UNIVERSITÄT KOBLENZ · LANDAU
Fachbereich 4: Informatik

Arbeitsgruppe Aktives Sehen

Entwicklung eines Gesichts für den Haushaltsroboter Lisa

Bachelorarbeit

zur Erlangung des Grades eines Bachelor of Science (B.Sc.)
im Studiengang Computervisualistik

vorgelegt von
Julian Giesen

Erstgutachter: Prof. Dr.-Ing. Stefan Müller
Institut für Computervisualistik, Universität Koblenz, AG Computergrafik

Zweitgutachter: Dipl.-Inform. Dominik Grüntjens
Institut für Computervisualistik, Univeristät Koblenz, AG Computergrafik

Koblenz, im Mai 2011

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

	Ja	Nein
Mit der Einstellung der Arbeit in die Bibliothek bin ich einverstanden.	☐	☐
Der Veröffentlichung dieser Arbeit im Internet stimme ich zu.	☐	☐

.....

(Ort, Datum) (Unterschrift)

Institut für Computervisualistik
AG Computergraphik
Prof. Dr. Stefan Müller
Postfach 20 16 02
56 016 Koblenz
Tel.: 0261-287-2727
Fax: 0261-287-2735
E-Mail: stefanm@uni-koblenz.de

UNIVERSITÄT KOBLENZ · LANDAU

Fachbereich 4: Informatik

Aufgabenstellung für die Bachelorarbeit
Julian Giesen
(207200084)

Thema: Entwicklung eines Gesichts für den Haushaltsroboter Lisa

In unserer heutigen Zeit gewinnen Haushaltshilferoboter immer mehr an Bedeutung. Sie erledigen Aufgaben wie zum Beispiel das Reinigen von Böden, das Mähen des Rasens oder helfen bei der Überwachung. Allerdings haben psychologische Studien gezeigt, dass die Ausstattung solcher Roboter mit Gesichtern die Aufmerksamkeit der Benutzer stärker bindet und ihnen zugleich mehr Intelligenz zugetraut wird. Der Haushaltsroboter Lisa der Arbeitsgruppe Aktives Sehen besitzt bisher ein rudimentär animiertes Gesicht, das auf einem Display am Roboter angezeigt wird.

Diese Arbeit befasst sich damit, ein verbessertes Gesicht für den Roboter Lisa zu entwickeln. Das neue Gesicht wird rein grafisch, unter Verwendung des vorhandenen Displays, umgesetzt. Ziel ist, das Gesicht mit einem Sprachsynthesizer zu synchronisieren, sodass es in der Lage ist, gegebene Phoneme in Bewegungen, sogenannte Viseme, umzusetzen.

Die inhaltlichen Schwerpunkte der Arbeit sind:

1. Recherche und Identifizierungen aktueller Technologien
2. Softwaretechnische Planung von Konzepten für Gesichtsysteme
3. Erstellung und Implementierung der Systeme
4. Anbindung des entwickelten Systems an das bestehende Lisa-System
5. Bewertung der Ergebnisse
6. Dokumentation

Koblenz, den 11.10.2010

– Julian Giesen – – Prof. Dr. Stefan Müller –

Inhaltsangabe

Haushaltsroboter gewinnen in heutiger Zeit immer mehr an Bedeutung. Sie finden ihren Einsatz zum Beispiel beim Reinigen von Böden, beim Mähen des Rasens oder helfen bei der Überwachung in Gebäuden oder Außenanlagen. Der Mehrheit dieser Roboter fehlt allerdings etwas Entscheidendes, nämlich ein Gesicht. In dieser Arbeit wird die Entwicklung eines virtuellen, austauschbaren Gesichtes festgehalten, das speziell für den Roboter Lisa der Arbeitsgruppe „Aktives Sehen" der Universität Koblenz entwickelt wurde. Psychologische Studien nach Krach u. a. (2008) haben hierzu gezeigt, dass die Ausstattung solcher Roboter mit Gesichtern die Aufmerksamkeit der Benuzter stärker bindet und ihnen zugleich mehr Intelligenz zugetraut wird.

Computergesichter oder auch Talkingheads finden ihren Einsatz nicht nur bei Robotern, sondern auch in Videospielen oder Animationsfilmen. Allerdings wird in diesen Bereichen die Bewegung und Mimik des Gesichtes meistens per Hand erstellt, was einen festen Bewegungsablauf liefert. Interessanter ist hier der Einsatz in barrierefreien Anwendungen. In diesem Bereich werden aus einem eingehenden Audiosignal die Phoneme extrahiert und den entsprechenden Visemen zugeordnet. Im Rahmen der vorliegenden Arbeit wird dieser Vorgang mit Hilfe des Text-to-Speech-Synthesizer *Festival* erzielt, wodurch sich die Bewegung der Lippen entsprechend des gegebenen Textes dynamisch verändert.

Aber nicht nur die korrekte Bewegung der Lippen spielt eine wichtige Rolle, sondern auch die Emotionen. Mit Hilfe einer bestimmten Emoticon-Arithmetik, die dem zu synthetisierenden Text beigefügt wird, ist die Visualisierung der Emotionen möglich.

Abstract

Domestic robots will become more important in current time. They will find use in everyday situations like cleaning the floor, mowing the lawn or helping to monitor buildings and outdoor installations. However, most of these robots lack in a “human” face, a crucial factor. In this bachelor thesis a virtual, exchangeable face for the robot Lisa of the study group “Active Vision” of the University Koblenz will be developed. Psychological studies by Krach et al. (2008) have shown that robots with features attract attention of the observers and are considered to be more intelligent.

Computerfaces or also Talkingheads are not only used with robots but also in videogames and animation movies. But the movement and mime of the face is usually constructed by hand that provides a fixed course of motions. More interesting is the usage in applications free of barrier. In this field the phonemes are extracted from an incoming audio signal and are assigned to the appropriate visemes. Within this bachelor thesis this occurrence is achieved with the help of the Text-to-Speech-Synthesizer Festival whereby the movement of the lips change dynamically according to the given text.

Not only is the correct movement of the lips of great importance, but also the showing of emotions. With the help of a special emoticon arithmetic which is added to the text that is to be synthesized the visualization of emotions is possible.

Danksagung

Ich möchte mich bei allen bedanken, die mir bei der Anfertigung dieser Arbeit geholfen haben.

Ein besonderer Dank geht an meine Betreuer Dominik Grüntjens, David Gossow und Herrn Professor Müller für ihre Vorschläge und Anregungen. Besonders bedanken möchte ich mich auch bei Susanne Thierfelder, für ihre Unterstützung zur Integration meines Talkingheadsystems auf dem Roboter Lisa.

Inhaltsverzeichnis

1 Einleitung

Ein Roboter als Aushilfe im Haushalt gewinnt in heutiger Zeit immer mehr an Bedeutung. Es gibt Roboter, die autonom den Boden reinigen, den Rasen mähen oder solche, die bei der Überwachung von Gebäuden und Außenanlagen helfen. Die meisten dieser Roboter sind allerdings nicht sehr menschenähnlich, besitzen also vor allem kein Gesicht. Allerdings haben Forschungen gezeigt, dass ein Roboter, der einem Menschen ähnlicher sieht, die Aufmerksamkeit des Benutzers stärker bindet und seine Wahrnehmung anpasst. Dadurch wird sein Verhalten und die Kommunikation mit den Roboter beeinflusst (vgl. Krach u. a., 2008). Das Gesicht ist dabei das primäre Erkennungsmerkmal für den Menschen, in dem man noch kleinste Veränderungen und Bewegungen wahrnimmt.

Es gibt Robotergesichter, die komplett in Hardware „gegossen" sind wie zum Beispiel *FloBi* von der Universität Bielefeld, siehe Abbildung 1(vgl. Lütkebohle u. a., 2010). Oder aber auch Robotergesichter, die rein grafisch mit Hilfe eines Displays angezeigt werden wie in Abbildung 2 der Roboter *Lisa* der Universität Koblenz.

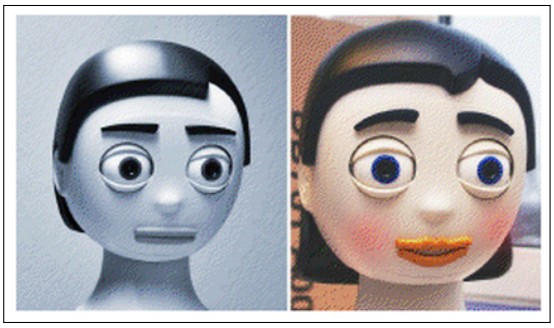


Abbildung 1: Hardware-Variante *FloBi*

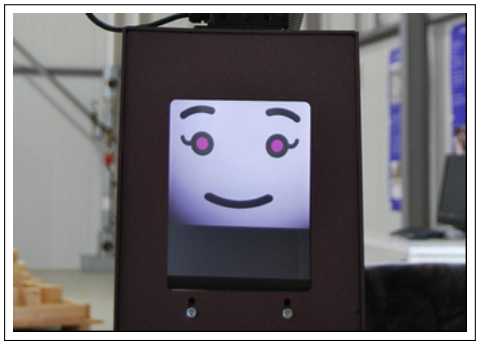


Abbildung 2: Display-Variante *Lisa*

Es ist von großer Bedeutung, ob ein Gesicht nur den Mund oder Kiefer auf und ab bewegt, oder ob Bewegungen im ganzen Gesicht zu erkennen sind. Ein hochdetailliertes Gesicht, das aufwendig erzeugt wurde und atemberaubend aussieht, wirkt dann nur leblos und unglaubwürdig. Daher ist es wichtig, dass ein vom Menschen erstelltes Gesicht glaubwürdige Bewegungen aufweist und zudem Emotionen zeigt. Vor allem sollten sich die Lippen entsprechend dem Gesprochenen bewegen. So kann auch ein sehr einfaches Gesicht lebendig wirken.

1.1 Motivation

Die vorliegende Arbeit legt ihr Interesse auf die Display-Variante. Gesichter in diesem Bereich werden Computergesichter oder auch Talkingheads genannt. Talkingheads finden ihren Einsatz nicht nur im Bereich der Robotik, sondern auch in Videospielen oder Animationsfilmen sind sie oft anzutreffen sowie in barrierefreien Anwendungen.

In Videospielen oder Animationsfilmen wird die Animation der Gesichter meist noch aufwendig von Hand erstellt. Dabei wird zu einem gesprochenen Text die Animation erstellt und es entsteht ein fester Bewegungsablauf. Interessant wäre also ein System, das die Animation entsprechend des Textes automatisch erzeugt. Assistenten für barrierefreie Programme sind hier zu nennen wie zum Beispiel das *SYNFACE Project*[1]. Dieses hat ein System entwickelt, das hörgeschädigten Personen das Telefonieren erleichtern soll. Dabei ist das Telefon mit dem Talkinghead-System verbunden und zum eingehenden Sprachsignal werden die Lippenbewegungen erzeugt, um so hörgeschädigten Menschen das Lippenlesen während des Telefonierens zu ermöglichen (vgl. Karlsson u. a., 2003).

1.2 Ziele der Arbeit

Der Roboter Lisa der Arbeitsgruppe „Aktives Sehen" der Universität Koblenz hat bisher ein rudimentär animiertes Gesicht. Der Mund geht beim Sprechen auf und zu und das Gesicht hat eine stetige Grundfreundlichkeit, ohne weitere Emotionen.

Ziel dieser Arbeit ist es, das bisherige Gesicht so zu erweitern, dass sich der Mund korrekt zum gesprochenen Text bewegt und neben der Grundfreundlichkeit auch weitere Gefühlsregungen zu erkennen sind.

Es wird ein neues System entwickelt, bestehend aus einer 3D Graphics Engine und einem Text-to-Speech(TTS)-System, mit dem es ggf. möglich ist, die Gesichter auszutauschen. Das TTS-System ist für die Generierung der Stimme und der Phoneme zuständig, die 3D-Engine zur Darstellung der Viseme und weiteren Animationen.
Neben der Entwicklung des neuen Talkingheadsystems ist es das Ziel, dieses System an das bisherige Lisa-System anzubinden.

[1] *SYNFACE Project*: `http://www.speech.kth.se/synface/` (zuletzt geprüft am 07.04.2011)

1.3 Aufbau der Arbeit

Die Arbeit gliedert sich in insgesamt 8 Kapitel. Auf das einleitende Kapitel folgen drei weitere Kapitel, die die Grundlagen zur Modellierung, Animation und Sprachsynthese einführen. Im fünften Kapitel wird der Ansatz und das Konzept des Talkinghead-Systems veranschaulicht. Anschließend folgen das sechste Kapitel über die Implementierung und eine Bewertung im siebten Kapitel. Im achten und letzten Kapitel schließt das Fazit sowie ein Ausblick auf die zukünftike Entwicklung des Talkingheadsystems die Arbeit ab.

2 Grundlagen der Gesichtsmodellierung

Dieses Kapitel gibt eine kurze Einführung in die grundlegenden Kategorien der Modellierung und befasst sich anschließend mit den Grundlagen zur 3D-Modellierung von Gesichtern. Vor allem bei der Modellierung von Gesichtern gibt es verschiedene Möglichkeiten. Man sollte deshalb schon vorab wissen, ob man ein menschliches Gesicht oder ein abstrahiertes Cartoongesicht erstellen will. Zur Erstellung eines menschlichen Gesichts gibt es zwei Methoden. Zum einen die Digitalisierung über physische Referenzen, wenn das gewünschte Model eine bestimmten Person repräsentieren soll. Zum anderen die Modifizierung eines bereits existierenden Models, wenn dieses schon mit kontrollierbaren Animationen ausgestattet ist. Um ein Cartoongesicht zu erstellen benötigt man nicht zwingend Digitalisierungsmöglichkeiten für bereits vorhandene Personen. Hier bietet es sich an, das Gesicht mittels eines CAD-Systems (computer-aided design) oder einer 3D-Grafik-Software zu erstellen, da hierbei dem Künstler alle Freiheiten erlaubt sind. Deshalb folgt eine kurze Vorstellung der grundlegenden Kategorien der Modellierung.

2.1 Grundlegende Kategorien der Modellierung

Zum Erstellen eines 3D-Modells wird ein entsprechendes Werkzeug benötigt, eine 3D-Grafik-Software. Es gibt derzeit verschiedene Programme auf dem Markt wie zum Beispiel die Open Source „3D content creation suite" *Blender*[2]. Blender bietet eine Fülle an Modelierungsfunktionen. Adam Watkins beschreibt in seinem Buch „3D Animation: From Models To Movies" drei Basiskategorien der zur Verfügung stehenden Funktionen zur Erstellung von Shapes (vgl. Watkins, 2001, S. 31ff), die im Folgenden kurz vorgestellt werden. Zum einen spricht er von Polygon Modellen, von Splinebasierten Modellen und zum anderen von „Extruding, Lathing und Skinning" , den sog. Sweeping Modellen nach Kerlow (2000). Auch Rick Parent erwähnt drei Methoden, welche die Geometrie eines Models beschreiben (vgl. Parent, 2002, S. 341f). Er unterteilt sie in ähnlicher Weise, erwähnt allerding nicht die Sweeping Modelle, sondern stellt die Subdivision Surface vor.

2.1.1 Polygon Modelle

Polygon Modelle werden in verschiedenen 3D Applikationen am häufigsten genutzt, da sie den Prozessor nur gering belasten wie Watkins in seinem Buch erwähnt:

[2]Blender: `http://www.blender.org/` (zuletzt aufgerufen am 29.04.2011)

„Polygon models are the least taxing on a computer's processing muscles. Because of this, most low-to mid-range 3D applications rely heavily or solely on polygon modeling."(Watkins, 2001, S. 31). Bei der Polygon Modellierung werden zweidimensionale Shapes, sog. Polygone, im virtuellen Raum zu einem dreidimensionalen Shape zusammengefügt. Dabei werden am häufigsten Quadrate oder Dreiecke verwendet. Abbildung 3 zeigt einen Würfel, zusammengesetzt aus sechs Quadraten. In Abbildung 4 ist eine Pyramide zu sehen, bestehend aus sechs Dreiecken (vgl. Watkins, 2001).

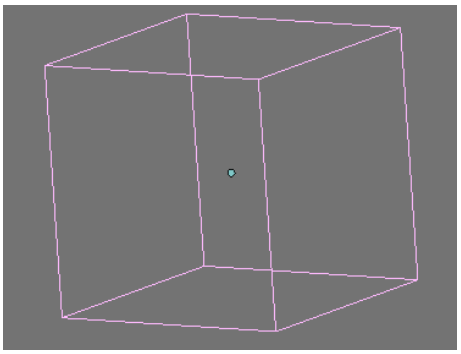


Abbildung 3: 3D Cube aus Polygonen

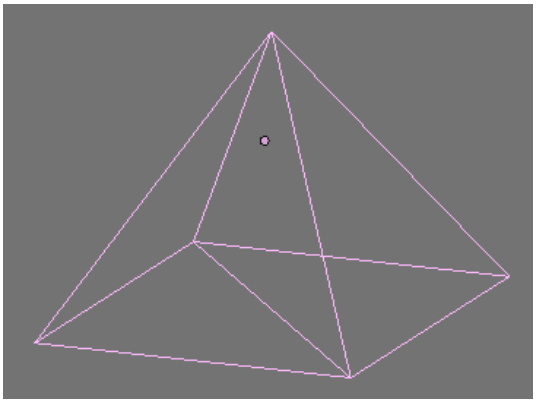


Abbildung 4: 3D Pyramide aus Polygonen

Solche Gebilde mit den vorgestellten Polygonen zu erstellen ist kein Problem. Will man aber nun abgerundete Objekte erstellen, gibt es ein Problem: zweidimensionale Shapes sind nicht biegsam. Das heißt, man muss auf die Verwendung vieler Polygone zurückgreifen, um zum Beispiel eine Kugel wie in Abbildung 5 zu erzeugen. Das treibt die Prozessorauslastung nach oben (vgl. Watkins, 2001). Mit dem Stand heutiger Grafikkarten lassen sich durchaus ohne Probleme mehr Polygone darstellen als früher, aber die Darstellung zu vieler Polygone kann den Rechner dennoch ins Stocken bringen.

3D-Grafik-Softwares stellen geometrische Primitive zur Verfügung. Das sind einfache, vorgefertigte 3D-Objekte mit einer festen Struktur wie Würfel, Kugeln oder Zylinder, die das Konstruieren von Modellen vereinfachen. Sie werden entweder als polygonale Strukturen oder als Spline Patches (s. Kapitel 2.1.2) erstellt. Sie können dann als einfache Shapes genutzt werden oder dienen als Basis für komplexere Shapes (vgl. Kerlow, 2000, S. 105ff).

2.1.2 Spline basierte Modelle und NURBS

Splines sind eine Sequenz von dreidimensionalen Eckpunkten, die durch eine Linie verbunden sind. Die Verbindung der Eckpunkte geschieht durch Interpolation. Splines lassen sich relativ frei anpassen und besitzen keine ungewollten harten Kanten (vgl. Watkins, 2001).

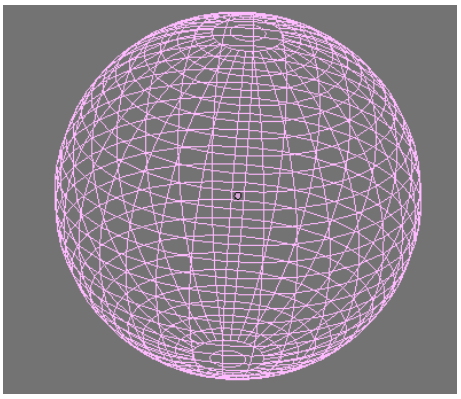


Abbildung 5: 3D Kugel aus Polygonen

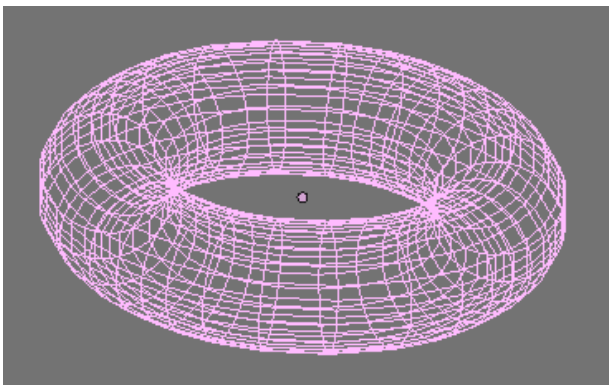


Abbildung 6: 3D Donut aus NURBS

Eine andere Variante sind NURBS (Non-Uniform Rational B-Splines). Der Vorteil von NURBS ist, dass sie abseits der Interpolationspunkte Kontrollpunkte haben, mit denen sich ein Spline leichter verformen lässt (vgl. Watkins, 2001). NURBS können Kurven oder aber auch Flächen sein und können somit zur Modellierung beliebiger Formen verwendet werden (Bsp. s. Abb. 6).

„Spline models typically use bicubic, quadrilateral surface patches, such as Bezier or B-Spline ...“(Parent, 2002, S. 342). Der Vorteil von Surface Patches ist die geringe Datenkomplexität im Gegensatz zu den polygonalen Techniken zur Generierung von glatten Oberflächen. Splines werden also häufig verwendet, wenn man eine glatte Oberfläche wünscht. Nachteile entstehen jedoch bei der Erstellung von Gesichtern. Bei der Standard Surface Patch Technologie wird das ganze Model aus einem rechteckigen Netz von Kontrollpunkten erstellt. Hierbei ist es schwierig, kleine Details und elegante Merkmale zu erzeugen und dabei die Komplexität gering zu halten, da man ganze Reihen und/oder Spalten von Kontrollinformationen modifizieren muss. Durch eine kleine Veränderung eines lokalen Bereiches der Oberflächen muss man also die ganze Oberfläche verändern (vgl. Parent, 2002).

2.1.3 Sweeping Modelle

Das Sweeping ist wohl die mächtigste abgeleitete Modelliertechnik. Die Hauptidee hinter allen Sweeping Techniken ist, sich eine zweidimensionale Ausgangsfigur zu definieren, die dann entlang eines bestimmten Weges ausgestrichen wird und ein Shape im dreidimensionalen Raum definiert (vgl. Kerlow, 2000, S. 108ff).

Extrusion ist eine solche Technik, um aus einem zweidimensionalen Shape einen dreidimensionalen Körper zu erzeugen. Durch Parallelverschiebung im Raum erhält man eine Dimensionserhöhung des entsprechenden Objektes. Man kann Shapes gerade extrudieren oder entlang einer Spline (vgl. Watkins, 2001).

Der Vorgang der Extrusion ist ähnlich wie bei einer Pastamaschine oder einem Fleischwolf, bei denen die Pasta oder das Fleisch durch verschieden geformte Pressbacken länger gepresst wird und eine neue Form annimmt (vgl. Kerlow, 2000).

Beim Lathing wird ein zweidimensionales Objekt um eine Achse aufgesponnen und ergibt dann ein symmetrisches dreidimensionales Shape. Ein Vorteil des Lathings ist, dass man das Shape während des Lathings rotieren kann (vgl. Watkins, 2001).

Ein Beispiel wäre der Donut aus Abbildung 6. Diesen könnte man auch durch Lathing erzeugen.

Bei Skinning wird ein Skelett aus zweidimensionalen Shapes erstellt und anschließend wird über das Skelett eine Art Haut gezogen. So kann man fast jede beliebige Form erzeugen, wobei es für manche Formen natürlich effizientere Methoden gibt (vgl. Watkins, 2001).

Mit dem Skinning ist es also leicht möglich gebogene Oberflächen zu erstellen. Das Überziehen der Haut funktioniert so, dass eine Sequenz der zweidimensionalen Shapes mit Kurven verbunden wird. Das Skinning ist besonders nützlich für die Erstellung von menschlichen Modellen, weil diese meist als eine Reihe von zweidimensionalen Konturen beschrieben sind (vgl. Kerlow, 2000, S. 127ff).

2.1.4 Subdivision Surfaces

„Subdivision Surfaces have the flexibility of polygonal meshes but without the faceted look typical of low resolution polygonal geometry.“(Kerlow, 2000).

Ein Subdivision Surface ist eine glatte Fläche, die aus einem Polygonnetz durch einen iterativen Prozess erzeugt wurde. Während des Prozesses wird das Netz ge-

glättet, indem seine Dichte erhöht wird. Subdivision Surfaces haben den Vorteil lokale Komplexität ohne globale Komplexität zu erzeugen. Sie bieten eine einfach zu benutzende, intuitive Schnittstelle zum Entwickeln neuer Models. Modelliert man Gesichter mit Subdivision Surfaces, kann es dennoch zu Problemen kommen, wenn man kleine Details oder elegante Merkmale hinzufügen möchte. Dennoch ist es eine gute Variante zur Erstellung von Cartoongesichtern (vgl. Parent, 2002).

Das Vorgehen bei Subdivision Surfaces ist folgendermaßen beschrieben und teilt sich in zwei Schritte auf: Teile jede Fläche in vier Bereiche und positioniere dann die Eckpunkte neu durch Mittelung der lokalen Gewichtspunkte. Das Verfahren kann beliebige Male wiederholt werden um die Detailstufe zu erhöhen (s. Abbildung 7).

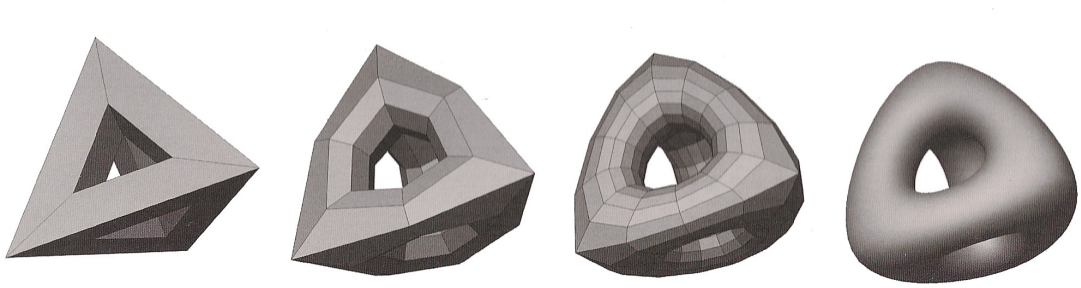


Abbildung 7: Ausgangsobjekt und drei nacheinanderfolgende Schritte der Subdivision Surface Methode

2.2 Gesichtsmodellierung von Hand

Mit einer 3D Modellierungssoftware geschieht die Modellierung generell von Hand. Da sich die Arbeit mit der Erstellung eines zeichentrickbasierten Gesichts befasst, wird hier eben diese Modellierung beschrieben. Es ist wichtig, vor Beginn der Modellierung zu bedenken, ob man ein animiertes Gesicht modellieren will. Deshalb muss darauf geachtet werden, dass während des Modellierens oder nachträglich verschiedene Formen hinzugefügt werden können. Oder wie Jason Osipa in seinem Buch „Stop Staring" schreibt: „The best point layout is always the one that, in wireframe, looks as if you can see your character's facial muscles."(Osipa, 2003, S. 76). Das in dieser Arbeit konstruierte Gesicht lehnt sich stark an die von Osipa beschriebene Vorgehensweise an, die im Folgenden kurz beschrieben ist.

