

Text Driven Talking Heads

Iain Brown

Undergraduate Dissertation
Computer Science
School of Informatics
University of Edinburgh
2015

Abstract

Table of Contents

1	Introduction	5
2	Background Information	7
2.1	Human Computer Interaction	7
2.2	Data and Task	7
2.2.1	Data Recordings	8
2.3	Euler angles	8
2.4	Poser Animation	9
3	Text-Driven Head Motion Synthesis	11
3.1	Hypotheses	11
3.1.1	Prosodic Features	11
3.1.2	Phrasing	12
3.1.3	Sentiment Analysis	12
3.1.4	Text Content	12
3.2	Text analysis with Festival	13
3.3	Head Motion Synthesis Systems	14
3.3.1	Basic	14
3.3.2	Random	14
3.3.3	Rule Based	14
4	Implementation	15
4.1	Basic System	16
4.1.1	Trigonometric Functions	16
4.2	Random System	17
4.2.1	Discrete Head Motions	17
4.2.2	Smoothing	17
4.2.3	Introduction of noise	18
4.3	Rule-Based System	18
4.3.1	Using the information from festival	19
4.3.2	Parametric Smoothing	19
5	Evaluation	21
5.1	Subjective Analysis	21
5.1.1	Design Overview	21
5.1.2	Implementation	22

5.1.3	Preliminary Results and feedback	22
5.1.4	Results	23
5.1.5	Conclusions	23
5.2	Objective Analysis	23
5.2.1	Results	23
6	Conclusions	25
6.1	System Overview	25
6.2	Discussion	25
6.2.1	Classification and Regression Trees	25
6.2.2	Data Driven System	25
6.3	Future Work	25
6.3.1	Expansion of techniques	25
6.3.2	Second Order Smoothing	25
	Bibliography	27

Chapter 1

Introduction

- Head motion is very important when it comes to human communication
- Dialogue is much harder to fully understand without the non-verbal information
- Generating Lifelike avatars in many applications, VR, video games, shopping assistant
- Realistic head motions are vital otherwise humans may feel weird interacting with an avatar
- Project aimed to create a system that synthesises natural head motions just from the text
- Without knowledge present in speech
- Using Various Natural Language Processing Techniques

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. [5]

Embodied Conversational Agents have had been found to enhance the interactions with computers [8] 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 [10], 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. [15]

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. [14]

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

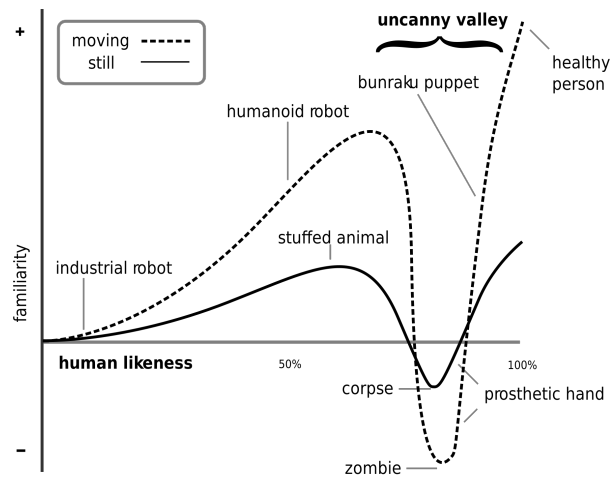


Figure 2.1: The Uncanny Valley effect

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.

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 software because of this. 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.

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



Figure 2.2: Euler Angles in Robotics

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. [7] Another potential issue is the Gimbal Lock [20]. 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, [19] allowing users to create scripts to manipulate objects and characters in the scene using Python. This was one of 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.

