call this subset of segments the initialisers and they are selected at random. Below we will discuss the clustering method but first we will show how we chose how many initialisers are needed.

We need the amount of initialisers chosen to be such that it produced similar clusters. This should be true no matter which initialisers are used provided they are randomly selected. In other words the number of initialisers must be high enough so that the clustering is consistent. To determine if there are enough initialisers for clustering we performed the clustering multiple times with different random selections of initialisers. We then see how similar the clustering is based on different initialisers. We then repeat this process for a increasing amounts of initialisers until the clustering does not increase in stability with more initialisers.

To determine how similar the clustering is when different initialisers are used we first complete the clustering with multiple sets of initialisers. Then we train a set of HMMs where each HMM is trained on all of the samples from one cluster. Between two clusters from two different sets of initialisers we calculate the mean cross entropy distance between all samples of both clusters using only the HMMs trained on the whole clusters. This gives us a measure of the distance between the two clusters. This process is then repeated between all the clusters of both sets of initialisers. We can then combine these distances with with the stable matching algorithm to determine cluster equivalence. The similarity of the clustering obtained by the two different sets of initialisers is then given by the total distance between only the matching clusters.

For each amount of segments selected for use as initialisers we generated 10 different sets of initialisers. We then calculated the similarity of the clustering at a selection of values for the different amount of initialisers. The results of this test are shown in Figure 5.4. For this experiment we used 5 clusters which was arbitrarily chosen. Based on this Figure 5.4 250 initialisers are more than sufficient to ensure the clustering is

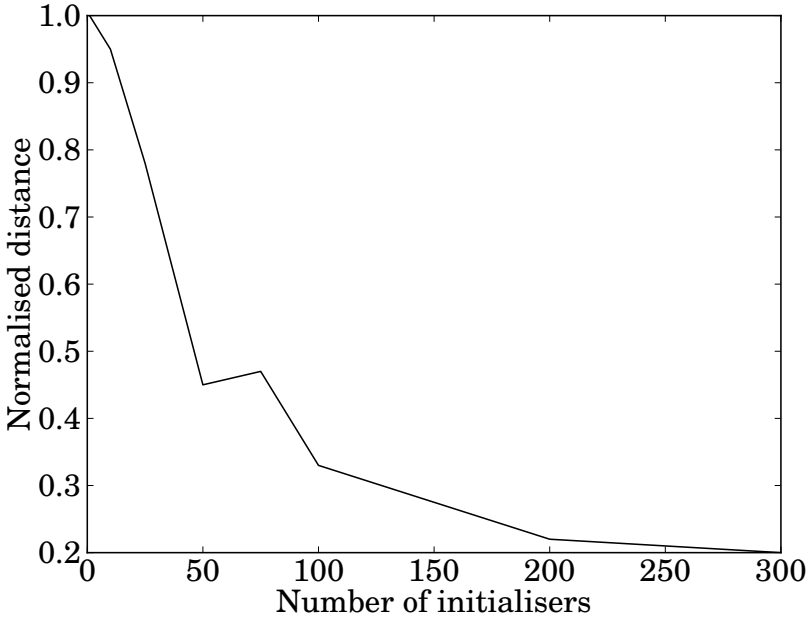


Figure 5.4: Total distances between equivalent clusters for different sets of segments (initialisers) as the amount of segments used for the clustering increases

independent of the selection of initialisers.

Once the cross-entropy distance between all the markers is calculated the next step is to group them using k-furthest neighbour clustering as recommended by Smyth (1997). This is a hierarchical clustering method. At each step the two closest clusters are combined until only k clusters remain. The distance between the clusters is defined as the furthest distance between any pair of points in each of the clusters hence the name (Gonzalez, 1985).

To determine the amount of clusters that would be used. At each iteration of the clustering algorithm the maximum intra-cluster and inter-cluster distance was calculated. This is shown in Figure 5.5. When the distance between clusters is smaller than the distance within samples of a cluster then too few clusters have been used. Thus Figure 5.5 shows that two or three clusters would be appropriate. We chose to use the higher number as it was likely that this cluster was catching outliers.

To find suitable templates first one HMM was trained for each cluster in the same man-

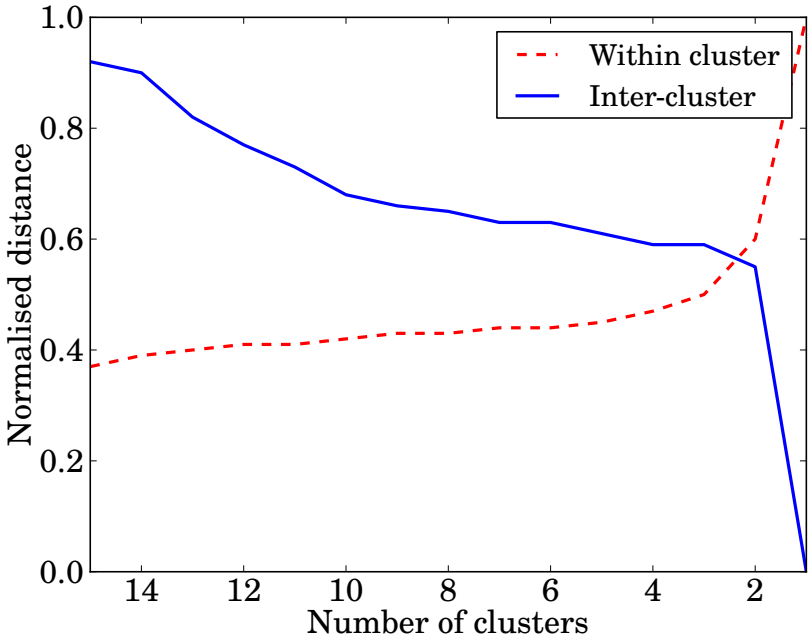


Figure 5.5: Maximum intra-cluster and inter-cluster distance for different number of clusters trajectory segments

ner we did for determining the amount of initialisers required. Then all the segments from the dataset were assigned to the cluster that gave the highest observation probability from the HMM. The segments were then plotted as 2D histograms which are shown in Figure 5.6. In this diagram the time and amplitude are on the horizontal and vertical axes respectively and darker areas indicates a higher frequency of a segments passing through that point in amplitude at that time. The space was divided into 150 bins in both axes.

Based on Figure 5.6 it is clear that all the angles can use the same templates. It is also apparent that only two templates are needed. It is possible to use a stored trajectory or alternatively a function as a template. If a banked trajectory is to be used then amplitude and time warping will reshape it to match the warping parameters. If a function is being used then it should be chosen so that it can easily produce a trajectory of the appropriate amplitude and duration. Based on Figure 5.6 the same two templates are defined for all angles, namely:

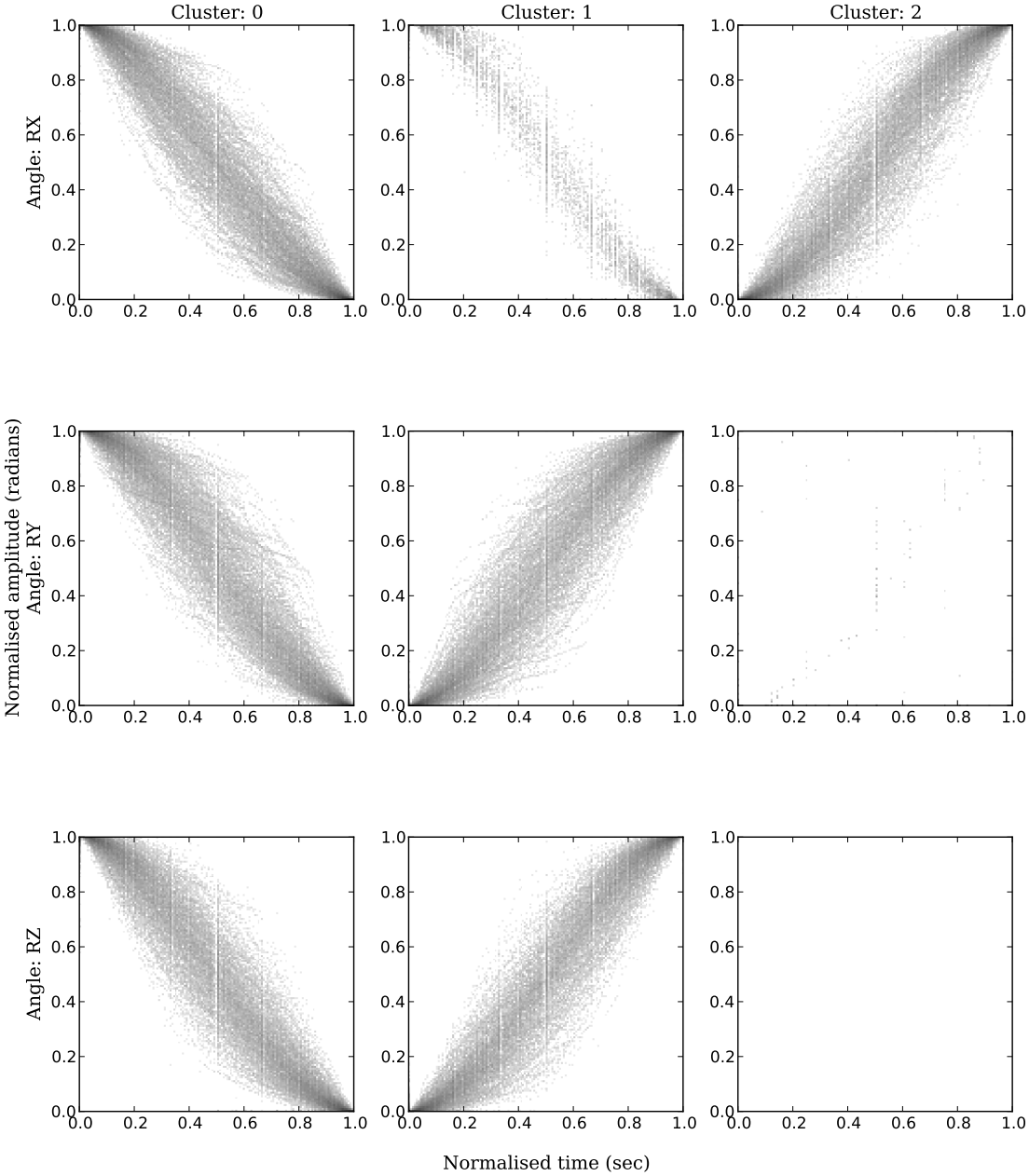


Figure 5.6: Segment trajectories assigned to different clusters, darker areas correspond to higher amounts of trajectories having a particular amplitude at the given time

Table 5.1: RMS Error for encoding with different templates, RMS amplitude of the original trajectory is given for comparison

Interpolation	RMS Error
Cosine	0.101
Piecewise linear	1.648
RMS Amplitude	0.205

$$\mathbf{g}_1^{(\kappa)}(t,\lambda_t^{(\kappa)}) = -\frac{a}{2}\cos\left(\frac{\pi t}{d}\right)+\frac{2c+a}{2}, \tag{5.1}$$

$$\mathbf{g}_2^{(\kappa)}(t,\lambda_t^{(\kappa)}) = \frac{a}{2}\cos\left(\frac{\pi t}{d}\right)+\frac{2c-a}{2} \tag{5.2}$$

By replacing the segments with motions generated from templates we are in effect encoding the trajectories. Thus a good objective test for these templates is to examine the error introduced by encoding with the chosen templates. In Table 5.1 we can see the RMS error when encoding with cosine templates, and as a comparison the RMS error for a piecewise linear template. Also given in the table is the RMS amplitude of the original motion trajectory to help interpret the results. From this table we can see the approximation error is low when using cosine templates and so they are an appropriate choice.

Thus far we have shown that template - warping can be used to used to parametrise head motion. As there are few fewer segments than frames this means that far less parameters need to be estimated from speech. Additionally by limiting the motion to the templates the head cannot make completely unreasonable motion. In the next section we will show the theoretical basis of synthesis using these warping parameters instead of a frame-wise based prediction.

5.4 Derivation of Template - Warping Based Synthesis

In this section we will show how template-warping based synthesis can be derived from the general optimisation problem of maximising the likelihood of the head motion given the speech features. The basic procedure is to split the head motion into segments (see Section 5.2), and use the warping parameters as surrogates for the motion within each segment. From there the warping parameters are optimised instead of the motion directly.