Ein wichtiger Punkt, den Osipa aufgreift, ist: „... to model for movement.“(Osipa, 2003, S. 76). Ein gut modelliertes Gesicht hat ein Point Layout bestehend aus konzentrischen Kreisen. Diese umkreisenden Point Layouts sind vor allem bei Mund und Augen zu erkennen. Die Kreise sollten möglichst perfekt sein (vgl. Osipa, 2003).

Allerdings ist hierbei darauf zu achten, dass man für die Bewegung modelliert, also sich Gesichtsausdrücke und Mundbewegungen verdeutlicht und das Gesicht so modelliert, dass die Möglichkeit besteht, die Gesichtsausdrücke und Mundbewegungen zu erzeugen. Es kann jede der Methoden zum Modellieren des Gesichtes verwendet werden, die in Kapitel 2.1 vorgestellt wurden.

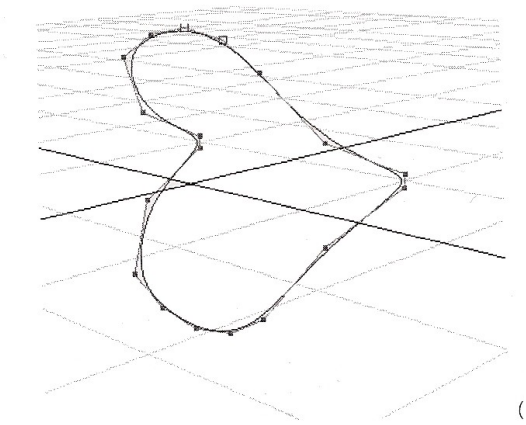


Abbildung 8: Grundgerüst für den Mund (NURBS)

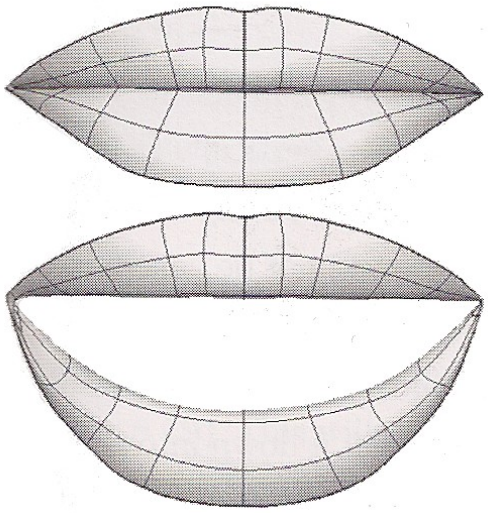


Abbildung 9: Mund geschlossen und offen

2.2.1 Erstellen des Mundes

Der Mund ist neben den Augen und Augenbrauen der wichtigste Teil des Gesichts. Es ist möglich, allein durch den Mund verschiedene Emotionen zu visualisieren wie zum Beispiel ein Lächeln für Freude oder abgesenkte Mundwinkel für Traurigkeit. Er ist aber vor allem relevant, um dem Gesicht beim Sprechen Ausdruck zu verleihen. Deshalb beginnt die Modellierung des Gesichtes auch mit dem Mund.

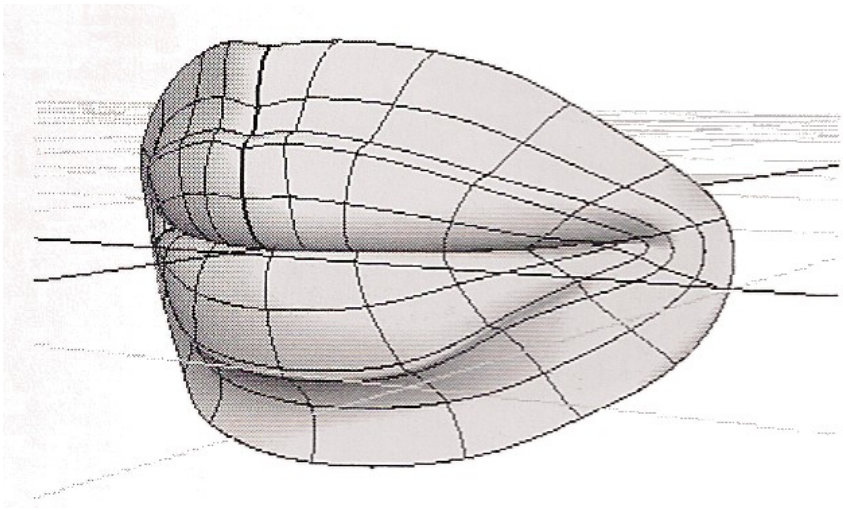


Abbildung 10: Erweiterter Mund

Als Erstes werden die Lippen modelliert. Dazu wird ein Kreis bestehend aus 18 Punkten verwendet. Zu Beginn ist es nicht von Bedeutung, ob der Kreis linear oder geglättet ist (s. Abbildung 8). Anschließend werden die Lippen so definiert, dass sie die entsprechende Lippenform bekommen (s. Abbildung 9). Hat man die Lippen modelliert, müssen diese zum Gesicht übergeleitet werden. Dazu extrudiert man bestimmte Bereiche und erweitert den Mund wie in Abbildung 10. Nun wird der umliegende Bereich des Mundes erstellt, zuerst die Nase und anschließend das Kinn, bis das Gesicht aussieht wie in Abbildung 11. Danach wird die Nase verfeinert. Als Nächstes werden Zähne und Zunge konstruiert, falls man diese für sein gewähltes Model benötigt, und anschließend mit dem Rest des Gesichtes verbunden. (vgl. Osipa, 2003)

2.2.2 Erstellen der Augen und Augenbrauen

„The brows and the eyes tell us what we need to know about a character's thoughts."(Osipa, 2003, S. 141). Die Augenbrauen dienen vor allem zur Anschauung der Emotionen. Auf- und Abbewegungen der Augenbrauen sagen alleine noch nicht viel über Emotionen aus, aber in Verbindung mit dem Zusammenpressen der Augenbrauen sind verschiedene Emotionen möglich. Hohe Augenlider signalisieren Aufmerksamkeit und niedrige Augenlider verstärken Emotionen. Die Augen vermitteln nur die Blickrichtung. Alles zusammen erzeugt dann aber eine ausdrucksstarke Emotion (vgl. Osipa, 2003).

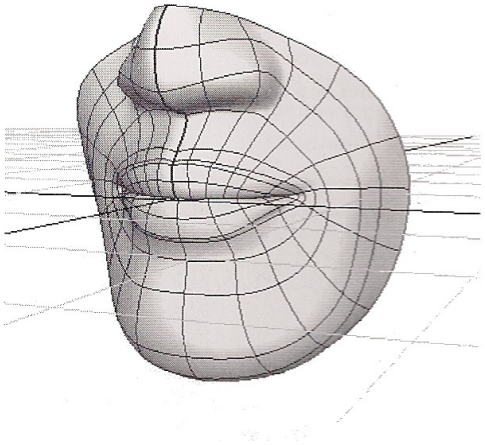


Abbildung 11: Mund mit Ansätzen von Nase und Kinn

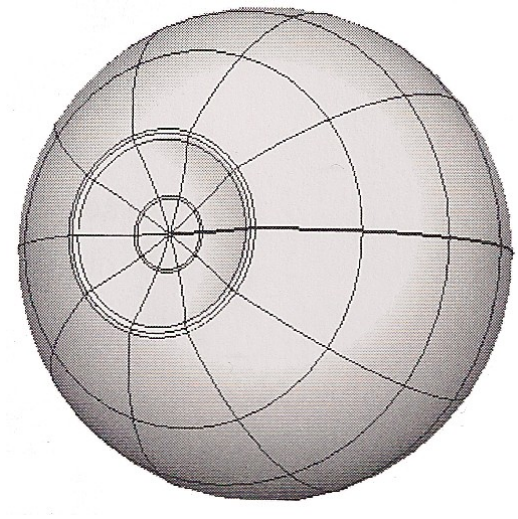


Abbildung 12: Augapfel mit Krater für Pupille und Iris

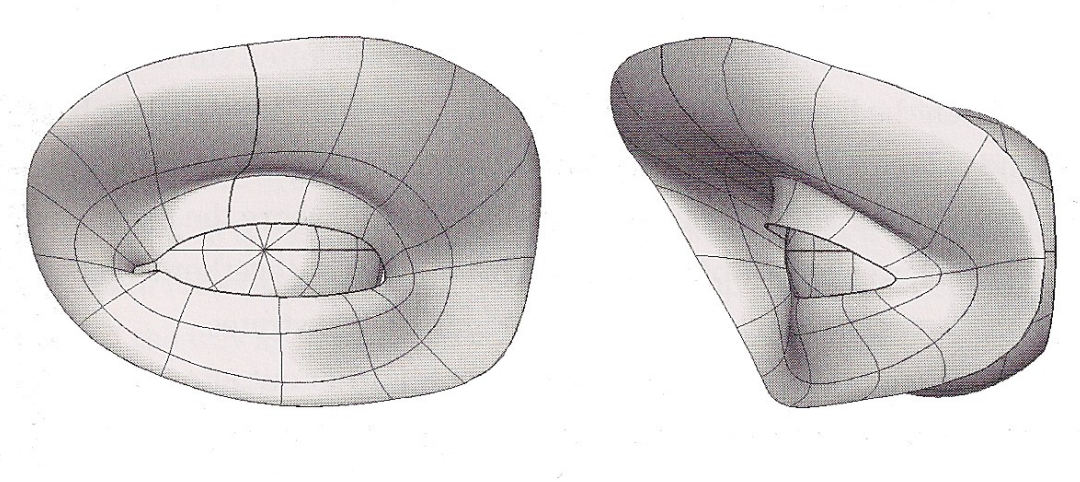


Abbildung 13: Augenhöhle mit Augenlid

Zu Beginn wird der Augapfel modelliert. Hierzu wird eine Kugel (z.B. NURBS Kugel) verwendet. An einer Stelle wird ein Krater in die Kugel „gemeißelt" für Pupille und Iris (s. Abbildung 12), die anschließend modelliert werden. Als nächster Schritt wird die Augenhöhle modelliert, wie in Abbildung 13 zu sehen ist. Das bisherige Auge wird gespiegelt und mit der Nase verbunden. Nach einigen Verfeinerungen sollte das Ganze wie in Abbildung 14 aussehen. Im Anschluss wird die Stirn hinzugefügt. Die Augenbrauen können in das Modell miteinfließen oder separat modelliert werden. Zuletzt wird der obere Teil des Schädels vervollständigt und sollte dann wie in Abbildung 15 aussehen (vgl. Osipa, 2003).

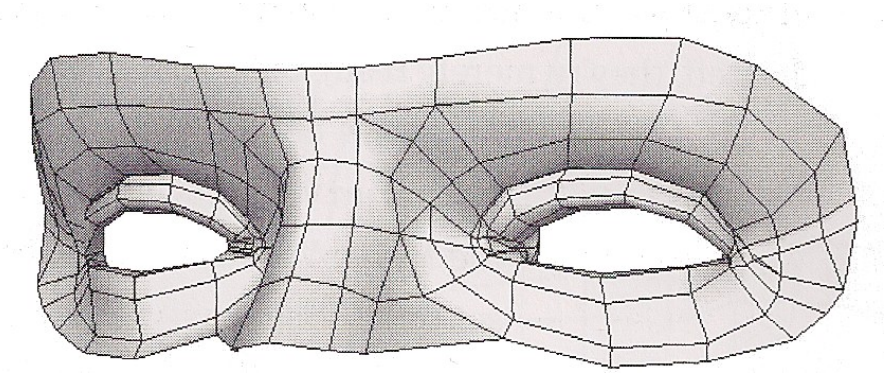


Abbildung 14: Augenmaske mit Ansätzen der Nase

2.2.3 Zusammenführen von Augenpartien und Mund

Die Augen und Augenbrauen haben wie der Mund und der Bereich um den Mund herum ihre eigenen Strukturen für die jeweiligen Bewegungen und die Darstellung der Emotionen (vgl. Osipa, 2003). Damit ein ganzes Gesicht entsteht, müssen jetzt beide Teile zusammengefügt werden.

Als Erstes wird der vordere Bereich des Gesichtes miteinander verbunden. Danach wird der Umriss des Kinns angepasst und der Nacken erstellt (s. Abbildung 16). Das Ohr wird aus einem Halbkreis wie in Abbildung 17 zu einem Ohr geformt und an die entsprechende Stelle des Kopfes angefügt (s. Abbildung 18) (vgl. Osipa, 2003).

Das Ergebnis ist nun ein animierbarer Kopf, der hier in den Abbildungen sehr menschlich aussieht, obwohl ein cartoon ähnliches Gesicht in dieser Arbeit erstellt wurde.

Während des Modelliervorgangs ist es allerdings freigestellt, ob man ein cartoonähnliches oder menschliches Gesicht erschafft. Die vorgestellten Schritte dienen allein dem Grundverständnis zur Modellierung eines Gesichts und können durchaus variiert werden. Eine genauere und ausführlichere Beschreibung der einzelnen Schritte zur Modellierung eines animierbaren Gesichtes sind im schon mehrfach erwähnten Buch „Stop Staring - Facial Modeling and Animation Done Right™“von Jason Osipa zu finden.

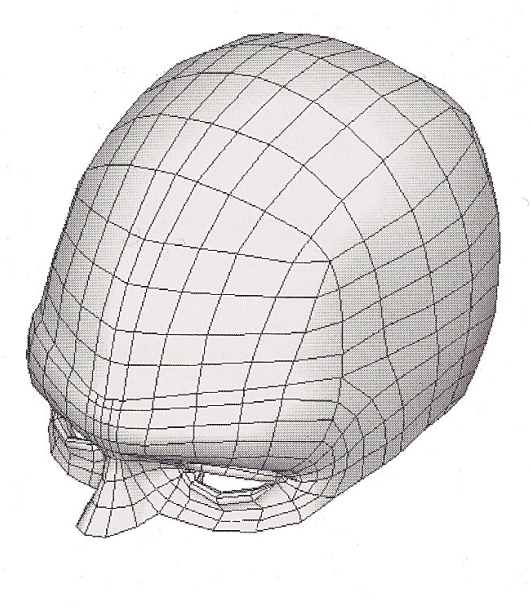


Abbildung 15: Fertiger oberer Teil des Kopfes

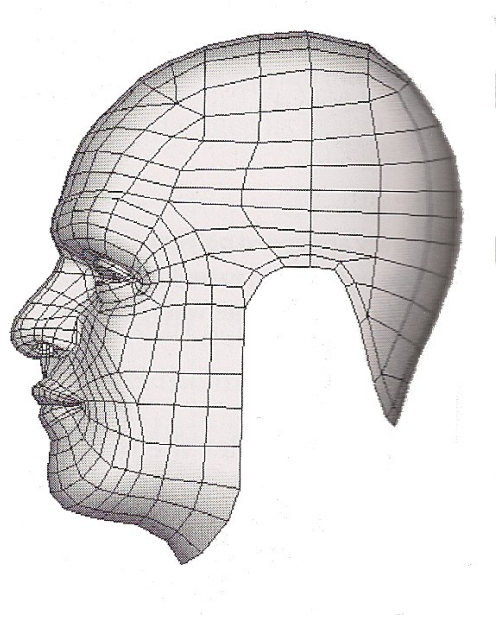


Abbildung 16: Oberer und unterer Teil des Kopfes zusammengesetzt

2.3 Zusammenfassung

Zur Modellierung dreidimensionaler Objekte gibt es eine Vielzahl an Techniken. Besonders weit verbreitet sind die Polygon und Splinebasierten Modelle. Die Techniken sind meist alle ohne Probleme kombinierbar, führen aber stets zu anderen Ergebnissen. Deshalb muss man vorab wissen, was das zu modellierende Objekt darstellen soll. So hat sich gezeigt, dass einige Verfahren besonders gut für die Modellierung von Gesichtern geeignet sind wie zum Beispiel das Skinning, andere dieses Ziel jedoch verfehlen.

Die von Jason Osipa vorgestellte Methode zur Modellierung von Gesichtern ist effizient und leicht umsetzbar. Da sie sich in die separate Modellierung von Mund und Augenpartien einteilt, besteht die Möglichkeit, sich auf die jeweiligen Bereiche ausgiebig zu konzentrieren. Ein weiterer Vorteil dieser Methode ist die Flexibilität. Es ist freigestellt, welche Modellierungstechniken man verwendet und mit welchem Teil man beginnt.

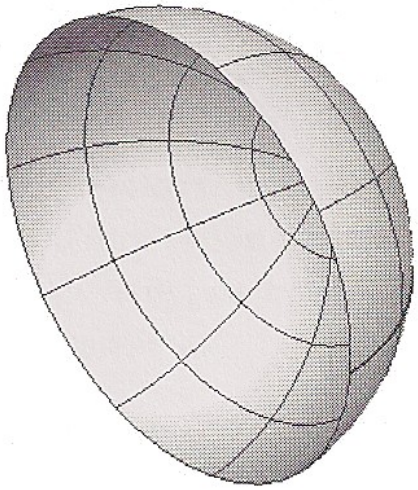


Abbildung 17: Ausgangsform zur Modellierung des Ohrs

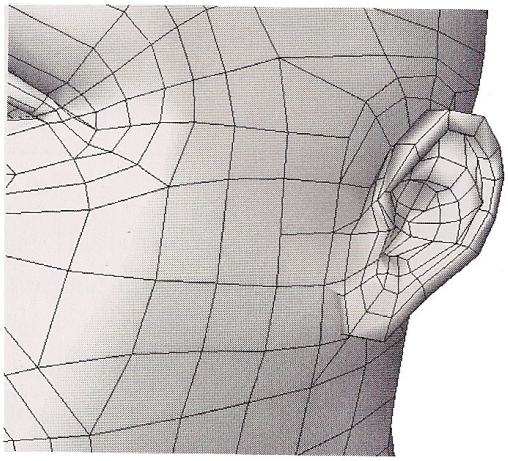


Abbildung 18: Kopf mit fertigem Ohr

3 Grundlagen der Gesichtsanimation

Dieses Kapitel soll einen Überblick über die verschiedenen Ansätze der Gesichtsanimation verschaffen. Im ersten Teil dieses Kapitels werden kurz die verschiedenen Ansätze vorgestellt. Im zweiten Teil wird das „klassische" Vorgehen bei der Animation von Gesichtern verdeutlicht.

Es gibt viele verschiedene Techniken zur Animation der Gesichtsmimik, hier wären das Morphing zwischen Key Poses, Blend Shapes, Motion Capture, Motion Dynamics Simulation und Goal-Oriented Techniques zu nennen. Oft werden verschiedene dieser Techniken miteinander kombiniert. Meist gelingt eine Kombination des Morphings und des Blend Shapes, das hier ähnlich der Interpolationstechnik (s. Kapitel 3.1.1) beschrieben ist. Die geöffnete und die geschlossene Mundposition sind die wichtigsten Shapes in der Gesichtsanimation, da sie die beiden zueinander gegensätzlichen extremsten Emotionen veranschaulichen (vgl. Kerlow, 2000, S. 354).

3.1 Ansätze der Gesichtsanimation

Es wurden verschiedene Ansätze entwickelt, wie das Facial Action Coding System (FACS), bei dem alle Gesichtsmimiken in eine Reihe von Grundgesichtsbewegungen dekonstruiert werden. Ein weiterer Ansatz sind Parameterized Models, bei dem das Gesichtmodel entsprechend seiner einfachen Handlungen parameterisiert wird und die Werte der Parameter über die Zeit kontrolliert werden. Der Muscle Model Ansatz teilt das Gesicht in die drei Muskeltypen linear, sheet und sphincter auf und erzeugt daraus die Bewegungen (vgl. Parent, 2002).

Da diese Ansätze eher realistisch wirkende Gesichter zu erzeugen versuchen, werden sie zur Vollständigkeit nur kurz beschrieben. Als einfachster Ansatz ist die Interpolationstechnik zu nennen, die die Interpolation zwischen Key Poses beschreibt, die von Hand erstellt werden. Außerdem ist diese Technik wohl am besten für die Erstellung eines Cartoongesichtes geeignet.

3.1.1 Interpolationstechnik

Der einfachste Ansatz der Gesichtsanimation ist es, eine Reihe von Key Poses, auch Shape Keys genannt, zu definieren. Es wird dann zwischen den Positionen der Eckpunkte der jeweiligen Poses interpoliert. Dadurch wird die mögliche Bewegung durch die Interpolation zwischen zwei Shape Keys limitiert (vgl. Parent, 2002, S. 347).

Jeder Shape Key hat ein Gewicht, welches zwischen `0.0` und `1.0` definiert ist. Ein mit `0.0` gewichteter Shape Key zeigt keine Veränderung in der Bewegung, wobei ein mit `1.0` gewichteter Shape Key die höchstmögliche Bewegung verursacht. Das Gewicht wird dann bei der Interpolation mit einbezogen.

Dadurch können Gesichtsposen produziert werden, die nicht direkt durch einen Key definiert sind. Je mehr Key Poses erstellt werden, desto mehr Gesichtsbewegungen sind produzierbar. Allerdings kann sich die Anzahl der Key Poses zu einer unkontrollierbaren Zahl erhöhen. Es stellt sich dabei also die Frage: Was sind die Grundbewegungen des Gesichtes? (vgl. Parent, 2002).

- „The open mouth or the close mouth positions are the most important shapes because they are the extremes that show emotion“(Kerlow, 2000, S. 355).

Es ist wichtig die Bewegung des Mundes beinahe perfekt zu visualisieren. Man sollte sich also eine Sammlung von Key Poses erstellen, mit denen man fast jede Mundbewegung durch Interpolation darstellen kann (Kerlow, 2000). Ein Ansatz hierzu ist es, Phoneme (s. Kapitel 4.1.1) auf Viseme (s. Kapitel 3.2.1) abzubilden. In Abbildung 19 sind die wichtigsten Phoneme und ihre zugehörigen Viseme bebildert.

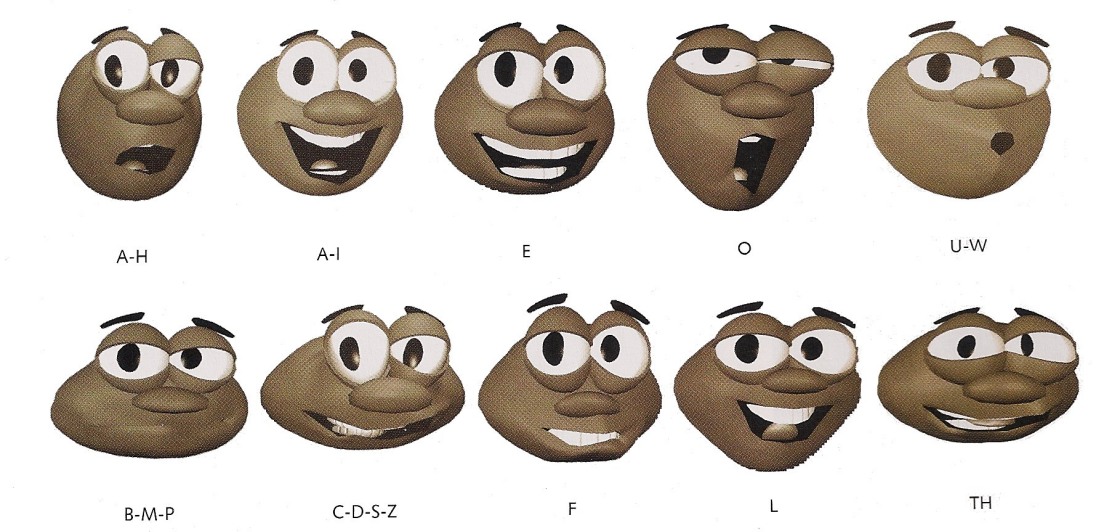


Abbildung 19: Basisphoneme und ihre dazugehörigen Viseme an einem Cartoongesicht veranschaulicht.

3.1.2 Parametrische Modelle

Schon in den siebziger Jahren versuchte man realistisch wirkende Gesichter zu synthetisieren. Parke (1974) entwickelte in dieser Zeit den parametergesteuerten Ansatz. „Dieser Ansatz gestattet unter Verwendung eines groben polygonalen Kopfmodells das Öffnen und Schließen der Augen und des Mundes.“(Jackèl u. a., 2006). Zur Erstellung des Polygonnetzes des Kopfes wird der Kopf eines Menschen mit photogrammetrischen Methoden erfasst oder direkt mechanisch vermessen. Die Augen- und Mundpartien werden durch die Parameter repräsentiert. Durch Interpolation sowie Rotations-, Translations- und Skalierungstransformationen erfolgt die Animation des Gesichtes, indem die Parameter verändert werden. So kann man anhand eines Parameters schon interessante Gesichtsausdrücke animieren. Ändert man beispielsweise den Parameter im Bereich des Unterkiefers, ist es möglich, den Rotationswinkel einzustellen und das Öffnen und Schließen des Mundes zu erzielen, zu sehen in Abbildung 20. Die Einfachheit dieses Ansatzes täuscht, denn man sollte mit dem Parametersatz vertraut sein, um den Zusammenhang zwischen den Parametern und der entsprechenden Gesichtsbewegung zu verstehen. Es gibt verschiedene Ansätze, die den Ansatz von Parke (1974) durch bessere Augenbewegungen oder phonembasierte Steuerung der Lippenbewegungen und visuellen Sprachartikulation erweitert haben (vgl. Jackèl u. a., 2006; Parent, 2002).

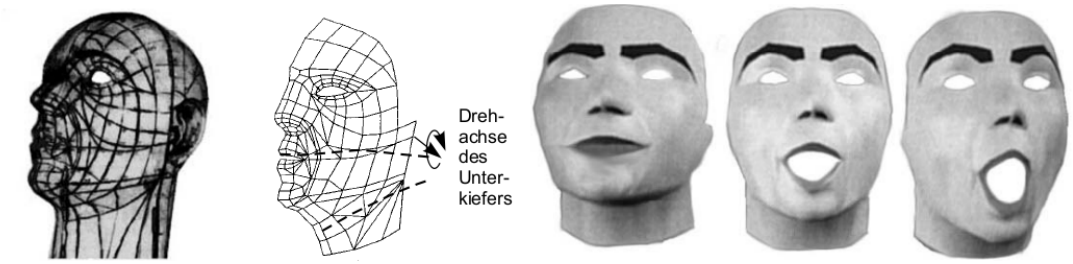


Abbildung 20: Öffnen des Mundes durch Koordinatenrotation und Veränderung eines Parameters nach Parke (1974)

3.1.3 Physikalische muskelgesteuerte Modelle

Ein erster Ansatz der physikalisch muskelbasierten Modelle wurden in den frühen achtziger Jahren von Platt und Balder entwickelt (vgl. Parke und Waters, 1996 u. 2008). Generell beruht dieses Modell auf der Mechanik und Anatomie von Haut und Mus-

keln im Gesicht und ihrem Zusammenwirken (vgl. Jackèl u. a., 2006). Im ersten Ansatz von Platt und Balder sind die Eckpunkte der Hauptpolygone mit simulierten Federn versehen. Anhand von simulierten Muskeln waren die Eckpunkte zusätzlich mit einer unterliegenden Knochenstruktur verbunden. Durch die elastische Eigenschaft der Muskeln konnten diese Kraft ausüben. Die Gesichtsmimik wurde durch Anwenden von Muskelkraft auf das elastische Hautmesh erzielt (vgl. Parke und Waters, 1996 u. 2008).

