5-30-2017

Facial Capture Lip-Sync

Victoria McGowen
vkm3473@rit.edu

Facial Capture Lip-Sync

by

Victoria McGowen

A Thesis Submitted
in
Partial Fulfillment of the
Requirements for the Degree of
Master of Science
in
Computer Science

Supervised by

Joe Geigel

Department of Computer Science

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

May 30 2017

The thesis “Facial Capture Lip Sync” by Victoria McGowen has been examined and approved by the following Examination Committee:

Joe Geigel
Associate Professor
Thesis Committee Chair

Alejandro Perez Sanchez
Assistant Professor
Thesis Committee Reader

Reynold Bailey
Associate Professor
Thesis Committee Observer

Acknowledgments

I would like to thank my committee and my advisor Joe Geigel. Thank you for taking a chance on me freshman year and approving me for my first CG class. Working together on Farewell to Dawn has been a joy and I will miss it. You also still owe me tatter tots.

I would like to thank Alejandro Perez Sanchez for meeting with me every Wednesday even before this thesis was proposed and creating the 3D assets used. Your insight was beyond helpful and the off topic chats were a well needed break from the stress.

I would also like to thank Daniel Simon for being the motion capture actor for the testing and experiment phase. Thank you for going all the way to the library when you didn't have Internet to sent me the videos during your holiday.

I want to thank my supportive roommates and friends who have watched every rendered iteration of this system, and were my guinea pigs when searching for the best dialogue. Sorry for fueling your nightmares for the past semester.

I would like to thank Cindy Wolfer and Hans-Peter Bischof in the CS office for helping me weasel my way into CS classes and complete my MS in half the time. I know I complicated things and I appreciate your patience during my disasters.

I want to send my sincerest gratitude to everyone who participated in my experimental survey and to those who reached out afterwards interested in the work. I received double the responses expected and I appreciate everyone who forwarded it along.

Finally, I would like to thank my mother for all her support, even when she didn't understand exactly what I was doing or why I was overloading with graduate level courses as an undergraduate student. But look mum, two degrees in five years! I did it!

Abstract

Facial Capture
Lip-Sync

Victoria McGowen

Supervising Professor: Joe Geigel

Facial model lip-sync is a large field of research within the animation industry. The mouth is a complex facial feature to animate, thus multiple techniques have arisen to simplify this process. These techniques, however, can lead to unappealing flat animation that lack full facial expression or eerie over-expressive animations that make the viewer uneasy. This thesis proposes an animation system that produces natural speech movements while conveying facial expression and compares them to previous techniques. This system used a text input of the dialogue to generate a phoneme-to-blend shape map to automate the facial model. An actor was motion captured to record the audio, provide speech motion data, and to directly control the facial expression in the regions of the face other than the mouth. The actor's speech motion and the phoneme-to-blend shape map worked in conjunction to create a final lip-synced animation that viewers compared to phonetic driven animation and animation created with just motion capture. In this comparison, this system's resultant animation was the least favorite, while the dampened motion capture animation gained the most preference.

Contents

List of Figures

Chapter 1

Introduction

With the increase of use of motion capture driven animation in the film and gaming industry, there has been a large number of animated films that have fallen victim to the uncanny valley phenomenon. The 2004 film, The Polar Express, and the 2007 film, Beowulf, are examples of films criticized for the ”dead-eyed” look of the characters [10, 12, 16]. The animation in these films are based on motion capture of human actors but failed to coherently animate the entire face. Individual facial features are targeted in a way that appears as if they are not connected to the rest of the face [10]. For example, a character is animated to smile but no stretching of the surrounding skin occurs. When most of the facial movements register just short of ”human-like”, the viewer can have a hard time determining if the animation is more cartoon-like or realistic, resulting in an uneasy viewing experience.

Even minor offsets in realistic motion and audio can result in unappealing animation. Speech animation on its own is a very complicated form of animation as any misalignment between the dialogue and the character's mouth movements can easily distract the viewer. If the mouth movements in a speech animation create a babbling affect, the character appears more zombie-like than human [32]. Speech animation techniques try to incorporate emotion detection to fully emote the character's face, but can easily fall short as there are no simple rules for human expression.

This thesis proposes a system that will produce realistic facial animation with a simple work flow and for the purpose of speech animation. The bulk of the processing needed for a motion capture animation system can be approximated by combining motion capture data with a speech driven animation technique. This lip-sync technique uses rules within

phonology to predict mouth shapes, or visemes, which, when combined with motion capture, will create a natural speech animation. The motion capture data will directly drive the upper regions of the face and cheeks to convey expression in the character. The aim of this system is to use the combination lip-syncing technique to produce better speech animation than that from just motion capture data or from speech data.

The rest of the paper is organized as follows. Chapter 2 describes the background and necessary vocabulary for the rest of the paper, and a discussion of previous works can be found in Chapter 3. Chapter 4 describes the design and implementation of the whole system in detail. The experiment used to test this version of the system is outlined in chapter 5 and its results are analyzed in chapter 6. Chapter 7 provides further discussion of the results and future work for this system.

Chapter 2

Background & Vocabulary

2.1 Phonology

Phonology is a sub-field of linguistics that focuses on the organization of phonemes. A phoneme is a distinct unit of sound that helps distinguish words from each other. There are approximately 44 phonemes in the English language, 26 for each of the letters in the alphabet and 18 for letter combinations [29]. Phonemes are commonly taught as the distinct sounds between vowels and consonants and are written with /'s surrounding the grapheme (ex. /$\bar{e}$/ for a long e sound). Graphemes are the graphical representation of a phoneme or groups of phonemes. For example, the English alphabet is a grapheme system [11]. The mouth shapes that occur during speech are known as visemes (Figure 2.1).

Figure 2.1: Example Preston Blair visemes used in animation [5]

Applications that convert spoken dialogue to visemes often also require the text of the speech to detect the appropriate phonemes. One such example application is the open-source animation tool Papagayo which converts a given audio file into Preston Blair animation visemes. This tool is used within the CIAS Film and Animation department at

Rochester Institute of Technology for its 2D animation courses. To convert a written text to its visemes, a grapheme-to-speech conversion is needed to first find the phonemes. These phonemes are saved to file, and then a mapping between phonemes and visemes can be done using a look-up-table. These systems are best when just providing the phonemes or visemes from the dialogue. There is no system available to accurately provide phonemes or visemes and their time of occurrence in an audio file at the moment. The systems that are used currently require a large data base for machine learning and still need a user to verify the results.

There is also no simple one-to-one conversion between visemes and phonemes as several phonemes have the same facial movement associated to them (i.e /t/ and /ch/). This is made more difficult as the current phoneme depends on the phonemes that occur before and after it. There are many published viseme to phoneme look-up-tables available today with minor variations from table to table. One such example of a phoneme to viseme LUT can be found in Figure 2.2. Each of the LUTs in current publishings base their mappings on different techniques, such as visual mouth movements, speech rules in linguistic, and a combination of the two from machine learning recognition algorithms ([8]).

Viseme	Visibility Rank	Occurrence [%]	TIMIT Phonemes
/A	1	3.15	/f/ /v/
/B	2	15.49	/er/ /ow/ /r/ /q/ /w/ /uh/ /uw/ /axr/ /ux/
/C	3	5.88	/b/ /p/ /m/ /em/
/D	4	.70	/aw/
/E	5	2.90	/dh/ /th/
/F	6	1.20	/ch/ /jh/ /sh/ /zh/
/G	7	1.81	/oy/ /ao/
/H	8	4.36	/s/ /z/
/I	9	31.46	/aa/ /ae/ /ah/ /ay/ /eh/ /ey/ /ih/ /iy/ /y/ /ao/ /ax-h/ /ax/ /ix/
/J	10	21.10	/d/ /l/ /n/ /t/ /el/ /nx/ /en/ /dx/
/K	11	4.84	/g/ /k/ /ng/ /eng/
/S	-	-	/sil/

Figure 2.2: Example of a phoneme to viseme map [22]

2.2 Phonetic Alphabets

There are many alphabetic systems for phonetic transcription. The most common systems are that of the International Phonetic Alphabet (IPA) and the Speech Assessment Methods Phonetic Alphabet (SAMPA). IPA is based on the Latin alphabet and was created as phonetic transcription standard between Britain and France [21]. The IPA alphabet consists of letters and diacritics to represent specific sounds in a language. This also carries over into accents within languages. For example, British English and American English IPA symbols differ slightly. One such instance is the representation of the sound of a R at the end of a word, as R is not as pronounced in British English as it is in American English. IPA can be used to transcribe over 30 different languages and the regional accents within them [21].

SAMPA is essentially the computer friendly version of IPA. It was developed by the European Strategic Program on Research in Information Technology in the late 1980s and it uses 7-bit ASCII notation [30]. As some IPA notation symbols do not transcribe well in ASCII, other signs not used by IPA were adapted, such as ”9” for the vowel sound in the French word ”neuf” (which so happens to mean nine). More examples of key differences between IPA and SAMPA can be seen in Fig 2.3.

