Text Driven Talking Heads

Iain Brown

Undergraduate Dissertation
Computer Science
School of Informatics
University of Edinburgh

2015

Abstract

Head motion is very important when it comes to human communication, a lot of information is given to us as head motions and dialogue is much harder to fully understand in the absence of these head motions. Generating avatars with realistic facial features has been an area of wide research in the past decade. Especially with the rise of applications like virtual reality, video games and online shopping assistants realistic head motions are vital to these interactions otherwise these application will fail. This project aimed to build a system which was capable of synthesising life-like natural head motions from text.

Acknowledgements

Table of Contents

Chapter 1

Introduction

Gestures are a huge part of communication [25] and provide information that adds meaning to communication. In the recent years we have a seen a large increase in embodied conversational agents as a form of human - computer interface which replicate realistic human features to make the interaction seem as natural as possible and provide a rich user experience. These systems need to synthesise facial features and facial gestures that appear natural to humans. Systems that synthesise realistic natural head motions have typically used speech as base, however it can be difficult and expensive to record said speech.The Text-Driven Talking Heads system was developed to synthesise these head motions. This project aimed to generate head motions solely from text by using many natural language processing to analyse the text and use that information to make head motions.

Chapter 2 outlines the background knowledge of this area of research as well as the techniques and tools used int he project. The related work is also contained in this chapter.

The hypotheses about how head motions relate to head motions and the theoretical implementation of the system is present in chapter 3.

Chapter 4 is the actual implementation of the system, explaining how the system was developed.

The 5th chapter contains the evaluation of the system in which two methods of evaluation are presented and experiments are conducted in order to evaluate the system.

The last chapter reviews the project and as well as discussing future improvements to the system.

Chapter 2

Background Information

2.1 Human Computer Interaction

Human Computer Interaction is a field of Computer Science in constant change. With the goal of enhancing Human Computer Interaction researchers have looked to the field of Embodied Conversational Agents, where an intelligent agent is mapped to a graphical animation or body to replicate the most natural of interactions: face to face dialogue. [14]

Embodied Conversational Agents have had been found to enhance the interactions with computers [17] and researchers are perpetually improving systems in order to increase user satisfaction. One of the biggest benefits to using an ECA is that the interaction is of a social nature, being more familiar to humans and aids the systems perceived trust worthiness, allowing a richer interaction between the user and the computer. Also the intelligibility of speech produced in noise is also improved when a speaker's face is visible [21], this means that because the user has a physical representation of the ECA they can more easily understand what the agent is saying rather than if it was just speech output. These results suggest that nonverbal gestures such as head movements play a more direct role in the perception of speech than previously known. [29]

There are some limitations to consider when talking about ECA's, mainly to do with what Masahiro Mori called the Uncanny Valley effect. Mori states that the familiarity with a robot or graphic representation of an avatar increases in correlation with the rise of human likeliness, but there is a distinct fall in familiarity before achieving a life-like human representation referred to as the Uncanny Valley which causes revulsion in humans. [28]

2.2 Data and Task

The task for the project was as follows: Using only text transcriptions of speech can we synthesise life-like head motions that seem realistic and natural to humans. The

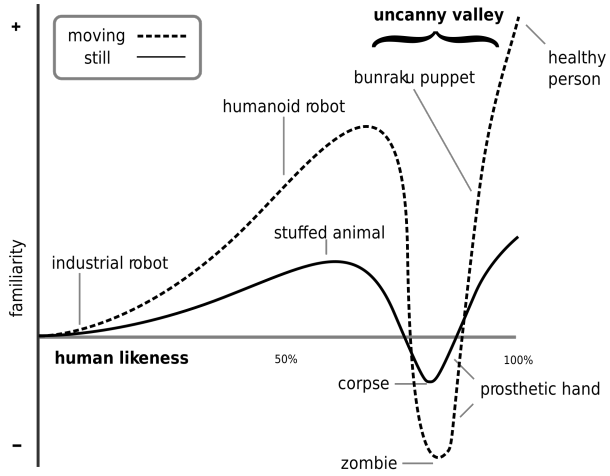

Figure 2.1: The Uncanny Valley effect

project aimed to investigate the correlation between head motions and the information present in speech unique to the speaker. The difficulty of the task comes from the lack of unique information about the speech, as we are only using the transcribed text. To tackle this we will be using a Text-to-Speech system called Festival.

2.2.1 Data Recordings

The data used for the project was data recordings from the Centre of Speech and Technology Research at Edinburgh University. The recordings consisted of optical motion capture sessions in which participants wore 4 reflective markers on their torso and a hat with 3 reflective markers. There were 7 V100:R2 cameras that tracked the reflective markers at a sampling rate of 100Hz. Participants were asked to read out transcriptions of fairy tales.

Figure 2.2: Setup of the mo-cap environment

2.3 Euler angles

Euler angles are three angles that represent the orientation of a rigid body in 3-Dimensional space, typically referred to as 'yaw', 'pitch' and 'roll'. Euler angles allows us to reduce the number of parameters normally used to represent rigid body orientation down to 3 parameters. It is a very popular method because of the reduced complexity and is commonly used in robotics and 3D animation. As the project focused on rigid head motions there was no translation taken into account when designing the 3D avatar head motions. This allowed the project to use Euler angles as the sole units of movement in the project.

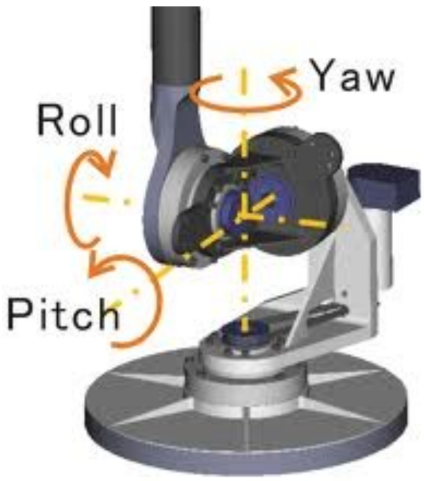

Figure 2.3: Euler Angles in Robotics

Euler angles do have some drawbacks. One of the most common issues animators experience with Euler angles is that different results can occur depending on the order of rotation. For example the rotation Roll x Yaw x Pitch will most likely have a completely different effect to the rotation Pitch x Yaw x Roll. [16] Another potential issue is the Gimbal Lock [38]. Euler angles were chosen because of their ease of use and these drawbacks were not an issue.

2.4 Poser Animation

3D animation and rendering software was used to visualise the generated head motions. The software used for this purpose was PoserPro 2012 by Smith Micro Software, a 3D animation tool with the emphasis on character creation and animation. PoserPro allows users to animate body parts of 3D characters and change their characteristic like rotation, translation and scale, meaning that the head of a character could be moved independently with ease.

With a scripting plug-in for Python called PoserPython, [37] allowing users to create scripts to manipulate objects and characters in the scene using Python. This was one of

