In the Neck of Time

by

Snehal Poojary

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Master
in Design Innovation with Computer Graphics.

Victoria University of Wellington
2016

Abstract

Numerous studies over the past decade have investigated to making human animation as realistic as possible, especially facial animation. Let's consider facial animation for human speech. Animating a face, to match up to a speech, requires a lot of effort. Most of the process has now been automated to make it easier for the artist to create facial animation along with lip sync based on a speech provided by the user. While these systems concentrate on the mouth and tongue, where articulation of speech takes place, very little effort has gone to understand and to recreate the exact motion of the neck during speech. The neck plays an important role in voice production and hence it is essential to study the motion created by it.

The purpose of this research is to study the motion of the neck during speech. This research makes two contributions. First, predicting the motion of the neck around the strap muscles for a given speech. This is achieved by training a program with position data of marker placed on the neck along with its speech analysis data. Second, understanding the basic neck motion during speech. This will help an artist understand how the neck should be animated during speech.

Acknowledgments

I would first like to thank my thesis advisors Assoc. Prof. John Lewis of the School of Engineering and Computer Science at Victoria University and Assoc. Prof. Douglas Easterly of the School of Design at Victoria University. They have helped me whenever I ran into a trouble spot or had a question about my research or writing. They both consistently allowed this paper to be my own work, but steered me in the right the direction whenever they thought I needed it.

I would also like to thank the experts who were involved in this research project: Dr. Cathy Ferguson, Otolaryngologist, Head and Neck Surgeon; and Evan Hazenberg, PhD candidate, Applied Linguistics and Language Studies. I would also like to thank the participants for this study. Without their passionate participation and input, this study could not have been successfully conducted.

I would also like to acknowledge Liviu Sas and Dr. Dilhani Bandaranayake as the second readers of this thesis, and I am gratefully indebted to them for their very valuable comments on this thesis. Figures from other sources are either copyright free or used under the fair use copyright principle.

Finally, I must express my very profound gratitude to my parents, sister and to my friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Contents

1	Introduction	1
1.1	Problem Statement	2
1.2	Objective	2
1.3	Approach	2
2	Methodology	5
3	Background and Literature Review	7
3.1	Background	7
3.1.1	Neck Anatomy	7
3.2	Literature Review	11
3.2.1	Neck Motion	11
3.2.2	Human speech analysis	13
4	Human speech analysis	15
4.1	Voiced Sound	15
4.2	Unvoiced Sound	16
4.3	Phonemes	16
4.3.1	Vowels	16
4.3.2	Consonants	16
4.4	Speech Graph	17
4.4.1	Wave Pattern	17
4.4.2	Spectrogram	17

4.4.3	Formants	18
4.4.4	Pitch	19
4.4.5	Intensity	20
5	Machine Learning	21
5.1	Supervised vs Unsupervised Learning	22
5.2	Decision Tree Method	22
5.2.1	Classification vs Regression	23
5.3	Random Forest	24
6	Implementation	25
6.1	Participants	25
6.2	Recital	26
6.3	Markers	28
6.4	Acquiring Training Data	39
6.4.1	Tracking	39
6.4.2	Audio data	39
6.4.3	Dealing with Incorrectly Labelled Phonemes	39
7	Building the System	41
7.1	Train and Predict	41
7.2	Virtual System	42
8	Results	45
8.1	Program results	45
8.2	Key point for Animators	49
9	Limitations and Future Work	55
9.1	Participants' data limitation	55
9.2	Program limitations	56
9.3	Future Work	56
10	Conclusions	59

Chapter 1

Introduction

One of the primary focal points of the computer graphics industry is simulation: creating the illusion of realism for media spanning games, film, training simulations as well as therapy. Most artists will tell you that the key to making any virtual object appear realistic is by paying attention to the tiny details. The more realistic detail you add to the virtual object, the more chances of you achieving the final goal. Consider facial animation, many studies have focused on trying to recreate the actual facial expressions for humans. Even micro expression are being added just to make the expression look as realistic as possible. In movies like *'Furious 7'* [38] we have all seen the brilliant results due to all the details added onto Brian O'Conner's character played by Paul Walker. In the last decade or so the facial animation for speech has vastly improved. Systems have been built to create realistic speech articulation for virtual characters. You can even give an audio input to the program and it will complete the lip sync animation along with speech articulation for any rigged character. These systems also provide you with a basic neck animation but what appears to be missing is the detail required for realistic motion of the neck. In the movie *'The BFG'* [35], the giants character has very realistic facial animation. Even though the character has some neck animation, it's still missing some realism. The motion caused by the larynx and the strap muscle during speech

is complicated but visible. This motion has been mostly ignored while animating a character. With this research, we are trying to study this motion and check if it is possible to recreate.

1.1 Problem Statement

Are we able study the motion caused by the neck while speaking and check if the motion is similar in all humans. If so, is it possible to recreate this motion for a given speech?

1.2 Objective

Our objectives are:

- To study the motion of the neck during speech.
- To build a system which creates the neck animation for any given speech

1.3 Approach

In order to solve the questions posed in the problem statement, we have used the following steps:

1. Conducted expert interviews to get an understanding of the various topics that was be required for this research.
2. Conducted a literature survey to determine to what extent other work had solved the problem.
3. Used our findings from the literature to determine the technique that worked for us to achieve our goals.

4. Filmed participants while speaking with markers placed on their neck.
5. Collected and conducted analysis on the video data.
6. Built a system that uses the data to achieve our goal.

For this study we applied machine learning to find the correlation between the marker data and the speech audio data. We will not review machine learning in detail.

Chapter 2

Methodology

Considering the different aspects of this research, we decided 'The Expert Interview' methodology would be the best approach. An expert interview is a qualitative empirical research designed to explore experts' knowledge of their field.

The neck is a complex structure, therefore understanding the anatomy of the neck will help with recreating the motion of the neck. It is important to learn how voice is produced and what parts of the neck are involved in this process. To have a better understanding of the anatomy and the working of the neck during a speech, an interview was conducted with Dr. Cathy Ferguson, Otolaryngologist, Head and Neck Surgeon. She explained how voice is produced with a detailed explanation of the functions and correlations of the larynx, the strap muscles and various other parts of the neck. This will be explained in detail in Chapter 3.