Arpabet is another example of a phonetic transcription alphabet that was encountered during this project. This alphabet was created by the Advanced Research Projects Agency in the 1970s and it is only used to represent phonetics in American English through ASCII characters [23]. Its use of ASCII characters makes it similar to SAMPA as it can be used as a map to digitally transcribe IPA symbols for American English. To represent a sound, letters or pairs of letters are used followed by a number, which is used as a stress indicator. Normal punctuation is included as a place holder for ”end of phrase” or ”end of sentence” notation. Conversion between IPA and Arpabet can be found in Fig 2.4.

IPA	ASCII	examples
ʌ	^	cup, luck
ɑː	a:	arm, father
æ	@	cat, black
ə	..	away, cinema
e	e	met, bed
ɜːʳ	e:(r)	turn, learn
ɪ	i	hit, sitting
iː	i:	see, heat
ɒ	o	hot, rock
ɔː	o:	call, four
ʊ	u	put, could
uː	u:	blue, food
aɪ	ai	five, eye
aʊ	au	now, out
oʊ/əʊ	Ou	go, home
eəʳ	e..(r)	where, air
eɪ	ei	say, eight
ɪəʳ	i..(r)	near, here
oɪ	oi	boy, join
ʊəʳ	u..(r)	pure, tourist

Figure 2.3: Difference between IPA and SAMPA for special characters [20]

2.3 Facial Animation Techniques

Computer graphics techniques used to animate the face differ from the full body animation as the face is a complex system of muscles. The oldest technique is that of key-framing specific facial features. This involves the animator positioning the facial features at certain frame intervals and interpolating the movement between the chosen frames to give the appearance of motion. For facial motion, key-framing can lead to inconsistent movements as there is no way to have an animator consistently place a facial feature in the exact position during every similar expression. For speech movements, key-framing can be very tedious to go back and edit individually if the character appears slightly off sync. Key-framing facial features also comes with the risk of constantly manipulating the neutral facial model, which decreases the integrity of the character.

A more commonly adapted technique for facial animation is using blend shapes. Blend shapes are copies of the a facial model, and each copy demonstrates a different facial expression or extreme facial feature movement, such as a raised eyebrow or a smile [9]. If the animator knows the expressions they want to use for their animation, they only need to create the corresponding blend shapes. These blend shapes are then weighted together

IPA Symbol	ARPAbet Symbol	Word	IPA Transcription	ARPAbet Transcription
[i]	[iy]	lily	['lɪli]	[l ih l iy]
[ɪ]	[ih]	lily	['lɪli]	[l ih l iy]
[eɪ]	[ey]	daisy	['deɪzi]	[d ey z i]
[ɛ]	[eh]	poinsettia	[pɔɪn'sɛriə]	[p oy n s eh dx iy ax]
[æ]	[ae]	aster	['æstɚ]	[ae s t axr]
[ɑ]	[aa]	poppy	['pɑpi]	[p aa p i]
[ɔ]	[ao]	orchid	['ɔrkɪd]	[ao r k ix d]
[ʊ]	[uh]	woodruff	['wʊdrʌf]	[w uh d r ah f]
[oʊ]	[ow]	lotus	['loʊɾəs]	[l ow dx ax s]
[u]	[uw]	tulip	['tulɪp]	[t uw l ix p]
[ʌ]	[uh]	buttercup	['bʌɾɚˌkʌp]	[b uh dx axr k uh p]
[ɝ]	[er]	bird	['bɝd]	[b er d]
[aɪ]	[ay]	iris	['aɪrɪs]	[ay r ix s]
[aʊ]	[aw]	sunflower	['sʌnflaʊɚ]	[s ah n f l aw axr]
[ɔɪ]	[oy]	poinsettia	[pɔɪn'sɛriə]	[p oy n s eh dx iy ax]
[ju]	[y uw]	feverfew	[fivɚfju]	[f iy v axr f y u]
[ə]	[ax]	woodruff	['wʊdrəf]	[w uh d r ax f]
[ɨ]	[ix]	tulip	['tulɪp]	[t uw l ix p]
[ɚ]	[axr]	heather	['hɛðɚ]	[h eh dh axr]
[ʉ]	[ux]	dude[1]	[dʉd]	[d ux d]

Figure 2.4: Difference between vowel sounds in IPA and Arpabet [23]

to create full-face expressions, and key-framed to create the animation. Blend shapes can be created to match viseme expressions for speaking animation as well, making dialogue animation easier. The movement between key-frames occur with the interpolated change between the key-framed sets of blend shapes. While blend shapes seem time consuming initially, once all desired blend shapes for the model are made, animation can be done quickly and consistently with no damage to the base neutral facial model.

As the face has many different features, audio driven animation has been used to animate just the mouth. The work flow requires the user to provide a face model with mouth blend shapes matching the visemes recognized by the system [4]. An audio file is then provided by the user for speech analysis. The system outputs the mouth movements in time with the audio to animate the character. While audio driven systems produce realistic lip-sync, further animation still needs to be done to give the appearance of expression as this technique is not a full facial animation technique. This speech analysis is still rudimentary as speech analysis techniques currently require a lot of training and need to be checked by a user. Thus incorrectly recognized visemes have a high chance of occurring.

2.4 Facial Motion Capture

Motion capture is another very common technique for achieving realistic facial computer animation. Capture systems are divided into two categories, marker and marker-less systems. Marker based systems provide higher quality capture, but come at a higher price point. For these types of systems, the actor wears a set of reflective facial markers, or dots, that are tracked during the performance. As these dots are reflective, the actor's face also needs to be illuminated evenly for an optimal capture, thus recordings occur in front of a light panel or with a head mounted rig with LEDs surrounding the recording camera. The number of markers needed depends on the system and the location of the markers can depend on the action being done by the actor [24]. There are two sets of marker classes, ones used to detect the movement of the head and another class used to detect the facial expressions. Due to markers needing to be re-positioned at the same specified points on the face, this type of system can take a while to set up and is cumbersome when having to re-shoot takes.

Marker-less systems do not rely on the actor wearing dots on their face during the capture. Instead, camera rigs with RGB cameras are put on the actor during recording. As even illumination on the face is also needed for this type of system, LED lights are often added surrounding the camera. The video stream is processed using computer vision algorithms to track key features in the face, such as the nostrils, lip corners, eyes, and eyebrows [24]. Since the recording device is a simple RGB camera, depth information of the actor's face is lost. There are also versions of markerless systems that use RGB-Depth cameras to fix this lack of depth issue. One such system is Faceshift. Faceshift relies on a RGB-Depth camera to track the user's expressions and head orientation. The user trains the software by scanning their face while making several extreme expressions, essentially creating blend shapes models [34]. While tracking, the system uses the scanned blend shape models to drive a generic facial model's corresponding blend shape weights. The blend shape weights are displayed in histogram form and the information can be used to animate a user provided model in real time. In relying more-so on the depth information of

the face, Faceshift is restricted to just the large facial features, and is unable to track the eye gaze and other more subtle facial movements [34]. Overall, markerless systems are much more affordable and more accessible to independent animators and small research groups, but are limited by the consumer level equipment.

Cameras used in either marker or marker-less systems are best if capable of at least 60 fps, as the higher frame rate allows detection of small movements in the eyes and lips. Every system type has their own mapping scheme when applying the data to a character model, so animation workflows between systems can have some variations.

2.5 Uncanny Valley

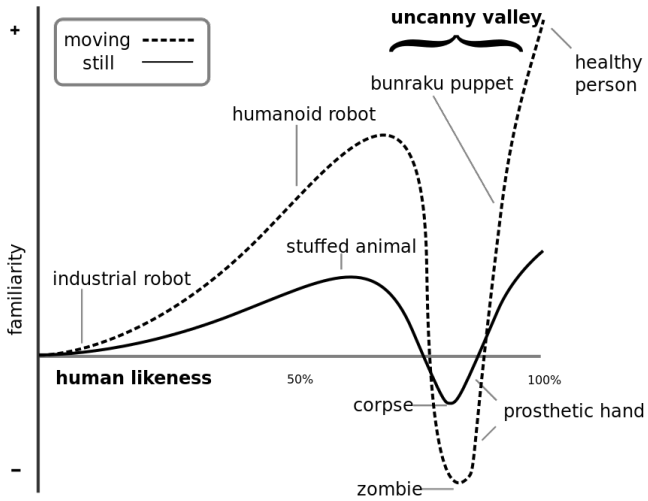


Figure 2.5: Simplified version of the human emotion response presented by Masahiro Mori [26]

The Uncanny Valley is a phenomenon where a computer character or a robot appears nearly human but causes a sense of unease when viewing it [26]. The phrase was first coined by Japanese roboticist Masahiro Mori in 1970 when he noticed a trend as humanoid robot design started to focus more on realism. Mori's observed relationship between human emotional response and human likeness can be seen below in Figure 2.5. An example Mori provides of something that would cause this uncanny feeling would be shaking a prosthetic

hand. Prosthetics are becoming more realistic with skin like material and details such as false finger nails. Because of this, at a glance, individuals may not notice anything off putting, but the act of shaking a realistic hand and finding that it is actually cold and made out of synthetic materials would cause the person to lose the sense of familiarity [26].