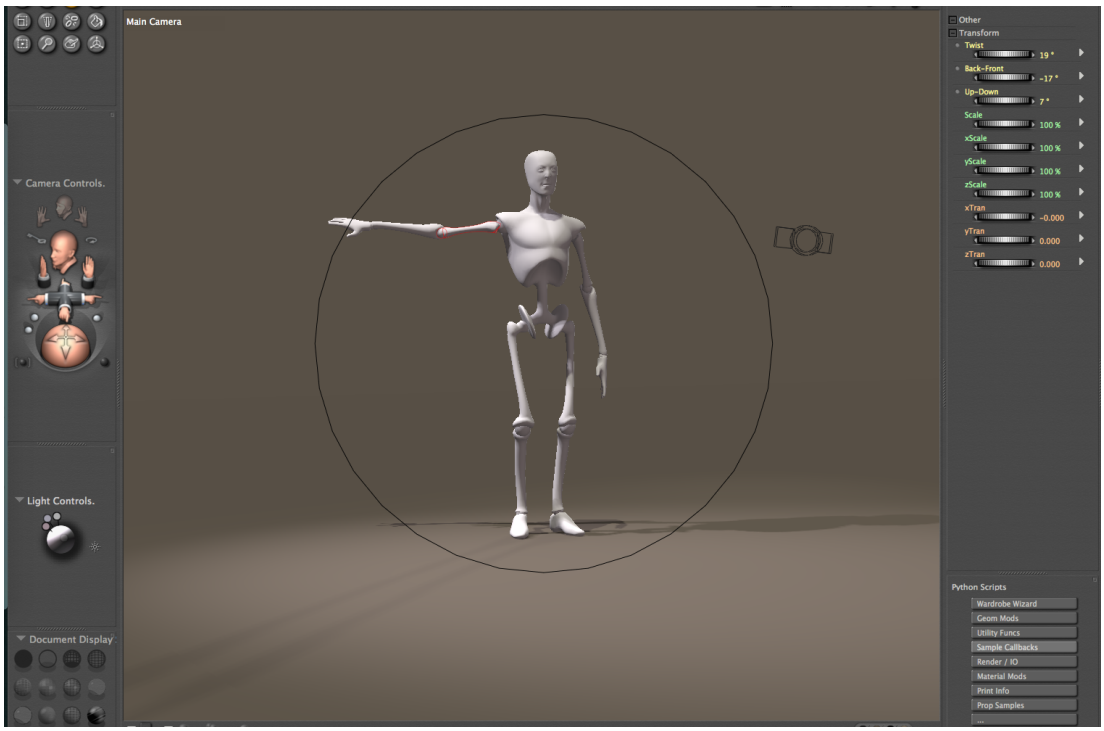

Figure 2.4: Overview of Poser Pro

the main reasons for choosing PoserPro over other animation and rendering software like Blender. Poser also uses Euler angles as it's unit of rotation.

2.5 Related Works

Creating systems that are capable of synthesising life-life human motions as facial animations is a large area of computer graphics research due to it's rapid increase in demand in computer games, feature films and communications.

2.5.1 Animation Synthesis Based on Speech

There are a number of systems that use the paralinguistic information in speech to synthesise head motions and other facial animations using prosodic information like intonation. These approaches are data-driven approaches that use machine learning techniques in order to train models representing head motions and facial animations.

Hidden Markov Models are finite state machines with probabilistic state transitions that can be trained on data which are used in many machine learning based classification tasks [33]. They are generative models of data and so can be used to synthesise similar data to what they have been trained on, because of this HMM's have been used in many of the systems as method of training data-driven systems [10]. HMM's are very

popular for data driven approaches as they eliminate the need to keep a large database of data as they can be trained fairly quickly [13].

G. Hofer designed the Speech Driven Head Motion Synthesis system [19] in which a HMM based system was modelled to synthesise natural head motions by extracting important information from the speech such as the F0 values and the Mel-Frequency Cepstral Coefficients which provide a unique fingerprint of the speech derived from the Fourier transform and is commonly used in signal processing. This approach produced head motions that were within a 70% accuracy to test data. A Ben-Yousse developed a system which improves upon the previous Speech Driven system by using articulacy features extracted the speech[9] to synthesise head motions.

Similarly C. Busso designed a system to synthesise head motions using using speech by training HMM's on data. [12] The system generated discrete vectors for the head notions and applied interpolation as a form of smoothing to create realistic head motions. The system showed that HMM's were capable of capturing the relationship between head motions and speech data.

Systems have also been developed that aim to synthesise facial movements to express emotions, rather than just synthesising natural head motions. Y. Cao designed the Expressive Speech-Driven Facial animation system [13] in which they used similar machine learning techniques to discover the mapping between F0 and head motions. The system takes in recorded speech, extracts key features such as the phoneme sequence and spectral envelope, performs emotion analysis and executes a graph search through their Anime data structure, designed to combine utterances with their recorded head motion. As outlined in the paper the system needs a lot of training data which is expensive to record.

2.5.2 Text Analysis for Facial Animation

All of the approaches to facial animation have used speech data as their primary source of information, but there are also systems in place that perform facial animation synthesis with just text as their primary source of information, without the need for speech. As speech is not always available these system use natural language processing techniques to analyse the text in order to synthesise facial animations like head motions of lip motions.

"May I talk to you? :)" [8] is a system that was designed to synthesise facial animation from just the text. The system uses a Text to Speech system which analyses the text and generates information about the text such as accents and pauses in order to synthesise non-verbal speech related facial expressions like lip motions and head movements. The system also allows the input of emoticons to indicate emotions (text representations of facial expressions) to dramatically alter the synthesised facial animation.

Similar to the previous system, much research has gone into synthesising lip motions solely using the text. In T. Masuko's paper [27], their team designed a data driven system using Hidden Markov Models to accurately generate lip motions from text for

Japanese. The approach is similar to those outlined in the previous section that also use HMM's.

2.6 Subjective Evaluation Approaches

This project used subjective evaluation to determine how natural the synthesised head motions were, according to humans. This was important because the project aimed to synthesise head motions that were natural for humans in order to be used in Embodied Conversational Agents, so humans had to deem the head motions as natural.

Subjective analysis is problematic due to evaluation being based on feelings and not measurable facts. This is a problem because 'feelings' and the evaluation itself can vary widely, even in the same person due to external factors like the time of day. Care must be taken when designing subjective evaluation environments and tests to mitigate variability and many tests need to be conducted in order to account for the this.

There are many different approaches to performing subjective analysis, each with their pros and cons.

2.6.1 Preference Tests

Preference tests are very simple to conduct. A participant is presented with two video samples and asked "Which do you prefer?". This is an example of an AB test using two samples. AB testing is very simple to conduct and so is very common when evaluating samples of video or audio. However, if there were more than two samples then the time needed to evaluate all samples would be significantly higher as each pair combination of samples would need to be conducted. ABX testing, [30] is where the participant is shown three videos (A, B and X) and is asked 'Is X more like A or B'. This approach is an alternative to AB but runs into the same issues with time constraints.

Preference testing is suitable for evaluating a small number of samples, simple to conduct and produces easy to understand results. It's main drawback is the time needed to perform the evaluations which increase quickly as the number of samples increase.

2.6.2 Mean Opinion Score - MOS

When evaluating 'Naturalness' where the concept is not so clearly defined, the MOS test is commonly used as the method of subjective analysis. [34]

The Mean Opinions Score (MOS) is a test to obtain the user's subjective measurement of quality in relation to a particular feature or concept. Generally, users will be asked to rate something on a numerical scale like 1 to 5, 1 being the worst or 'bad' and 5 being the best or 'very good'. This method is fairly trivial and is a very common form of subjective evaluation due to is ease. Results from MOS testing are easier to analyse than preference tests due to their values being numerical.

2.6.3 MUSHRA : Multiple Stimuli with Hidden Reference and Anchor

Multiple Stimuli with hidden Reference and Anchor is another methodology for subjective evaluation more commonly referred to as MUSHRA [35]. In a MUSHRA test, the subject is shown multiple stimuli are available at the same time which the subject can switch between and go back to and then gives a rating from 1-100 depending on how the subject was asked to rate the stimuli. Anchors are used in the evaluation, hidden in the test as one of the stimuli to serve as a reference point for each candidate. These anchors are supposed to score highly on the test. Subjects that score these particular stimuli are usually deemed as unreliable. Typically a MUSHRA test is used for evaluating audio quality [7] where the subject is given 5 audio clips to listen to and rate the quality on a scale of 1 to 100.

The main advantage of MUSHRA over the other two method of subjective evaluation is that many different stimuli can be evaluated at once meaning MUSHRA can scale up the number of samples easily with little change in the time taken. This also means that MUSHRA requires less participants in order to obtain significant results in comparison to MOS even though they are similar in their design.

Chapter 3

Text-Driven Head Motion Synthesis

The goal of this project was to develop a system that was capable of synthesising life-like, realistic head motions from transcribed speech. This project aimed to developed a system that was capable of synthesising head motions given just the text. In order to to this hypotheses about the relation of head motion events and speech content were introduced. The system built on these hypotheses, linking them together in order to synthesis the final head motions. Text itself is very limited in information, so in order to have richer input data the transcribed text would be passed through a Text to Speech system which would apply natural language processing techniques in order to synthesise speech.