Waters entwickelte in den späten achtziger Jahren ein dynamisches Gesichtsmodell mit zwei verschiedenen Muskeltypen: den linearen Muskeln und den sphincter Muskeln. Ähnlich wie bei Balder und Platt wird bei diesem Ansatz ein Massefedermodell für Haut und Muskeln verwendet. Allerdings hat Waters Modell gerichtete Eigenschaften, was es unabhängig von der unter der Haut befindlichen Knochenstruktur macht (vgl. Parke und Waters, 1996 u. 2008).

Zusätzliche Erweiterungen dieses Modells unterscheiden drei Muskeltypen, die für das Gesicht modelliert werden müssen: linear, sheet, sphincter. Der lineare Muskel zieht sich in eine bestimmte Richtung zusammen. Der sheet Muskel zieht sich ähnlich dem linearen Muskel zusammen, ist aber ein paralleles Feld von Muskeln, das kontrahiert. Der sphincter Muskel zieht sich nur in einem Punkt zusammen. Der Benutzer kann dann die Aktivität der Muskeln ansprechen, um Gesichtsanimationen am Gesichtmodell hervorzurufen. Die Muskeln können entweder an der Oberfläche des Gesichtes oder unter der Haut, zum Beispiel an den Knochen, befestigt sein (vgl. Parent, 2002).

In diesem Zusammenhang ist das Facial Action Coding System (FACS) zu betrachten, mit dem sich die Interaktionen der einzelnen primären Muskelaktionen beschreiben lassen.

3.1.4 Facial Action Coding System

Das Facial Action Coding System wurde Ende der siebziger Jahre von Ekman und Friesen entwickelt und war zuerst nicht für Gesichtsanimationen gedacht. Allerdings hat es sich weit als solches verbreitet. FACS beschreibt die meisten Gesichtsmuskelbewegungen und deren Effekt auf die Gesichtsmimik. Diese Bewegungen werden Action Units (AUs) genannt. Ihre Kombination ermöglicht die Beschreibung aller Gesichtsausdrücke. Es gibt 46 AUs und jede AU ist eine minimale Bewegung, die sich nicht weiter unterteilen lässt. Dabei kann sich eine AU aus den Bewegungen mehrerer Muskeln

zusammensetzen. Zudem kann ein Muskel an mehr als einer AU beteiligt sein. FACS beschränkt sich auf die generellen Gesichtsbewegungen und vernachlässigt zu feine Veränderungen im Gesicht.

„FACS is concerned only with the description of facial motions, *not* in inferring what the motion means."(Parke und Waters, 1996 u. 2008). Es beschreibt also die Mimik, aber generiert sie nicht. FACS eignet sich besonders für die Bewegungen von Augenbrauen, Stirn und Augenlidern, aber beinhaltet nicht die Bewegungen des unteren Teils des Gesichtes. Demnach beschreibt FACS auch keine Sprache. Man hat nicht die Möglichkeit individuell Phoneme zum Ausdruck zu bringen. Dennoch wird FACS in vielen Animationssystemen verwendet und mit anderen Techniken kombiniert, aber meist werden nicht alle 46 AUs abgebildet (vgl. Parent, 2002; Parke und Waters, 1996 u. 2008).

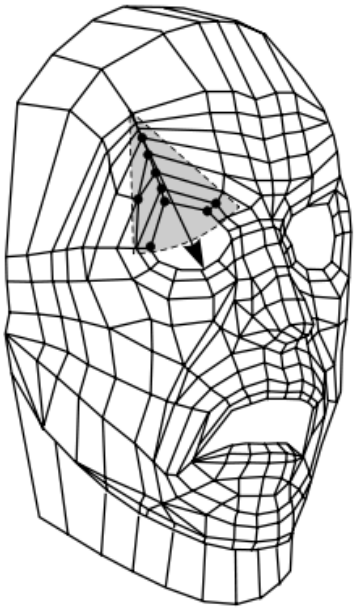


Abbildung 21: Erzeugung der AU „Outer Brow" mit einem linearen Muskel

3.2 Klassische Gesichtsanimation

Die klassische Gesichtsanimation knüpft an der in Kapitel 2.2 vorgestellten Modellierung von Hand an. Dabei ist darauf zu achten, dass man für die Bewegung modelliert. Hierzu werden die einzelnen Bereiche noch einmal aufgegriffen und mit den für die Animation nötigen Shapes und deren Verwendung erweitert.

Animatoren verlassen sich bei ihrer Arbeit meist auf Intuition, weswegen eine streng wissenschaftliche Betrachtung des Vorgangs schwer fällt. Das Vorgehen eines Animators ist zudem noch von Person zu Person unterschiedlich. Der Animator Jason Osipa (2003) teilt sein Vorgehen in sechs Schritte auf:

1. Animation des Mundes / der Lippen
2. Kopfbewegungen
3. Augenbewegungen
4. Bewegung der Augenlider
5. Bewegung der Augenbrauen
6. Finesse

Mund und Lippen. Die Animation von Mund und Lippen wird durch die Viseme (s. Kapitel 3.2.1) erzielt. Dies kann als Vorarbeit durch den Animator geleistet werden, aber auch zur Laufzeit im Animationssystem geschehen.

Kopfbewegungen. Die Kopfbewegungen unterstreichen die Lautstärke und Tonhöhe beim Sprechen und bestärken zusätzlich verschiedene Emotionen. Beim plötzlichen Ansteigen von Lautstärke und Tonhöhe folgt nach Osipa ein Anheben des Kopfes. Sinken Lautstärke und Tonhöhe wieder, bewegt sich der Kopf in seine Ausgangsposition zurück. Kopfbewegungen nach unten können eine Betonung ebenso verdeutlichen.

Augenbewegungen. Augenbewegungen hängen meist vom Kontext des Gesprochenen ab. Während eines Dialogs fokussieren sie oft den Dialogpartner, in einer Gruppe wandert der Blick. Durch einen Blick nach oben kann zum Beispiel Nachdenklichkeit symbolisiert werden.

Augenlider. Durch Augenlider können Emotionen ausgedrückt werden. Bei aufgerissenen Augen (Pupille und Iris sind komplett sichtbar) wirkt das Gesicht alarmiert. Sind Pupille und Iris durch die Augenlider teils verdeckt, wirkt das Gesicht schläfrig.

Augenbrauen. Bewegungen der Augenbrauen können Emotionen unterstreichen und verstärken. Dabei ist das Zusammenspiel mit den anderen Bereichen des Gesichtes wichtig.

Finesse. Osipa kontrolliert in einem letzten Schritt noch einmal alle Elemente und rundet ihr Zusammenspiel ab. Dies ist in Animationssystemen, die die Bewegungen zur Laufzeit generieren, allerdings nicht möglich.

3.2.1 Viseme

Der Mensch kann viele Laute von sich geben, die er mit dem Mund erzeugt und dabei unbedingt eine visuelle Veränderung des Mundes fordern. Diese Laute, die nur durch den Mund mit bestimmten Ausprägungen erzeugt werden können, nennen sich Viseme. Die einfachsten vier Viseme sind:

- Mund offen
- Mund geschlossen
- Mund/Lippen weit oder breit
- Mund/Lippen schmal

Osipa beschreibt diese vier Viseme als zwei deutlich getrennte Sprachkreisläufe, die Auf- und Zubewegung im Kiefer als Ersten und die Breit- und Schmalbewegung der Lippen als Zweiten. Beide Zyklen müssen nicht unbedingt zur selben Zeit ablaufen und auch nicht von einem Extrem ins andere übergehen. Die Öffnen- und Schließenbewegung kommt bei fast jedem Ton vor, der erzeugt wird. Die Schmal- und Breitbewegung steht eher in Verbindung mit der Art des Tons, der erzeugt wird. Zur Veranschaulichung ist der Satz „Why are we watching you?“ in seine einzelnen Wörter zerlegt worden und zusammen mit den beiden Sequenzen für die Breit- und Schmal- sowie Auf- und Zubewegung aufgelistet (vgl. Osipa, 2003).

Wort	**Breit/Schmal-Sequenz**	**Offen/Geschlossen-Sequenz**
why	schmal, breit	zu, auf, zu
are	keine Veränderung	zu, auf, zu
we	schmal, breit	zu, leicht auf
watching	schmal, etwas breit	zu, auf, zu, leicht auf, zu
you	schmal	zu / keine Veränderung

Die zwei vorgestellten Sprachzyklen sind die Grundlage für alle weiteren Viseme. Es besteht die Möglichkeit, Viseme auf Phoneme abzubilden. Für jedes Phonem ein Visem zu erstellen macht aber wenig Sinn, da es zum Beispiel in der Englischen Sprache ungefähr 38-45 Phoneme (vgl. Osipa, 2003; Parke und Waters, 1996 u. 2008, S. 9; S. 262) gibt und trotz verschiedener Töne kein Unterschied in der visuellen Rückmeldung zu erkennen ist. Daher ist es sinnvoll, visuell nicht unterscheidbare Phoneme für das entsprechende Visem zusammenzufassen. Abbildung 19 auf Seite 16 gibt einen Überblick über verschiedene Viseme neben Offen/Geschlossen sowie Breit/Schmal, die jeweils Gruppen von Phonemen zugeordnet sind. Osipa unterteilt die Viseme noch genauer, ähnlich der in Abbildung 19 gezeigten Viseme. Dabei geht er von neun Grundformen des Mundes aus, mit denen alle Viseme gebildet werden können. In Abbildung 22 sind die neun Grundformen dargestellt. Das „Smile" ist hierbei die Grundform „Weit". Die durch die Grundformen erzeugbaren Viseme sind in sechs Viseme-Laute und fünf Nicht-Viseme-Laute unterteilt:

Viseme-Laute	**Nicht-Viseme-Laute**
- B / M / P / Geschlossen	- L / N
- EE / Weit	- D / SH / TH / NG / J (Weiches G) / H
- F / V	- AW / OH / UH
- OO / Schmal	- EH / AH / UH
- IH / T / S	- Hartes G / K
- R	

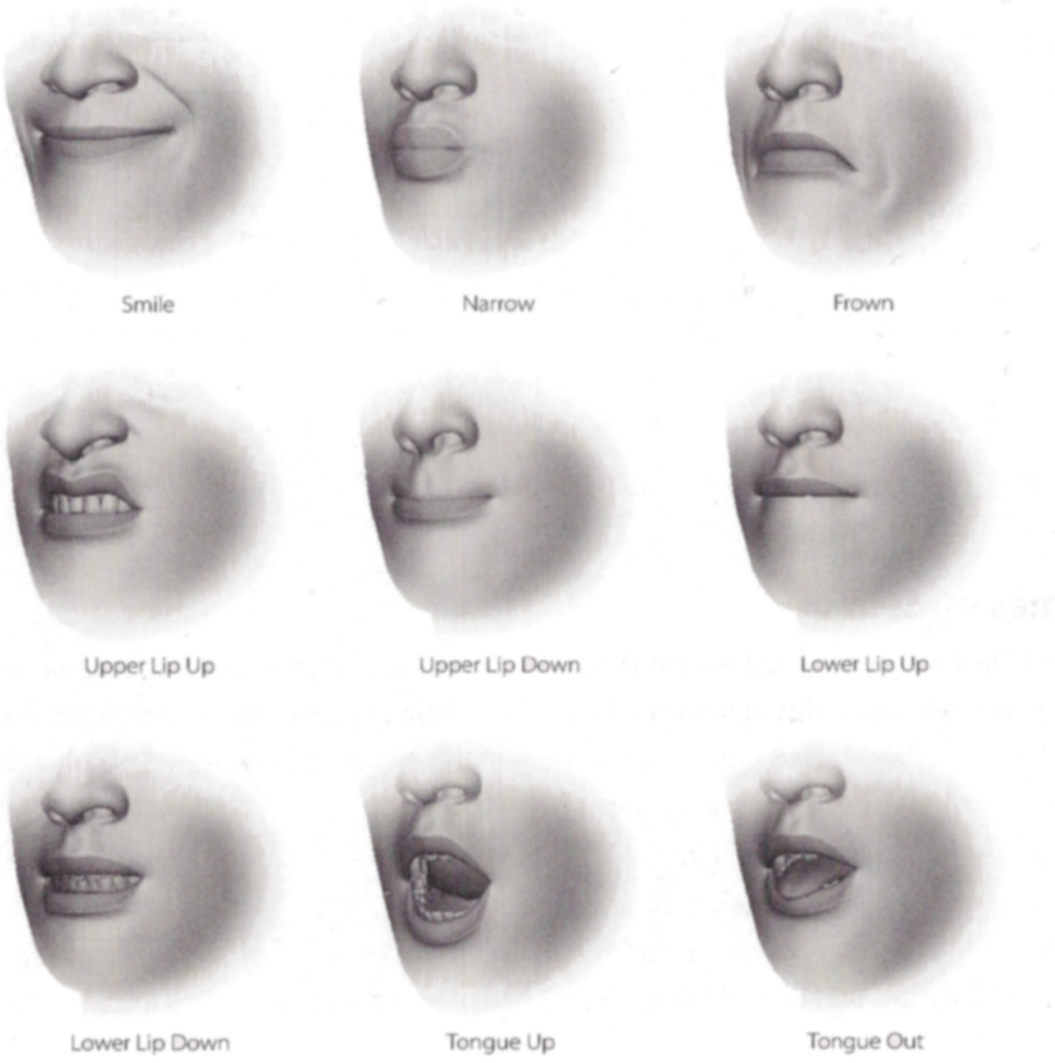


Abbildung 22: Neun Grundformen des Mundes zum Bilden aller vorhandenen Viseme

Ein Ansatz nach Walther beschreibt ebenfalls die Verbindung zwischen Visemen und Phonemen. Die Viseme sind die visuell unterscheidbaren Phonemklassen. Ein weiterer Ansatz nach Nitchie diskutiert die Lippenbewegungen basierend auf 18 Visemen unter Verwendung der Lippen, Zähne und Zunge. Die verschiedenen Viseme nach Nitchie sind in Abbildung 23, und eine Auflistung und Beschreibung dieser Viseme in Abbildung 24, zu sehen.

Viseme-Laute

B / M / P / Geschlossen: Dieses Visem hat viele Formen. Wenn der Mund geschlossen ist, kann es ein Lächeln sein, der Mund könnte schmal oder die Mundwinkel gesenkt sein. Wichtig ist bei diesem Visem also, dass der Mund geschlossen ist.
EE / Weit: Dieses Visem beeinflusst normalerweise nur die Weite, aber in manchen Situationen können auch die Zähne zu sehen sein oder der Mund ist offen. Ein Lächeln ist zum Beispiel gleichzeitig ein geschlossenes und ein weites Visem.
F / V: Das „F / V“ Visem basiert auf der Position der Lippen. Es gibt verschiedene Variationen, zum Beispiel wird die obere Lippe etwas hochgeschoben und der Kiefer wird weiter geschlossen als das geschlossene Visem. Gibt es bei dem Gesicht keine Lippen, kann das Visem als halb-geschlossen betrachtet werden.
OO / Schmal: Das Visem für schmal ist nicht kombinierbar mit Lächeln oder gesenkten Mundwinkeln. In seiner Grundform ist es geschlossen.
IH / T / S: Generell ist bei diesem Visem der Mund weiter. Es kommt aber auf den Kontext an, so dass der Mund in manchen Situationen auch schmaler sein kann.
R: „R“ ist das Gegenteil von „IH / T / S“ und somit also schmaler. Besitzt das Gesicht eine Zunge, sollte man sie bei diesem Visem benutzen. Ein R kann mit jeder Mundform ausgesprochen werden. Es gibt aber keine „*korrekte*“ Form eines Visems, nur eine korrekte Form in Bezug zu anderen Formen.
(vgl. Osipa, 2003, S. 49ff)

Nicht-Viseme-Laute

L / N: L und N werden hauptsächlich mit Hilfe der Zunge ausgesprochen. Allerdings ist, wie schon erwähnt, die Verwendung der Zunge nicht immer notwendig oder wünschenswert. Die Form des Mundes ist beliebig, allerdings muss der Mund etwas geöffnet sein. Diese Nicht-Viseme-Laute lassen sich den Visemen IH oder R zuordnen.

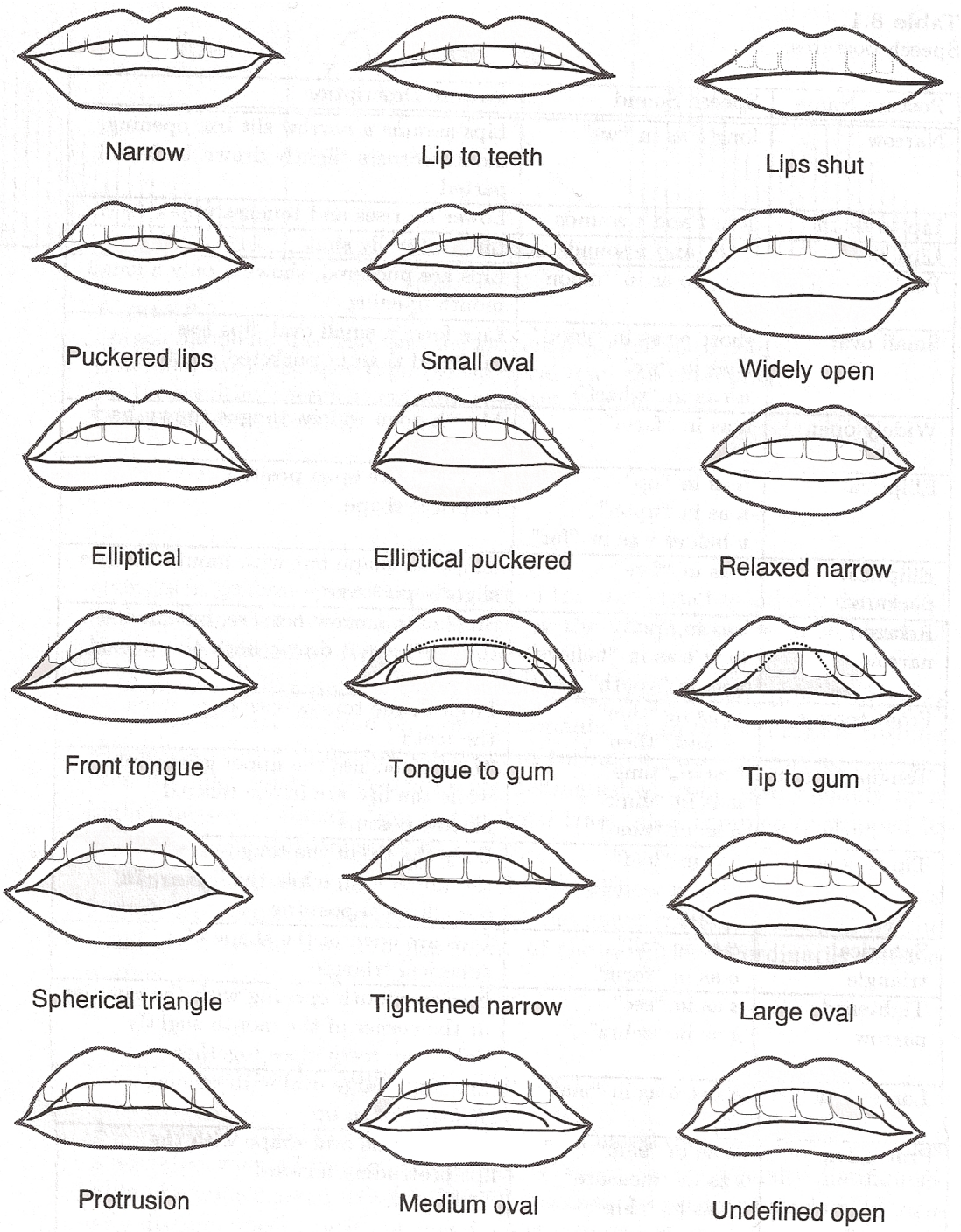


Abbildung 23: Die verschiedenen Viseme nach Nitchie

Posture Name	Speech Sound	Posture Description
Narrow	long *e* as in "we"	Lips assume a narrow slit like opening, mouth corners slightly drawn back and parted
Lip to teeth	{em f and *v* sounds	Lower lip rises and touches upper teeth
Lips shut	*b*, *m*, and *p* sounds	Lips naturally shut
Puckered lips	long *oo* as in "moon"	Lips are puckered, showing only a small mouth opening
Small oval	short *oo* as in "good" *w* as in "we" *wh* as in "wharf"	Lips form a small oval, lips less puckered than in puckered posture
Widely open	*a* as in "farm"	Mouth open widely, tongue drawn back slightly
Elliptical	*u* as in "up" *u* as in "upon", *u* before *r* as in "fur"	Intermediate open position with elliptical shape
Elliptical puckered	*r* as in "free"	Elliptical shape but with mouth corners slightly puckered
Relaxed narrow	*i* as in "pit" long *e* as in "believe" *y* as in "youth"	Similar to narrow posture, but mouth corners are not drawn back and parted
Front tongue	*th* as in "thin" and "then"	Front of the tongue visible between the teeth
Tongue to gum	*t* as in "time" *d* as in "dime" *n* as in "nine"	Tongue touches the upper gum while while the lips are in the relaxed narrow posture
Tip to gum	*l* as in "leaf"	Only the tip of the tongue touches the upper gum while the lips are in the elliptical posture
Spherical triangle	*a* as in "all" *o* as in "form"	Lips are open in the shape of a spherical triangle
Tightened narrow	*s* as in "see" *z* as in "zebra"	Narrow mouth opening with the muscles at the corner of the mouth slightly tightened, teeth close together
Large oval	short *a* as in "mat"	Lips form large oval with corners slightly drawn up
Protrusion	*sh* as in "ship" *s* as in "measure" *ch* as in "chip" *j* as in "jam" *g* as in "gentle"	Mouth forms oval shape with the lips protruding forward
Medium oval	short *e* as in "let" *a* as in "care"	Similar to the elliptical posture but with the mouth corners further apart
Undefined open	*k* as in "keep" *g* as in "go" *nk* as in "rank" *ng* as in "rang"	Mouth open with shape similar to the closest associated vowel

Abbildung 24: Tabelle mit Visemen, gesprochenem Ton und Beschreibung

D / SH / TH / NG / J (Weiches G) / H: Diese nicht-Viseme können genauso wie „IH / T / S“ behandelt werden. Beim TH kommt die Zunge ins Spiel. Die wichtigste Veränderung bei allen Nicht-Viseme-Lauten dieser Kategorie ist, dass der Mund weiter wird.
AW / OH / UH: Diese Laute lassen sich durch eine Kombination aus variierenden Anteilen von schmalen und offenen Visemen darstellen.
EH / AH / UH: Bei diesen Lauten handelt es sich um das Zwillingspaar zu „AW / OH / UH“. Statt mit schmalen Visemen kann man diese durch eine Kombination von weiten und offenen Visemen darstellen.
Hartes G / K: Diese Laute haben keine Form. Sie werden im Rachen erzeugt.
(vgl. Osipa, 2003, S.52ff)

3.2.2 Stellungen der Augenbrauen

Augenbrauen untermalen Emotionen, wie in Kapitel 3.2 thematisiert. Deshalb gibt es auch einige Variationen von Position und Bewegung der Augenbrauen. Osipa unterteilt die verschiedenen Stellungen in zwei Oberkategorien. Zum einen die „Realistischen Stellungen der Augenbrauen“, die bei einem menschlichen Gesicht von Nutzen sind und zum anderen die „Stilisierten Stellungen der Augenbrauen“ für Cartoongesichter. Die verschiedenen Stellungen der Augenbrauen sind insgesamt weniger als die des Mundes und in ihrem Rahmen auch einfacher.

Realistische Stellungen der Augenbrauen

Die realistischen Stellungen der Augenbrauen setzen sich aus 6 verschiedenen Stellungen zusammen:

1. Augenbrauen außen oben
2. Augenbrauen unten
3. Augenbrauen mittig oben
4. Augenbrauen mittig unten
5. Augenbrauen zusammengedrückt
6. Blinzeln

Die erste Stellung beschreibt eine Variante der Erhebung der Augenbrauen, bei der sich aber fast nur der äußere Teil der Brauen bewegt. Mit dieser Form kann man das Gesicht unbegeistert wirken lassen. Die zweite Form ist das Gegenteil der ersten. Hierbei werden die Augenbrauen außen gesenkt. Allerdings existiert diese Form in der Realität nicht, denn ein Mensch kann seine Augen nicht senken ohne sie zusammenzudrücken. Dennoch benötigt man diese Form, denn die Augenbrauen zusammenzudrücken, ohne sie zu senken, gelingt. Die Form „Augenbrauen mittig oben“, Abbildung 25 (c), sorgt dafür, dass der mittlere Bereich der Augenbrauen sich erhebt. Ähnlich wie die zweite Form ist diese Form nicht möglich, ohne das Zusammendrücken der Augenbrauen. Sie symbolisiert Traurigkeit. Die vierte Stellung ist das Gegenteil der dritten Stellung und lässt das Gesicht böse wirken. Das Zusammendrücken der Augenbrauen stellt das Gesicht nachdenklich dar. Zudem wird diese Stellung für die zweite und dritte Stellung benötigt um sie realistisch erscheinen zu lassen. Die letzte Stellung lässt das menschliche Gesicht blinzeln. In Abbildung 25 sind die Stellen der einzelnen Stellungen der Augenbrauen zur Veranschaulichung hervorgehoben.

