Paula Dornhofer Paro Costa

Two-Dimensional Expressive Speech Animation

Animação 2D de Fala Expressiva

Campinas

2015

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Paula Dornhofer Paro Costa

Two-Dimensional Expressive Speech Animation

Animação 2D de Fala Expressiva

Thesis presented to the School of Electrical and Computer Engineering of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor, in the area of Computer Engineering.

Tese apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutora em Engenharia Elétrica, na Área de Engenharia de Computação.

Supervisor: Prof. Dr. José Mario De Martino

Este exemplar corresponde à versão final da tese defendida pela aluna Paula Dornhofer Paro Costa, e orientada pelo Prof. Dr. José Mario De Martino

---

Campinas

2015

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Elizangela Aparecida dos Santos Souza - CRB 8/8098

C823t

Costa, Paula Dornhofer Paro, 1978-
Two-dimensional expressive speech animation / Paula Dornhofer Paro Costa. – Campinas, SP : [s.n.], 2015.

Orientador: José Mario De Martino.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Computer animation. 2. Digital image processing. 3. Statistical modeling. 4. Computer graphics. I. De Martino, José Mario,1958-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Animação 2D de fala expressiva
Palavras-chave em inglês:
Animação por computador
Computação - Processamento de imagens
Métodos estatísticos
Avatares
Área de concentração: Engenharia de Computação
Titulação: Doutora em Engenharia Elétrica
Banca examinadora:
José Mario De Martino [Orientador]
Alberto Barbosa Raposo
Soraia Raupp Musse
Léo Pini Magalhães
Hélio Pedrini
Data de defesa: 23-02-2015
Programa de Pós-Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE DOUTORADO

Candidata: Paula Dornhofer Paro Costa

Data da Defesa: 23 de fevereiro de 2015

Título da Tese: "Two-Dimensional Expressive Speech Animation (Animação 2D de Fala Expressiva)"

Prof. Dr. José Mario De Martino (Presidente): _____

Prof. Dr. Alberto Barbosa Raposo: _____

Profa. Dra. Soraia Raupp Musse: _____

Prof. Dr. Léo Pini Magalhães: _____

Prof. Dr. Helio Pedrini: _____

Abstract

The facial animation technology experiences an increasing demand for applications involving virtual assistants, sellers, tutors and newscasters; lifelike game characters, social agents, and tools for scientific experiments in psychology and behavioral sciences. A relevant and challenging aspect of the development of talking heads is the realistic reproduction of the speech articulatory movements combined with the elements of non-verbal communication and the expression of emotions. This work presents an image-based, or 2D, facial animation synthesis methodology that allows the reproduction of a wide range of expressive speech emotional states and also supports the modulation of head movements and the control of face elements, like the blinking of the eyes and the raising of the eyebrows. The synthesis of the animation uses a database of prototype images which are combined to produce animation keyframes. The weights used for combining the prototype images are derived from a statistical active appearance model (AAM), which is built from a set of sample images extracted from an audio-visual corpus of a real face. The generation of the animation keyframes is driven by the timed phonetic transcription of the speech to be animated and the desired emotional state. The keyposes consist of expressive context-dependent visemes that implicitly model the speech coarticulation effects. The transition between adjacent keyposes is performed through a non-linear image morphing algorithm. To evaluate the synthesized animations, a perceptual evaluation based on the recognition of emotions was performed. Among the contributions of the work is also the building of a database of expressive speech video and motion capture data for Brazilian Portuguese.

Keywords: facial animation; talking head; image-based animation; 2D animation; expressive speech animation.

Resumo

O desenvolvimento da tecnologia de animação facial busca atender uma demanda crescente por aplicações envolvendo assistentes, vendedores, tutores e apresentadores de notícias virtuais; personagens realistas de videogames, agentes sociais e ferramentas para experimentos científicos em psicologia e ciências comportamentais. Um aspecto relevante e desafiador no desenvolvimento de cabeças falantes, ou "talking heads", é a reprodução realista dos movimentos articulatórios da fala combinados aos elementos de comunicação não-verbal e de expressão de emoções. Este trabalho apresenta uma metodologia de síntese de animação facial baseada em imagens, ou animação facial 2D, que permite a reprodução de uma ampla gama de estados emocionais de fala expressiva, além de suportar a modulação de movimentos da cabeça e o controle de elementos faciais tais como o piscar de olhos e o arqueamento de sobrancelhas. A síntese da animação utiliza uma base de imagens-protótipo que são processadas para obtenção dos quadros-chave da animação. Os pesos utilizados para a combinação das imagens-protótipo são derivados de um modelo estatístico de aparência e formas, construído a partir de um conjunto de imagens de treinamento extraídas de um corpus audiovisual de uma face real. A síntese das poses-chave é guiada pela transcrição fonética temporizada da fala a ser animada e pela informação do estado emocional almejado. As poses-chave representam visemas dependentes de contexto fonético que implicitamente modelam os efeitos da coarticulação na fala visual. A transição entre poses-chave adjacentes é realizada por um algoritmo de metamorfose não-linear entre imagens. As animações sintetizadas aplicando-se a metodologia proposta foram avaliadas por meio de avaliação perceptual de reconhecimento de emoções. Dentre as contribuições deste trabalho encontra-se a construção de uma base de dados de vídeo e captura de movimento para fala expressiva em português do Brasil.

Palavras-chave: animação facial; cabeça falante; animação baseada em imagens; animação 2D; animação de fala expressiva.

Contents

To Ana Luiza and Laura.

Acknowledgements

I would like to thank CNPq for the financial support to the present work.

It is with immense gratitude that I acknowledge Prof. Dr. José Mario De Martino for introducing me to this fascinating research field that lead me to a maze of exciting computer engineering problems spiced with a touch of human uniqueness. I feel grateful for his thrust, his patience, his ideas, inputs and suggestions that were fundamental to complete the current work.

The audiovisual corpus presented in the current work is the result of the diligent work of numerous professionals. I owe my deepest gratitude to:

- the actors and actresses that participated in the capture sessions: Bruna Duarte, Carolina Holly, Daniel Tonsig, Eduardo Bordinhon, Gabriel Bito, Inês Fabiana, Luiz Nunes, Luiz Terribele, Rogério Kowaski and Sheila Faermann;
- the staff of the RTV Unicamp, with special thanks to: Jorge Luis Calhau, Wanderlei Paré and Luiz Henrique Tadeu;
- the staff of the Center for Information Technology Renato Archer Motion Capture Laboratory, with special thanks to Maria de Fátima de Gouveia, for her valuable friendship and encouragement.

I would like to express my gratitude to professors Hélio Pedrini, Léo Pini Magalhães and Plínio Almeida Barbosa for their review and helpful comments regarding the present work.

I would also like to thank the colleagues of department and the many acts of kindness of Rodrigo Mologni.

I wish to thank my parents and brother for all the help they provided during the last years, for their patience and for being careful grandparents and a playful uncle.

I would like to thank my beloved husband for supporting my choices, for not letting me down and for being an amazing father.

I wish to thank my daughters, Ana Luiza and Laura, for their spirit of cooperation and for their awakening “hugs and kisses of energy”.

1 Introduction

In 1991, Mark Weiser defined the term "ubiquitous computing" and described a scenario where the interaction between humans and computing devices would be less like the desktop interface paradigm (keyboard, mouse, windows and icons on a display) and more like the way humans interact with the physical world and other humans (WEISER, 1991). A scenario where computing technology will "disappear" and users, independently of their age, level of education or physical abilities, would use it without thinking, focusing their primitive goals on communicating, accessing information and performing everyday tasks.

Two decades later, while users may experience ubiquitous access to the information through computer screens installed on tables, walls or street totems and through cheap, mobile, low-power and multi-purpose computers with small-sized displays, there is still a long way to the deployment of natural interfaces capable of making the devices transparent to the users.

Among the initiatives to make the human-computer interfaces more intuitive and efficient, interactive talking heads arise as an alternative to the traditional WIMP (Windows, Icons, Mouse, Pointing Devices) interfaces. Combined with the technologies of natural language processing, artificial intelligence, speech synthesis and recognition, the synthesis of talking heads enable the implementation of interfaces that are based on our natural face-to-face communication mechanisms, which can be considered particularly advantageous to users that are not well familiarized with technology, illiterate children, adults with low literacy level and users with physical disabilities.

Since the pioneering work of Parke (1972), the facial animation synthesis technology has been evolving together with the advances in computer graphics which was favored, among other factors, by the advances in computer hardware and the easier access to sophisticated capturing equipments like high definition digital video cameras, motion capture systems and, more recently, RGB-D sensors.

Presently, computational systems involving talking heads are rapidly emerging as commercial products, assuming the role of virtual assistants for desktop computers, TV equipments or mobile devices; virtual sellers for web pages; newscasters; healthcare personal agents; training tutors; virtual guides and virtual characters of games or movies (MAGNENAT THALMANN; THALMANN, 1995; PANDZIC, 2002; COSATTO *et al.*, 2003). Other particular applications of facial animation include tools for: conducting controlled experiments in psychology and behavioral sciences (COHN, 2010), language learning (DEY *et al.*,

2010), training perception and production of speech for individuals with hearing loss (MASSARO; LIGHT, 2004), and supporting the therapy of individuals with developmental disabilities like autism (BOSSELER; MASSARO, 2003; MENDI; BAYRAK, 2013).

1.1 Challenges of Videorealistic Synthesis of Facial Animation

The problem of synthesizing videorealistic facial animations — i.e. animations that that can be confounded with the video of a real person — is not trivial and presents numerous challenges.

The human mechanisms of expressing and recognizing facial expressions are part of our set of most primitive survival skills. Studies show, for example, that human neonates as young as three days imitate facial expressions and are also capable of discriminating the facial expressions of happiness, sadness and surprise (FIELD *et al.*, 1982). Around seven months of age, small babies are ready to make the cognitive association of a fearful expression to a signal of threat (PELTOLA *et al.*, 2009). In other words, humans are trained since the birth to recognize emotions and to detect even the slightest changes or imperfections in the face.

Additionally, the careless modeling of the static and dynamic aspects of the human face may result in unexpected perceptual phenomena like the McGurk's effect and the "uncanny valley". The McGurk's effect can be observed, for example, when the video recording of the lips articulation of the syllable [ga], being dubbed with the speech audio of the syllable [ba], is recognized by human observers as the syllable [da] (MCGURK; MACDONALD, 1976).

The term "uncanny valley" was coined by Masahiro Mori, a robotics professor at the Tokyo Institute of Technology, in 1970 (MORI *et al.*, 2012). Based on informal observations, Mori argued that the affinity of human observers with robots and toys increases with realism. However, Mori also pointed out that when a robot or a toy look and behave almost, but not exactly, like a real human, a revulsion reaction arises. Figure 1.1 presents the graph hypothesized by Mori of the emotional response of subjects against anthropomorphism of a robot or a toy. More recently, the implications of the "uncanny valley" have also being considered in the field of computer graphics and facial animation (PARKE; WATERS, 1996; TINWELL *et al.*, 2011).

In the pursuit of videorealistic facial animation synthesis, a key aspect is the modeling of the human face and the head. From the literature, it is possible to identify two main strategies: the model-based and the image-based approaches.

In model-based, or 3D, facial animation, the head and the face are typically described

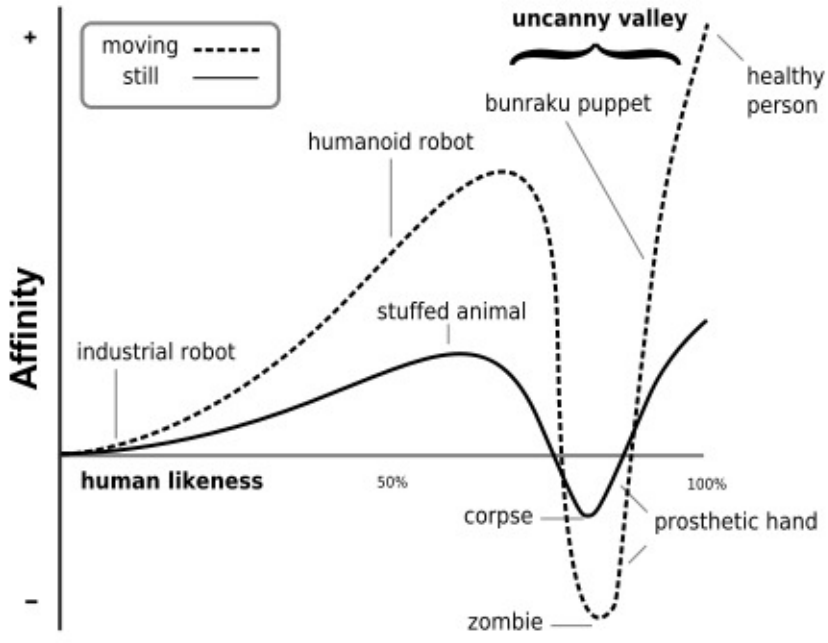
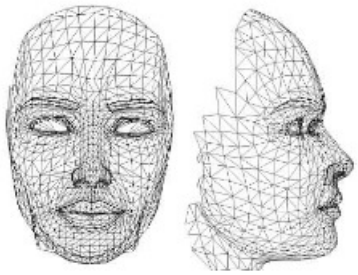


Figure 1.1 – Emotional response of subjects against anthropomorphism of a robot (MORI *et al.*, 2012). Source: Adapted from "Mori Uncanny Valley" by Smurrayinchester - self-made, based on image by Masahiro Mori and Karl MacDorman.

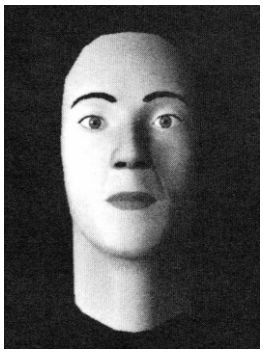
by a tridimensional polygonal mesh to which is mapped the texture information of the many visually distinguishable elements of the face like the skin texture, the eyes, the eyebrows, the lips, the hair, etc. (Figure 1.2(a)). Despite the fact that advanced and modern 3D face modeling techniques are successful in synthesizing high quality images, a more careful observation of 3D talking heads rapidly reveals that the synthesized face is “artificial” (Figures 1.2(c)-1.2(e)). In order to obtain natural looking faces, 3D models require sophisticated animation control strategies, to reproduce, for example, the plastic deformations of the mouth dynamics during speech or the observable changes in the skin texture. Such implementations typically involve special apparatus to capture motion and to scan the skin texture and they are implemented at high computational costs (ALEXANDER *et al.*, 2010).

In image-based facial animation systems, the animation is synthesized through the appropriate processing, sequencing, concatenation and presentation of image samples of a real face (Figura 1.2(b)). In the present work, the image-based approach is also referred to as 2D facial animation, since its output is rendered in two dimensions (MATTHEYSES; VERHELST, 2015). As can be observed in Figures 1.2(f) and 1.2(g), the photographic nature of the manipulated images in the 2D approach results in a human-like modeling of the appearance of the face elements and the skin texture. In this sense, 2D facial animation can



(a)

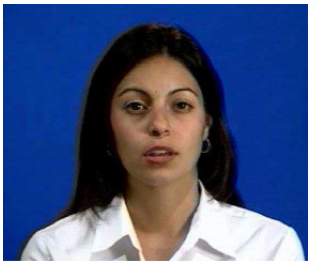
(b)



(c)

(d)

(e)



(f)

(g)

Figure 1.2 – (a) 3D model of a head and face. Source: extracted from (COSTA, 2009). (b) Image-based face model. Source: extracted from (COSTA, 2009). (c) 3D model of the pioneering work of Parke (1972). (d) GRETA 3D head model. Source: extracted (BEVACQUA *et al.*, 2007). (e) A high-quality 3D rendering is presented in (CAO *et al.*, 2005). (f) Example of 2D animation frame. Source: adapted from (EZZAT *et al.*, 2002). (g) 2D animation frame example from (LIU; OSTERMANN, 2011).

be considered inherently photorealistic. On the other hand, 2D facial animation provides limited control of head orientation and movements.

Another important aspect of videorealistic facial animation is the proper representation of the speech articulatory movements in harmony and synchronized with speech. In the last two decades, many facial animation researchers have been proposed different approaches to model the visual speech dynamics including all the complex interactions that happen among the articulatory patterns of different sounds of a language. Such models focus on the animation of neutral speech and they are reviewed in greater detail in Chapter 2.

More recently, following the advances in Computer Graphics and the emergence of Affective Computing paradigm, the facial animation research has also focused on the synthesis of expressive speech, i.e. the speech accompanied by expression of emotions.

1.2 Towards Interactive Conversational Systems

The automated synthesis of videorealistic talking heads capable of expressing emotions remains a challenging problem in computer graphics. Even the most sophisticated models capable of rendering impressive high quality face images are not capable of inspiring user empathy and thrust without the appropriate modeling of the non-verbal communication signals, a task that still requires human intervention.

Moreover, there is no consensus concerning a computational model of emotions and many works adopt a categorical approach which is influenced, for example, by the study of the psychologist Paul Ekman regarding the universality of six basic human facial expressions (anger, happiness, surprise, disgust, fear and sadness). Other researchers adopt a dimensional modeling of emotions, in which the facial expressions are points of a multidimensional continuous space, without no clear specification of how or when those facial expressions are triggered and what would be the valid paths in that space.

The present work points the limitation of such models for the reproduction of everyday dialogues and the implementation of intelligent virtual assistants or embodied conversational agents (ECAs). As an ultimate goal, the present work aims at the implementation of ECAs that would be capable of instilling thrust and empathy in users. Such embodied agents could be imagined involved in situations like: communicating the delay of a service, congratulating a good performance, calling the attention to a wrong procedure, comforting someone in pain, raising the morale of a discouraged individual, etc.

Thus, the present work focuses on the following four main problems:

- the realistic synthesis of the facial appearance;
- the realistic reproduction of the speech articulatory movements;
- the synthesis of facial expressions compatible with everyday conversational interactions;
- the expression of emotions.

1.3 Methodology

The present work describes a 2D expressive speech animation synthesis methodology, that allows the reproduction of a wide range of expressive speech emotional states and also supports the modulation of limited head movements and control of face elements (like the blinking of the eyes or the raising of the eyebrows). The synthesis of the animation depends on the model parameters provided by an expressive speech face model database, which is derived from images extracted from an audiovisual corpus of a real face. The inputs that drive the animation are: the timed phonetic transcription of the speech to be animated and the emotion label corresponding to the desired emotional state.

The problems depicted in the previous section are addressed in the current work through the combination of three main design choices:

- in order to guarantee the photorealism of the synthesized animations, the work adopts an image-based or 2D, face modeling approach;
- the work implements a rule-based visual speech synthesis strategy, based on context-dependent visemes, associated to a non-linear morphing visemes synthesis approach, which is proved to be a robust model for visual speech (DE MARTINO *et al.*, 2006; COSTA; DE MARTINO, 2013; COSTA; DE MARTINO, 2010b; COSTA; DE MARTINO, 2010a);
- the work adopts the Ortony, Clore and Collins (OCC) model of emotions, which embraces a complex but still concise vocabulary of twenty-two emotions, and the specification of the appraisal processes associated to them (ORTONY *et al.*, 1988; COSTA; DE MARTINO, 2014b; COSTA; DE MARTINO, 2014a).

The methodology involves:

- the building of an expressive speech corpus for Brazilian Portuguese (Chapter 3);

- the processing, analysis and the modeling of the image samples extracted from the corpus (Chapters 3 and 4);
- the definition of a synthesis framework (Chapter 5);
- and the validation of the proposed synthesis methodology through a perceptual evaluation of emotion recognition (Chapter 6).

1.4 Contributions

The main contributions of the present work are:

- The proposal and implementation of a 2D expressive speech animation synthesis methodology, which achieves emotion recognition rates that are comparable with the rates obtained by real video.
- The presentation of a rule-based animation synthesis methodology compatible with the Ortony, Clore and Collins (OCC) model of emotions.
- The description of an expressive speech face model that is based on a "shape-independent" representation of expressive visemes (visual phonemes) of the language.
- The creation of an annotated expressive visemes images database that is public for research and private studies.
- The creation of a multimodal expressive speech corpus for Brazilian Portuguese.
- The presentation of an evaluation protocol and the results analyzes of a perceptual evaluation of emotions recognition.

The above contributions were also partially reported and resulted in the following publications:

- COSTA, P. D. P.; DE MARTINO, J.M.. 2D Expressive Speech Talking Head Based on the OCC Model of Emotions. In: Proc. of the 27th Conference on Computer Animation and Social Agents, CASA 2014, Houston, TX, USA, 2014.
- COSTA, P. D. P. ; DE MARTINO, J. M.. Expressive Talking Head for Interactive Conversational Systems. In: 9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2014, Lisbon. Proceedings of the DCVISIGRAPP, 2014. Lisbon: Scitepress, 2014. v. 1. p. 20-24.

- COSTA, P. D. P.; DE MARTINO, J. M.. Assessing the Visual Speech Perception of Sampled-Based Talking Heads. In: Proc. of the 12th International Conference on Auditory-Visual Speech Processing, 2013.

- COSTA, P. D. P. ; DE MARTINO, J. M.; GOUVEIA, M. F.. Captura de Movimento Aplicada à Pesquisa de Agentes Conversacionais Expressivos. In: Proceedings of LatinDisplay 2012 IDRC 2012, São Paulo, 2012.

- COSTA, P. D. P.; DE MARTINO, J. M.; GOUVEIA, M. F.. Towards Interactive Conversational Talking Heads. In: Proceedings of the 3rd Symposium on Facial Analysis and Animation. New York, NY, USA: ACM, 2012.

- COSTA, P. D. P. ; DE MARTINO, J. M. . Context dependent visemes: a new approach to obtain realistic 2D facial animation from a reduced image database. In: 23rd SIBGRAPI Conference on Graphics, Patterns and Images 2010 Workshop of Thesis and Dissertations, 2010, Gramado. 23rd SIBGRAPI Conference on Graphics, Patterns and Images, 2010.

- COSTA, P. D. P.; DE MARTINO, J. M. Compact 2D Facial Animation Based on Context-dependent Visemes. In: Proceedings of the SSPNET 2nd International Symposium on Facial Analysis and Animation. New York, NY, USA: ACM, 2010. (FAA '10).

1.5 Organization

The following chapters describe in details the guidelines, the design choices and the implementation of the proposed synthesis methodology. The text is organized as follows:

- Chapter 2 discusses the state of the art of speech synchronized facial animation systems, focusing on the synthesis of expressive speech. The chapter also provides an introduction to the modeling of emotions.

- Chapter 3 describes the building of an expressive speech corpus for Brazilian Portuguese and the generation of an image database of expressive visemes that is applied to derive the expressive speech face model.

- Chapter 4 details the processing of modeling the shape and the appearance of the image samples and the creation of the expressive speech face model through the projection of the database images on the appearance space obtained.

- Chapter 5 describes how the elements of the methodology are combined to synthesize 2D expressive speech animation.
- Chapter 6 presents the results of an emotions recognition perceptual evaluation performed to validate the proposed synthesis methodology.
- Chapter 7 presents the conclusions of the work and discusses the envisioned future work.

The appendices and the annexes attached to this text are auxiliary material that provides complementary information on specifics aspects of the evaluation processes and details of the expressive speech image database constructed and used in this thesis.

2 Expressive Speech Animation: A Review

This chapter discusses the state of the art of speech synchronized facial animation systems, focusing on the synthesis of expressive speech, i.e. the speech accompanied by the expression of emotions. The first section of this chapter provides a historical perspective on the study of emotions throughout the centuries and places the current work into the context of the emergent interest on affective computing systems. In order to introduce the vocabulary necessary to review the literature of expressive talking heads, Section 2.2 provides an overview of models of emotions from a computer engineering perspective. Section 2.3 presents existent strategies to model the speech gestures in facial animation systems and the literature concerning expressive talking heads is reviewed according to the approach adopted to combine the visual synthesis of speech gestures with the facial emotional signaling. The final section of this chapter discusses how the current work relates with the existing approaches.

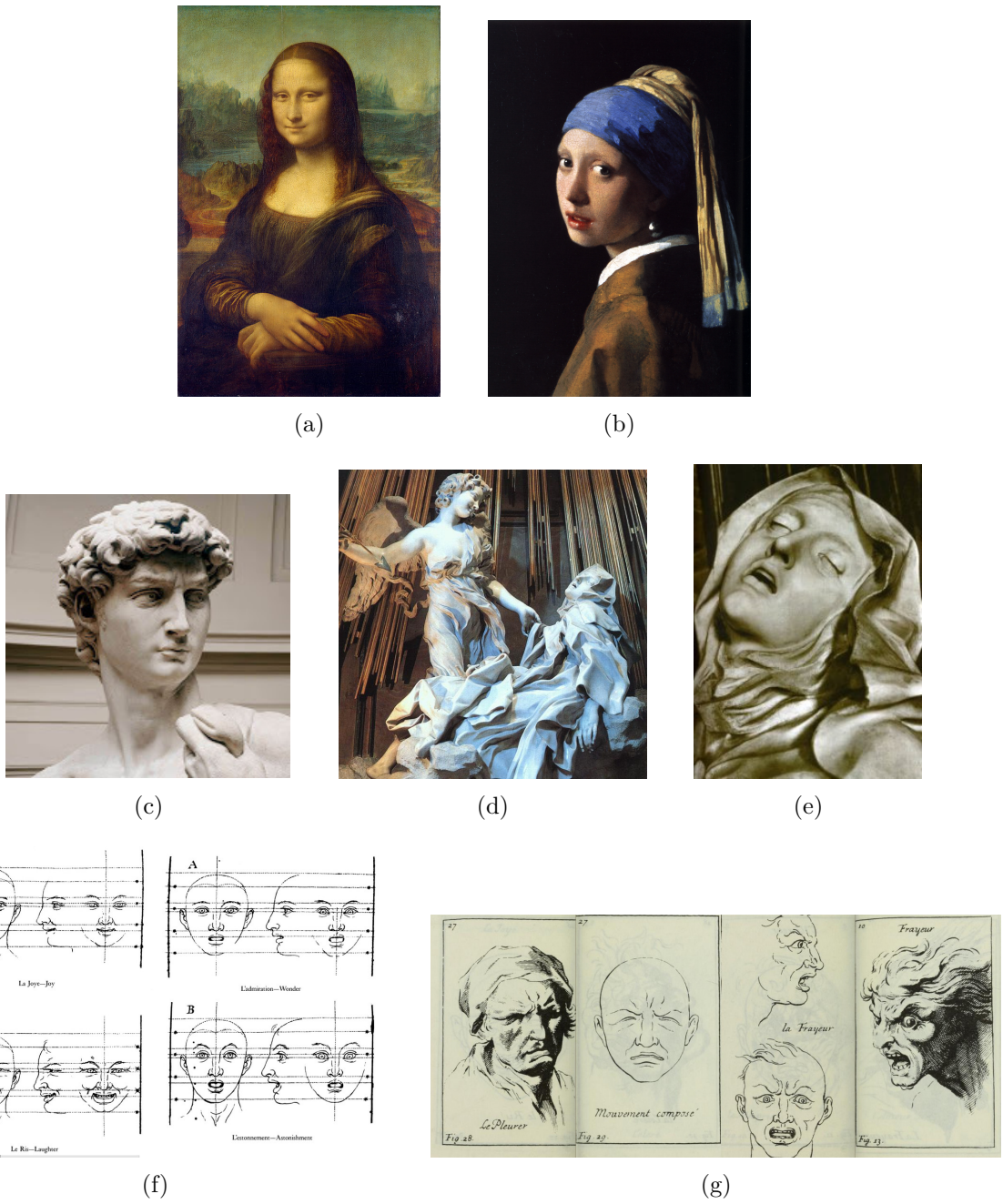
2.1 Theories of Emotions: A Historical Perspective

Emotions, how they arise, how they are expressed and how they can be characterized, have been intriguing artists, philosophers and scholars for many centuries.