In general, synthesising head motion can be phrased as a likelihood optimisation problem which finds the best head motion $\mathbf{Y} = [Y_t]_{t=1}^{T}$ given speech features $\mathbf{X} = [X_t]_{t=1}^{T}$ where Y_t is a tuple of the head motion angles (φ, θ, ψ) of the head at time t, and X_t is the speech features vector. Note that this derivation will hold true for rotation vector components instead of angles, but we will use angles here purely because the symbols are more distinct. Formally the optimisation problem can be expressed as:

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} P(\mathbf{Y} \,|\, \mathbf{X}), \tag{5.3}$$

where $\mathbf{Y}^*$ is the estimated head motion. This is equivalent to

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \prod_{t=0}^{T} P(Y_t \,|\, Y_{t-1}, Y_{t-2}, \ldots, Y_0, \mathbf{X}). \tag{5.4}$$

When considering head motion as a series of segments let the time index of the start of segment τ be t_τ, where $\tau \in [0, T']$ and $t_{T'+1} = T$. This means that (5.5) becomes

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \prod_{\tau=0}^{T'} \prod_{t=t_\tau}^{t_{\tau+1}} P\Big(Y_t \,|\, Y_{t-1}, Y_{t-2}, \ldots, Y_0, \mathbf{X}\Big). \tag{5.5}$$

One could split this into optimising each angle separately:

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \prod_{\tau=0}^{T'} \prod_{t=t_\tau}^{t_{\tau+1}} P\Big((\varphi_t, \theta_t, \psi_t) \,|\, (\varphi_{t-1}, \theta_{t-1}, \psi_{t-1}), \ldots, (\varphi_0, \theta_0, \psi_0), \mathbf{X}\Big). \tag{5.6}$$

This can be re-expressed in terms of templates:

$$\mathbf{g}^{(K)}(t) = \left(\mathbf{g}_i^{(\varphi)}(t, \boldsymbol{\lambda}_t^{(\varphi)}),\ \mathbf{g}_j^{(\theta)}(t, \boldsymbol{\lambda}_t^{(\theta)}),\ \mathbf{g}_k^{(\psi)}(t, \boldsymbol{\lambda}_t^{(\psi)})\right), \tag{5.7}$$

noting that the template choice can be different for each angle.

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \prod_{\tau=0}^{T'} \prod_{t=t_\tau}^{t_{\tau+1}} P\left(\mathbf{g}^{(K)}(t) \,|\, \mathbf{g}^{(K)}(t-1), \ldots, \mathbf{g}^{(K)}(0), \mathbf{X}\right). \tag{5.8}$$

However, the movement within a segment is completely determined by the warping parameters within that segment. So (5.8) can be changed to reflect that only the warping parameters and the template choice need to be optimised and that they are constant for each segment.

$$\Lambda^{(K)}(t) = \left(i, \lambda_t^{(\varphi)},\ j, \lambda_t^{(\theta)},\ k, \lambda_t^{(\psi)}\right) \tag{5.9}$$

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \prod_{\tau=0}^{T'} P\left(\Lambda^{(K)}(t_\tau) \,|\, \Lambda^{(K)}(t_{\tau-1}), \ldots, \Lambda^{(K)}(0), \mathbf{X}\right). \tag{5.10}$$

We now take the Markov assumption and assume that each segment parameters only depend on a segment of the speech features $\mathbf{X}_\tau$. This allows for sequential optimisation:

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \prod_{\tau=0}^{T'} P\left(\Lambda^{(K)}(t_\tau) \,|\, \Lambda^{(K)}(t_{\tau-1}), \mathbf{X}_\tau\right). \tag{5.11}$$

However based on Bayes theorem we can state that this equivalent to

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \prod_{\tau=0}^{T'} \frac{P\left(\Lambda^{(K)}(t_\tau), \Lambda^{(K)}(t_{\tau-1}), \mathbf{X}_\tau\right)}{P\left(\Lambda^{(K)}(t_{\tau-1}), \mathbf{X}_\tau\right)}. \tag{5.12}$$

But if we are operating sequentially, $P\left(\Lambda^{(K)}(t_{\tau-1}), \mathbf{X}_\tau\right)$ is constant when optimising for segment τ, so this term can be ignored and (5.12) becomes

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \prod_{\tau=0}^{T'} P\left(\Lambda^{(K)}(t_\tau), \Lambda^{(K)}(t_{\tau-1}), \mathbf{X}_\tau\right). \tag{5.13}$$

Additionally there is a requirement that head motion is continuous so c is determined for all segments by the segments that have already been synthesised. This means that only the template choice, amplitude, and duration must be optimised. Also in general only one angle's parameters are changing so (5.13) becomes:

$$\mathbf{Y}^* = \arg\max_{\mathbf{Y}} \prod_{\tau=0}^{T'} P\left(i_\tau^{(\kappa)}, d_\tau^{(\kappa)}, a_\tau^{(\kappa)}, \Lambda^{(K)}(t_{\tau-1}), \mathbf{X}_\tau\right). \tag{5.14}$$

Where κ is the angle for which new warping parameters are needed.

This is the problem that will be solved in the remainder of this chapter.

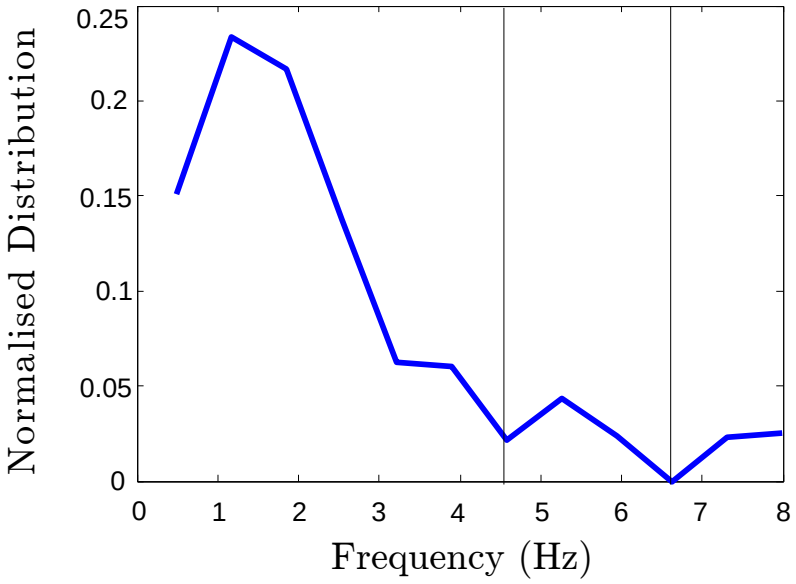


Figure 5.7: Distribution of the speed of head motion segments, lines indicate the boundary between fast, medium, and slow movement

5.5 Template Warping Based Synthesis System

We seek to estimate warping parameters from speech using a probabilistic model. A GMM is a reasonable model to use initially, especially because other researchers have successfully used a GMM based synthesiser before, and GMMs are easy to implement and train. We can also use some prior information. First we know that the two templates alternate, so we do not have to predict which template to use. Second, we found that there is some improvement when splitting the segments into clusters based on their speed.

This second piece of prior knowledge is based on the fact that both Hadar et al. (1983) and McClave (2000) found that head motion can be split into fast, medium, and slow movements. To compare to the Hadar and McClave we express the speed as frequency and show the distribution of the speed of head motion segments in Figure 5.7. The diagram indicates the boundaries between slow, medium, and fast motion which are in line with Hadar et al. (1983) and McClave (2000).

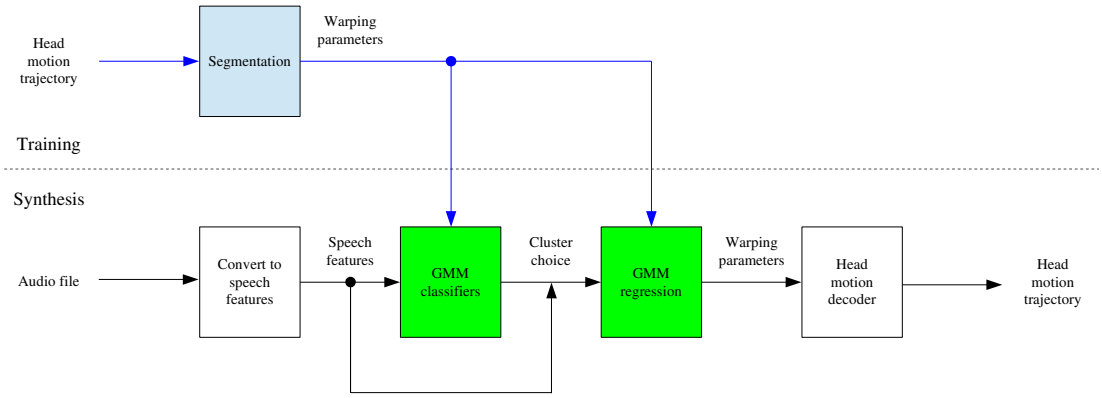


Figure 5.8: Synthesis and training process for GMM based template - warping head motion synthesis; blue lines indicate that those dependencies are used for training only

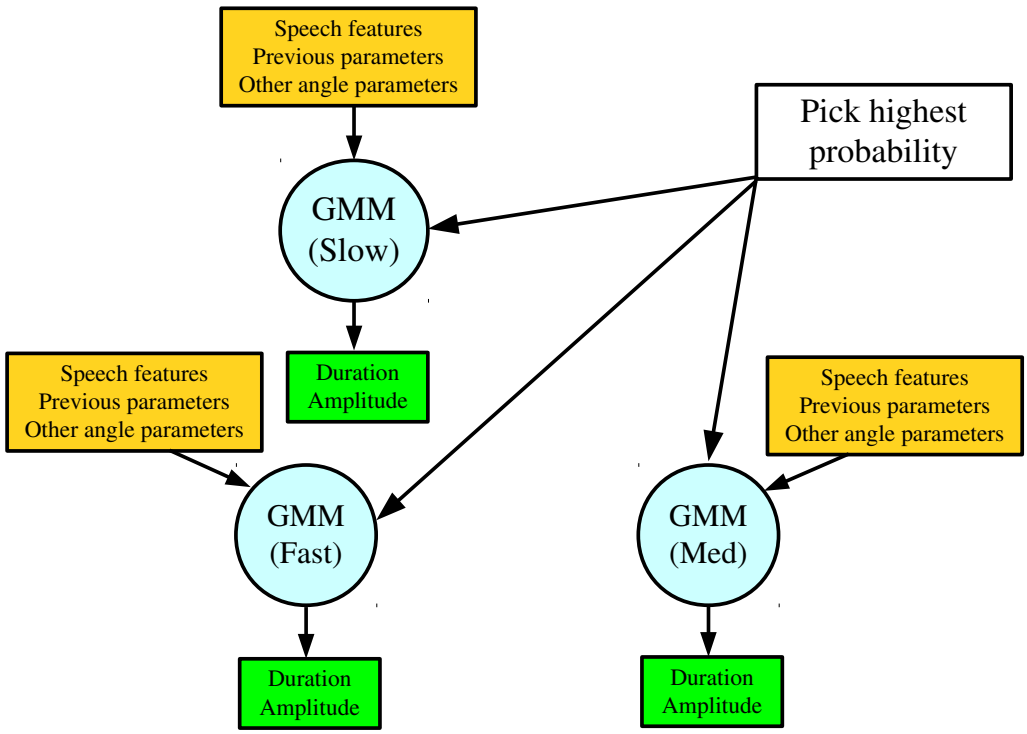


Figure 5.9: Template - warping parameter selection with a GMM based predictor for head motion synthesis

After splitting the segments into the speed categories the warping parameters are synthesised with a GMM trained on each segment. The training and synthesis procedure are illustrated in Figure 5.8. In detail the training procedure is as follows:

1. Segment the training data
2. Split the segments into different speed categories with the splits at 4.5 Hz and 6.5 Hz.
3. Create training vectors which are comprised of output warping parameters, input warping parameter, and speech features. Input warping parameters are the previous warping parameters for the current angle and the warping parameters for the other angles at the time of change. Speech features will be discussed in the evaluation section.
4. Fit a GMM to the training vectors from each speed class using expectation maximisation.

In the synthesis step we first use the input features to find the best speed category based on the observation probability of the GMMs from each category. Then we find the duration and amplitude based on the marginal distribution of the GMM which is given by equations (2.7) and (2.8). The warping parameters can be calculated by maximising the likelihood using the procedure in Toda et al. (2007, 2008) (see Section 2.2.1). One alternative, though not technically correct in terms of the original optimisation problem, is to minimise the mean square error of the prediction using equation (2.9). This which is computationally faster as it has a closed form solution. Another option is to generate a stochastic output by sampling from the marginal distribution. Mapping inputs to outputs using GMMs is called Gaussian mixture regression or simply Gaussian regression.

Diagrammatically the input output relationship of this system is shown in Figure 5.9 and Figure 5.8. The complete procedure for synthesis past the first segment is:

1. Determine which angle needs a new segment to be estimated
2. Create an input vector of input warping parameters as above and speech features
3. Calculate the observation probability of the three GMMs given the speech features and warping parameters from the other angles and the previous warping parameters.
4. Pick the GMM with the highest observation probability and calculate the marginal distribution of the amplitude and duration.
5. Calculate the best fit for amplitude and duration by maximising their likelihood in the marginal distribution.
6. Warp the appropriate template with the warping parameters and append onto the trajectory for the current angle.

Note one can generate a stochastic output with the following modifications:

- **Step 4**: Instead of picking the GMM with the highest likelihood, we can generate a multinomial distribution from the three observation probabilities and then sample from that distribution.
- **Step 5**: Rather than maximising the the marginal distribution one can sample from it to find the warping parameters.

One could either use both modifications or just one to obtain different levels of stochasticity.

5.5.1 Evaluation of Synthesis System

This section will focus on the evaluation of template - warping synthesis as it has been presented so far. First we will present some objective results. We will then show the results of a subjective evaluation where we compared template - warping synthesis to

motion capture and the system proposed by Le et al. (2012). We must first discuss some modifications that were needed so that the comparison to the Le et al. (2012) would be fair. Additionally before the evaluation can be performed we will need to pick some speech features. We chose to use the OpenSmile estimation of pitch and loudness (Eyben et al., 2010), the reasons for this choice are explained below.