After gaining an understanding of the anatomy, the next step was to determine how to read a human speech audio. This would be used to find any correlations with the movement of the various parts of the neck involved in speech. It would also be used to build the system that animates the neck for any given speech. We spoke to Evan Hazenberg, PhD candidate, Applied Linguistics and Language Studies who have a good knowledge about Linguistics. This will be further explained in Chapter 4.

Chapter 3

Background and Literature Review

This chapter is divided into two parts. First being the background which covers the neck anatomy. The second part focuses on the literature review that prepared and framed my approaches and research methods.

To study the motion around the neck region during speech and to recreate the same, one needs to consider these aspects:

1. Neck anatomy.
2. Neck animation.
3. Audio analysis/human speech analysis.

3.1 Background

3.1.1 Neck Anatomy

As shown in Fig. 3.1, the human vocal system consists of different structures each of which causes different vibrations and disturbances in the air as it passes through them. Each structure adds a certain feature to the human sound[19].

The structures are:

- the nasal cavity
- the mouth cavity
- the tongue
- the glottis (voice-box)
- the trachea and lungs and
- the velum(soft palate at the back of the mouth)

Figure 3.1: Human sound system [26]

As we are only looking at the neck region during a speech, we will be concentrating on the glottis which is a structure in the neck. We look at the anatomy of the neck in order to determine the motion around the neck region during a speech. After speaking to Dr. Ferguson, we focused on the understanding of the biology and function of the neck in voice production using Gray's anatomy[15].

There are two types of muscles in the neck that are important in the production of speech.

- Intrinsic muscles
- Extrinsic muscles

Intrinsic muscles

Figure 3.2: Intrinsic muscles: larynx [31]

The phonation process which is the production of speech sounds, occurs at the larynx [31] which is part of the intrinsic muscles. The larynx contains

the vocal folds. The vocal cords consist of folds of ligament extending from the thyroid cartilage in the front to the arytenoid cartilages at the back, as seen in Fig. 3.2. The gap between these folds is called the glottis. The intrinsic muscles of the larynx are the ones that move the vocal cord and produce the basic sound. Voiced sounds are created by vibrations of the vocal folds. During a “normal” mode of vibration, the vocal cords open as shown in Fig. 3.3 and close completely during the cycle and generate puffs of air roughly triangular in shape when airflow is plotted against time. For normal breathing, the arytenoids are spaced well apart. They come together when sound is produced. For unvoiced consonants, the folds may be completely open (such as when producing “s”, “sh”, and “f” sounds) or partially open (for “h” sounds) [36]. The vocal folds do not merely rotate in axial plane to open and close. When the glottis opens, each vocal fold changes shape by flexing at the vocal process, which is the junction between the membranous and cartilaginous portions [40].

Extrinsic Muscles

The extrinsic muscles which are the strap muscles, directly surround the larynx. The strap muscles are so-called because they are small, flat, and strap-like. More technically, they are called the infrahyoid muscles [21]. Infra means below and hyoid refers to the hyoid bone, so this group of muscles is below the hyoid bone. Fig. 3.4 shows the four individual pairs of muscles that make up the strap muscles are:

- the sternohyoid which extends from sternum to hyoid bone
- the omohyoid which extends from shoulder to hyoid bone
- the sternothyroid which extends from sternum to thyroid cartilage, aka the side of the larynx
- the thyrohyoid which extends from the thyroid cartilage to the hyoid bone

Figure 3.3: Intrinsic muscles: larynx [14]

Together these four pairs of muscles suspend the larynx and serve to lower the hyoid bone and larynx during speech or swallowing. They work in opposition to the suprahyoid muscles, the muscles above the hyoid bone which lifts the whole structure.

3.2 Literature Review

3.2.1 Neck Motion

The intrinsic laryngeal muscles, which both originate and insert on laryngeal cartilages are responsible for nearly all motion of the vocal folds.

Figure 3.4: Muscles of the neck. Anterior view. [14]

However, the extrinsic muscles which connect the larynx to other structures, can also influence shape and motion of the vocal folds [31]. The motion caused by extrinsic muscles are the ones we see around the neck region during speech. The raising and depressing of the extrinsic muscles is related to the tone of the voice. If you want to talk in a deep voice you lower the larynx and if you want to talk in a high pitch you raise the larynx (you tighten it).

Many muscles in the neck that are attached to the larynx and/or hyoid bone have voice and speech-related contractions due to their role in controlling the vertical position of the larynx in the neck and, to some degree, the position of the tongue. In their seminal paper, “The function of strap

muscles in speech: Pitch lowering or jaw opening?" Atkinson and Erickson [2] explored the sternohyoid muscle, one of several extrinsic laryngeal muscles, and its role in controlling voice frequency. The authors of this paper conducted an electromyographic experiment on a speaker under normal and clenched jaw conditions. The researches relied on this method in order to investigate the sternohyoid muscle, one of several extrinsic laryngeal muscles, and its role in controlling voice frequency[2]. The speaker was asked to say "Bev loves Bob" with emphasis on various words. To summarise the experiment, the utterance having the same pitch pattern regardless of articulatory difference shows very similar sternohyoid activity. Utterances having the same articulatory and jaw opening gestures (but different pitch patterns) show very different sternohyoid activity. They reached the conclusion that sternohyoid activity is more closely related to pitch effects than to jaw opening effects [2]. This leads to an understanding that the neck motion can be studied by concentrating on the strap muscle. Furthermore, in "A muscle model for animating three-dimensional facial expression" Waters [39] describes using muscle simulations to animate a model of a human face. The author derived a method for a simple modelling of a muscle process that is suitable for the animation of a number of divergent facial types. The same concept could be applied to the neck. We determine the muscle motion of the neck which will be used to build our system to recreate this motion.

3.2.2 Human speech analysis

Human speech analysis is a well defined area of research. There are many techniques to read audio. We used spectrograms [9] to read the audio files used throughout our experiments and data collection. We will be looking at this in detail in Chapter 4.