The Uncanny Valley has been a topic within the computer graphics field ever since Mori's publication of the theory as modern technology has enabled computer generated characters to move and look more human-like. In facial animation, this phenomenon can occur when part of the animated face does not react or move with the rest of the facial expression. One example would be a poor animation of the eyes, which could created a glossed over effect. There has been research as to what facial features provoke the greatest discomfort responses in viewers. Dill et al. surveyed over 200 people asking about their level of comfort looking at still images and videos of different characters with varying human likeness [14]. For videos, the regions of the face reported to be the most provoking were the eyes and the mouth, with the mouth region being reported the most, accounting for 37.85% of the total responses [14]. In Tinwell et al. uncanny valley in dialogue synchronization and human likeness in voice was tested [33]. It was found that the human likeness in the voice and the comfort of viewing the model speaking were related. The more familiar the character appeared, the more comfortable the audio sounded to the viewer. When the dialogue was said at a different speed or pitch than expected, or it was out of sync, viewers were more likely to report discomfort [33].

Chapter 3

Related Work

All text-to-speech applications today use a grapheme-to-phoneme conversion. The CMU Pronouncing Dictionary is a command line package that converts graphemes from text files or command line strings into Arpabet transcribed phonemes [25]. As this tool uses Arpabet, it works best for dialogues with standard American vocabulary and North American English pronunciations. This dictionary features 39 phonemes, and knows over 134000 words and their pronunciations. Using this literal dictionary approach is very common in text-to-speech programs and is effective for one language specific applications if the text does not contain made-up or foreign words. An example of a linguistic rule based grapheme-to-phoneme system is Google's text-to-speech. This system produces the appropriate phonemes depending on the pronunciation rules for the given language and through deep learning [17]. This type of approach requires large databases and frequent use, but are able to guess pronunciations of unknown words quiet well. For a simple application with a single language, rule based grapheme-to-phoneme systems can be excessive, but are less likely to break over an obscure word.

Phoneme-to-visemes conversion tables have been available within the linguistic field since the 1970s. In 1971, Janet Jeffers and Margaret Barley released a viseme-to-phoneme map with 11 visemes mapped to 43 phonemes and a silent viseme [22]. Their findings were entirely based on linguistics and their mapping merged two phonemes, /hh/ and /hv/, as both phonemes have, arguably, the same mouth shape. This trend of combining /hh/ and /hv/ has continued in later maps. In 2000, Neti et al. combined linguistic findings with a decision tree based on IBM's ViaVoice database [27]. This map also consisted of 43 phonemes but

divided into 12 viseme classes and a silent viseme. In 2004, an entirely database driven approach was created [18]. This map has 52 phonemes mapped to 14 viseme classes with a silent viseme. The mapping was done through analyzing visual features in frames, so there are some reported inconsistencies. All of these approaches are a many-to-one map, meaning that some of the proposed phonemes are mapped to the same viseme class. Each phoneme-to-viseme map available has been tailored to the developers task, but the 1971 conversion map by Jeffers and Barley has been a favorite for years [8].

Previous research for lip-sync animation have used motion capture data to teach speech models articulation rules to then animate a 3D facial model [7, 13]. This technique requires large data sets of audio and video information recorded through marker based capture systems, and were not able to be performed in real-time. The motion capture data still needed to be reviewed and cleaned before running the data through the system to ensure a generally appealing final animation [13]. Due to the large amount of post-processing and the tedious set ups, this approach to lip-sync animation is not easily accessible but works very well for large studio productions.

A team at Disney Research also used a video-based learning approach and developed a system that dynamically created visemes [31]. This system had a final total of 150 visemes and were mapped to phonemes in a graph-like structure. This system takes co-articulation into account, so visemes are recycled over similar phonemes. As the visemes are dynamically created, the proposed advantage of this process is the compatibility among differently shaped facial models [31]. While this might have worked for a large scale studio that handles a large number of different types of facial models, having a system that creates 150 visemes for a language such as English, with 44 agreed phonemes, is beyond excessive. There is no need to work with 150 visemes for a single model speaking one language when 11-13 visemes would work.

Video trained lip-sync animation has been used to ”rewrite” video footage to make the speaker appear to be saying words they were not recorded saying [6]. Computer vision was used on the original footage to track the facial movements, the mouth shapes were matched

to visemes, and then were synced to the new audio with an video processing algorithm. Processing was used on intermediate frames to smooth the changed facial movements. This technique has been used in Hollywood by recycling old footage of historical persons and making them appear to be speaking in sync with the film's dialogue.

Chapter 4

Design & Implementation

4.1 System Design

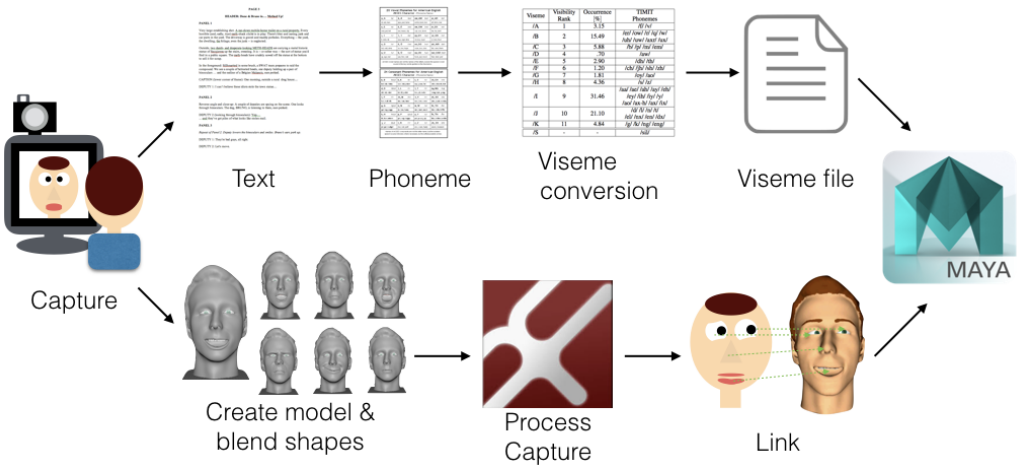


Figure 4.1: Overview of system workflow. The upper branch describes the grapheme-to-viseme portion and the lower branch describes the model and capture portion of the system. The viseme and capture data will be combined in Autodesk Maya to achieve the speech animation.

An outline of the system work flow can be seen in Figure 4.1. To achieve a smooth animation with motion capture and viseme-to-blend shape animation, a text dialogue is provided to complete the grapheme-to-viseme part of the system. This dialogue consists of words found in the American English language as the grapheme-to-phoneme conversion is a dictionary based system. A dictionary conversion was used because of its simplicity and this system is only being tested with one language, American English. A phoneme-to-viseme look-up-table was written in Python to match the linguistic map used. The visemes

are saved out to file to be read by Autodesk Maya during the final animation phase. The grapheme-to-viseme steps previously described are outlined on the upper branch of the diagram.

In accordance to the lower branch in the workflow in Figure 4.1, an actor is recorded saying and acting the text. The motion capture is recorded using a consumer level RGB camera with capabilities of up to 120 fps. The capture data was processed using a facial capture program, and the processed xml files would be used later in the workflow. A humanoid face model is the target model type for this system as it allows for easier conversion between capture data and model. Blend shapes for the test model were made to match the visemes used in the viseme mapping along with blend shapes for the upper region of the face.

With the recordings and model completed, the motion capture data was linked to the face model in Autodesk Maya. The capture data and the visemes-blend shape file are used together to create a lip-synced animation. By predetermining the order of the blend shapes from the viseme-blend shape file, the system knows what the actor is supposed to be mouthing. This also helped result in a smooth and more recognizable mouth shape even if imperfections occurred during capture, such as the actor mumbling or if the camera is of very poor quality or lighting changes. The capture data from the actor drove the upper features of the face with the additional blend shapes to give the character a full-face animation. The resulting animation was exported as an image sequence and needed to have the actor's audio added once combined into a video file.

4.2 Dialogue & Phoneme-to-Viseme Mapping

To help test this system, the famous poem *O Captain! My Captain!* by Walt Whitman was chosen [28]. The reason behind this choice is familiarity. Students from North America are often taught this poem in school as it is about the death of Abraham Lincoln, and a large audience are also familiar with this poem due to its claim to fame in the 1989 film *Dead Poets Society* with Robin Williams. The 3D model chosen also looks the part of a poet, and

a test pool of viewers said the dialogue did not seem obscure being spoken by the model. A portion of the poem was chosen because it featured all of the visemes in the viseme map at least once in a single stanza (Table 4.1). The portion of the poem chosen for testing the system is as follows:

O Captain! my Captain! our fearful trip is done,

The ship has weatherd every rack, the prize we sought is won,

The port is near, the bells I hear, the people all exulting,

While follow eyes the steady keel, the vessel grim and daring;

But O heart! heart! heart!