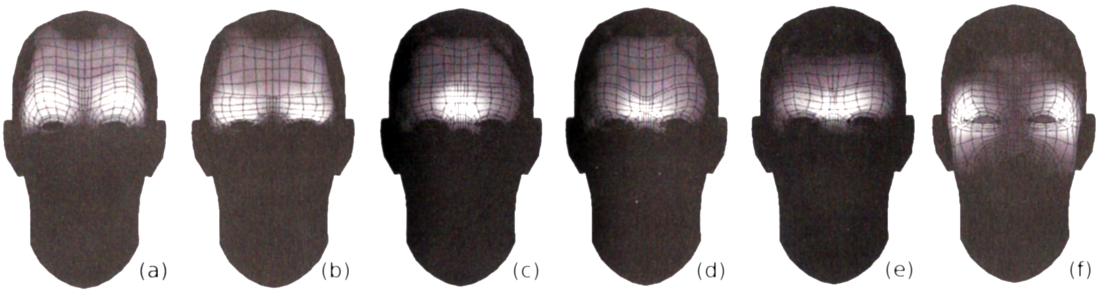


Abbildung 25: Sechs Stellungen der menschlichen Augenbrauen: (a) Augenbrauen außen oben; (b) Augenbrauen unten; (c) Augenbrauen mittig oben; (d) Augenbrauen mittg unten; (e) Augenbrauen zusammengedrückt; (f) Blinzeln

Stilisierte Stellungen der Augenbrauen

Die stilisierten Stellungen der Augenbrauen unterteilen sich in dieselben 6 Stellungen wie die realistischen Stellungen der Augenbrauen. Die „Augenbrauen außen oben“ können hier generell als erhobene Augenbrauen angesehen werden. Die gesenkten Augenbrauen haben bei Cartoongesichtern einen gewünschten Effekt. Die dritte Stellung symbolisiert Traurigkeit, auch ohne das Zusammendrücken der Augenbrauen. Die nächste Stellung steht für Wut. Das Zusammendrücken ist bei Cartoongesichtern nicht zwingend notwendig (s. Abbildung 26 (e)), kann bei manchen Gesichtern aber

hilfreich sein. Das Blinzeln kann bei Cartoongesichtern freudige Augen bedeuten. In Abbildung 26 sind Beispiele der Stellungen der Augenbrauen für stilisierte Gesichter.

Abbildung 26: Sechs Stellungen der stilistischen Augenbrauen: (a) Augenbrauen außen oben; (b) Augenbrauen unten; (c) Augenbrauen mittig oben; (d) Augenbrauen mittg unten; (e) Augenbrauen zusammengedrückt, hier neutral; (f) Blinzeln

Abriss

Es ist also notwendig, zwischen den Stellungen der Augenbrauen menschlicher Gesichter und Cartoongesichtern zu unterscheiden. Durch die in Kapitel 3.1.1 vorgestellte Interpolationstechnik kann jede Stellung miteinander kombiniert und verschiedene Augenbrauenbewegungen erzeugt werden.

3.3 Zusammenfassung

Die Animation von Gesichtern bietet also eine große Bandbreite an Techniken und Ansätzen. Das in den siebziger Jahren entwickelte Parametrische Modell eignet sich besonders für die Erzeugung und Animation menschlicher Gesichter. Auch der muskelbasierte Ansatz für hauptsächlich menschliche Gesichter ermöglicht das Verändern der Gesichtsmimik. Mit FACS als Grundlage oder verschiedenen Motion Capture Techniken können täuschend echte Gesichter und Bewegungen erzeugt werden. Dennoch bleibt die realistische Gesichtsanimation eine der größten Herausforderungen in der Computergrafik.

Die Interpolationstechnik, die in dieser Arbeit eine große Rolle spielt, eignet sich besonders für die Erzeugung von Animationen für stilistische Gesichter und ist deshalb auch wegen ihrer Einfachheit in ihrer Verwendung weit verbreitet. Die in der Interpolationstechnik verwendeten Shape Keys sind eine Option die Viseme abzubilden und zu animieren. Die richtigen Viseme sind der Schlüssel zum Erfolg einer überzeugend wirkenden Mundbewegung. Zur Erzeugung von Emotionen ist besonders das Zusammenspiel von Mund und Augenbrauen ausschlaggebend. Daher sollte neben der synchronen Mundbewegung zur Sprachausgabe auch die Emotion durch Zusammenarbeit von Mund und Augenbrauen zum Ausdruck gebracht werden, denn besonders Cartoongesichtern wird durch Emotionen Leben eingehaucht.

4 Grundlagen der textgesteuerten Sprachsynchronisierung

Gesichtsanimationen und Sprache können entweder direkt von einem Animator synchronisiert werden oder automatisch durch eine Software. Die wohl interessantere Methode ist es, einen Text einzugeben und automatisch Sprache und Gesichtsanimationen gleichzeitig synchronisieren zu lassen. Dieses Vorgehen wird im folgenden Kapitel näher betrachtet und erläutert. Zunächst wird eine Einführung in die Sprachsynthese einschließlich der Phoneme vorgenommen. Anschließend werden noch weitere Verfahren der Sprachsynthese und Ansätze beschrieben.

Die textgesteuerte Sprachsynchronisierung erwartet die Eingabe eines Textes, der dann automatisch in die dazugehörigen phonetischen Symbole übersetzt wird. Dabei nutzt das System zusätzliche Informationen wie Sprachrhythmus und Betonung sowie eine Reihe von Regeln, die eine genaue Modellierung der natürlichen Sprache ermöglichen (vgl. Parke und Waters, 1996 u. 2008).

Ein System zur Sprachsynthese ist das Opensource *Festival Speech Synthesis System*[3] der Universität Edinburgh. Dieses bietet ein generelles Framework zur Erstellung von Sprachsynthese Systemen und bietet die Synchronisierung von Text zu Sprache (vgl. Festival, 12.04.2011).

4.1 Sprachsynthese

„Speech synthesis is a process which artificially produces speech for various applications, diminishing the dependence on using a person's recorded voice."(Furui, 2001, S. 213). Die Sprachsynthese ermöglicht es, eine Maschine sprechen zu lassen. Dabei ist es notwendig, die menschliche Sprache bezüglich der Phoneme (s. Kapitel 4.1.1) und der Prosodie (s. Kapitel 4.1.2) zu analysieren. Methoden wie zum Beispiel das Linear Predictive Coding (s. Kapitel 4.2) haben bei der Forschung der Sprachsynthese geholfen.
Man kann drei verschiedene Sprachsynthese Methoden unterscheiden:

1. **Waveform Kodierung**
   Sprachwellen einer aufgenommenen menschlichen Stimme werden nach der Waveform Kodierung gespeichert oder direkt nach der Aufnahme zur Erzeugung einer gewünschten Nachricht verwendet.

[3]Festival: `http://www.cstr.ed.ac.uk/projects/festival/` (zuletzt aufgerufen am 27.04.2011)

2. **Analyse-Synthese Methoden**
   Sprachwellen einer aufgenommenen menschlichen Stimme werden anhand der Analyse-Synthese Methode in Parametersequenzen transformiert und gespeichert. Anschließend wird ein Sprachsynthesizer verwendet, in dem dieser die erzeugten Parameter zu einer Nachricht verbindet.

3. **Synthese nach Regeln** (Text-to-Speech Synthesis)
   Sprache wird basierend auf phonetischen und linguistischen Regeln von Buchstabensequenzen oder von Phonemsequenzen und prosodischen Eigenschaften erzeugt.

Jede dieser drei Methoden hat ihre Vor- und Nachteile und man sollte die Methode entsprechend des Vorhabens wählen (vgl. Furui, 2001).

In dieser Arbeit wurde die dritte Methode, die Synthese nach Regeln, gewählt und ist in Kapitel 4.3 näher beschrieben.

4.1.1 Phoneme

Phoneme bilden die kleinste Einheit um zu beschreiben, wie Sprache linguistische Bedeutung vermittelt. Dabei ist ein Phonem eine Gruppe von ähnlichen, aber nicht identischen Lauten, die sich von einem zum anderen entsprechend ihres Kontexts unterscheiden. Ein Phonem ist kein Laut, sondern eine Abstraktion, die ein Menge von Lauten beschreibt. Zum Beispiel hört sich das Phonem /p/ in den Wörtern „pit“ und „spit“ ähnlich an, aber dennoch unterscheiden sich die Wörter aufgrund der umgebenden Laute. Man kann also sagen, ein Phonem ist eine Gruppe von Lauten, die sich für den Sprecher gleich anfühlen, die nicht zur Unterscheidung zwischen Wörter genutzt werden können und sich entsprechend ihres Kontexts unterscheiden (vgl. Fallside und Woods, 1985).

4.1.2 Prosodie

Prosodie oder auch Satzrhythmus ermöglicht es, eine synthetische Sprache natürlicher klingen zu lassen. Prosodie umfasst die Stimmlage, Lautstärkeschwankungen und andere Aspekte der Sprache, die über Phonemsequenzen und Aussprache hinausgehen. Es ist nachgewiesen worden, dass korrekte Prosodie das Verständnis der Sprache erleichtert. Prosodie hängt zudem vom Geschlecht des Sprechers ab. Auch der Akzent ist ein prosodisches Merkmal (vgl. Schroeder, 1999). Die Prosodie wird in dieser Arbeit nicht weiter berücksichtigt.

4.2 Linear Predictive Coding

Der Begriff Linear Predictive Coding (LPC) wurde zum ersten Mal von Wiener verwendet. LPC ist ein extrem effizienter Ansatz zur Sprachsynthese. Es dient als Grundlage für viele moderne Sprachausgabesysteme und kann entweder zur Sprachsynthese oder zur Analyse von natürlicher Sprache genutzt werden. Verschiedene Variationen dieser Form der Analyse oder Synthese werden im Allgemeinen für die Sprachkompression verwendet. Das Besondere am LPC ist, dass durch eine geringe Anzahl an Parametern, die Sprachwelle und das Spektrum effizient und präzise repräsentiert werden können.

Eine künstliche Sprache kann erzeugt werden, wenn das Reizungssignal eine synthetisch generierte oder zufällige Impulsfolge ist. Im Fall der synthetischen Reizung ist es einfach, die Sprache zu beschleunigen oder zu verlangsamen. Durch die Anregungsfunktion wird nur die Stimmlage kontrolliert und somit gibt es keine Tonhöhenverschiebung.

Als zuverlässigster Analyseansatz ist das Restdifferenzsignal zwischen der originalen Sprache und der Ausgabe des Linear Prediction Filters als Reizungssignal zur Synthese zu benutzen. Das Restdifferenzsignal approximiert eine unkorrelierte Störung für Konsonanten und geflüsterte Vokale. Zudem nähert es für stimmhafte Vokale eine Impulsfolge an. Zusammen mit dem Linear Prediction Filter kodiert das Restdifferenzsignal einen Großteil der Information in der originalen Sprache. Die neusynthetisierte Sprache ist besonders gut verständlich und behält den originalen Tonfall und Rhythmus, zudem ist sie besonders gut für Computeranimationen geeignet (vgl. Furui, 2001; Parke und Waters, 1996 u. 2008).

4.3 Sprachsynthese nach Regeln

Die Sprachsynthese nach Regeln ist eine Methode, um jedes Wort oder jeden Satz basierend auf einer Sequenz phonetischer Symbole oder Buchstaben zu erzeugen. Dabei werden zum Beispiel Phoneme gespeichert und nach bestimmten Regeln miteinander verbunden. Zur selben Zeit werden Tonlage und Amplitude kontrolliert (vgl. Furui, 2001).

„Voice sound changes result directly from movements of the vocal articulators including the tongue, lips, and jaw rotation which in turn cause changes in facial posture.“(Parke und Waters, 1996 u. 2008, S. 272). Als Ergebnis dieser Gesichtsbewegungen können Programme für die Sprachsynthese durch Informationen für jede Sprechweise ergänzt werden, um die entsprechenden Parameter für das Gesicht zu kontrollieren. Hierzu

gibt es zwei Ansätze unter Verwendung von Phonemen:

Bei dem Ansatz nach Pearce et al. wird das Phonemskript direkt vom Animator spezifiziert. Zum Beispiel lässt sich der Satz „*Speak to me now bad kangoroo*“ in folgende phonetische Sequenz zerlegen:

s p ee k t u m in ah uu b aa d k aa ng g uh r uu

Der Ansatz nach Hill et al. generiert die Phonemsequenz aus einem eingegebenen Text. Allerdings ist es schwierig, natürlichen Rhythmus und deutliche Artikulation zu erreichen, wenn die Sprache aus einem Phonemskript erzeugt wurde. Die Qualität der synthetisierten Sprache kann aber durch Hinzufügen von Tonhöhen, Zeit und Lautstärke Annotationen im Phonemskript verbessert werden.

Bei der Verwendung von Sprachsynthese Algorithmen wird die Phonemsequenz zur Generierung der Sprache genutzt. Zudem wird die Phonemsequenz verwendet um ein parameterisiertes Gesichtsmodell zu kontrollieren. Demnach werden die Sprachsynthese Algorithmen so erweitert, dass nicht nur die unterschiedlichen Parameter für die Sprachsynthese erzeugt, sondern auch die sichtbaren Attribute der Lautbildung eines gerenderten Gesichts generiert werden. Dadurch ist eine perfekte Synchronisation zwischen dem Gesichtsausdruck beim Sprechen und der generierten Sprache garantiert. Beide Prozesse werden unabhängig voneinander generiert und anschließend zusammengefügt, um die abschließende Sprachanimation zu bilden (vgl. Parke und Waters, 1996 u. 2008).

4.4 Waters' Echtzeit Ansatz

Zusammen mit Levergood entwickelte Waters ein Echtzeit synchronisiertes Visual Speech System, das auf der Software Version des *DECtalk*[4] Sprachsynthesizer basiert. DECtalk ist eine kommerzielle Variante des Festival Sprachsynthesizers und wurde in den achtziger Jahren entwickelt. DECtalk ist bekannt durch Stephen Hawking, der viele Jahre einen DECtalk Synthesizer zum Sprechen benutzte. Die Software erwartete die Eingabe des zu sprechenden Textes, der anschließend zur lexikalischen Analyseeinheit des DECtalk Synthesizers geschickt wurde. Diese Einheit produzierte eine zeitlich

[4]DECtalk für Linux: `http://www.fonixspeech.com/dectalk_linux.php` (zuletzt aufegerufen am 28.04.2011)

festgelegte Phonemsequenz und weitere Kontrollparameter zur Generierung der Sprachausgabe. Waters und Levergood fingen die zeitlich festgelegte Phonemsequenz ab und generierten daraus eine parallel ablaufende Echtzeit Gesichtsanimation, die sich automatisch mit der generierten Sprache synchronisierte. Dieser Ansatz ist in Abbildung 27 noch einmal verdeutlicht (vgl. Parke und Waters, 1996 u. 2008).

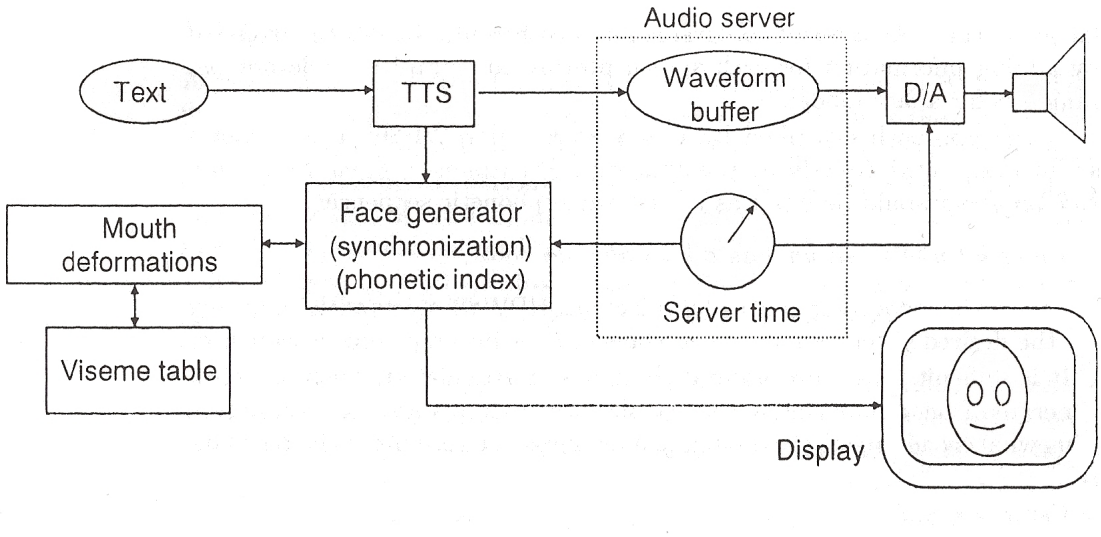


Abbildung 27: Waters' Echtzeit Ansatz der synchronisierten visuellen Sprachsynthese

4.5 Zusammenfassung

Es gibt verschiedene Ansätze der Sprachsynthese. Der in diesem Kapitel vorgestellte Ansatz Sprache nach Regeln zu erzeugen ist einer der interessantesten. Durch Verbindung von Phonemen und anderen lingustischen Merkmalen sowie der Prosodie ist es möglich, eine künstliche Sprache zu erzeugen und eine Maschine sprechen zu lassen. Die vorgestellten Grundlagen sind wichtig, um einem computeranimierten Gesicht eine Stimme zu geben. Mit dem Wissen über Sprachsynthese und Viseme kann dann das Gesicht beim Sprechen überzeugend wirken.

5 Konzept

In diesem Kapitel wird das Konzept des in der vorliegenden Arbeit entwickelten Talkingheads für den Roboter Lisa der Arbeitsgruppe Aktives Sehen der Universität Koblenz vorgestellt. Zuerst werden die Ziele und Anforderungen festgehalten und verschiedene schon bestehende Ansätze eines Talkingheads vorgestellt. Anschließend wird die grundlegende Vorgehensweise des entwickelten Systems beschrieben, um dann anknüpfend auf dessen Aufbau und Struktur einzugehen.

5.1 Ziele und Anforderungen

Der Roboter Lisa der Arbeitsgruppe Aktives Sehen der Universität Koblenz benötigt ein neues Gesicht. Das vorherige Gesicht basiert auf *OpenGL*[5], eingebunden in *Qt*[6]. Es wurde durch verschiedene Zeichenroutinen direkt mit OpenGL erzeugt. Durch eigene Interpolation wurden die Animationen des Gesichtes ermöglicht. Augen- und Kopfbewegungen waren annehmbar. Allerdings waren die Bewegungen des Mundes beim Sprechen zufällig. Der Mund ging beim Sprechen nur auf und zu. Die Stimme bekam Lisa durch Festival. Dazu wurde ein Text eingegeben, aus dem Festival durch Extrahieren phonetischer Merkmale die Stimme erzeugt. Der gesprochene Text wurde unterstützend in einem kleinen Fenster unter dem Gesicht angezeigt. Emotionen waren kaum sichtbar. Es war eine generelle Grundfreundlichkeit zu erkennen und beim „Aufwachen" des Roboters war ein Lächeln zu sehen. Das Gesicht wurde mit Hilfe eines Displays, der am Roboter angebracht ist, angezeigt.

Das neue Gesicht wird diesen Display als Anzeige übernehmen. Es soll ein grundsätzlicher „Facelift" stattfinden, allerdings soll das Aussehen des Gesichts des Roboters unverändert bleiben und nur die Funktionen überholt werden. Eine wichtige neue Funktion ist, dass das neue Gesicht austauschbar sein soll. Es soll also die Möglichkeit geben, ein neues Gesicht zu entwerfen und es ohne große Probleme beliebig zu wechseln. Die Augen- und Kopfbewegungen sollen beibehalten werden. Die Mundbewegungen verlangen allerdings eine Überholung. Dazu sollen die zufälligen Mundbewegungen beim Sprechen soweit angepasst werden, dass sich der Mund synchron zum Gesprochenen bewegt. Lisas Stimme soll dabei möglichst ähnlich bleiben. Da *Festival* anhand von phonetischen Merkmale die Stimme erzeugt, kann die Stimme beibehalten werden. Auch

[5] OpenGL: `http://www.opengl.org/` (zuletzt aufgerufen am 29.04.2011)

[6] Qt Product: `http://qt.nokia.com/products/` (zuletzt aufgerufen am 29.04.2011);
Qt Dokumentation: `http://doc.qt.nokia.com/` (zuletzt aufgerufen am 29.04.2011)

die Anzeige des gesprochenen Textes wird als Hilfestellung beibehalten. Zusätzlich soll das überarbeitete Gesicht die Eigenschaft besitzen, Emotionen auszudrücken. Die Emotionen sollen an den zu synthetisierenden Text anfügbar und während des Sprechens aktiv sein sowie auch darüber hinaus.

Zusammengefasst lassen sich folgende Ziele und Anforderungen formulieren:

1. Das neue Gesicht ist austauschbar.
2. Das Talkinghead System erhält als Eingabe einen String, der mit Hilfe von Festival in seine phonetischen Merkmale zerlegt wird.
3. Der Mund bewegt sich synchron zum Gesprochenen, indem die phonetischen Merkmale zur synchronen Mundanimation verwendet werden.
4. Das Gesicht kann Emotionen zeigen, die dem Eingabestring als Emoticons beigefügt werden.

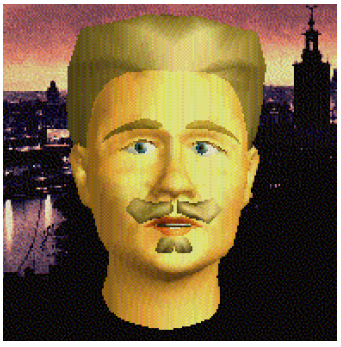
5.2 Vorhandene Ansätze

Im Laufe der Jahre wurden viele ähnliche Ansätze eines Talkingheads entwickelt. Hier sind das *August Dialogue System* des Centre for Speech Technology Schweden (Gustafson u. a., 1999; Lundeberg und Beskow, 1999), ein Facial Animation System von Albrecht u. a. (2002) aus Saarbrücken und das Text-to-Audio-Visual Speech System von Niswar u. a. (2009) zu nennen.

Das August Dialogue System des Centre for Speech Technology Schweden (Gustafson u. a., 1999; Lundeberg und Beskow, 1999) (s. Abbildung 28(a)) ist ein multi-modal sprechendes Dialogsystem mit einem animierten Vertreter, der zur Interaktion dient. Der Talkinghead kommuniziert über synthetische Sprache, Mimik und Kopfbewegungen. Er kann Emotionen zeigen und nutzt prosodische Merkmale. Darüber hinaus hat der Talkinghead eine Gedankenblase, die zusätzliche Informationen anzeigt. Dieser Ansatz basiert auf parametrisierten, verformbaren 3D Gesichtsmodellen, die durch ein Text-to-Speech Framework kontrollierbar sind. Dabei generieren entsprechende Regeln angewandt auf den Text die Parameter für das Gesicht. Es wird hier eine ähnliche Methode zur Parameterisierung verwendet wie schon bei Parke (1974).

Das Facial Animation System von Albrecht u. a. (2002) (s. Abbildung 28(b)) erzeugt Echtzeitanimationen mit Sprachsynchronisation und nonverbalen, sprachbedingten Gesichtsbewegungen durch einfache Texteingabe. Dazu wird als Text-to-Speech Komponente *Festival* verwendet, das den eingegebenen Text linguistisch analysiert und ein Sprachsignal aus phonetischen Informationen erzeugt. Die phonetischen Informationen werden zusätzlich dazu genutzt, die Sprachsynchronisation für die Gesichtsanimation zu steuern. Darüber hinaus fließen prosodische Merkmale ein, die die Mimik des Talkingheads beeinflussen. Außerdem verwendet dieser Ansatz Emoticons (Bsp.: `:-))` die in XML übersetzt werden und emotionale Gesichtsbewegungen auslösen.

Der Ansatz des Text-to-Audio-Visual Speech System von Niswar u. a. (2009) (s. Abbildung 28(c)) beschreibt einen 3D Talkinghead, der aus einem 3D Visemedatensatz konstruiert wird. Dazu wird der Visemedatensatz statistisch analysiert, um optimale lineare Parameter zur Kontrolle der Lippen- und Kieferbewegungen für den Talkinghead zu erhalten. Die erhaltenen Parameter entsprechen den low-level MPEG-4 FAPs (Facial Animation Parameters). MPEG-4 ist ein weiterer Kodierungsstandard wie das FACS (s. Kapitel 3.1.4), allerdings sind die FAPs eher muskelbasiert. Das parametrisierte Modell ist mit einem Text-to-Speech System verbunden, um audiovisuelle Sprache durch eingegebenen Text zu erzeugen. Zusätzlich sind während des Sprechens Kopfbewegungen und Augenzwinkern animiert.



(a) August Dialogue System des Center for Speech Technology Schweden. (Gustafson u. a., 1999)

(b) Facial Animation System von Albrecht u. a. (2002).

(c) Text-to-Audio-Visual Speech System von Niswar u. a. (2009).

Abbildung 28: Verschiedene Talkinghead Konzepte.

Der in dieser Arbeit entwickelte Ansatz orientiert sich an dem von Albrecht u. a. (2002) vorgestellten Konzept. Dieses verwendet eine Text-to-Speech Komponente, die gleichzeitig für Sprachsignal und Sprachsynchronisation der Gesichtsanimation zuständig ist. Zusätzlich werden Emoticons, die über den Text eingegeben werden, verarbeitet. Es werden allerdings keine weitere Kopfbewegungen wie bei Albrecht u. a. (2002) durch prosodische Merkmale erzeugt.

Die beiden anderen Ansätze verwenden parametrische Modelle, die in dieser Arbeit nicht ihren Einsatz finden. Zudem gibt es noch weitere ähnliche Ansätze, die hier nicht aufgeführt sind.

5.3 Grundsätzliches Vorgehen

Die grundlegende Vorgehensweise des Talkingheadsystems ist folgendermaßen beschrieben und ist vergleichbar mit Waters' Echtzeit Ansatz wie in Kapitel 4.4 vorgestellt:

Das Talkingheadsystem erhält als Eingabe einen Text in Form eines Strings. Dieser String wird von dem Text-to-Speech System Festival linguistisch analysiert und synthetisiert. Durch die Analyse des Textes erhält man verschiedene lingustische Merkmale wie zum Beispiel die Phoneme und auch Zeitstempel. Die komplette Sprachsynthese erzeugt eine Stimme für den Talkinghead und gibt den Text akustisch wieder. Vorher werden mit Hilfe der Phoneme die Viseme ausgewählt und entsprechend der mitgelieferten Zeitstempel zu einer Animationsequenz zusammengesetzt. Nebenbei wird der Text vom Talkinghead auf Emoticons überprüft. Vorhandene Emoticons werden als zusätzliche Key Frames zu den bestehenden Key Frames, die aus den Visemen erzeugt wurden, hinzugefügt. Zwischen den generierten Key Frames wird dann interpoliert. Anhand der Zeitinformation der Phoneme ist eine synchrone Wiedergabe von Sprache und Gesichtsanimation gegeben.

Abbildung 29 zeigt ein Aktivitätsdiagramm des oben beschriebenen Vorgehens.