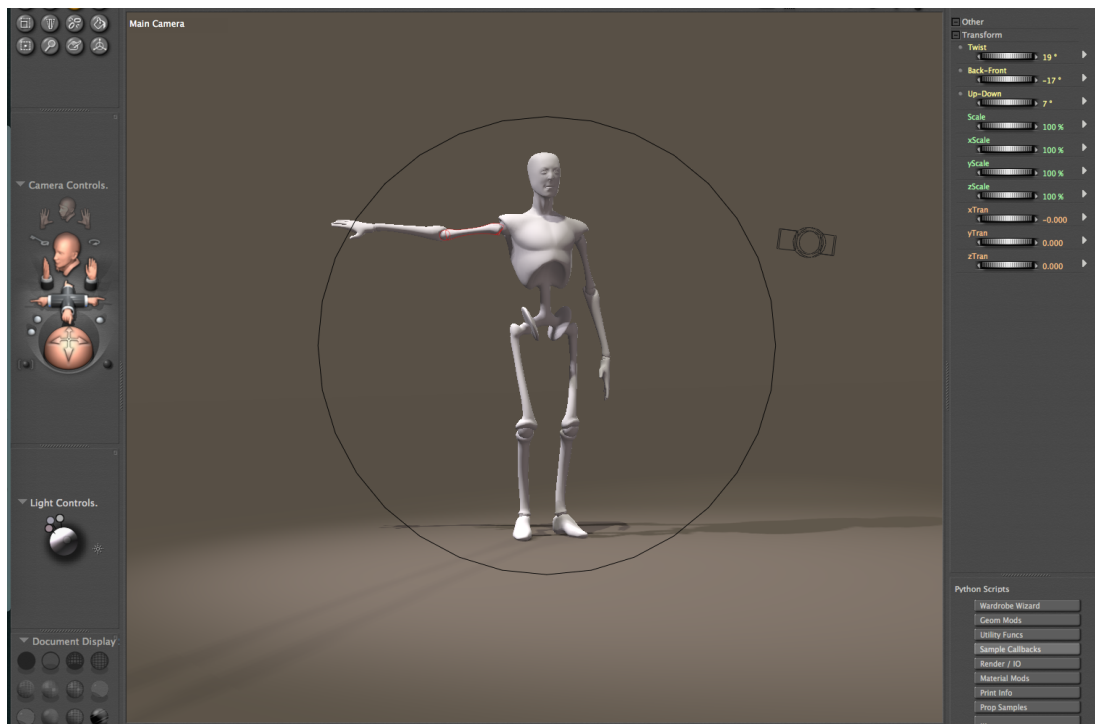


Figure 2.3: Overview of Poser Pro

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 develop a system that was capable of synthesising head motions given just the text. In order to test these 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 synthesise 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. [12] 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 reinforcement 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 these 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 [18], 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 [15]. 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 [21]. Sudden changes in the frequency should be reflected in head motions. For example in the intonation falls the head should lower accordingly [13].

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. [11].

3.1.3 Sentiment Analysis

Sentiment analysis is a popular area of natural language processing. It is often used to gauge reviews for products[17] and films[16] 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. [10] 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 from other areas of the text rather than altering the head motions to fit an emotion.

3.1.4 Text Content

There are two types of gestures that relate to speech. [9] 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. [6]

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.

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.

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 [3], a way of symbolically representing intonation. These symbols are useful as it they are conceptually easy to understand and reduce the number of parameters needed to understand intonation.

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.