3.1 Hypotheses

It was found that when two people have a conversation, head motions are more prevalent in the dialogue if the two people do not have a close relationship. [23] In dialogue the nod is commonly understood as a gesture of affirmation and agreement. A possible explanation for this effect could be that because there is no pre-existing relationship humans subconsciously overcompensate their head motions as a method of positive reenforcement to aid in building a foundation for this new relationship. In the context of Embodied Conversational Agents the system aimed to synthesise head motions that made the interaction between the user and the ECA as natural as possible in relation to head motions, to incorporate this hypotheses into the system the main head motion should primarily be a nodding motion and head motions should overcompensate similar to an interaction with a new person with the aim of making the interaction feel more like a face to face conversation and provide a more comfortable experience.

3.1.1 Prosodic Features

Prosody is the rhythm, stress and intonation of speech. As english is the language domain for the project and english is a stress timed language [36], prosodic features in

speech are very important to the meaning of speech and can change the meaning of the underlying text.

Head motions correlate strongly with prosody present in speech [29]. My first hypothesis relates to the rate of change of the fundamental frequency present in the synthesised speech. This is indicated in the Festival output as values around 120 Hz for male voices and 210Hz for female voices [39]. Sudden changes in the frequency should be reflected in head motions. For example in the intonation falls the head should lower accordingly [**?**].

3.1.2 Phrasing

Phrasing plays a huge role in how something is said. It has been found that nodding head motions frequently occur at the end of phrases or at strong phrase boundaries, especially if the speaker is confident in what they are saying. [22].

3.1.3 Sentiment Analysis

Sentiment analysis is a popular area of natural language processing. It is often used to gauge reviews for products[32] and films[31] due to the availability of data and ease of tagging the reviews as either positive or negative.

Sentiment or emotion in speech has various effects on head motions. In a study It was found that the absence of head motion can be easily identified as 'neutral' emotion. [21] whereas participants in the study found it difficult to differentiate between head motions that were typical of 'happy' and 'sad' emotions. This study shows that emotive analysis on the text should be treated as an 'intensifier' of the underlying head motions derived form other areas of the text rather than altering the head motions to fit an emotion.

Negation is very commonly associated with head shakes [24]. For example, it is very common for people to respond to a question by shaking their head and saying "No". The head shake is strongly associated with negation. Sentiment analysis can be used to determine wether a sentence is positive or negative, however there are many other factors that relate to head shakes. Kendon states that head shakes can be used as an affirmation [25] in many positive sentiment sentences.

For example in the sentence "I have never saw anything more beautiful in my life" has a positive sentiment as the subject is talking about something beautiful but the subject will shake their head on the word 'never' to enforce the positive affirmation. As pointed out in another paper [40] negation statements are very ambiguous and sentiment alone cannot capture the overall meaning and that the lexical structure may play a large role.

3.1.4 Text Content

There are two types of gestures that relate to speech. [18] The first are motor movements which are typically simple, brief, repetitive and have a high correlation with prosodic features. The other type are lexical movements, gestures that help the speaker mentally perform lexical lookups subconsciously. These gestures are very different to motor movements and are much longer, more complex and relate more to the lexical information in the speech. To portray these ideas in this project, unique words that are not common should cause the avatar to tilt their head. This is similar to the theory that eye movements can aid with memory recall. [15]

3.2 Text analysis with Festival

Festival applies various Natural Language Processing techniques to the text to generate information so that it can synthesise speech. There are many steps in this pipeline, each adding a little bit more information to the text before Festival can then apply signal processing to generate audio.

The first stage in the Text to Speech pipeline is the text processing. Festival breaks the text up into more suitable units for processing, for example expanding abbreviations. Then Parts-of-Speech tags are assigned to the units, these POS tags indicate what type of word it is and how it relates to the overall structure of the sentence allowing for phrase break prediction and are represented as a series of capital letters. For example ”NNP” is a pos tag indicating that the word is a singular proper noun. POS tags are useful for the project to determine which words are the subject of the utterance and show where the emphasis should lie in the synthesised speech, which will be reflected in the synthesised head motions.

Phrase break prediction assigns a break strength to each unit, which highlights where the phrases are in the sentences. Phrase break predictors are usually taggers trained on annotated data and are accurate. This is shown in Festival as a series of tags with their assigned values. The ’pbreak’ indicates what level of break has been assigned to this word, for example BB for a big break, B for a normal break or NB for no break.

```
pbreak_index 0 ;
pbreak_index_score 0 ;
pbreak B ;
blevel 3 ;
```

As outlined in the hypothesis section, head motions frequently occur at the end of phrases, where there will be a strong break strength like a big break or normal break. The phrase break value will be very useful to determine these breaks to synthesise head motions.

Festival generates pronunciations by performing syllabification, breaking the sentences and words into syllables and looking up phonemes in a lexicon to determine their pronunciations. Using this information the system can generate ToBi markers [5], a way

of symbolically representing intonation. These symbols are useful as they are conceptually easy to understand and reduce the number of parameters needed to represent intonation. A typical marker will indicate the tone in terms of two letters : 'L' for a low intonation and 'H' for a high intonation, these can be combined to represent a rise from low to high intonation such as 'L-H%' or vice versa. This means that sudden rises or falls in intonation can be identified easily when synthesising head motions.

Similarly to phrase break predictors, Festival uses duration predictors that have been trained on annotated data using classification and regression trees to produce duration information for each phoneme.

Now that the system has a linguistic specification of the sentence like the phone sequence, phone duration and pitch contour signal processing can be performed to generate the final speech output.

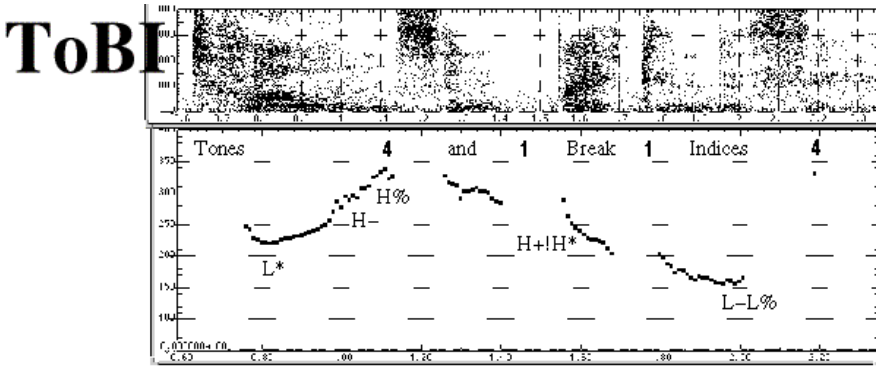

Figure 3.1: ToBi Markers

3.3 Head Motion Synthesis System

3.3.1 Outline

3.3.2 Basic

3.3.3 Random

3.3.4 Rule Based

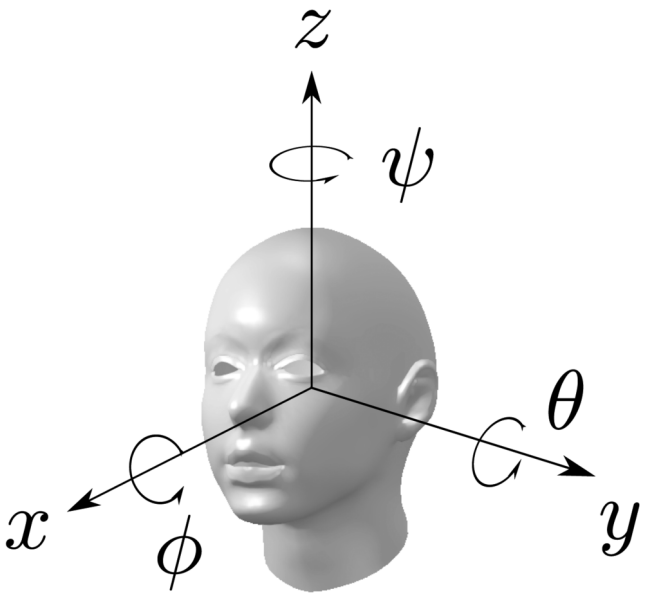

Figure 3.2: How Euler Angles affect head rotation.