O the bleeding drops of red,

Where on the deck my Captain lies,

Fallen cold and dead.

Table 4.1: Number of occurances of each viseme in the poem.

Viseme	# Occurance
F/V	7
ER	30
BMP	16
AW	1
TH	10
CH	1
OY	3
S	13
EH	68
G	48
NG	12
Neutral	78

To convert the dialogue into phonetic descriptions, an open sourced speech synthesizer application, eSpeak, was used on the dialogue file. eSpeak is a compact application developed and released in the mid-90s with support for over 50 languages [15]. eSpeak translates command line text input to the International Phonetic Alphabet (IPA) along with notation

for stressed and unstressed syllables and pauses. The language for transcription was picked to be American English and the stress notations were ignored during transcription.

A Python file phoneme2viseme.py was used to interact with the eSpeak command line application and to convert the resultant IPA notation to visemes. For this project, the phoneme-to-viseme map proposed by Janet Jeffers and Margaret Barley was used as it is still one of the highest ranked maps for accuracy (Figure 2.2) [22, 8]. The Python standard library subprocess was used to call eSpeak on the dialogue file and the IPA notation was saved as a massive string to a variable. Through string parsing and the use of a predetermined dictionary of viseme definitions, the viseme map was written to file.

It was found that having the timing of each word in the dialogue for the chosen audio file produced the best distribution of visemes. The audio was processed using IBM's Watson Text-to-Speech online application. Watson's Speech-to-Text application provides word recognition and approximate timings of that word found in a .wav file [19]. Some of the timings were edited by hand as the application is still in its demo phase and it provided a wide time range for words said quickly. The time range was divided amongst the word's phonemes and these time stamps were added to the saved viseme mapped file.

4.3 3D Assets & Viseme-to-Blend Shape Mapping

A male humanoid character model was created and rigged by Professor Alejandro Perez Sanchez from the 3D Digital Design department in CIAS. For the model, Professor Sanchez made 11 blend shapes with the same mouth shapes of the 11 visemes in the Jeffers and Barley map (Fig 4.2). Four blend shapes for each of the eyes (looking left, looking right, wide, and blinking), four for each eyebrow (left, right, raised and lowered) were also created to give control to the upper facial features. A few extra mouth blend shapes were added for shapes such as ”puff”, ”pucker”, and ”mouth press” to give more life to the character when controlled by motion capture. Examples of some of these extra blend shapes can be seen in Figure 4.3.

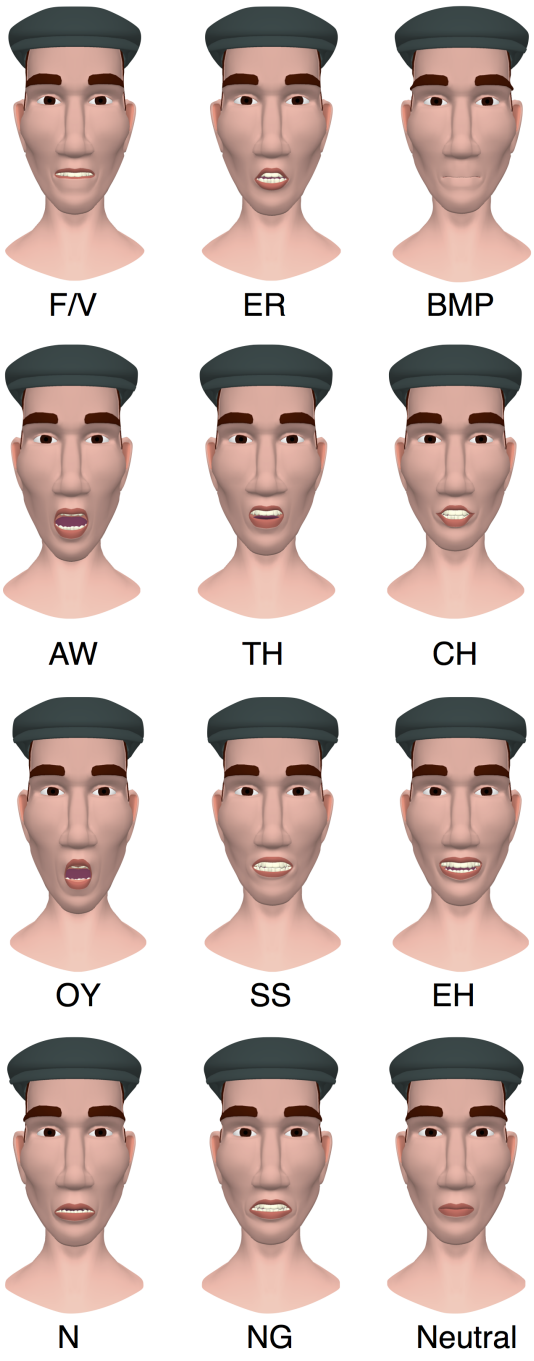


Figure 4.2: Blend shapes specific for the phonemes in the Jeffers and Barley map.

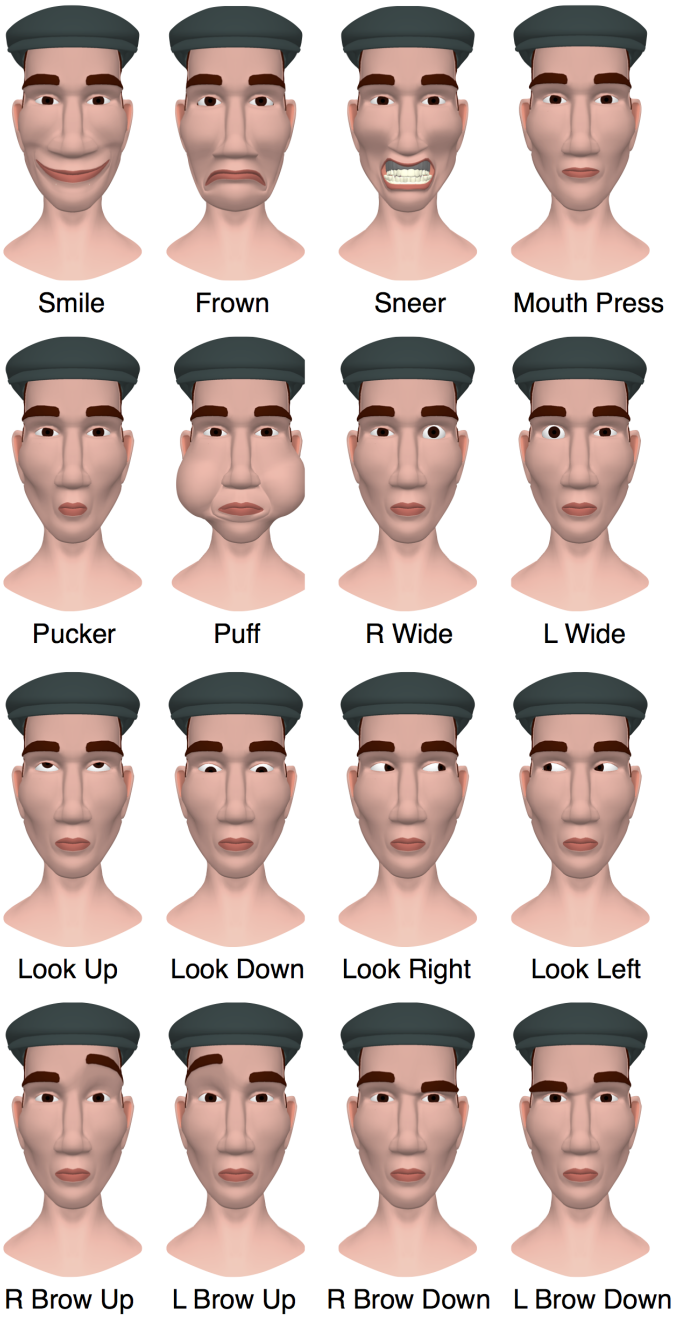


Figure 4.3: Blend shapes for the rest of the face.

The viseme-to-blend shape mapping was handled by Autodesk Maya through the python

console. As there are 11 visemes in the Jeffers map and 11 viseme blend shapes, the mapping from the viseme files was almost one-to-one. The Python file viseme2blendshape.py was used to read in the viseme mapped file and change the model to match the current viseme using the model's provided controllers. To access the blend shapes for the visemes, the Python code manipulated the attributes of the mouth controller on the model's rig (Fig 4.4). As the blend shapes can be make the model look rather extreme in their 100% state (as shown in Figure 4.3 for the Puff, Smile, and Frown blend shapes), the highest percentage used on a blend shape attribute was 80%, with most blend shape attributes being set to around 70%. Some of the viseme mouth shapes were more complex than others, thus some of the extra mouth blend shapes were used in addition to that specific viseme blend shape. These additional mouth blend shapes' attributes were edited as needed. Once the attributes were changed for the new viseme, the mouth controller was key framed for the given time stamp.