In the seminal paper 'Interactive Speech-driven Facial animation' Hodgkin-

son [19] has achieved a similar goal for facial animation. The author analyses and uses a speech to automatically create 3D facial animation of the speech articulation for a 3D model. This paper refers to the various aspects that would also be used for our study such as, human speech analysis, wave theory and phonemes. This paper also mentions important distinctions in sound classification, which categorises whether the sound is voiced or unvoiced. This was used in creating the virtual system, which was able to differentiate the various classes of sound and then match it to a predefined neck shape to create an animation for that particular voice speech.

Chapter 4

Human speech analysis

The complexity of human brains involves substantial anatomy and neurology dedicated to audio processing. The problem we faced for this study is that we needed to rely on computers to do the processing for us. To learn the key features of human speech analysis, we spoke to our linguistics experts.

4.1 Voiced Sound

Voiced sound is produced when the vocal folds in the larynx are vibrating during the production of a particular consonant. Voiced sounds, when viewed on a spectrogram exhibit high amounts of energy clustered around certain frequency levels [6]. You can tell if a sound is voiced or not by putting your hand gently on your throat. You can feel this vibration by placing your fingers on the side of your larynx, and alternately pronouncing a sustained [zzzzzz] and [ssssss]. The vibration you feel during the voiced [z] will cease during voiceless [s] [42].

4.2 Unvoiced Sound

Unvoiced sounds are consonants produced without vibration of the vocal folds. Unvoiced sounds exhibit more of a general low-intensity cloud of energy spread across many different frequencies. Pitch is not an important characteristic of unvoiced sounds [6].

4.3 Phonemes

A phoneme is a single unit of sound that has meaning in any language. The most common way to to classify human speech is to break up the sound into phonemes [12] [7] [10]. Phonemes are initially classified into voiced or unvoiced [9], which approximately map to the concept of vowel and consonant respectively[19].

4.3.1 Vowels

Vowel is a speech sound which is produced by comparatively open configuration of the vocal tract with vibrations of the vocal cords but without audible friction. It is a unit of the sound system of a language that forms the nucleus of a syllable.

4.3.2 Consonants

Consonants is a basic speech sound in which the breath is at least partly obstructed and which can be combined with a vowel from a syllable. While analysing consonants it is not enough to look at the position of the tongue. One also needs to needs to look at the airflow from the lungs [34].

4.4 Speech Graph

A speech graph is a visual representation of a speech. It allows one to read the human voice for various qualities. This section will be dedicated to explaining the various features a speech graph comprises of.

4.4.1 Wave Pattern

Wave patterns are created when samples from the microphone are plotted. As shown in Fig. 4.1 the x-axis represents time and the y-axis represents the amplitude.

Figure 4.1: Wave file from a male participant using Praat [5].

4.4.2 Spectrogram

Spectrograms [9] are used to show spectral density. It is a graph where the x-axis is usually time, the y-axis is frequency, and the amplitude of the component of the wave of the given frequency is represented by the brightness (or darkness) of the dot (pixel) on the graph. The louder the

sound at a given frequency and time, the brighter (or darker - depending on the graph palette) the appropriate pixel would be. This can be seen in Fig. 4.2

Figure 4.2: Spectrogram from an audio file for female participant using Praat [5].

4.4.3 Formants

Formants are distinctive frequency components of the acoustic signal produced by speech. The information that humans require to distinguish between speech sounds can be represented purely quantitatively by specifying peaks in the amplitude/frequency spectrum[9]. The formant with the lowest frequency is called F1, the second F2, and the third F3.

The vibrating vocal folds apply energy to the air in the vocal tract at multiple harmonic frequencies. Some of these frequencies will happen to match or come close to a resonant frequency of the particular vocal tract shape assumed at the moment. Those resonant frequencies (formants) will be amplified as the repeating waves reinforce each other. Frequencies produced

by the vocal folds that do not come close to a resonant frequency of the vocal tract will be damped out – their energy will be randomly dispersed. Thus the vocal tract is a filter[42]. Fig. 4.3 shows the various formants represented by red dots.

Figure 4.3: Formants represented as red dots from an audio file for a male participant using Praat [5].

4.4.4 Pitch

Pitch or frequency, in speech is the relative highness or lowness of a tone as perceived by the ear. Pitch depends on the number of vibrations per second produced by the vocal cords. Pitch is the main acoustic correlate of tone and intonation. Pitch of the sound is referred to in Hertz (Hz). Pitch of a man's voice fall under low frequency, whereas woman's voice is of the high pitch type. Fig. 4.4 shows pitch data represented by blue lines for a male participant. The pitch is sometimes clearly defined and sometimes not. The voiced portions of speech are frequently nearly periodic and may be assigned a pitch experimentally. Unvoiced consonants, gen-

erated partly (and often entirely) by air turbulence, don't have a readily assignable pitch (and neither does silence.) [28]

Figure 4.4: Pitch from an audio file for a male participant using Praat [5].

4.4.5 Intensity

The intensity represents the loudness of the sound. It is measured in decibels (dB). Fig. 4.5 shows the speech intensity for a male participant.

Figure 4.5: Intensity from an audio file for a male participant using Praat [5].

Chapter 5

Machine Learning

To be able to recreate the neck motion one has to find the link between the strap muscle and the speech. As it is highly complex to establish the link, machine learning was utilised. Machine learning is programming computers to optimize a performance criterion using example data or past experience. It may make predictions of the future, or be descriptive, gaining knowledge from the data, or both [1] [24]. It is used for pattern recognition and computational learning theory in artificial intelligence. In our problem, the idea is to find patterns and correlations between the formants, pitch, intensity from audio data and the relative tracked marker data. We are also looking to automatically predict the values to use for the animation for any given speech.

Broadly speaking, machine learning methods can be divided into two categories:

- Supervised learning
- Unsupervised learning

5.1 Supervised vs Unsupervised Learning

Supervised learning is when you are given a dataset which already has an output for the input data. Hence giving it an idea of the relationship between input and output data [20].

In unsupervised learning we have no idea of what the output should look like. You can derive structures from data without knowing the effect of the variables [13]. A standard example is clustering data into groups. In this case, the criterion used for clustering, such as the distance between data points, is independent of the data.