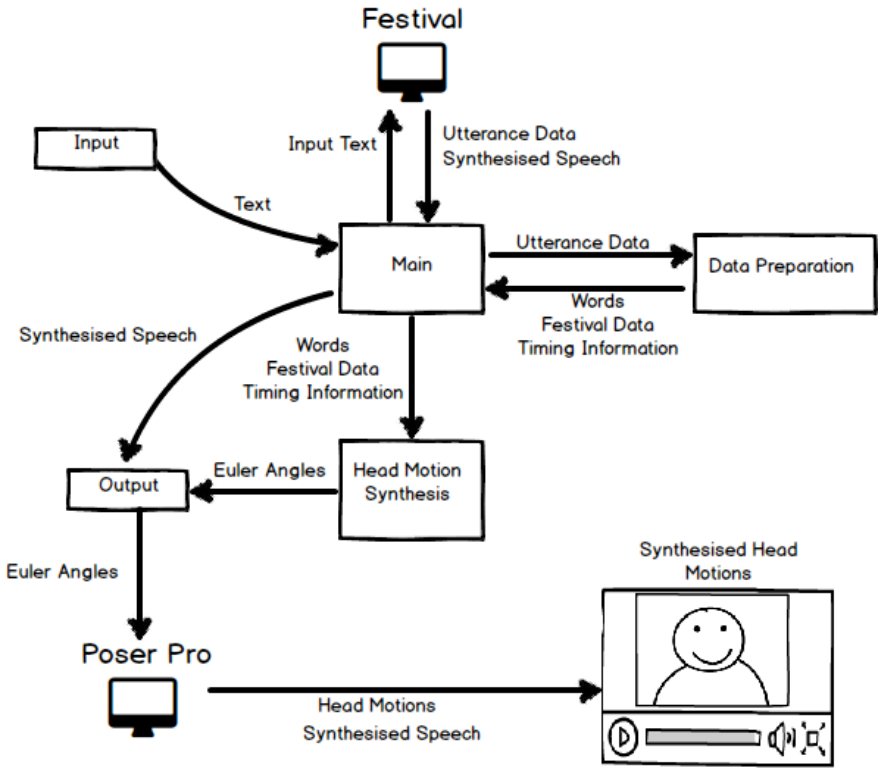

Figure 3.3: Outline of System Architechture

Chapter 4

Implementation

The system was implemented using Python, a high-level programming language widely used for many purposes including scripting and large scale software development. Python was the clear choice for many reasons including powerful libraries such as NLTK (Natural Language Toolkit) [2] and it's scripting compatibility with Poser.

The initial difficulty of the project was that there were a lot of individual components to tie together, like Festival and Poser. The subprocess module in Python allows the script to spawn new processes, connect to their input and output and retrieve their return codes. This was invaluable in the project as it allowed the Python script to call Festival with parameters that could change with each run.

```
fest_location = DIR+'preparation/text2utt.sh'
festival = subprocess.Popen(
                [fest_location,text_file],
                stdout=subprocess.PIPE
                )
utterance = festival.stdout.read()
return utterance
```

The data received from the Festival output was a large block of text containing information about the utterance it had processed (See Figure 4.1). This data was fed into a text processing module implemented from scratch to extract the important information regarding the utterance and build a dictionary using these elements to link the individual words with their properties like parts-of-speech tags and phrase break strength.

Normally Festival is run as an interactive interface. The system used a LISP script called "text2utt.sh" that came with the installation, allowing it to run batch commands in "Text to Speech mode" without entering an interactive state. This was perfect for the project but in order to extract duration information correctly and save the outputted speech to audio files the script had to be altered by adding in the following code.

```
1. (utt.save.words utt outfile 'est_ascii)
2. (save_waves_during_tts)
```

The line of code (save_waves_during_tts) meant that Festival saved all synthesised speech to wave files. An issue that raised from this was that it generated an audio

```
id _23 ; name grandmother ;
         pos_index 8 ;
         pos_index_score 0 ;
         pos nn ;
         phr_pos n ;
         phrase_score -13.43 ;
         pbreak_index 0 ;
         pbreak_index_score 0 ;
         pbreak B ;
         blevel 3 ;
```

Figure 4.1: An Excerpt from the festival analysis output

file for each sentence, which was not suitable for the Text-Driven Talking Heads System. This problem was solved by implementing a function in the preparation stage called c̈ombine_audio_filesẅhich created a new file that was the concatenation of all the sentences. This was a suitable solution as the produced output sounded quite natural.

The Poser Script setMotion.py was taken from another Project that used Euler Angles to rotate a character's limbs, the only alterations made were to set he active body part to the head, add the speech to the scene and to load in the output of the Head Motion Synthesis by hardcoding the name and location of the output file.

The head motion synthesis was developed as three separate modules, increasing in complexity and building on what was successful from the previous head motion synthesis methods.

4.1 Basic System

4.1.1 Trignomietric Functions

As outlined in chapter 3 the most common occurring head motion in dialogue is the nod which was reflected in the analysis of the recorded motion data. This implementation of this hypothesis is the baseline system. It aimed to synthesise a natural nodding motion distributed across the length of the utterance.

The nodding motion is a smooth repetitive oscillation of the head along one axis and visualisation of the data recordings (Shown in figure 4.2) shows that there are large sections of data which are very similar to a sine wave's oscillation so the first system used the trigonometric sine function as it's model.

The file basic_predict.py takes in the dictionary containing all the information retrieved from festival and calculates the number of frames needed for the output rotations and calculates the angle change per frame so that the motion from the first frame to the last frame represents one complete oscillation of the sine formula.

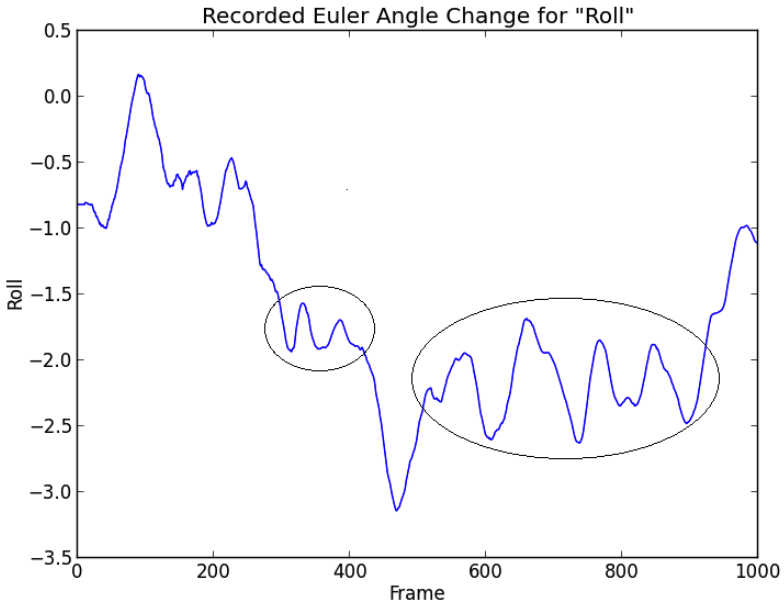

Figure 4.2: Visualisation of Euler change from data recordings

The prediction system adds the rotation information for each Euler angle to the dictionary of utterance data and returns said dictionary. That dictionary is then passed to output.py, which is a generic function that also accepts the utterance dictionary and a filename so that there was no need to re-implement and output function for each system and also allowed for comparison between the systems without necessarily saving them to file allowing rapid prototyping.