Humans represent emotions in paintings and sculptures since prehistory, but the Renaissance period can be highlighted by the work of great artists that dedicated their lives to the realistic representation of the expression of emotions in the human face and body, resulting in masterpieces that are still admired in the present day (see Figure 2.1).

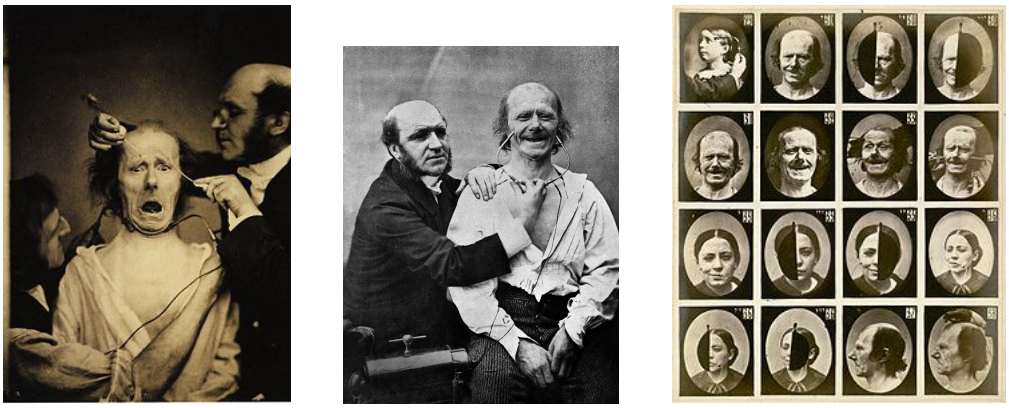
The seventh century witnessed the rise of enlightenment, or the "Age of Reason", and the origins of emotions and their expression started to be explored sistematically. From a dialogue established through correspondence with the Princess Elisabeth of Bohemia regarding moral and ethics questions, the French philosopher, physicist and mathematician René Descartes wrote the *"Passions de l'âme"* (Passions of the soul), published in 1649; a philosophical treatise that explains the mechanisms of "the passions" — better known as *emotions* in the modern period (BOS, 2010). Descartes' work explores how emotions are originated in the Cartesian definition of soul and how they are expressed (see Figure 2.1(f)) (DESCARTES, 1989).

From the beginning of the eighteenth century, the guide to draw facial expressions entitled "Méthode pour apprendre à dessiner les passions" ("A method to learn to design



(a) (b) (c) (d) (e) (f) (g)

Figure 2.1 – (a) *"Mona Lisa"*, c. 1503-1506, Leonardo Da Vinci, Paris; (b) *"Girl with a Pearl Earring"*, c. 1665, Johannes Vermeer, The Hague; (c) Detail of the head of *"David"*, 1501-1504, Michelangelo Buonarroti, Florence; (d) *"Ecstasy of Saint Teresa"*, 1647-1652, Gian Lorenzo Bernini, Rome; (e) Detail of the head of Saint Teresa; (f) Illustrations from the book of Descartes, *"Passions of the Soul"*(1649); (g) Illustrations from the book of Charles Le Brun, *"Méthode pour apprendre à dessiner les passions"*(1702). Sources: (a),(b),(d) Wikimedia Commons; (c) From the website: <http://sfnowak.com/2013/02/08/cant-escape-god/>; (e) Photograph by Fredrick Holland Day; (f) Extracted and adapted from the digital version of the book available at: <http://books.google.com.br/books?id=Ca9aBAAAQBAJ>; (g) Extracted and adapted from the digitized version of the original available at: <https://archive.org/details/methodepourappre00lebr>



(a) (b) (c)

Figure 2.2 – (a),(b) Duchenne applies electrical stimulation to his patient's face; (c)A variety of expressions were studied by Duchenne. Source: Extracted from (BOULOGNE; CUTHBERTSON, 1990).

the passions") (LE BRUN, 1702), by the French painter Charles Le Brun, can be considered the first systematic description to the graphical representation of facial expressions (see Figure 2.1(g)).

In 1824, it was published "Essays on The Anatomy and Philosophy of Expression", from the Scottish surgeon, anatomist and neurologist Sir Charles Bell. In the book Bell mixes his studies in anatomy with his interests in arts and theological philosophy. Bell believed that the animals had more limited expressions than humans because human facial muscles were given by God to express human emotions (BELL, 1824).

Characterizing a controversial field of study — defined by many as a pseudoscience — Physiognomy also played an important role in the history of the theory of emotions (GEISSLER, 1845). In the nineteenth century, the French neurologist Duchenne de Boulogne became interested in the possibility of judging the character and the mental traits of individuals through the study of their facial features. In 1862, Duchenne published *"Mecanisme de la physionomie Humaine"*(The Mechanism of Human Facial Expression), where he presented the results of the studies he conducted applying electrical stimulation to his patients' facial muscles in order to "create" and "mimic" facial expressions (BOULOGNE; CUTHBERTSON, 1990). Duchenne was capable of correlating expressive gestures of the face to individual and groups of muscles, also demonstrating how they may combine to generate a variety of facial expressions(see Figure 2.2).

In 1872, the British naturalist Charles Darwin published *"The Expression of the Emotions in Man and Animals"*, in which he calls out the correlation between different mental states and their corresponding behavioral patterns. Darwin's work is remarkably thorough.

He based his observations on the anatomy knowledge available at that time (including the Bell's and Duchenne's works); the information obtained from interviews with psychiatrists; the analysis of photographs of actors, babies and children; and the results of a questionnaire about emotional expressions that was sent worldwide to different ethnic groups. Darwin concluded that the expression of the "states of the mind", or emotions, in the face and the body are genetically determined and they are compatible with the mechanisms described in the evolution theory. Darwin also explicits his position against Bell, establishing a parallel between the behavior on humans and animals, stating: "the young and the old of widely different races, both with man and animals, express the same state of mind by the same movements" (DARWIN, 1998).

The twentieth century saw the flourishing of many theories of emotions, specially in the fields of neuroscience and psychology (GENDRON; BARRETT, 2009):

- James-Lange and Cannon-Bard theories explored the role of the nervous system in the rise and the expression of emotions (CANNON, 1987);
- Freud explored the correlation between the expression of emotions and their reflection on the health condition of an individual;
- Ekman, Izard, Plutchik introduced the idea of basic categorical models of emotions (EKMAN, 1971), (IZARD, 1977), (PLUTCHIK, 1991);
- Mehrabian e Russell (1974) proposed a multidimensional theory of emotions;
- Arnold (1960), Lazarus (1966), Ortony *et al.* (1988), Scherer (1999) and others directed their attention to the appraisal process that leads to the rise of emotions.

The evolution of the intelligent systems and the possibility of developing more sophisticated human-computer interfaces (HCI), enabled the rise of Affective Computing research field (PICARD, 1995), an interdisciplinary field involving computer engineering, cognitive science and psychology. Affective computing studies are dedicated to the development of systems and devices capable of recognizing, interpreting, processing and simulating human emotions, with the objective of simulating empathy of the machine to the users. Without brushes and paint, computer scientists can be considered the modern version of the artists of Renaissance, trying to reproduce the human emotions with great level of realism.

The following section provides a review of discrete, dimensional and cognitive models of emotions. In this review, the models that describe the physiological mechanisms of the emotions and those that are based on a philosophical representation of the mind are excluded, since their computational implementation is still subject of study.

2.2 Models of Emotions

In the present work, the term *emotion* is used to describe a short-term episode, observed as a change in the functioning of an organism, that is caused by some triggering event; which can be external (such as an action of others, or the acknowledgment of an event) or internal (such as thoughts, memories or sensations) (SCHERER, 1999).

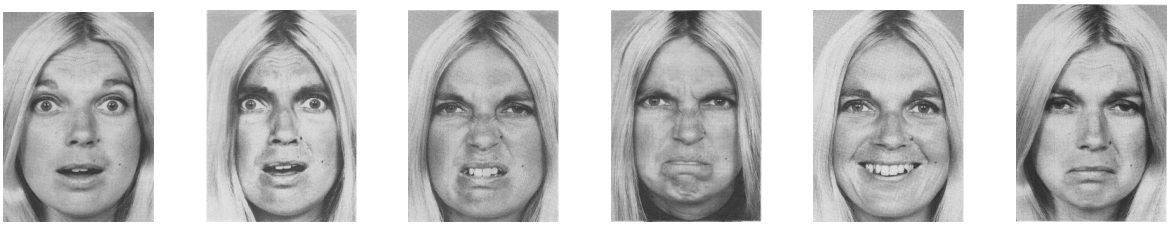
Throughout the last decades, a number of efforts have been made to the development of computational model of emotions, i.e. computer programs capable of simulating some aspects of human emotions. As pointed by Marsella *et al.* (2010), most implementations are ad-hoc adaptations of psychological models of emotions, which are designed to attend the specific needs of applications in different fields like psychological research; human computer interaction; and artificial intelligence (AI).

The following sections present an overview of the models of emotions that have been inspiring the implementation of expressive computational models for talking heads. A model of emotions provides a set of interrelated concepts, definitions and propositions that describes how emotions can be differentiated from each other. In this work, the models are organized into three different categories:

- Categorical models: identify a set of basic emotions that are cross-culturally recognizable and distinguishable by facial expression and biological processes.
- Dimensional models: identify dimensions like valence and intensity that are used to represent a continuum of emotions.
- Appraisal models: differentiate emotions by the appraisal process that triggers them.

2.2.1 Categorical Models of Emotions

Categorical models of emotions are characterized by a finite set of basic emotions, alternatively named primary or fundamental emotions. A common explanation shared by the categorical models is that some emotions seems to exist in all cultures and they can be universally recognized by characteristic facial expressions (ORTONY; TURNER, 1990), (IZARD, 2007). They are typically rooted to biological functions related to the survival needs of the species and they are also present in some higher animals. For example, in an anger expression, a natural reaction to a threatening situation, the eyes are narrowed to increase the focus, the muscles are tense and ready to react, the pupils are dilated and the face is truly “red with anger”, signals of the increased heart rate, blood pressure and the higher levels of adrenaline in the blood.



(a) Surprise (b) Fear (c) Disgust (d) Anger (e) Happiness (f) Sadness

Figure 2.3 – The "big six" facial expressions of Ekman e Friesen (1975). Source: Images extracted from (EKMAN; FRIESEN, 1975).

Descartes states that there are only six simple and basic "passions": wonder, love, hatred, desire, joy, sadness. He also adds (DESCARTES, 1989): "All the others are either composed from some of these six or they are species of them." Instead of choosing a short list of basic emotions, Darwin worked with a list of "states of mind" like: anger, terror, hatred, jealousy, surprise, fear, astonishment, shame, disdain, disgust, guilt, pride, love, joy among others.

Among the psychologists that investigated the existence of basic emotions, it is important to highlight the work of Ekman and Friesen (EKMAN; FRIESEN, 1971), (EKMAN; FRIESEN, 1975). They conducted a cross-cultural study spanning several countries, including isolated communities like Papua New Guinea. The results showed the existence of six facial expressions that are universally recognizable by humans (Figure 2.3): anger, sadness, fear, surprise, happiness and disgust.

Later, Ekman updated his original study and added the "contempt" as a universal facial expression (EKMAN; FRIESEN, 1986). More recently, Ekman proposes a new set of basic emotions: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure and shame (EKMAN, 1999). Nevertheless, the original "big six" emotions of Ekman still represent an important reference for the development of computational systems, animated agents and robots.

It is also important to remind the contribution of Ekman e Friesen (1978) to the popularization of the Facial Action Coding System (FACS), an anatomically based system used to measure or parametrize the movements that are visually discernible in the face. FACS describes the facial movements on the basis of unique action units (AUs) that describe the position and action of muscles groups, the head and the eyes (EKMAN; ROSENBERG, 1997). FACS became popular among psychologists and animators for the analysis and synthesis of human expressions respectively (the application of FACS on the synthesis of facial expressions

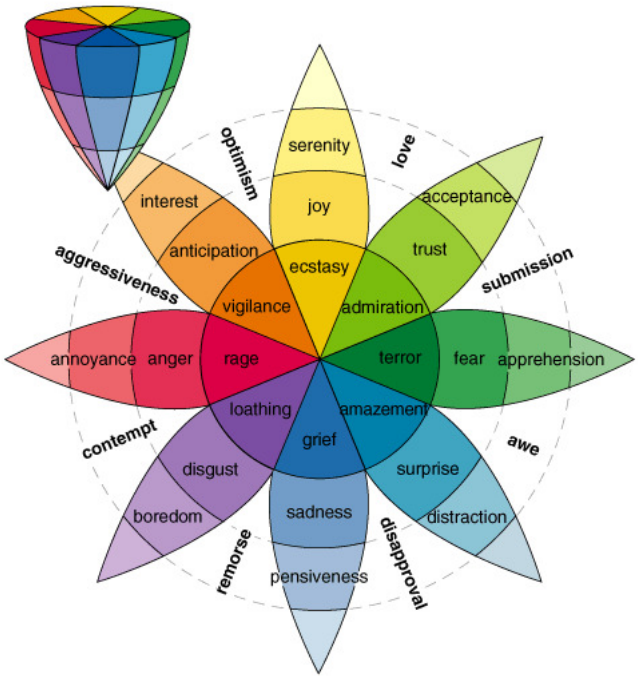


Figure 2.4 – Plutchik's wheel of emotions. Source: Extracted from (PLUTCHIK, 2001).

is exemplified in Section 2.3).

Another example of categorical model of emotions is the wheel of emotions proposed by Plutchik (1984). Plutchik organized the primary emotions in a fashion similar to a color wheel where similar emotions are placed together and opposites are placed 180 degrees apart, like complementary colors (see the inner circle in Figure 2.4). Other emotions are considered combinations of the primary emotions (represented by the petals in the Figure 2.4). The Plutchik's wheel may also be represented as a three-dimensional cone, where the height of the cone is used represent the intensity of the emotions (see the detail at the top-right corner in Figure 2.4).

2.2.2 Dimensional Models of Emotions

An alternative description of emotions is provided by researchers that model the various emotional states as a continuum in a multidimensional space.

The intensity dimension, already mentioned in the Plutchik's model for example, is frequently present in the dimensional models.

In Schlosberg (1954), the intensity dimension is named *activation*. He explains the continuum of activation levels using the analogy of a sleeping man that is suddenly awaken by the alarm clock ring. The sleeping condition refers to the lower levels of activation; the muscles are

relaxed and the cerebral cortex is relatively inactive. The alarm clock ring is a strong stimulus, sending the man to the other extreme of activation level. His whole body becomes alert and prepared to react efficiently to the environment. Besides activation, the three-dimensional model of Schlosberg includes the analysis of facial expressions regarding the Pleasantness-Unpleasantness they express and the indication of the level of Attention-Rejection towards something. The resulting model is illustrated as a three-dimensional cone (SCHLOSBERG, 1954).

Russell e Mehrabian (1977) introduced Pleasure, Arousal and Dominance (PAD model) as three independent emotional axes to describe human emotions (MEHRABIAN, 1996). Arousal corresponds to activation. The pleasure axis was conceived as a continuous range that describes the level of pleasure of a person, from the extreme pain or unhappiness to the extreme happiness. The dominance dimension embraces the social restrictions that someone can face and takes into account the internal interpretation process that someone perform to evaluate if his/her behavior is well accepted or not in a particular situation or environment. The dominance axis range from "control" to "submissiveness".

Whissell (1989) organizes a dictionary of affect in which each emotion term is associated to an angular value derived from the Plutchik's wheel of emotions (Section 2.2.1); and activation and evaluation values, derived from statistical studies.

2.2.3 Appraisal Models of Emotions

The appraisal models of emotions focus on the evaluation, or appraisal, process that leads to the elicitation of emotions.

Arnold (1960) was the pioneer to use the term "appraisal" (SCHERER, 1999), arguing that situations or objects are evaluated according to three dichotomies: good vs. bad (or beneficial vs. harmful), present vs. absent, and the relative difficulty to attain vs. to reject.

According to Lazarus (1966), the appraisal process happens in two stages. The *primary appraisal* evaluates if the emotion elicitator has a positive or negative significance for one's well being. The *secondary appraisal* evaluates the ability to deal with the consequences of the event.

Ortony, Clore and Collins proposed a model that associates cognitive meanings to the logical operations involved in the appraisal process (ORTONY *et al.*, 1988), (SCHERER, 1999). The model comprehends the definition of 22 emotion types organized in a structure that is commonly referred as the OCC model, named after the initials of the its authors. Figure 2.5 presents the structure of the OCC model, which is hierarchically organized in three branches that divides the emotions regarding consequences of events (e.g. happy-for,

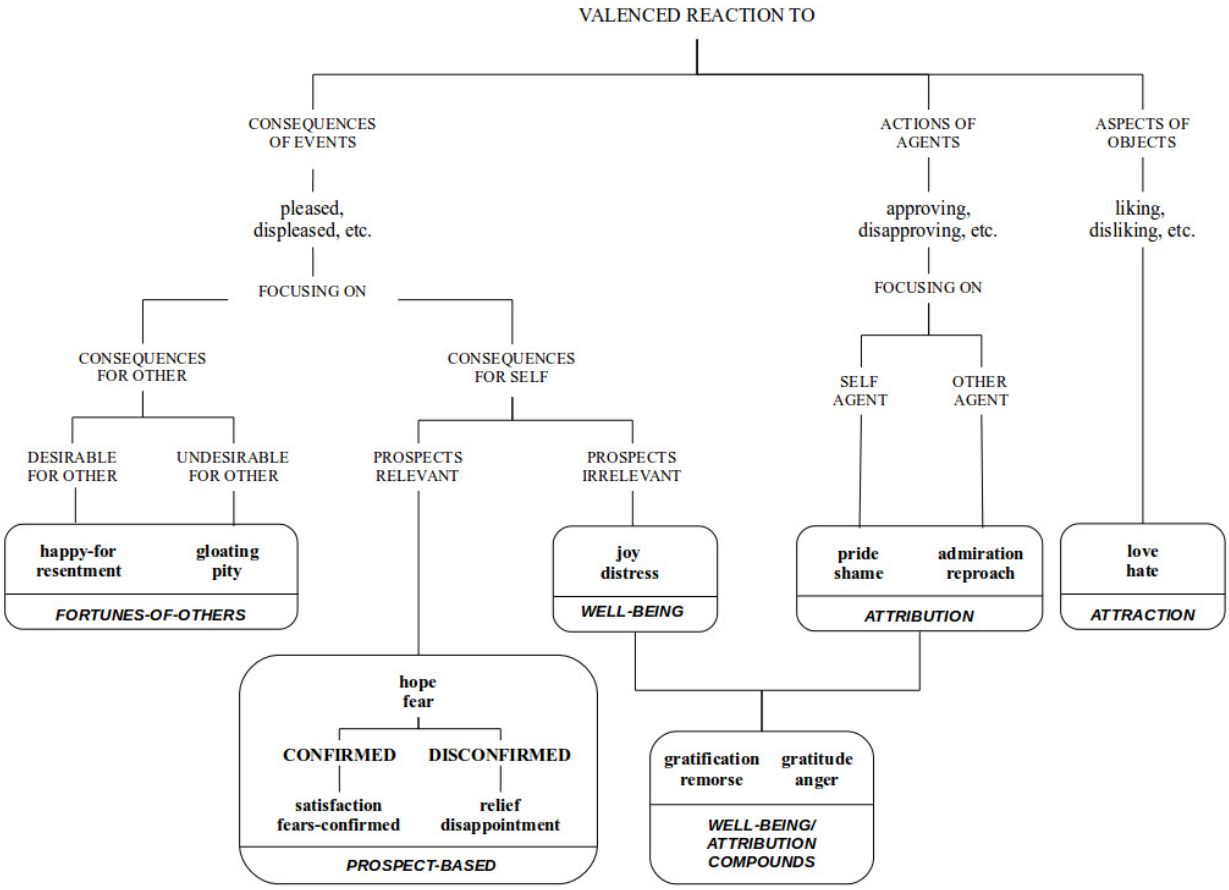


Figure 2.5 – Ortony, Clore and Collins (OCC) model structure of emotions. Adapted from (ORTONY *et al.*, 1988).

resentment, relief), actions of agents (e.g. pride, shame, reproach) and aspects of objects (love and hate). In the figure, the evaluated valence of the emotion is expressed by terms like: pleased vs. displeased, approving vs. disapproving, etc.. The structure describes the appraisal process that occurs from the individual's perspective (e.g. "CONSEQUENCES FOR SELF", "SELF AGENT") but it also includes the interpretation of the others' perspective (e.g. "CONSEQUENCES FOR OTHER", "DESIRABLE FOR OTHER", "OTHER AGENT"). The emotion type "pity", for example, is the cognitive meaning associated to the logical operations that results in being "displeased about an event presumed to be undesirable for someone else".

Scherer (2001) proposed the Sequential Check (SEC) model to distinguish emotions in which four types of information are appraised: the relevance of the event for the individual; the beneficial or harmful implications of the event; the potential of coping with the consequences of the event; and the normative significance with respect to self-concept and to social norms

and values.

2.2.4 Discussion

Previous sections described different categories of models of emotions that provide different answers, at different levels of depth, to the questions: how the emotions arise, how they are expressed and how they can be characterized.

A key aspect of the adoption of models of emotions aiming the synthesis of expressive talking heads, is the translation of a vocabulary of emotions (or points in a dimensional space of emotions) into facial expressions and settings of parameters that control the animation dynamics. The animation of anger, for example, requires facial expressions that typically includes the lowered eyebrows, the thinned lips and the flared nostrils, but also it requires the modulation of speech articulatory movements to properly represent the increased speed and energy typically observed in angry speech. Similarly, variations in intensity, like in angry versus furious speech, may also be modeled.

Categorical models provide a straightforward mapping of emotions into facial expressions. In such models, the basic emotions and their characteristic facial expressions are considered inatte to all human beings, turning them recognizable cross-culturally (ORTONY; TURNER, 1990). In particular, the psychologists Ekman and Friesen provided detailed descriptions of the "big six" facial expressions using the Facial Action Coding System (FACS) (EKMAN, 1971; EKMAN; FRIESEN, 1978). Their contribution still have great influence on several initiatives to model emotions in talking heads (see Table 2.1). However, it is important to observe the limited usefulness of stereotypical facial expressions for the reproduction of everyday dialogue episodes (WEHRLE; KAISER, 2000). To emphasize this, we could imagine the situation where a virtual airport assistant needs to inform a passenger that his/her flight was canceled. In this case, the expression of any of the "big six" plain emotions by a virtual assistant seems to be inappropriate. In order to visually communicate that the system is empathetic with the user's frustration, it becomes necessary the presentation of more complex and subtle facial expressions.

While the dimensional models presented in Section 2.2.2 do not provide a detailed description of how points in a multidimensional space of emotions can be mapped to facial expressions, facial animation researchers have been proposing models to make this association (ZHANG *et al.*, 2007; COURGEON *et al.*, 2008; ARELLANO *et al.*, 2008). Such continuous representation of facial expressions enable smooth trajectories between expressive keyposes and the synthesis of an unlimited number of different facial expressions through the control of "emotional dials". However, like the categorical approach, the dimensional model

of emotions does not provide any mechanism to associate events or situations to locations in the expressive space. Without this, the expressive response of a computational system to a specific situation becomes a designer choice. In other words, the dimensional space of emotions provides us many paths from facial expression A to the B, but it does not tell us, for example, in which conditions we should leave from a neutral state to a happy state. Additionally, dimensional models are abstract representations of emotions. Therefore, visual models have to ensure that the infinite paths between points A an B in the space of emotions, do not lead to unrealistic facial expressions.

Appraisal models of emotions provide a connection between stimuli and emotional reaction. For this reason, they are considered advantageous to the implementation of Embodied Conversational Agents (ECAs) (MARSELLA *et al.*, 2010). Figure 2.6 shows the architecture of an ECA system in which an *Intelligent System* is responsible for processing user inputs and environmental stimuli. In this architecture, the *Intelligent System* implements the appraisal process that leads to specific emotional responses to be expressed by an animated talking head, which is synthesized by the *Facial Animation System*. Some efforts have been made to describe the facial expressions associated to appraisal models emotional states (WEHRLE; KAISER, 2000; KSHIRSAGAR, 2002).

Finally, it is important to observe that the emergent interest on affective computing interfaces has been motivating the discussion around the definition of standard vocabularies of emotions. This is the case, for example, of the W3C (World Wide Web Consortium) EmotionML (Emotion Markup Language) initiative (BURKHARDT; SCHRÖDER, 2014). The vocabularies of emotions provided by the specification include: the "big six" basic emotions of Ekman; the OCC emotion types; the PAD dimensions; the Scherer's appraisals, among others. The language is conceived for application on manual annotation of data and implementation of recognition and synthesis systems.

2.3 Talking Head Synthesis Strategies

Since the pioneering work of Parke (1972), the synthesis of talking heads has become a vigorous branch of research on facial animation.

In addition to the photorealistic representation of the human face, the development of speech synchronized facial animation systems requires attention to the following aspects:

- the accurate lip sync with speech audio;
- the proper reproduction of the speech articulatory movements of a target language, including its coarticulation patterns;

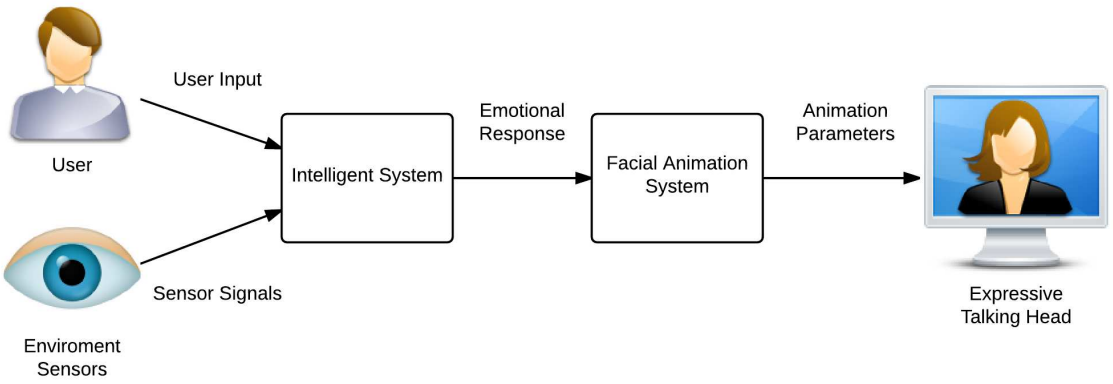


Figure 2.6 – Appraisal models provide a straightforward approach to the implementation of Embodied Conversational Systems (ECA). The *Intelligent System* is responsible for processing user inputs and environmental stimuli. It also implements the appraisal process that leads to a specific emotional response to be expressed by the animated talking head, which is synthesized by the *Facial Animation System*.