As previously mentioned we will be comparing template - warping to the system proposed by Le et al. (2012). There are a few reasons for this. Firstly it is a state of the art system that has outperformed other methods, including ones based on HMMs which will be important when we show some improvements to template - warping synthesis.

Secondly the Le et al. (2012) system works in a similar manner to template -warping. By this we mean it optimises the next part of the trajectory based only on the previous part of the trajectory and speech features. A key difference is that our system takes into account the position of the other angles in addition to their dynamics. To give Le et al. (2012) a fair comparison this too should be taken into account. This can be achieved by optimising

$$(\varphi_t^*, \theta_t^*, \psi_t^*) = \underset{(\varphi,\theta,\psi)}{\arg\max}\, P(\varphi, \theta, \psi, \mathbf{v}, \mathbf{a}, p_t, l_t), \tag{5.15}$$

for all time (see Section2.3). Simultaneous optimisation of multiple variables is very computationally expensive and thus would not be suitable for the stated goal of the Le et al. (2012) system of having real time operation. However, once it is included it in fact can be seen as analogous to a frame-wise version of template - warping synthesis, and hence a fairer comparison.

Continuing in the theme of making a fair comparison between the proposed system and the Le et al. (2012) system, the same speech features were used in both. We followed their choice of the OpenSmile estimation of pitch and loudness (Eyben et al., 2010). It is an open question as to how many previous speech feature samples to use in addition to the current frame. We chose to only use the speech features at the segment boundaries so that the Le et al. (2012) system would not be penalised for not taking

Table 5.2: Confusion matrix for GMM based speed category recognition for all speakers, columns give the prediction and rows the true value.

	Slow	Medium	Fast	Number of samples in original
Slow	81%	8%	11%	30180
Medium	47%	46%	6%	3471
Fast	67%	8%	25%	1374

Table 5.3: Mean Forced CCA results for synthesis systems for GMM based Template - Warping Synthesis.

Condition	FCCA Score
Motion capture	0.33
Template - Warping Synthesis	0.23
Le et al. (2012)	0.21

into account the history of the speech features. In other words only one time sample of speech features were used for estimation for both systems.

The first criteria by which template - warping synthesis can be judged is how well it can identify the speed category. The results of this test is presented as a confusion matrix which is shown in Table 5.2. To generate the confusion matrix the system was given the correct segment boundaries and then it attempted to recognise the next segment's speed category.

A key observation from Table 5.2 is that the system does not always predict the same category, if it did it would mean that the the input features would not have any effect on the prediction. On the other hand it does not predict fast movements very well, it is better at identifying medium speeds but classifies segments as slow movements far too often.

Using the FCCA method we compared our system to the one proposed by Le et al.

(2012). The results are given in Table 5.3. We can see that Template - Warping Synthesis outperforms the Le et al. (2012) system slightly. This indicates that Template - Warping Synthesis should produce better subjective results.

For subjective evaluation we duplicated the methodology reported in the Le et al. (2012) paper. This was a simple A/B forced choice test (see Section 4.2.1). The types of animation used in this test were generated from the modified Le et al. (2012) system described above, the proposed GMM based Template - Warping Synthesis, and motion capture for use as a baseline. For all pairs both types of animation were shown an equal amount of times on the left and the right, in an order randomised for differently for each participant to remove effects from the ordering bias. The audio for both samples was the same. The participants saw each pair 10 times for a total of 30 comparisons. Each sample was between 10 and 15 seconds and cut to be at the end of a sentence. 20 participants took part.

The results from the A/B test are given in Table 5.4. It is clear that the template based system outperforms the frame-based system and is comparable to motion capture. Interestingly we obtained worse results for the modified Le et al. (2012) system than the original paper, only being preferred to the other systems once or twice per participant. The difference between our results and the original results could be explained by the fact that our experiment used longer samples for evaluation. Alternatively it could be because of the differences in the speech feature set, or the available training data.

It should be noted that when asked afterwards, participants did report difficulty in choosing between the majority of the samples. This is one of the issues of an A/B forced choice evaluation and affirms our recommendation for using a MOS or MUSHRA test as they show relative levels of preference and are not winner takes all. In this case participants stated that while the animation quality was very similar, occasionally some movements would happen that appeared to be ‘jerky’. This description only applied to animations from Le et al. (2012). In general, the animations produced by Basic

Table 5.4: Percentage of times system in the row was preferred to the system in the column in an A/B comparison. Each system was shown equal times left and right.

Condition	Motion capture	Template - Warping	Le et al.
Motion capture		**53**%	**99**%
Basic Template - Warping Synthesis	47%		**97**%
Le et al. (2012)	1%	3%	

Template - Warping Synthesis were considered to be more smooth. This is probably a result of optimising over longer stretches of time than the Le et al. (2012) system, and this slight difference gave it an edge when participants were being forced to choose between the two systems. Obviously motion capture would not appear to be jerky unless there were tracking errors.

Despite the good results in the subjective evaluation there are a number of areas where improvements can be made. Additionally there are some issues with the training method that can be addressed. How we solve these problems will be discussed in the next section, but we will identify them here.

The first problem is that the system requires prior knowledge on the characteristics of the motion which is intuitive to humans. The speed categories are the biggest culprit. In this case a metric was chosen that was understandable to humans i.e. the speed of the head motion. The same requirement for prior knowledge would be true if instead of grouping by speed we grouped by amplitude into small, medium, and large motions. Instead rather than relying on what can be determined by humans, it would be better to cluster the data in a way that is more appropriate for recognition which would be learned from data. In other words rather than splitting based on speed, the clusters should be chosen to minimise confusion and in a manner with minimal supervision. Though this has the disadvantage that the groupings may no longer be human understandable.

Secondly, while it was important for a fair comparison to the Le et al. (2012) system, the choice of speech features heavily influences the performance of the system. The speech features should be chosen more carefully as discussed in Section 2.2.4. Additionally the current method for recognising clusters with a GMM and one sample of speech features is very simplistic. HMMs have been shown to be better than GMMs for recognition tasks in general, and are widely used in automatic speech recognition in particular. So replacing GMM classifiers with HMMs for recognition seems reasonable.

The rest of the chapter will then focus on how these issues were addressed, and the performance of the resulting system when using the methodology we established in Chapter 4.

5.6 Improved Template Warping Based Modelling

In this section we will discuss how we improved upon the synthesis system already shown in Section 5.5. For clarity we will refer to the system proposed in Section 5.5 as Basic Template - Warping Synthesis, and the system presented in this section as Hierarchical Template - Warping Synthesis because here we propose an improved system which based on a hierarchical model with three layers.

In the hierarchical model the lowest level predicts the template warping parameters: duration and amplitude, the second level picks which cluster to use, and the third level biases the transitions between clusters. To relate it to an analogous HMM based speech recognition model, the middle layer would be the set of HMMs, and the higher level would be the language model. This particular combination of multiple layers with different purposes is part of the novelty of our system.

The middle and top layers of Hierarchical Template - Warping Synthesis are there to reduce the high levels of confusion that was found with Basic Template - Warping Syn-

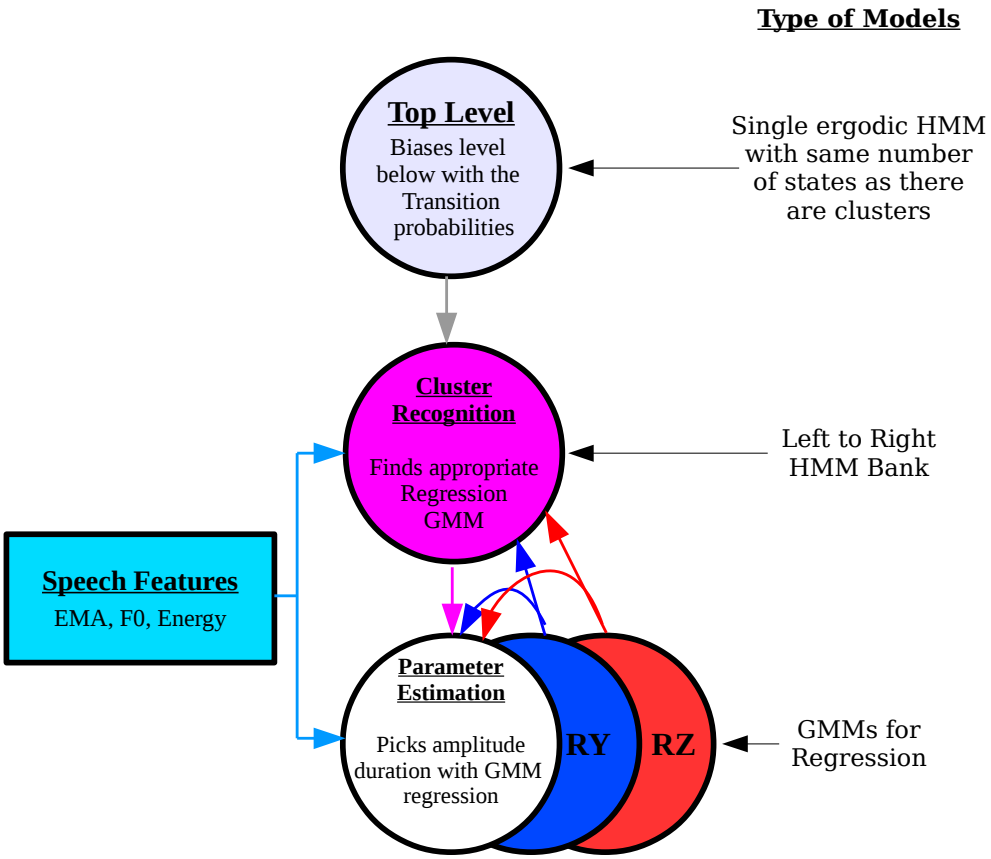


Figure 5.10: Hierarchical Template - Warping Synthesis parameter selection dependencies and purpose and types of model used in each layer for head motion synthesis

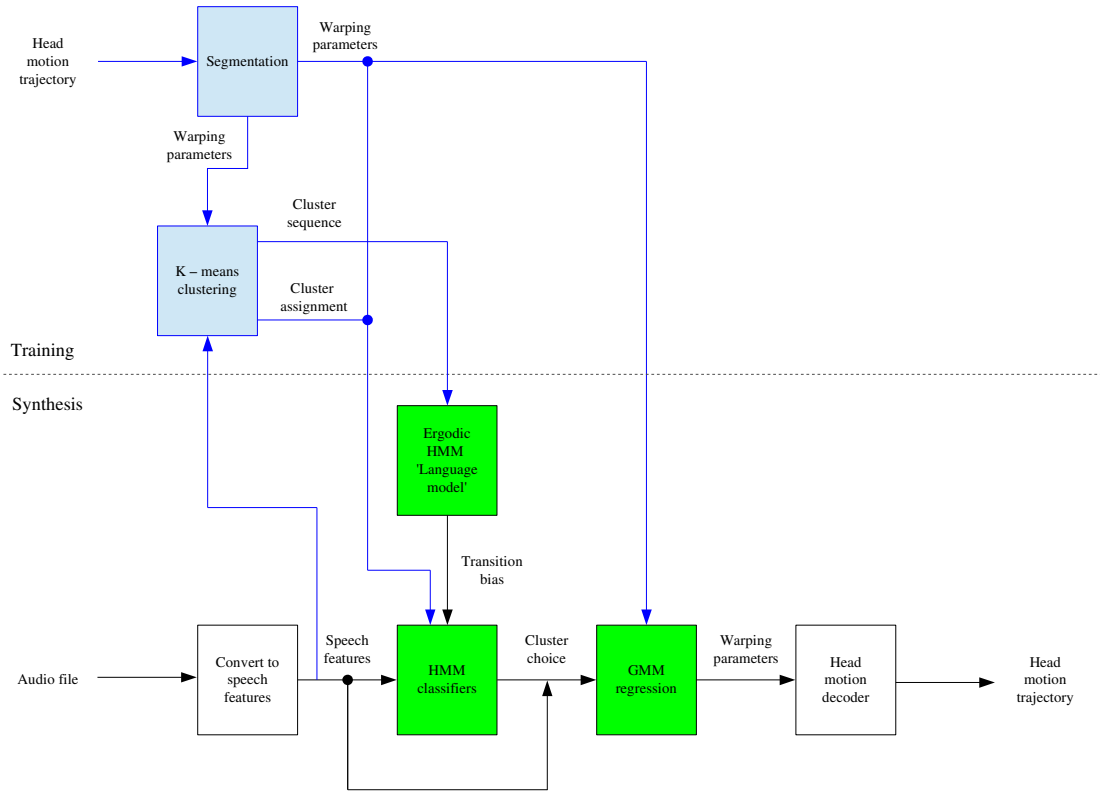


Figure 5.11: Synthesis and training process for GMM based template - warping head motion synthesis; blue lines indicate that those dependencies are used for training only

thesis. The lower level in Hierarchical Template - Warping Synthesis provides a direct link between the speech features and the output using Gaussian regression like in Basic Template - Warping Synthesis. All of this should provide a better model of the probabilities needed for solving the optimisation problem in (5.13). To help clarify the following explanation we have included Figure 5.10 which shows the dependencies of each layer, the purpose of that layer, and the type of model used in that layer of the hierarchy.