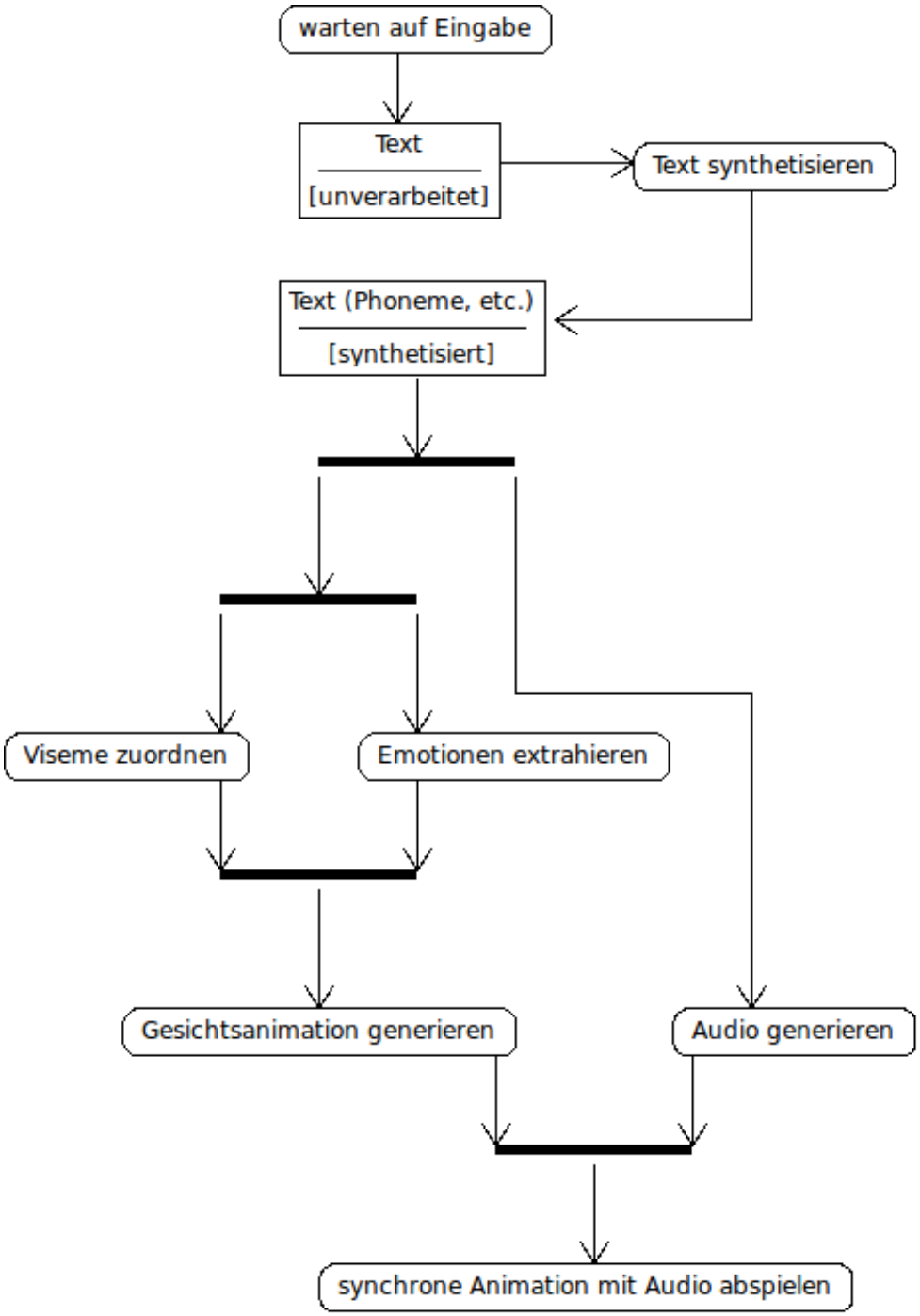


Abbildung 29: Aktivitätsdiagramm des Talkingheadsystems

5.4 Aufbau und Struktur

Der Aufbau strukturiert sich in zwei Teile. Im ersten Teil wird der Aufbau der Gesichtsmodellierung beschrieben, bei der darauf geachtet wird, für die Bewegung bzw. Animation zu modellieren. Der zweite Teil umfasst die Struktur des Talkingheadsystems mit seinen Modulen und deren Funktionalität.

5.4.1 Aufbau der Gesichtsmodellierung und -animation

Bei der Modellierung eines Gesichts für das in dieser Arbeit entwickelte Talkingheadsystem müssen einige Regeln beachtet werden, damit das erstellte Modell mit dem System kompatibel ist. Dazu ist im Anhang ein Template (Vorlage) zu finden, das den Vorgang mit der Open Source 3D Content Creation Suite *Blender* genaustens beschreibt. In diesem Kapitel wird konkret der Aufbau des neuen Gesichts für den Roboter Lisa beschrieben.

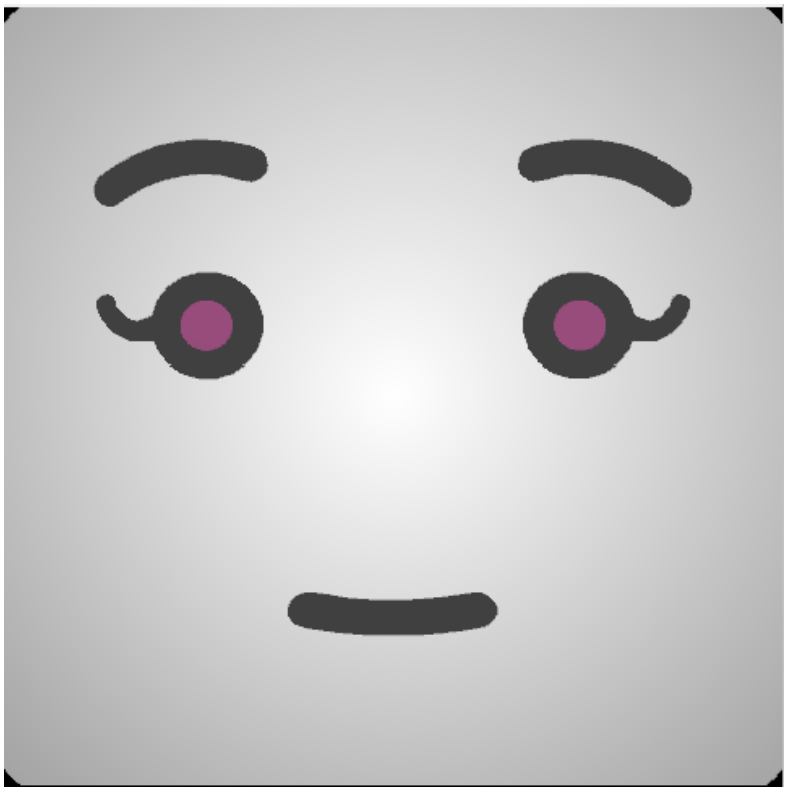


Abbildung 30: Neues Gesicht des Haushaltsroboters Lisa

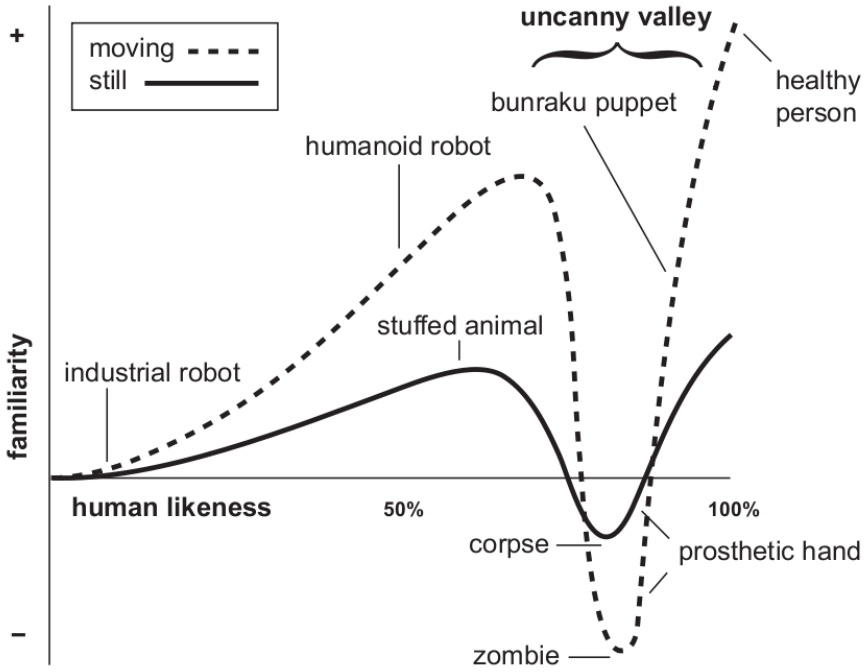


Abbildung 31: Der Effekt des Uncanny Valley nach Mori

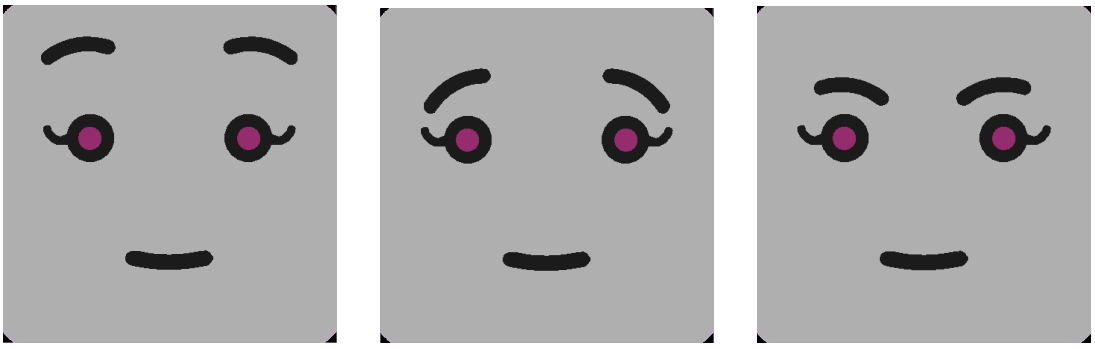
Das neue Gesicht ist in Abbildung 30 zu sehen. Es ist ein weiblich wirkendes Gesicht, da der Roboter auch einen weiblichen Namen trägt. Lisa (Lisa Is a Service Android) ist zudem ein Haushaltsroboter. Das Gesicht ist nicht realistisch gestaltet, sondern cartoonartig gehalten. Der Grund für die Wahl eines Cartoongesichts liegt den Forschungen der Osaka University nach dem „Uncanny Valley" zugrunde. Abbildung 31 beschreibt das Uncanny Valley nach dem Roboter Designer Mori. Es sagt aus, dass Roboter, die dem Menschen sehr ähnlich sehen, vertraulicher wirken, allerdings nur bis zu einem bestimmten Punkt, an dem schon dezente Veränderungen des Aussehens ein Gefühl von Seltsamkeit erzeugen. Sollte dieses Gefühl bei einem Roboter eintreten und dieser bzw. Teile davon *bewegen* sich noch zusätzlich, wird die Seltsamkeit verstärkt (vgl. MacDorman, 2005).

Man müsste also ein Gesicht erstellen, das entweder genauso aussieht wie ein Mensch und sich auch so bewegt, oder eben ein Cartoongesicht. Die Wahl fiel also nicht nur, weil in dieser Arbeit die Technik des Motion Capturing nicht zur Verfügung stand, auf ein Cartoongesicht, sondern auch aus dem Grund, dass einem Roboter nicht die Intelligenz eines Menschen zugetraut werden soll. Dennoch sollte ein Vertrauen zum Roboter ermöglicht werden, um angenehm mit ihm zu interagieren und kommunizie-

ren und das Wirken seiner Handlung erwartungsgemäß zu präsentieren. Deshalb ist ein Cartoongesicht mit seinen einfachen Bewegungen und seiner Simplizität in diesem Fall die bessere Wahl. Das Gesicht hat einen freundlichen Gesichtsausdruck und eine helle sowie freundliche Gesichtsfarbe. Es sind ein Mund zum Sprechen vorhanden, Augen die ab und zu blinzeln und Augenbrauen, mit denen Emotionen verstärkt werden können.

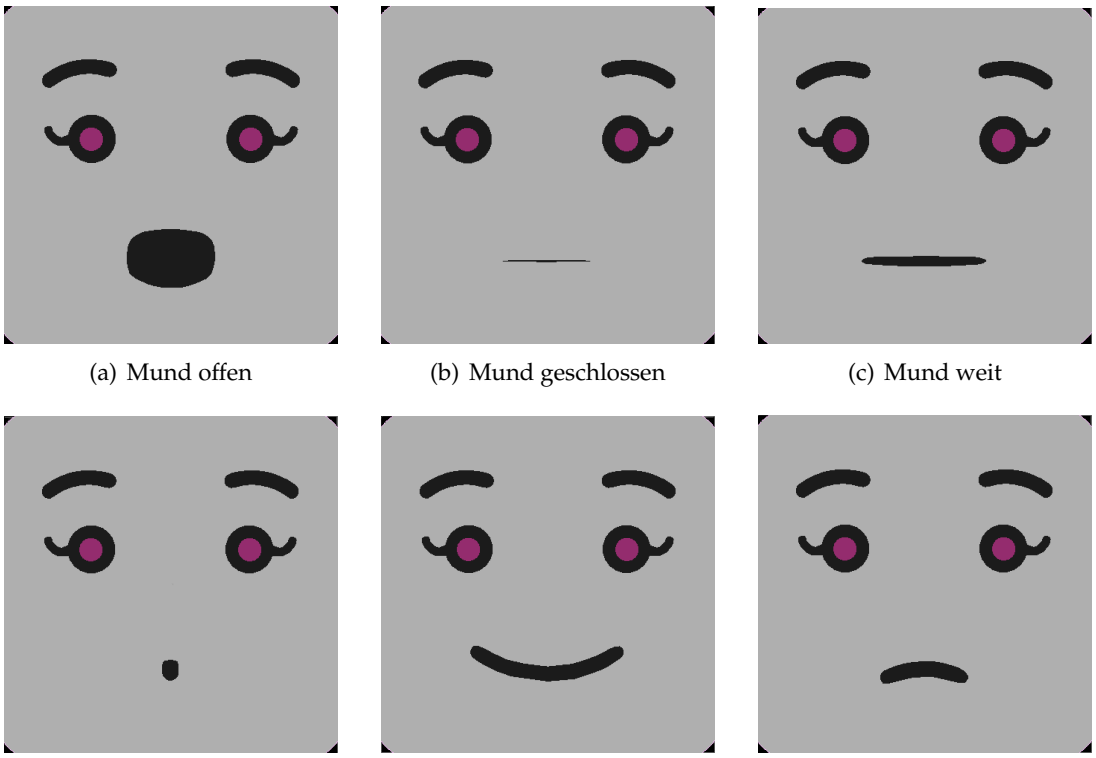
Da während der Modellierung für das Gesicht schon auf die Animationen geachtet werden muss, ist es notwendig, Shape Keys zu erstellen. Die Augenbrauen haben verschiedene Formen, wie in Kapitel 3.2.2 beschrieben. Es sind die Shape Keys „Augenbrauen nach oben“, „Augenbrauen mittig oben“ und „Augenbrauen mittig unten“ vorhanden (s. Abbildung 32). Das Zwinkern der Augen ist als feste Animation modelliert. Der Mund hat vier Shape Keys zum Sprechen, die vier wichtigsten Viseme (s. Kapitel 3.2.1), „Mund offen“, „Mund geschlossen“, „Mund weit“ und „Mund schmal“. Emotionen werden durch die Shape Keys „Mundwinkel nach oben“ und „Mundwinkel nach unten“ ermöglicht. In Abbildung 33 sind die verschiedenen Shape Keys aufgelistet. Zusätzlich sind noch zwei feste Animationen modelliert worden, die die Bewegung des Kopfes beeinflussen. Der Kopf bewegt sich gleichmäßig, aber dennoch erwartungsgemäß hoch und runter, um eine Atmung zu simulieren. In gewissen Abständen erfolgt eine Rotation des Kopfes, um eine Gelassenheit des Roboters auszustrahlen.

Die festen Animationen und die Interpolation der Shape Keys erfolgt im entwickelten System, dessen Aufbau und Struktur im folgenden Unterkapitel vorgestellt wird.



(a) Augenbrauen nach oben (b) Augenbrauen mittig oben (c) Augenbrauen mittig unten

Abbildung 32: Formen der Augenbrauen von Lisa



(a) Mund offen (b) Mund geschlossen (c) Mund weit

(d) Mund schmal (e) Mundwinkel nach oben (f) Mundwinkel nach unten

Abbildung 33: Die vier wichtigsten Viseme (a - d) und die Mundformen für die Emotionen (e - f), veranschaulicht am Gesicht des Roboters Lisa

5.4.2 Aufbau des Talkingheadsystems

Das Talkingheadsystem gliedert sich in drei Hauptmodule. In dem Modul `TalkingHead` werden das Modell und die Animationen verwaltet. Das Modul `FestivalSynthesizer` ist zuständig für die Erzeugung der phonetischen Merkmale, einschließlich Sprache und Stimme. Das `SpeechOutDisplay` zeigt den gesprochenen Text unter dem Talkingheadfenster an. In Abbildung 34 ist eine Übersicht der Module dargestellt.

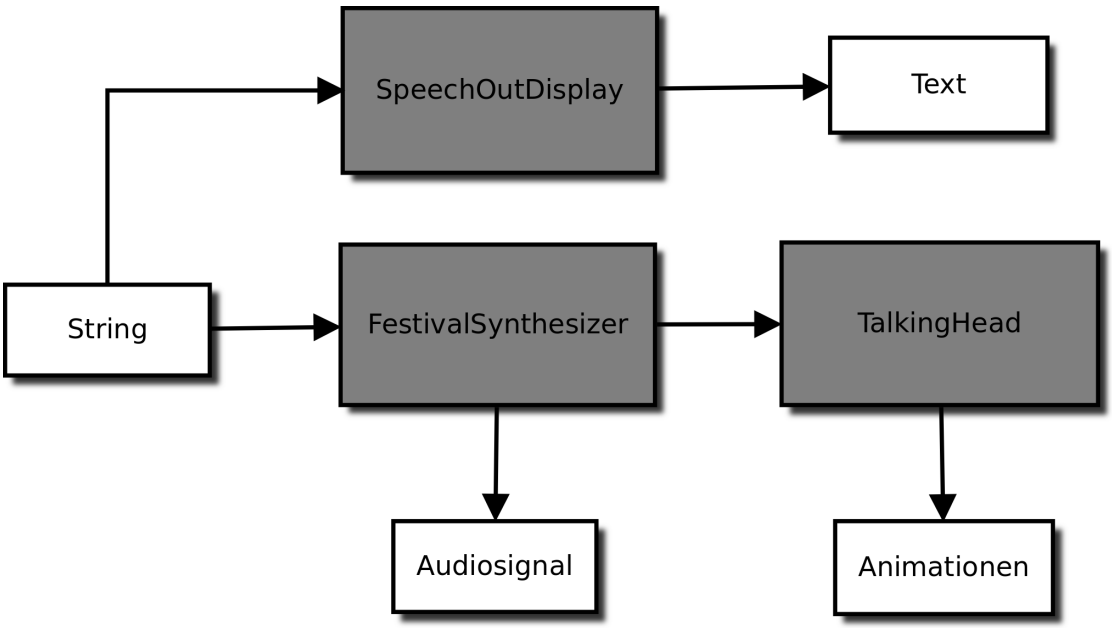


Abbildung 34: Übersicht der Hauptmodule des Talkingheadsystems

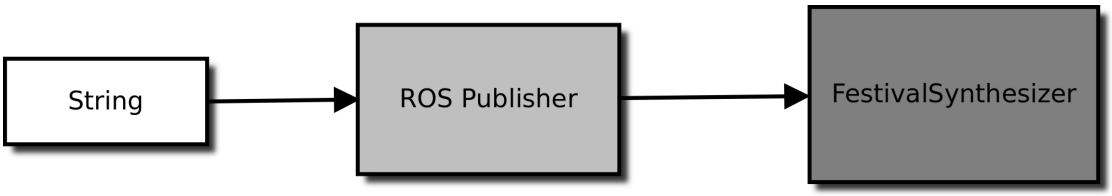
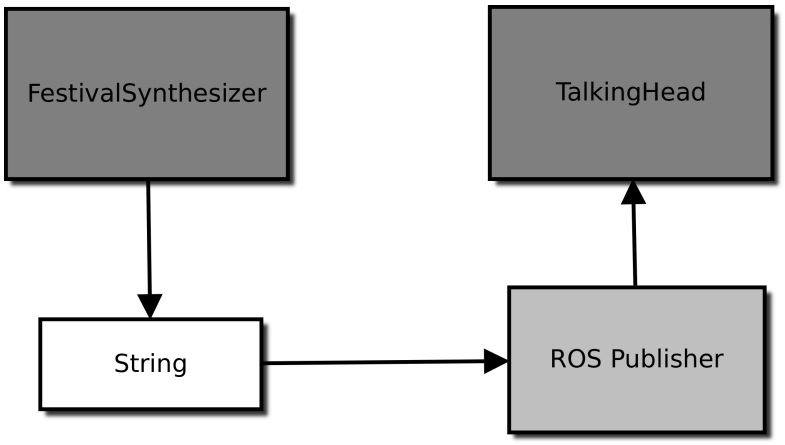


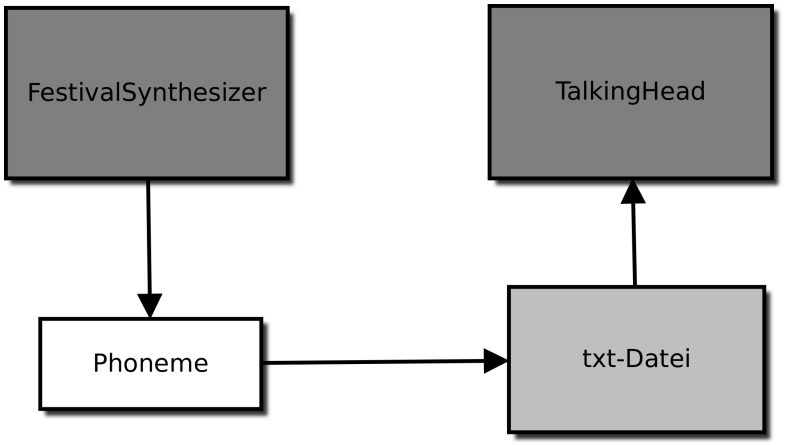
Abbildung 35: Nachrichtenfunktion des *ROS*s - Eingabe eines Strings

Im Schritt zwischen `FestivalSynthesizer` und `TalkingHead` werden die Phoneme und der String weitergereicht (s. Abbildung 36(a)). Hierzu dient die Nachrich-

tenfunktion des *Robot Operating System*[7] (*ROS*). Wie schon der zu Beginn eingegebene String wird dieser über die Nachrichtenfunktion des ROS gesendet (s. Abbildung 35). Aufgrund der Komplexität der Festival API erfolgt die Weitergabe der Phoneme nicht über den Nachrichtendienst des ROS, sondern die erzeugten Phoneme werden in einer txt-Datei gespeichert und dem Talkinghead zugänglich gemacht (s. Abbildung 36(b)). Dem `SpeechOutDisplay`-Modul wird der String ebenfalls über die ROS-Nachrichtenfunktion gesendet, ähnlich wie in Abbildung 35, nur eben zum anderen Modul.



(a) Nachrichtenfunktion des *ROS*s - Von `FestivalSynthesizer` zu `TalkingHead`



(b) Speichern der aus dem String synthetisierten Phoneme in einer txt-Datei

Abbildung 36: Zwischenschritt zwischen `FestivalSynthesizer` und `TalkingHead`

[7] ROS: `http://www.ros.org/` (zuletzt aufgerufen am 29.04.2011)

Audiosignale werden innerhalb von Festival erzeugt. Die Erstellung der Animation gelingt mit der 3D Bibliothek *Ogre3D*[8]. Texte werden anhand eines Qt-Widgets angezeigt. Der Aufbau der drei Hauptmodule wird im Folgenden detaillierter beschrieben.

Aufbau des `TalkingHead`-Moduls

Das `TalkingHead`-Modul ist der Kern des Systems. Hier wird mit Hilfe von Ogre3D das erstellte Gesichtsmodell eingebunden und animiert. Das in *Blender* erstellte Gesichtsmodell wird anhand des *OgreMeshExporters*[9] als Mesh-Datei exportiert und anschließend im `TalkingHead`-Modul verwendet. Der Aufbau des `TalkingHead`-Moduls teilt sich in drei verschiedene Bereiche:

- Erstellen der Szene
- Erstellen der Animationen
- Abspielen der Animationen

Zum Anzeigen des Gesichts wird ein Fenster benötigt und eine Szene, in der es platziert wird. Im `TalkingHead`-Modul wird festgelegt, wie das System konfiguriert ist. Dazu gehören die Auflösung der Anwendung, die Einstellung des Antialiasing und VSync (Vertikale Synchronisation), aber auch die Einstellung des FBO (Framebuffer Object). Anschließend wird mit diesen Einstellungen das Fenster erstellt. Da die TalkingHead-Szene mit einem Textfeld gekoppelt ist, geschieht die Fenstererstellung mit Hilfe von Qt. Sobald das Fenster erstellt wurde, wird die Szene eingerichtet.

Zuerst wird das Mesh geladen und ein Entity erstellt, dem das Mesh zugewiesen wird. Beim Laden des Mesh werden anhand der Anzahl der Submeshes, dieselbe Anzahl an Animationen generiert. Diese Animationen erzeugen zu Beginn keine Bewegungen. Erst nachdem die Keyframes bekannt sind, lässt sich eine Bewegung mit diesen Animationen abspielen. Nachdem das Entity erzeugt wurde, werden die festen Animationen geladen. Zum Schluss kommt noch Licht ins Dunkle, indem die Beleuchtung der Szene eingerichtet wird. Sobald das Fenster und die Szene erstellt sind, ist die Grundlage für die Anzeige des Gesichts geschaffen.

[8] Ogre3D: `http://www.ogre3d.org/` (zuletzt aufgerufen am 29.04.2011)
[9] OgreMeshExporter: `http://www.ogre3d.org/tikiwiki/Blender+Exporter` (zuletzt aufgerufen am 29.04.2011)

Wie schon erwähnt, werden die Animationen während der Erzeugung der Szene erstellt. Die festen Animationen erzeugen auch schon beim Start der Anwendung Bewegungen. Das sind Animationen wie das Blinzeln, die Atmungsanimation und das Kopfrotieren. Die Bewegungen des Mundes und der Augenbrauen werden erst zur Laufzeit generiert. Dabei wartet das Talkingheadsystem auf einen String, der zuerst vom `FestivalSynthesizer`-Modul bearbeitet wird und dann an das `TalkingHead`-Modul weitergereicht wird. Die aus dem Text generierten Phoneme und Zeitstempel, werden durch eine txt-Datei dem `TalkingHead`-Modul zugänglich gemacht. Anhand dieser Informationen und einer im `TalkingHead`-Modul integrierten Phonem-Visem-Tabelle kann das Modul dann die entsprechenden Phoneme den Visemen zuordnen und die Emotionen verarbeiten. Hinsichtlich der Zeitstempel wird für jedes Visem und jede Emotion ein Key Frame erzeugt. Die Key Frames werden dann zu einer Kette zusammengefügt und ergeben eine Animation. Die Animation wird anschließend synchron zum Gesprochenen abgespielt.