Consider an example where we want to teach a computer to distinguish pictures of cats and dogs. Tags have been added to the pictures stating if it was a 'cat' or a 'dog'. We can use this data to "supervise" our algorithm in learning the right way to classify images. Once our algorithm learns how to classify images we can use it on new data and predict labels ('cat' or 'dog' in our case) on previously unseen images. For unsupervised learning example, consider that the images were not tagged. You can still split this data into 2 categories. You can employ unsupervised learning to separate the images in two groups based on some inherent features(characteristics) of the pictures.

Supervised learning was used for our system as we had the motion data of the neck in the form of tracked marker data. There is also a relation between the marker data and the audio data comprising of formants, pitch, intensity and vowels. This method will be used for predicting the marker data to animate the neck on the basis of the new audio data.

5.2 Decision Tree Method

A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. It is used to predict the output as shown in Fig. 5.1 The decision tree maps a tree-like model of decisions and their

possible consequences [33]. The two main types of decision trees are [22] [37]:

- Classification tree
- Regression tree

Figure 5.1: Decision tree.

5.2.1 Classification vs Regression

A classification tree is used to separate the dataset into classes belonging to the response variable. Usually the response variable has two classes: Yes or No (1 or 0). You can have more than two classes but there will always be a finite number of classes.

A regression tree is used when the response variable is numeric or continuous. Since our target variable does not have classes, we fit a regression model to the target variable each of the independent variable. Then for each independent variable the data is split at several split points. At each split point, the "error" between the predicted value and the actual values

is squared to get a "Sum of Squared Errors (SSE)". The split point errors across the variables are compared and the variable/point yielding the lowest SSE is chosen as the root node/split point. This process is recursively continued. Regression trees are applicable for prediction type of problems as opposed to classification and hence used for this study.

5.3 Random Forest

Unfortunately, decision trees have high variance [17][18]. Bootstrap aggregation (bagging) is a general procedure that can be used to reduce the variance for those algorithm that have high variance. Applying bagging to decision trees results in random forests [8]. A random forest is simply an average of a number of decision trees, that are each trained with different random splits, or different random subsets of the training data.

Random forest changes the algorithm for the way that the sub-trees are learned so that the resulting predictions from all of the sub-trees have less correlation. When a new input is entered into the system, it is run down all of the trees. The result may either be an average or weighted average of all of the terminal nodes that are reached, or, in the case of categorical variables, a voting majority.

For this study we have used scikit-learn machine Learning [27] in Python [32] to write our program. It's a simple and efficient tool for data mining and data analysis.

Chapter 6

Implementation

One of the most important components of this study was to observe and find repetitive motion patterns that can be associated with its audio. This was achieved by filming both male and female participants, with high-speed video cameras, while they recited a few passages and words. The participants had markers on key points of the neck, for us to be able to track the motion and collect this data. All the data was later compared, to look for similar motion patterns for the same speech. Later the audio file of the speech would be analysed to access data such as formants, pitch and intensity. This data was used to recreate the motion of the neck in a 3D environment during speech using a program written by us.

6.1 Participants

The data collection technique we judged as most appropriate relied on markers placed on the strap muscles of the neck. We wanted to have optimal conditions for data collection, therefore participants had to meet a number of characteristics and conditions to be considered eligible for inclusion. This was primarily achieved by only considering male and female participants between the ages 20 - 45 years. Conscious effort was made to only have participants of this age group as they would have op-

timal voice and skin conditions. Until puberty male and female larynx are similar in size. After puberty, the larynx in the female increases only slightly. In the male it undergoes considerable increase; all the cartilages are enlarged and the thyroid cartilage becomes prominent in the middle line of the neck, while the length of the rima glottidis is nearly doubled [14]. Skin tends to get loose with age and hence we excluded participants with wrinkles on their neck. Also, voice tends to tremble with age and hence we would not get the best speech voice data if older age groups were included. Participants with slender necks were chosen in order to capture the muscle movement and to avoid tracking movement caused by a double chin. A double chin is a layer of subcutaneous fat located at the anterior neck covered by sagging or stretched skin, creating a redundancy, making the person appear as if he/she has a second or third chin [25]. Our expert suggested that we should film participants who have a common accent in order to maintain similar linguistic patterns. Accents typically differ in quality of the voice, pronunciation and distinction of vowels and consonants which will be required for the speech audio analysis. Hence we filmed participants speaking with a New Zealand accent in order to have common speech patterns.

6.2 Recital

Each participant was asked to recite the alphabet, numbers from 1 to 20, vowels and the words: yes, no, okay, cat, banana oil, cut, father, together. After this they were asked to recite two passages, the short version of ‘The Rainbow passage’ [11] and ‘The grandfather passage’ [16]. These passages are two of the most common standard reading passages used to test an individual’s ability to produce connected speech. These passages represent a diverse inventory of English phonemes appearing both in isolation and also nested within an array of phonotactically improbable clusters [30]. These passages offer an ideal mode of eliciting speech and reading errors

and hence they have been used in speech evaluation and also for testing language recognition software.

The Rainbow Passage [11]

"When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow."

The Grandfather Passage [16]

"You wished to know all about my grandfather. Well, he is nearly ninety-three years old. He dresses himself in an ancient black frock coat, usually minus several buttons; yet he still thinks as swiftly as ever. A long, flowing beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks his voice is just a bit cracked and quivers a trifle. Twice each day he plays skillfully and with zest upon our small organ. Except in the winter when the ooze or snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, "Banana Oil!" Grandfather likes to be modern in his language."

Only two participants had to recite an additional passage which would be used to predict motion for the neck markers. This was to check if the system we built was able to predict motion without being trained on the same speech.

The North Wind and the Sun [23]

"The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed

that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveller fold his cloak around him, and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two."

6.3 Markers

Figure 6.1: Experimenting with marker placement.