- the presentation of non-verbal signaling related, for example, to speech prosody (like the intonational movement of the head), and physiological factors (like the blinking of the eyes);

- and finally, the appropriate modulation of the above elements when the speech is accompanied by the expression of emotions, the so called *expressive speech*.

Throughout the years, many approaches have been proposed to synthesize talking heads. In a first moment, most works devoted their attention to the proper modeling of visual speech, i.e. the modeling of the articulatory movements that are visible on the face and that are related solely to the production of speech (no emotion) (NOH; NEUMANN, 1998; DENG; NOH, 2007; MATTHEYSES; VERHELST, 2015). Many approaches are concerned, at some level, with two key aspects: the modeling of visemes and the modeling of *coarticulation* effects.

Visemes, or visual phonemes, are the visual distinctive facial displays associated to the various phonemes of a language. Many approaches adopt a many-to-one phonemes to visemes mapping, through the identification of homorganic groups of the language — groups of different phones that share the same place of articulation, turning them not distinguishable by visual cues alone. The phones [p], [b] and [m], for example, are articulated with both lips (bilabials) and they are frequently associated to a unique viseme.

Figure 2.7 – Example of anticipatory coarticulation, in which the articulation pattern of phone [s] is modified by the following sounds in the word. Source: Extracted and adapted from (WINN *et al.*, 2013).

However, the identification of the visemes alone is not enough to model the dynamics of visual speech. In order to obtain a realistic modeling of speech articulatory movements, the visual speech model has also to take into consideration the effects of coarticulation. Coarticulation arises when the typical articulation pattern of a speech segment is modified by the interaction with nearby segments. Figure 2.7 shows an example of different mouth configurations to pronunciate the “s” sound in the words “sea” and “sue”. This is an example of the so called anticipatory coarticulation, in which the articulation pattern of the “s” sound anticipates the following sounds in the word.

Over time, the level of videorealism observed on neutral speech talking heads has been significantly improved. Part of this success was reached due to the technological evolution that resulted in increased processing and storage capacity of computers and the availability of technologies like motion capture (mocap), which enabled the acquisition and analysis of large amounts of audiovisual data from real performances. Additionally, as depicted in Section 2.1, the beginning of this century has observed an emerging interest on the development of talking heads that are also capable of expressing emotions, the so called expressive speech talking heads.

The following subsections present a review of the literature of talking heads. The works are classified according to the strategy used to simulate the facial dynamics during speech, based on the categories proposed by Mattheyses e Verhelst (2015): rule-based systems

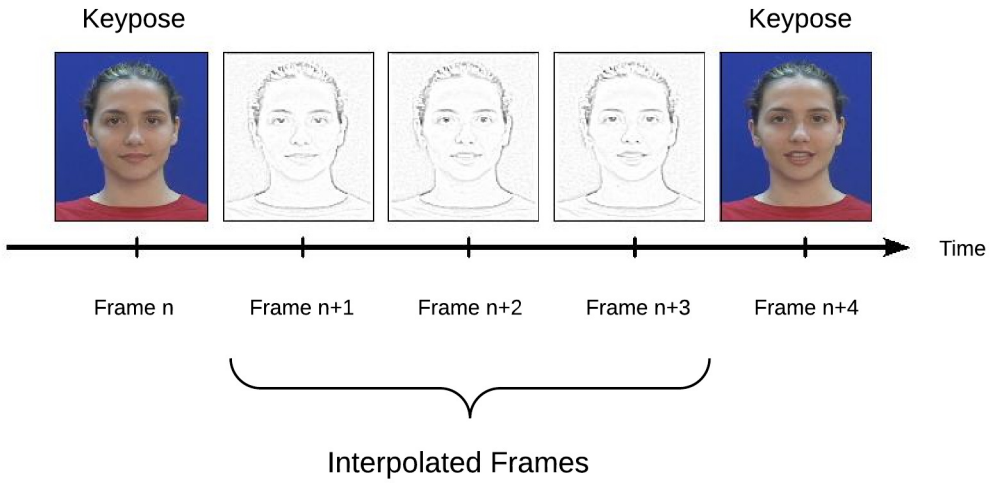


Figure 2.8 – Example of keyframe animation for an image-based talking head. The keyposes are determined following predefined rules. The intermediate frames between two adjacent keyposes are synthesized according to an interpolation function.

(Section 2.3.1), concatenative systems (Section 2.3.2) and systems based on statistical prediction (Section 2.3.3). Moreover, the review is divided into neutral and expressive speech systems, with special focus in the latter. The review of neutral speech systems also includes works that do not comprehend the visual modeling of emotions, but introduce some sort of non-verbal signaling to the facial animation, for instance: the blinking of the eyes; smiling patterns; or some sort of visual prosody modeling that associates the movement of the head to intonational patterns. On the other hand, works that model the facial expressions alone, like the works from Zhou e Lin (2005), Pighin *et al.* (2006), Zhang *et al.* (2008) and Theobald *et al.* (2009), are excluded from the review of expressive speech animation systems. Section 2.3.4 discusses how the different synthesis strategies described throughout this section are related to the resulting level of videorealism of the facial animation.

2.3.1 Rule-based Systems

Rule-based visual speech models apply predefined rules to determine the values of 3D or 2D head models control parameters at particular frames of the animation. Therefore, rule-based systems fit in the keyframe-based animation paradigm, in which the intermediate frames between adjacent keyposes are synthesized according to an interpolation function (Figure 2.8).

2.3.1.1 Neutral Speech Systems

Basic implementations of neutral speech talking heads analyzes the timed phonetic transcription of the speech to be animated and apply conversion rules for a straightforward phonemes to key-visemes mapping (HILL *et al.*, 1988), (SCOTT *et al.*, 1994), (EZZAT; POGGIO, 1998), (GOYAL *et al.*, 2000). Such systems do not model coarticulation and consequently, they present a low level of videorealism.

In order to create more realistic transitions between key-visemes, different strategies have been proposed to model coarticulation. Starting with 3D models, Cohen e Massaro (1993) proposed that the transition between adjacent visemes could be modeled as an overlapping of raising and falling negative exponential functions, called dominance functions. This approach was also adopted by (LE GOFF; BENOIT, 1996), (ALBRECHT *et al.*, 2002b), (KING; PARENT, 2005), (BESKOW; NORDENBERG, 2005). In the work of Beskow (1995) the coarticulation parameters are derived from the morphologic, syntactic and phonetic analysis of an input text. The parameters are then used to modify the key-visemes. In Revéret *et al.* (2000), the interpolation rules were derived from the Linguistics study on coarticulation of Öhman (1967) for VCV (Vowel-Consonant-Vowels) sequences. DE MARTINO *et al.* (2006) combines a non-linear transition function strategy with the identification of visemes that are dependent of phonetic-context. In Costa e DE MARTINO (2013), the context-dependent visemes approach is adapted to a 2D model, where Radial Basis Functions (RBF) are used to warp images on a non-linear morphing transition between adjacent keyposes.

2.3.1.2 Expressive Speech Systems

Most relevant expressive talking heads proposed in the literature can be classified as rule-based systems. First initiatives were concentrated on building a database of emotional facial expressions that are further blended with neutral visemes produced by the visual speech synthesis step.

The pioneering work of Pelachaud *et al.* (1996) not only was one of the first to propose a complete set of rules to model visual speech coarticulation effects, but it also added to the 3D system, mechanisms to model speech prosody and emotional expression. The system reads an input file that contains: the sequence of phonemes and their timing; the emotion label and its intensity; and an intonational structure that controls the periodic blinking of the eyes and some head movements. The system then computes the lip shapes according to the coarticulation rules. The specified affect, or emotion, corresponds to a facial expression which serves as a base with which the lip shapes are blended. The facial expressions modeled by the system are the six facial expressions of Ekman, defined in terms of action units (AUs)

as specified by the Facial Action Coding System (EKMAN; FRIESEN, 1978) (see Section 2.2.1).

A similar visual parametrization of emotions is found in the MPEG-4 Face and Body Animation International Standard (ISO/IEC 14496-2:2004). In MPEG-4, the "big six" emotions of Ekman are defined by archetypal profiles of Face Animation Parameters (FAPs) - a description of a set of facial actions that are closely related to muscle actions - that produces the visual representation of an emotion (PANDZIC; FORCHHEIMER, 2002).

Tsapatsoulis and colleagues proposed a strategy to model nonarchetypal expressions that is based on the angular and activation values of the Whissel's dictionary of affect(Section 2.2.2) (TSAPATSOULIS *et al.*, 2002). The authors proposed a methodology to blend archetypal MPEG-4 FAPs profiles definitions to generate new profiles. Following this paradigm, the emotion term "guilty", for example, has angular measure 102.3 degrees and activation level 4; lying between the archetypal emotion terms "afraid" (70.3 degrees, activation 4.9) and "sad" (108.5 degrees, activation 3.8). In this case, the FAPs profile definitions of "affraid" and "sad" are blended to generate a new profile that is claimed to correspond to the "guilty" facial expression.

In "Byrne", a virtual RoboCup soccer commentator, the emotional expression is driven by an input FACSML script; an adaptation of the FACS annotation system to a Standard Generalized Markup Language (SGML). Byrne is capable of expressing the Ekman's "big six" with control of their intensity (BINSTED; LUKE, 1999).

In the work of Albrecht *et al.* (2002a), prosodic information is derived from the analysis of the input text and it is further transformed into parameters to control the blinking of the eyes, the eye gaze and the head model movement. Emoticons are used as annotation tags of a script language to drive the reproduction of the six facial expressions of Ekman. The same authors, in a more recent work, proposed to model a wider range of emotions adopting a dimensional model of emotions based on activation-evaluation axes (see Section 2.2.2). The values in the space of emotions are mapped to parameters of a physics-based deformation model of a polygonal head mesh (ALBRECHT *et al.*, 2005).

Kshirsagar (2002) proposed a multilayer structure that allows the definition of the personality, the mood and the emotions of a virtual agent (Figure 2.9). In her work, personality is defined as a distinguishing characteristic of individuals, that does not suffer changes over time. The personality is implemented as an initial setup of parameters. The mood is defined as a state of mind, resulting from a cumulative effect of emotions, whose dynamics is also affected by the personality definition. Emotions are states of mind that lasts for short periods of time. The systems uses a MPEG-4 compliant model and receives as input an

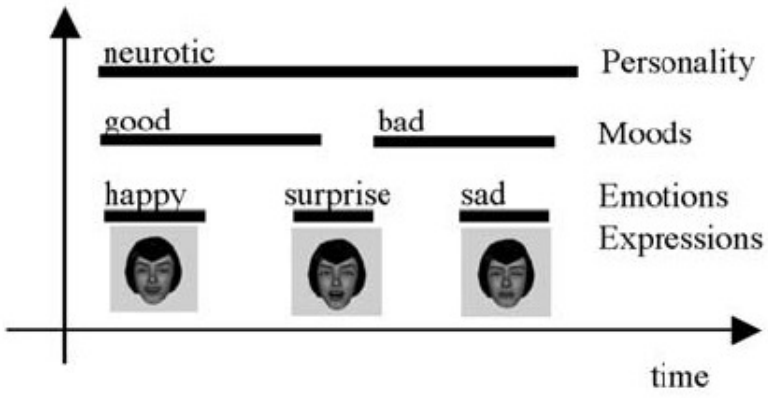


Figure 2.9 – Multilayer structure of personality, mood and emotions. Source: Extracted from (KSHIRSAGAR, 2002).

AIML (Artificial Intelligent Mark-up Language) file, which is an XML based language that include emotional tags. The six basic expressions of Ekman are used to represent emotional states. The synthesis of the emotional facial expressions is affected by the initial personality setup and by the current mood of the agent. The mood is relatively stable over time, but a sequence of emotional states affects the probability of mood transition. The final animation is rendered combining the emotional expressions with the sequence of visemes obtained from the analysis of the speech to be animated.

GRETA is an Embodied Conversational Agent (ECA) platform, based on a 3D head model compliant with MPEG-4 standard, that has been subject of several improvements towards the synthesis of believable agents. In the work of Pelachaud e Bilvi (2003), the implementation is based on a library of facial expressions associated to specific meanings, like "happiness" or "surprise". The animation is synthesized blending the facial expressions to the lips shapes associated to the animation of speech. In the implementation described by Bevacqua *et al.* (2007), the face model is divided into smaller regions of control in such a way that the Ekman's expressions profiles can be manipulated to represent more complex facial expressions that arise, for example, when a felt emotion should not be displayed for some reason and needs to be "masked" with other facial expression (see Figure 2.10). The emotional expression of GRETA is driven by an input APML script (Affective Presentation Markup Language); a XML-language whose tags can be translated into a facial expression, head movements and gaze change.

LUCIA is a MPEG-4 facial animation engine that implements a modified version of Cohen-Massaro coarticulation model to model visual speech (see Section 2.3.1) (COSI *et al.*, 2004; LEONE *et al.*, 2012). The system also receives as input an APML script containing

(a) Anger (b) Superposition of sadness and anger (c) Sadness masked by anger (d) Sadness

Figure 2.10 – Facial expressions of the ECA GRETA. Source: Images extracted from (OCHS *et al.*, 2005).

emotional tags associated to the six facial expressions of Ekman. The authors describe an integrated software environment with aiding tools and methods to develop talking heads.

Queiroz *et al.* (2009) proposed a framework in which the facial animation is driven by an input script that follows the Face Description Language (FDL) format. The system architecture has three main modules: a *lip synchronization module*, which implements a phonemes to viseme mapping; a *facial expressions module*, which produces facial expressions from a predefined database of emotions; and an *eye behavior module* that generates the eye animation. The system implements a categorical model of emotions, in which each emotion is defined as a set of FAP values. New facial expressions, corresponding to different emotions, can be added to the emotional database.

The technological evolution for processing and storing videos and the greater accessibility to motion capture (mocap) systems, enabled the acquisition of large amounts of accurate data of natural head movements and facial expressions, offering the alternative of modeling speech articulatory movements and emotional facial expressions all together. Systems that adopt this data-driven strategy build a visual expressive speech model for each emotion label. During synthesis, the emotion input drives the selection of the appropriate learned model.

Beskow and Nordenberg (BESKOW; NORDENBERG, 2005) used an optical mocap system to track 29 markers attached to a subject's face. The subject produced short sentences playing happiness, anger, sadness, surprise and the neutral expressions. The captured data was converted to MPEG-4 FAPs of a 3D model and principal component analysis (PCA) was performed for dimensionality reduction. From the segmentation of the synchronized audio, the phonemes were mapped to 25 visemes categories for each expression and a Cohen-Massaro coarticulation model was trained for each one of the five emotions. The resulting models are used to generate trajectories of the principal components, that are further converted to FAPs

trajectories, resulting in the synthesis of expressive speech for the corresponding emotion.

In the work of Deng *et al.* (2006), the authors describe a methodology to derive a speech coarticulation model and an expressive space from motion capture data. In the work, an actress play 4 different emotional states (neutral, happy, angry and sad) with several markers attached to her face, in front of an optical mocap system. For each emotional state, the actress repeats the same sentences, that are designed to contain the most frequently diphones used in English. Although the collected corpus have samples of expressive speech, the visual speech and the emotional facial expression models are built independently. The methodology is divided in two separated phases. The first phase consists of training a speech coarticulation model. For this purpose, it is considered only the markers around the actress mouth (10 markers) and the data captured for neutral speech. From the segmentation of the audio, a database of diphones and triphones transitions is created. The model training process consists of finding the coefficients of time weighting polynomials that best describe the transitions among the phones. The second phase, is the modeling of a three-dimensional PIEES (Phoneme Independent Expressive Eigenspace). This three-dimensional space presents two characteristics: (1) the facial expression variations during the utterance of a sentence are represented by a continuous trajectory in the PIEES; (2) the trajectories associated to an emotional state occupies a characteristic region of the PIEES. To obtain the PIEES orthogonal axes directions, first, a phoneme-based time warping and resampling operation is performed in order to align the expressive mocap data with the neutral speech data. Second, the neutral motion is subtracted from the expressive motion data; a mean to obtain pure expressive motion signals, or Phoneme-Independent Expressive Motion Signals (PEMS). A PCA is applied to the obtained PIEMS and the first three principal components represent the PIEES axes. The input to the talking head synthesis is a sequence of speech phonemes and the desired emotional state output. The phonemes are mapped to a set of 13 visemes blendshapes and the neutral speech synthesis consists of applying the appropriate weighting polynomials transitions. The integration of emotional signaling with neutral visemes consists of defining a continuous space in the PIEES that will guarantee a smooth variation of facial expressions. The trajectory in the PIEES space is defined following a texture synthesis inspired approach: small fragments of trajectory are concatenated accordingly to grow a longer trajectory. Synthesized speech and expressive markers trajectories are blended together on the marker level. Finally, the synthesized marker motions are mapped to a 3D head model.

Instead of using mocap data, Zhang, Jia and colleagues used annotated data from the Japanese Female Facial Expression (JAFFE) database (LYONS *et al.*, 1998) to propose what they call an emotional text-to-audio-visual-speech (ETTAVS) (ZHANG *et al.*, 2010), (JIA *et al.*, 2011). Another novelty aspect of their work is the adoption of the dimensional (Pleasure-

Arousal-Dominance) PAD model of emotions (Figure 2.11, Section 2.2.2). The methodology consists of creating, first, a pseudo facial expression database. The JAFFE database consists of 213 expression images with 10 Japanese females posing three or four examples of seven basic expressions: neutral, happy, sad, surprise, angry, disgust and fear. Each image is manually annotated with 18 facial feature points in accordance to MPEG-4 facial definition points (FDPs). A group of trained annotators attributed values for the Pleasure, Arousal and Dominance dimensions for each expression in the JAFFE database, using a specific questionnaire for expression annotation and evaluation. Secondly, the facial feature points provided by the JAFFE database are mapped to Partial Expression Parameters (PEPs); an intermediate layer of facial expression representation that is created to reduce the complexity of dealing with high dimensional MPEG-4 FAPs. In summary, the pseudo facial expression database consists of a set of facial expressions that are represented by PAD values and associated PEPs. The information from the database is used to train a polynomial mapping model between PAD values and PEPs. The input to the ETTAVS system is a text message annotated with PAD values corresponding to the desired expressive output. Complementary modules are responsible for synthesizing the emotional speech audio and a sequence of parametrized Chinese visemes MPEG-4 FAPs. In parallel, the PAD values provided as input are mapped to PEPs, that are further mapped to MPEG-4 FAPs. Following, the visemes FAPs are blended to the expressive FAPs. The animation flow continuity is obtained through a modulation step — driven by prosodic features extracted from the speech audio — that guarantees smooth facial expressions transitions.

Liu and colleagues presented an expressive speech-driven 2D morphing-based talking head system, that is built upon Adaboost classifiers (LIU *et al.*, 2011). An audiovisual corpus was collected from a female speaker with sentences expressed in five emotions: neutral, anger, happiness, sadness and surprise. Following, 60 image frames were extracted from the corpus, corresponding to 12 Chinese visemes represented in each one of the emotional states. The feature vectors used to train the classifiers were built as a combination of acoustic, articulatory and prosodic features extracted from the speech audio and 58 landmarks manually annotated on the visemes images. Six classifiers were trained: an emotion classifier; and five phonemes classifiers, one for each emotional stimulus. The synthesis is driven by the new speech to be animated and it consists of four steps. First, the emotion classifier determines an emotion label for the speech. Second, the phoneme classifier corresponding to the previous determined emotion is used to determine the sequence of phonemes in the speech. Third, the emotion label and the sequence of phonemes are used to index the database of expressive visemes. Finally, the animation is synthesized following a morphing between keypose visemes.

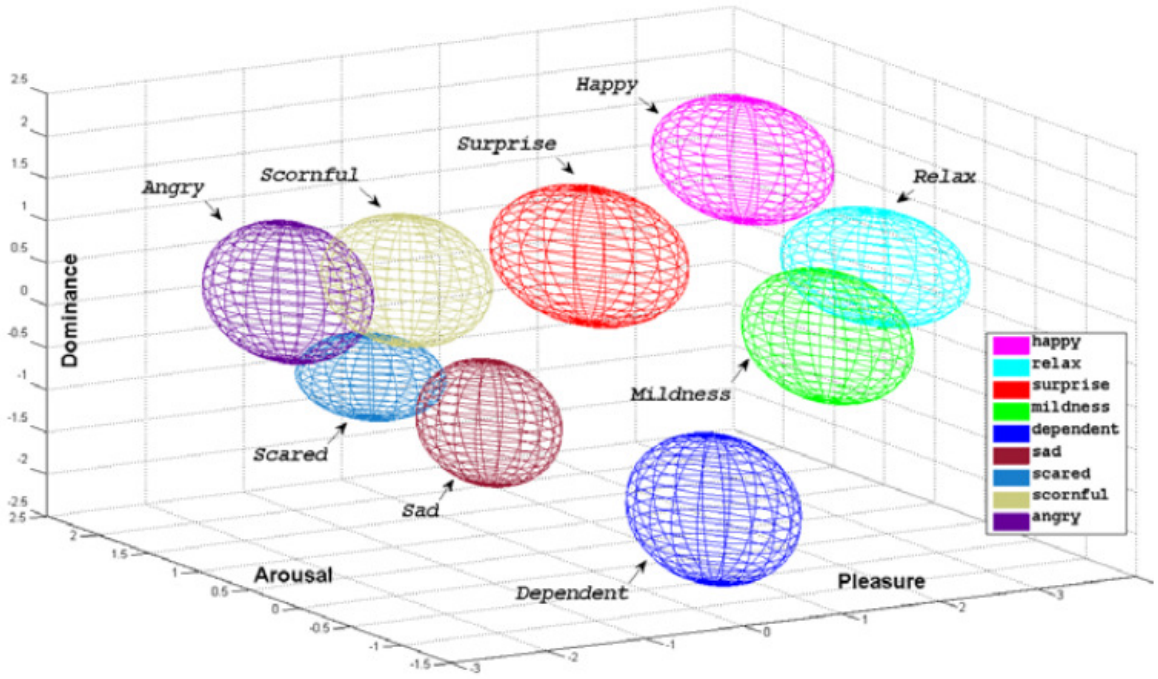


Figure 2.11 – The distribution of 9 emotional states/words in PAD emotion space used by Zhang *et al.* (2010). Source: Extracted from (ZHANG *et al.*, 2010).

2.3.2 Concatenative Systems

An approach proposed to obtain more natural visual speech modeling is to reduce the number of transitions between keyposes and consequently, the errors imposed by the artificial models that implement them. This alternative approach consists of concatenating segments of performances recorded from a subject (Figure 2.12).

The first step in this strategy is the building of a database of collected data (corpus) composed of 3D information obtained from face markers tracked by a motion capture (mocap) system; or 2D frames extracted from a recorded video. In a preprocessing stage, the corpus is segmented and indexed for future reuse. Different works may adopt different sizes and types of segments like: pairs of phones (diphones), sequences of three phones (triphones) or syllables. The size of the segment — or the number of phonetic transitions it comprehends — is directed related to the capability of the system to model coarticulation. Larger segments are capable of conveying more accurate reproduction of transitions, reducing the number of artificial transitions on the animated speech. However, the collection of all possible phonetic combinations grows quickly with the size of the segment. For this reason, concatenative synthesis quality is typically limited by the variety of transition samples captured in the corpus.

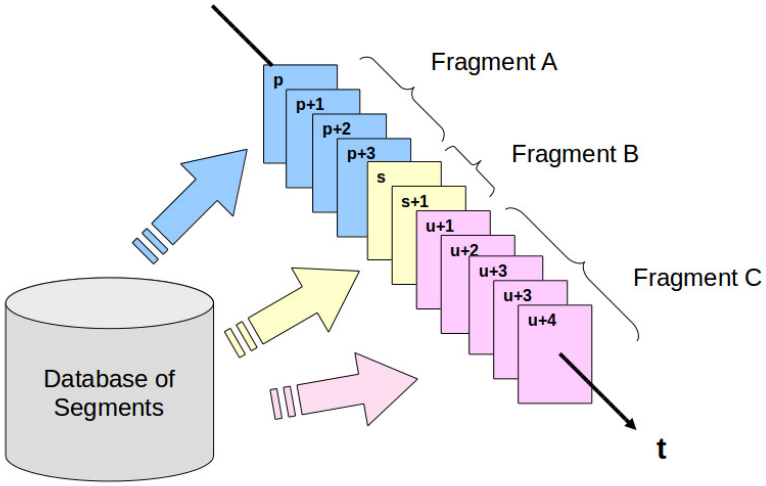


Figure 2.12 – Illustration of a concatenative synthesis approach. A database of segments extracted from a corpus allows the combination of sequences of frames to synthesize the final animation.

2.3.2.1 Neutral Speech Systems

Concatenative synthesis approach showed the potential of 2D models. An iconic work of this category is Video Rewrite (BREGLER *et al.*, 1997). The methodology proposed by Bregler *et al.* (1997) consists of building a database of triphones from the video of a subject, without restrictions concerning the content of the uttered sentences or the background scenario. The animation of new speech content is obtained concatenating and stitching together the appropriate triphone sequences from the database. Instead of concatenating triphone video units, Kshirsagar e Magnenat-Thalmann (2003) propose to concatenate syllables, defining their visual counterparts, named visyllables. Theobald *et al.* (2004) select triphone units from a database of trajectories on an Active Appearance Model (AAM) space. Jiang *et al.* (2008) propose a speech-driven system, in which the analysis of input speech audio provides information to select and concatenate an appropriate sequence of divisemes (visemes pairs) stored in a database.

The concatenative approach for 3D head models is typically based on mocap datasets. Edge *et al.* (2004) organize the motion database into fragments of sentences, words and diphones. In the work of Deng *et al.* (2005), mocap data is used to derive models of coarticulation for diphones an triphones. During synthesis, a dynamic programming technique is used to search for optimal combinations of diphones and triphones.

Some concatenative systems adopt a dynamic length of segments or a unit-selection approach. In such systems, the synthesis is performed trying to find the minimum cost tran-

sitions path in the space of samples formed by the database. In Cosatto e Graf (2000), for example, the system seeks to identify the best sequence of variable-length video sequences from a large image database. Liu e Ostermann (2009) implement a unit-selection algorithm performing a frame-by-frame selection from the database. In the work of Mattheyses *et al.* (2010) variable-length video segments are selected from a database of images that are projected on an AAM space.