As mentioned the lowest level of the model is similar to the originally proposed Basic Template - Warping Synthesis in that GMM regression will predict the amplitude and duration. The difference here is that only one GMM will be picked which we will refer to the cluster GMM. Then the prediction from that regression used. While in Basic Template - Warping Synthesis GMMs were able to give fairly good predictions for the warping parameters, and hence can still be used at this stage. Additionally GMMs are fast to train and run quickly, so the fact that they are able to provide good predications means that it is suitable to keep using them. However, the confusion of picking the correct GMM was too high in Basic Template - Warping Synthesis. Hence we introduced the other two layers of the model.

With the second layer and top layer of Hierarchical Template - Warping Synthesis we wish to reduce the confusion of the GMM selection. One way is to use ASR based techniques in the middle layer to recognise which GMM we should be using. The other is to cluster the GMMs used for regression in an unsupervised manner that the middle layer would find easier to recognise. For this process we clustered using K - nearest neighbours, but instead of only taking duration into account we use the entire set of input and output that is being used for regression i.e. both the speech features and the warping parameters from all angles. This should result in a clustering which will have lower confusion. The disadvantage of including all of the features is is that the clusters will no longer have a meaning understandable to humans. In this application

this lack of a human meaning should not be an issue.

To further improve the cluster recognition the higher layer was also added. In essence this is an ergodic HMM that controls the likelihood of changing from one cluster to another. This has the effect of controlling the slightly longer term trajectory while the middle layer's HMMs control the short term motion. In the final structure each of the HMMs from the second layer is tied to a state, and the cluster selection is dependant on both the observation probabilities of the HMMs for each of the clusters and the transition probabilities from the top level HMM. An important difference between the two layers is the middle layer is a left - to - right HMM, as these is often used in speech recognition and synthesis, while the top layer is an ergodic HMM, so that each cluster can be selected after any other cluster. This would be analogous to an HMM based language model.

Rather than implement our own HMM training and synthesis systems, we used the library of tools called The HMM-based Speech Synthesis System (HTS) (Black et al., 2007) which is a modification of The Hidden Markov Model Toolkit (HTK) [1]. These systems are widely used in speech technology research and are available for free online.

The next three sections will cover some of the details of this modelling process. First we will discuss at a theoretical level which techniques we are employing, then the implementation details. With the implementation details we shall include some relevant objective measures that explain the choices we made for the structure of that part of the system. After that we will cover the synthesis and training process in greater detail which we have illustrated in Figure 5.11. We will then present the objective and subjective evaluation done according to the recommendations in the previous chapter. As even Basic Template - Warping Synthesis outperforms the system proposed by Le et al. (2012) and their system outperformed several HMM based systems, there was no

[1] `http://htk.eng.cam.ac.uk/`

need to compare back to other HMM based systems. This is provided that Hierarchical Template - Warping Synthesis outperforms Basic Template - Warping Synthesis.

5.6.1 Clustering

As mentioned previously using human understandable clusters of head motion, such as the fast, medium, and slow speeds from Basic Template - Warping Synthesis, is desirable, but may reduce cluster recognition accuracy. Instead a machine learning based clustering that takes into account all the input and output data should increase recognition accuracy. However, there are numerous problems associated with clustering. The three that must be addressed here are the clustering method, the number of clusters, and the distance measure. Seeing as there are many clustering methods it is best to pick one that is reasonable, and if it does not improve results then try another.

One of the easiest methods to implement and understand is k-means clustering. In this method the data is divided into a predetermined number of clusters, k, hence the name. The following explanation is adapted from (Izenman, 2013, pp. 423 – 424).

The general procedure is to first determine k 'centroids' in the data. A centroid is the mean vector of the data that was assigned to the cluster. Centroid initialisation is often done randomly. Here lies one of the problems of k means, the initialisation heavily affects the quality of the output. This is because k means tends to find local optima. The optimal clustering is the one that minimises the maximum distance from a centroid to any of the cluster's data points. There are many ways to mitigate this, for example initialising the centroids to divide the range evenly or trying multiple initialisations sometimes called Monte - Carlo sampling. This is the approach we used.

Once the centroids have been chosen each data point is assigned to the cluster that minimises the distance of the new point to the centroid. Euclidean distance is the most common distance measure, and was the one chosen for the seminal paper (MacQueen,

1967), but other methods are possible. As it is common practice we chose to use this measure. If there had been no improvement using clusters found using Euclidean distance then other distance measures could have been tried.

The next step is to re-estimate the centroid's position. This is done by finding the mean position in the cluster of all the assigned data, which is why it is called *k*-means. There are some modifications where the data points are weighted in the mean calculation. The assignment and update steps are repeated until convergence

5.6.2 Cluster Recognition

HMMs are a popular machine learning method used for speech recognition. The cluster recognition step of Hierarchical Template - Warping Synthesis is in essence a type of speech recognition. Thus it is probable that using HMMs would increase the accuracy of the cluster recognition compared to GMMs as they have done for ASR.

In speech recognition and syntheses it is very common to use left - to - right HMMs (see Section 2.2.2) because of they have a higher dependency on temporal ordering than ergodic HMMs. Though technically left - to - right HMMs are a special case of ergodic HMM, the increase in temporal ordering dependence comes from the initialisation, not the limits of the model.

In a left - to - right HMM transitions to previous states are not allowed. This means that the state sequence can be thought of as the mean trajectory through the data that the model was trained on. On the other hand ergodic HMM the state means and variances are not forced to follow the mean trajectory of the training observation sequences because any sequence of state is allowed. A simple example to explain the difference would be: Given the input is a single sinusoid, a left - to - right HMM would be forced to discriminate between a peak then a trough, and a trough then peak, while an ergodic HMM may not find a difference depending on the initialisation and number of states.

We gain a large advantage by switching to HMMs to recognise clusters. A GMM operates on a fixed number of variables, and does not take into account any previous observations, i.e. it is trying to recognise a cluster based on a fixed snapshot in time. In this application the GMM would have to use a fixed number of speech feature observations, in other words it would have to take a fixed window in time of speech features, for instance only the last 10 samples. An HMM on the other hand can take into account any amount of observations, so in this application we can recognise based on a changed number of samples relative to the length of the segment.

Because one can calculate the observation probability of an arbitrary length sequence they are a powerful tool when the sequence length is variable. In speech recognition it is difficult to find segment boundaries in the input data, and hence the need for language models. On the other hand in template - warping synthesis (or for that matter speech synthesis) we have the advantage of having the segmentation being generated at the same time as the trajectory. For template - warping synthesis if we treat the system as causal (output is not dependant on future values), either a fixed window length, or all of the speech features observed during the previous template can be used as the signal over which the observation probabilities are calculated. If it is not causal then the predetermined end point is lost, but the start should still be dependent on the previous segment.

At a simple level, in speech recognition, HMMs are used by building a bank of HMMs each trained on different speech segments, for example phones or word. Then at recognition time the HMM bank picks the model with the highest observation probability for a segment of audio. That audio is recognised as having the same meaning (phone or word) as the HMM that had the highest probability. We can take a similar approach for cluster recognition by training a different HMM for each cluster of warping parameters and speech features and then at synthesis time finding the HMM with the highest observation probability given the input speech features.

As mentioned before, the segmentation is very difficult in speech recognition. A language model helps determine which segments are likely to follow each other, which in turn limits the search space. In our system because we have reduced the search space already by only having a few clusters, a full language model is not necessary. Also as mentioned previously there is a natural segmentation, so that is not needed from a language model either. However, a more simple "language model" than is used in ASR should help to reduce the confusion as some models are more likely to follow others.

In our system we used a very simple language model analogue. An ergodic HMM was trained for each angle separately to create a bias in the cluster selection process. Because this in essence means that we have an HMM whose state observation probabilities are governed by another HMM we have created a hierarchical HMM (Fine et al., 1998). Additionally because the HMM for each angle is separate it is a parallel system.

This hierarchical and parallel system vastly reduces the training space. Because the angles are estimated separately there are far few conditions to train for and thus we avoid the curse of dimensionality. For example if there are only three clusters like fast, medium, and slow, and all the angles are estimated together, one would need 27 different models to represent every combination of clusters. On the other hand by treating the angles separately we only need 9 models. In fact for k clusters, a united system needs k^3 models for three models, while the structure we propose only needs $3k$.

5.7 Implementation Details

In the following sections we will discuss the implementation details that were chosen for the synthesis system. Additionally we will discuss what objective test were used to guide the decisions, and give the results where appropriate. Objective scores were

calculated by leaving one sample out of the training and then calculating the error on the left out sample. This was then repeated so each of the five samples from each speaker was left out once, a technique known as multiple cross-validation.

As a side note the systems in this research were predominantly implemented in Python 2.7.3. GMMs modelling was implemented with Sci-kit learn a Python module for machine learning (Pedregosa et al., 2011). The Hidden Markov Model Toolkit (HTK) is a library of software tools for HMMs, a modification of the library is The HMM-based Speech Synthesis System (HTS) (Black et al., 2007), this was used for the HMM based parts of the system. As before OpenSmile (Eyben et al., 2010) and Poser [2] were used for speech feature extraction, and animation respectively. Most of the statistical analysis was done using the Scientific Python module [3].

5.7.1 Speech Feature Selection

Previously we have discussed that there are many potential speech features to use for head motion synthesis. Initially when comparing to Le et al. (2012) it was most appropriate to use the same features as they specified in their paper. Now it would be better to chose the speech features that best match our system.

Considering the vast array of speech features available and that any combination could be used it is important to select only a few. In Section 2.2.4 we covered some of the options available. In this research we wanted to find the features with the highest correlation to head motion. To that end we considered the correlation of MFCCs, F0, Energy, LSPs, another form of LPC called Log Area Ratios (LAR), and predicted EMA values. We also included the first derivatives.

The author of this thesis collaborated on the research published by Ben Youssef et al.

[2] http://my.smithmicro.com/poser-3d-animation-software.html
[3] http://www.scipy.org/

The work on speech feature selection was previously published in (Ben Youssef et al., 2013a), the author assisted with the implementation, testing and writing of the paper.

(2013a) in this area. We found that EMA features predicted by HMM based speech inversion had higher correlation with head motion than standard speech features, such as MFCCs. This was confirmed using multiple speakers from the dataset published by Richmond et al. (2011) which had ground truth for the EMA features, head motion and audio. This makes sense as the head will move to reduce stress on the articulators (McClave, 2000). Thus we will use them as the speech features for our system.

5.7.2 Clustering

In our implementation of k-means clustering we chose to follow common practice and use Euclidean distance, and unweighted means. To form the feature vector we chose to use a few samples of the acoustic features prior to the change. Because we wanted to use Euclidean distance the number of acoustic points had to be fixed.

As has already been stated we clustered based on the warping parameters and the speech features, in this case predicted EMA features. For the speech features we used the 10 previous samples before the transition. As the dataset is sampled at 100 Hz and the speech features and the motion capture frames were synchronised this means that 100 ms of speech was being used to cluster, we we considered to be more than sufficient based on our experience with head motion synthesis.

To determine the number of clusters there are a number of approaches. We chose to use the approach presented by Hamerly and Elkan (2003) called G-means. This algorithm assumes that the data exists in Gaussian distributed clusters, and tests for how well it fits this assumption. This test is highly appropriate as we will be fitting GMMs to perform the regression using the data from each of the clusters.

The G-means test was run on all the segments from all the angles separately, and on all the speakers separately. The results were that either three or four clusters are appropriate depending on the angle and speaker. To simplify implementation we used

Table 5.5: Example cluster information for warping parameters using k means clustering and taking into account the speech features. Note that clusters 2 and 4 are similar despite objective measures showing that there should be four clusters in the data.

Warping Cluster	Mean Duration	Mean Amplitude
1	55.11	0.09
2	24.22	0.03
3	40.69	0.06
4	28.27	0.05

four clusters for all angles and all speakers. Apart from having a slightly more complex model than needed there is no real disadvantage to using only one more cluster.

In Table 5.5 we give an example of the warping parameter means for the segments assigned to the different clusters. This example is from one speaker, and one angle. This example was chosen specifically because the *G*-means test showed that there should be four clusters here. What is important to note is that while there is not much difference between Clusters 2 and 4 in terms of their warping parameters. However, the clustering showed that once you take into account the speech features these are distinct groups of segments, and thus the system will be able to predict the warping parameters more easily if they each have their own cluster.