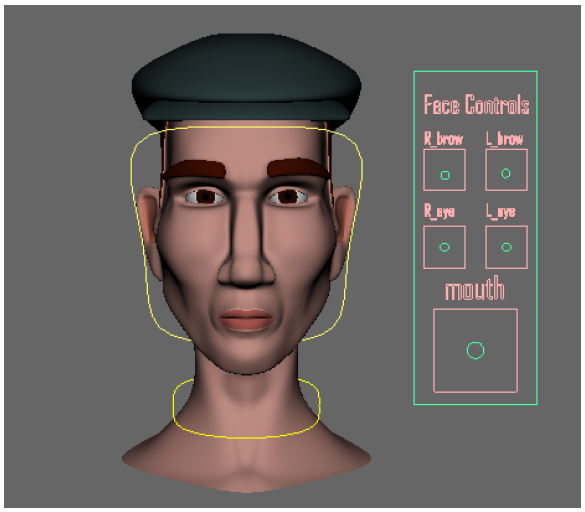


Figure 4.4: Character model with rig in Autodesk Maya Scene.

4.4 Motion Capture & Final Animation

For the capture system, Faceware Tech's personal edition of Analyzer and Retargeter were used [2]. Faceware's Analyzer is developed using Matlab and requires a video file input

[1]. The video is then dissected into frames, and the actor's facial features are tracked and visually outlined. The facial features tracked by Analyzer include the eyes, eyebrows, pupils, nostrils, and lips. Faceware's Retargeter is a plugin available for Autodesk Maya and Autodesk MotionBuilder [3]. Retargeter takes the data exported from Analyzer and connects the tracked facial features to an input CG model through a character setup process for animation. The character setup process can be seen in Figure 4.5.

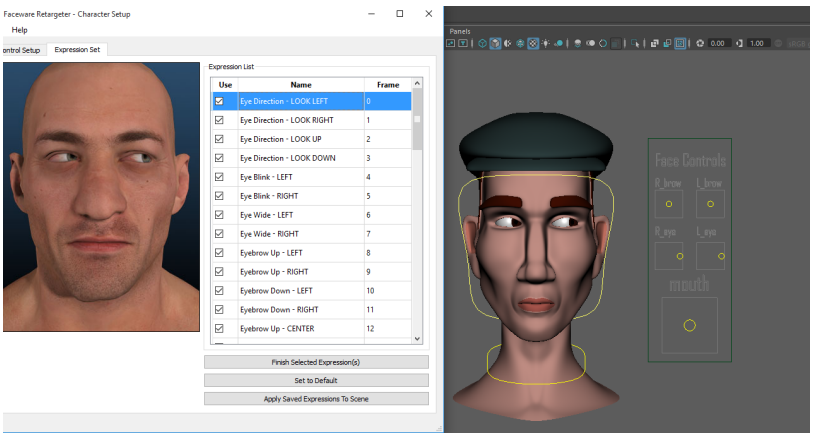


Figure 4.5: Faceware's Retargeter character setup menu.

This version of Faceware's capture suite does not work in real-time, and it allows the user to only have access to their "Auto Track" features. The workflow for the Faceware Tech suite used is outlined in Figure 4.7. After recording, the videos of the actor were first processed through Analyzer and were given a neutral frame to better the tracking results (Fig 4.6). An xml file listing the facial feature placements per frame was created for use within Retargeter. Retargeter is a Maya plug-in that works directly with the model within the project scene. Once the model's rig controls were connected to Retargeter, an "Auto Solve" option was used to allow Retargeter to key frame the full face of the model for the duration of the capture video.

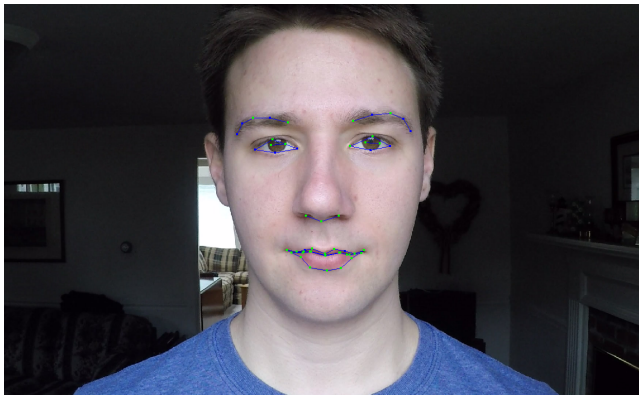


Figure 4.6: Actor's facial features outlined to create a neutral frame in Faceware's Analyzer.

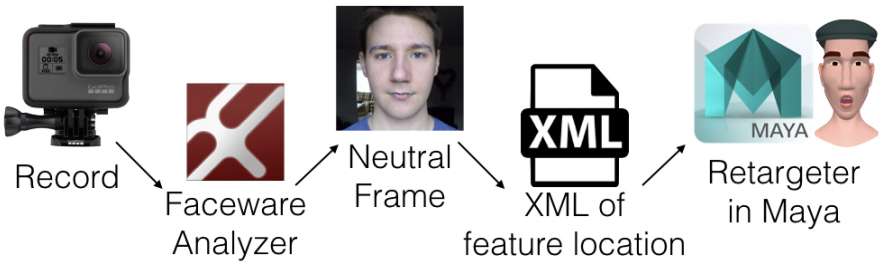


Figure 4.7: Faceware workflow.

As Regtargeter key frames at every frame in the video, the key frames that were deemed unnecessary (those not matching the mapping time stamp) were deleted. These extra key frames produced a lot of noise in the animation. The values for the viseme blend shape attributes were combined with the values of the attributes edited by Retargeter to create a combination of linguistic based and motion capture influenced mouth movements. This keeps the model life-like movements in the lips, but slightly accentuates the mouth shape to that of the desired viseme. The animation was then rendered out as an image sequence. The image sequence was combined into a .mp4 video file using the command line tool ffmpeg and the audio was added to create the final animation video.

Chapter 5

Experiment

An IRB approved subjective comparison survey was made open to faculty and students to provide their feedback on this lip-sync technique. A single actor was used to record the chosen dialogue in one take. The selected actor was male to better match with the test male character model. Recordings were done with a GoPro Hero 5 and at frame rates of 24, 30, 60, and 120 fps. The videos were about 30 seconds longs and the audio was extracted using QuickTime for the animations.

The animations used for the comparison survey were based off of the 60fps video and were animated as follows:

1. Motion capture technique
2. Phonetic mapping technique
3. Proposed combination technique

These animations used the same character model with the same set of blend shapes. The break down of the viseme blend shapes used in each animation technique can be seen in Table 5.1. These comparison recordings looked exactly the same aesthetically. The background of the animations were a neutral grey as well as to not cause any visual distractions from the lip-sync. The videos only featured the character model in the very center of the scene. An example of all three videos at the same frame used for the survey can be seen in Figure 5.1. As the version of Faceware used only allowed for their auto features, the capture from all videos actually missed the actor's upper lip. As facial pose training was

not an enabled option, some manual manipulation was performed on the resultant motion capture influenced animations through Maya's Graph Editor. The only fix done was adding a "dampening" filter on the attributes for the mouth to shrink the negative and positive peaks to a more contained range, between -0.5 and +0.6. These values were determined subjectively for the test animations.

The survey was hosted through Google Forms and the videos were hosted through Youtube (Appendix D). Close captioning was provided for each video to allow the user to directly compare the mouth shapes to the movement expected for the dialogue. The videos were shown in a random order to the viewer to avoid a bias, and they were named Dan, Sam or Bob to avoid numbering. Sam was the motion capture animation, Dan was the phonetically mapped animation, and Bob was the combined animation. The user's age, frequency of watching animated films, frequency of watching animated television shows, previous experience working on animations, and length in years of animation work were asked to see if there was a connection between animation expertise and facial animation preferences.

Table 5.1: List of the visemes present in each animation technique.

Viseme	Motion Capture	Phonetic Map	Combined
F/V	X	X	X
ER		X	X
BMP	X	X	X
AW		X	X
TH		X	X
CH	X	X	X
OY	X	X	X
S		X	X
EH		X	X
G		X	X
NG		X	X
Neutral	X	X	X

(a) Phonetic Mapping - Dan

(b) Motion Capture Only - Sam

(c) Current System - Bob

Figure 5.1: Example of the three different techniques compared in the survey. The still is from frame 2092 in the animation.