Markers were approximately placed on the skin above the strap muscles. It was important to observe and gently feel the neck to locate the strap muscles. We would often have to engage the participants in conversation in order for the strap muscles to be visible. As shown in Fig. 6.1 we ex-

perimented with various marker positions in order to get the best motion data [41]. In the end 19 markers were placed for the front camera and additional 5 markers for the profile. Fig. 6.2 shows the approximate marker placement [41]. Fig. 6.3 shows the muscles directly under the skin. Fig. 6.4 shows the strap muscles [41]. Fig. 6.5 shows the larynx [41]. Markers 1-3 were placed in the centre above the sternohyoid muscles. The larynx lies opposite the 2nd and 3rd cervical vertebra in children and females. It is somewhat lower for an adult male, where it lies opposite the 3th-6rd cervical vertebrae. The larynx would be approximately situated around marker 1 for female participants and around marker 2 for male participant. Markers 4-9 were placed on the collar bone. Markers 10-13 were placed above platysma muscle at the point where it overlay the sternocleidomastoid muscle. Markers 14 and 15 were placed on the platysma muscle just north of the collar bone. Markers 16 and 17 were placed on trapezius muscle. Markers 18 and 19 were placed at the base of sternocleidomastoid muscle next to the collar bone.

Figure 6.2: Layer 1 shows the approximate position of the markers [41].

Figure 6.3: Layer 2 shows the muscles under the skin [41].

Figure 6.4: Layer 3 shows the strap muscles [41].

Figure 6.5: Layer 4 shows larynx [41].

6.4 Acquiring Training Data

6.4.1 Tracking

Each participants video is used for tracking the markers. After the markers are tracked, we looked for a neutral position for the neck, i.e. when the participants are not speaking. We exported the x and y position values for all trackers for the entire length of the video along with its neutral position and its relative position . The position of these markers was used along with the speech pitch and formants to identify the relation between the speech and the neck motion.

6.4.2 Audio data

After collecting the video data, we extracted the audio file. This audio file went through two phases. In the first phase for the speech analysis we extracted wave pattern, spectrogram, formants, pitch and intensity from the audio file [5]. We then created a transcript of the audio file in order to get the vowels. This was then placed along with its respective formant, pitch and intensity and time frame [29].

6.4.3 Dealing with Incorrectly Labelled Phonemes

Unfortunately, pronunciation of words are different for different accents. One example is the utterance of the word “tomato”. Some cultures pronounce the word as “/t/ /ow/ /m/ /ah/ /t/ /ow/” and others pronounce it as “/t/ /ow/ /m/ /ey/ /t/ /ow/”. Since we only have all the participants reciting the same thing with a New Zealand accent, we already know what vowels would required to be tweaked in order for it match the pronunciation. As advised by our expert, Fig. 6.6 shows the revised phonemes for some of words recited by the participants.

Word	Phonemes
air	EHR
form	AO
gold	AO
according	AO
arch	AHR
all	AO
old	AO
short	AO
more	AO
organ	AO
years	IYR
beard	IYR
our	AW
answer	AHR

Figure 6.6: Words with changed phonemes.

Chapter 7

Building the System

In this section we look at finally building the system. This involves two steps.

- Train and predict system using the collected data.
- Create a virtual system that reads predicted data

7.1 Train and Predict

We have already discussed the use of supervised learning for training our data. We used regression as a decision tree to predict the output.

We first needed to get our training data ready. We are training the system to recognise the marker data based on the audio data. We aligned the marker data for all the 19 markers against all the vowels extracted from audio data along with the formants, pitch, intensity. Fig. 7.1 shows part of the training data.

We checked our results for two predictions. First by excluding the marker data for one male and one female participant from the training data-set for only 'The grandfather passage, We were able to make a prediction for 'the grandfather passage' for only these two participants. Remember that the system has already been trained with the data collected from eight

F0_Hz	F1	F2	F3	Intensity	Time_s	relative_x_001	relative_y_001	relative_x_002	relative_y_002	relative_x_003	relative_y_003
0	928.8	1498.5	2356.3	28.06235	2.069757	-0.005324781	0.006486714	-0.005505562	0.002270699	-0.00534749	-0.001718536
0	542.6	1404.7	1550	40.12503	2.315091	-0.003778756	-4.46E-05	-0.00423193	-0.000787199	-0.004422069	-0.001988053
190.9793	563.5	1566.3	2200.7	44.09102	2.464424	-0.002762198	-0.001857638	-0.003458381	-0.002376676	-0.003958225	-0.003685653
195.4394	932.6	1580.6	1781.9	50.73953	3.296424	-0.004131377	0.004429549	-0.004426777	0.002905577	-0.004361689	-0.000251919
184.9167	577.7	1601.3	2351.7	47.05686	4.288424	-0.004747272	0.009330839	-0.005131781	0.008081496	-0.005117595	0.004923046
184.9167	577.7	1601.3	2351.7	47.07046	4.299091	-0.004747272	0.009330839	-0.005131781	0.008081496	-0.005117595	0.004923046
176.2645	467.8	1549.4	2648.8	40.12114	5.344424	-0.001744926	0.008183837	-0.002363205	0.00740847	-0.002894998	0.00418736
0	412.1	1523.4	2514.1	38.18932	7.125757	0.000252604	0.002402097	-0.000783741	0.002264202	-0.001590073	-0.000301689
0	412.1	1523.4	2514.1	38.54017	7.136424	0.000252604	0.002402097	-0.000783741	0.002264202	-0.001590073	-0.000301689
194.8198	526	1644	2424.8	41.20196	7.883091	0.002679825	0.00440219	0.001481116	0.003347993	0.000472188	6.89E-05
0	1024.4	2078.4	3144.6	34.86952	8.821757	0.00307554	0.00120756	0.001979053	0.001516968	0.001136959	-0.000634357
204.0012	914.3	1458.5	1995.2	47.50434	9.792424	0.004312873	0.012634218	0.003051996	0.008395761	0.002306044	0.00274089
179.2024	873.1	1533.1	1889.6	45.8851	11.75509	0.002893329	0.007605553	0.001915395	0.006815165	0.001228511	0.003086939
165.6073	838.8	1525.4	2019.9	46.60704	12.72576	0.001275241	0.008943975	0.00041008	0.007914007	-0.000162065	0.004087865
165.6073	838.8	1525.4	2019.9	47.56438	12.73642	0.001275241	0.008943975	0.00041008	0.007914007	-0.000162065	0.004087865
184.6636	940.1	1534.3	1827.3	53.60188	13.57909	0.000472844	0.011030018	-0.000336826	0.008265167	-0.000770271	0.003670782

Figure 7.1: Example of training data.

other participants who recited ‘The grandfather passage’. For the second prediction we predict the marker for The North wind and the sun passage. The system has not been trained for this passage at all. The aim is to see if it is able to predict values to create similar neck motion for a speech that the system has not been trained for.