5.7.3 Gaussian Regression

When training the GMMs we need to determine how many mixture components are needed. A very common method is by minimising the BIC as defined in Section 2.2.1. In Figure 5.12 we have plotted the BIC for different structures of the GMM, we have also included different covariance structures. This is the mean BIC from all the speakers and angles. From the figure there is not much difference between number of mixture components, though the covariance has a large impact. Nevertheless the BIC is

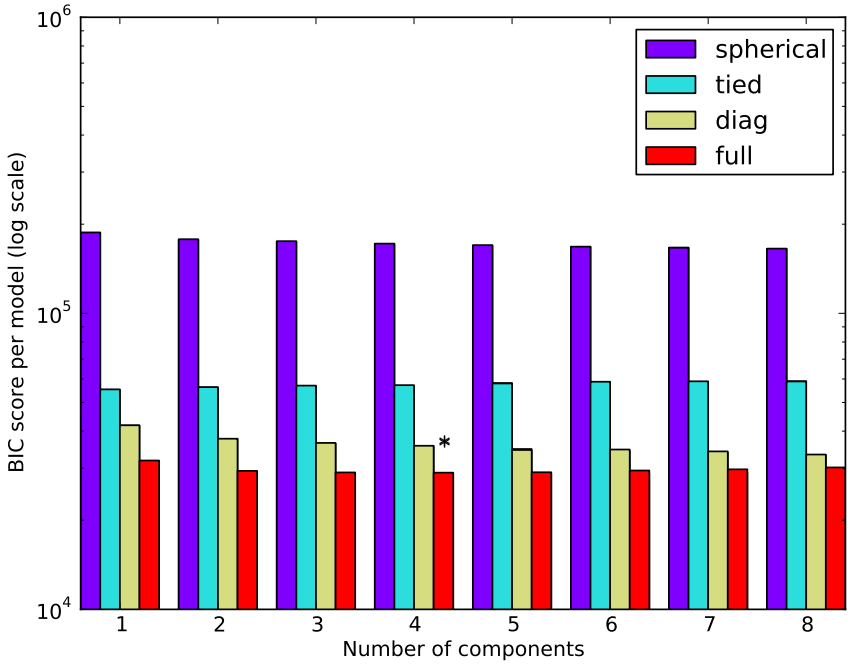


Figure 5.12: Bayesian Information Criteria for different GMM structures calculated on all available segments.

minimised at four components. This confirms our choice of four clusters above.

In this particular case we are training the GMMs in separate clusters and the number of clusters was specifically chosen by the *G*-means algorithm so that the data points in each cluster have a Gaussian distribution. Thus we will only use one mixture component in each of the regression GMMs.

For each cluster we wish to train a GMM. We chose a Monte-Carlo approach when training the GMMs to avoid local optima. In this approach the GMMs are initialised multiple times at random and the best GMM is used. In this case each GMM was initialised 100 times.

To check that the GMMs were in fact able to predict the warping parameters we calculated the RMS error of the predicted duration and amplitude given the correct cluster. To check that the speech features were in fact having an impact we calculated the error

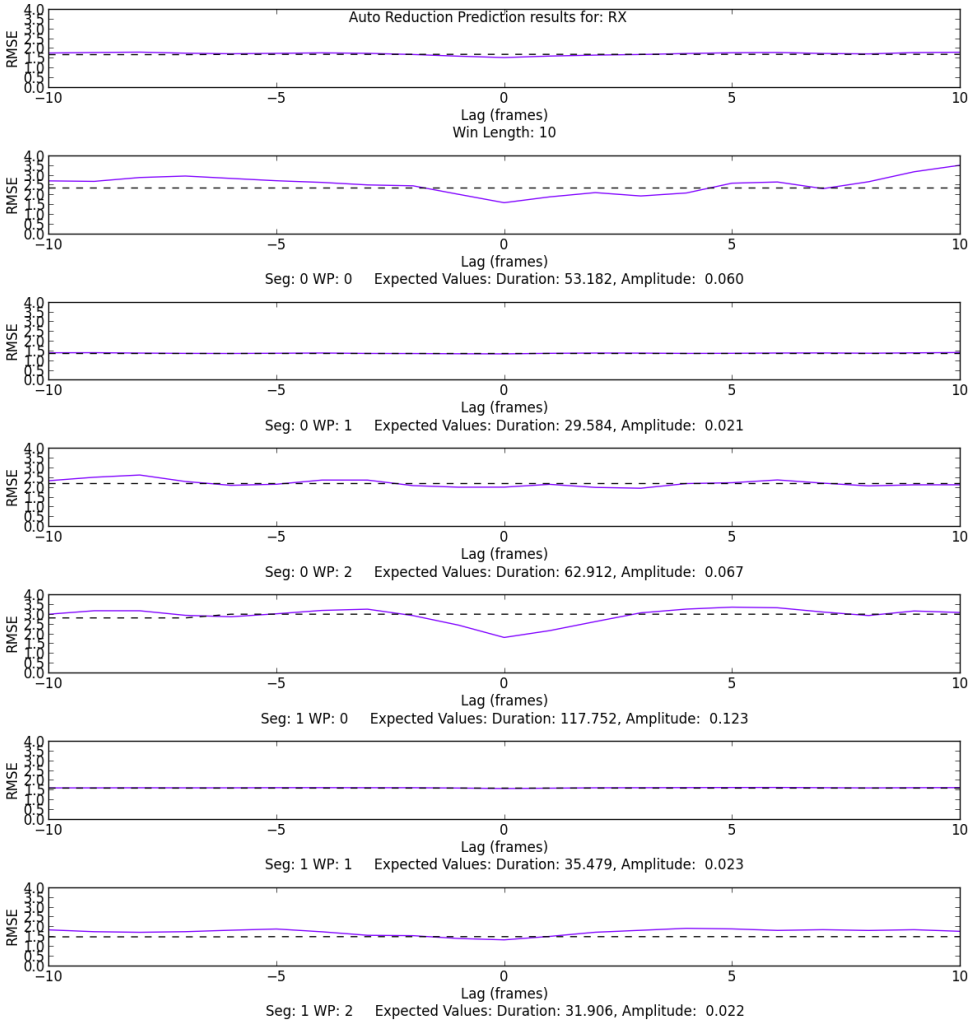


Figure 5.13: Prediction error for different offsets in time, *Seg* refers to whether the segment is rising (0) or falling (1), *WP* refers to which cluster is being used. The dashed line is the error calculated for inputs held at zero. Note that different subfigures refer to different clusters.

if the speech features were out of sync by a few samples. We also calculated the error for when the inputs were held at zero. Six sample clusters are shown in Figure 5.13. This figure was generated from a single speaker and angle, as it would be difficult to find compatible clusters between speakers and angles. Other speakers and angles do follow similar patterns though.

As we can see in Figure 5.13 the speech features do have an effect. However, some clusters are more predictable than others. We can see the predictability of each cluster because each of the subfigures show the results of this test for a different cluster.

There does not seem to be a pattern as to which clusters will have predictable warping parameters, based on the warping parameters themselves, nor when including the speech features. The mean of the distance between the segments and the centroids of the clusters, which is an approximation of how well the data fitted the clusters, also does not seem to correlate ($\rho < 0.1$ for all samples) to predictability. We measured predictability by calculating the peak decrease in RMS error. Despite this the GMM regression still provides no worse predictions than noise and performs better than noise most of the time.

5.7.4 Cluster Recognition HMMs

Like the regression GMMs there are several factors to consider when designing HMMs. These include the structure of the covariance matrix (full or diagonal), the number of states, and the structure of the transition matrix. As there are too many combinations to show the detailed results for, we are only going to show the evaluation results for the final structure. First though we will discuss how we decided upon the structure.

While it would be technically possible to train completely different structures for each angle, and each template, this makes it impractical to do this analysis. Each training takes several hours and considering the amount of features that must be fixed it was

decided that all templates and angles would have the same structure. As such the results were calculated as the mean of the results from all angles and templates. To make the clusters comparable between the templates, angles, and speakers we indexed the clusters so that Cluster 4 had the most segments in training and Cluster 1 had the least. Another method would be to find the most similar clusters. However, we thought that seeing as the purpose of these HMMs was to distinguish between the clusters, it would make more sense to index the clusters so that it would always be more important to predict Cluster 4 correctly as that should be the one that was observed the most often.

With regards to the number of states, we examined the total confusion from three to ten emitting states. Three states is the lowest meaningful amount to use, and well before ten states the results had stopped significantly changing. The results did improve with the inclusion of more states, however, beyond four states the confusion matrices had stabilised. To be safe we chose to use five states. For simplicity we kept the number of states in the middle layer and top layer the same.

As mentioned previously we need to use a left-to-right HMM to recognise the clusters in the middle layer, and an ergodic transition matrix in the top layer. As the top layer has a multi-modal distribution for observations it is not meaningful to discuss its covariance matrix, as it does not have one. On the other hand the lower layer has GMMs to describe the observation probabilities. It is standard practice in speech recognition to use diagonal covariance and so we tried both diagonal and full covariance. The difference in recognition accuracy was not large, but the full covariance significantly slowed the training and synthesis and so diagonal covariance was used. The speech features were the same as those used for the regression, but instead of using a fixed window of speech features each of the HMMs were trained in a manner more similar to speech recognition, each observation was one frame of speech features and their first and second derivatives, and the current warping parameters of the other angles. The observation vector length was over the entire previous segment.

The final confusion matrix is given in Table 5.6 which we can compare to Table 5.7. Table 5.7 is the confusion found by just using the observation probability of the regression GMMs alone, which in effect is Basic Template - Warping Synthesis. We see that there is an improvement in accurately predicting Clusters 3 and 4, which are the clusters with the most segments in the source data, and so are arguably the most important to correctly recognise. While the prediction accuracy for the other two clusters is lower compared to the GMM recogniser they are less frequent in the data. Also despite having higher confusion on Clusters 1 and 2, the HMM based classifier still predicts the correct cluster most of the time.

5.8 Synthesis Process

At synthesis time the procedure for Hierarchical Template - Warping Synthesis remains similar to Basic Template - Warping Synthesis before the cluster recognition improvements. We pick a cluster based on the observation probability of the HMM cluster recognisers, biased by the transition probabilities of the top level HMM. In other words we do a Viteribi decoding with the observation probabilities dictated by lower level HMMs.

Once the cluster and template are chosen the regression GMM can be used to find the warping parameters. As before we find the marginal distribution of the regression GMM given the speech features and warping parameters of the other angles at the point of change and then either calculate the warping parameters that maximise the marginal likelihood, which is theoretically correct, or sample from the resulting distribution, which results in more varied head motion that should still be synchronised with the speech features.

Once the warping parameters for the template are chosen the synthesiser does an amplitude and time warp on that template and appends it onto the trajectory of that angle.

Table 5.6: Warping parameter cluster confusion matrix for HMM based classifier. Clusters are arranged in increasing order of amount of segments in source data.

	0	1	2	3
1	**57.3**%	5.4%	8.3%	28.9%
2	10.5%	**51.5**%	10.7%	27.3%
3	4.9%	2.2%	**67.2**%	25.7%
4	7.0%	0.0%	0.0%	**93.0**%

Table 5.7: Warping parameter cluster confusion matrix for GMM based classifier. Clusters are arranged in increasing order of amount of segments in source data.

	1	2	3	4
1	**64.5**%	35.0%	0.6%	0.0%
2	9.5%	**89.7**%	0.8%	0.0%
3	2.7%	39.9%	**57.4**%	0.0%
4	7.0%	**56.3**%	1.4%	35.2%

The system then proceeds to find the next point of change in any angle and calculate the warping parameters for at that point of change, then does the amplitude and time warp again to the trajectory using those warping parameters and appending it to that angle's trajectory. This process is then repeated until the end of the speech sample.

In practice it is slightly easier to actually work to one sample before the point of change, as this does not affect the distribution significantly and there is no need to take into account the unlikely but not impossible circumstance of two angles changing simultaneously. At the start of the synthesis process we used the mean starting parameters from the dataset to initialise the trajectory. This system is causal like Basic Template - Warping Synthesis and so is still suitable for a live agent, though on a laptop it is slightly slower than real time to synthesise the trajectory and render the avatar using Poser. This means that currently a powerful desktop with a graphics accelerator would be needed. Obviously as computers advance in power this will no longer be a concern.

5.8.1 Example

The theoretical explanation above covers the process in an abstract, general way. To assist with understanding the synthesis process the following subsection will cover an example. Though Figure 5.10 and Figure 5.11 should also assist. We will assume that part of the trajectory already exists. This situation is illustrated in Figure 5.14, where the black circle indicates the position in time we need to estimate the warping parameters for the trajectory of the blue angle.

At this point the Viterbi decoding algorithm is used with the cluster recognition HMMs (middle layer) to determine the observation probabilities of each cluster. This will take into account the warping parameters of the trajectories of the other two angles at this time point, the previous warping parameters of the red angle, and the speech features, as indicated in Figure 5.15. The top layer then is invoked to bias the selection of which

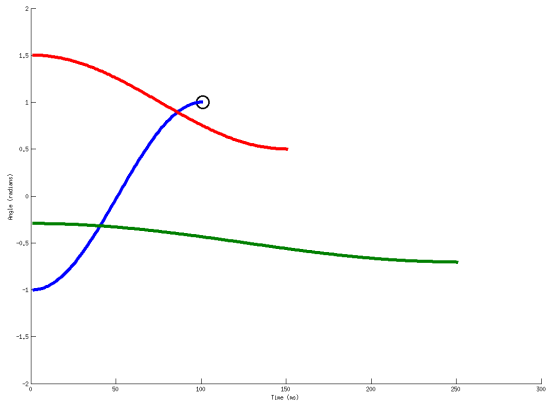


Figure 5.14: Initial trajectory, black circle indicates the point at which the next warping parameters are needed

cluster is to be used.

If we were using Basic Template - Warping Synthesis, the cluster with the highest probability from the regression GMM would be selected. The warping parameters are then estimated from this GMM.

Once the cluster is chosen, warping parameters are estimated from the regression GMM. This is done by finding the marginal pdf of the warping parameters with equations (2.7) and (2.8), then either maximising the likelihood or sampling from the marginal pdf. As the previous template of that angle was an upwards movement, the next template would be the downwards motion. The down template is warped and appended onto the trajectory. This results the trajectory shown in Figure 5.16.