As mentioned, the basic system only took one axis into account and so only altered the ”Roll” Euler angle. This system although simple, showed promising results and provided the workflow to have a working system that synthesised head motions, saved them into a .head file containing the Euler angle change for each frame which can be read in by the PoserPython script and assigned to character to be rendered out.

4.2 Random System

The basic system itself produced a natural nodding motion, but after having compared the output of the basic system with the motion recordings, especially comparing different individuals speaking the same sentence there were a low correlation between what the participants were saying and how their head movements changed. To represent this finding the second system moved away from trigonometric functions.

$$B_{P_0}(t) = P_0$$
$$B(t) = B_{P_0P_1..P_n}(t) = (1-t)B_{P_0P_1..P_n-1}(t) + tB_{P_1P_2..P_n}(t)$$

Figure 4.3: Recursive Bezier Definition

4.2.1 Discrete Head Motions

The next system was designed to assign random discrete numbers to each word in the utterance for all Euler angles and apply a form of smoothing to make the random assignment seem smooth and natural even though it was completely random.

Each word in the utterance was assigned Euler values, ranging from -10 to +10 for ”Roll” and -5 to +5 for the ”Pitch” and ”Yaw” as the ”Roll” relates to nodding motions which are found more frequently head motions. The random function was implemented using Python’s random library.

4.2.2 Smoothing

To synthesise smooth head motions between random points multiple interpolation algorithms were considered. Spherical Linear Interpolation (Slerp) was the first that was considered and was already used in similar research [11]. There was difficulty when trying to implement Slerp due to the choosing Euler angles as the unit of rotation. Euler angles are difficult to apply interpolation [16], Slerp commonly uses Quartinions which are more complex. Another method of interpolation which was derived from Slerp which was considered was the Bezier Interpolation algorithm, invented by Pierre Bezier a french mathematician was much simpler to implement than Slerp.

A recursive Bezier function was implemented which, given a list of points return a formula representing the a smooth interpolation between said points. Unlike Slerp, Bezier’s smoothing algorithm doesn’t take the interpolated line to the given points which works well given this scenario. It helped to smooth out the randomness and generate natural motions. (Figure 4.5)

The random system calculated a Bezier function for each of the Euler angles and returned a dictionary containing the Euler angle change for each frame of the animation which was passed to output.py.

4.2.3 Introduction of noise

The output of the second system when mapped to a character in poser showed promise, the motions were smooth and looked like they could have been recorded from actual participants, however the motions produced were deemed too smooth and approached the uncanny valley, looking unnatural.

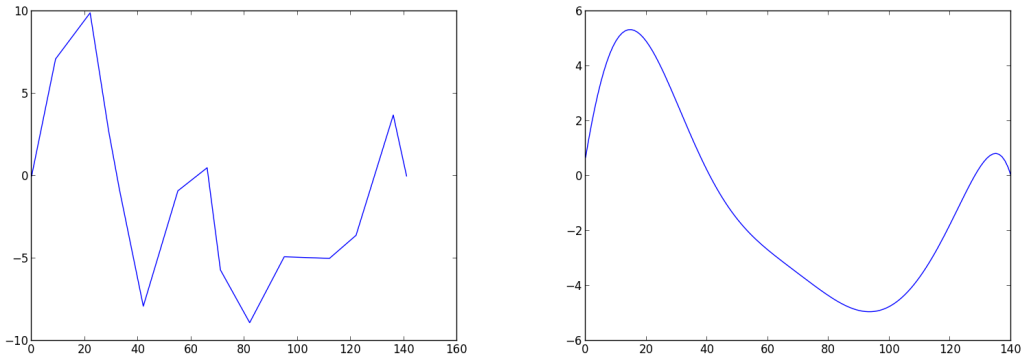

Figure 4.4: Applying Bezier Smoothing to the discrete points

Figure 4.5: Frames from the random system visualisation

To counteract this issue a probability based assignment was introduced which adds or subtracts a very small percentage with the purpose of adding noise to the output. This was done with the use of python's in built random library, functions in the library are based on the random() function which generates a random float uniformly in the range of 0.0 to 1.0. The values used for noise ranged from -0.01 to +0.01. The introduction of noise reduced the feeling of the uncanny valley in the initial results.

4.3 Rule-Based System

The Rule-Based system uses the output from Festival to apply manually written rules in order to synthesise head motions. Having considered the initial results of the previous two system the third built upon those ideas integrating with the rules derived from the multiple hypotheses outlined in chapter 3.

As there if not a 1 to 1 mappings between head motions and the words found in speech, the rule-based system removes stop words : words that are very common or are short function words like 'the', 'and' and 'which'. The system used the list of stop words from the NLTK library.

4.3.1 Using the information from festival

4.3.1.1 Intionation

Initially, the system was designed to store all of the information regarding the fundamental frequency (F0) from the Festival output. This was so that the system contained complete information about the intonation but this proved to be fairly complicated.

```
151 id _113 ; f0 102.921 ; pos 0.22 ;
```

F0 values were only available to the individual phonemes of each word, with only timing value to relate it back to where it belonged. In order to store all of the F0 values a lot of pre-processing was done to carefully assign each value to it's correct word in the proper order. During this pre-processing the system computing the minimum, maximum and average values for comparison.

However, the information was difficult to work with due to the need for excessive pre-processing and so the system used the ToBi intonation tags from the Festival output in order to alter head motions based on the intonation. The ToBi tags proved to be very useful as they represented an overview for the intonation so the pre processing was reduced significantly.

The ToBi tags were reduced down to solely their letter indicators 'L' and 'H' as their combination provided sufficient information regarding the intonation changes. For example, 'L-L' was reduced to 'LL'. Depending of the values of the ToBi tags a 'shift' value was added to the Euler angle change. If the tags were all 'L' a shift value was subtracted, similarly if the tags were all 'H', the shift value was added.

However, if the two ToBi tags were different for that word, the previous value would be 'shifted' to match the pattern and the current value would be shifted with an extra value. The default shift value was 5.0 change in the Euler angle, with an Extra 2.0 in the case of 'L-H' tags or 'H-L'. The intonation changes affected only the Roll Euler angle which influenced nodding motions.

4.3.1.2 Break Strength

```
pbreak BB; pbreak B; pbreak NB;
```

As nodding motions were found at the end of phrases and phrase boundaries, the system was design to reset the Roll angle back to it's origin on strong phrase boundaries. That being phrase break values of BB as they were the strongest. This simple rule produced good results as the synthesised nodding motions returned to first position upon finishing the video, showing that the avatar is done speaking.

4.3.1.3 Uniqueness

As outlined in the hypotheses the system should reflect lexical lookups of unique words by tilting the head. Head tilts are reflected by the Pitch Euler angle.

The system uses the NLTK to build a Frequency Distribution of all the words in the Brown corpus [1]. This is so that each word in the utterance can be compared against this Frequency Distribution to assign a uniqueness value to that word. The system assigns a positive or negative value of 3 if the word has a uniqueness score of less than 5. The uniqueness score is assigned by the 'query_corpus' method that divides the number occurrences of the word in the Brown corpus by the number of the most common word.

4.3.1.4 Sentiment Scaling

The rule based system performs sentiment analysis on the utterance. As NLTK does not come with an in built sentiment classifier other options were considered.

Sentiment Classifier [3] is a sentiment analysis tool based on NLTK and the SentiWordNet that assigns a positive and negative score to words in order to classify sentences.

```
>>> from senti_classifier import senti_classifier
>>> sentences = ['The movie was the worst movie',
'It was the worst acting by the actors']
>>> pos_score, neg_score = senti_classifier.polarity_scores(
    sentences)
>>> print pos_score, neg_score
0 0
>>>
```

The example shown above was a recreation of the example on the package's website. The sentences received scores that did not match the example on their website. This error was due to package constraints not being satisfied. Sentiment Classifier requires NLTK, Numpy and the SentiWordNet corpus. TextBlob [4] was an alternative that did not rely on such package constrains and was far more lightweight than Sentiment Classifier. It assigns a sentiment score to a string of text between 1 and -1, 1 being an extremely positive sentiment and -1 being an extremely negative sentiment. An example of TextBlob is presented below.

```
>>> from textblob import TextBlob
>>> tb = TextBlob("I am very happy because of the nice weather")
>>> tb.sentiment
Sentiment(polarity=0.8, subjectivity=1.0)
```

TextBlob was chosen to analyse the sentiment in the rule based system. The information, both the actual sentiment and the polarity was added to the information dictionary for reference.