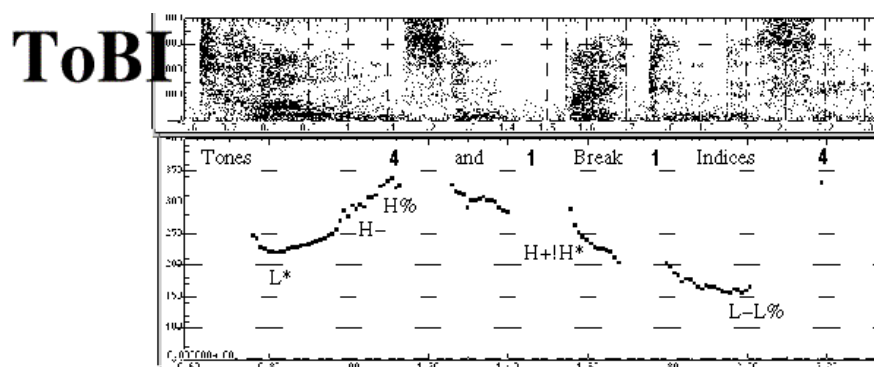


Figure 3.1: ToBi Markers

3.3 Head Motion Synthesis Systems

3.3.1 Basic

3.3.2 Random

3.3.3 Rule Based

Chapter 4

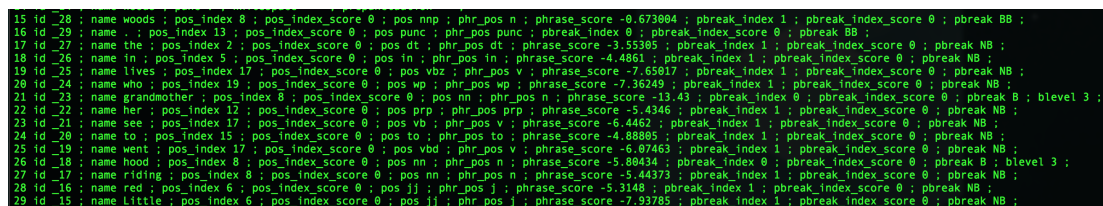
Implementation

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.



```
15 id _28 : name woods ; pos_index 0 ; pos_index_score 0 ; pos nnp ; phr_pos n ; phrase_score -0.673004 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak BB ;
16 id _29 : name . ; pos_index 13 ; pos_index_score 0 ; pos punc ; phr_pos punc ; pbreak_index 0 ; pbreak_index_score 0 ; pbreak BB ;
17 id _27 : name the ; pos_index 2 ; pos_index_score 0 ; pos dt ; phr_pos dt ; phrase_score -3.55305 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
18 id _26 : name in ; pos_index 5 ; pos_index_score 0 ; pos in ; phr_pos in ; phrase_score -4.4861 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
19 id _25 : name lives ; pos_index 17 ; pos_index_score 0 ; pos vbz ; phr_pos v ; phrase_score -7.65017 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
20 id _24 : name who ; pos_index 19 ; pos_index_score 0 ; pos wp ; phr_pos wp ; phrase_score -7.36249 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
21 id _23 : name grandmother ; pos_index 8 ; pos_index_score 0 ; pos nn ; phr_pos n ; phrase_score -13.45 ; pbreak_index 0 ; pbreak_index_score 0 ; pbreak 0 ; level 3 ;
22 id _22 : name her ; pos_index 12 ; pos_index_score 0 ; pos prp ; phr_pos prp ; phrase_score -5.4346 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
23 id _21 : name see ; pos_index 17 ; pos_index_score 0 ; pos vb ; phr_pos v ; phrase_score -6.4462 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
24 id _20 : name to ; pos_index 15 ; pos_index_score 0 ; pos to ; phr_pos to ; phrase_score -4.88885 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
25 id _19 : name went ; pos_index 17 ; pos_index_score 0 ; pos vbd ; phr_pos v ; phrase_score -6.07463 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
26 id _18 : name hood ; pos_index 8 ; pos_index_score 0 ; pos nn ; phr_pos n ; phrase_score -5.80434 ; pbreak_index 0 ; pbreak_index_score 0 ; pbreak 0 ; level 3 ;
27 id _17 : name riding ; pos_index 8 ; pos_index_score 0 ; pos nn ; phr_pos n ; phrase_score -5.44373 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
28 id _16 : name red ; pos_index 6 ; pos_index_score 0 ; pos jj ; phr_pos j ; phrase_score -5.3148 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
29 id _15 : name little ; pos_index 6 ; pos_index_score 0 ; pos jj ; phr_pos j ; phrase_score -7.93785 ; pbreak_index 1 ; pbreak_index_score 0 ; pbreak NB ;
```

Figure 4.1: An Excerpt from the festival analysis output

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 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 `combine_audio_files` which 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 the 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 Trigonometric 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 so we represented this using the trigonometric sine function.

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.

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 an output function for each sys-

tem 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 didn't seem to be any kind of 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.

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.

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 [4]. 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 [7], Slerp commonly uses Quaternions 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.

$$B(t) = \sum_{i=0}^n \binom{n}{i} (1-t)^{n-i} t^i P^i$$

Figure 4.2: Recursive Bezier Definition

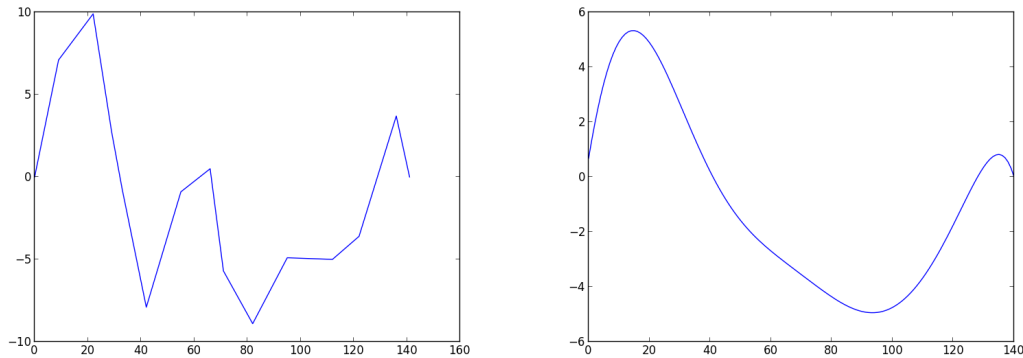


Figure 4.3: Applying Bezier Smoothing to the discrete points

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.