Now the system will move onto the next point where warping parameters must be estimated, this is the circled point on the red angle's trajectory in Figure 5.17. This whole process is then repeated until the end of the speech feature trajectory.

Table 5.8: Mean Forced CCA results for synthesis systems.

Condition	FCCA Score
Motion capture	0.33
Hierarchical Template - Warping Synthesis	0.29
Basic Template - Warping Synthesis	0.23
Head motion shaped noise	0.09

Table 5.9: Mean Opinion Scores for different synthesis systems, rescaled for each subject to a 0 to 1 scale.

Condition	Score
Motion capture	0.57
Hierarchical Template - Warping Synthesis	0.53
Basic Template - Warping Synthesis	0.52
Head motion shaped noise	0.39

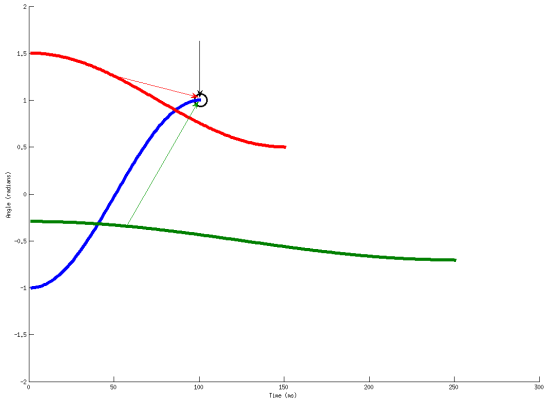


Figure 5.15: Trajectory with input information highlighted, warping parameters from red and green angles, and speech features in black

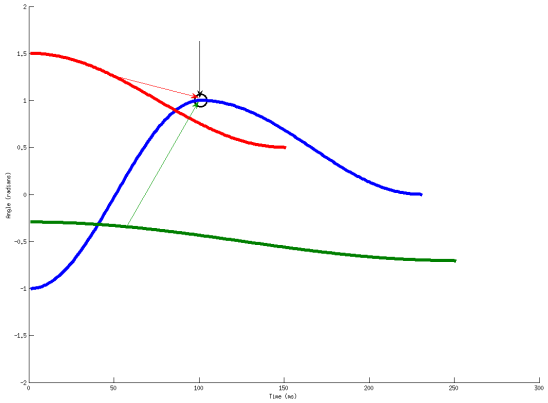


Figure 5.16: Trajectory after next template has been appended

5.9 Evaluation

To evaluate the method we first trained several speaker dependant models that each left out one of the trajectories for that speaker, then repeated for all speakers. Using these models we synthesis the trajectory that was omitted during training for that speaker. We then conducted an objective and a subjective test for both Basic Template - Warping Synthesis and Hierarchical Template - Warping Synthesis. Recalling that Basic Template - Warping Synthesis outperformed other state of the art systems, including ones based on HMMs we felt it was only important that Hierarchical Template - Warping

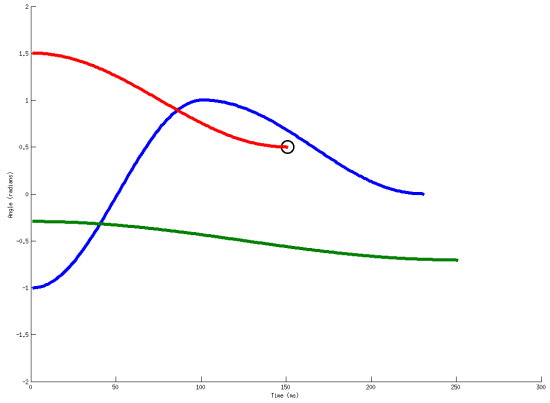


Figure 5.17: Next point at which warping parameters need to be estimated, black circle on red trajectory

Synthesis outperformed Basic Template - Warping Synthesis.

For objective evaluation we performed the Forced CCA method proposed in Section 4.3. The acoustic features used were the predicted EMA features as these were the ones used for training. As has already been stated they have higher correlation than other speech features. The mean results of the FCCA tests on each of the samples are given in Table 5.8. Which shows that Hierarchical Template - Warping Synthesis has a closer match to the original motion capture Basic Template - Warping Synthesis, and both are better than head motion shaped noise.

We then conducted the subjective evaluation in the same manner as recommend in Section 4.2. The types of motion used in this test were motion capture, Hierarchical Template - Warping Synthesis, Basic Template - Warping Synthesis, and head motion shaped noise. Two speakers were chosen at random for this evaluation, one male and one female. A total of ten samples were used for the training phase, four for each type of synthesised motion and six motion capture samples. Then 32 samples were evaluated, eight samples of each motion type. The order was randomised differently for each participant.

For this evaluation 24 volunteers were recruited who performed the test at home. As

our previous tests showed that once normalised the demographics we had tested against did not affect the results, they were not gathered. The scores are given in Table 5.9.

5.10 Analysis and Discussion

In the proposed system we have treated head motion synthesis as an optimisation problem. Under this basic assumption we have moved from trying to solve for the ideal trajectory at every point in time to a system that treats each angle individually, though not independently. The system predicts small segments of data rather than building the trajectory frame by frame. The advantage of dealing with each angle individually is that it reduces the degrees of freedom of the model in a way that does not impact on the final result.

Our rule based method for segmentation removes the need of manual annotation, while still restricting the motion in a reasonable manner. When encoding with the templates that were generated we have a very small error. Template selection was driven by data in terms of number, but in the end were hand crafted based on observation of the trajectories that make up the clusters in the data. By using these templates of motion we can reduce the number of parameters that need to be generated for a segment of the trajectory T samples long from $3T$ to 6 warping parameters that will describe the whole motion, noting that T is mostly much greater than 6.

Using the past warping parameters and acoustic features allowed us to build a synthesis system based on GMM regression. By splitting the segments into clusters we can more accurately predict the warping parameters. While we tried hand crafting the clusters based on the speed of the motion, we found better objective results based on clusters that were learned from data.

To improve cluster recognition we moved from a simple GMM based system which we called Basic Template - Warping Synthesis to a hierarchical HMM based model we

called Hierarchical Template - Warping Synthesis. Hierarchical Template - Warping Synthesis was better able to predict the clusters that each segment belonged to, and unlike Basic Template - Warping Synthesis, Hierarchical Template - Warping Synthesis always predicted the correct cluster more often than any other single cluster. In fact it was correct the majority of the time.

By adding the HMM to the system we have in effect added a language model like those used in speech recognition. The use of a head motion language model should help to improve system reliability compared to other state of the art systems. While an HMM is a very simple language model, there are not many types of head motion to chose from in this system. Thus the HMM should be adequate for head motion synthesis in this case.

Using our objective measure of the final synthesised trajectories rather the components of the systems, both Basic Template - Warping Synthesis and Hierarchical Template - Warping Synthesis showed promise. Hierarchical Template - Warping Synthesis did outperform Basic Template - Warping Synthesis as expected from the improvements in cluster recognition. Furthermore, both Basic Template - Warping Synthesis and Hierarchical Template - Warping Synthesis were able to outperform head motion shaped noise and actually showed dependence on the speech features. This test is often overlooked, but it provides confidence that the synthesised trajectory will be synchronised with speech.

Ultimately the most important test is a subjective one. We carried out the test as per the guidelines from the last chapter. Part of the subjective testing was to compare to another state of the art system. Because we wished to limit the number of comparisons being done at once we only compared the state of the art system with to Basic Template - Warping Synthesis. Basic Template - Warping Synthesis outperformed the other system and had similar performance to motion capture in a direct A/B forced choice test.

Once it was established that Basic Template - Warping Synthesis would outperform a system that itself had outperformed other HMM based systems it was no longer needed to compare Hierarchical Template - Warping Synthesis with the other system. This reduced the number of conditions we were testing against, bringing the length of test down to a reasonable level.

In this final evaluation it was shown that despite the fact that Hierarchical Template - Warping Synthesis outperformed Basic Template - Warping Synthesis in objective tests, the results of a subjective test show no real difference in preference. It is important to note two things. The first is that Basic Template - Warping Synthesis is already comparable to motion capture, and hence there is not much room for improvement. Secondly the fact that Hierarchical Template - Warping Synthesis system performs better in objective measures indicates that it would probably be more consistent in obtaining good results with other speakers.

It might also be the case that the limits of the presentation may be obscuring the differences in head motion quality. If participants are too distracted by the black box, or the quality of the rendering, they may not be able to tell the difference in the naturalness of the head motion. However, as was discussed, it would be impossible to add in lip synchronisation and / or eye movement without loosing confidence that any improvements in a subjective evaluation would be because of changes to the head motion. Though this does imply that Hierarchical Template - Warping Synthesis may perform better as part of a system with eyes and lips than Basic Template - Warping Synthesis, as the objective results show that it should be closer to the original head motion.

As a final note, we suspect that the improvements our synthesisers show is due to the fact that we have included speech features at every step of the way. Whereas other researchers first work with head motion trajectories and then add the speech features afterwards.

Chapter 6

Discussion and Conclusion

6.1 Overall Achievements

In this thesis we have presented three contributions to the field of head motion synthesis. We have explained how we gathered a large corpus of data, and performed some analysis on this dataset. We have shown the basis of our recommendations for both subjective and objective testing methodology. We have also described and tested a new system for head motion synthesis.

In the following sections we will present some conclusions that we have drawn based on the findings in the rest of thesis. We will also speculate as to what future work could be done based on the research we have conducted and then presented in the thesis.

6.2 Data Collection

We have collected one of the largest publicly available corpora of motion capture of the head and upper body, synchronised with speech. In this thesis we explained the collection methodology. We also gave the demographic information of the participants

and the basic statistics.

What we found when examining the data is that read speech has greater variation than free speech. We also found that there is a dependence on the task in the motion. In other words samples of free speech are more similar to other samples of free speech than to samples of read speech, and the same holds true for read speech samples compared to free speech samples. However, this was overshadowed by the speaker dependence. By this we mean that samples were more different between speakers than between tasks. Though when considering samples from only one speaker the task dependence was still observed. This supports our assertion that head motion is task dependent.

Interestingly the difference between the speakers does not seem to correlate much with speaking rate. Though there is some dependence, it is not very significant.

6.3 Evaluation Methodology

There are two types of evaluation methodology that were examined in this research: subjective and objective. The more interesting is subjective evaluation as this is the metric that reflects how end users will react to the synthesised head motion in practice. On the other hand objective testing is still important as it is far more cost effective to run objective testing when determining the effect of changes to synthesis systems.

In this thesis we provided our motivation for using a modified version of a MUSHRA test in place of an A/B comparison. This is primarily to address the issues of determining exactly how good the relative quality is, rather than just a binary ‘better’ or ‘worse’ decision. The need for increasing the detail in the results was particularly emphasised when comparing our new model to the state of the art. While participants informally reported that the quality was only a little better, the test results implied that the improvement using the new system was vast. By combining with methodology from MOS testing, we were able to develop a version of MUSHRA testing that can

be used for evaluating head motion synthesis. This MUSHRA test gives a much better indication of the relative performance of multiple systems. In this case the stimuli are animations of the different types of head motion. For the adapted version of the MUSHRA test only one stimulus is shown at a time along with a reference, but among the stimuli are reference and anchor samples.

When we examined subjective testing we looked at a few factors that could bias the results. The first was the type of rendering used for animation. Our experiments showed that people were able to distinguish between the original motion capture and head motion that was completely desynchronised from the speech. Participants gave more consistent results when using realistic models than when using either untextured but realistically shaped 3D models or silhouettes. We also found during MOS tests that the 10th through 40th samples were the ones that gave the most reliable difference and that a reference sample should be provided. Finally we found that participants did not need to be native speakers and that the evaluation could be carried out online.

With regards to objective measures, we found that the current practice of using Canonical Correlation Analysis (CCA) is not suitable for evaluating synthesised head motion. Instead we proposed a modification to the CCA process that gives objective results that are lower when not evaluating the original motion capture and gave the lowest results when evaluating noise. We called this analysis technique Forced CCA (FCCA). Though at this point FCCA does not predict the results of a subjective test it does at least give some indication if the head motion has a similar dependence on the speech features as the original motion capture.

6.4 Template - Warping Synthesis System

With regards to our proposed synthesis systems there are a few aspects to discuss. First we split the head motion trajectory into three parallel models, one for each angle. For

each angle we could automatically segment the trajectory using a simple rule. We then with clustering in a semi-supervised manner were able to reduce the segments into a few types motion which we called templates. We were able to approximate these templates with a sinusoidal function. With these motion templates we could amplitude and time warp to reconstruct the original trajectory with a low encoding error. We called the amplitude and duration the warping parameters.

We then proposed a system based on Gaussian Mixture Models (GMM)s that could predict the amplitude and time for each template based on speech features. This involved using GMMs to determine a cluster of warping parameters and then using GMM base regression to predict the warping parameters. This system outperformed other state of the art systems in a subjective test. We named this approach Basic Template - Warping Synthesis.

We then sought to improve this method by including a hierarchical Hidden Markov Model (HMM) based method. Firstly instead of hand crafting the clusters of warping parameters we used a machine learning based clustering technique that not only considered the warping parameters but also the acoustic features too. Then we used a bank of Left - to - Right HMMs which were weighted by an ergodic HMM to recognise the cluster, while still using GMM based regression to pick the warping parameters.