7.2 Virtual System

For the virtual system, we looked at animating a rigged 3D model with 19 controllers on approximately the same place as the markers on the participants. Fig. 7.2 shows the 3D model in a neutral pose. We read the predicted values and set keys at the time frame when the vowel occurred. At this stage we have only added animation keys for the speech at the vowels. We still need to consider the breaths during speech. While observing the participant’s speech, we noticed that they breathe between sentences and when there is a break during a sentence. In the audio file you can visibly notice breath as it shows as an amplitude larger than complete silence. It is also smaller compared to sound due to speech. This breathing pattern needed to be added to the system. We first located the silence in the audio file. We then calculated the length of the silence and accordingly place keys for silence and breath. For silence the neck motion

would go back to its rest position. The breath was shown by doing the following steps:

- controllers from 1-3, 10-13, 18-19 move slightly down towards the collarbone
- controllers from 4-9, 16-17 move up towards the center of the neck to show the raised collar bone
- controllers 14-15 move slightly down and outwards

Figure 7.2: Neutral pose of 3D female model using Maya [3]

Chapter 8

Results

The participants were filmed using 2 Blackmagic URSA EF cameras (frontal and profile) with a resolution of 1920x1080 at 80 fps [4].

8.1 Program results

We first start with comparing the results of our program with the marker data that was tracked. Fig. 8.1 shows the tracked marker data. Our program has been trained with tracked marker data from all participants for the 'The grandfather passage'. The black line represents the tracked marker data for the passage recited by a female participant. In Fig. 8.2 the red line represents the predicted data for the same participant. One can see that the predicted data is very similar to the marker data. The mid section of the figure shows the amplitude marker data(black line) further away from the predicted data(red line) but the predicted data dips down or raises up just like the original data. The amplitude might not match entirely but they are still relatively similar.

The result shows that the predicted data is similar to the actual marker data and therefore we were able to predict values for both a speech that the program was trained with and a new speech that the program has

never been trained with. This enables us to use the data to realistically animate our model so that the neck looks like it should while speaking

Figure 8.1: Marker results for 'The grandfather' passage.

We now try and predict data for 'The North Wind and the Sun' passage. Our system has not been trained with marker data for this passage at all. The predicted data along with the added breath details as discussed in Chapter 7.2 is used to animate the rigged female 3D model. Our system only sets keys for predicted vowels data and breath data compared to the marker data which sets a key at every frame. A video of the model animated with the predicted values for 'The North Wind and the Sun' has been submitted along with this paper. Fig. 8.3 shows the 3D model pose breathing during a speech. Fig. 8.4 and Fig. 8.5 shows the 3D model during different parts of the speech. In the video one can clearly see the movement around the strap muscle that would have occurred while speaking.

Figure 8.2: Prediction results for 'The grandfather' passage.

Figure 8.3: Neutral pose of 3D female model using Maya [3].

Figure 8.4: Neck pose of 3D female model during speech using Maya [3].

Figure 8.5: Neck pose of 3D female model during speech using Maya [3].

8.2 Key point for Animators

Figure 8.6: Female neutral neck pose.

For realistic animation, an animator's key job is to try and recreate the motion they observe. Consider an example of someone animating a basic human walk. This is something that animators have recreated for a realistic version and for a cartoon version. Extensive study has been done to replicate different types of walks. The posture and walk of a character who is happy will be different from that of a character who is sad. Let's consider a happy walk for a realistic looking character and cartoon character. Both walks would be similar except that the action would be exaggerated for emphasis in the cartoon character. The basic principle of the walk would be applied onto both characters. Similarly, we developed a few basic principles that would help animators animate the neck region. We focused on at describing muscle movement. Fig. 8.6 shows a neutral pose for a female participant.

Figure 8.7: Female neck pose for breath during a speech.

The most visible neck motion while speaking happens when the participant inhales. As shown in Fig. 8.7 when inhaling, the collar bone represented by markers 4-9 along with the Trapezius muscle represented by markers 16-17 tends to raise up. The Sternohyoid muscle represented by 1-3 tends to move down. The Sternocleidomastoid muscle represented by 18-19 tends to move down and inward and forms a triangular dip between them. Omohyoid muscle represented by markers 10-15 move slightly down and outwards. It makes the neck expand a little as you inhale. Female breath during speech is more prominent compared to the male. They tend to raise their collarbones higher than the male participants. So one could always exaggerate the female inhaling motion during a speech.

The strength of a voice comes from air that one exhales. The air stream provides the energy for the vocal folds in the voice box to produce sound. When one speaks you can constantly see 2 repetitive motion. First being a more relaxed pose for the neck as shown in Fig. 8.6. Second, is when you

Figure 8.8: Female neck pose showing stretched strap muscle during a speech.

see the strap muscles stretched a bit as shown in Fig. 8.8

As mentioned earlier, the larynx in the male is positioned lower compared to that of the female. The larynx is constant moving while speaking. One will often notice the larynx moves up and down while speaking. The movement of the larynx is more visible on the male neck due to the adams apple, compared to the female neck. Fig. 8.9 and Fig. 8.10 shows the larynx of male participants during a neutral pose and while speaking, respectively.

Another prominent neck motion is when you can see the omohyoid muscle bulge as seen in Fig. 8.11. This motion happens for barely a second and then returns back to its neutral pose where you don't notice it at all.

Figure 8.9: Male neutral neck pose.

Figure 8.10: Male neck pose during a speech.

Figure 8.11: Female neck pose showing Omohyoid during a speech.