Aufbau des `FestivalSynthesizer`-Moduls

Im `FestivalSynthesizer`-Modul wird der zu sprechende Text auf seine phonetischen Bestandteile analysiert und eine Stimme erzeugt. Beim Start der Anwendung wird der Klang der Stimme definiert. Zur Ausgabe des Audiosignals wird das Sound System *PulseAudio*[10] initialisiert. Das Modul wartet nun, bis ein String eintrifft. Ist der String eingetroffen, werden zu dem Text die Phoneme mit Zeitstempeln generiert und in die txt-Datei gespeichert, damit sie für das `TalkingHead`-Modul zur Verfügung stehen, das diese Informationen zur Generierung der Animationen benötigt. Nachdem die Informationen weitergereicht wurden, wird die Sprache synthetisiert und über das PulseAudio Sound System zusammen mit den synchronen Mundbewegungen ausgegeben.

Aufbau des `SpeechOutDisplay`-Moduls

Das `SpeechOutDisplay`-Modul zeigt den Text, der gesprochen wird, in einem Qt-Fenster unterhalb des Talkinghead-Fensters an. Dazu wird beim Start das Layout des Widget bestimmt, sprich Größe des Fensters und Schriftart des Textes und Ähnliches. Das Modul wartet dann solange auf den Text, bis dieser ankommt. Anschließend wird der Text im definierten Fenster angezeigt.

[10] PulseAudio: `http://www.pulseaudio.org/` (zuletzt aufgerufen am 29.04.2011)

5.5 Emoticon-Arithmetik

Das neue Gesicht für Lisa hat die Funktion Emotionen anzuzeigen. Die Emotionen müssen bei der Eingabe des Textes mitgeliefert werden. Dazu dient eine entsprechende Emoticon-Arithmetik. Diese Arithmetik ist noch in Arbeit und nicht vollkommen ausgereift. Aktuell gibt es nur vier Emotionen, die das Gesicht zeigen kann. Dementsprechend gibt es auch nur vier Zeichen in der Arithmetik, die folgend aufgelistet sind:

Emoticon	**Emotion**
:)	lächeln, fröhlich, freundlich
:(	traurig
(:(	böse, wütend, zornig
.	neutral, Standardemotion

Fügt man eine Emoticon an den zu sprechenden Text an, wirkt an dieser Stelle die entsprechende Emotion. Die Emotion ist solange aktiv, bis ein anderes Emoticon im Text auftritt und zur entsprechenden Emotion wechselt. Damit das Gesicht am Ende des Sprechens einen neutralen Gesichtsausdruck annimmt, muss darauf geachtet werden ein „.“ an das Ende des Satzes einzufügen. Emoticons können nicht direkt hintereinander verwendet werden, sondern müssen immer mit einem „space“ hinter dem Wort stehen (z.B. „Hello :)“). Das Emoticon „.” muss wie gewöhnlich direkt am letzten Wort des Satzes angefügt werden.

6 Implementierung

In diesem Kapitel wird die Implementierung des Talkingheadsystems beschrieben. Dazu werden zu Beginn die verwendeten Bibliotheken genannt, anschließend wird auf die im Kapitel 5.4 vorgestellten Module, ihre zugehörigen Klassen und deren Funktionen eingegangen. Zusätzlich werden alle weiteren Klassen aufgezählt, die am Talkingheadsystem beteiligt sind, dazu gehören die Klassen `MainWindow`, die die Verwaltung der Fenster übernimmt und `QtRosNode`, die für die Anbindung an das bestehende Lisasystem verantwortlich ist. Abschließend wird die Integration an das bestehende Lisasystem anhand der Kommunikation über das ROS beschrieben.

6.1 Eingesetzte Bibliotheken

Zur Entwicklung des Talkingheadsystems wurden verschiedene Bibliotheken und Software eingesetzt. Um die Liste komplett zu halten, werden neben den verwendeten Bibliotheken für die Implementierung auch die für die Modellierung verwendeten Programme kurz aufgelistet und beschrieben.

6.1.1 Software - Modellierung

Blender

Die Modellierung des neuen Gesichts für Lisa wurde durch die Open Source 3D Content Creation Suite *Blender*[11] realisiert. Blender ist frei für jedes Betriebssystem nutzbar. Für den Zweck in der vorliegenden Arbeit wurde die Blenderversion 2.49b für Linux verwendet, da zu gegebener Zeit die neuere Version 2.54 nur als Betaversion zur Verfügung stand. Mit Blender lassen sich dreidimensionale Modelle erstellen, die direkt mit Blender gerendert oder für andere Anwendungen exportiert werden können. Blender ist eine mächtige 3D Content Creation Suite, weshalb nicht auf jede Einzelheit des Autorentools eingegangen werden kann. Alle in dieser Arbeit verwendeten Gesichtsmodelle wurden mit Blender erstellt.

OgreMeshExporter

Damit das mit Blender erstellte Gesichtsmodell für Ogre3D zur Verfügung steht, muss das Modell in eine Mesh-Datei exportiert werden. Dazu dient der *OgreMeshExporter*[12].

[11] Blender: `http://www.blender.org/` (zuletzt aufgerufen am 29.04.2011)

[12] OgreMeshExporter: `http://www.ogre3d.org/tikiwiki/Blender+Exporter` (zuletzt aufgerufen am 29.04.2011)

Mit dem OgreMeshExporter ist es möglich, Meshobjekte mit Eckpunktfarben, mehrfache Materialien, UV Texturen und Blend Modes zu exportieren. Zudem können Key Frame Animationen von relativen Meshformen als Pose- oder Morphanimationen exportiert werden. Besonders der zweite Punkt ist wichtig, um in der Implementierung die in der Modellierung erstellten Animationen und Shape Keys ansprechen zu können. Der OgreMeshExporter unterstützt den Export weiterer Elemente, die hier nicht interessant sind. Auch aufgrund des OgreMeshExporters fiel die Wahl der Blenderversion auf Version 2.49b, da der Exporter zu gegebener Zeit nur für diese Version zur Verfügung stand.

6.1.2 Bibliotheken - Implementierung

Ogre3D Library

Ogre3D[13] (Object-Oriented Graphics Rendering Engine) ist eine szenenorientierte, flexible 3D Open Source Engine. Die Klassenbibliothek von Ogre3D abstrahiert alle Details der grundlegenden Systembibliotheken wie Direct3D und OpenGL und bietet eine Schnittstelle basierend auf Weltobjekten und anderen intuitiven Klassen. Ogre3D unterstützt Materialien, Shader, Meshes, Animationen und einiges mehr. Für die vorliegende Arbeit wurde Ogre3D als Basis zur Visualisierung genutzt, da es eine einfache Verwaltung der Meshes und Animationen liefert. Demnach konnten die in Blender erstellten Modelle leicht in das System eingebunden werden, auch die Interpolation der Animationen wird von Ogre3D schon berechnet.

Festival API

Festival[14] ist ein allgemeines freies Framework zur Erstellung von Sprachsynthese-Systemen. Es steht eine C++-API zur Verfügung und ermöglicht mit der *Edinburgh Speech Tools*[15] Bibliothek für Low-Level-Architektur einen Scheme (SIOD) basierten Kommandozeilen Interpreter. Alle APIs basieren auf dem Scheme Interpreter (Scheme ist eine Programmiersprache ähnlich wie Lisp). Aufgrund dessen und weil die C++-API nur einige der wichtigsten Funktion von Festival zur Verfügung stellt, werden in dieser Arbeit neben der C++-API auch verschiedene Scheme Kommandos verwendet. Mit

[13] Ogre3D: `http://www.ogre3d.org/` (zuletzt aufgerufen am 29.04.2011)

[14] Festival: `http://www.cstr.ed.ac.uk/projects/festival/` (zuletzt aufegrufen am 29.04.2011)

[15] Edinburgh Speech Tools: `http://www.cstr.ed.ac.uk/projects/speech_tools/` (zuletzt aufgerufen am 29.04.2011)

Festival werden phonetische Merkmale des Textes analysiert und dem Talkingheadsystem zur Verfügung gestellt, außerdem wird durch Festival die Stimme synthetisiert.

PulseAudio SimpleAPI

PulseAudio[16] ist eine freie, plattformunabhängige Sound-Zwischenanwendung und wird in dieser Arbeit in Kombination mit der Sprachsynthese verwendet, um den erzeugten Audiostream abzuspielen. Die SimpleAPI[17] wurde für Anwendungen mit grundlegender Soundwiedergabe oder -aufnahme erstellt und reicht für die Sprachsynthese mit Festival aus.

Qt SDK

Qt[18] ist eine plattformunabhängige Anwendung und ein UI Framework. Das Qt Software Development Kit beinhaltet die Qt C++ Klassenbibliothek zur Entwicklung mit Qt. Qt ist ein mächtiges Werkzeug zur Erstellung von GUIs. In der vorliegenden Arbeit wird Qt zur Verbindung verschiedener Anzeigen genutzt und erleichtert die Einbindung in das bestehende Lisasystem.

ROS

Das Robot Operating System (*ROS*[19]) stellt eine Bibliothek und Werkzeuge zur Erstellung von Roboteranwendungen zur Verfügung. ROS bietet Hardware Abstraktionen, Treiber, Bibliotheken, Visualisierungen (Ogre3D), Nachrichtenaustausch, Paketmanagement und mehr. ROS wird bereits auf dem bestehenden Lisasystem verwendet. Damit ist eine einfache Integration des Talkingheadsystems in das Lisasystem gegeben. ROS dient im Talkingheadsystem vor allem zur Kommunikation der Module untereinander und mit den Modulen des Lisasystems.

[16] PulseAudio: `http://www.pulseaudio.org/` (zuletzt aufgerufen am 29.04.2011)

[17] PulseAudio SimpleAPI: `http://0pointer.de/lennart/projects/pulseaudio/doxygen/simple.html` (zuletzt aufgerufen am 29.04.2011)

[18] Qt: `http://qt.nokia.com/` (zuletzt aufgerufen am 29.04.2011)

[19] ROS: `http://www.ros.org/` (zuletzt aufgerufen am 29.04.2011)

6.2 Talkingheadsystem

Das Talkingheadsystem setzt sich aus fünf Klassen zusammen. Dazu zählen zum einen die Klasse `FestivalSynthesizer`, die für die Sprachsynthese verantwortlich ist, die Klasse `TalkingHead`, die das Gesicht visualisiert und die Klasse `SpeechOutDisplay` die den gesprochenen Text anzeigt und zum anderen die Klasse `MainWindow`, die `TalkingHead` und `SpeechOutDisplay` integriert sowie `QtRosNode`, die die Kommunikation über das ROS ermöglicht.

Das Talkingheadsystem wurde unter Verwendung der Programmiersprache C++ implementiert. In Abbildung 37 ist eine Übersicht der Klassen als Klassendiagramm gezeigt. Unwichtige Attribute und Funktionen wurden zur besseren Übersicht weggelassen. Das gilt für alle weiteren UML-Klassen, die folgen. Der Würfel in Abbildung 37 repräsentiert das ROS. Alle Klassen kommunizieren über das ROS miteinander, bis auf `MainWindow`. Für die Klasse `MainWindow` steht die Klasse `QtRosNode` zur Verfügung, die die Kommunikation mit dem ROS übernimmt. Im Folgenden werden ausgewählte Methoden der verschiedenen Klassen und deren Funktion beschrieben.

6.2.1 TalkingHead

Die Klasse `TalkingHead` ist der Kern des Talkingheadsystems. Hier wird die Ogre3D Bibliothek eingebunden und verwendet. Dabei kümmert sich die Klasse um die Visualisierung und Animation des Gesichts. Zur Übersicht ist in Abbildung 38 die UML-Klasse von `TalkingHead` zu sehen. Anhand der in der UML-Klasse eingetragenen Attribute und Methoden wird die Funktion der Klasse beschrieben.

`TalkingHead` wird zur Integration in Qt als QtWidget definiert. Die Klasse initialisiert zu Beginn eine Phonem-Visem-Map. Das geschieht durch den Aufruf der Methode `initPhoneMap()`. Anhand dieser Tabelle können später die richtigen Viseme entsprechend der Phoneme zur Animation ausgewählt werden.

Die Methode `createAnimations( std::String meshFile )` lädt die vorhandene Mesh-Datei aufgrund des in `meshFile` übergebenen Namens und generiert entsprechend der Anzahl der Submeshes dieselbe Anzahl an Animationen. Diese Animationen haben eine Länge von 0 Sekunden. Die Animationen werden angelegt, um zur Laufzeit einen AnimationTrack zu erstellen, der den Animationen zugewiesen wird. Der AnimationTrack besteht aus Key Frames, die im Fall der Mundbewegungen ein Paar aus Visem und Zeitstempel bilden.

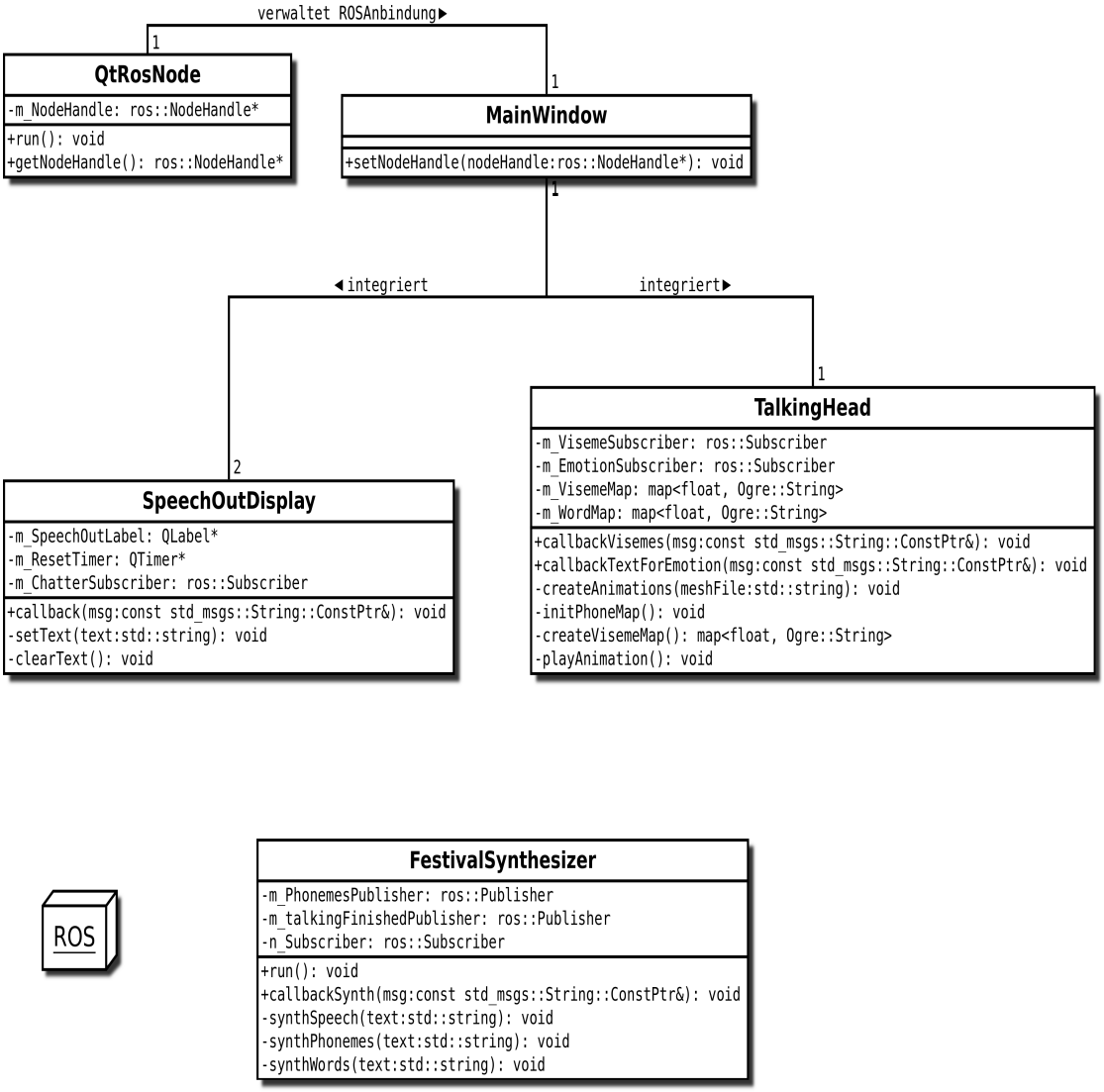


Abbildung 37: Klassendiagramm des Talkingheadsystems

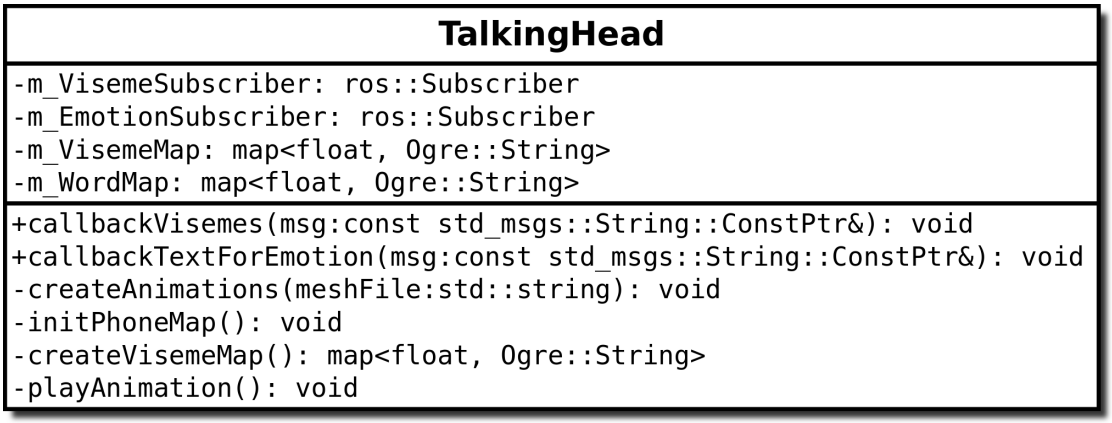


Abbildung 38: UML-Klasse von `TalkingHead`

Visem und Zeitstempel werden anhand der Methode `map<float, Ogre::String> createVisemeMap()` zugewiesen und in `m_VisemeMap` gespeichert. Die zu Beginn initialisierte Phoneme-Visem-Map hilft als Referenz, um für die erhaltenen Phoneme und den zugehörigen Zeitstempeln, eine Map zu erstellen, die die Viseme mit zugehörigem Zeitstemel enthält. Mit Hilfe dieser Map können dann die Key Frames und der AnimationTrack erzeugt werden.

Die Methode `playAnimation()` erzeugt die AnimationTracks. `playAnimation()` nimmt die erstellte Viseme-Zeitstempel-Map und erstellt einen AnimationTrack für die Mundbewegung, der anschließend abgespielt wird. Zusätzlich werden in dieser Methode auch die Emotionen des Mundes als Key Frames dem AnimationTrack beigefügt. Die Interpolation zwischen den Key Frames behandelt Ogre3D mit einer Spline-Interpolation. Das gilt auch für die Bewegung der Augenbrauen, die in dieser Methode in einem AnimationTrack realisiert wird. Die Augenbrauen bewegen sich zur Verstärkung der Emotionen. Deshalb wird für die Augenbrauen nur ein AnimationTrack anhand der Emotionen erstellt.

Die Emotionen werden durch die synthetisierten Wörter und deren Zeitstempel ermittelt. Dazu wird eine Map, ähnlich der Viseme-Zeitstempel-Map, in `m_WordMap` gespeichert. `m_WordMap` wird in den zuletzt genannten Methoden für die Visualisierung der Emotionen ausgewertet.

Die Verbindung des ROS mit dem Talkingheadsystem ist von großer Bedeutung, da die Kommunikation der Klassen untereinander über das ROS geregelt ist. Dazu hat die Klasse `TalkingHead` zwei Subscriber der ROS-API. Anhand dieser Sub-

scriber kann `TalkingHead` bestimmte Topics abonnieren. `TalkingHead` abonniert das Topic „chatter“ mit `m_EmotionSubscriber` und das Topic „festivalPhonemes“ mit `m_VisemeSubscriber`. Durch das Abonnieren wird `TalkingHead` benachrichtigt, wenn etwas zu dem Topic publiziert wurde. Kommt eine Nachricht in Form eines `std_msgs::String::ConstPtr` zum entsprechenden Topic an, wird eine callback-Methode aufgerufen. `TalkingHead` hat zwei Methoden, trifft ein String zum Topic „festivalPhonemes“ ein, wird die Methode `callbackVisemes( const std_msgs::String::ConstPtr& msg )` aufgerufen, die den Animationsprozess der Mund- und Augenbewegungen in Gang setzt. Das heißt, an dieser Stelle wird die Methode `playAnimation()` aufgerufen und die Animation abspielt. Eine Nachricht zum Topic „chatter“ ruft die Methode `callbackTextForEmotion( const std_msgs::String::ConstPtr& msg )` auf, die die Animation der Emotionen behandelt.

6.2.2 FestivalSynthesizer

Die Klasse `FestivalSynthesizer` ist für die Sprachsynthese im Talkingheadsystem verantwortlich. In Abbildung 39 ist eine UML-Klasse des `FestivalSynthesizer` abgebildet. Die aufgezeigten Attribute und Methoden werden nun näher beschrieben.

`FestivalSynthesizer` muss zu Beginn Festival initialisieren, um auf dessen Funktionen Zugriff zu haben. Dazu wird die Methode `initFestival()` aufgerufen, die die Initialisierungsmethoden der Festival-API ausführt und verschiedene Scheme-Kommandos zur Definition der Stimme evaluiert. Außerdem wird in dieser Methode PulseAudio initialisiert.

Die Methoden `synthPhonemes( std::string text )` und `synthWords( std::string text)` erhalten den eingegebenen Text und synthetisieren diesen entsprechend den Phonemen und Wörtern. Dabei werden die Phoneme sowie die Wörter in einer txt-Datei mit zugehörigem Zeitstempel gespeichert. Diese txt-Dateien stehen dann der Klasse `TalkingHead` zur Verfügung.
Die Methode `synthSpeech( std::string text)` synthetisiert den eingegebenen Text zur Sprache und gibt die erzeugte Wave an den PulseAudio-Stream, der das Audio abspielt.

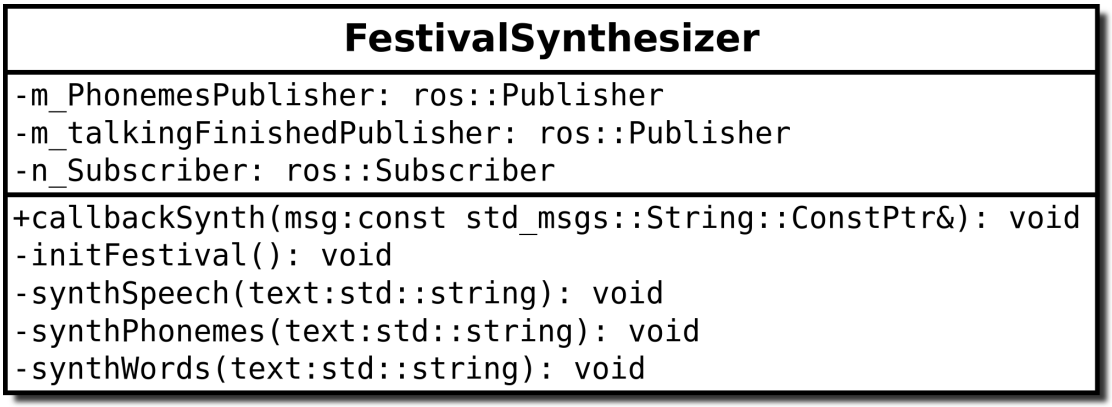


Abbildung 39: UML-Klasse von `FestivalSynthesizer`

`FestivalSynthesizer` hat wie `TalkingHead` einen Subscriber, `m_Subscriber`. Mit diesem Subscriber abonniert `FestivalSynthesizer` das „chatter“ Topic. Dadurch erhält `FestivalSynthesizer` den zu synthetisierenden Text. Wenn der Text zum Topic publiziert wurde, wird die Methode `callbackSynth( const std_msgs::String::ConstPtr& msg )` aufgerufen und signalisiert den Beginn der Synthese. Sobald die Synthese zu den Phonemen und Wörtern abgeschlossen ist, publiziert `FestivalSynthesizer` anhand des Publishers `m_PhonemesPublisher` zum Topic „festivalPhonemes“ und benachrichtigt `TalkingHead`, dass die Phoneme und Wörter mit den Zeitstempeln in den txt-Dateien zur Verfügung stehen. Zusätzlich überprüft `FestivalSynthesizer`, ob der Audiostream beendet ist. Sollte dies der Fall sein, publiziert `FestivalSynthesizer` mit dem Publisher `m_talkingFinishedPublisher`, dass das Gesicht zu Ende gesprochen hat und benachrichtigt somit das Lisasystem, um die Spracherkennung zu regulieren.

6.2.3 SpeechOutDisplay

Die Klasse `SpeechOutDisplay`[20] zeigt neben den Animationen den gesprochenen Text in einem Fenster an. Dazu werden die Methoden und Attribute, die in Abbildung 40 zu sehen sind, näher beschrieben.

`SpeechOutDisplay` ist als QtWidget definiert. Es hat einen QLabel `m_SpeechOutLabel`, das den Text anzeigt. Ein QTimer `m_ResetTimer` bestimmt die Dauer der Anzeige des Textes.

[20] Programmcode aus der ursprünglichen Implementierung von David Gossow

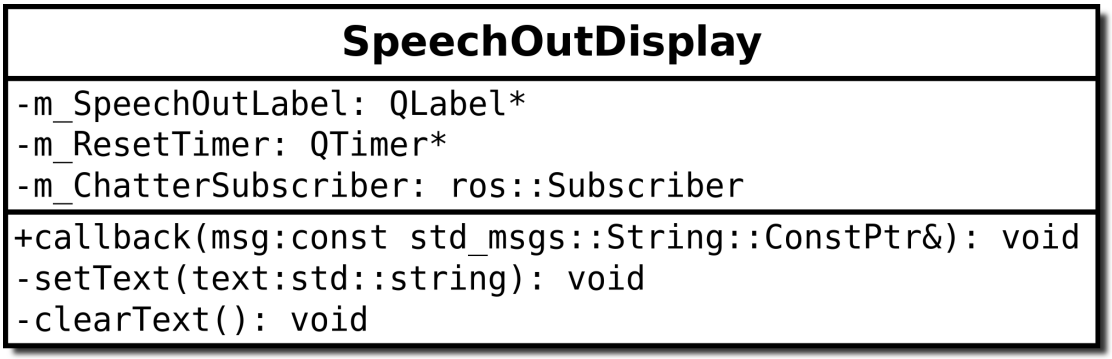


Abbildung 40: UML-Klasse von `SpeechOutDisplay`

Die Methode `setText( std::string text)` setzt den Text des `m_SpeechOutLabel` auf den gesprochenen Text und zeigt ihn somit an. Die Methode `clearText()` wird entsprechend `m_ResetTimer` in einem bestimmten Intervall aufgerufen, um den angezeigten Text wieder zu löschen.