Chapter 6

Results

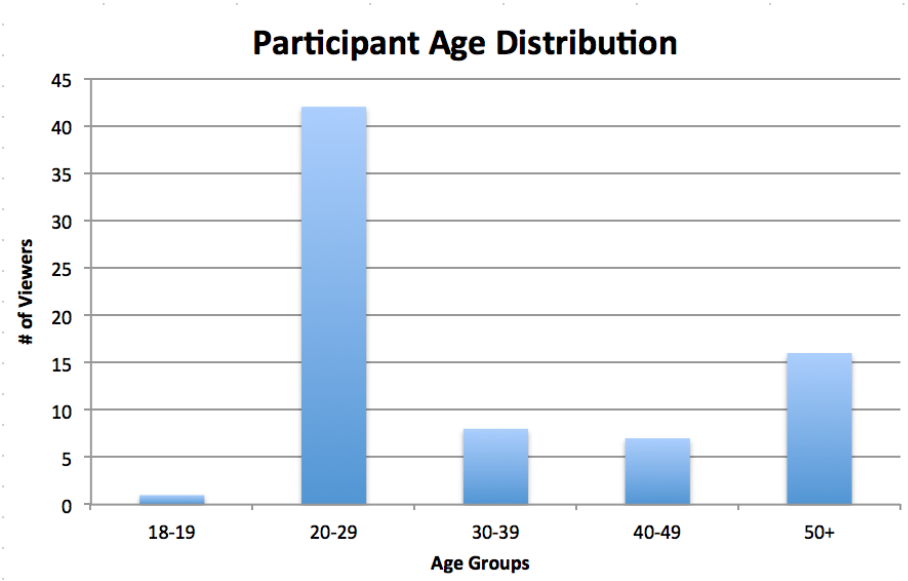


Figure 6.1: Age distribution of survey participants.

A total of 74 individuals partook in the survey. The survey was advertised on Facebook, through emails, and through the 3D Digital Design Facebook community group. There was at least one individual from every age group and the age group distribution can be seen in Figure 6.3. As to be predicted, the largest age group is that of the 20-29 year olds with 42 participants, as the survey was targeted towards college students in computer graphics and in 3D digital design.

The results from the survey show that the combined technique did not gain the most preference (Fig 6.2). It actually was the least preferred out of the three videos. Interestingly enough, the motion capture animated video was the top choice, followed by the phonetically mapped video. This could be because, while the motion capture technique was the one technique that did not have the most sharp or clear mouth shapes, the mouth moved with the most precise timing with the audio. The motion capture technique's more neutral mouth movements also better mimicked real life as people do not purposefully enunciate during casual conversation. However, that same reason could have caused the opposite effect as animation usually features more dramatic expression. The phonetic technique produced the most clear mouth shapes and had the least amount of noise in the animation. This technique produces the closest to traditional lip-sync animation and would thus appear the most familiar to viewers. This would explain why it was the second highest ranked choice with 22 viewers selecting it.

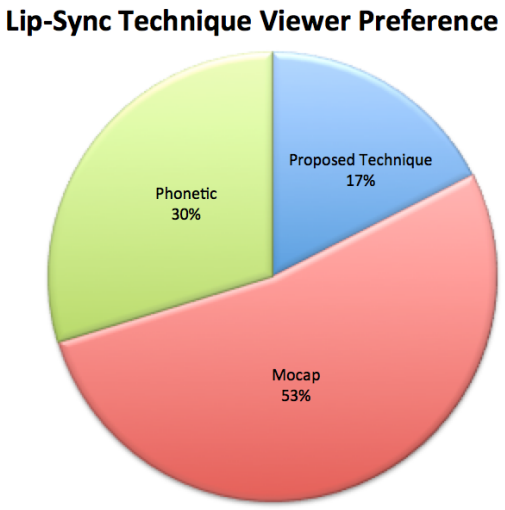
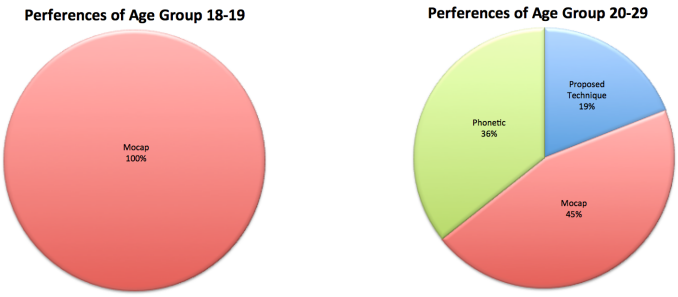
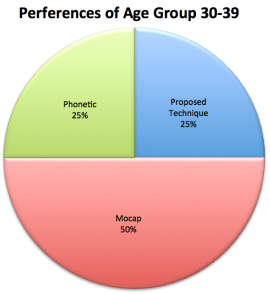


Figure 6.2: Lip Sync Technique Preference Final Results.

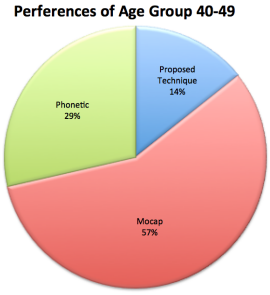


(a) 18-19 Years Old

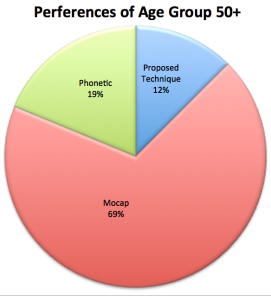
(b) 20-29 Years Old



(c) 30-39 Years Old



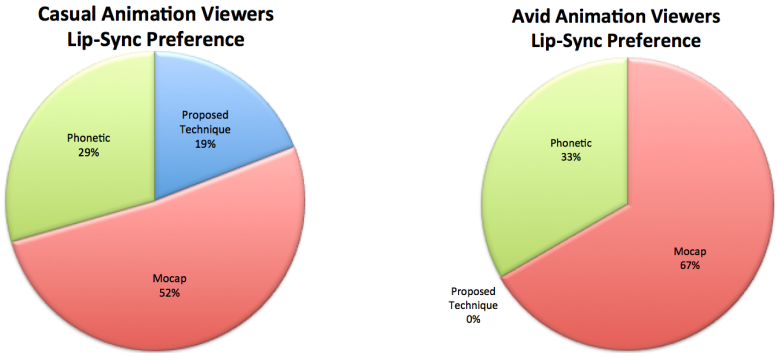
(d) 40-49 Years Old



(e) 50+ Years Old

Figure 6.3: Preference of animation technique between age groups.

When dividing the responses into their corresponding age groups, the trend of the motion capture animation being the top preference followed by the phonetically mapped animation continued. There was only one participant who was 18-19 years of age, thus the 18-19 year chart in Fig 6.3a cannot provide any conclusive answers about that age group, although it still follows the preference trend.

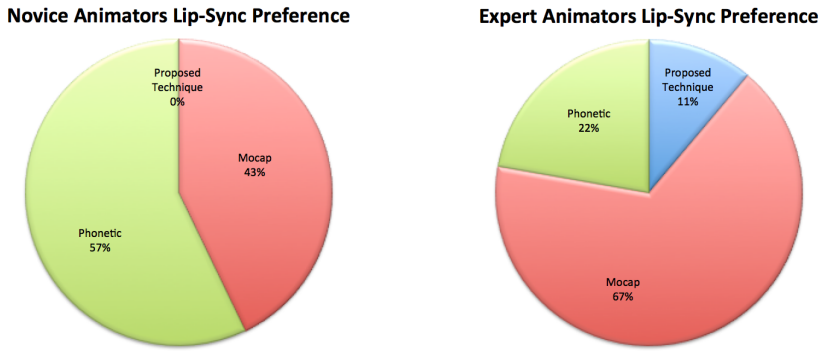


(a) Casual Viewers (b) Avid Viewers

Figure 6.4: Preference of animation technique between casual and avid animated movie and TV show viewers.

Further division amongst participants was done to see if a differing trend occurred. Casual and avid animation viewers were determined by the participant's answers about their animated film and television show viewing habits. Those who answered "Occasionally" or "Once a month or more" to the movie and television show questions were deemed casual viewers while those who answered "Once a week or more" or "Daily" to were categorized as avid viewers. There were six participants whom fit the avid animation viewer criteria and 67 casual viewers. As seen in Figure 6.4, both viewer groups showed the same trends as the collective participation group with the motion capture technique being the highest rated followed by the phonetic mapping technique. It is interesting to note that none of the avid viewers preferred the combined technique.

Expertise among the participants was determined by the question asking whether they had worked on an animation or on part of an animation work flow, and the duration in years of said experience. To those who answered "yes" to having prior experience, the number of years categorized them as either expert or novice. Experts were those with 4+ years and novices were those with 1-3 years of experience. These year ranges are biased to deem professors and individuals in the industry as experts and any students as novices. There were seven participants who were identified as novice animators and nine who were identified as



(a) Novice Animators (b) Expert Animators

Figure 6.5: Preference of animation technique between novice and expert animators.

experts. From the percent preference figures in Figure 6.5, it can be seen that novice animators prefer the phonetically mapped animation over the motion capture animation. With the assumption that novices are students or self taught individuals, this preference towards the phonetic technique may be because of the way introductory animation courses teach lip animation. These courses often teach lip sync animation with mouth shapes similar to the exaggerated shapes in the Preston Blair mouth set. For someone with this assumed education background, exaggerated mouth shapes are expected and thus would look the most appealing, while someone with more experience may find the phonetically influence exaggeration too distracting.