The system proposed by Liu e Ostermann (2011) explores the smiling during speech. It implements a unit-selection algorithm, which selects and concatenates mouth image segments from a database of smiling lip shapes.

2.3.2.2 Expressive Speech Systems

The work of Cao *et al.* (2005) can be highlighted as an example of concatenative expressive speech talking head. The corpus built by the system is based on an optical mocap system with 8 infrared cameras used to track 109 markers that were attached to a professional actor's face. The actor uttered several sentences displaying five emotional states: frustrated, happy, sad, angry and neutral. The proposed methodology consists first, of creating a database of *animes*. According to the authors, "an anime captures a phoneme instance and contains a phoneme label, the associated motion segment and other audio information". The anime is the dynamic counterpart of a viseme: "Unlike a viseme that captures a single frame of facial pose, anime holds a number of motion frames." The database of animes is obtained, first, through the PCA of the mocap data in order to obtain a reduced number of principal components. Second, the synchronized speech audio is segmented, making possible the association between phonemes and motion frames. Third, a prosody feature vector is extracted from each phone segment. The last piece of information to integrate the anime, is the emotion label associated to the uttered sentence. The sequence of animes is organized in a graph. The talking head synthesis follows a concatenative unit-selection strategy, as illustrated in Figure 2.13. The input audio is processed by an emotion classifier that determines the emotion to be synthesized and by an automatic segmentation system that provides the sequence of phones to be animated. The synthesis process consists then, in finding the minimum cost path of transitions in the anime graph. After defining the path, the animation flow is made smooth through time warping and blending operations between adjacent animes. Finally, the principal components of the motion capture are mapped to a 3D head model. The head texture is defined from photographs from the actor's face.

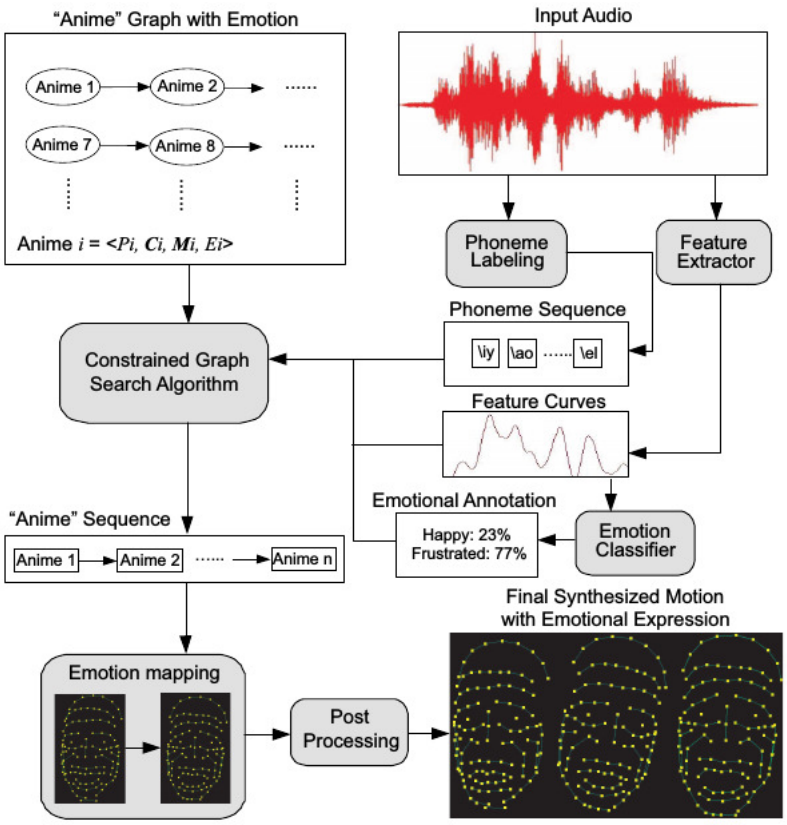


Figure 2.13 – Synthesis estrategy proposed by Cao *et al.* (2005). Source: Extracted from (CAO *et al.*, 2005).

2.3.3 Systems Based on Statistical Prediction

The synthesis based on statistical prediction applies machine learning techniques in order to derive a mathematical model from the analysis of collected data (corpus) from a real subject, also called the training dataset. An initial training phase consists of statistically modeling the correlation among the corpus speech parameters, the corresponding static facial displays properties and the dynamics of the observed transitions between states. The feature vectors that are used to train the model typically consist of a combination of phonemes/visemes labels associated to audio speech parameters (like Mel-frequency cepstral coefficients and log-F0) and visual parameters derived from the 3D coordinates of mocap markers or image features extracted from recorded video frames. Given a new target input sequence of phonemes/visemes or given the features extracted from a a speech to be animated, the trained model predicts the most likely transitions of states, generating new feature vectors, implementing a synthesis-by-analysis pipeline.

2.3.3.1 Neutral Speech Systems

In Voice Puppetry, Brand proposes to transfer the visual speech modeling derived from a collected corpus to static images or photographies (BRAND, 1999). The methodology is based on a recorded video used as training data to build a finite state machine in which each state has an output probability distribution over facial configurations and their corresponding acoustic features. The proposed approach models coarticulation through a Hidden Markov Model (HMM), which is used for predicting facial configuration sequences. Ezzat and colleagues used samples from a recorded corpus to build a Multidimensional Morphable Model (MMM). The MMM model is capable of conveying mouth appearance and shape variation from the corpus. In this space, each phoneme tends to form a cluster, modeled as a multidimensional Gaussian function. The magnitude of the variance of each phoneme implicitly models coarticulation effects. Clusters with higher variances correspond to phonemes that are more sensitive to undergoing visual effects of coarticulation (EZZAT *et al.*, 2002).

Cosker e Marshall (2004) explore the correlation among speech, articulatory movements and the non-verbal signaling through the combination of visual parameters provided by an active appearance model (AAM), with speech signal parameters (such as Mel-frequency cepstral coefficients or pitch). The parameters are used to train an HMM speech-driven synthesis model.

Mattheyses e Verhelst (2015) call the attention to the fact that some systems implement an hybrid approach, in which the statistical prediction model is applied as an improved unit-selection algorithm on systems that adopt the concatenative synthesis approach. In the work of Govokhina *et al.* (2006), for example, an HMM is trained from mocap data. During synthesis, the HMM provides a trajectory planning that is used to select the sequence of diphones that synthesizes the final animation. In the work of Wang e Soong (2014), an HMM is built from an audiovisual database of a person lip movement. Again, the HMM is used to select, from the original database, an optimal sequence of lip images, that are lately stitched to a background head video.

2.3.3.2 Expressive Speech Systems

Recently, machine learning techniques have also been applied to explore the correlation between the speech signal, the visible articulatory movements and the facial expressions.

The work of Tao *et al.* (2009) is an hybrid system that derives an statistical prediction model to select the sequence of segments for a concatenative synthesis approach. The system maps mocap data of expressive speech (happiness, sadness, anger and surprise) to the same content with neutral expression. They combine the neutral speech mocap data with the

acoustic and prosodic features extracted from the corresponding expressive speech, to create a vector of features used to train a fused Hidden Markov Model (HMM). The model drives the appropriate selection of subsequences of mocap data stored in a database, that are further concatenated and control the animation of a 3D head model.

Anderson *et al.* (2013) present a 2D text-to-visual speech synthesizer, a system capable of synthesizing both speech audio and facial animation in parallel. The training data was extracted from a video corpus of a female face uttering sentences in six emotional states: neutral, angry, happy, sad, tender and fearful. A subset of the corpus video frames was used to build an Active Appearance Model (AAM), from which is possible to express a facial image in terms of shape and appearance parameters. The final feature vector used to train the system is a combination of speech parameters (like Mel-cepstral coefficients and log-F0) and the corpus video frames AAM shape and appearance parameters. The set of feature vectors built from the corpus is the input to the Cluster Adaptive Training (CAT) process. The CAT methodology is an extension to the Hidden Markov Model (HMM). The output model is linear combination expression. The interpolation weights can be interpreted as a space of emotions. Trajectories in that space controls the expressiveness in the synthesised speech and facial animation, that may also result in new emotions, not seen in the corpus.

A recent study conducted by Ding and colleagues, proposes the application of deep neural networks (DNN) to train a speech-driven talking head system (DING *et al.*, 2014). The work focus on the evaluation of synthesized non-verbal communication expressed by the head movement, using as training data a large audiovisual database collected from the NBC English broadcast news. The study presented encouraging results to the exploration of this technique for emotional visual speech synthesis.

2.3.4 Discussion

Table 2.1, provides a summary of the expressive speech talking head systems reviewed in this section organized according to their main characteristics. The first column classifies the systems according to the synthesis strategy.Following, the second column of Table 2.1 indicates the head and face modeling approach used by the systems. The third column classifies the model of emotions according to the categories presented in Section 2.2. The fourth column refers to the set of emotions modeled by each system as discussed in Section 2.3. Finally, the last column indicates the associated references.

The comparison among different talking heads synthesis methodologies is a difficult task for many reasons, including:

- there is no universal accepted criteria to assess the level of videorealism of a facial

animation system output;

- a facial animation can be analyzed from different perspectives such as: the level of speech intelligibility it provides, how well it reproduces the structure of a real face (level of photorealism), its capability of expressing emotions, the level of comfort, thrust or empathy inspired in the user, etc.;

- some works do not provide access to demo videos or do not report any evaluation;

- many methodologies are built upon corpora material that are not made public or accessible for alternative implementation tests;

- some implementations depend on third-party modules like text-to-speech synthesizers or speech recognition systems that can influence the talking head system output;

- numerical comparisons between synthesized and ground-truth signals are easier to implement but they do not necessarily provide useful feedback about the quality perceived by the user and vice versa.

Nevertheless, it is possible to state that the artificial modeling of transitions between adjacent keyframes on rule-based systems may result on a perceived low level of videorealism. In the works of Scott *et al.* (1994) and Ezzat e Poggio (1998), for example, the transition between key-visemes is linear and the speech coarticulation effects are not modeled, resulting in a poor representation of the visual speech dynamics. One advantage of rule-based systems is their flexibility, since the same set of rules can be applied to different face models.

On concatenative systems, the need for sophisticated transition models required by the rule-based systems is overcome by samples of real speech transitions. However, this approach requires large databases of recorded video or mocap data in order to guarantee an embracing set of transitions samples. Besides, new face models require the building of new corpora. Concatenative talking heads are suitable for specific context applications, in which is possible to define a finite combination of words and sentences to be synthesized.

Finally, systems based on statistical prediction convert large datasets on compact statistical models, while delivering fair videorealism levels; however, as every data-driven speech synthesis approach, the output quality is dependent on the properties of collected data, including: the variability of features captured by the corpus, the quality of the recordings and the robustness of the features extraction process.

Synthesis Strategy	Head Model	Model of Emotions (Type)	Model of Emotions	References
Rule-based	3D	Categorical	Ekman's "big six"	Pelachaud *et al.* (1996)
			Ekman's "big six"	Binsted e Luke (1999)
			Ekman's "big six"	Albrecht *et al.* (2002a)
			Ekman's "big six"	Kshirsagar (2002)
			> Ekman's "big six"	Pelachaud e Bilvi (2003)
			Ekman's "big six"	Cosi *et al.* (2004)
			happiness, anger, sadness, surprise	Beskow e Nordenberg (2005)
			happiness, anger, sadness	Deng *et al.* (2006)
			> Ekman's "big six"	Bevacqua *et al.* (2007)
			> Ekman's "big six"	Queiroz *et al.* (2009)
		Dimensional	Whissel's dictionary of affect	Tsapatsoulis *et al.* (2002)
			PAD (Pleasure, Arousal, Dominance)	Zhang *et al.* (2010), Jia *et al.* (2011)
	2D	Categorical	happiness, anger, sadness, surprise	Liu *et al.* (2011)
		Appraisal	OCC (Ortony, Clore and Collins)	**THIS WORK**
Concatenative	3D	Categorical	frustration, happiness, sadness, anger	Cao *et al.* (2005)
Statistical Prediction	3D	Categorical	happiness, anger, sadness, surprise	Tao *et al.* (2009)
	2D	Dimensional	Not specified (AAM space)	Anderson *et al.* (2013)

Table 2.1 – Summary of expressive speech talking heads reviewed in Section 2.3. The works from Pelachaud e Bilvi (2003), Bevacqua *et al.* (2007) and Queiroz *et al.* (2009) are extensible frameworks which enable the definition of facial expression beyond the Ekman's six basic facial expressions.

2.4 Concluding Remarks

This chapter presented a historical perspective about the modeling of human emotions and the modern demand for affective computing systems (Section 2.1). It also presented a classification of the most important theories of emotion from the perspective of computational models of emotions (Section 2.2).

In Section 2.3, the review of the literature of talking heads showed that the techniques to synthesize speech synchronized facial animations have been evolved to obtain a more realistic representation of the visible speech articulatory dynamics. The same techniques have been applied to the synthesis of expressive speech talking heads and it is possible to observe that the majority of these systems adopts 3D head models. Three-dimensional head models are characterized by their flexibility, providing great control to the animator entity. They enable not only the setting of a wide range of speech articulatory poses, but also the

setup of an unlimited number of facial expressions through the control of elements like: the head movement, the blinking of the eyes, the eye gaze control, the eyebrow movement and the skin texture characteristics.

The review also showed the great influence of the Ekman's "big six" model of emotions in the systems implemented so far (see Table 2.1). As discussed in Section 2.2.4, this approach has limited capability to reproduce everyday dialogue episodes. Like the categorical approach, the dimensional models of emotions do not provide any mechanism to associate events or situations to points in the space of emotions.

The following chapters present a 2D rule-based expressive talking head that is based on the Ortony, Clore and Collins (OCC) appraisal model of emotions. As presented in Section 2.2.3, the OCC model offers a more complex but still concise and clear vocabulary of 22 emotions; it embraces mechanisms to control the intensity of emotions and, most important, it enables the integration of the talking head with any intelligent system capable of simulating an appraisal process triggered by an event, an user action or an object — a key aspect of the development of embodied conversational agents (ECAs) (COSTA *et al.*, 2012; COSTA; DE MARTINO, 2014a). Moreover, the image-based synthesis methodology adopted in this work comprehends the inherent modeling of the face elements structure and appearance, an important characteristic of videorealistic talking heads. Following a keyframe synthesis strategy, the generation of the animation keyposes is driven by the timed phonetic transcription of the speech to be animated and the desired emotional state. The keyposes consists of expressive context-dependent visemes that implicitly model the speech coarticulation effects. The transition between adjacent keyposes is performed through a non-linear image morphing algorithm.

3 Building an Expressive Corpus for Brazilian Portuguese

Among the contributions of the present work is the building of a comprehensive expressive speech corpus, obtained under controlled conditions. The corpus provides multimodal samples of expressive speech including: high quality speech audio, digital high definition (HD) video recordings and three-dimensional motion capture (mocap) data.

The present chapter is divided into three sections. The first section (Section 3.1) describes the profile of the participants, the methodology designed to build the corpus and the technical setup of the recording sessions. Section 3.2 describes the creation of *CH-Unicamp*: an annotated database of expressive visemes images played by a female face[1]. *CH-Unicamp* is the training database used to build the active appearance model (AAM) described in Chapter 4. Section 3.3 presents the concluding remarks of the chapter.

3.1 Expressive Speech Brazilian Portuguese Corpus

The building process of audiovisual speech corpora may adopt two strategies:

- the capture of material under controlled conditions, typically performed in a lab or in a TV studio, with a predefined set of utterances;
- or the selection of material recorded for different purposes, without restrictions concerning the speech content, the scenario background, the informant characteristics or the nature and the quality of the recorded signal (like video excerpts extracted from TV programs or personal videos made public in the web).

In the present work, the corpus was captured under controlled conditions, enabling the recording of audio, video and motion capture data of predefined utterances that contain samples of all Brazilian Portuguese language phonemes, in particular phonetic contexts of interest. Additionally, the same utterances were produced by different subjects, making possible the study of inter-subjects variability of expressions.

A drawback of recordings in a controlled environment is the fact that the performances are not, typically, spontaneous expressions of emotions. For this reason, the subjects chosen

[1] The CH letters refer to the actress' initials.

to participate in the experiments were experienced actors and actresses with the appropriate training to play different emotion scripts with the proper speech articulation and the adequate acting style for video recording. The profile of the participants is detailed in Section 3.1.1.

The built corpus is the result of two experiments. The first experiment required the actors to represent the twenty-two emotions of the Ortony, Clore and Collins (OCC) model of emotions (Section 3.1.3). For that purpose, twenty-two recording scripts were designed coherently with the cognitive state description given by the model ensuring that each speech has occurrences of all Brazilian Portuguese context-dependent visemes (visual phonemes), described in Section 3.1.2. Each speech segment was recorded with a neutral expression, followed by the actor's representation of the corresponding emotional state. In a second experiment, the objective was to obtain data to investigate how the expressive signals observed in the face are modulated by different personality traits (Section 3.1.4). For that purpose, we adopted the "Big Six" emotions of Ekman and the actors were asked to represent them with three different extroversion characteristics: shy, neutral/balanced and extroverted. In this experiment, the speech segments were the same for all emotions and personalities. Each experiment consisted in a motion capture session and a video recording session, with synchronous capture of audio. The technical setup of the recording sessions is detailed in Section 3.1.5.

3.1.1 Participant's Profile

Four actresses and six actors, aged between 20 and 60, volunteered to participate in the experiments. They were recruited through e-mail messages sent to schools of arts and through a homepage that explained the nature and the objective of the experiments. All of them signed a document giving the permission of use of their images for research purposes.

One actress and three actors that participated in the experiments were professionals with several years of experience. The remaining actresses and actors were students from the last year of the undergraduate school of arts at the University of Campinas. All of them are native Brazilian Portuguese speakers.

The professionals were randomly divided into two groups. The first group, composed of two actresses and four actors, participated in the OCC emotions experiment. The second group, composed of two actresses and two actors, participated in the personality trait experiment. Figure 3.1 presents pictures of actors and actresses with markers attached to their faces during the motion capture session.

Figure 3.1 – Actors and actresses with markers attached to their faces during the motion capture session.

3.1.2 Context-Dependent Visemes

In order to obtain representative samples of the speech articulatory movements, all the utterances captured to build the corpus were designed to contain samples of all the context-dependent visemes of Brazilian Portuguese.

De Martino et al. (2006) identified that due to coarticulation, some visemes, or "visual phonemes", may present perceptible variations in their dynamics depending on the phonetic context in which they are produced. The mapping of these variations for the most common phonetic contexts of a language gives origin to the definition of context-dependent visemes (CDVs). The authors developed a methodology that can be applied to identify the CDVs of any language. The methodology involves, first, the identification of the language homorganic groups, i.e. groups of different phonemes that share the same place of articulation, which are not distinguishable by visual cues alone. Second, with the help of a motion capture apparatus, the visual motor pattern for each homorganic group is measured and analyzed under different phonetic contexts. Finally, with the help of a clusterization algorithm, it is possible to identify different visemes that can be associated to the same homorganic group depending on the phonetic context in which they are produced.

Following this methodology, Tables 3.1 and 3.2 present the CDVs identified for Brazilian Portuguese (DE MARTINO *et al.*, 2006). The first columns of the tables show respectively, the consonantal and vocalic homorganic groups identified for the language. In Table 3.2, for example, it is possible to observe that the oral vowels [i], [e], [a], [o] and [u] are grouped with the nasal vowels [ĩ], [ẽ], [ɐ̃], [õ] and [ũ]; since they are not distinguishable by

Homorganic Groups	Context-Dependent Visemes	Phonetic Contexts
[p,b,m]	$< p_1 >$	[pi] [pa] [ipɪ] [ipɐ] [ipʊ] [apɪ] [apɐ] [apʊ] [upɐ]
	$< p_2 >$	[pu] [upɪ] [upʊ]
[f,v]	$< f_1 >$	[fi] [fa] [ifɪ] [ifɐ] [ifʊ] [afɪ] [afɐ]
	$< f_2 >$	[fu] [afʊ] [ufɪ] [ufɐ] [ufʊ]
[t,d,n]	$< t_1 >$	[ti] [tu] [itɪ] [itɐ] [itʊ] [atɪ] [atʊ] [utɪ] [utɐ] [utʊ]
	$< t_2 >$	[ta] [atɐ]
[s,z]	$< s_1 >$	[si] [sa] [isɪ] [isɐ] [asɪ] [asɐ]
	$< s_2 >$	[su] [isʊ] [asʊ] [usɪ] [usɐ] [usʊ]
[l]	$< l_1 >$	[li] [ilɪ] [alʊ] [ulɪ] [ulɐ]
	$< l_2 >$	[la] [ilɐ] [alɪ] [alɐ]
	$< l_3 >$	[lu]
	$< l_4 >$	[ilʊ] [ulʊ]
[ʃ,ʒ]	$< ʃ_1 >$	[ʃi] [ʃa] [iʃɪ] [iʃɐ] [iʃʊ] [aʃɪ] [aʃɐ] [aʃʊ] [uʃɪ] [uʃɐ]
	$< ʃ_2 >$	[ʃu] [uʃʊ]
[ʎ,ɲ]	$< ʎ_1 >$	[ʎi] [ʎa] [iʎɪ] [iʎɐ] [aʎɪ] [aʎɐ]
	$< ʎ_2 >$	[ʎu] [uʎɪ] [uʎɐ]
	$< ʎ_3 >$	[iʎʊ] [aʎʊ] [uʎʊ]
[k,g]	$< k_1 >$	[ki] [ikɪ] [ikɐ] [akɪ] [ukɪ] [ukɐ]
	$< k_2 >$	[ka] [akɐ]
	$< k_3 >$	[ku] [ikʊ] [akʊ] [ukʊ]
[ɣ],[ɾ]	$< ɣ_1 >$	[ɣi][1] [ɣa][1] [iɣɪ] [iɣɐ] [aɣɪ] [aɣɐ] [uɣɐ]
	$< ɣ_2 >$	[ɣʊ][1] [iɣʊ] [aɣʊ] [uɣɪ] [uɣʊ]

Table 3.1 – Consonantal context-dependent visemes (adapted from (DE MARTINO *et al.*, 2006)). The phonetic symbols are from the International Phonetic Alphabet (INTERNATIONAL PHONETIC ASSOCIATION, 1999).

visual cues alone.

The second columns of Tables 3.1 and 3.2 show the CDVs; and the third column, the phonetic contexts they represent. For the sake of simplicity, the context-dependent visemes are named with the first phone of each homorganic group. For example: two context-dependent visemes were identified in the homorganic group [p,b,m] and the [p] phone is used

[1] Phonetic context not valid for [ɾ] phone. The "tap" is never observed in the beggining of words in Brazilian Portuguese.

in their names ($< p_1 >$ and $< p_2 >$) since it is the first phone of the group.

It is possible to observe that not all consonantal phonetic contexts that exist in Brazilian Portuguese are represented in Table 3.1. However, based on the knowledge of the language, it is possible to map any phonetic context encountered in the language to one of the CDVs shown in Table 3.1. The rules applied to perform this mapping is out of the scope of this text but they are detailed in (COSTA, 2009).

Homorganic Groups	**Context-Dependent Visemes**	**Phonetic Contexts**
[i,ĩ]	$< i_1 >$	All contexts except [tit] and [ʃiʃ].
	$< i_2 >$	[tit] and [ʃiʃ].
[e,ẽ]	< e >	All contexts.
[ɛ]	< ɛ>	All contexts.
[a,ɐ̃]	< a >	All contexts.
[ɔ]	< ɔ>	All contexts.
[o,õ]	$< o >$	All contexts.
[u,ũ]	$< u >$	All contexts.
[ɪ]	< ɪ>	All contexts.
[ɐ]	< ɐ>	All contexts.
[ʊ]	< ʊ>	All contexts.

Table 3.2 – Vocalic context-dependent visemes (adapted from (DE MARTINO *et al.*, 2006)). The phonetic symbols are from the International Phonetic Alphabet (INTERNATIONAL PHONETIC ASSOCIATION, 1999).

3.1.3 OCC Emotion Experiment

As discussed in Chapter 2, the OCC model of emotions presents an alternative to the Ekman's six emotions vocabulary. The OCC model comprehends a vocabulary of twenty-two emotion types that correspond to cognitive "meanings" associated to the logical operations involved during the appraisal process of the emotions (SCHERER, 1999). The OCC model presents a clear appraisal structure; a comprehensive yet simple vocabulary of emotions; and it also provides parameters like valence and emotion intensity. For these reasons, the OCC model is considered well suited and adaptable for the development of affective computational systems, justifying its adoption in the building of the present expressive speech corpus (BARTNECK, 2002), (STEUNEBRINK *et al.*, 2009), (KSHIRSAGAR, 2002), (EGGES *et al.*, 2004).

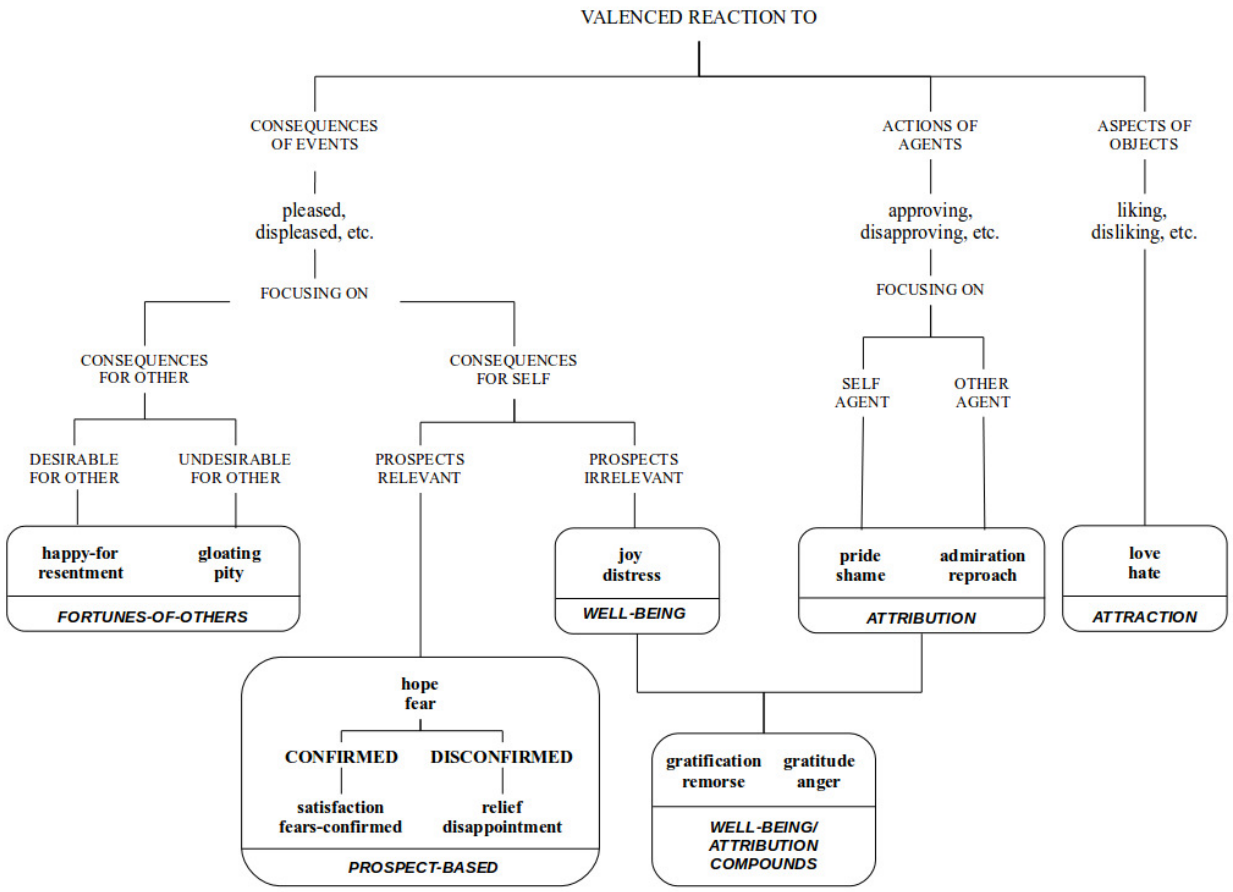


Figure 3.2 – Ortony, Clore and Collins (OCC) model structure of emotions. Adapted from (ORTONY *et al.*, 1988).