Zur Anbindung an das ROS hat `SpeechOutDisplay` einen Subscriber, `m_ChatterSubscriber`, der das „chatter“ Topic abonniert. Sobald ein Text zum Topic publiziert wurde, ruft `SpeechOutDisplay` die Methode `callback( const std_msgs::String::ConstPtr& msg )` auf, die anschließend die `setText( std::string text )`-Methode aufruft, um den gesprochenen Text anzuzeigen.

6.2.4 MainWindow

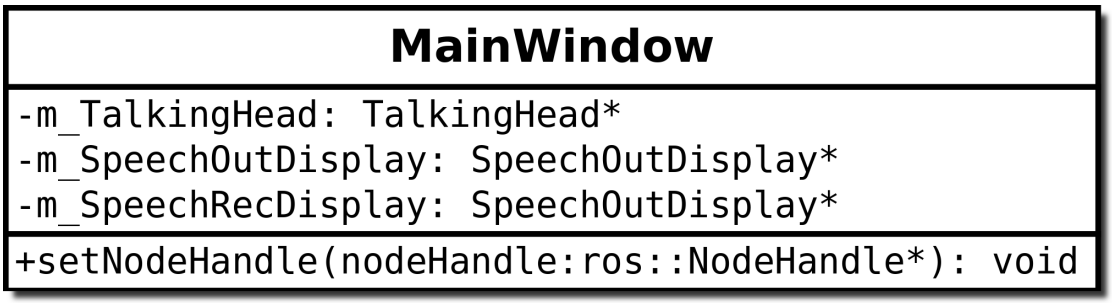


Abbildung 41: UML-Klasse von `MainWindow`

Die Klasse `MainWindow` ist als QWidget definiert und dient der Verbindung einer Instanz von `TalkingHead` und zwei Instanzen von `SpeechOutDisplay`. Die Instanz

des TalkingHead zeigt das Gesicht, wobei die Instanzen des SpeechOutDisplay jeweils einen Text anzeigen. Es gibt das Objekt m_SpeechOutDisplay, das den gesprochenen Text des Talkingheads ausgibt und das Objekt m_SpeechRecDisplay, das den Befehl anzeigt, den man Lisa gegeben hat.

Die Methode setNodeHandle(ros::NodeHandle* nodeHandle) ermöglicht den Instanzen von TalkingHead und SpeechOutDisplay das Abonnieren verschiedener Topics. nodeHandle ist hierbei der ros::NodeHandle von der Klasse QtRosNode.

6.2.5 QtRosNode

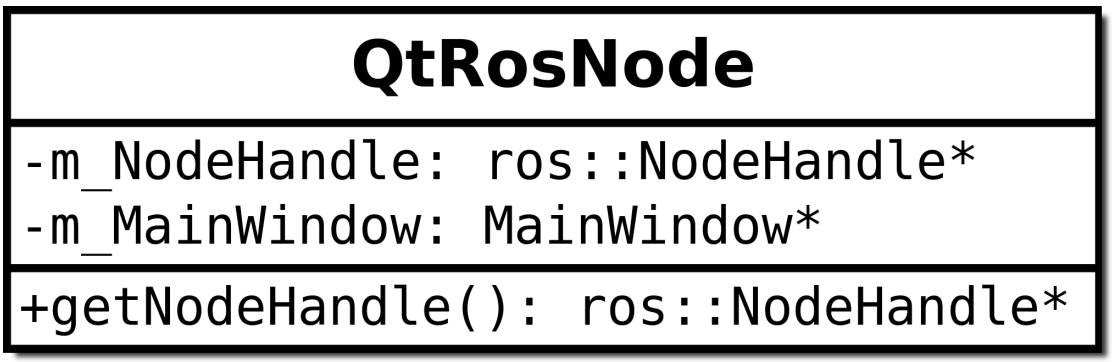


Abbildung 42: UML-Klasse von QtRosNode

Die Klasse QtRosNode verwaltet den ros::NodeHandle zur Kommunikation mit dem ROS. Dazu wird der ros::NodeHandle, m_NodeHandle, angelegt. Dieser ist dann über den Getter ros::NodeHandle getNodeHandle() abrufbar. In der Klasse MainWindow wird er benötigt, um Instanzen von TalkingHead und SpeechOutDisplay das Abonnieren verschiedener Topics zu ermöglichen. Zusätzlich hat die Klasse eine Instanz des MainWindow, m_MainWindow, und aktualisiert in einer Schleife dessen Geometrie.

6.3 Integration in das Lisasystem - Nachrichtenaustausch mit ROS

Wie in der Beschreibung der einzelnen Klassen zu lesen ist, geschieht ein Großteil der Kommunikation über das ROS. Dadurch ist es möglich, das Talkingheadsystem in das Lisasystem zu integrieren. Dieser Vorgang wird in diesem Kapitel noch einmal aufgegriffen und detailliert beschrieben. Abbildung 43 verdeutlicht den Ablauf der Kommunikation des Talkingheadsystems über das ROS in einem Diagramm.

Legende zu Abbildung 43	
grüner Pfeil	Nachricht abonnieren
blauer Pfeil	Nachricht publizieren

In der Mitte der Abbildung 43 ist der roscore zu sehen. Über diesen Kern können alle ROS-Nodes miteinander kommunizieren. Zudem verwaltet roscore die rostopics wie hier die Topics „chatter“, „festivalPhonemes“ und „talkingFinished“. Die aufgezeigten Systeme haben jeweils NodeHandles, mit denen sie zu den verschiedenen Topics publizieren oder von diesen abonnieren können. Das Lisasystem setzt den Nachrichtenaustausch in Gang. Dazu muss vom Lisasystem aus über den NodeHandle eine Nachricht zum Topic „chatter“ publiziert werden. Während dieses Vorgangs schaltet das Lisasystem die Spracherkennung aus. Die NodeHandles des Talkingheadsystems abonnieren das „chatter“ Topic. Somit erhält `FestivalSynthesizer` den Text zum Synthetisieren und `SpeechOutDisplay` den Text zum Anzeigen. Nach der Synthese publiziert `FestivalSynthesizer` zum Topic „festivalPhonemes“, das über den NodeHandle von `QtRosNode` abonniert wird und benachrichtigt `TalkingHead` über die vorhandenen Phoneme, um die Animationen zu erzeugen. `FestivalSynthesizer` prüft, ob die Sprachausgabe beendet ist, und publiziert anschließend zum Topic „talkingFinished“. Das Lisasystem, das dieses Topic abonniert, kann nun die Spracherkennung wieder einschalten. Damit ist der Ablauf einer gesendeten Nachricht beendet und der nächste Text kann über das ROS gesendet werden.

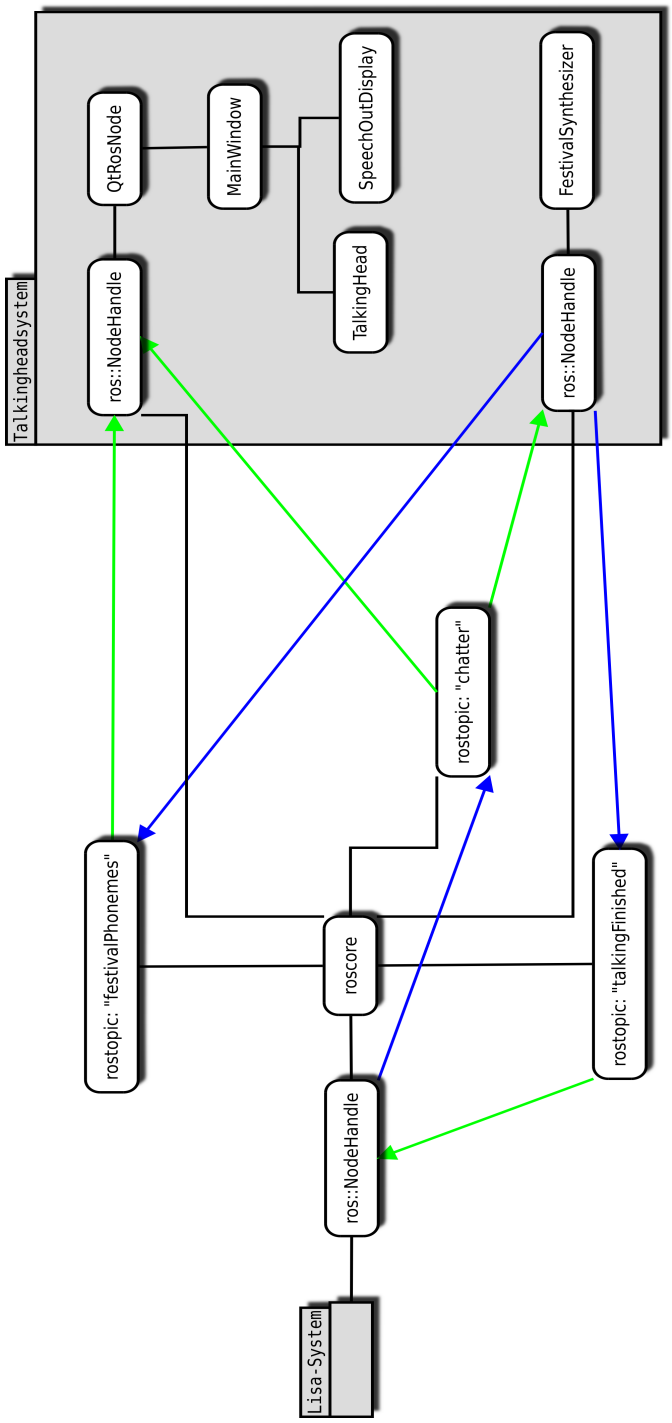


Abbildung 43: Ablauf der Kommunikation des Talkingheadsystems über das ROS

7 Evaluation

Das Talkingheadsystem wurde einer Evaluation unterzogen, um einen Vergleich zum „alten" Gesicht zu erhalten. Einleitend wird der Verlauf der Evaluation beschrieben, worauf anschließend die Vergleichsgegenstände erläutert werden. Zuletzt werden die Ergebnisse vorgestellt, ausgewertet und abschließend die Bewertung resümiert.

7.1 Ablauf

Die Evaluation wurde in einer Halle der Wehrtechnischen Dienststelle (WTD) Koblenz durchgeführt, wo der Haushaltsroboter Lisa zu Hause ist. Dazu kamen die Probanden einzeln in die Halle, um an der Evaluation teilzunehmen. Alle Testpersonen saßen auf einer Couch mit einem davor stehenden Couchtisch. Insgesamt nahmen 13 Testpersonen teil. Die Probanden wurden in zwei Gruppen aufgeteilt. Vor der Evaluation war den Probanden nicht bekannt, dass mehrere Gruppen existieren oder welcher Gruppe sie angehören. Auch über den Vergleich der Gesichter wussten die Probanden vorher nichts. Die erste Gruppe sollte das alte Gesicht bewerten, entsprechend die zweite Gruppe das neue Gesicht. Allen Probanden wurde zu Beginn der Bewertung ein Fragebogen (s. Anhang A) vorgelegt. Zunächst war die erste Seite dieses Fragebogens auszufüllen, um Angaben zur befragten Person aufzunehmen. Nachdem die personenbezogenen Fragen beantwortet waren, begann der zweite Teil der Evaluation. Aus technischen Gründen verlief der zweite Teil beim alten und beim neuen Gesicht unterschiedlich.

Den Probanden der ersten Gruppe wurde das alte Gesicht vorgeführt. Dazu wurde das Gesicht auf dem integrierten Display des Roboters angezeigt. Der Roboter fuhr autonom von einer entfernten Position zu dem Couchtisch und drehte sich zu dem auf der Couch sitzenden Probanden, sodass dieser das Gesicht des Roboters ohne Probleme sehen konnte. Daraufhin sprach der Roboter folgenden Text:

„Hello, my name is Lisa. I hope you are sitting comfortable and enjoy your stay. I come from the University of Koblenz. I am a servicerobot and I serve the Working Group Active Vision. I like my work, but sometimes I am really sick of it. They send me from A to B, from B to C and back to A again. That makes me a little bit angry when they command me. But that's my work. When I am in need I will help. I can for example grab objects and recycle them with my little helper Gigo. Oki doki. Thank you for your patience. It was nice to meet you."

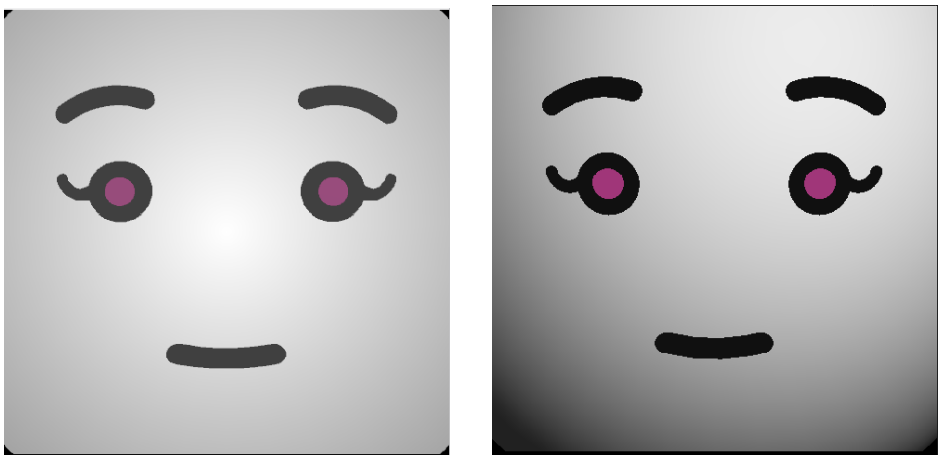
Der Proband sollte das Gesicht des Roboters beim Sprechen beobachten und anschließend den zweiten Teil des Fragebogens beantworten.

In der zweiten Gruppe wurde das neue Gesicht vorgeführt. Da es, wie oben schon erwähnt, zu technischen Problemen während der Evaluation kam, konnte der Teil nicht genauso durchgeführt werden wie beim alten Gesicht. Die autonome Fahrt von der entfernten Position viel leider aus. Deshalb musste der Roboter direkt vor dem Probanden platziert werden. Der Roboter stand dann an demselben Platz wie bei der ersten Gruppe und sprach auch denselben Text mit zusätzlichen Emoticons:

„Hello :), my name is Lisa. I hope you are sitting comfortable :) and enjoy your stay. I come from the University of Koblenz. I am a servicerobot :) and I serve the Working Group Active Vision. I like :) my work, but sometimes. I am really :(sick of it. They :(send me from A to B, from B to C (:(and back to A again, that makes me a little bit (:(angry when they command me. But that's my work. When I am in need :) I will help. I can for example grab objects and recycle them with my little :) helper Gigo. Oki doki. Thank you for your patience :), it was nice to meet you.“

Anschließend mussten auch die Probanden der zweiten Gruppe den zweiten Teil des Fragebogens beantworten.

7.2 Vergleich „altes" und „neues" Gesicht



(a) Altes Gesicht (b) Neues Gesicht

Abbildung 44: Altes Gesicht und Neues Gesicht im direkten Vergleich

Das alte und das neue Gesicht unterscheiden sich in verschiedenen funktionalen Aspekten. Das Aussehen beider Gesichter ist relativ ähnlich, wie in Abbildung 44 zu sehen. Die Evaluation wurde darauf ausgelegt, beide Gesichter zu vergleichen und damit das neue Gesicht zu bewerten. Somit gibt es folgende Nullhypothese und Alternativhypothese:

H_0: *„Das neue Gesicht ist schlechter oder genauso gut wie das alte Gesicht.“*

H_1: *„Das neue Gesicht ist besser als das alte Gesicht.“*

Es gilt also H_0 zu widerlegen. Zu den verschiedenen Aspekten wurden zusätzliche Null- und Alternativhypothesen aufgestellt, die im jeweiligen Unterkapitel zu finden sind. Die Analyse der einzelnen Aspekte versucht detaillierter vorzugehen, um die Nullhypotese H_0 zu widerlegen.

Die funktionalen Aspekte, die verglichen wurden, waren:

- akustisches Verständnis
- Überzeugung der Aktionen
- Veränderung der Stimmungslage
- erwartungsgemäße Natürlichkeit
- wirkendes Interesse und Wohlbefinden

Im nächsten Unterabschnitt werden die ausgewerteten Ergebnisse der Befragung vorgestellt und analysiert.

7.3 Ergebnisse

Es nahmen 13 Personen mit einem Durchschnittsalter von 25 Jahren an der Befragung teil. Insgesamt waren es 8 Männer und 5 Frauen. Der Bildungsabschluss verteilte sich auf die Allgemeine Hochschulreife und einen Abschluss an einer Universität. Sechs der Probanden gaben an Talkingheads schon zu kennen, fünf gaben das Gegenteil an und zwei gaben an mit Talkingheads schon interagiert zu haben. Bis auf zwei Personen benutzen alle einen PC privat und auf der Arbeit, die zwei übrigen Personen nur privat. Es gab wie schon erwähnt zwei Gruppen. Die erste Gruppe, bestehend aus 4 männlichen und 2 weiblichen Probanden, bewerteten das alte Gesicht. Die zweite Gruppe mit 4 Männern und 3 Frauen evaluierten das neue Gesicht. Demnach ergaben sich zwei unabhängige Stichproben.

Die erzielten Ergebnisse werden nun anhand der Aspekte ausgewertet und analysiert, um einen genauen Vergleich der Stichproben abzudecken.

Anhand der Nullhypothese H_0 wird entschieden, ob die Alternativhypothese H_1 akzeptiert werden kann oder nicht. Es gilt also die Nullhypotese zu widerlegen. Dies kann nicht zu 100% gelingen, sondern immer nur mit einer Irrtumswahrscheinlichkeit. Mithilfe des *Mann-Whitney-U Tests* (Bortz und Döring, 2009, S. 150) ist es möglich, die zwei Stichproben hinsichtlich ihrer zentralen Tendenz zu vergleichen. Mit anderen Worten ist der Mann-Whitney-U Test eine Überprüfung der Signifikanz der Übereinstimmung beider Verteilungen. Man unterscheidet einseitige und zweiseitige Tests. Bei einseitigen Tests geht man von einer gerichteten Hypothese (Vergleich ob z.B. ein Verfahren besser ist als das andere) aus. Zweiseitige Tests ermitteln nur, ob ein Unterschied vorliegt (ungerichtete Hypothesen). Liegt die Irrtumswahrscheinlichkeit unter 5%, bei einem zweiseitigen Test, oder unter $2,5\%$ bei einem einseitigen Test, kann man davon ausgehen, dass eine statistische Signifikanz vorliegt. Liegt sie unter 1%, ist die Signifikanz sehr hoch (vgl. Bortz und Döring, 2009).

Dieses Verfahren wird auf die Stichproben angewendet und ermöglicht das eventuelle Verwerfen der H_0. Dazu werden zuerst die verschiedenen Aspekte einzeln analysiert und zum Schluss die komplette Befragung.

Akustisches Verständnis

Als Erstes soll das akustische Verständnis des jeweiligen Gesichtes ausgewertet werden. Die Sprache und Stimme beider Gesichter haben sich nicht unterschieden. Der Hintergedanke hierbei war zu prüfen, ob die Bewegungen des Mundes zum Verstehen des Gesprochenen beitragen oder nicht. Es wird angenommen, dass das neue Gesicht

anhand seiner zum Gesprochenen synchronen Mundbewegungen besser zu verstehen ist als das alte Gesicht. Demzufolge ergeben sich die Unternullhypothese U_1 H_0 und -alternativhypothese U_1 H_1, die zu bewerten sind:

U_1 H_0: „Das neue Gesicht ist akustisch schlechter oder genauso gut zu verstehen wie das alte Gesicht.“

U_1 H_1: „Das neue Gesicht ist akustisch besser zu verstehen als das alte Gesicht.“

Nun gilt es eine Prüfsumme U zu ermitteln, um anhand dieser die Signifikanz zu bestimmen. Dazu wird eine Tabelle aufgestellt und die ermittelten Werte in einer Rangfolge eingetragen. Die Rangsumme R_n und R_m der jeweiligen Zeilen (Gesichter) wird dann zur Berechnung der Prüfsumme U verwendet. Als minimale Prüfsumme wurde $U = 35$ ermittelt. Anhand der Testpersonen für das alte Gesicht mit $n = 6$ und der Testpersonen für das neue Gesicht mit $n = 7$ lässt sich die kritische Prüfsumme $U_* = 6$ auslesen. Bei dieser kritischen Prüfsumme beträgt das Signifikanzniveau α (maximaler Wert der Irrtumswahrscheinlichkeit) für einen einseitigen Test $2,5\%$. Es ist zu sagen, dass die Signifikanz der Stichproben mit $2,5641\%$ über dem Signifikanzniveau liegt. Es besteht also eine $2,5641\%$ige Wahrscheinlichkeit die Nullhypothese fälschlicherweise abzulehnen. Daraus folgt, dass sich die Stichproben signifikant nicht unterscheiden und die Nullhypothese in diesem Fall nicht verworfen werden darf. Man kann also nicht davon ausgehen, dass das neue Gesicht besser zu verstehen ist, als das alte Gesicht.

Anzumerken sei hier, dass die Geräuschkulisse nicht kontrollierbar war und somit eventuell störende Hintergrundgeräusche die Probanden bei der Bewertung beeinflusst haben. Ein Proband schrieb dazu als Anmerkung: *„... leider war die Stimme verzerrt und etwas zu schnell an manchen Stellen. Dadurch, dass es etwas gequetscht klang, fiel die Schnelligkeit der Stimme noch mehr ins Gewicht.“*

Überzeugung der Aktionen

Die Aktionen des Roboters sollten die Probanden überzeugen. Allerdings gab es bei diesem Teil, wie schon erwähnt, ein technisches Problem, wodurch der Roboter mit dem neuen Gesicht nicht alle Aktionen ausführen konnte. Dennoch wurden insgesamt die Aktionen des Roboters unabhängig mit beiden Gesichtern bewertet. Die Aktionen

des neuen Gesichts stehen im Vergleich zu den Aktionen des alten Gesichts. Es erschließt sich also die Annahme, dass die Aktionen des Roboters mit dem neuen Gesicht die Probanden überzeugt haben. Somit ergeben sich die Unternullhypothese U_2 H_0 und -alternativhypothese U_2 H_1:

U_2 H_0: „Die Aktionen des Roboters mit neuem Gesicht sind nicht überzeugend.“

U_2 H_1: „Die Aktionen des Roboters mit neuem Gesicht sind überzeugend.“

Hinsichtlich der Bewertungen ergab sich eine Wahrscheinlichkeit von unter 1% (Signifikanzniveau für den zweiseitigen Test), genauer 0.4%, die Nullhypothese irrtümlich abzulehnen. Also kann man sagen, dass zu $99,6\%$ die Nullhypothese abgelehnt werden kann, da eine hohe statistische Signifikanz vorliegt. Die Aktionen des Roboters mit dem neuen Gesicht sind also überzeugend.

Veränderung der Stimmungslage

Ein weiterer Aspekt, der zu untersuchen ist, ist die Veränderung der Stimmungslage. Das alte Gesicht hat nicht die Funktionalität Emotionen zu zeigen, im Gegensatz zu dem neuen Gesicht. Allerdings strahlt es nach eigenen Beobachtungen eine Grundfreundlichkeit aus, was von den Probanden bestätigt wurde. Der aus der Frage: „Das Gesicht strahlt eine Grundfreundlichkeit aus.” ermittelte Median beider Gesichter ist gleich und bildet mit dem Wert[21] 4 die Antwort „trifft zu” ab. Es wurden weitere Fragen bezüglich der Stimmungslage gestellt, ob diese zu erkennen war und welche Stimmungslagen zu erkennen waren. Dazu wurde angenommen, dass das neue Gesicht verschiedene Emotionen zeigen kann. Somit ergaben sich die Unternullhypothese U_3 H_0 und -alternativhypothese U_3 H_1:

U_3 H_0: „Das neue Gesicht kann keine Emotionen zeigen.“

U_3 H_1: „Das neue Gesicht kann verschiedene Emotionen zeigen.“

Mit einer Wahrscheinlichkeit von unter 5% bezüglich eines zweiseitigen Tests (genauer $3,7\%$) ergibt sich ein geringfügig signifikanter Unterschied zwischen neuem und altem

[21] ein hoher Wert (5) bedeutet „trifft absolut zu”, ein niedriger Wert (1) bedeutet „trifft überhaupt nicht zu“

Gesicht. Die Nullhypothese kann also abgewiesen werden, da mit ungefähr $96,3\%$iger Wahrscheinlichkeit die Alternativhypothese gilt. Es ist also bewiesen, dass das neue Gesicht verschiedene Emotionen zeigen kann. In Abbildung 45 ist die Bewertung der verschiedenen Emotionen in einem Diagramm verdeutlicht. Interessant ist hierbei zu erwähnen, dass die Emotionen „überrascht" und „verzweifelt" während der Evaluation nicht gezeigt wurden und diese, wie aus dem Diagramm zu entnehmen, von den Probanden bei beiden Gesichtern auch nicht erkannt wurden. Das Diagramm verdeutlicht auch, dass die Emotionen beim neuen Gesichts eher erkannt wurden als beim alten Gesicht. Zum Beispiel ergab sich für die Emotion „Traurig" der gemittelte Wert 4 beim neuen Gesicht, hingegen beim alten Gesicht nur der Wert 2.