Chapter 9

Limitations and Future Work

In this chapter we are going to discuss the various problems and limitations that were encountered during this research.

9.1 Participants' data limitation

Participants that fit a few criteria, as explained in chapter 6.1 were considered for this study. This results in a very controlled database. Data from participants with wrinkled skin or with a double chin have not been included. Due to timing and scope, a purposeful sample of only 10 participants (five female and five male) were considered for our training database. To train the system more effectively it would be better to have a larger database with more participants. It would also be ideal if the participants recited more passages than just the "The Rainbow passage" and "The Grandfather passage". Training the program with more passages would create a wider range of database which would lead to an accurate prediction.

Jaw movements were completely ignored while creating the virtual animation as the aim was to only concentrate on the neck. The platysma muscle which is placed directly below the skin. It is responsible for pulling down lower face and mouth. Even though the markers have tracked the

motion around the neck region, markers were not placed on the line where the jaw and the neck meet. For future research we may consider tracking the lower Jaw.

As the motion for neck can be erratic for certain parts of the speech. For example while tracking the adam's apple for the male participant. There also have been cases when filming participants with longer jaws where marker 1 would get hidden behind the jaw. In such cases the tracking software would lose the marker and stop tracking or the tracking box would slide a bit away from the marker. This had to be manually corrected and hence it could introduce a slight error in the training data.

Participant were filmed in a neutral environment. We have not looked at the data when one would be expressing emotions such as anger. Through observation we have noticed the difference in some of the motion when one is speaking with a regular tone and when one is screaming.

9.2 Program limitations

One of the limitations of the program occurs during the extraction of pitch and intensity data from the audio file. The data is not available for every time frame. Due to this a zero was added for the missing values.

Through observation we have noticed that the neck motion is more visible during vowels and hence in our current program we have only trained the system for vowels and have used to set keys on vowels.

9.3 Future Work

Our goal for this study was to build a system that recreates the neck motion to add secondary detail to the character. As seen in chapter 8.1, our results show that our system was able to predict values that were used to animate the neck. This could be used as a base for future research in this area.

We would like to continue this study to build a better system. One key step would be to add consonant data. In our current system, the audio analysis data consist of only vowels. By adding the consonant data we will be able to train our program with more data which will work towards creating realistic neck motion.

Another key factor about this study was to build a system that animators could use to create neck animation during speech. Currently, there are a few manual steps during the audio data extraction. The speech transcript has to be manually matched to the speech audio file before we can extract the pitch, intensity and formant data via Darla [29]. In the future, one could work towards making the process completely automated. They could also build a program that would work in real time. This would save time and provide a lot of flexibility to the user.

We work with different characters depending on the what the project requires. It's not necessary that the character would always have a slender neck. Consider the movie BFG [35], where there were two characters with different features. The main giant character, BFG is a thin old male with lots of wrinkles on his neck. On the other hand, the character Fleshlumpeater has a double chin. Both these characters would have subtle differences on their neck motion. Our current system has been trained with data collected from participants with a slender neck. For future research, one could be looking at participants beyond this criteria as it would add to the realism of the animation for characters who could have a double chin or wrinkles on their neck. They could build a program that lets the user add the additional details to their animation depending on their requirement.

Chapter 10

Conclusions

In the last decade or so, a lot of study has been dedicated towards creating realistic animation. Most of this study has been concentrated on the main features like the facial animation of the characters. We have seen that realism of a character is possible if the animation follows the biological motion. Even though the quality of facial animation has vastly improved, there is still some realism lacking. According to us this is mainly related to the secondary animation. The human eye is well trained to differentiate between a real character and a virtual character. Adding realistic secondary animation to the character works towards making the character look even more realistic. To be able to recreate this realistic detail one needs to determine the actual motion created by a character.

For this study we wanted to concentrate on the neck as this secondary motion is an integral part of facial animation. Through basic observation one can notice that there is a lot of motion around the neck region when one speaks. Strap muscles are stretched and relaxed during voice production which is visible on the skin layer. Our approach was to recreate this motion by understanding what goes on under the skin layer. As discussed in chapter 3, we studied the neck anatomy along with the role played by the larynx and strap muscles in voice production.

Our initial problem statement was to determine if we could study motion of the neck while speaking and recreate the neck motion for a virtual model for any given speech. To be able to recreate this realistic neck motion we studied the actual motion around the neck region during speech. As seen in chapter 6.3, participants were filmed with marker strategically placed on their neck above the strap muscles. This was used to track the movement of the strap muscle during speech as shown in chapter 6.4.

In chapter 7.1, we combined the marker data with the speech audio data and trained our system. We were able to predict values for both, a speech that the program was trained with and a new speech that the program have never been trained with. The results for our program shows that the predicted data is similar to the actual marker data. Chapter 7.2 shows how we used this data to animate our model which does make the neck look like it would while speaking.

The video data of the participants speaking was used for observing the motion of the neck region during speech. In chapter 7.2, we observed breath patterns and incorporated this in a system that controls the motion of the virtual system. We have also come up with a list of key animation solutions to make it easy for animators to add secondary motion to the neck. This will help animators to add realistic animation details to the neck region.

We conclude that our program will add realistic details to the neck motion. Beyond this, animators can further add details as suggested in chapter 8.2.

Bibliography

- [1] ALPAYDIN, E. *Introduction to machine learning*. MIT press, 2014.
- [2] ATKINSON, J., AND ERICKSON, D. The function of strap muscles in speech: Pitch lowering or jaw opening. *J. Acoust. Soc. Am.* (1976).
- [3] AUTODESK. Maya 2016. <http://www.autodesk.com/education/free-software/maya>. Accessed: 2016.
- [4] BLACKMAGIC DESIGN. *Blackmagic URSA*, 2014. EF.
- [5] BOERSMA, P., AND WEENINK, D. Praat Version 6.0.19. <http://www.praat.org/>. Accessed: 2016.
- [6] BRADBURY, J. Linear predictive coding. *Unpublished* (2000).
- [7] BREGLER, C., COVELL, M., AND SLANEY, M. Video rewrite: Driving visual speech with audio. 353–360.
- [8] BREIMAN, L. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- [9] CARMELL, T. Spectrogram reading, Mar. 1997.
- [10] DIXON, N. R., AND MARTIN, T. B., Eds. *Automatic Speech and Speaker Recognition*. John Wiley & Sons, Inc., New York, NY, USA, 1979.
- [11] FAIRBANKS, G. *Voice and articulation: drillbook*. Harper & Brothers, 1940.