Figure 3.2, presents the structure of emotions of the OCC model. The structure is hierarchically organized into three branches that divides the emotions concerning consequences of events (e.g. happy-for, resentment, relief), actions of agents (e.g. pride, shame, reproach) and the attractiveness of things (love and hate). In the figure, the evaluated valence of the elicitator is expressed by terms like: pleased vs. displeased, approving vs. disapproving, etc. The structure describes the appraisal process that occurs from the individual's perspective (e.g. "CONSEQUENCES FOR SELF", "SELF AGENT") but it also includes the interpretation of the others' perspective (e.g. "CONSEQUENCES FOR OTHER", "DESIRABLE FOR OTHER", "OTHER AGENT").

Each OCC emotion type has a formal specification. "Joy", for example, is specified as follows (ORTONY *et al.*, 1988):

JOY EMOTIONS
TYPE SPECIFICATION: (pleased about) a desirable event
TOKENS: contented, cheerful, delighted, ecstatic, elated, euphoric, feeling good, glad, happy, joyful, jubilant, pleasantly surprised, pleased, etc.
VARIABLES AFFECTING INTENSITY:
(1) the degree to which the event is desirable
EXAMPLE: The man was pleased when he realized he was to get a small inheritance from an unknown distant relative.

The example shows that each emotion specification has five major components:

- **Type identification:** a convenient label for the emotion type, e.g. "Joy emotions";
- **Type specification:** a sentence describing the type of reaction (enclosed in parenthesis) followed by the eliciting conditions of the emotion type, e.g. "(pleased about) a desirable event";
- **Tokens:** a list of words that share the same type specifications;
- **Variables affecting intensity**: a description of the variables that affects the intensity of emotion type;
- **Example:** a prototypical example.

The type specifications of all the 22 OCC emotions as described in (ORTONY *et al.*, 1988) are shown in Table 3.3.

Following these specifications, twenty-two Brazilian Portuguese texts were designed to be played by the actors during the sessions. All the texts suggest a dialogue situation: the character played by the actor is talking (with emotion) about something to one or more imaginary characters.

The texts were intentionally designed to not elicit exaggerated or stereotyped expressions of emotions. For "Fear", for example, the objective was not to refer to a situation of terror or panic, which is rarely observed in dialogue situations. Instead, the text refers to a situation where the character fears the serious consequences of a contract not being signed.

Additionally, all the texts were designed to contain samples of Brazilian Portuguese context-dependent visemes, as described in Section 3.1.2.

OCC Emotion	Type Specification
Joy	(pleased about) a desirable event
Distress/Sadness	(displeased about) an undesirable event
Happy-for	(pleased about) an event presumed to be desirable for someone else
Pity	(displeased about) an event presumed to be undesirable for someone else
Gloating	(pleased about) an event presumed to be undesirable for someone else
Resentment	(displeased about) an event presumed to be desirable for someone else
Hope	(pleased about) the prospect of a desirable event
Fear	(displeased about) the prospect of an undesirable event
Satisfaction	(pleased about) the confirmation of the prospect of an undesirable event
Fears-confirmed	(displeased about) the confirmation of the prospect of an undesirable event
Relief	(pleased about) the disconfirmation of the prospect of an undesirable event
Disappointment	(displeased about) the disconfirmation of the prospect of a desirable event
Pride	(approving of) one's own praiseworthy action
Shame	(disapproving of) one's own blameworthy action
Admiration	(approving of) someone else's praiseworthy action
Reproach	(disapproving of) someone else's blameworthy action
Gratification	(approving of) one's own praiseworthy action and (being pleased about)the related desirable event
Remorse	(disapproving of) one's own blameworthy action and (being displeased about) the related undesirable event
Gratitude	(approving of) someone else's praiseworthy action and (being pleasedabout) the related desirable event
Anger	(disapproving of) someone else's blameworthy action and (being displeased about) the related undesirable event
Love	(liking) an appealing object
Hate	(disliking) an unappealing object

Table 3.3 – OCC emotions type specifications.

As an example, the framed boxes below show the text that was played by the actors to express "Fear". The first box shows the original text in Brazilian Portuguese, followed by the text translation to English. Table 3.4 shows excerpts from the text from which is possible to extract samples of context-dependent visemes. The complete set of texts for each emotional state, is reproduced in Annex A.

"Fear" (Brazilian Portuguese)

Lucas... Tulha... Estou muito preocupado... Se não conseguirmos este contrato, tudo que realizei e pelo qual batalhei nesta vida pode ser arrasado. Sem este contrato ficarei sem dinheiro para pagar o que devo para o Lilo e o Juliano. Eles me tomarão a casa e o carro. Nunca mais poderei olhar com orgulho para minha família. E o pior, é que já passei por dificuldades no passado e sei que, nessas horas, muitos dos que se dizem

meus amigos, simplesmente sumirão... Sei que estarei sozinho e não terei para quem pedir ajuda.

"Fear" (Translation to English)

Lucas... Tulha... I'm very worried... If we don't get the contract, all that I've realized and for which I have struggled during my whole life can be destroyed. Without this contract, I won't have the money to pay what I owe to Lilo and Juliano. They will take my home and my car. I will never be able to look again to my family with pride. And the worst part is that I've already been in difficulties in the past and I know that, in such situations, many who call themselves "friends", will simply disappear... I know that I will be alone and no one will help me.

During the recording sessions, the OCC emotion experiment required that the actors, first, enunciate the text with a neutral expression. Following, the actors played the text according to the specified emotional state. Few directions were provided to the actors, restricted to the type specification of each emotion. Each session of the OCC emotion experiment had an average duration of 90 minutes.

3.1.4 Personality Trait Experiment

The speech rate (typically expressed in terms of words per minute) has great influence on the visual motor patterns of speech segments. Similarly, factors like the personality of the informant influence the visual motor patterns of expressive speech segments.

In order to provide experimental material to the study of such phenomena, the present experiment required the actors to play different personality traits. The personality traits were modeled as three levels of activation (as defined in Chapter 2, Section 2.2.2, the *activation* level refers to the readiness for action). In the first level of activation, the actors were instructed to play a character that does not clearly expresses his emotions through actions. The actors and the actresses that participated in this experiment, associated this description to a character that is shy. The second description corresponds to a balanced level of activation. The balanced level was associated to a pattern of expression that is considered "polite" or "socially acceptable". Finally, the third level of activation was described as a character that clearly expresses his emotions through actions like smiling widely, showing anger with no restrictions or almost crying when expressing sadness.

Context-Dependent Visemes	Text Excerpts	Phonetic Transcription
$< p_1 >$	"(...) conseguirei pagar (...)"	kõsegiɾeɪ pagɐɾ
$< p_2 >$	"(...) o pior (...)"	u piɔɾ
$< f_1 >$	"(...) minha família (...)"	mĩɲɐ famiʎɾɐ
$< f_2 >$	"(...) contrato ficarei (...)"	kõtaɾatʊfikaɾeɪ
$< t_1 >$	"Tulha (...)"	tuʎɐ
$< t_2 >$	"(...) batalhei (...)"	bataʎeɪ
$< s_1 >$	"(...) passado (...)"	pasadʊ
$< s_2 >$	"(...) simplesmente sumirão (...)"	sĩpelesmẽtɪsumiɾãʊ
$< l_1 >$	"(...) Juliano (...)"	ʒuliãnʊ
$< l_2 >$	"(...) realizei (...)"	realizeɪ
$< l_3 >$	"Lucas (...)"	lukɐs
$< l_4 >$	"(...) Lilo (...)"	lilʊ
$< ʃ_1 >$	"(...) ajuda (...)"	aʒudɐ
$< ʃ_2 >$	"(...) o Juliano (...)"	u ʒlɪfãnʊ
$< ʎ_1 >$	"(...) família (...)"	famiʎɐ
$< ʎ_2 >$	"Tulha (...)"	tuʎɐ
$< ʎ_3 >$	"(...) sozinho (....)"	soziɲʊ
$< k_1 >$	"Lucas (...)"	lukɐs
$< k_2 >$	"(...) linda casa (...)"	lĩdɐ kazɐ
$< k_3 >$	"(...) preocupado (...)"	peɾeʊkupadʊ
$< ɾ_1 >$	"(...) arrasado (...)"	aɣazadʊ
$< ɾ_2 >$	"(...) carro (...)"	kaɣʊ
$< i_1 >$	"(...) Lilo (...)"	lilʊ
$< i_2 >$	"(...) dizem (...)"	dizẽ
< e >	"(...) devo (...)"	devʊ
< ɛ>	"(...) pior é que (...)"	piɔɾ ɛ ki
< a >	"(...) casa (...)"	kazɐ
< ɔ>	"(...) sozinho (...)"	soziɲʊ
$< o >$	"(...) poderei (...)"	poderɐɪ
$< u >$	"Lucas (...)"	lukɐs
< ɪ>	"Sei que (...)"	seɪ ki
< ɐ>	"(...) casa (...)"	kazɐ
< ʊ>	"(...) Lilo (...)"	lilʊ

Table 3.4 – Examples of excerpts from the "Fear" text from which is possible to extract samples of context-dependent visemes.

A unique phonetically rich text was designed for this experiment. In this case, the text was designed to be neutral in order to enable the interpretation of different emotions. The actors and actresses first uttered the text with neutral expression. Following, for each activation level, the same text was played to express the six emotions of Ekman. Each session

of the personality trait experiment had an average duration of 60 minutes. The uttered text in Brazilian Portuguese is presented in the following frame.

> Não é possível, você tem certeza disso? Prepare tudo. Filho xarope! Lavou a unha? Filha, rúcula para a pata. Passarinho, cuidado com a asa. Gato vamos! Lilo, Kika, Luku, puxem o cavalo! O que é isso? Chuva!

3.1.5 Motion Capture and Video Sessions

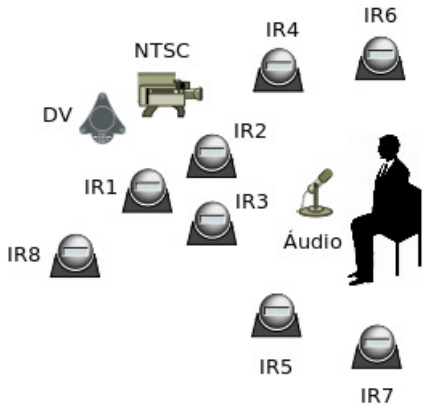
The experiments described in Sections 3.1.3 and 3.1.4 were performed twice: on a motion capture (mocap) session and a video recording session.

The mocap sessions took place at the Center for Information Technology Renato Archer – CTI, in Campinas (Brazil) (COSTA; DE MARTINO, 2012).

Sixty-three reflective markers were distributed over the actors' face and head (Figure 3.3). The markers were tracked by a "Vicon™" system, at capture rate of 120 fps (frames per second). The mocap setup was composed of eight infrared cameras plus a digital video camera. Additionally, the experiment setup included a high resolution LCD display that was placed in front of the subject to display the utterances; and a high fidelity microphone, attached to a video camera synchronized to the motion capture system (Figures 3.4(a) and 3.4(b)).

Figure 3.3 – Sixty three reflective markers were attached to the actors' face and head during the motion capture session.

In the video recording session, the actor repeated the performance in front of a chroma-key background, without markers, makeup or accessories in his head and face. The video was recorded using a HD 1920×1080 pixels, NTSC 29.97 fps digital video camera. The audio recording was synchronous to the video recording. A teleprompter device was used to show



(a) Cameras Setup

(b) Motion Capture Session

Figure 3.4 – The mocap session setup included eight infrared cameras (IR1 to IR8), a digital video camera (DV) and an analogue video camera (NTSC) to capture the audio.

the utterances to the participant. The video sessions were performed in the TV studio of the University of Campinas (RTV Unicamp). Figure 3.5 shows a picture of the video session setup.

Figure 3.5 – Picture of the TV studio during the recording of a video session.

3.2 CH-Unicamp Expressive Viseme Images Database

Part of the expressive speech corpus presented in Section 3.1, was processed to create the *CH-Unicamp* expressive viseme images database. Among the actors and actresses performances captured to build the corpus, the video material of a female actress playing the twenty-two OCC emotions was selected (the CH letters are her initials). The selection was

Figure 3.6 – The timed phonetic transcription of the speech audio, makes possible the association of small sequences of video frames to intervals of production of phones. In the figure, for example, the frames (n+3) and (n+4) are visemes associated to the production of phone [f], while the frames (n+2) and (n+5) represent the transitions between adjacent phones. A representative viseme for a phone is selected observing the inflection point of the mouth articulators. For [f], the final configuration of the mouth articulators are reached at frame (n+4). For [ɛ], the maximum excursion of the lips to produce the vocalic sound is presented by frame (n+6).

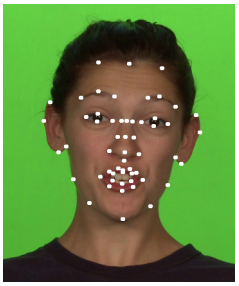


Figure 3.7 – The coordinates of fifty-six feature points are associated to each image in the *CH-Unicamp* database.

made after a visual inspection of the video, to guarantee that the chosen face would not present any particular characteristic that would deviate the attention of an observer, such as: scars, uncommon spots in the facial skin or deep wrinkles of age in the forehead.

For each OCC emotion, the actress video material was divided in two: part of the material was processed to create the image database, characterizing a training dataset; and the complementary part was reserved for cross-validation tests, characterizing a test dataset. In the present work, the test dataset is used to generate test stimuli for a perceptual evaluation (Chapter 6).

The audio tracks extracted from the video clips were manually segmented using Praat

(a scientific computer software application for the analysis of speech in phonetics) (BOERSMA; WEENINK, 2001). The segmentation process resulted in the timed phonetic transcription of the speech utterances, making possible the association of small sequences of video frames to intervals of production of phones. For each phone of interest, a representative viseme is selected observing the inflection point of the visible mouth articulators (lips, tongue and teeth). Figure 3.6 shows a sequence of phones and their corresponding visemes. The fricative phone [f], for example, is produced touching the lower lip in the superior teeth and letting the air flow through the mouth. In the picture, the frame (n+2) represent the onset of the mouth articulators. Their final configuration is reached at frame (n+4). Frame (n+5) shows the release of the lips to prepare the following vocalic sound. In this case, the frame (n+4) would be selected to be the representative viseme for the [f] phone, in the presented phonetic context.

Thirty-four expressive visemes were selected for each OCC emotion and also for the neutral expression speech resulting in 782 facial images: (22 OCC emotions + 1 neutral expression)× 34 visemes. The thirty-four visemes selected for each emotion correspond to representatives of the 22 consonantal CDVs shown in Table 3.1, plus the 11 vocalic CDVs of Table 3.2, plus a silence viseme.

The coordinates of 56 feature points were obtained using a semi-automatic software tool developed specially for the present work. The tool estimates the location of various facial feature points through the use of native detection routines that are are available in the OpenCV image processing software library (BRADSKI; KAEHLER, 2008). Following, each image is visually inspected and some points are marked manually to guarantee the accuracy of point identification. The feature points adopted in the current work characterize a subset of the MPEG-4 standard facial feature points (PANDZIC; FORCHHEIMER, 2002). They were chosen to delineate the facial and head elements like eyebrows, eyes, nose, lips, ears and chin, providing the information of "shape" of the facial image (Figure 3.7).

The *CH-Unicamp* database is composed of 782 facial images and a CSV (Comma Separated Values) spreadsheet file which carries information about each image in the database. The facial images have dimensions 1920 × 1080 pixels and they are saved as RGB uncompressed PNG files. The CSV spreadsheet file has the following structure:

- The file has no header.
- Each row carries the information of an specific image.
- Each column corresponds to an specific information, as follows:
  - First column: viseme image filename.

CSV File Column Number	**Feature Point Identifier**	**CSV File Column Number**	**Feature Point Identifier**
10	RM2	38	RN3
11	LM2	39	LN3
12	CUM	40	RN2
13	CLM	41	LN2
14	VMR	42	N2
15	VML	43	N4
16	RM1	44	HC
17	LM1	45	HR
18	RM3	46	HL
19	LM3	47	CJAW
20	REB1	48	RJAW2
21	REB2	49	LJAW2
22	REB3	50	RJAW1
23	LEB1	51	LJAW1
24	LEB2	52	REAR1
25	LEB3	53	REAR2
26	REYE1	54	REAR3
27	REYE3	55	REAR4
28	REYE2	56	LEAR1
29	REYE4	57	LEAR2
30	LEYE1	58	LEAR3
31	LEYE3	59	LEAR4
32	LEYE2	60	RM4
33	LEYE4	61	RM5
34	RN1	62	CIUM
35	LN1	63	CILM
36	N1	64	LM4
37	N3	65	LM5

Table 3.5 – Specification of the columns and the feature points they represent.

Feature Point Identifier Prefix	Description
RM	Right Mouth
LM	Left Mouth
REB	Right Eyebrow
LEB	Left Eyebrow
REYE	Right Eye
LEYE	Left Eye
RN	Right Nose
LN	Left Nose
N	Nose
HC	Head Center
HR	Head Right
HL	Head Left
CJAW	Center Jaw
RJAW	Right Jaw
LJAW	Left Jaw
REAR	Right Ear
LEAR	Left Ear
CUM	Center Upper at Mouth
CLM	Center Lower at Mouth
VMR	Right V of Mouth (Superior Lips)
VML	Left V of Mouth (Superior Lips)
CIUM	Center Internal Upper Mouth
CILM	Center Internal Lower Mouth

Table 3.6 – Description of the feature points prefixes.

– Second column: emotion associated to the image.

– Columns 3 to 7: phonetic context from which the viseme was extracted, phones are represented using Praat notation.

  * Column 3: second phone at left;
  * Column 4: first phone at left;
  * Column 5: phone represented by the image viseme (center phone, phone of interest);
  * Column 6: first phone at right;
  * Column 7: second phone at right.

– Column 8: Corresponding context-dependent viseme.

– Column 9: Word or phrase corresponding to the context where the image viseme was captured

– Columns 10 to 65: pair of x and y coordinates of feature points, as specified in Tables 3.5 and 3.6. Format used: (x;y).

Free copies of *CH-Unicamp* can be requested for research and private study only.

3.3 Concluding Remarks

The present chapter described the building of a comprehensive expressive speech corpus obtained under controlled conditions, which provides multimodal samples of expressive speech. The proposed methodology to build the corpus can be applied to any language.

In the present work, just a part of the expressive speech corpus was processed to create the *CH-Unicamp* expressive viseme images database. In the following chapters, the *CH-Unicamp* database characterizes the samples database of the expressive talking head synthesis methodology that is subject of the present work.

The processing and the annotation of the remaining corpus material is part of the future work and it characterizes a relevant contribution for future studies involving Brazilian Portuguese expressive speech. The captured material enables the study of the mechanisms of the production of speech accompanied by the expression of emotions through the access to:

- Shape variables: like the position and the relative distances of facial elements (such as the eyebrows, the lips and the eyes).
- Dynamic variables: provide the amplitude of excursion, acceleration profile and average speed of facial feature points.
- Appearance variables: enable the analysis of changes of color and brightness of the skin and the detection of wrinkles.

4 Expressive Speech Modeling

Among the existing approaches to model the human face, the present work focuses on an image-based, or 2D, approach. Image-based models are derived from a training set of real face images that are typically extracted from video of an audiovisual corpus.

Since the introduction of Eigenfaces by Turk and Pentland (TURK; PENTLAND, 1991), various statistical learning methods have been applied to model facial variations on images such as independent component analysis (ICA) and Kernel PCA (Principal Component Analysis) (ZHAO *et al.*, 2003). These algorithms were first applied for face recognition and they focus on modeling the appearance, or the texture, of the face, through the statistical analysis of the pixel values of sample images. For this reason, they present limited efficacy to model the shape variations caused, for instance, by speech articulation, the expression of emotions and pose variations. Cootes *et al.* (2001) proposed Active Appearance Model (AAM) as a generic object recognition technique which takes into consideration not only the appearance of objects but also their shapes. The linear deformable model provided by AAM makes them suitable for application on facial animation synthesis. In the works of Liu e Ostermann (2009) and Mattheyses *et al.* (2010) for example, AAMs were applied to improve the videorealism of 2D neutral speech animations systems that employ a concatenative synthesis strategy. In such systems, the shape and appearance parameters are included in the process of selecting the best sequence of units from a database of thousands of images which are concatenated to generate the final animation. Speech-driven or VTTS (visual text-to-speech) systems like those proposed by Cosker e Marshall (2004), Anderson *et al.* (2013) and Jiang *et al.* (2013), explore the correlation among speech, articulatory movements and the non-verbal signaling through the combination of visual parameters provided by the AAM with speech signal parameters, such as Mel-frequency cepstral coefficients and pitch. Among these works, only Anderson *et al.* (2013) address the speech accompanied by the expression of emotions.

The present chapter describes the process to build an expressive speech face model that is based on the AAM paradigm. Figure 4.1 provides an overview of the process. The expressive speech face model is derived from the statistical analysis of the variation of shape and appearance parameters extracted from the facial images of the *CH-Unicamp* database. Sections 4.1.1 to 4.1.5 detail how the shape and appearance vectors are built and the required processing steps to proceed with the statistical analysis of the data. Sections 4.1.6 and 4.1.7 describe the implemented PCA and how the shape and appearance models are obtained. Section 4.1.8 discusses how the appearance model can be applied to the synthesis of images

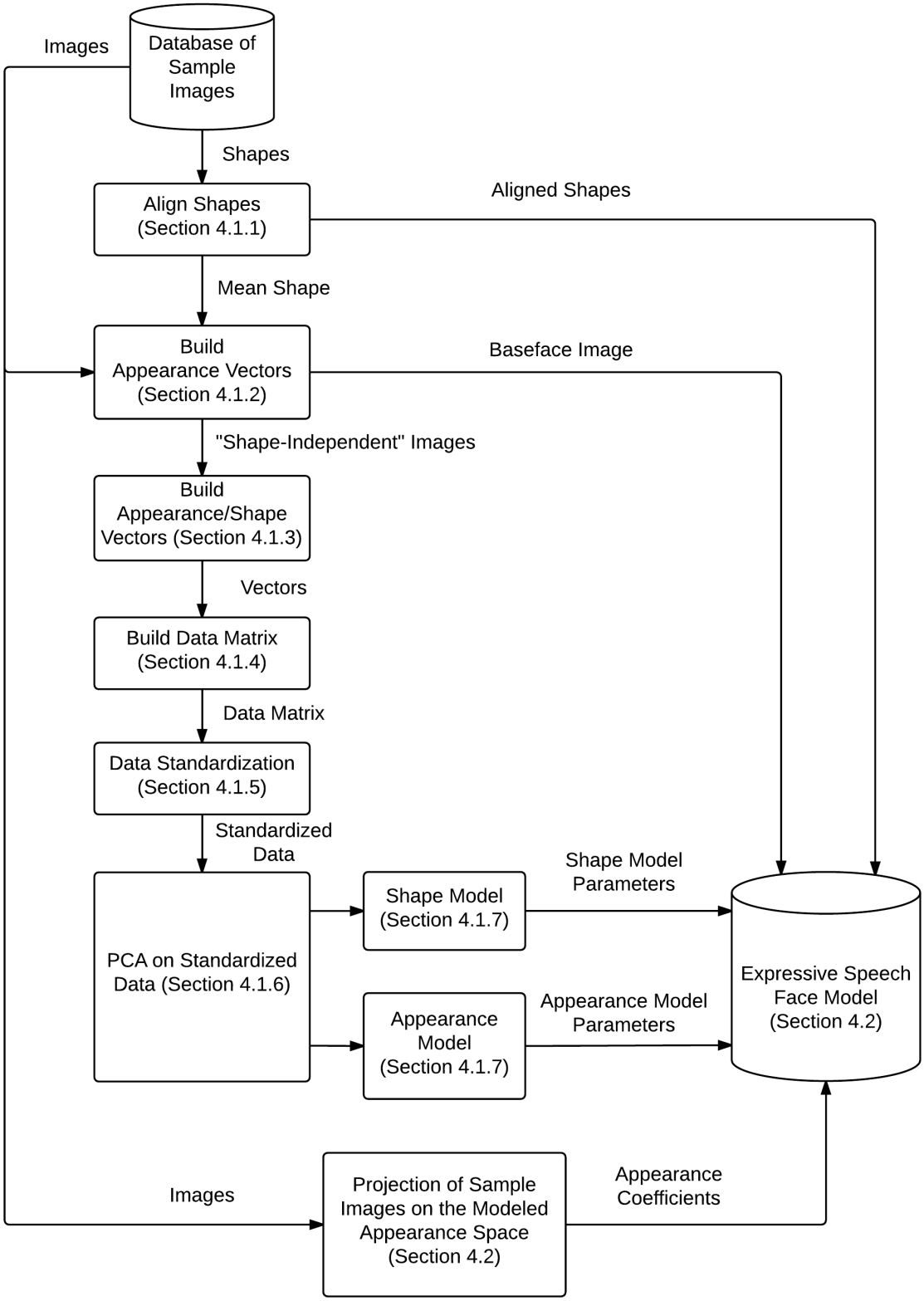


Figure 4.1 – Process to build the expressive speech face model

with facial expressions that were not originally present in the samples database. Section 4.2 describe the combination of data parameters that characterizes the expressive speech face model proposed in the present work. Section 4.3 presents the concluding remarks of the chapter.

4.1 Data Analysis

CH-Unicamp is a database of viseme image samples composed of:

- $m = 782$ RGB uncompressed images, $R \times C$ pixels ($R = 1080, C = 1920$);
- and the x and y-coordinates of $k = 56$ facial feature points associated to each image of the database.