As outlined in the hypotheses, negation cannot be captured effectively using purely sentiment analysis, and can often be used in positive sentences. To account for these findings the system used the twitter negation corpus [6] which builds upon [26][20] and contains a list of around 5000 negative words captured from social media platform Twitter.

To use this corpus NLTK's Plaintext Corpus reader was used to read in the text file and store in into a query-able format. When the system iterates over each word, the

$$Y = K * (1 - e^{-t/T})$$

K = Steady State Gain
T = Time Constant
t = time

Figure 4.6: First Order Equation Definition

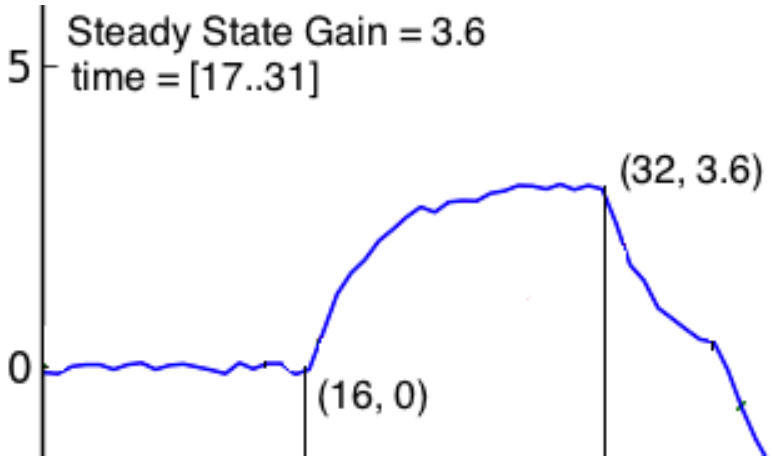

Figure 4.7: Graph showing Parametric smoothing between points.

negation corpus is queried to determine if the word in the utterance is present in the negation corpus. If present, the Yaw angle that controls head shakes will be assigned to the value of 3.0 and the previous word's Yaw value is reduced by 3.0 to produce a shake motion. If not present the system assign a uniform random number from -1 to 1.

4.3.2 Parametric Smoothing

One of the drawbacks from the initial reviews of the random system was that Bezier smoothing does not look like natural head motions. Having compared the output from the random system to the data recordings, the videos seemed to be more sharp and discontinuous, initially while smoothing out toward the end of the motion. The observed effect was similar to that of a first order differential equation, commonly used in electronics to represent the power in a circuit. Rising very quickly to begin with but slowing and smoothing off before reaching the desire level of output.

The parametric smoothing system used the first order differential equation shown above the determine three curve between two points. The equation requires several parameters to generate the correct, which was calculated for each pair of points as shown in Figure 4.7. The Time Constant was set to a value of 4.0. The system builds the formula and calculates the y values for each point in between the two points passed into the method 'curve_between_points()'. This y values represents value of the Euler angle change.

The parametric smoothing provided good results and reflected the discontinuous nature of the recorded head motions.

Chapter 5

Evaluation

To evaluate the Text-Driven Head Motion System I performed both subjective analysis and objective analysis. The aim for the project was to develop a system which generates life-like talking heads with head motions that seem realistic and natural so it was important for humans to evaluate if the head motions were natural or not. Having a unit of measurement describing how close to the original head motions was also necessary in evaluating the system.

5.1 Subjective Analysis

5.1.1 Design Overview

The subjective evaluation was derived from the MUSHRA methodology outlined in chapter 2. It was designed with many factors taken into account. To effectively isolate the head motions for evaluations and make sure that participants were not affected or influenced by other factors, care was taken to ensure that as much as possible would remain constant, only changing the head motions.

Volunteers were shown 5 different video clips and were asked to evaluate how natural they felt the head motions were. Of the 5 videos, 3 were synthesised from the Text Driven Talking Heads system and 2 videos were taken from real recordings. Participants were not told anything about the videos and so did not know if the videos were synthesised or real. This was to eliminate any potential bias that could be introduced by telling the participants some of the videos were real and some were not and made sure that the participants focused on the task.

It was important for participants to be able to review their evaluation and update their choices as their subjective idea of what natural head motions are could easily change after watching the videos. Participants were encouraged to watch the videos more than once and update their evaluation are necessary until they were content with their evaluation.

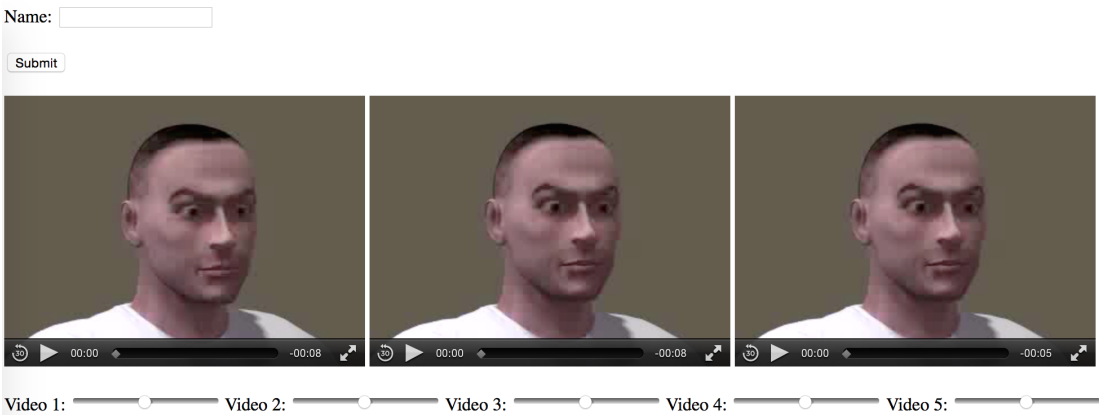

Figure 5.1: Preliminary Evaluation platform

The preliminary evaluation mapped the head motions to a character called 'Dom Casual' from the the Poser Standard Library as shown in figure 5.1.

As there had no be no indication which video was which, the video were simply labelled as number 1 through 5. The key is shown below.

Video 1 Basic system

Video 2 Desmond recordings

Video 3 Rule-based system

Video 4 Jane recordings

Video 5 Random system

5.1.2 Implementation

The evaluation platform was implemented in a web-based format using HTML5, Python and Flask, a Python framework used for web development. The website (shown in figure 5.1), showed the 5 videos side by side with sliders allowing the users to click and drag the slider to a level they felt reflected how natural the video above that slider was. This approach accomplished two things: having sliders ranging from 0 to 100 allowed for much greater precision than asking participants to rate the video from 1-5 or 1-10 and also helped users generate numeric feedback without the need to arbitrarily assign a numerical value.

As outlined in the design overview, only once participants were content with their evaluation would the results be recorded. The 'submit' button commits the users evaluation to a text file. This was suitable for the purpose of the evaluation system as it very simple to implement and had little development overhead.

5.1.3 Preliminary feedback

Preliminary evaluations were carried out to to test the evaluation platform. A small number of volunteers were asked to perform the evaluation and were asked a series of questions about the environment.

1. How difficult was the task? What were the areas of difficulty?
2. Did you find the 3D avatars creepy?
3. Would the task be easier or more diffuclt with longer videos?

This was to try and improve the user experience before the final evaluation. The questions chosen were to address some key concerns regarding the environment. The participants were to feel comfortable during the evaluation process and if they videos they were evaluating were in the uncanny valley, many participants would feel discomfort which could negatively impact results. Gauging if participants felt the videos were too short or too long was important as well, the video needed to be of appropriate length so users had the right amount of information to effectively evaluate the videos.