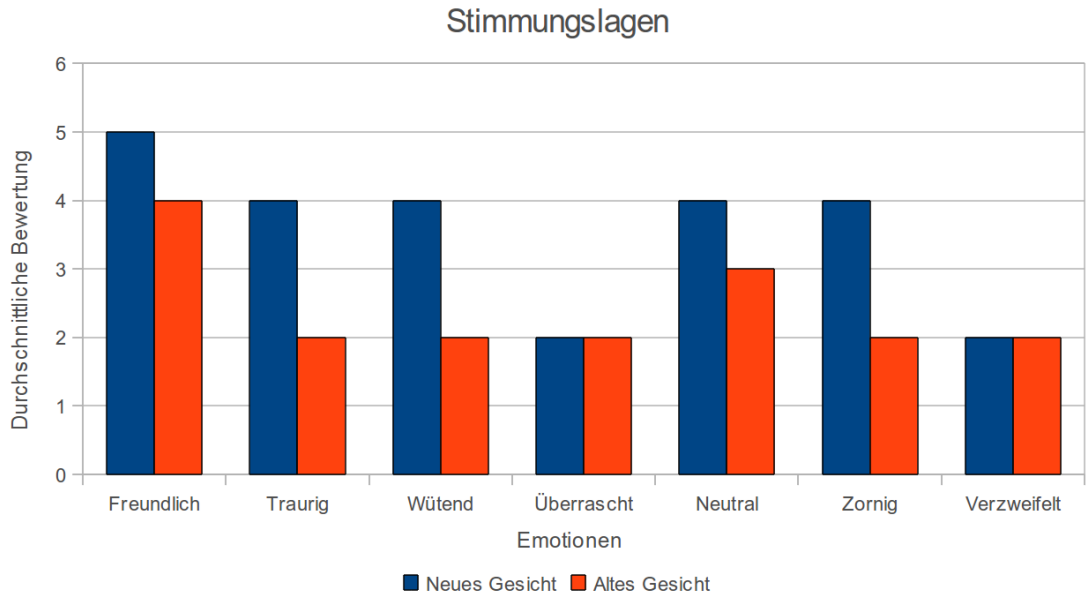


Abbildung 45: Vergleich der verschiedenen Stimmungslagen zwischen neuem und altem Gesicht

Erwartungsgemäße Natürlichkeit

Die Natürlichkeit des Gesichts ist interessant, da ein Cartoongesicht gewählt wurde, und eventuell der Effekt des Uncanny Valley (MacDorman, 2005) (s. Kapitel 5.4.1, S. 42) hätte auftreten können. Beim Vergleich zwischen altem und neuem Gesicht ergab sich die Annahme, dass das neue Gesicht erwartungsgemäß natürlich wirkt und somit die Aufstellung der zwei Unterhypothesen erforderte:

U_4 H_0: *„Das neue Gesicht wirkt unerwartet künstlich.“*

U_4 H_1: *„Das neue Gesicht wirkt erwartungsgemäß natürlich.“*

Das Ergebnis ist überraschend. Die approximierte Wahrscheinlichkeit beträgt 50% für beide Gesichter. Es besteht zu 100% kein signifikanter Unterschied der Natürlichkeit zwischen den Gesichtern. Die Nullhypothese U_4 H_0 kann damit nicht abgelehnt werden. Allerdings ergab sich bei der Mittlung der Bewertung der Natürlichkeit ein Wert von 4 für das neue Gesicht und ein Wert von 3 für das alte Gesicht. Zudem machten Probanden des alten Gesichts folgende Anmerkungen:

„Die Auf- und Abbewegung des Gesichtes wirkte leicht irritierend.“
„Das Gesicht scheint permanent rauf und runter zu schaukeln.“

Das lässt darauf schließen, dass das alte Gesicht nicht so natürlich wie das neue Gesicht wirkt und folglich ein Mehrwert in Bezug auf die Natürlichkeit durch das neue Gesicht erzielt wurde. Man kann dennoch den Effekt des Uncanny Valley nicht ausschließen, da die Nullhypothese nicht widerlegt wurde.

Wirkendes Interesse und Wohlbefinden

Es wurde jeweils eine Frage zum wirkenden Interesse des Gesichtes und dem Wohlbefinden während der Interaktion mit dem Gesicht gestellt. Im Mittel ergaben sich fast dieselben Werte für das neue und das alte Gesicht. Es traf für alle Probanden zu, dass die Interaktion mit dem Gesicht zufriedenstellend ist. Auch beide Gesichter wirken interessant auf die Probanden, das neue Gesicht lag mit einem Wert von 5 (trifft absolut zu) allerdings höher als das alte Gesicht mit einem Wert von 4 (trifft zu). Man kann also sagen, dass die Kommunikation zwischen Mensch und dem Roboter Lisa nicht abschreckend auf den Benutzer wirkt.

Bewertung der gesamten Evaluation

Betrachtet man die gesamte Evaluation, steht noch die Analyse der zu Anfang aufgestellten Hypothese aus. Vor der Analyse werden die zuvor analysierten Aspekte zusammengefasst und mit in die Gesamtbewertung einbezogen.

Zuerst wurde das akustische Verständnis betrachtet. In diesem Fall war das Ergebnis eher unbefriedigend aufgrund der unkontrollierbaren Geräuschkulisse und man konnte nicht sagen, dass das neue Gesicht anhand der synchronen Mundbewegungen zum Gesprochenen besser zu verstehen war als das alte Gesicht. Die Bewertung der Überzeugung der Aktionen fiel deutlich positiver für das neue Gesicht aus. Mit einer Wahrscheinlichkeit von $99,6\%$ wurde die Nullhypothese dieses Bereiches widerlegt und es war möglich zu sagen, dass die Aktionen des Roboters mit dem neuen Gesicht überzeugend sind. Im Bereich der Emotionen konnte das neue Gesicht auch punkten. Der angewandte Test ergab, dass ein geringfügiger signifikanter Unterschied zwischen den Gesichtern besteht und dass die Emotionen des neuen Gesichtes im Gegensatz zum alten Gesicht deutlich erkannt wurden. Die Natürlichkeit des Gesichts konnte nur teilweise bestätigt werden. Hingegen waren wirkendes Interesse und Wohlbefinden eindeutig nachweisbar.

Die Analyse der gesamten Ergebnisse, wie in Abbildung 46 zu sehen, ergab, dass sich die Stichproben geringfügig signifikant voneinander unterscheiden. Für einen hier durchzuführenden einseitigen Test ergab die Irrtumswahrscheinlichkeit $1,31\%$ und liegt somit unter dem Signifikanzniveau von $2,5\%$. Dadurch ist die Nullhypothese H_0 widerlegt. Das heißt entsprechend der Alternativhypothese H_1, dass das neue Gesicht besser ist als das alte. Auch die durchschnittlichen Bewertungen verdeutlichen dies (s. Abbildung 46). Die oben durchgeführte Analyse zeigt, dass das Gesicht nicht in allen Bereichen einen Mehrwert aufweist, aber anhand der abschließenden Analyse der gesamten Bewertung doch ein genereller Mehrwert besteht.

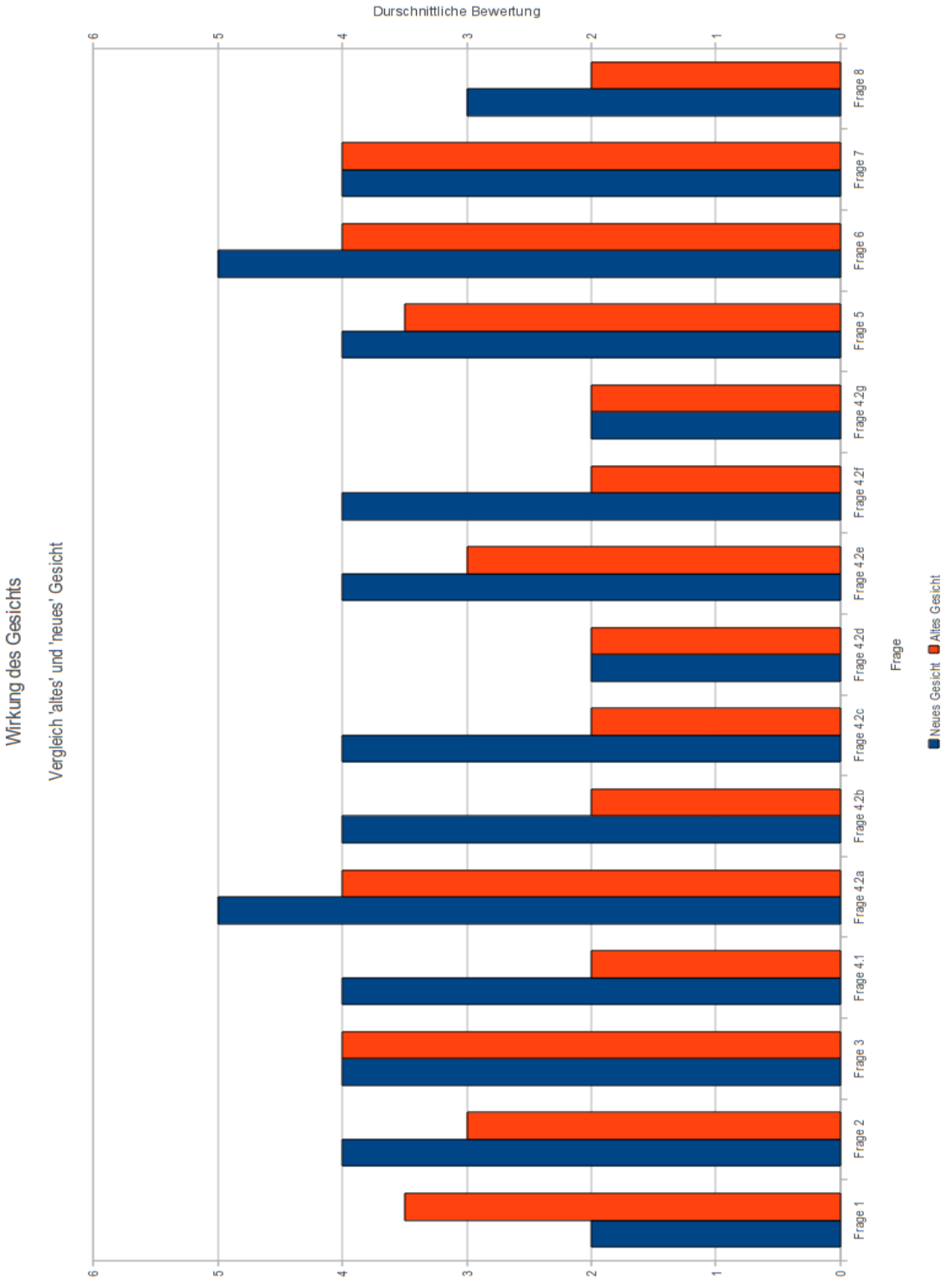


Abbildung 46: Ergebnisse der Evaluation des Vergleichs zwischen altem und neuem Gesicht

Tabelle zu den Fragen in Abbildung 46	
Frage 1	Ich konnte Lisa immer verstehen.
Frage 2	Während der Aktion wirkte das Gesicht überzeugend.
Frage 3	Das Gesicht strahlte eine Grundfreundlichkeit aus.
Frage 4.1	Es war eine Änderung der Stimmungslage zu erkennen.
Frage 4.2	Folgende Stimmungslagen waren zu erkennen.
Frage 4.2a	Freundlich
Frage 4.2b	Traurig
Frage 4.2c	Wütend
Frage 4.2d	Überrascht
Frage 4.2e	Neutral
Frage 4.2f	Zornig
Frage 4.2g	Verzweifelt
Frage 5	Das Gesicht wirkte natürlich.
Frage 6	Das Gesicht wirkte interessant auf mich.
Frage 7	Ich fühlte mich bei der Interaktion mit dem Gesicht wohl.
Frage 8	Die Handlungen des Roboters waren unerwartet.

7.4 Zusammenfassung

Zusammenfassend ist zu sagen, dass mit der Entwicklung des neuen Gesichts eine grundsätzliche Verbesserung für den Roboter entstanden ist. Anhand der durchgeführten Evaluation ist bewiesen, dass das Gesicht zwar akustisch nicht besser zu verstehen ist, aber durchaus seine Vorteile durch die dargestellten Emotionen hat. Auch die Aktionen des Roboters wirken durch das neue Gesicht überzeugender. Zudem hat der Roboter durch das neue Gesicht nicht an Natürlichkeit verloren und weckt beim Benutzer ein allgemeines Interesse. Außerdem fühlen sich die Benutzer während der Interaktion mit dem Roboter wohl.

8 Fazit und Ausblick

8.1 Fazit

Das Ergebnis dieser Arbeit zeigt, dass ein einfaches Cartoongesicht, angebracht an einem Roboter, durch Hinzufügen von Emotionen lebendig und überzeugend wirken kann.

Anhand der durchgeführten Evaluation lässt sich festhalten sagen, dass für den Roboter Lisa ein neues und besseres Gesicht entwickelt wurde. Das neue Gesicht kann jetzt Emotionen zeigen und erlaubt es, aus einem Text Phoneme zu synthetisieren, um daraus Viseme zu erzeugen, die dann eine synchrone Animation des Mundes zum gesprochenen Text ergeben. Zusätzlich wird der gesprochene Text zum besseren Verständnis unter dem Gesicht angezeigt. Positiv ist auch, dass das Gesicht bei Bedarf ausgetauscht werden kann. Mithilfe des erstellten Templates lassen sich verschiedene Gesichter erzeugen, die für den Roboter genutzt werden können. Das Talkingheadsystem wurde mithilfe des ROS erfolgreich in das vorhandene Lisasystem integriert. Somit hat Lisa ein neues, aber dennoch vertrautes Gesicht.

8.2 Ausblick

Das entwickelte Talkingheadsystem ist trotz der positiven Ergebnisse noch nicht vollständig ausgereift. Das System könnte zum Beispiel durch ein anderes Text-To-Speech System eine bessere und deutlicher zu verstehende Stimme bekommen. Zusätzlich kann die Mimik des Gesichts durch prosodische Merkmale verbessert werden. Darüber hinaus muss die Emoticon-Arithmetik überarbeitet werden. Dazu müssen neue Emoticons und die entsprechenden Emotionen hinzugefügt und die Verwendung von aufeinander folgenden Emoticons ermöglicht werden. Hierzu könnte der Anzatz von Albrecht u. a. (2002) betrachtet werden, bei dem die Emoticons in XML übersetzt und anschließend für die Emotionen ausgewertet werden. Eine weitere Verbesserung könnte durch die Überarbeitung des Gesichtmodells erzielt werden.

Ein Ziel wäre, das System so weiterzuentwickeln, dass das System unabhängig vom Lisasystem auch für andere Robotersysteme einsetzbar ist. Diesen Vorteil hat das Talkingheadsystem bis jetzt nicht.

Trotz noch bestehender Verbesserungsmöglichkeiten hat das in dieser Arbeit entwickelte Talkingheadsystem durchaus Potential und eine Grundlage geschaffen, die durch Projekte der Arbeitsgruppe Aktives Sehen weiterentwickelt werden sollte.

Abbildungsverzeichnis

Literatur

[Albrecht u. a. 2002] ALBRECHT, Irene ; HABER, Jörg ; KÄHLER, Kolja ; SCHRÖDER, Marc ; SEIDEL, Hans peter: "May I talk to you? :-)" — Facial Animation from Text. In: *Proceedings of the 10th Pacific Conference on Computer Graphics and Applications (Pacific Graphics 2002)*, 2002, S. 77–86

[Bortz und Döring 2009] BORTZ, Jürgen ; DÖRING, Nicola: *Forschungsmethoden und Evaluation - für Human- und Sozialwissenschaftler*. 4. Springer, 2009

[Fallside und Woods 1985] FALLSIDE, Frank ; WOODS, William A.: *Computer Speech Processing*. Prentice-Hall International, 1985

[Festival 12.04.2011] FESTIVAL: *The Festival Speech Synthesis System*. 12.04.2011. – URL `http://www.cstr.ed.ac.uk/projects/festival/`

[Furui 2001] FURUI, Sadaoki: *Digital Speech Processing, Synthesis, and Recognition - Second Edition, Revised and Expanded*. 2. Ausgabe. Marcel Dekker, 2001 (Signal Processing and Communications Series)

[Gustafson u. a. 1999] GUSTAFSON, Joakim ; LUNDEBERG, Magnus ; LILJENCRANTS, Johan: Experiences from the development of August - a multi-modal spoken dialogue system. In: *ESCA Workshop on Interactive Dialogue in Multi-Modal Systems (IDS-99)*, 1999, S. 61–64

[Jackèl u. a. 2006] JACKÈL, Dietmar ; NEUNREITHER, Stephan ; WAGNER, Friedrich: Gesichtsanimation. In: *Methoden der Computeranimation*. Springer Berlin Heidelberg, 2006 (eXamen.press), S. 117–140. – URL `http://dx.doi.org/10.1007/3-540-33407-6_5`. – ISBN 978-3-540-33407-1

[Karlsson u. a. 2003] KARLSSON, Inger ; FAULKNER, Andrew ; SALVI, Giampiero: SYNFACE – a talking face telephone. In: *The Eurospeech Special Event on "Spoken Language Technology in E-inclusion"*, 2003

[Kerlow 2000] KERLOW, Isaac V.: *The Art of 3-D Computer Animation and Imaging*. John Wiley & Sons, Inc., 2000

[Krach u. a. 2008] KRACH, Sören ; HEGEL, Frank ; WREDE, Britta ; SAGERER, Gerhard ; BINKOFSKI, Ferdinand ; KIRCHER, Tilo: Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. In: *PLoS ONE* 3 (2008), Juli, Nr. 7

[Lundeberg und Beskow 1999] LUNDEBERG, Magnus ; BESKOW, Jonas: DEVELOPING A 3D-AGENT FOR THE AUGUST DIALOGUE SYSTEM. In: *Auditory-Visual Speech Processing (AVSP'99)*, 1999

[Lütkebohle u. a. 2010] LÜTKEBOHLE, Ingo ; HEGEL, Frank ; SCHULZ, Simon ; HACKEL, Matthias ; WREDE, Britta ; WACHSMUTH, Sven ; SAGERER, Gerhard: The Bielefeld Anthropomorphic Robot Head ''Flobi''. In: *2010 IEEE International Conference on Robotics and Automation*. Anchorage, Alaska : IEEE, Mai 2010

[MacDorman 2005] MACDORMAN, Karl F.: Androids as an Experimental Apparatus: Why Is There an Uncanny Valley and Can We Exploit It? In: *Proceedings Of the CogSci 2005 Workshop: Toward Social Mechanisms of Android Science*, 2005

[Niswar u. a. 2009] NISWAR, Arthur ; ONG, Ee P. ; NGUYEN, Hong T. ; HUANG, Zhiyong: Real-time 3D Talking Head from a Synthetic Viseme Dataset. In: *VRCAI '09 Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry*, 2009

[Osipa 2003] OSIPA, Jason: *Stop Staring - Facial Modeling and Animation Done Right ™*. Sybex, 2003

[Parent 2002] PARENT, Rick: *Computer Animation - Algorithm and Techniques*. Academic Press, 2002 (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling). – ISBN 1-55860-579-7

[Parke 1974] PARKE, Frederic I.: *A parametric model for human faces*, University of Utah, Dissertation, 1974

[Parke und Waters 1996 u. 2008] PARKE, Frederic I. ; WATERS, Keith: *Computer Facial Animation*. 1. u. 2. Ausgabe. AK Peters, Ltd, 1996 u. 2008

[Schroeder 1999] SCHROEDER, Manfred R.: *Computer Speech - Recogbition, Compression, Synthesis*. Springer, 1999

[Watkins 2001] WATKINS, Adam: *3D Animation: From Models to Movies*. Charles River Media, 2001. – ISBN 1-58450-023-9

A Fragebogen

Evaluationsfragebogen der Bachelorarbeit:

Entwicklung eines Gesichts für den Haushaltsroboter Lisa

Julian Giesen

Dieser Fragebogen wird im Rahmen einer Bachelorarbeit durchgeführt zum Thema „Entwicklung eines Gesichts für den Haushaltsroboter Lisa". Der Fragebogen gliedert sich in 2 Teile. Zuerst werden Angaben zur befragten Person aufgenommen. Zum Schluss gibt es Fragen zur Präsentation des Roboters Lisa bezogen auf das Gesicht. Die Aufgabe des Befragten besteht nun darin, die Aktionen und Reaktionen des Roboters Lisa zu beobachten und im Anschluss die beiliegenden Fragen zu beantworten.

Dabei ist Folgendes zu beachten:

- Bitte füllen Sie den kompletten Fragebogen aus.
- Bitte lesen Sie jede Frage sorgfältig durch.
- Beantworten Sie bitte jede Frage ehrlich.
- Alle hier gesammelten Daten werden sicher, vertraulich und anonym aufbewahrt.

Teil 1 – Angaben zur Person

Alter: _____ Geschlecht: □ Frau □ Mann

Welches ist Ihr höchster Bildungsabschluss? _____

Kannten Sie Computergesichter (sog. "Talking Heads") schon vorher?
□ Ja □ Nein

Wenn ja, haben Sie mit Computergesichtern schon in irgendeiner Weise interagiert oder sie beobachtet? □ Ja □ Nein

In welchen Bereichen verwenden Sie einen Computer/Laptop?
□ Privat □ Arbeit □ Ich benutze keinen PC □ Andere _____

Teil 2 – Wirkung des Gesichtes

Der Roboter Lisa hat sich Ihnen soeben vorgestellt und Ihnen etwas gebracht. Was haben Sie beobachtet? Beurteilen Sie nur das Gesicht, nicht das Aussehen des ganzen Roboters. Bitte kreuzen Sie Entsprechendes an. Möchten Sie zu einer Frage keine Angabe machen, kreuzen Sie nichts an. 0 oder 1 Kreuz pro Frage!

1. Ich konnte Lisa immer verstehen.

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

---

2. Während der Aktion wirkte das Gesicht überzeugend.

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

---

3. Das Gesicht strahlte eine Grundfreundlichkeit aus.

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

4.1. Es war eine Änderung der Stimmungslage zu erkennen.

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

4.2. Folgende Stimmungslagen waren zu erkennen:

a) Freundlich

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

b) Traurig

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

c) Wütend

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

d) Überrascht

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

e) Neutral

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

f) Zornig

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

g) Verzweifelt

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

---

5. Das Gesicht wirkte natürlich.

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

---

6. Das Gesicht wirkte interessant auf mich.

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

---

7. Ich fühlte mich bei der Interaktion mit dem Gesicht wohl.

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

---

8. Die Handlungen des Roboters waren unerwartet.

trifft absolut zu	trifft zu	neutral	trifft nicht zu	trifft überhaupt nicht zu
□	□	□	□	□

Wenn Ihnen noch zusätzlich etwas bezüglich des Gesichtes aufgefallen ist, können sie dies hier vermerken.

Ende des Fragebogens

Vielen Dank für Ihre Teilnahme

B Template

Template zur Modellierung

Dieses Template dient zur Modellierung eines Talkingheads für das Talkingheadsystem, welches in der Bachelorarbeit „Entwicklung eines Gesichts für den Haushaltsroboter Lisa“ von Julian Giesen entwickelt wurde. Die Modellierung erfolgt mit der Blender Version 2.49b und benötigt zum Export des Modells in das mesh-Dateiformat das OgreMeshExporter Plugin. Es wird bei dieser Vorlage nicht beschrieben, *wie* man ein Gesicht mit Blender modelliert, sondern *was* dabei zu beachten ist, damit das Gesicht und die Animationen im Talkingheadsystem verwendet werden können.

Benötigte Software:

- *Blender 2.49b*[1]
- *OgreMeshExporter aka Blender Exporter*[2]

Vorbereitung:

1. Blender 2.49b ist gestartet
2. das OgreMeshExporter-Plugin wurde installiert
3. Das Menü „3D View“ ist unter dem Reiter Ansicht (View) auf „Front“ gesetzt

Modellierung:
Für das Aussehen des Gesichtes sind alle Freiheiten gesetzt. Es sollten folgende Punkte beachtet werden, damit das modellierte Gesicht reibungslos in dem Talkingheadsystem angezeigt wird.

- *WICHTIG!* Alle Skalierungen, Rotationen und Translationen müssen im „Edit Mode“ vollzogen werden.
  - Ogre erkennt später in der Anwendung nur Koordinatentransformationen des Meshs und *nicht* des Objekts
- *WICHTIG!* Alle Normalen müssen nach außen zeigen.
  - Wenn die Normalen nach innen zeigen, ist in Ogre die Oberfläche nicht zu sehen
- Das Mesh sollte in der Mitte des Koordinatensystems sein. Blickrichtung emtlang der +Y-Achse. Blickrichtung des Talkingheads entlang der -Y-Achse. Up-Vektor ist die +Z-Achse.
- Die insgesamte Größe des Meshs, also die komplette Größe des Talkingheads, sollte auf 5 skaliert sein. (Shortcut „s“ für skalieren, dann „5“ für die Skalierung)
- Es müssen Shape Keys erstellt werden. Folgende Shape-Keys müssen wie folgt benannt werden und müssen den gegebenen Effekt erzeugen:
  - mouth-rest: Mund in Ruheposition
  - mouth-wide: Mund weit / breit
  - mouth-narrow: Mund eng / schmal
  - mouth-open: Mund offen
  - mouth-close: Mund geschlossen
  - mouth-smile: Mund lächelnd
  - mouth-sad: Mund traurig
  - eyebrows-rest: Augenbrauen in Ruheposition
  - eyebrows-up: Augenbrauen hoch
  - eyebrows-down: Augenbrauen mittig unten
  - eyebrows-sad: Augenbrauen außen unten
- Es sollten weitere Shape Keys angelegt werden um Animationen des Atmens, Blinzelns/Zwinkerns und Rotierens erstellen zu können (optional)
  - Später wichtig für den Export
- *WICHTIG!* Materialien direkt bei der Erstellung benennen. Nachträgliches benennen der

1 Download Blender 2.49b: http://www.blender.org/development/release-logs/blender-249/
2 Download OgreMeshExporter: http://www.ogre3d.org/tikiwiki/Blender+Exporter

Materialien kann zu Fehlern führen. (Sicherheitskopie anfertigen)
- Die Augenbrauen und der Mund sollten jeweils unterschiedliche Materialien haben und sich zu den anderen Materialien der Objekte unterscheiden
- Die Materialien sollten nach dem englischen Namen des entsprechenden Organs im Gesicht benannt werden. Zum Beispiel: Material von Mund = Mouth

Export:
- Einstellungen des Plugins (Buttons die ausgewählt sein müssen):
  - Animation Settings of „“: `Pose`
  - Material Settings: `Export Materials` (.material Namen angeben); `Rendering Materials`
  - `Export Meshes` (Pfad zum Speichern auswählen)
  - `Fix Up Axis to Y`
  - `Require Materials`
  - `Skeleton name follow mesh`
  - `Apply Modifiers`
  - `OgreXMLConverter`
  - Unter dem Reiter „Preferences“
    - Location: `Auto`
    - Export options: `Edge Lists`; tangent; 3 component; `Reorganise vertex buffers`; `Optimise animations`
- Es müssen mindestens die drei Animationen mit folgendem Namen unter `Pose` erstellt werden:
  - blink
  - breathe
  - rotate
- Falls der Export fehlschlägt, Blender mit Administratorrechten starten und den Export wiederholen

Sollten bei der Erstellung Probleme auftauchen oder Fragen aufkommen, stehe ich gerne zur Verfügung:

jgiesen@uni-koblenz.de

Weitere Hilfe zu Blender oder dem Blender Exporter findet man hier:

http://www.blender.org/forum/

http://www.ogre3d.org/forums/
http://www.ogre3d.org/tikiwiki/OGRE+Exporters