The statistical analysis of the data is performed with the objective of obtaining shape and appearance models that are capable of expressing the diversity of facial configurations present in the database. The following sections describe the algorithm implemented to obtain such models.

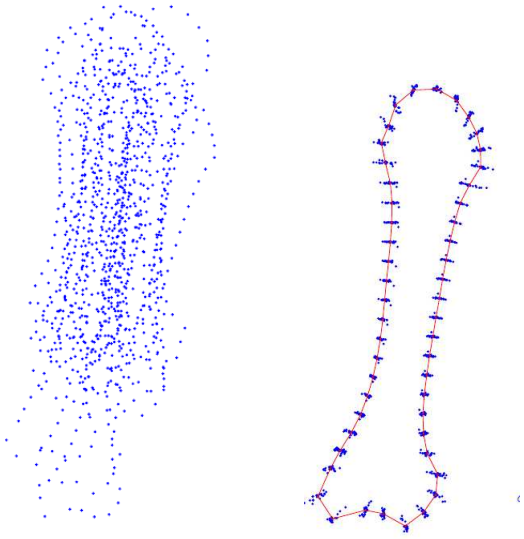
4.1.1 Shape Alignment

Following the concept described by Cootes *et al.* (2001), the "shape" of a facial image is a set of feature points which are typically defined to delineate the eyes, the eyebrows, the nose, the mouth and the outline of the face.

The shape vector $\mathbf{s}$ is defined concatenating the x and y-coordinates of the k facial feature points associated to each image in the training dataset, resulting in a column vector with $2k$ elements, as shown in Equation 4.1.

$$\mathbf{s} = \begin{bmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \\ \vdots \\ x_k \\ y_k \end{bmatrix} \tag{4.1}$$

The first step in the analysis process is to align the shapes of the database and to obtain a "mean shape" representation ($\bar{\mathbf{s}}$).



(a) (b)

Figure 4.2 – (a) A set of unaligned shapes of bones. (b) Mean shape and point distribution after alignment. Source: Image adapted from (STEGMANN, 2000).

Cootes *et al.* (1995) indicate that the alignment of the shapes is necessary to obtain a true representation of the feature point distribution. Stegmann (2000) illustrates this problem using shapes from hand metacarpals radiographs. As shown in Figure 4.2, the alignment process establishes a coordinate reference to which all shapes are aligned, filtering out the variations of location, rotation and scale.

The alignment of the m shapes of the database was performed following the iterative algorithm proposed by Cootes *et al.* (1995). The authors propose the alignment of a pair of shapes through the computation of the minimum squared weighted distance of two shapes.

Consider $\mathbf{s_i}$ and $\mathbf{s_j}$, two shape vectors that need to be aligned. Consider now that the shape $\mathbf{s_j}$ is rotated by θ, scaled by h, and translated by (tx, ty), giving origin to the transformed vector $\mathbf{s_{jtrans}}$. The quadratic distance between $\mathbf{s_i}$ and the transformed vector $\mathbf{s_{jtrans}}$ is given by Equation 4.2.

$$D = (\mathbf{s_i} - \mathbf{s_{jtrans}})^T(\mathbf{s_i} - \mathbf{s_{jtrans}}) \tag{4.2}$$

The alignment of $\mathbf{s_i}$ and $\mathbf{s_j}$ may be performed choosing the appropriate values of θ, h, and (tx, ty) that minimize the distance D.

In some cases, some feature points may be considered more "stable" than others as they are not easily moved by muscular actions. In facial images, for example, the points

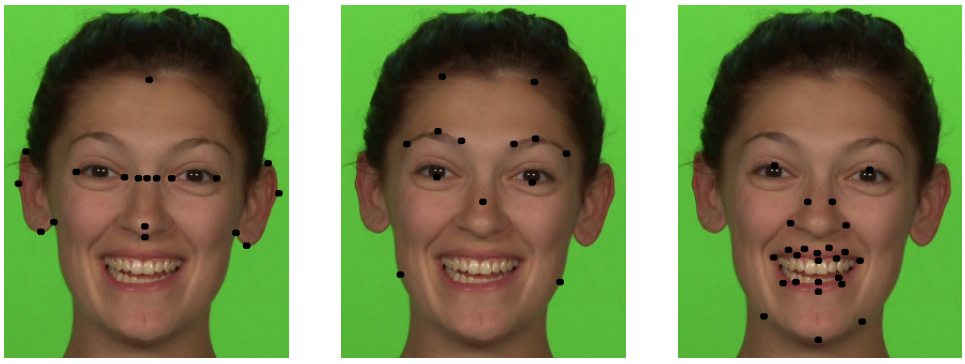
located in the corners of eyes or in the nose, are less subject to variations in relation to other points than the points around the lips. For this reason, Equation 4.2 is adapted to include a diagonal weighting matrix $\mathbf{W}$, as a mechanism to increase the significance of more stable points in the definition of the rotation, scale and translation parameters, as shown in Equation 4.3.

$$D_{weighted} = (\mathbf{s_i} - \mathbf{s_{jtrans}})^T \mathbf{W} (\mathbf{s_i} - \mathbf{s_{jtrans}}) \tag{4.3}$$

In the present implementation, the feature points were divided into three different sets associated to three weighting values, which were empirically defined:

- **Anchor points**: landmarks that have full weight value $w = 1$ during the alignment process. They are located in regions of the face that are not deformed by the speech articulatory movements or the expression of emotions, like the points in the ears, the tip of the nose, the corners of eyes and the top points that delineate the face (Figure 4.3(a)).
- **Intermediate weight points**: landmarks that are not heavily affected by face deformation during expressive speech, like the points in the forehead, in the eyebrows, around the nose and a pair of points in the jaw (Figure 4.3(b)). Their weight is $w = 0.5$.
- **Dynamic points**: landmarks that are heavily influenced by the speech dynamics and the expression of emotions. The points around the lips are in this category (Figure 4.3(c)). A low weight value ($w = 0.2$) is attributed to them to guarantee that they have limited influence in the shapes alignment.

The problem of minimizing $D_{weighted}$ can be solved through a set of equations (COOTES *et al.*, 1995).



(a) Anchor points (b) Intermediate weight points (c) Dynamic points

Figure 4.3 – Illustration of the feature points according to their classification as anchor points, intermediate weight points or dynamic points.

Adopting the above strategy to align a pair of shapes, the alignment of the m shapes in the database is obtained through the following iterative algorithm:

- Randomly select one of the shapes in the dataset to be the default normalizing shape (avoid outliers).
- Align each shape in the database to the default normalizing shape.
- **Repeat:**
  - Compute the mean shape from the aligned shapes.
  - Normalize the orientation, scale and origin of the current mean to the default normalizing shape.
  - Realign all the shapes with to the current mean.
- **Until the process converges.**

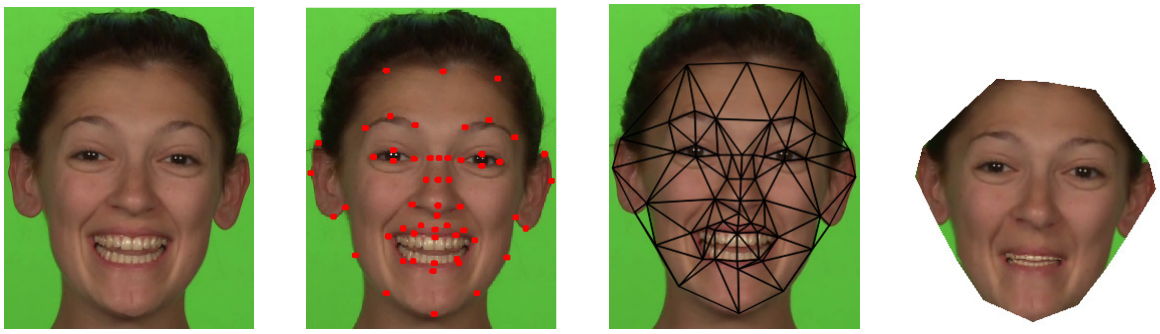
The mean normalization to a default scale and pose during each iteration is necessary to guarantee that the algorithm converges, avoiding that the mean shrinks, rotates or slide off to infinity. In the present work, the mean shape obtained after each iteration is compared to the previous iteration mean shape, before the normalizing steps. The sum of the differences of the shapes coordinates is used as criteria of convergence. Summed differences as low as 0.0001 result in the algorithm converging after 5 iterations. Alternative strategies for shape normalization and for convergence criteria are discussed by Cootes and colleagues (COOTES *et al.*, 1995).

The whole process results in a mean shape vector $\bar{\mathbf{s}}$ and a set of aligned shapes associated to each image of the database.

4.1.2 Appearance Vectors

The "appearance", also called by some authors the "texture" of an image region of interest (ROI), refers to the values of the pixels that lie inside the ROI.

Supposing a ROI containing q pixels, the appearance vector $\mathbf{a}$ is built concatenating their p_i, $i = 1...q$ pixel values. While the original implementation of AAM deals with grayscale images, in the present work the desired face model requires the processing of the color information provided by the Red (R), Green (G) and Blue (B) image planes. Thus,



(a) (b) (c) (d)

Figure 4.4 – (a) An image from the database. (b) Original shape information. (c) Delaunay triangulation. (d) Warped image to the mean shape $\bar{\mathbf{s}}$ and ROI used to generate the appearance vector.

each pixel is expressed as a triple of values: $\mathbf{p_i} = (p_{iR}, p_{iG}, p_{iB})$, $i = 1...q$. Therefore, the appearance vector, with $3q$ elements, has the structure shown in Equation 4.4.

$$\mathbf{a} = \begin{bmatrix} p_{1R} \\ p_{2R} \\ p_{3R} \\ \vdots \\ p_{qR} \\ p_{1G} \\ p_{2G} \\ p_{3G} \\ \vdots \\ p_{qG} \\ p_{1B} \\ p_{2B} \\ p_{3B} \\ \vdots \\ p_{qB} \end{bmatrix} \tag{4.4}$$

In order to perform the statistical analysis of the appearance distribution in the database, it is necessary to perform an "appearance alignment" procedure, filtering out the variation that is caused by shape variation. This procedure is performed warping all the database images to the mean shape $\bar{\mathbf{s}}$ (STEGMANN, 2000). The warping of images was implemented using a piecewise affine transformation (GLASBEY; MARDIA, 1998). In this

strategy, the Delaunay triangulation algorithm was applied to generate a mesh of triangles, having the shape feature points as the vertices of the triangles. In the present work, the Delaunay triangulation and piecewise affine warping transformations were implemented using native functions of the SciPy scientific toolbox for Python language (JONES *et al.*, 2001). Figure 4.4 illustrates intermediate results of the process. Figures 4.4(a) and 4.4(b) show an image from the database and the shape information associated to it, respectively. Figure 4.4(c) shows the triangulated mesh used as reference for the piecewise affine warping. Figure 4.4(d) shows the resulting image warped to the mean shape and also, the ROI that is considered to build the appearance vector (Equation 4.4).

As a result of the warping to the mean shape process, it is possible to compute the "shape-independent" appearance vectors of all images in the database.

The mean appearance vector obtained from the dataset analysis is defined by Equation 4.5.

$$\bar{\mathbf{a}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{a_i} \tag{4.5}$$

where $\mathbf{a_i}$ is the appearance vector of i-th image of the database and m the number of images in the database.

4.1.3 Building the "Appearance/Shape" Vectors

The AMM shape and appearance modeling strategy is based on the Principal Component Analysis (PCA) of the training dataset. PCA is a data analysis technique that identifies, in a multidimensional space, the directions in which the data presents the greater variance. PCA technique also provides information about orthogonal directions that characterize the data principal components, i.e. the most relevant projection axes to express the data variability. PCA is implemented performing the analysis of the statistics of a samples dataset, in particular the principal components are determined through the computation of the eigenvectors of the dataset covariance matrix. The underlying concepts and the many implementations and applications for PCA are explored in details by Jackson (2003).

The statistical analysis implemented in the present work differs from the implementation proposed by Cootes *et al.* (2001), which is described in details by Stegmann (2000). In the present work, the three-stage combined principal component analysis (PCA) proposed by Cootes *et al.* (2001) is substituted by a unique PCA step of the combined standardized shape and appearance data. The results of the comparative study described in Appendix B, show that this approach is efficient to explore the correlation that exists between the shape and appearance variables and it is simpler to implement, resulting in reduced processing time to compute the model.

Appendix B also presents results showing that, compared to the situation in which a unique full face model is used to reconstruct images, the piecewise modeling approach improves the quality of the reconstruction. The piecewise modeling is implemented building different models to different facial regions. In the present work, this approach is adopted to implement a coarse-to-fine synthesis strategy. The face is divided into three regions that can be overlaid: the full face; the cheeks+lips region; and the lip region (Figure 4.5). During synthesis, the full face is synthesized first; then the cheeks+lips synthesized texture is overlaid on the full face; finally, the lips are superimposed on the face. This approach guarantees that the principal components of the regions that suffer great variation in appearance during expressive speech, like the lips, carry the most relevant information about its variation modes. On the other hand, regions like the forehead, the eyebrows and the eyes, that present a lower number of variation modes, can be modeled taking advantage of the information provided the full face facial elements.

Figure 4.5 – Regions of the face that are modeled separately.

The combined "appearance/shape" vector $\mathbf{c}$ is constructed concatenating the appearance vector ($\mathbf{a}$) of a ROI with the shape vector ($\mathbf{s}$) of the same image, as shown in Equation 4.6.

$$\mathbf{c} = \begin{bmatrix} \mathbf{a} \\ \mathbf{s} \end{bmatrix} \tag{4.6}$$

When adopting the piecewise approach, a different model is computed for each ROI and the shape vector can be structured to reflect, or not, the regions segmentation; i.e. for a specific region, only the feature points that lie inside the ROI could be used to build $\mathbf{c}$. However, empirical tests showed that there is no observable difference between the use of the full face shape vector or its subset version. Since the shape vector is many times smaller than

the appearance vector of any region, it was adopted the full sized shape vector in all cases, without observable loss of performance.

Thus, the number of elements of $\mathbf{c}$ is $n = 3q + 2k$, where q is the number of RGB pixels that lie inside the ROI and k is the number of shape feature points, represented by their x and y-coordinates.

4.1.4 Building the Data Matrix

Considering the vector structure defined in Equation 4.6, the $\mathbf{c_i}$ "appearance/shape" vector of the i-th image of the database can be represented by Equation 4.7, where i is the index of the image in the database ($i = 1, 2, ..., m$) and j identifies the element of the "appearance/shape" vector ($j = 1, 2, ..., n$).

$$\mathbf{c_i} = [c_{i1}, c_{i2}, c_{i3}, ..., c_{ij}]^T \tag{4.7}$$

From Equation 4.7, we define the data matrix $\mathbf{C}$, in which the data samples are organized in the rows of the matrix and the variables are represented by the columns.

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix} = \begin{bmatrix} \mathbf{c_1^T} \\ \mathbf{c_2^T} \\ \vdots \\ \mathbf{c_m^T} \end{bmatrix} \tag{4.8}$$

Matrix $\mathbf{C}$ has m rows (number of samples) and n columns ($n = 3q + 2k$).

Taking into consideration the "appearance/shape" vectors of all the images from the dataset, the mean "appearance/shape" vector ($\bar{\mathbf{c}}$) is defined by Equation 4.9, where:

- $\bar{\mathbf{a}}$ is defined in Equation 4.5;
- $\bar{\mathbf{s}}$ is computed in Section 4.1.1;
- and the notation $\bar{c}_j$ ($j = 1, 2, ..., n$) is adopted to refer to an individual element of vector $\bar{\mathbf{c}}$.

$$\bar{\mathbf{c}} = \begin{bmatrix} \bar{\mathbf{a}} \\ \bar{\mathbf{s}} \end{bmatrix} = [\bar{c}_1, \bar{c}_2, ..., \bar{c}_n]^T \tag{4.9}$$

4.1.5 Data Standardization

The "appearance/shape" vectors ($\mathbf{c_i}$, $i = 1, 2, ..., m$) mix variables that are in different units and present different characteristics of variance. While the RGB pixel values assume values from 0 to 255, the range of values for x and y-coordinates of the shape vector elements depends on the dimensions of the original image. For example, on an image with dimensions 1920×1080 pixels, x may assume values between 0 and 1919 and y may assume values between 0 and 1079.

The problem is that PCA based on covariance matrix is sensitive to the units of measurement used for the variables. If there are large differences between their variances, those variables whose variances are largest will tend to dominate the first principal components (JOLLIFFE, 2002). The solution is to make the variance the same, through the use of standard units, in a process called data standardization.

Consider, for example, one of the columns of matrix $\mathbf{C}$, that represents a set of m observations of the j-th variable, expressed as the vector $\mathbf{t}$ represented by Equation 4.10.

$$\mathbf{t} = [t_1, t_2, ..., t_m]^T \tag{4.10}$$

To put $\mathbf{t}$ in standard units, Equation 4.11 is applied.

$$(t_i)_{std} = \frac{t_i - \bar{t}}{\sigma}, i = 1, 2, ..., m; \tag{4.11}$$

In Equation 4.11, $\bar{t}$ and σ are the mean and the standard deviation obtained from the samples, as defined by Equations 4.12 and 4.13, respectively:

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i; \tag{4.12}$$

$$\sigma = std(\{t_1, t_2, ..., t_n\}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})^2}. \tag{4.13}$$

From Equation 4.11, we obtain the vector $\mathbf{t_{std}} = [\mathbf{t_{1std}}, \mathbf{t_{2std}}, ..., \mathbf{t_{nstd}}]^T$ of standardized data, that has unit variance.

Considering the data matrix $\mathbf{C}$ (Equation 4.8), the standard deviation σ_j of a specific column j is computed according to the Equation 4.14.

$$\sigma_j = std(\{c_{1j}, c_{2j}, ..., c_{mj}\}) \tag{4.14}$$

From Equations 4.9 and 4.14, we define the standardized $m \times n$ data matrix $\mathbf{H}$ (Equation 4.16), with elements h_{ij} computed using Equation 4.15.

$$h_{ij} = \frac{c_{ij} - \bar{c}_j}{\sigma_j} \tag{4.15}$$

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ h_{m1} & h_{m2} & \cdots & h_{mn} \end{bmatrix} \tag{4.16}$$

4.1.6 PCA implementation

The PCA of the standardized data matrix $\mathbf{H}$ consists of computing the eigenvectors of its covariance matrix $Cov(\mathbf{H})$, which, in matrix notation, can be written as Equation 4.17.

$$Cov(\mathbf{H}) = \frac{1}{m}\mathbf{H^T H} \tag{4.17}$$

$Cov(\mathbf{H})$ has dimensions $n \times n$ ($n = 3q+2k$). Considering the typical sizes of the regions of interest, determining the eigenvectors of such a large matrix is a demanding computational task. Turk e Pentland (1991) calls the attention to the fact that since $m < n$ there will be only up to m meaningful eigenvectors (the remaining eigenvectors will have associated eigenvalues of zero). The authors show that the non-zero eigenvectors of $Cov(\mathbf{H})$, can be determined through the computation of the eigenvectors of the "minor product" $\mathbf{HH^T}$, resulting in the processing of a much smaller matrix, with dimensions $m \times m$.

The PCA analysis of $\mathbf{H}$ results in a set of m orthonormal vectors $\mathbf{h_i}$ and their associated eigenvalues λ_i ($i = 1, 2, ..., m$). A relevant property of PCA is that the eigenvalues λ_i indicate the proportion of the total data variability, or the explained variance ratio (EVR), that is accounted for by each direction of projection defined by the vectors $\mathbf{h_i}$. When the vectors $\mathbf{h_i}$ are considered in the descending order of their corresponding eigenvalues, they are named the principal components (PCs) of the data. In other words, when $\lambda_1 > \lambda_2 > ... > \lambda_m$, $\mathbf{h_1}$ is the first principal component of the data; $\mathbf{h_2}$ is the second principal component of the data; an so on.

The $\mathbf{h_i}$ vectors can be decomposed into appearance and shape principal components as shown in Equation 4.18.

$$\mathbf{h_i} = \begin{bmatrix} \mathbf{e_i} \\ \mathbf{f_i} \end{bmatrix} \tag{4.18}$$

where:

- $\mathbf{e_i}$ are vectors with $3q$ elements, representing the principal components of appearance;
- $\mathbf{f_i}$ are vectors with $2k$ elements, representing the principal components of shape.

4.1.7 Shape and Appearance Models

From the PCA analysis results, the appearance and shape diversity in the samples dataset is expressed by the appearance and shape linear models of Equations 4.19 and 4.20, respectively, where:

- $\bar{\mathbf{a}}$ and $\bar{\mathbf{s}}$ are the mean appearance and shape vectors, as defined in Equation 4.5 and Section 4.1.1, respectively;
- $\mathbf{D_a}$ ($3q \times 3q$) and $\mathbf{D_s}$ ($2k \times 2k$) are diagonal matrices, with standard deviation values σ_j (Equation 4.14) in the main diagonal, following the structure presented in Equations 4.21 and 4.22, respectively;
- α_i and β_i are the coefficients of the linear combination of appearance and shape principal components, respectively;
- $\mathbf{e_i}$ and $\mathbf{f_i}$ are the principal components of appearance and shape models, respectively.

$$\mathbf{a} = \bar{\mathbf{a}} + \mathbf{D_a} \sum_{1}^{m} \alpha_i \mathbf{e_i} \tag{4.19}$$

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{D_s} \sum_{1}^{m} \beta_i \mathbf{f_i} \tag{4.20}$$

$$\mathbf{D_a} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{3q} \end{bmatrix} \tag{4.21}$$

$$\mathbf{D_s} = \begin{bmatrix} \sigma_{3q+1} & 0 & \cdots & 0 \\ 0 & \sigma_{3q+2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} \tag{4.22}$$

4.1.8 Considerations About the Appearance Model

The vectors $\mathbf{e_i}$ of Equation 4.19 can be considered the vectorized version of prototype images or eigenfaces. Such names provide an intuition of what those vectors represent: images, that when properly combined, result in new facial images that were not originally present in the image database (Figure 4.6).

(a) Mean appearance. (b) Eigenface $\mathbf{e_1}$. (c) Eigenface $\mathbf{e_2}$. (d) Eigenface $\mathbf{e_3}$.

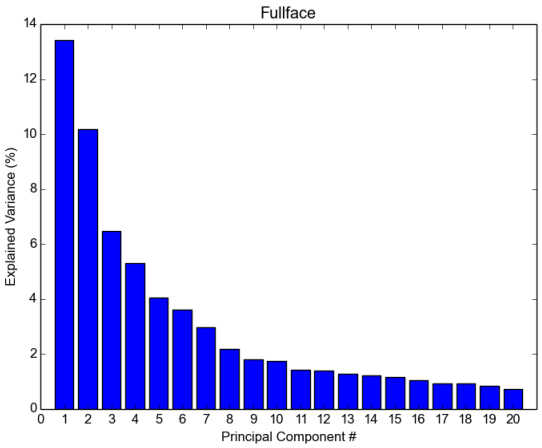
Figure 4.6 – Illustration of the mean appearance image ($\bar{\mathbf{a}}$) and the first three principal components or eigenfaces: $\mathbf{e_1}$, $\mathbf{e_2}$ and $\mathbf{e_3}$.

Equations 4.19 and 4.20 can be viewed as models capable of performing two tasks: synthesis or analysis. Given as input a set of weight parameters α_i and β_i, a unique result for the linear combination of eigenvectors $\mathbf{e_i}$ and $\mathbf{h_i}$ is provided and new facial images can be synthesized for the given "appearance/shape" configuration. In the synthesis mode, the output of the appearance model is a "shape-independent" image, i.e. an image warped to the mean shape. The final synthesized facial image is obtained warping the "shape-independent" image to the shape computed from Equation 4.20.

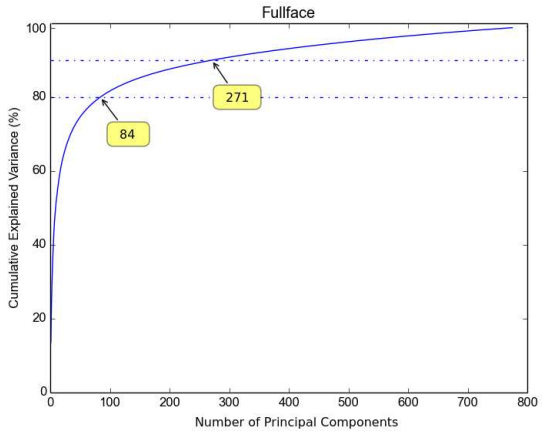
Conversely, the models can also be applied to perform analysis: given an input facial image and its corresponding feature points, it is possible to determine the α_i and β_i parameters that represent the position of the input image in the multidimensional spaces defined by the orthonormal vectors $\mathbf{e_i}$ and $\mathbf{h_i}$. The projection of a facial image onto the appearance space requires, first, its warping to the mean shape.

In the present work, the analysis mode is applied to project the entire *CH-Unicamp* database onto the modeled appearance space. In this manner, the expressive visemes are coded using a "shape-independent" representation expressed by the α_i coefficients resulting from the projection operation. The synthesis mode is applied to generate the final animation keyposes, as detailed in Chapter 5.