- [12] FAROOQ, O., AND DATTA, S. Phoneme recognition using wavelet based features. *Information Sciences* 150, 1–2 (2003), 5 – 15. Recent Advances in Soft Computing.
- [13] GENTLEMAN, R., AND CAREY, V. Unsupervised machine learning. In *Bioconductor Case Studies*. Springer, 2008, pp. 137–157.
- [14] GRAY, H. *Anatomy of the human body*. Philadelphia : Lea & Febiger, 1918.
- [15] GRAY, H. *Henry Gray (1821–1865). Anatomy of the Human Body*. @ONLINE, Oct. 1999.
- [16] HALL, J. W., AND MUELLER, H. G. *Audiologists' Desk Reference: Diagnostic audiology principles, procedures, and protocols*, vol. 1. Cengage Learning, 1996.
- [17] HO, T. K. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on* (1995), vol. 1, IEEE, pp. 278–282.
- [18] HO, T. K. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence* 20, 8 (1998), 832–844.
- [19] HODGKINSON, W. Interactive speech-driven facial animation. *J. Acoust. Soc. Am.* (2005), 22.
- [20] KOTSIANTIS, S. B., ZAHARAKIS, I., AND PINTELAS, P. Supervised machine learning: A review of classification techniques, 2007.
- [21] LEY, D. *Vibrant Voice Technique* @ONLINE, May 2015.
- [22] LIAW, A., AND WIENER, M. Classification and regression by randomforest. *R news* 2, 3 (2002), 18–22.

- [23] MCNAMEE, G. *The North Wind and the Sun and Other Fables of Aesop*, vol. 3. Daimon, 2015.
- [24] MICHALSKI, R. S., CARBONELL, J. G., AND MITCHELL, T. M. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.
- [25] MOS. *Mosby's Medical Dictionary*. Elsevier Health Sciences, 2013.
- [26] PANJA, A. Human vocal system and properties, Oct. 2009.
- [27] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [28] PUCKETTE, M. *Acoustics for Musicians and Artists*, Nov. 2014.
- [29] REDDY, S., AND STANFORD, J. DARLA: Completely Automated Vowel Extraction. <http://darla.dartmouth.edu/semi>. Accessed: 2016.
- [30] REILLY, J., AND FISHER, J. L. Sherlock holmes and the strange case of the missing attribution: A historical note on “the grandfather passage”. *Journal of Speech, Language, and Hearing Research* 55, 1 (2012), 84–88.
- [31] ROBERT OSSOFF, STANLEY SHAPSHAY, G. W., AND NETTERVILLE, J. *The Larynx*. Lippincott Williams and Wilkins, 530 Walnut Street, Philadelphia, USA, 2003.
- [32] ROSSUM, G. Python reference manual. Tech. rep., Amsterdam, The Netherlands, The Netherlands, 1995.

- [33] SAFAVIAN, S. R., AND LANDGREBE, D. A survey of decision tree classifier methodology.
- [34] SAKEL, J. *Study skills for linguistics*. Routledge, 2015.
- [35] SPIELBERG, S. The bfg. Walt Disney Studios Motion Pictures, 2016.
- [36] STANDFORD. *Speech Production @ONLINE*, Jan. 1999.
- [37] SVETNIK, V., LIAW, A., TONG, C., CULBERSON, J. C., SHERIDAN, R. P., AND FEUSTON, B. P. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences* 43, 6 (2003), 1947–1958.
- [38] WAN, J. Furious 7. Universal Pictures, 2015.
- [39] WATERS, K. A muscle model for animation three-dimensional facial expression. *SIGGRAPH Comput. Graph.* 21, 4 (Aug. 1987), 17–24.
- [40] WOODSON, G. E. Configuration of the glottis in laryngeal paralysis. i: Clinical study. *The Laryngoscope* 103, 11 (1993), 1227–1234.
- [41] WWW.HEALTHLINE.COM.
- [42] ZSIGA, E. C. *The Sounds of Language : an introduction to phonetics and phonology*. LHoboken : Wiley-Blackwell, 2013.

List of Figures

3.1	Human sound system [26]	8
3.2	Intrinsic muscles: larynx [31]	9
3.3	Intrinsic muscles: larynx [14]	11
3.4	Muscles of the neck. Anterior view. [14]	12
4.1	Wave file from a male participant using Praat [5].	17
4.2	Spectrogram from an audio file for female participant using Praat [5].	18
4.3	Formants represented as red dots from an audio file for a male participant using Praat [5].	19
4.4	Pitch from an audio file for a male participant using Praat [5].	20
4.5	Intensity from an audio file for a male participant using Praat [5].	20
5.1	Decision tree.	23
6.1	Experimenting with marker placement.	28
6.2	Layer 1 shows the approximate position of the markers [41].	31
6.3	Layer 2 shows the muscles under the skin [41].	33
6.4	Layer 3 shows the strap muscles [41].	35
6.5	Layer 4 shows larynx [41].	37
6.6	Words with changed phonemes.	40
7.1	Example of training data.	42

7.2	Neutral pose of 3D female model using Maya [3]	43
8.1	Marker results for 'The grandfather' passage.	46
8.2	Prediction results for 'The grandfather' passage.	47
8.3	Neutral pose of 3D female model using Maya [3].	47
8.4	Neck pose of 3D female model during speech using Maya [3].	48
8.5	Neck pose of 3D female model during speech using Maya [3].	48
8.6	Female neutral neck pose.	49
8.7	Female neck pose for breath during a speech.	50
8.8	Female neck pose showing stretched strap muscle during a speech.	51
8.9	Male neutral neck pose.	52
8.10	Male neck pose during a speech.	52
8.11	Female neck pose showing Omohyoid during a speech. . . .	53