By including the higher level HMM into the system we have in effect created a language model. In speech recognition a language model helps determine the likelihood of moving from one model to another, and the higher level HMM serves that same function. In this case, instead of recognising phonemes or words, we are recognising the appropriate type of head motion. However, the principle of biasing model selection remains the same. Additionally we created the 'language' for the HMM by grouping head motion types in an unsupervised manner. This should improve performance over hand chosen groups which use only one feature to split the clusters, as the clusters are learned using the full feature set that the HMM will be using for recognition. Due to

the fact the system now had a hierarchy of models we named it Hierarchical Template - Warping Synthesis.

The use of the HMM improved the objective measures of the synthesis, but it does not improve the subjective results. However, as it does not degrade the results, it should still probably be included. This is because the objective tests imply that it is more reliable. On the other hand the HMM based system is slower, so it may be that in live systems the system that only utilised GMMS would be the better choice.

6.5 Future Work

As there are three parts to this research there are three directions that the research could take from here. It would be interesting to capture several hours of data from one speaker rather than the approximately 20 minutes we have from many speakers. It would also be important to capture more types of speech, perhaps a better way to solicit free speech could be investigated. The results of the statistical analysis would be interesting to compare to dialogue data to see how similar they are.

In terms of the evaluation method, more work could be conducted on the subjective testing. Firstly a true MUSHRA test could be performed on a very large screen which would necessitate the use of a laboratory. The results from the MUSHRA test can be compared to the MOS testing we used. Secondly one could investigate how long the samples used for the evaluations should be.

With regards to the objective measure the next goal would be to provide a better predictor of subjective evaluation results, rather than a measure that is only capable of telling if the synthesised motion is as good as motion capture. Two possible approaches would either be modifying FCCA further or investigating other statistical measures that can be utilised with or adapted to multivariate temporal data.

With regards to the synthesis system there are a few areas that would be of interest. The first is trying to improve the subjective results to be exactly 50 % preference compared to motion capture in an A/B test, or alternatively to have exactly the same score on a MOS or MUSHRA test. This probably would mean that the probabilistic model would need to be improved or changed. For instance deep belief networks are showing promise in other speech technology research fields and might be suitable to replace the GMM / HMM based system we used in this research.

The second avenue of improvement for the synthesis system would be to try and determine the templates automatically, instead of just using cosine interpolation. This may involve using more parameters than just the duration and amplitude. If through a different method of clustering there are more than two templates one could pick the template as well as the cluster with a probabilistic model. This is as opposed to the current system which considers template choices to be equivalent to a flip - flop or switch.

Finally the synthesis system could be tested in other domains. The immediately obvious application would be facial animation, but perhaps it would be useful to approach completely different areas such as financial modelling with the same basic approach. By the basic approach we mean first segment using a simple rule which is based based on the data, then cluster the segments to find a small number of templates. Next create a probabilistic model to predict the warping parameters of the templates, possibly clustering will help improve the prediction accuracy. Throughout this procedure treat all variables in the data stream separately but with cross dependencies in the probabilistic model.

6.6 Concluding Remarks

This thesis presented our research in the field of head motion synthesis. We showed a novel approach to synthesis where we split the trajectory into parallel streams, segmented with a simple rule, created templates of motion and then warped them to synthesis new motion. We also can give recommendations as to how to conduct a subjective test: Use approximately 30 to 40 pairs of video, ignoring the first 10, and animate with a realistic avatar. Additionally we can state that CCA is not a suitable measure for head motion synthesis but FCCA provides a better measure of the quality of the output. Finally we explained how we collected a large dataset of head motion synchronised with speech, based on this we made some observations about the nature of the dependencies in the data. Namely head motion is very speaker dependant, and to a lesser extent task dependent. Also the difference between speakers does not seem to correlate strongly to speaking rate.

Overall by covering these areas we have contributed to the field in a structured and experimentally supported manner. In doing so we not only met the original objective of finding a new method for mapping speech to head motion, but also improved the supporting areas.

Appendices

Appendix A

Cross Entropy Distance Tables

Table A.1: Mean cross entropy distance of samples between speakers

	Ally	Carla	Irene	Jane	Nadine	Natalie	Nicole	Rebecca	Robin	Desmond	Gary	Mark	Marvin	Ray	Sam
Ally	0.72	0.83	0.83	0.83	0.82	0.82	0.84	0.83	0.83	0.83	0.81	0.82	0.83	0.81	0.83
Carla	0.83	0.74	0.86	0.87	0.86	0.86	0.87	0.86	0.87	0.86	0.85	0.85	0.86	0.84	0.86
Irene	0.83	0.86	0.76	0.86	0.85	0.85	0.87	0.86	0.86	0.85	0.84	0.85	0.86	0.83	0.86
Jane	0.83	0.87	0.86	0.78	0.86	0.86	0.87	0.86	0.87	0.86	0.85	0.86	0.87	0.84	0.86
Nadine	0.82	0.86	0.85	0.86	0.76	0.85	0.86	0.85	0.86	0.85	0.84	0.85	0.86	0.83	0.85
Natalie	0.82	0.86	0.85	0.86	0.85	0.77	0.86	0.85	0.86	0.85	0.84	0.85	0.86	0.83	0.85
Nicole	0.84	0.87	0.87	0.87	0.86	0.86	0.77	0.87	0.87	0.87	0.85	0.86	0.87	0.84	0.87
Rebecca	0.83	0.86	0.86	0.86	0.85	0.85	0.87	0.73	0.86	0.85	0.84	0.85	0.86	0.83	0.86
Robin	0.83	0.87	0.86	0.87	0.86	0.86	0.87	0.86	0.76	0.86	0.85	0.85	0.86	0.84	0.86
Desmond	0.83	0.86	0.85	0.86	0.85	0.85	0.87	0.85	0.86	0.77	0.84	0.85	0.86	0.83	0.85
Gary	0.81	0.85	0.84	0.85	0.84	0.84	0.85	0.84	0.85	0.84	0.75	0.83	0.84	0.82	0.84
Mark	0.82	0.85	0.85	0.86	0.85	0.85	0.86	0.85	0.85	0.85	0.83	0.74	0.85	0.83	0.85
Marvin	0.83	0.86	0.86	0.87	0.86	0.86	0.87	0.86	0.86	0.86	0.84	0.85	0.78	0.84	0.86
Ray	0.81	0.84	0.83	0.84	0.83	0.83	0.84	0.83	0.84	0.83	0.82	0.83	0.84	0.72	0.83
Sam	0.83	0.86	0.86	0.86	0.85	0.85	0.87	0.86	0.86	0.85	0.84	0.85	0.86	0.83	0.75

Table A.2: Mean cross entropy distance of samples between speakers only considering free speech samples

	Ally	Carla	Irene	Jane	Nadine	Natalie	Nicole	Rebecca	Robin	Desmond	Gary	Mark	Marvin	Ray	Sam
Ally	0.61	0.81	0.81	0.81	0.80	0.80	0.82	0.80	0.81	0.81	0.78	0.80	0.81	0.77	0.80
Carla	0.81	0.65	0.87	0.87	0.86	0.85	0.88	0.86	0.87	0.86	0.83	0.85	0.87	0.82	0.86
Irene	0.81	0.87	0.69	0.86	0.86	0.85	0.87	0.86	0.86	0.86	0.83	0.85	0.87	0.82	0.86
Jane	0.81	0.87	0.86	0.69	0.86	0.85	0.88	0.86	0.87	0.86	0.83	0.85	0.87	0.82	0.86
Nadine	0.80	0.86	0.86	0.86	0.68	0.84	0.87	0.85	0.86	0.85	0.83	0.84	0.86	0.81	0.85
Natalie	0.80	0.85	0.85	0.85	0.84	0.67	0.86	0.84	0.85	0.85	0.82	0.83	0.85	0.80	0.84
Nicole	0.82	0.88	0.87	0.88	0.87	0.86	0.66	0.87	0.88	0.87	0.84	0.86	0.88	0.82	0.87
Rebecca	0.80	0.86	0.86	0.86	0.85	0.84	0.87	0.64	0.86	0.85	0.82	0.84	0.86	0.81	0.85
Robin	0.81	0.87	0.86	0.87	0.86	0.85	0.88	0.86	0.65	0.86	0.83	0.85	0.87	0.81	0.86
Desmond	0.81	0.86	0.86	0.86	0.85	0.85	0.87	0.85	0.86	0.68	0.83	0.84	0.86	0.81	0.85
Gary	0.78	0.83	0.83	0.83	0.83	0.82	0.84	0.82	0.83	0.83	0.64	0.82	0.83	0.79	0.83
Mark	0.80	0.85	0.85	0.85	0.84	0.83	0.86	0.84	0.85	0.84	0.82	0.62	0.85	0.80	0.84
Marvin	0.81	0.87	0.87	0.87	0.86	0.85	0.88	0.86	0.87	0.86	0.83	0.85	0.70	0.82	0.86
Ray	0.77	0.82	0.82	0.82	0.81	0.80	0.82	0.81	0.81	0.81	0.79	0.80	0.82	0.58	0.81
Sam	0.80	0.86	0.86	0.86	0.85	0.84	0.87	0.85	0.86	0.85	0.83	0.84	0.86	0.81	0.64

Table A.3: Mean cross entropy distance of samples between speakers only considering read speech samples

	Ally	Carla	Irene	Jane	Nadine	Natalie	Nicole	Rebecca	Robin	Desmond	Gary	Mark	Marvin	Ray	Sam
Ally	0.67	0.85	0.84	0.86	0.84	0.85	0.85	0.85	0.85	0.85	0.85	0.84	0.85	0.84	0.85
Carla	0.85	0.58	0.86	0.87	0.86	0.86	0.87	0.86	0.86	0.86	0.86	0.86	0.86	0.85	0.86
Irene	0.84	0.86	0.63	0.86	0.85	0.86	0.86	0.86	0.86	0.85	0.85	0.85	0.85	0.84	0.85
Jane	0.86	0.87	0.86	0.70	0.86	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.86	0.86	0.87
Nadine	0.84	0.86	0.85	0.86	0.64	0.86	0.86	0.86	0.86	0.85	0.85	0.85	0.85	0.85	0.86
Natalie	0.85	0.86	0.86	0.87	0.86	0.69	0.87	0.86	0.87	0.86	0.86	0.86	0.86	0.85	0.86
Nicole	0.85	0.87	0.86	0.87	0.86	0.87	0.65	0.87	0.87	0.86	0.86	0.86	0.86	0.86	0.87
Rebecca	0.85	0.86	0.86	0.87	0.86	0.86	0.87	0.58	0.86	0.86	0.86	0.86	0.86	0.85	0.86
Robin	0.85	0.86	0.86	0.87	0.86	0.87	0.87	0.86	0.65	0.86	0.86	0.86	0.86	0.85	0.86
Desmond	0.85	0.86	0.85	0.86	0.85	0.86	0.86	0.86	0.86	0.68	0.85	0.85	0.85	0.85	0.86
Gary	0.85	0.86	0.85	0.86	0.85	0.86	0.86	0.86	0.86	0.85	0.68	0.85	0.86	0.85	0.86
Mark	0.84	0.86	0.85	0.86	0.85	0.86	0.86	0.86	0.86	0.85	0.85	0.64	0.85	0.85	0.86
Marvin	0.85	0.86	0.85	0.86	0.85	0.86	0.86	0.86	0.86	0.85	0.86	0.85	0.69	0.85	0.86
Ray	0.84	0.85	0.84	0.86	0.85	0.85	0.86	0.85	0.85	0.85	0.85	0.85	0.85	0.67	0.85
Sam	0.85	0.86	0.85	0.87	0.86	0.86	0.87	0.86	0.86	0.86	0.86	0.86	0.86	0.85	0.65

Bibliography

Alpert, M. I. and Peterson, R. A. (1972). On the Interpretation of Canonical Analysis. *Journal of Marketing Research*, 9(2):187 – 192.

André, E., Bevacqua, E., Heylen, D., Niewiadomski, R., Pelachaud, C., Peters, C., Poggi, I., and Rehm, M. (2011). Non-verbal Persuasion and Communication in an Affective Agent. In Cowie, R., Pelachaud, C., and Petta, P., editors, *Emotion-Oriented Systems*, Cognitive Technologies, pages 585–608. Springer Berlin Heidelberg.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2001). The infinite hidden markov model. In *Advances in Neural Information Processing Systems*, pages 577 – 584.

Ben Youssef, A., Shimodaira, H., and Braude, D. (2014). Speech driven talking head from estimated articulatory features. In *in Proc. Acoustics, Speech and Signal Processing (ICASSP)*, pages 4573–4577.

Ben Youssef, A., Shimodaira, H., and Braude, D. A. (2013a). Articulatory Features for Speech-Driven Head Motion Synthesis. In *Interspeech*, pages 2758 – 2762.

Ben Youssef, A., Shimodaira, H., and Braude, D. A. (2013b). Head Motion Analysis and Synthesis over Different Tasks. In *Intelligent Virtual Agents*, pages 285–294. Springer Berlin Heidelberg.

Bengio, Y. and Frasconi, P. (1995). An Input Output HMM Architecture. In *Advances in Neural Information Processing Systems*, pages 427 – 434. MIT Press.