The preliminary sessions provided good feedback on the system. Many participants stated that the avatar used for the videos was very distracting, particularly with the eyes of the avatar. Two of the six participants reported that they feel the avatar definitely influenced their evaluation. Participants reported that the length of the videos was suitable as they could re-watch the videos and re-evaluate the video.

5.1.4 Final Evaluation

Having taken on board the feedback from the preliminary results, the evaluation platform was tweaked. The head motion videos were re-exported from Poser with a different avatar. Rather than using a particular character from Poser Pro's library, the basic model called 'Andy' was used that lacks facial details like hair and appears as an emotionless manikin. This helped the participants focus solely on the head motions and not be distracted by the previous character's expression or facial features. The final subjective evaluation was performed on 15 participants.

5.1.5 Results

The subjective evaluation led to interesting results and results showed a definite trend of which videos they felt were most natural. Interestingly, the head motions that were taken from recorded data were deemed to be fairly natural usually scoring around 60-70. These videos were added into the subjective evaluation as anchors for the evaluation and were intended to score highly but they scored less than expected. A possible reason for this could be that the recorded data does not necessarily align with what most people deem to be natural head motions.

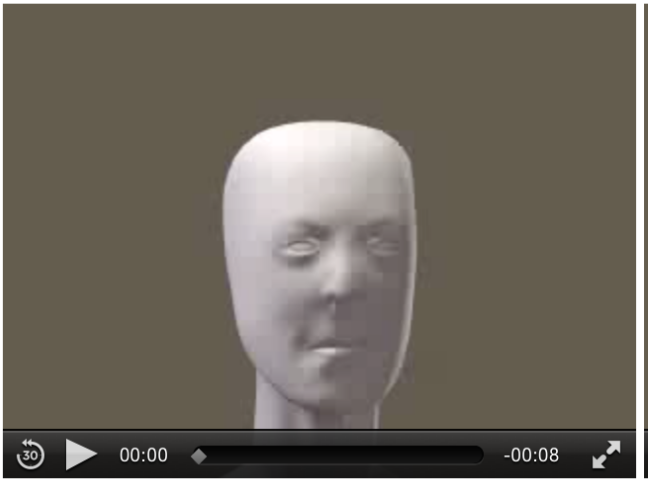

Figure 5.2: Avatar used for final Evaluation

The basic system that used a sine wave to generate a smooth nodding motion scored relatively well against the anchors, usually on par with one or both of the anchors. The basic system scored a mean of 59, which was between the mean score of the two anchors.

The rule based system scored 29.93 in the evaluation and was deemed un-natural by almost all of the participants. Many participants reported that the head motions were very 'jerky' and looked un-natural because of this. This does not necessarily mean that the hypotheses made regarding introducing more discontinuous head motions is incorrect and could be because parameters used for smoothing were not optimal in this scenario.

The random system scored highest overall with mean value of 71.6, including the videos taken from the data recordings. With no relation to the content of the text this is a very surprising result.

5.1.6 Conclusions

The main factor which seemed to influence how natural participants deemed the head motions was how smooth the head motions were. The rule based system performed poorly in the evaluation because of it's jerky, discontinuous nature and the random system scored very highly due to it's smooth motions. These results also highlight a potential issue with the evaluation method: participants are given any information about the text the head motions were synthesised or taken from (in the case of the data recordings) and how the head motions relate to what is being said is not a factor.

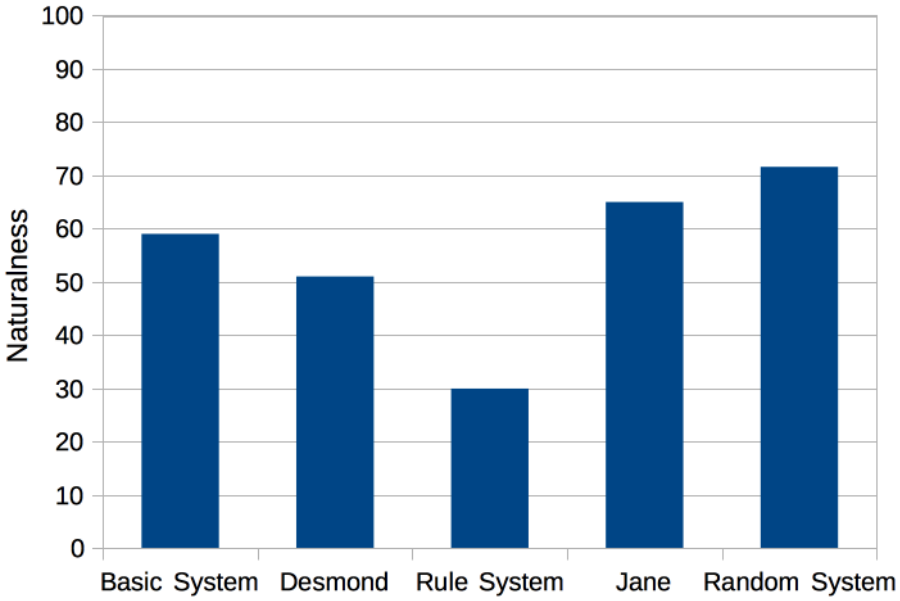

Figure 5.3: Mean 'naturalness' of each video

5.2 Objective Analysis

To evaluate the system objectively the output of a system was compared with both data recordings used for evaluation previously. The Euler angles were periodically sampled every 10 frames and the distance between the two head positions was calculated using the Euclidian distance measure. The values were then normalised using their sampling rate so that different sampling rates could be used and compared. The neutral position was also used for the objective evaluation, that being the point (0,0,0). The sampling rate used for the objective evaluation was every 10th frame.

$$Euclidian(x,y) = \sqrt{(x1-y1)^2+(x2-y2)^2+...+(x_n-y_n)^2}$$

5.2.1 Results

Euler Distance	No movement	Sine	Random	Rule
Desmond	6.3	7.02	6.4	6.94
Jane	17.68	18.08	18.8	18.6

The results from the objective evaluation show that the two head motions that are most similar in terms of raw distance measure are that of the origin (no Euler change) and the Desmond Euler angles. The distance between the head position and the origin was seemed to be a large factor when comparing the distance. The distance between

Desmond and the origin is very close to the distance between Desmond and all the synthesised motions and the same is true for Jane.

The approach taken to measure the head motion similarity is naive, and does not take into account the possibility of head motions being very similar in gesture like a nod but having a low score because of a large distance in the other Euler angles due to tilts and shakes present in the head motions.

Chapter 6

Discussion

6.1 Conclusions

The goal of this project was to build a system that is capable of synthesising life-like, natural head motions from text and that is what this project has achieved. Results from the evaluation showed that the random system with bezier smoothing was deemed the most natural of head motions, even over the real life recordings with the basic sine wave system coming scoring very similar results to the data recordings. This may be due to the fact that the environment in which the data recordings took place was a not a typical environment where head motions would be observed like in a real life conversation resulting in the real life recordings being scored worse than expected. This suggests that without adding in some kind of context to the evaluation, head motions that are smooth will be deemed more natural by humans.

The Rule based system performed poorly in the evaluation achieving a naturalness score of 29 due to it's jerky head motions even though this system implemented the hypotheses outlined in chapter 3 the system fell short because of it's smoothing. This supports the findings regarding natural head motions being perceived as smooth. With some tweaks to parameters and the type of smoothing used this system could perform much better, especially if the participants were given more content in the evaluation for example being played the synthesised speech with the head motions or simply shown the text and asked which video do they feel best represents the head motions.