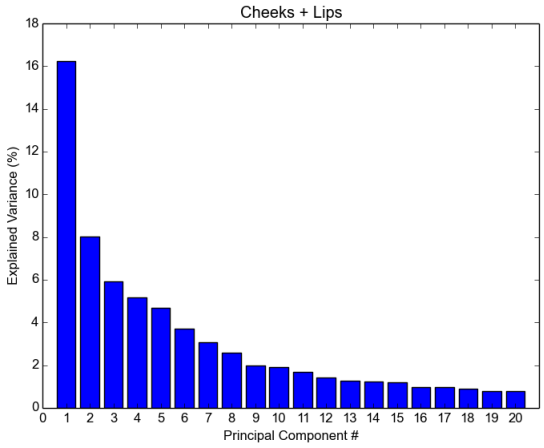
As discussed in Section 4.1.6, PCA makes possible to identify the principal components that are more relevant to express the data variability and to discard the less relevant components, in a process called dimensionality reduction. When no dimensionality reduction is performed, the number of prototype images resulting from the PCA analysis is equal to the number of samples (m) used to derive the model. Figure 4.7 shows the Explained Variance Ratio (EVR) profiles of the principal components of appearance models for the full face; cheeks+lips, and lip regions of Figure 4.5. The barplots of Figures 4.7(a), 4.7(c)



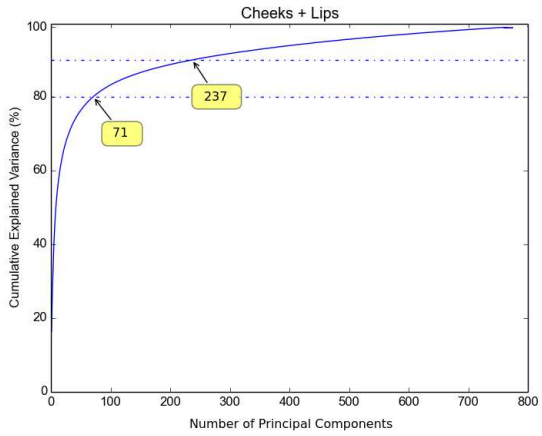
(a) Full Face EVR x PCs



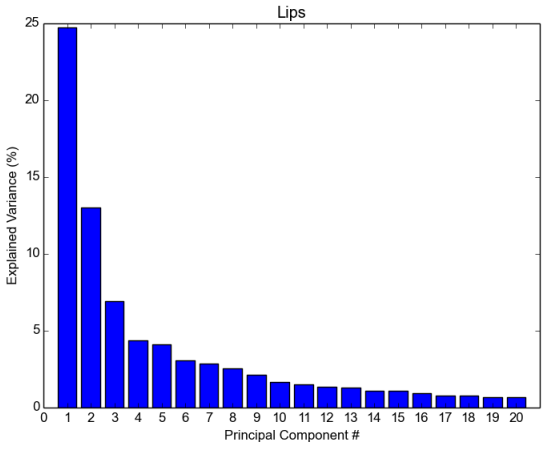
(b) Full Face Cumulative EVR



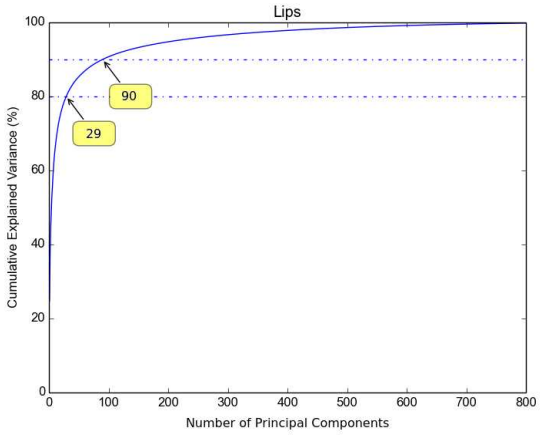
(c) Cheeks + Lips EVR x PCs



(d) Cheeks + Lips Cumulative EVR



(e) Lips EVR x PCs



(f) Lips Cumulative EVR

Figure 4.7 – The barplots on the left represent the variance ratio explained by the ranked principal components. The graphs on the right present the cumulative explained variance ratio considering the sorted principal components.

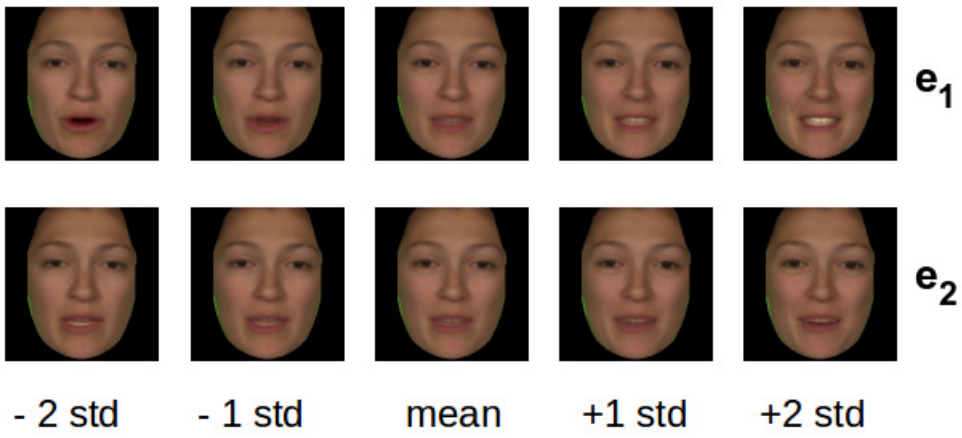


Figure 4.8 – First two modes of variation (principal components $\mathbf{h_1}$ and $\mathbf{h_2}$).

and 4.7(e) represent the variance ratio explained by each individual principal component. In Figure 4.7(e) for example, the barplot shows that the first ranked principal component is capable of representing almost 25% of the variance observed in the data distribution. The plots of Figures 4.7(b), 4.7(d) and 4.7(f) present the cumulative explained variance ratio considering the sorted principal components. Each plot highlights the number of principal components that is necessary to express 80% and 90% of the data variance. In Figure 4.7(f) for example, 80% of the data variance is explained by the first 29 components, while 90% of the data variance is explained by the first 90 principal components (in all cases, the number of principal components is n=782).

Figure 4.8 illustrates how the principal components express different modes of variation of the data. The first row of the figure illustrates the range of facial images that can be generated varying only the coefficient α_1 of the first principal component ($\mathbf{e_1}$) of the model represented in Equation 4.19 and discarding all other components. The extremes of variation represented in the figure correspond to two standard deviations (std) from the mean. The second row of the figure illustrates the second mode of variation ($\mathbf{e_2}$), when all other components were discarded.

4.2 Expressive Speech Face Model

A key aspect of the current work is the building of an expressive speech face model based on the "shape-independent" representation of the appearance of the *CH-Unicamp* database facial images. The underlying concepts of this model can be applied to any language or any set of expresssive visemes.

The expressive speech face model is characterized by a database that combines the following sets of data (see Figure 4.1):

- **Aligned Shapes**: the set of aligned shapes vectors associated to each image of the *CH-Unicamp* database as computed in Section 4.1.1.
- **Baseface Image**: the baseface image is obtained choosing an image from the aligned shapes and warping it to the mean shape $\bar{\mathbf{s}}$; in the present implementation, it was selected the image whose shape, after alignment, presents the minimum euclidean distance to the mean shape.
- **Shape Model Parameters**: the model parameters of Equation 4.20, including: the mean shape vector $\bar{\mathbf{s}}$, the diagonal elements of matrix $\mathbf{D_s}$ and the principal components $\mathbf{f_i}$.
- **Appearance Model Parameters**: the model parameters of Equation 4.19, including: the mean appearance vector $\bar{\mathbf{a}}$, the diagonal elements of matrix $\mathbf{D_a}$ and the principal components $\mathbf{a_i}$.
- **CDVs Appearance Coefficients**: a set of "appearance weights" (α_i in Equation 4.19), computed each region of interest, that represents the projection of all expressive context-dependent visemes images from the *CH-Unicamp* database onto the multidimensional space defined by the principal components of Equation 4.19.

The expressive speech face model can be configured to meet different requirements regarding the desired animation quality versus the size of the expressive speech face model database (Figure 4.1). The full quality configuration of the model is characterized when no dimensionality reduction is operated in the model and all the obtained principal components are kept. In this case, if m images were used to build the model, the number of prototype images in the expressive speech face model database is also m. However, an interesting aspect of this modeling approach is the possibility of coding the original images from the training database using a compact representation. The compression is performed choosing the first most relevant p principal components of the model. In this case, the images from the database are also projected on a reduced number of axes, resulting in a more compact representation of the face model.

Appendix C presents the results of a subjective perceptual evaluation conducted to investigate the impact of dimensionality reduction, in the perceived video image quality. The assessment was based on the presentation, side by side, of two synchronized videos: a full quality video reconstructed with all principal components and a lower quality version of

the same video, which was reconstructed with a reduced number of principal components. Forty-nine observers were asked to compare the lower quality video to its full quality version, and state if the differences they perceived could be considered: “imperceptible”, “perceptible but not annoying”, “slightly annoying” or “very annoying”. The results proved the feasibility of a compact face model representation through the reduction of the number of principal components used to synthesize facial images. For the fullface and cheeks+lips region, for example, the assessment results showed that approximately 80% of the participants classified as “imperceptible” or “perceptible but not annoying” videos reconstructed with less than half of the original number of principal components. On the other hand, the results also showed that the dimensionality reduction should be avoided in the lip region, since the observers presented increased sensitivity to detect lower quality reconstructions in this facial region.

4.3 Concluding Remarks

The present chapter described the building of the expressive speech face model that is part of the synthesis methodology described in the following chapter (Chapter 5). The face modeling is a one time offline analysis process of a set of context-dependent viseme sample images.

In the described implementation, the *CH-Unicamp* database was used both to derive the shape and appearance models; and additionally, as a source of CDVs samples to implement the expressive speech face model. This approach was adopted to guarantee that the training set had appearance and shape samples of all the expressive CDVs that are relevant to the synthesis process. This methodology was adopted to reject the statement that possible problems on the visual quality of the synthesized animations could be attributed to the lack of representation of specific facial configurations in the original database of samples.

However, this implementation strategy is not unique. An alternative implementation would be using a set of images to train the model that is different from the set of expressive context-dependent visemes recorded to build the expressive speech face model. Thus, additional image samples of CDVs can be added to improve the face model at any time or the dimensionality reduction can also be obtained with a smaller training dataset. Another strategy of implementation is to derive the shape and appearance models using sample images of a face, and use CDVs samples from a *different face* to build the expressive speech face model. The mapping between different “appearance” spaces is discussed in (THEOBALD *et al.*, 2009).

The chapter also refers to the perceptual evaluation results presented in Appendix C regarding the sensitivity of observers to the deterioration of video quality caused by the

dimensionality reduction of the appearance reconstruction model. The evaluation provided encouraging results to future implementations of more compact implementations of the expressive speech face model and it also provided guidelines regarding safeguarded levels of dimensionality reduction that could be implemented. Finally, the evaluation results also brought new evidences that the lip region is the most critical region in the model. Future improvements of the model should take this information into consideration.

5 Expressive Speech Animation Synthesis

This chapter presents a methodology to synthesize expressive talking heads based on the expressive speech face model described in Chapter 4. The diagram of Figure 5.1 provides an overview of the methodology, showing the main processing steps involved in the synthesis of expressive speech animation.

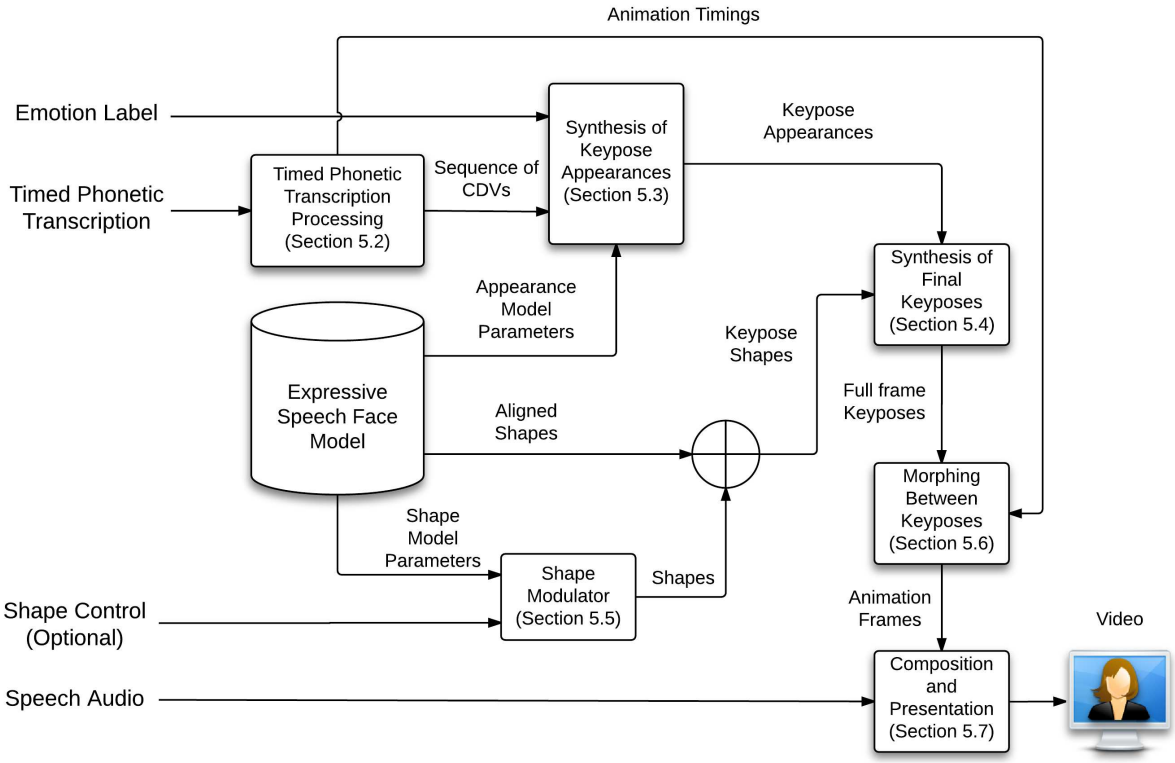


Figure 5.1 – Expressive Speech Animation Synthesis Process

The synthesis process requires four types of inputs (Figure 5.1):

- Emotion label: the indicator of the emotion to be expressed by the talking head. In the present work, the emotion label is a tag corresponding to one of the twenty-two emotions of the OCC model of emotions, such as: “joy”, “anger”, “hope”, etc..
- Speech audio: the audio file of the speech to be articulated by the synthetic talking head, which can be recorded or synthetic speech.[1]

[1] The synthesis methodology does not require expressive speech audio (neutral speech can also be animated)

- Timed phonetic transcription: an input text file informing the sequence of phones that compose the input speech audio, and their duration. The timed phonetic transcription can be obtained:
  - through manual segmentation with the aid of audio/speech analysis software applications such as Praat or Adobe™Audition;
  - through automatic segmentation provided, for example, by software libraries typically used in the development of speech recognition systems like HTK (Hidden Markov Model Toolkit);
  - as a byproduct of automatic speech synthesis.
- Shape control: optionally, the synthesis methodology may also process shape control input parameters that are used to implement a more sophisticated control over the head orientation and other facial elements, like the eyes and the eyebrows (Section 5.4).

Following the definition provided in Chapter 2, Section 2.3.1, the present methodology implements a rule-based keyframe animation synthesis strategy (Figure 5.2), in which the facial control parameters at particular frames of the animation are derived from the timed phonetic transcription of the speech to be animated and the emotion to be synthesized. The keyposes are synthesized based on the appearance model parameters of the expressive speech face model database, together with the default aligned shape information provided by the model or by an external "shape modulator" entity. The synthesis of the intermediate frames between adjacent keyposes is performed through a non-linear image morphing algorithm. The final step of the methodology consists of mixing together the animation frames and audio speech which results in the presentation of an expressive talking head video.

The following sections describe in details the methodology. Section 5.1 describes the timed phonetic transcription processing to obtain the sequence of context-dependent visemes that defines the keyposes of the animation and the temporal information that is necessary to define the animation frames timings. Section 5.2 describes the synthesis of the keyposes appearance information. The synthesis of final keyposes are described in Section 5.3. The shape modulator is discussed in Section 5.4. Section 5.5 presents the image morphing algorithm implemented. The composition of the final animation is discussed in Section 5.6. Finally, Section 5.7 presents the concluding remarks of the chapter.

and does not provide any mechanism to check if the speech audio is compatible with the emotion label input. As presented in Chapter 2, Section 2.3, some expressive talking heads systems adopt audio speech emotion classifiers to determine the emotion label from the automatic analysis of speech signal. This implementation is compatible with the methodology described in this chapter.

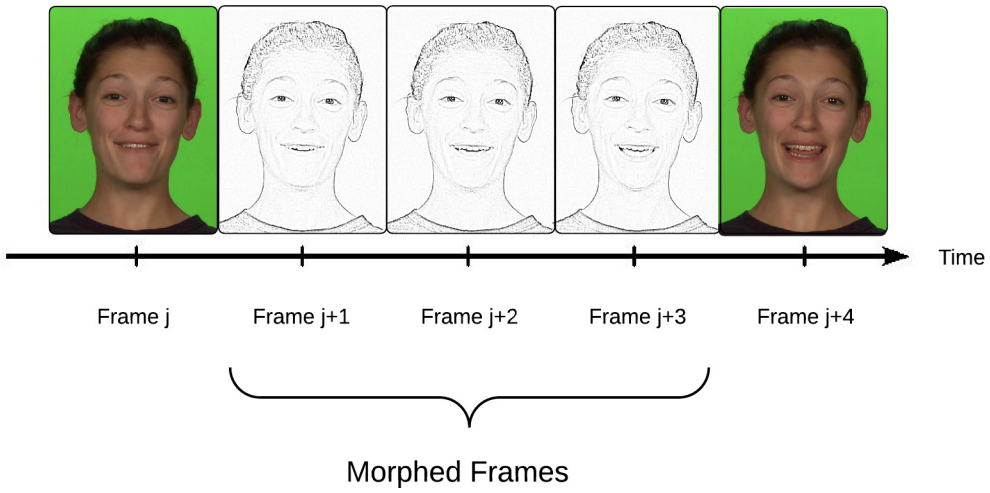


Figure 5.2 – Keyframe animation synthesis strategy. The keyposes are synthesized following a predefined set of rules. The intermediate frames are generated through image morphing algorithm. Source: Adapted from (COSTA, 2009).

5.1 Timed Phonetic Transcription Processing

The timed phonetic transcription of the speech audio provides the sequence of phones (speech sound segments) that are involved in speech production. The timed phonetic transcription can be expressed as a sequence of phonetic symbols or an alternative textual representation. It also provides the phones intervals and their frontiers in time. Figure 5.3 illustrates the timed phonetic transcription of a speech represented by the sequence of k phones, $F_i, i = 1, 2, ..., k$. The frontiers of each phone F_i are defined by the time instants t_{i-1} and t_i. The interval corresponding to the acoustic production of phone F_i is defined by the difference $(t_i - t_{i-1})$.

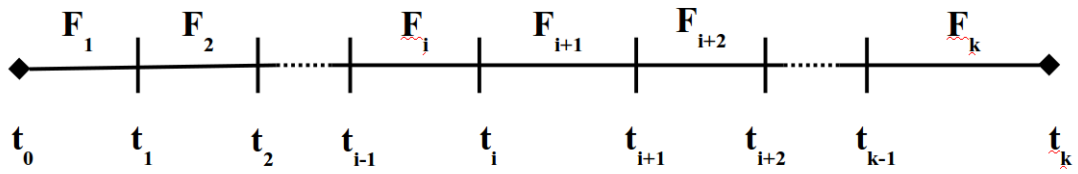


Figure 5.3 – Illustration of the information provided by the timed phonetic transcription of a speech audio represented by a sequence of k phones.

The following sections describe the processing performed to obtain the outputs shown in Figure 5.1: the animation timings (Section 5.1.1) and the sequence of CDVs (Section 5.1.2).

5.1.1 Extracting Animation Timings

From the temporal information provided by the timed phonetic transcription, it is possible to define the duration of the animation, the number of frames necessary to animate the speech and the time instants associated to the keyposes. The keyposes, or key-visemes, are chosen to represent points of inflection in the trajectory of the visible speech articulators — basically, the lips, the teeth and the tongue. In other words, only one key-viseme is associated to each phone interval and it represents the final excursion of the articulators trajectory to configure the visual representation of a phone. The frame immediately subsequent to a key-viseme defines the start of a new trajectory towards the next adjacent keypose (Figure 5.2).

When analyzing real speech, the dynamics of speech articulators is complex and, as discussed in Chapter 2, the effects of coarticulation among neighboring speech segments affects considerably the typical articulatory patterns of phones. For this reason, the articulatory targets, or the points of inflection between speech segments, may occur any time during a phone interval. As a simplification of this dynamic behavior, the adopted strategy associates the key-visemes to the time instants that correspond to the half period of a phone interval.

Considering a phone F_i, the i-th key-viseme is associated to the time instant:

$$K_i = t_{i-1} + \frac{t_i - t_{i-1}}{2} = \frac{t_{i-1} + t_i}{2} \tag{5.1}$$

5.1.2 Conversion from Phones to Context-Dependent Visemes

In the present work, the keyposes of animation correspond to synthesized images of expressive context-dependent visemes (CDVs). The synthesis of such keyposes depends on the conversion of the sequence of phones provided by the timed phonetic transcription to a sequence of Brazilian Portuguese CDVs, presented in Chapter 3, Section 3.1.2, and again in Tables 5.1 and 5.2.

The conversion from phones to Brazilian Portuguese CDVs, involves:

- **Conversion of the vocalic phones:** Table 5.2 shows that most vocalic visemes are independent of phonetic context, with the only exception being the phone [i]. The vocalic phones [e],[ɛ],[a],[ɔ],[o],[u],[ɪ],[ɐ] and [ʊ] present in the phone sequence are converted to the corresponding visemes shown in the second column of Table 5.2. If a vocalic phone [i] is encountered in the sequence, the triphone in which it is the central phone is analyzed, i.e the first phone at left of [i] and the first phone at right are considered in the analisys. If the phonetic contexts [tit] or [ʃiʃ] are identified, the [i] phone is substituted

[1] Phonetic context not valid for [ɾ] phone. The "tap" is never observed in the beggining of words in Brazilian Portuguese.

Homorganic Groups	Context-Dependent Visemes	Phonetic Contexts
[p,b,m]	$< p_1 >$	[pi] [pa] [ipɪ] [ipɐ] [ipʊ] [apɪ] [apɐ] [apʊ] [upɐ]
	$< p_2 >$	[pu] [upɪ] [upʊ]
[f,v]	$< f_1 >$	[fi] [fa] [ifɪ] [ifɐ] [ifʊ] [afɪ] [afɐ]
	$< f_2 >$	[fu] [afʊ] [ufɪ] [ufɐ] [ufʊ]
[t,d,n]	$< t_1 >$	[ti] [tu] [itɪ] [itɐ] [itʊ] [atɪ] [atʊ] [utɪ] [utɐ] [utʊ]
	$< t_2 >$	[ta] [atɐ]
[s,z]	$< s_1 >$	[si] [sa] [isɪ] [isɐ] [asɪ] [asɐ]
	$< s_2 >$	[su] [isʊ] [asʊ] [usɪ] [usɐ] [usʊ]
[l]	$< l_1 >$	[li] [ilɪ] [alʊ] [ulɪ] [ulɐ]
	$< l_2 >$	[la] [ilɐ] [alɪ] [alɐ]
	$< l_3 >$	[lu]
	$< l_4 >$	[ilʊ] [ulʊ]
[ʃ,ʒ]	$< ʃ_1 >$	[ʃi] [ʃa] [iʃɪ] [iʃɐ] [iʃʊ] [aʃɪ] [aʃɐ] [aʃʊ] [uʃɪ] [uʃɐ]
	$< ʃ_2 >$	[ʃu] [uʃʊ]
[ʎ,ɲ]	$< ʎ_1 >$	[ʎi] [ʎa] [iʎɪ] [iʎɐ] [aʎɪ] [aʎɐ]
	$< ʎ_2 >$	[ʎu] [uʎɪ] [uʎɐ]
	$< ʎ_3 >$	[iʎʊ] [aʎʊ] [uʎʊ]
[k,g]	$< k_1 >$	[ki] [ikɪ] [ikɐ] [akɪ] [ukɪ] [ukɐ]
	$< k_2 >$	[ka] [akɐ]
	$< k_3 >$	[ku] [ikʊ] [akʊ] [ukʊ]
[ɣ],[ɾ]	$< ɣ_1 >$	[ɣi][1] [ɣa][1] [iɣɪ] [iɣɐ] [aɣɪ] [aɣɐ] [uɣɐ]
	$< ɣ_2 >$	[ɣʊ][1] [iɣʊ] [aɣʊ] [uɣɪ] [uɣʊ]

Table 5.1 – Consonantal context-dependent visemes (adapted from (DE MARTINO *et al.*, 2006)). The phonetic symbols are from the International Phonetic Alphabet (INTERNATIONAL PHONETIC ASSOCIATION, 1999).

by the $<i_2>$ viseme. For all other phonetic contexts involving [i], it is substituted by the $<i_1>$ viseme.

- **Conversion of the consonantal phones:** based on the linguistic analysis performed by De Martino et al. (2006), the third column of Table 5.1 lists the phonetic contexts that discriminates the Brazilian Portuguese visemes. Thus, the conversion of consonantal phones to CDVs is performed mapping the phonetic contexts in which the consonantal phones occur to the phonetic contexts of Table 5.1. The present work adopts

Homorganic Groups	Context-Dependent Visemes	Phonetic Contexts
[i,ĩ]	$< i_1 >$	All contexts except [tit] and [ʃiʃ].
	$< i_2 >$	[tit] and [ʃiʃ].
[e,ẽ]	< e >	All contexts.
[ɛ]	< ɛ>	All contexts.
[a,ɐ̃]	< a >	All contexts.
[ɔ]	< ɔ>	All contexts.
[o,õ]	$< o >$	All contexts.
[u,ũ]	$< u >$	All contexts.
[ɪ]	< ɪ>	All contexts.
[ɐ]	< ɐ>	All contexts.
[ʊ]	< ʊ>	All contexts.

Table 5.2 – Vocalic context-dependent visemes (adapted from (DE MARTINO *et al.*, 2006)). The phonetic symbols are from the International Phonetic Alphabet (INTERNATIONAL PHONETIC ASSOCIATION, 1999).

the mapping strategy proposed by Costa (2009), that implements a pentaphone (a sequence of five phones) context analysis. In other words, the two phones at left and the two phones at right of the consonantal phone to be converted are analyzed. The mapping implementation consists of a table that correlates the possible Brazilian Portuguese pentaphones structures to the phonetic contexts listed in the third column of of Table 5.1. The mapping table was derived from the analysis of the linguistic characteristics of Brazilian Portuguese. The mapping table and the underlying concepts of the linguistic analysis are detailed in (COSTA, 2009).

5.1.2.1 Example of Conversion from Phones to Context-Dependent Visemes

In this section, it is presented an example of phones to context-dependent conversion.

Consider the utterance of the Brazilian Portuguese sentence: *"Que notícia excelente!"* (What wonderful news!).

The phonetic transcription of the utterance results in the following sequence of phones:

k i n o t i s e l ẽ t ɪ

The following alternative notation is adopted to discriminate repeated phones:

$\mathbf{k}_1$ $\mathbf{i}_1$ $\mathbf{n}_1$ $\mathbf{o}_1$ $\mathbf{t}_1$ $\mathbf{i}_2$ $\mathbf{s}_1$ $\mathbf{e}_1$ $\mathbf{l}_1$ $\mathbf{\tilde{e}}_1$ $\mathbf{t}_2$ $\mathbf{ɪ}_1$

The second column of Table 5.3 presents the result of the analysis performed to translate the sequence of phones above to Brazilian Portuguese CDVs. The consonantal phones are mapped to the phonetic contexts of Table 5.1 following the linguistic analysis detailed in (COSTA, 2009).

Phone	Context-Dependent Viseme	Justificative
k_1	$<k_1>$	Mapped to context [ki] (Table 5.1).
i_1	$< i_1 >$	Not in context [tit] or [ʃiʃ] (Table 5.2).
n_1	$< t_1 >$	Mapped to context [itʊ] (Table 5.1).
o_1	$< o >$	All contexts (Table 5.2).
t_1	$< t_1 >$	Mapped to context [utɪ] (Table 5.1).
i_2	$< i_1 >$	Not in context [tit] or [ʃiʃ] (Table 5.2).
s_1	$< s_1 >$	Mapped to context [isɐ] (Table 5.1).
e_1	$<$ e $>$	All contexts (Table 5.2).
l_1	$< l_2 >$	Mapped to context [alɐ] (Table 5.1).
$\tilde{e}_1$	$<$ e $>$	All contexts (Table 5.2).
t_2	$< t_1 >$	Mapped to context [atɪ] (Table 5.1).
$ɪ_1$	$<$ ɪ$>$	All contexts (Table 5.2).