Benoît, C., Grice, M., and Hazan, V. (1996). The SUS Test: A Method for the Assessment of Text-to-Speech Synthesis Intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18(4):381–392.

Bickmore, T. and Cassell, J. (2005). Social dialongue with embodied conversational agents. In van Kuppevelt, J., Dybkjær, L., and Bernsen, N. O., editors, *Advances in Natural Multimodal Dialogue Systems*, volume 30 of *Text, Speech and Language Technology*, pages 23 – 54. Springer Netherlands.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Black, A., Zen, H., and Tokuda, K. (2007). Statistical Parametric Speech Synthesis. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1229–IV–1232.

Boksem, M. A., Meijman, T. F., and Lorist, M. M. (2005). Effects of Mental Fatigue on Attention: An ERP study. *Cognitive Brain Research*, 25:107 – 116.

Borg, I. and Groene, P. J. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer.

Braude, D. A., Shimodaira, H., and Ben Youssef, A. (2013a). Template-Warping Based Speech Driven Head Motion Synthesis. In *Interspeech*, pages 2763 – 2767.

Braude, D. A., Shimodaira, H., and Ben Youssef, A. (2013b). The University of Edinburgh Head-Motion and Audio Storytelling (UoE-HAS) Dataset. In *Intelligent Virtual Agents*, pages 466 – 467.

Brkic, M., Smid, K., Pejsa, T., and Pandzic, I. S. (2008). Towards Natural Head Movement of Autonomous Speaker Agent. In Lovrek, I., Howlett, R., and Jain, L., editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 5178 of *Lecture Notes in Computer Science*, pages 73 – 80. Springer Berlin Heidelberg.

Buchholz, S. and Latorre, J. (2011). Crowdsourcing Preference Tests, and How to Detect Cheating. In *Interspeech*, pages 3053 – 3056.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation*, 42(4):335 – 359.

Busso, C., Deng, Z., Grimm, M., Neumann, U., and Narayanan, S. (2007). Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1075–2007.

Busso, C., Deng, Z., Neumann, U., and Narayanan, S. (2005). Natural Head Motion Synthesis Driven by Acoustic Prosodic Features. *Computer Animation and Virtual Worlds*, 16(3-4):283 – 290.

Carnegie Mellon University (2013). CMU Graphics Lab Motion Capture Database - Behavior Informatics Project. http://mocap.cs.cmu.edu/.

Cassell, J. and Thorisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538.

Choi, K., Luo, Y., and Hwang, J.-N. (2001). Hidden Markov Model Inversion for Audio-to-Visual Conversion in an MPEG-4 Facial Animation System. *Journal of VLSI Signal Processing*, 29:51– 61.

Chuang, E. and Bregler, C. (2005). Mood swings: expressive speech animation. *ACM Transactions on Graphics*, 24:347.

Coxon, A. P. M. (1982). *The User's Guide to Multidimensional Scaling*. Heinemann Educational Books.

de Jong, N. H. and Wempe, T. (2009). Praat Script to Detect Syllable Nuclei and Measure Speech Rate Automatically. *Behavior Research Methods*, 41(2):385–390.

de Ruiter, J. P. (2000). *Language and Gesture*, chapter The production of gesture and speech, pages 284 – 311. Cambridge University Press.

Diebel, J. (2006). Representing Attitude: Euler Angles, Unit Quaternions, and Rotation Vectors.

Eyben, F., W ollmer, M., and Schuller, B. (2010). openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. ACM Multimedia (MM), ACM*.

Ferré, G., Bertrand, R., and Blache, P. (2007). The CID Video Corpus: A Multimodal Resource for Gesture Studies. In *Third ISGS conference 'Integrating Gestures'*, Evanston, Illinois, United States.

Field, A. (2013). *Discovering Statistics Using IBM SPS Statistics*. Sage, fourth edition.

Fine, S., Singer, Y., and Tishby, N. (1998). The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, 32:41 – 62.

Gonzalez, T. F. (1985). Clustering to Minimize the Maximum Intercluster Distance. *Theoretical Computer Science*, 38(0):293 – 306.

Graf, H. P., Casatto, E., Strom, V., and Huang, F. J. (2002). Visual Prosody: Facial Movements Accompanying Speech. In *Proc. 5th International Conf. on Automatic Face and Gesture Recognition*, pages 381–386.

Graham, J. A. and Argyle, M. (1975). A Cross-Cultural Study of the Communication of Extra-verbal Meaning by Gesture. *International Journal of Psychology*, 10:57 – 67.

Hadar, U., Steiner, T., Grant, E., and Rose, F. (1983). Kinematics of Head Movements Accompanying Speech During Conversation. *Human Movement Science*, 2(1-2):35–46.

Hadar, U., Steiner, T. J., and Rose, F. C. (1985). Head Movement During Listening Turns in Conversation. *Journal of Nonverbal Behavior*, 9(4):214–228.

Hamerly, G. and Elkan, C. (2003). Learning the k in k-means. In *In Neural Information Processing Systems*, page 2003. MIT Press.

Helén, M. and Virtanen, T. (2010). Audio Query by Example Using Similarity Measures between Probability Density Functions of Features. In *EURASIP Journal on Audio, Speech, and Music Processing*, pages 1 – 12.

Hill, H. and Johnston, A. (2001). Categorizing Sex and Identity from the Biological Motion of Faces. *Current Biology*, 11(11):880–885.

Hofer, G. and Shimodaira, H. (2007). Automatic Head Motion Prediction from Speech Data. In *Proceedings of Interspeech*, pages 722 – 725.

Hofer, G. O. (2009). *Speech-driven Animation Using Multi-modal Hidden Markov Models*. PhD thesis, University of Edinburgh.

Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321 – 377.

Ishi, C. T., Ishiguro, H., and Hagita, N. (2013). Analysis of Relationship Between Head Motion Events and Speech in Dialogue Conversations. *Speech Communication*, 57:233 – 243.

Izenman, A. J. (2013). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.

Kawahara, H., Masuda-Katsuse, I., and de Cheveign, A. (1999). Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based {F0} extraction: Possible role of a repetitive structure in sounds1. *Speech Communication*, 27(34):187 – 207.

Kita, S. (2009). Cross-cultural Variation of Speech-Accompanying Gesture: A Review. *Language and Cognitive Processes*, 24:145 – 167.

Kita, S. and Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48(1):16 – 32.

Kittur, A., Chi, E. H., and Suh, B. (2008). Crowdsourcing User Studies with Mechanical Turk. In *SIGCHI conference on human factors in computing systems*, pages 453 – 446.

Krauss, R. M. and Hadar, U. (1999). *Gesture, Speech, and Sign*, chapter The role of speech-related arm/hand gestures in word retrieval., pages 93 – 116. Oxford University Press.

Kruskal, J. (1964). Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika*, 29(2):115–129.

Kuratate, T., Munhall, K. G., Rubin, P. E., Vatikiotis-Bateson, E., and Yehia, H. (1999). Audio-Visual Synthesis of Talking Faces for Speech Production Correlates. In *Eurospeech'99*, volume 3, pages 1279 – 1282.

Ladd, D., Faulkner, D., Faulkner, H., and Schepman, A. (1999). Constant "segmental anchoring" of F0 movements under changes in speech rate. *The Journal of the Acoustical Society of America*, 106(3 Pt 1):15431554.

Lambert, Z. V. and Durand, R. M. (1975). Some Precautions in Using Canonical Analysis. *Journal of Marketing Research*, 12(4):468 – 475.

Le, B. H., Ma, X., and Deng, Z. (2012). Live Speech Driven Head-and-Eye Motion Generators. *IEEE Transactions on Visualization and Computer Graphics*, 18:1902 – 1914.

Lee, S. P., Badler, J. B., and Badler, N. I. (2002). Eyes Alive. *ACM Trans. Graph.*, 21:637–644.

Levine, S., Theobalt, C., and Koltun, V. (2009). Real-time prosody-driven synthesis of body language. *ACM Trans. Graph.*, 28(5):172:1–172:10.

Linde, Y., Buzo, A., and Gray, R. (1980). An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28(1):84–95.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif. University of California Press.

Massaro, D. W., Cohen, M. M., Beskow, J., Daniel, S., and Cole, R. A. (1998). Developing and Evaluating Conversational Agents. In *Proceedings of Workshop on Embodied Conversation Characters (WECC)*.

McClave, E. Z. (2000). Linguistic Functions of Head Movements in the Context of Speech. *Journal of Pragmatics*, 32(7):855 – 878.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schröder, M. (2012). The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing*, 3(1):5 – 17.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Bateson, E. V. (2004). Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science*, 15:133–137.

Orden, K. F. V., Jung, T.-P., and Makeig, S. (2000). Combined Eye Activity Measures Accurately Estimate Changes in Sustained Visual Task Performance. *Biological Psychology*, 52:221 – 240.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825 – 2830.

Persson, J., Welsh, K. M., Jonides, J., and Reuter-Lorenz, P. A. (2007). Cognitive Fatigue of Executive Processes: Interaction between Interference Resolution Tasks. *Neuropsychologia*, 45:1571– 1579.

Phillips, C. L., Parr, J. M., and Riskin, E. A. (2008). *Signals, Systems, and Transforms*. Pearson Prentice Hall, fourth edition.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77:257 – 286.

Reips, U.-D. (2002). Standards for Internet-Based Experimenting. *Experimental Psychology*, 49(4):243 – 256.

Rett, J., Faria, D., Neves, A., and Simplicio, C. (2007). HID-Human Interaction Database. paloma.isr.uc.pt/hid/.

Richmond, K., Hoole, P., and King, S. (2011). Announcing the Electromagnetic Articulography (Day 1) Subset of the mngu0 Articulatory Corpus. In *Interspeech*, pages 1505 – 1508.

Sako, S., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). HMM-based Text-to-audio-visual Speech Synthesis. In *ICSLP2000*.

Sargin, M. E., Aran, O., Karpov, A., Ofli, F., Yasinnik, Y., Wilson, S., Erzin, E., Yemez, Y., and Tekalp, A. (2006). Combined Gesture-Speech Analysis and Speech Driven Gesture Sythesis. In *IEEE International Conference on Multimedia and Expo*.

Sargin, M. E., Erzin, E., Yemez, Y., Tekalp, A. M., Erdem, A. T., Erdem, C., and

Özkan, M. (2007). Prosody-Driven Head-Gesture Animation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages II–677 – II–680.

Sargin, M. E., Yemez, U., EnginErzin, and Tekalp, A. M. (2008). Analysis of Head Gesture and Prosody Patterns for Prosody-Driven Head-Gesture Animation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1330 – 1345.

Seyama, J. and Nagayama, R. S. (2007). The Uncanny Valley: Effect of Realism on the Impression of Artificial Human Faces. *Presence: Teleoperators and Virtual Environments*, 16(4):337 – 351.

Smyth, P. (1997). Clustering Sequences with Hidden Markov Models. In *Advances in Neural Information Processing Systems*, pages 648 – 654.

Söderkvist, I. and Wedin, P.-Å. (1993). Determining the Movements of the Skeleton Using Well-configured Markers. *Journal of Biomechanics*, 26:1473 – 1477.

Suwita, A., Böcker, M., Mühlbach, L., and Runde, D. (1997). Overcoming Human Factors Deficiencies of Videocommunications Systems by Means of Advanced Image Technologies. *Displays*, 17(2):75 – 88.

Toda, T., Black, A., and Tokuda, K. (2007). Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222 – 2235.

Toda, T., Black, A. W., and Tokuda, K. (2008). Statistical Mapping between Articulatory Movements and Acoustic Spectrum using a Gaussian Mixture Model. *Speech Commun.*, 50(3):215 – 227.

Tokuda, K. T., Zen, H. Z., and Kitamura, T. (2004). Reformulating the HMM as a Trajectory Model. In *Proceedings of Beyond HMM Workshop on Statistical Modeling Approach for Speech Recognition*.

Tosato, D., Spera, M., Cristani, M., and Murino, V. (2012). Characterizing Humans

on Riemannian Manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints):1.

Turk, A., Scobbie, J. M., Geng, C., Macmartin, C., Bard, E. G., Campbell, B., Diab, B., Dickie, C., Dubourg, E., Hardcastle, B., Hoole, P., Kainada, E., King, S., Lickley, R., Nakai, S., Pouplier, M., Renals, S., Richmond, K., Schaeffler, S., Wiegand, R., White, K., and Wrench, A. (2010). An Edinburgh Speech Production Facility.

Wagner, P., Malisz, Z., and Kopp, S. (2014). Gesture and Speech in Interaction: An overview . *Speech Communication*, 57(0):209 – 232.

Wolters, M. K., Isaac, K. B., and Renals, S. (2010). Evaluating Speech Synthesis Intelligibility using Amazon Mechanical Turk. In *7th Speech Synthesis Workshop (SSW7)*.

Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E. (2002). Linking Facial Animation, Head Motion, and Speech Acoustics. *Journal of Phonetics*, 30:555 – 568.

Zen, H., Tokuda, K., and Kitamura, T. (2007). Reformulating the HMM as a Trajectory Model by Imposing Explicit Relationships Between Static and Dynamic Feature Vector Sequences. *Computer Speech and Language*, 21(1):153 – 173.

Zielinski, S., Hardisty, P., Hummersone, C., and Rumsey, F. (2007). Potential Biases in MUSHRA Listening Tests. In *Audio Engineering Society Convention 123*.