Chapter 7

Conclusions

7.1 Current Status

As can be seen from the results of the comparison survey, no one likes Bob, the animation from the combined technique. The industry standard phonetic keyframe technique and motion capture technique had a much larger preference with the motion capture technique reigning supreme in almost all subgroups. Even though the motion capture technique does not follow any traditional rules of facial animation (exaggerated mouth shapes, keyframing only at peak audio locations, etc), and the actor's upper lip location was greatly miscalculated, its perfect timing with the audio and subtle shapes made it the most likable to the average viewer. A speculation as to why the phonetic animation was still a contender amongst viewers is that some were more comfortable with an animation that looked the most traditional. The phonetic animation had the cleanest mouth transition and the most exaggerated mouth shapes, giving it an classic MGM or Warner Brothers cartoon vibe.

7.2 Discussion

As the two industry used animation techniques varied so much in their viewer preference and response, it is difficult to create a well-liked and cohesive combined animation workflow. This was the first full test of this system, so there were some limitations. Further improvements need to be made to the combined technique before re-evaluating against motion capture only and phonetic only animations. As the most subtle technique ranked the highest in preference, the weight values given to the mouth blend shapes during the

phonetic pipeline of the workflow could be lowered as they are currently between 60% and 70%. A test between several phonetically mapped animations at different percent points could be tested against each other to figure out what viewers prefer. This, however, is still very subjective and will easily change depending on the model and blend shapes used.

7.2.1 Lip-Sync

A comment by survey participants as to why they preferred the motion capture animation compared to the others was that it had better sync. While the audio and audio analysis for the animations were the same, a speculation as to why participants said this is that the motion capture technique animated the mouth not only when it was speaking. The motion capture technique was able to animate every breath and pause better than the phonetic or the combined techniques because this technique involved animating at every frame instead of just keyframing the peak changes according to the audio analysis. Because of this, the phonetic and the combined techniques were not able to include every slight pause in the middle of a word and the mouth might have appeared to be moving between blend shapes when the actor was taking a breath. While this was a problem for some, one participant did mention that the combined technique had what they perceived to be the best lip sync. They commented that, ”I was most convinced of the vocal movements matching that was said with bob”. This progresses the idea that the observation of better lip-sync is preference based as there is no consensus if one animation had better timing than the other.

7.2.2 Mouth

Another possible culprit in having participants believe that one lip-sync was better than the other were the teeth. According to survey feedback, some participants stated an issue with ”too much teeth” in the phonetic and combined animations. One such participant commented on the teeth causing a distraction when picking between videos, ”[...] lots of teeth, so the sync might have been sharper but the shapes were distracting”. Having the lips move without the teeth following as one would expect could have created a disconnect between

the mouth movements and the audio for viewers. This could be due to the way the teeth location values are connected to the mouth rig controllers, and maybe some changes need to be done to the teeth in each viseme blend shape. The teeth also need to be animated separately when doing the combined system as a better approach is needed when determining where the teeth should be. Maybe more weight should be given to the phonetic animation data for the teeth placement as the phonetic approach produces more exaggerated movements instead of doing equal weighting between the values from the two techniques. This equal weighting approach might have caused the teeth location values to average out, and appear mostly in the center of the mouth throughout the entire animation.

Along with the ”too much teeth” comment, it was stated that some of the uncanniness of the motion capture animation and the combined animation came from the fact that neither one ever had a neutral, closed mouth. One such participant said that their choice was determined by this fact because, ”[...] That was what made the others feel unnatural”. This can be entirely related to the tracking error of the actor’s upper lip from the motion capture data. To improve the motion capture data, the actor should have worn make up to increase contrast in needed areas. In this case, the tracking would have improved if the actor’s lips had more contrast to their skin. Even with image enhancements, the lips were rather fair and thin, so if the actor were to have lip stick or something to outline the lip region, tracking of the upper lip might have improved. That being said, this is known limitation to markerless facial capture systems and several takes with different conditions always need to be collected to avoid these types of instances.

Another limitation to the motion capture system could be its predefined list of required expressions in the Faceware character setup. In referring back to Table 5.1, it can be seen that the motion capture animation actually lacked the direct use of 7 of the 13 viseme blend shapes. Because Faceware did not specifically ask for these expressions, the motion capture technique might have not been able to recreate the most authentic mouth shape, thus creating very subtle, almost mumbling mouth movements. This would have also impacted the combined animation as there was no weighting influence from the capture data for

those viseme blend shapes. As there was no possibility of dampening of the full blend shape weight of the phonetic map by the motion capture data, the combined animation's viseme blend shapes have a varying peaking points, making it easier to notice one mouth shape being more exaggerated than another and causing viewer discomfort.

7.2.3 Uncanny Valley

Furthermore from the comments, the combined technique was said to be uncanny. The noise from the motion capture data and the exaggerated mouth shapes of the phonetic map technique were too much. The eyes were also stated to cause some distraction and discomfort. One participant noticed that, "the eye motion was very distracting from the lip movement. With [the phonetic animation], there was no eye movement, so it was easier to focus on the lip movement". The capture data for the upper region of the face was not smoothed, and should be filtered in the future to decrease the noise. The next iteration of this system should maybe avoid including the upper region of the face all together to avoid distraction until the combined lip-sync technique is improved enough to gain more viewer preference. It is also good to note that, while most who commented preferred the motion capture animation, many mentioned that they wish there was some exaggeration in the mouth movements of that animation. One participant was tied between the motion capture animation and the combined animation saying that, "[Sam] seemed the most natural [...] I liked Bob too - only slightly over the top and it was a hard choice between Sam and Bob". Another participant remarked wanting some exaggeration in the motion capture animation but "not to the extent or level as the other two". This means that while this version of the system did not work, a combined motion capture and phonetic mapping technique could be liked in the future.

7.3 Future Work

There is the possibility that this type of technique will never be preferred among viewers. The two techniques the combined system is based off of are so different in their final results

that combining the data from the two may always result in uncanny animation. However, in this early stage, it is important to keep testing this type of technique. With more testing of each subprocess, the combined technique could prove to be a viable option for small scale animation projects and student work, even as a beginning step in the workflow to then be manipulated as the animator sees fit. As the main issue with the current system from the survey comments seemed to be that the face was too exaggerated, the weight percentage of the blend shape attributes for the phonetic map needs to be lowered for this current character model. The weight percentage of the blend shape attributes can easily be adjusted to viewer and user preference, so there is a possibility of creating an appeasing animation with just one value change with this current system. Meaning, this system could be packaged as is and given to animators to adjust as needed. The next phase to test the system will be to create a final animation that is as subtle as the motion capture only animation, but with minor instances of recognizable phonetic mouth shapes, and do an evaluation comparison survey again. If the upper facial features are still going to be incorporated, the motion capture data for the eyebrow and eye needs to be processed as the noise caused distraction to many viewers.

Bibliography

[1] Analyzer 3.0. http://facewaretech.com/products/software/analyzer/, 2017.

[2] Faceware. http://facewaretech.com/, 2017.

[3] Retargeter 5.0. http://facewaretech.com/products/software/retargeter/, 2017.

[4] AUTODESK. Audio-driven facial animation workflow. https://knowledge.autodesk.com/support/motionbuilder/learn-explore/caas/CloudHelp/cloudhelp/2017/ENU/MotionBuilder/files/GUID-CAD62D87-DCE5-4DD2-8F57-EA5039D29C80-htm.html, Nov 2016.

[5] BLAIR, P. *Advanced Animation*. Walter F. Foster, 1994.

[6] BREGLER, C., COVELL, M., AND SLANEY, M. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques* (1997), ACM Press/Addison-Wesley Publishing Co., pp. 353–360.

[7] CAO, Y., TIEN, W. C., FALOUTSOS, P., AND PIGHIN, F. Expressive speech-driven facial animation. *ACM Trans. Graph. 24*, 4 (Oct. 2005), 1283–1302.

[8] CAPPELLETTA, L., AND HARTE, N. Phoneme-to-viseme mapping for visual speech recognition. In *ICPRAM (2)* (2012), pp. 322–329.

[9] CHUANG, E., AND BREGLER, C. Performance driven facial animation using blendshape interpolation. *Stanford University* (2002).

[10] CLINTON, P. *Polar Express a creepy ride*. CNN, November 2004.

[11] COULMAS, F. *The Blackwell Encyclopedia of Writing Systems*. Blackwell Publishing, 1999.

[12] DARGIS, M. *Do You Hear Sleigh Bells? Nah, Just Tom Hanks and Some Train*. NY Times, November 2004.

[13] DENG, Z., NEUMANN, U., LEWIS, J. P., KIM, T.-Y., BULUT, M., AND NARAYANAN, S. Expressive facial animation synthesis by learning speech coarticulation and expression spaces. *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (Nov. 2006), 1523–1534.

[14] DILL, V., FLACH, L. M., HOCEVAR, R., LYKAWKA, C., MUSSE, S. R., AND PINHO, M. S. Evaluation of the uncanny valley in cg characters. 511–513.

[15] DUDDINGTON, J. espeak text to speech. http://espeak.sourceforge.net/, June 2007.

[16] GALLAGHER, D. F. *Digital Actors in Beowulf Are Just Uncanny*. NY Times, November 2007.

[17] GOOGLE. Cloud speech api beta. https://cloud.google.com/speech/.