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 reduced the feeling of the uncanny valley and lead to good 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.

$$Y = K.A * (1 - e^{-t/\tau})$$

Figure 4.4: First Order Equation Definition

4.3.1 Using the information from festival

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.

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 evaluation system 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 as necessary until they were content with their evaluation.

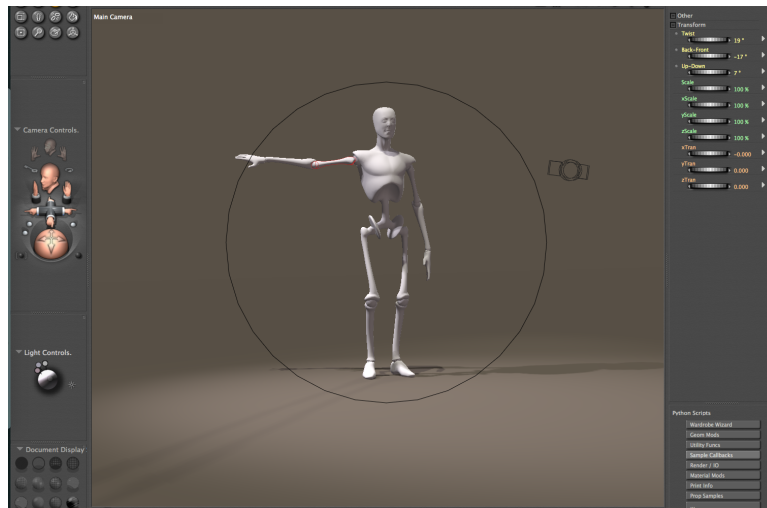


Figure 5.1: Evaluation Platform

5.1.2 Implementation

The evaluation platform was implemented in a web-based format using HTML5, Python and Flask [1], 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 Results and feedback

Preliminary evaluations were carried out 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 difficult 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.

”Everything was fairly easy.” - Participant 1

”Another quote” - Participant 2

5.1.4 Results

- Taken on board suggestions from preliminary feedback
- 15 volunteers
- Good results for Random system
- Good results for rule-based system
- Bad results for ones taken from data

5.1.5 Conclusions

5.2 Objective Analysis

- Calculate difference in Euler angles
- Compare each system to two recordings due to lack of transcriptions

5.2.1 Results

Chapter 6

Conclusions

6.1 System Overview

6.2 Discussion

6.2.1 Classification and Regression Trees

6.2.2 Data Driven System

6.3 Future Work

6.3.1 Expansion of techniques

6.3.2 Second Order Smoothing

Bibliography

- [1] Flask - a python micro framework. <http://flask.pocoo.org>.
- [2] Natural language toolkit. <http://www.nltk.org>.
- [3] Tobi : Symbollically representing intonation. <http://www.ling.ohio-state.edu/~tobi/>.
- [4] 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.
- [5] Justine Cassell. *Embodied conversational agents*. MIT press, 2000.
- [6] 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.
- [7] Martin Likkholm Erik B. Dam, Martin Koch. Quaternions, interpolation and animation, 1998.
- [8] 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.
- [9] Yihsiu Chen Frances H. Rauscher, Rovert M. Krauss. Gesture, speech and lexical access: The role of lexical movements in speech production. *Psychological Science*, 1996.
- [10] Sasha N. Ilnyckyj. Communication of emotional states through rigid head motion in speakers and singers.
- [11] 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.
- [12] 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.
- [13] Adam Kendon. *Gesture: Visible action as utterance*. 2004.

- [14] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *Robotics & Automation Magazine, IEEE*, 19(2):98–100, 2012.
- [15] 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.
- [16] 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.
- [17] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
- [18] Peter Roach. On the distinction between stress-timed and syllable-timed languages. *Linguistic controversies*, pages 73–79, 1982.
- [19] Smith Micro Software. Poserpython for poser. http://www.smithmicro.com/support/faq-graphics/downloads/PoserPython_8_Methods_Manual.pdf.
- [20] Jonathan. Strickland. What is a gimbal – and what does it have to do with nasa? <http://science.howstuffworks.com/gimbal.htm>, 2015.
- [21] Hartmut Traunmuller and Anders Eriksson. The frequency range of the voice fundamental in the speech of male and female adults.