6.2 Future Work

6.2.1 Better Smoothing

As previously mentioned the rule based system's smoothing was deemed very unnatural by all the participants in the subjective evaluation. To counteract this the smoothing should be improved upon. The parameters used for the parametric smoothing may have been sub-optimal for the conditions. The utterances used were very short and there

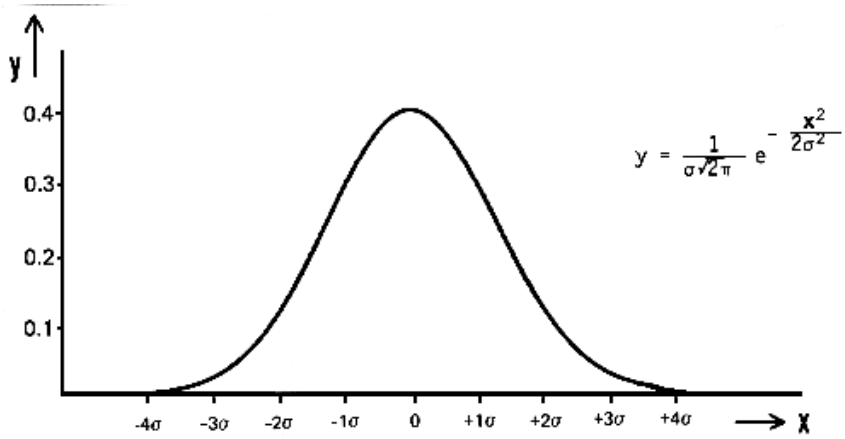

Figure 6.1: The bell curve of the gaussian distribution

were many points that needed to be smoothed between over a short distance, which explains why the movements were so jerky. To tackle this in the future the smoothing should no necessarily smooth directly between neighbouring points but smooth over a group, similar to the bezier smoothing.

A suitable distribution for this could be the Gaussian distribution, as shown in figure 6.1 it produces a smooth curve rising to a point and then falling back down to where it began. The gaussian could be very useful as the rise begins slowly and increases as we get closer to the mean value, which could reduce the jerky nature of the head motions produced from the rule based system whilst maintaining it's individual motions that would be reduced using the bezier smoothing algorithm.

6.2.2 Evaluation Redesign

The subjective evaluation platform that was implemented was very useful when obtaining results, t did however have some flaws. The evaluation of the head motions were not merited in terms of their relation to the text they were synthesised from as participants were only shown the head motions, not the text they were generated from or the synthesised speech from the text. If the experiments were to be performed again, context should be given to the participants. The videos could have the synthesised speech play as well as the head motions so that the participants will be able to evaluate if the head motions match what the avatar is saying.

6.2.3 Data Driven System

Due to time constraints the project never reached a data driven system. Many existing system's are using Hidden Markov Models as a tool to develop a data-driven approach to synthesising realistic head motions as outlined previously in the related works section. The next system which was planned to be implemented was a data driven approach to synthesising head motions.

Using Hidden Markov Models, the relation between the head motions and text information generated from festival would be mapped. A direct mapping between the head motions and the individual words would be a naive approach but it could capture some of the hypotheses outlined especially in regards to negation and using head shakes to represent words like ”no and never”. [24]

Rather than using the individual words another approach would be to use the structural information present in the text information from festival like the Parts-of-speech tags and phrase break information. As we know head motions occurs more frequently at phrase boundaries [22] training HMM's using the structural information and the head motions could build a good model of the data. This approach could provide good results as it is far more general than the naive individual words approach and there would be more data to use as the occurrence of pos tags would be far higher than individual words providing a larger training set.

Bibliography

[1] Brown corpus. http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/.

[2] Natural language toolkit. http://www.nltk.org.

[3] Sentiment classifier. http://pythonhosted.org//sentiment_classifier/.

[4] TextBlob : Simple text processing. http://textblob.readthedocs.org/en/dev/.

[5] Tobi : Symbollically representing intionation. http://www.ling.ohio-state.edu/~tobi/.

[6] Twitter negation corpus : Twitter text mining in R. https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/blob/master/data/opinion-lexicon-English/negative-words.txt.

[7] The mushra audio subjective test method. *White Paper WHP 038*, 2002.

[8] I. Albrecht, J. Haber, K. Kahler, M. Schroder, and H.-P. Seidel. "may i talk to you? : -) " - facial animation from text. In *Computer Graphics and Applications, 2002. Proceedings. 10th Pacific Conference on*, pages 77–86, 2002.

[9] David A. Braude Atef Ben-Youssef, Hiroshi Shimodaira. Articulatory features for speech-driven head motion synthesis. pages 0–1, 2013.

[10] Matthew Brand. Voice puppetry. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 21–28, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[11] C. Busso, Zhigang Deng, M. Grimm, U. Neumann, and S. Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1075–1086, March 2007.

[12] Carlos Busso, Zhigang Deng, Ulrich Neumann, and Shrikanth Narayanan. Natural head motion synthesis driven by acoustic prosodic features. *Computer Animation and Virtual Worlds*, 16(3-4):283–290, 2005.

[13] Yong Cao, Wen C. Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Trans. Graph.*, 24(4):1283–1302, October 2005.

[14] Justine Cassell. *Embodied conversational agents*. MIT press, 2000.

[15] Stephen D Christman, Kilian J Garvey, Ruth E Propper, and Keri A Phaneuf. Bilateral eye movements enhance the retrieval of episodic memories. *Neuropsychology*, 17(2):221, 2003.

[16] Martin Likkholm Erik B. Dam, Martin Koch. Quartionions, interpolation and animation, 1998.

[17] Mary Ellen Foster. Enhancing human-computer interaction with embodied conversational agents. In *Proceedings of the 4th International Conference on Universal Access in Human-computer Interaction: Ambient Interaction*, UAHCI'07, pages 828–837, Berlin, Heidelberg, 2007. Springer-Verlag.

[18] Yihsiu Chen Frances H. Rauscher, Rovert M. Krauss. Gesture, speech and lexical access: The role of lexical movements in speech production. *Pyschological Science*, 1996.

[19] Gregor Hofer, Hiroshi Shimodaira, and Junichi Yamagishi. Speech-driven head motion synthesis based on a trajectory model. 2007.

[20] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[21] Sasha N. Ilnyckyj. Communication of emotional states through rigid head motion in speakers and singers.

[22] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of inter- and intra-speaker variability of head motions during spoken dialogue. In *AVSP*, pages 37–42, 2008.

[23] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication*, 57:233–243, 2014.

[24] Adam Kendon. Some uses of the head shake. *Gesture*, 2(2):147–182, 2002-01-01T00:00:00.

[25] Adam Kendon. Gesture: Visible action as utterance. 2004.

[26] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM, 2005.

[27] Takashi Masuko, Takao Kobayashi, Masatsune Tamura, Jun Masubuchi, and Keiichi Tokuda. Text-to-visual speech synthesis based on parameter generation from hmm. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 6, pages 3745–3748. IEEE, 1998.

[28] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *Robotics & Automation Magazine, IEEE*, 19(2):98–100, 2012.

[29] Callan DE Kuratate T Vatikiotis-Bateson E. Munhall KG1, Jones JA. Visual prosody and speech intelligibility: head movement improves auditory speech perception. pages 133–137, 2004.

[30] W. A. Munson and Mark B. Gardner. Standardizing auditory tests. *The Journal of the Acoustical Society of America*, 22(5), 1950.

[31] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[32] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[33] L. Rabiner and B.H. Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, Jan 1986.

[34] Junichi Yamagishi Rasmus Dall and Simon King. Rating naturalness in speech synthesis: The effect of style and expectation. *In Proceedings of Speech Prosody,*, 2014.

[35] ITUR Recommendation. Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra). *International Telecommunications Union, Geneva*, 2001.

[36] Peter Roach. On the distinction between stress-timedand syllable-timedlanguages. *Linguistic controversies*, pages 73–79, 1982.

[37] Smith Micro Software. Poserpython for poser. `http://www.smithmicro.com/support/faq-graphics/downloads/PoserPython_8_Methods_Manual.pdf`.

[38] Jonathan. Strickland. What is a gimbal – and what does it have to do with nasa? `http://science.howstuffworks.com/gimbal.htm`, 2015.

[39] Hartmut Traunmuller and Anders Eriksson. The frequency range of the voice fundamental in the speech of male and female adults.

[40] Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 60–68. Association for Computational Linguistics, 2010.