[18] HAZEN, T. J., SAENKO, K., LA, C.-H., AND GLASS, J. R. A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In *Proceedings of the 6th international conference on Multimodal interfaces* (2004), ACM, pp. 235–242.

[19] IBM. Speech to text. https://speech-to-text-demo.mybluemix.net/.

[20] IPA. American ipa chart. http://www.keywordsuggests.com/d2OzGY7k*8b*t4X0lm63HirRZlxD

[21] IPA. *The International Phonetic Association Handbook.* Cambridge University Press, 1999.

[22] JEFFERS, J., AND BARLEY, M. *Speechreading (Lipreading)*. Charles C Thomas Pub Ltd., 1971.

[23] JURAFSKY, D., AND MARTIN, J. H. *Speech and Language Processing 2nd Ed.* Stanford, Jan 2009.

[24] KITAGAWA, M., AND WINDSOR, B. *Mocap for Artists*. Focal Press, March 2008.

[25] LENZO, K. The cmu pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[26] MORI, M., AND MACDORMAN, K. F. *The Uncanny Valley*, vol. 7 of *Energy*. IEEE, 1970.

[27] NETI, C., POTAMIANOS, G., LUETTIN, J., MATTHEWS, I., GLOTIN, H., VERGYRI, D., SISON, J., AND MASHARI, A. Audio visual speech recognition. Tech. rep., IDIAP, 2000.

[28] PECK, G. *Walt Whitman in Washington, D.C.: The Civil War and Americas Great Poet*. The History Press, 2015.

[29] REITHAUG, D. Orchestrating success in reading. *The National Right to Read Foundation* (2002).

[30] SCIENCES, U. P. . L. Sampa - computer readable phonetic alphabet. http://www.phon.ucl.ac.uk/home/sampa/, 2015.

[31] TAYLOR, S. L., MAHLER, M., THEOBALD, B.-J., AND MATTHEWS, I. Dynamic units of visual speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Aire-la-Ville, Switzerland, Switzerland, 2012), SCA '12, Eurographics Association, pp. 275–284.

[32] TINWELL, A. *The Uncanny Valley in Games and Animation*. A K Peters/CRC Press, January 2014.

[33] TINWELL, A., GRIMSHAW, M., AND WILLIAMS, A. Uncanny behaviour in survival horror games. *Journal of Gaming & Virtual Worlds 2*, 1 (2010), 3–25.

[34] WEISE, T., BOUAZIZ, S., LI, H., AND PAULY, M. Realtime performance-based facial animation. *ACM Trans. Graph. 30*, 4 (July 2011), 77:1–77:10.

Appendix A

User Manual

1. Pick a dialogue, then save to text file.
2. Download and install eSpeak command line tool.
3. Record actor saying the dialogue with a rgb camera at 60fps.
4. Extract audio from video.
5. Run audio through IBM Watson's Speech-to-Text beta program and tweak time range of words as necessary.
6. Write words and their corresponding time frames to file.
7. Using phoneme2viseme.py, using the dialogue file and the file with the time frames included, convert dialogue to American English phonemes.
8. Using phoneme2viseme.py, map the phonemes to their corresponding visemes with the provided dictionary. Write to file with their time frames included.
9. Create or choose a pre-existing facial model.
10. Create blend shapes for the model that match the desired visemes and some extras to control the eyes, eyebrows, jaw, etc.
11. Create a rig or controller for these blend shapes for easier attribute editing in Autodesk Maya.

12. Process the footage with Faceware Analyzer. Create a neutral face frame, and process for Retargeter.

13. The capture data is read into Maya and connected to the model through the Retargeter plug-in.

14. The viseme mapping is read into Maya.

15. Using viseme2blendshape.py in the Maya Python console, convert the visemes from the file to control their corresponding blend shapes.

16. Run the ”Auto Pose” feature in Retargeter and delete unnecessary frames to help remove noise.

17. The capture data and viseme-blend shape sequence are used together to animate the mouth on the model. Both are given equal weight percentages for the blend shape attributes.

18. Render final animation out to image sequence files.

19. Combine image sequence to video file.

20. Add the audio.

Appendix B

Code Listing & Data

All code and data from surveys can be found on the provided disk submitted with this thesis report. On that disk, the folder ”McGowen_lipSyncThesis” contains all of the assets needed to recreate this system with the test model used, the .csv files of the survey results, and the full spread sheet with the results analyzed with the graphs found in this report. In the included ”readMe.txt” file, step by step instructions for installing and running this system can be found.

Appendix C

Participant's Comments

- ”I thought Dan was the best facial animation because it closed its mouth. That was what made the others feel unnatural.”
- ”I was most convinced of the vocal movements matching what was said with bob”
- ”Dan and Bob had bigger mouth movements making lip sync look worse and over emphasizing words with lips.”
- ”None of the lip syncs are very convincing, especially in the ”heart heart heart” sections. You need better shapes and must remember to hold on a few shapes as they merge into other ones instead of returning to a resting position between sounds. Bob has the most expression because of the face movement, but the eyebrow jitter is distracting rather than helpful.”
- ”[Sam] seemed the most natural- [Dan was] toooo exaggerated- i liked Bob too- only slightly over the top and it was a hard choice between Sam and bob.”
- ”With Sam and Bob, the eye motion was very distracting from the lip movement. With Dan, there is no eye movement, so it was easier to focus on the lip movement. After looking them all, I felt the least exaggerated movement (Sam) seemed to be most in sync with the soundtrack. The mouth motions also appeared more accurate.”
- ”I didn't think any of them were particularly convincing or appealing. ”Sam” was the most subtle, thus the most realistic. ”Dan” and ”Bob” were both way too exaggerated

and unrealistic. Sam could have used a tiny bit more exaggeration in certain areas, but not to the extent or level as the other two."

- "A big gap between subtle mouth shapes and exaggerated ones, a mix of the two would be ideal. Lips had a very hard time making convincing shapes in Dan and Bob, lots of teeth, so the sync might have been sharper but the shapes were distracting."

Appendix D

Survey Form

Animation Techniques Survey

The purpose of this research project is to compare three lip-sync animation techniques. This research is being conducted by Victoria McGowen at Rochester Institute of Technology.

Your participation is completely voluntary. You may choose not to participate and you may withdraw at any time. If you choose not to participate or withdraw from the survey, you will not be penalized. There are no known risks or discomforts associated with this survey. Your responses will be kept strictly confidential, and the digital data will be stored in secure computer files after submission. To help protect your confidentiality, the survey will not contain information that will personally identify you. Your name and contact information will not be stored. Your responses will be used to help improve facial animation techniques and work flows.

If you have any questions about this research study, please contact Victoria McGowen at vkm3473@rit.edu

* Required

Electronic Consent

Clicking the "agree" button below indicates that:

- you agree to the information above.
- you voluntarily agree to participate in the study.
- you are at least 18 years of age.

I agree to participate in the study *

- ○ agree
- ○ disagree

NEXT

Animation Techniques Survey

* Required

Survey Questions

The purpose of this research project is to compare three lip-sync animation techniques. This research is being conducted by Victoria McGowen at Rochester Institute of Technology.

Your participation is completely voluntary. You may choose not to participate and you may withdraw at any time. If you choose not to participate or withdraw from the survey, you will not be penalized. There are no known risks or discomforts associated with this survey. Your responses will be kept strictly confidential, and the digital data will be stored in secure computer files after submission. To help protect your confidentiality, the survey will not contain information that will personally identify you. Your name and contact information will not be stored. Your responses will be used to help improve facial animation techniques and work flows.

If you have any questions about this research study, please contact Victoria McGowen at vkm3473@rit.edu

This survey will take approximately 5 minutes.

What is your age group? *

- ◯ 18-19
- ◯ 20 - 29
- ◯ 30 - 39
- ◯ 40 - 49
- ◯ 50+

How often do you watch animated movies? *

- ◯ Never
- ◯ Occasionally (Less than once a month)
- ◯ Once a month or more
- ◯ Once a week or more
- ◯ Daily

How often do you watch animated TV shows? *

- ◯ Never
- ◯ Occasionally (Less than once a month)
- ◯ Once a month or more
- ◯ Once a week or more
- ◯ Daily

Have you ever worked on an animated film? (Directed, lighting, asset creation, etc) *

- ◯ Yes
- ◯ No

If yes to the question above, how many years experience do you have working with animation?

- ◯ 0-1 year
- ◯ 2-3 years
- ◯ 4+ years

BACK NEXT

Animation Techniques Survey

* Required

Videos Comparison

You will find 3 videos below. In them, the same 3D character model is lip-synced to the same audio. They are in no particular order. Closed captioning has been enabled, and viewing the videos in a quiet space or with headphones is advised.

Please pick which of the three videos has the most convincing and visually appeasing lip-sync.

Video 1 - Sam

Video 2 - Dan

Video 3 - Bob

Of the 3 videos above, which had the more convincing and visually appeasing lip-sync? *

- ◯ Sam
- ◯ Dan
- ◯ Bob

Any comments about what you noticed? (Optional)

Your answer

BACK SUBMIT