Table 5.3 – Example of conversion of phones to context-dependent visemes.

5.2 Synthesis of Keypose Appearances

The synthesis of the animation keypose appearance follows the coarse-to-fine approach described in Chapter 4, Section 4.1.3:

1. the full face appearance is synthesized (Figure 5.4(a));
2. the appearance of the cheeks+lips ROI (region of interest), shown in Figure 5.4(b), is synthesized and overlaid to the full face appearance obtained in the previous step;
3. the lips ROI (Figure 5.4(c)) is synthesized and added to the full face appearance;
4. the resulting appearance is stitched to a baseface, shown in Figure 5.4(d).

Figure 5.4(e) shows the final result of the synthesis of a keypose appearance. As described in Chapter 4, Section 4.2, the baseface is a full sized video frame warped to the mean shape. Figure 5.4(d) shows the region into which the synthesized keypose appearance is stitched to the video frame. The composition of two images I_1 and I_2 was implemented following an alpha blending mask operation: $I_1(\alpha) + I_2(1-\alpha)$, with α ranging gradually from 0 to 1 in the borders of the regions to be combined. As an example, Figure 5.4(f) shows

the alpha blending mask for the cheeks+lips region. The masks for each region of interest were automatically computed through the definition of interpolation curves that connect the feature points that delimit each region.

The appearance model parameters, for each keypose and each ROI, are retrieved from the expressive speech face model database given the pair of indexing keys: emotion label and context-dependent viseme.

The appearance model parameters consists of (Chapter 4, Section 4.1.7):

- a set of l of principal components or prototype images $\mathbf{e_i}$, with $i = 1, 2, ..., l$;
- the original data standard deviation information $\mathbf{D_a}$;
- a set of l, weight coefficients α_i;
- the mean appearance vector $\bar{\mathbf{a}}$.

From these definitions, the ROI's keypose appearance $\hat{\mathbf{a}}$, is computed through the equation:

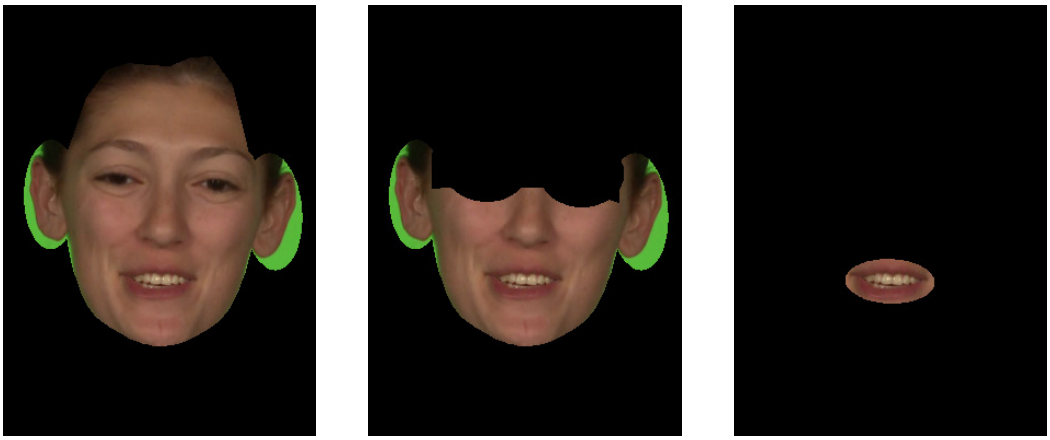
$$\hat{\mathbf{a}} = \bar{\mathbf{a}} + \mathbf{D_a} \sum_{i=1}^{l} \alpha_i \mathbf{e_i} \tag{5.2}$$

The result of the synthesis of keyposes appearance is a sequence of full sized video frames, corresponding to expressive context-dependent visemes in the "shape-independent" form, i.e. with the frame warped to the mean shape.

5.3 Synthesis of Final Keyposes

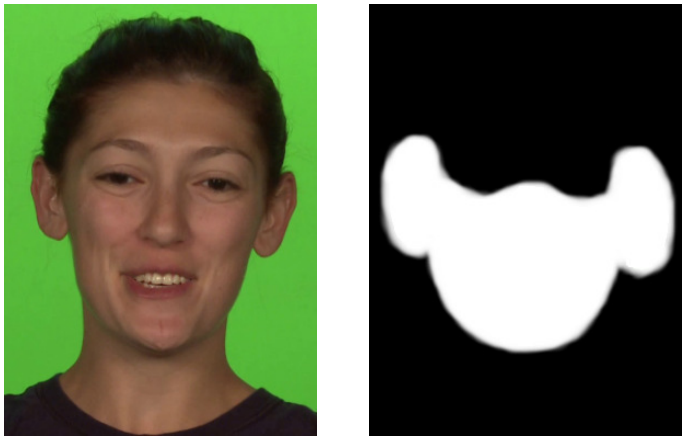
The synthesis of keypose shape process depicted in Figure 5.1 consists of warping the "shape-independent" appearance keyposes obtained in Section 5.2 to the final shapes that the keyposes should assume (more details about the adopted warping algorithm are provided in Section 5.5).

The process receives as input a sequence of shape vectors, $\mathbf{s_t}$ associated to the occurrences of the keyposes. In the case when no external shape control is performed, the final synthesized keypose shapes are the aligned shapes associated to each expressive context-dependent viseme in the expressive speech face model database, resulting from the shape alignment process described in Chapter 4, Section 4.1.1. In this synthesis mode, the expressive visemes are presented by the talking head accordingly with the specified emotion label, however, the final animation is characterized by the lack of head movements, the blinking of



(a) (b) (c)

(d)



(e) (f)

Figure 5.4 – (a) Full face; (b) Cheeks+lips ROI; (c) Lips ROI; (d) Baseface; (e) Final synthesized keypose appearance; (f) Alpha blending mask for the cheeks+lips region.

the eyes and the raising or the frowning of the eyebrows accompanying the expressive speech. In other words, there is no programmed non-verbal signaling outside the lips region.

5.4 Shape Modulator

Optionally, shape control parameters may be provided to a shape modulator module which processes the external input to derive the head orientation or to command specific actions of facial elements like the blinking of the eyes and the raising or frowning of the eyebrows. In the present work, the shape modulator is not implemented in its fullest potential. As suggested in Figure 5.1, the shape modulator may use the shape model available in the the expressive speech face model, to synthesize new keypose shapes. In this case, the external shape control inputs may be the s_i coefficients of the shape model and new shapes can be obtained as a result of the linear combination of l principal components, as expressed by Equation 5.3 (Chapter 4, Section 4.1.7).

$$\hat{\mathbf{s}} = \bar{\mathbf{s}} + \mathbf{D_s} \sum_{i=1}^{l} \beta_i \mathbf{f_i} \tag{5.3}$$

5.5 Morphing Between Keyposes

The input to the process that synthesizes the final animation frames is the sequence of synthesized keyposes (warped to the final shapes), and their corresponding timings (Section 5.1, Equation 5.1). From the speech audio duration, it is obtained the timings of the animation frames. The animation rate of 30 frames per second (fps) is adopted and the first animation frame is associated to $t = 0$.

The synthesis process consists of generating the animation frames between two adjacent keyposes through the image morphing algorithm. Given two keyposes K_{source} and K_{target}, the morphing algorithm determines the functions that warp image K_{source} to K_{target} in a forward direction; and the function that warps K_{target} to K_{source} in the backward direction (Figure 5.5). The transformation in each direction is divided into intermediate steps, corresponding to the frames between two adjacent key-poses. The results of the transformations in both directions are cross-dissolved, following a proportion that is a function of time (WOLBERG, 1998). Considering a normalized interval between two adjacent keyposes, the cross-dissolving operation can be stated as:

$$F = K_f(t)(1-t) + K_b(t)(t) \tag{5.4}$$

Where:

- F is the final synthesized animation frame;
- $K_f(t)$ is the warped image in the forward direction at t;
- $K_b(t)$ is the warped image in the backward direction at t;
- t is the normalized time variable for any interval between two adjacent keyposes, $0 \leq t \leq 1$.

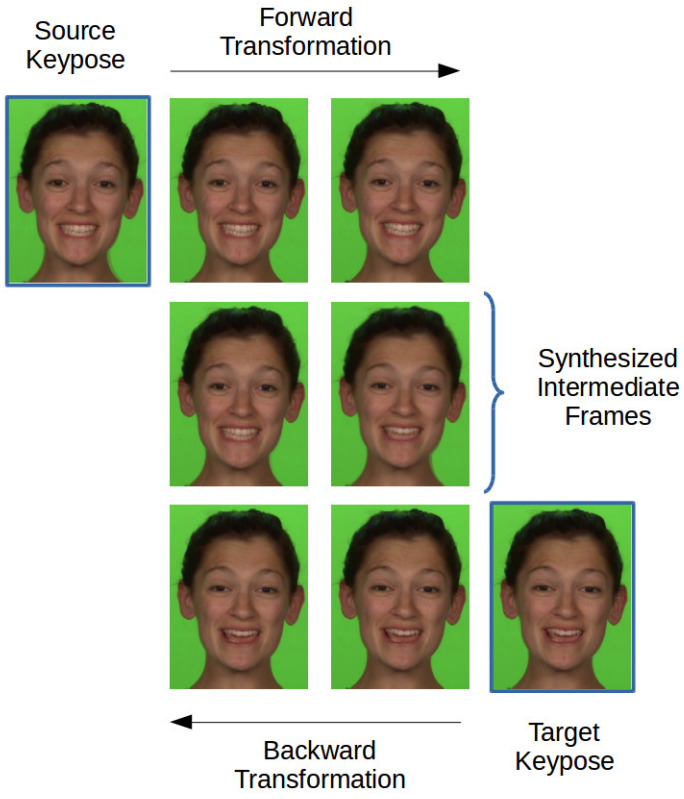
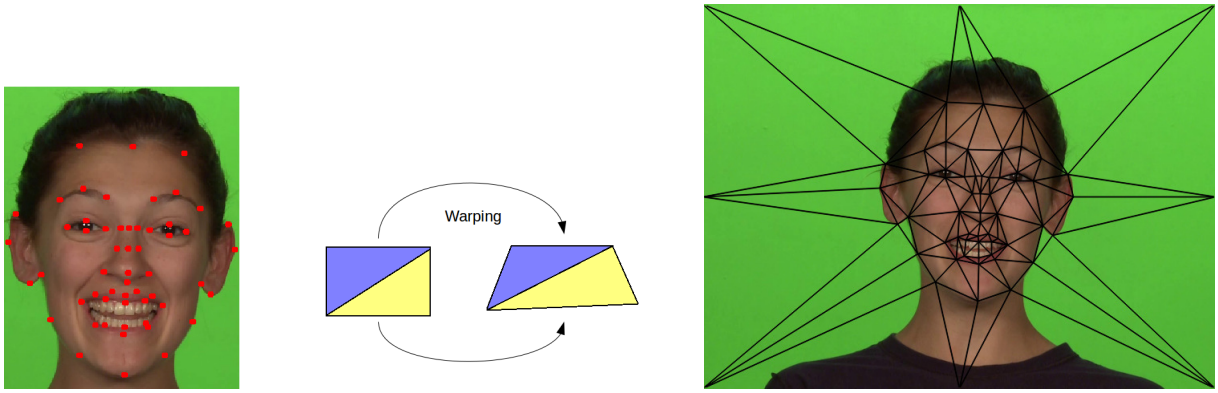


Figure 5.5 – Illustration of the morphing from a source to a target keypose. The first row of images illustrates the forward direction of transformation from the source image towards the target. The third row of images illustrates the backward direction of transformation, from the target to the source. The row in the middle shows the final synthesized frames resulting from the cross-dissolve operation between images from the forward and backward directions of transformation. Source: Adapted from (COSTA, 2009).

Warping is guided by the definition of a correspondence map of feature points between the source and target images (the feature points are shown in Figure 5.6(a)). From the correspondence map it is possible to compute a warping function that defines the spatial relationship between all points in both images. In the present work, the warping of images



(a) (b) (c)

Figure 5.6 – (a) Feature points considered for building the correspondence map; (b) Illustration of the piecewise warping approach, based on a mesh of triangles; (c) Triangulation used as reference for the warping.

was implemented following a piecewise affine transformation strategy (GLASBEY; MARDIA, 1998). In this strategy, the Delaunay triangulation algorithm is used to generate a mesh of triangles, having the shape feature points as the vertices of the triangles (Figure 5.6(c) shows that some points in the borders of the full frames to stabilize the image). The same triangulation is applied both for the source shape and the target shape. Following, for each triangle of the mesh, an affine transformation is computed to map the pixels from the source to the target mesh of triangles (Figure 5.6(b)). In the present work, the piecewise affine warping and the Delaunay triangulation were implemented using native functions of the SciPy scientific toolbox for Python language (JONES *et al.*, 2001).

The trajectory defined by the feature points during morphing greatly affects the visual perception of the speech articulatory movements. De Martino et al. (2006) proposed the feature points follow a non-linear smooth interpolation curve in time as a mechanism to model the complex and variable dynamic of transitions between articulatory targets observed in real speech.

The trajectory of the points during the successive warping operations, shown in Figure 5.5, is then modeled through an Hermite parametric interpolation curve (FOLEY, 1990), which provides G^0 continuity between successive frames of a sequence and ensures derivative equal to zero at the time instants associated to the keyposes.

For each x and y-coordinates of a specific feature point, the interpolation curve is

obtained through the solution of the following set of equations:

$$\begin{bmatrix} x_i(t) \\ y_i(t) \end{bmatrix} = \begin{bmatrix} Sx_i & Tx_i \\ Sy_i & Ty_i \end{bmatrix} \begin{bmatrix} 2 & -3 & 0 & 1 \\ -2 & 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix}, 0 \leq t \leq 1$$

Where:

- $x_i(t)$ and $y_i(t)$ are the Hermite functions for the x and y-coordinates of the i-th feature point, where $i = 1, .., k$ (Chapter 4, Section 4.1.1).
- Sx_i e Sy_i are x and y-coordinates, respectively, of the corresponding feature point in the source keypose image;
- Tx_i e Ty_i are x and y-coordinates, respectively, of the corresponding feature point in the target keypose image;
- t is the independent parametric variable normalized in relation to the time interval between the two keyposes.

5.6 Composition and Presentation

The morphing between keyposes is the last step in the synthesis of the facial animation frames. Following, the synthesized animation frames are mixed with the speech audio. In the present implementation, this operation was performed using primitive functions of OpenCV image processing library in association with the FFmpeg multimedia software.

5.7 Concluding Remarks

In the previous sections, the expressive speech animation synthesis methodology presented in Figure 5.1 was detailed. Together with the expressive speech face model presented in Chapter 4, the proposed methodology enables the synthesis of a wide range of facial expressions associated to a robust modeling of the speech articulation ensured not only by the implicit modeling of coarticulation provided by the context-dependent visemes but also by the non-linear image morphing between keyposes.

A key aspect of the presented methodology is the “shape-independent” synthesis of the expressive context-dependent visemes, described in Section 5.2. This “shape-independent” representation of the visual phonemes is capable of carrying relevant information about the

speech articulators, such as the visible presence or not of the teeth and the tongue during the production of the visemes; the thinning of the lips; the so called "bright in the eyes" and other subtle changes in the skin texture in the cheeks or in the forehead, for example. Additionally, and most important in the context of the present work, this representation is capable of expressing the variation of such appearance elements for different emotional expressions. Together with the shape modulation of the keyposes, that is not fully implemented in the present work, the proposed synthesis methodology characterizes a flexible approach with the potential of synthesizing expressive videorealistic talking heads.

As a "proof of concept" of the envisioned implementation involving a "shape modulator", the following chapter (Chapter 6) presents the results of a subjective perceptual of emotions in which "copies" of shapes of a real video frames are provided to modulate the shape of the keyposes.

6 Emotion Recognition Evaluation

Assessing the visual quality of facial animations is a key aspect to compare and evaluate different synthesis approaches. However, there is no universally accepted criteria to measure the visual quality of talking heads or their level of videorealism. In the present work, videorealism is understood as the capability of a facial animation being confused with the video of a real face.

In the literature, it is possible to identify five main approaches to evaluate talking heads:

- **Subjective tests:** human observers express their opinions about aspects such as "synchronicity with speech", "smoothness" and "precision" of animations (BREGLER *et al.*, 1997), (COSATTO; GRAF, 2000), (ALBRECHT *et al.*, 2005), (BEVACQUA *et al.*, 2007), (JIA *et al.*, 2011), (LIU; OSTERMANN, 2012), (ANDERSON *et al.*, 2013).

- **Visual "Turing tests":** inspired by the test envisioned by Turing (1950), the visual version of the so called "Turing test" consists of a subjective test where people are asked to distinguish synthesized animations from recorded videos of a real face (BRAND, 1999), (EZZAT *et al.*, 2002).

- **Ground-truth trajectory comparisons:** objective measurements are obtained comparing the trajectories of facial feature points in the facial animation to reference trajectories that are measured on recorded video (BESKOW, 2004; DENG *et al.*, 2006; TAO *et al.*, 2009; LIU *et al.*, 2011; LIU; OSTERMANN, 2012).

- **Speech intelligibility tests:** essentially involve the presentation of unimodal auditory and bimodal audiovisual contents (syllables, words or sentences) under different conditions of acoustic degradation (SUMBY; POLLACK, 1954). Participants are then asked to recognize the utterances in the test. The speech intelligibility results obtained by animated faces are compared to the results obtained by the recorded video of a real face. Speech intelligibility tests are frequently applied to evaluate and compare visual speech coarticulation models (OUNI *et al.*, 2007), (COSTA; DE MARTINO, 2013).

- **Emotion recognition tests:** in order to evaluate expressive talking heads, participants are asked to choose, among a vocabulary of emotions, the emotion expressed by a talking face. The results of the synthetic expressive facial animation are compared to

the results obtained by the recorded video of a real face (BESKOW; NORDENBERG, 2005; ANDERSON *et al.*, 2013).

Subjective and Turing tests assess both verbal and non-verbal communication aspects of a talking head. Subjective tests require careful guidance of the participants since they are asked to judge abstract criteria like "smoothness" and "precision". Frequently, the subjective test results present discrepancies caused by individual reactions; since a viewer can be influenced by the facial or voice characteristics of the animation (COSATTO *et al.*, 2003). In Turing tests, a high level of confusion between the facial animation and real video is an indicator that the tested model shows a high level of videorealism. However, poor results in Turing tests typically do not provide useful feedback information to improve the synthesis model.

On the other hand, objective measures extracted from the comparison between original and synthesized trajectories of facial feature points provide information about how well the control model predicts the trajectories, but it is not obvious how they relate to the perceived quality of the animations (BESKOW, 2004).

Speech intelligibility and emotion recognition tests turn possible to focus on specific aspects of the synthesis model like the proper reproduction of the speech articulatory movements or the realistic representation of the emotional facial expressions. Additionally, they provide objective measures of user perception which can be used to compare different models and keep track of model improvements.

This chapter presents the results of a perceptual evaluation performed to validate the synthesis methodology presented in Chapter 5. Given the focus of the present work on the synthesis of expressive speech, the emotion recognition test strategy was adopted. The participants were asked to recognize the emotions expressed by a real face during speech through the observation of muted video excerpts with few seconds of duration. The results were compared to the recognition rates obtained by the participants when the same emotional utterances are animated using a head model built upon image samples from the same face. Additionally, the participants were asked to evaluate the perceived valence of the emotional stimuli as negative, neutral or positive.

Aiming to validate the present synthesis methodology, the results of the emotions evaluation presented in this chapter are complemented by the speech intelligibility test results reported in (COSTA; DE MARTINO, 2013). The work compares the speech intelligibility scores obtained by the context-dependent viseme synthesis strategy (Chapter 5) to the results obtained by real video; and a simple linear morphing viseme approach that does not include the modeling of speech coarticulation effects. Costa and De Martino (2013) show

that,in comparison to the simple linear morphing viseme strategy, the context-dependent viseme approach greatly improves the speech intelligibility of audio heavily degraded by noise. Additionally, in situations where audio is intelligible, the results obtained by the context-dependent visemes approach are not statistically distinguishable from those obtained by real video.

The present chapter is organized as follows: Section 6.1 details how the stimuli for the expressive speech evaluation were generated; Section 6.2 describes the adopted evaluation protocol; Section 6.3 describes the population of participants; Section 6.4 presents the results of the evaluation and finally, Section 6.5 presents the concluding remarks of the chapter.

6.1 Test Stimuli

The emotion recognition test was performed presenting to the participants two types of stimuli: recorded video of a real face (real video) and synthetic facial animation (facial animation).

The procedure to generate the real video stimuli is summarized as follows:

1. The video material of a female actress playing the twenty-two OCC emotions was selected from the built corpus (Chapter 3). The selection was made after a visual inspection of the video material to guarantee that the chosen face would not present any particular characteristic that could deviate the attention of the observer or influence his/her judgement like: scars, uncommon spots in the facial skin or deep wrinkles of age in the forehead.

2. For each OCC emotion, the actress video material was divided into two: part of the material was used to generate test stimuli (test dataset) and the complementary part was used to extract the visemes to build the face model following the methodology described in Chapter 3, Section 3.2.

3. From the test dataset, twenty-two representative sentences (one for each OCC emotion) were selected to be presented during the test. The selected video fragments have duration between 2 to 4 seconds. As described in Chapter 3, the sentences extracted from the corpus are in Brazilian Portuguese.

4. The audio track from the selected videos was discarded and the final real video stimuli consists of twenty-two muted video clips representing the vocabulary of OCC emotions.

After the generation of the real video test stimuli, the procedure to generate facial animation stimuli is summarized as follows:

1. A talking head face model is built upon the training set video material, following the methodology described in Chapter 5.

2. The audio tracks extracted from the test stimuli video clips were manually segmented using Praat (BOERSMA; WEENINK, 2001). The objective of the segmentation was to obtain — for each phrase representing one of the OCC emotions — the timed phonetic transcription file that guides the facial animation synthesis. Thus, the test facial animations are the synthetic animated versions of the same utterances of the real video test stimuli.

3. In order to simulate the trajectory of the head in the facial animation, the facial animation synthesis process was also guided by the definition of a sequence of target shapes that informs the position that the head should assume at specific time instants. The feature points that define a target shape were manually marked in the real video frames that are associated to the half period instant of each phone in the speech (Figure 6.1). The shape modulation process is covered in Chapter 5.

4. The facial animation test stimuli were synthesized with full quality (no dimensionality reduction of the appearance model) following the methodology described in Chapter 5. No audio track was combined with the generated sequence of animation frames.

Figures 6.2 and 6.3 present examples of frames extracted from the stimuli video clips: the top rows are real video snapshots and the bottom rows are the corresponding synthetic frames. Section 6.2 describes the evaluation protocol applied to present the stimuli to the participants and to obtain their votes.

6.2 Evaluation Protocol

The evaluation was guided by a test application running on a desktop computer in a quiet room dedicated to the experiment, under the supervision of a researcher.

Each subject was firstly welcomed in the test room and quickly interviewed to obtain some basic personal data and to check if the participant presented any visual condition or problem that is not corrected through the use of eyeglasses or eye lenses. The subject was then asked to read a set of written instructions that introduce the evaluation and the operation of the test application (the original Brazilian Portuguese test instructions are presented in

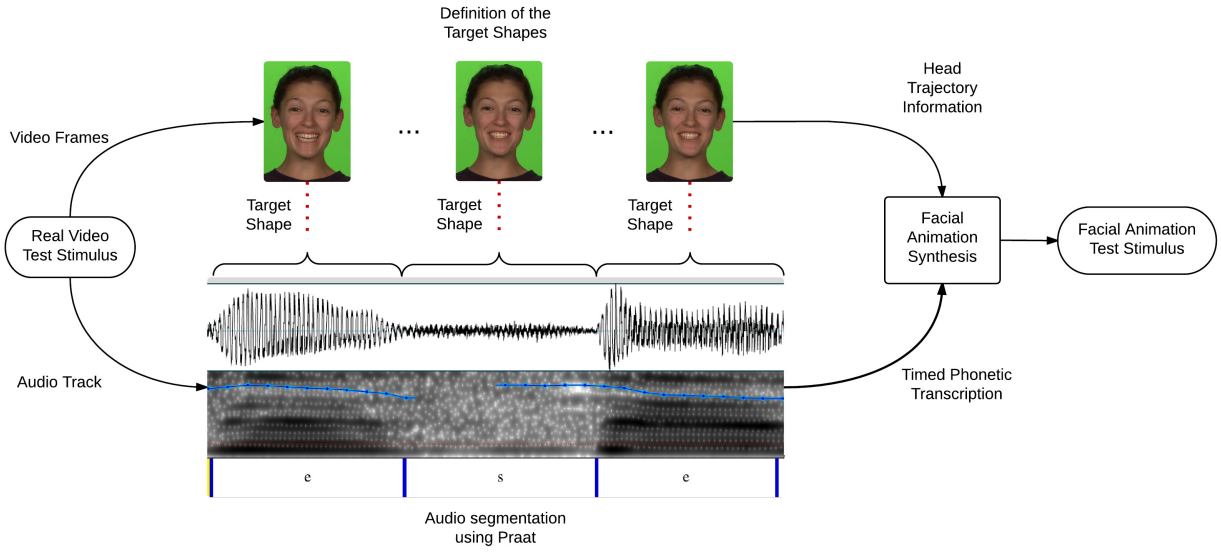


Figure 6.1 – The facial animation test stimuli animate the same utterances of the real video test stimuli. The figure illustrates the audio segmentation process and the definition of target shapes to simulate the head movements of the facial animation, based on the manual annotation of real video frames associated to the half period instant of phones.

Annex E). Following, the test supervisor presents a test application demo. The subjects interacted with the test application demo until they felt comfortable to proceed with the evaluation.

The test application was designed to select a random sequence of presentation of stimulus types (real video first and then facial animation, or vice versa). For each stimulus type, the order of presentation of the twenty-two sentences was also randomly selected. The height of the face on the screen was about 8 cm. The Java application developed to guide the evaluation has the following flow of execution (the interface of the test application is in Brazilian Portuguese):

1. A new test is started; the name of the participant is registered and a new test result file is created.

2. In the background, the application randomly selects which type of stimulus will be presented first: real video or facial animation.

3. The application selects a random sequence of presentation for the twenty-two video clips corresponding to the OCC emotions.

Figure 6.2 – Examples of frames extracted from real video stimuli (top row) and facial animation stimuli (bottom row). (a) "Happy For", (b) "Joy", (c) "Hope", (d) "Satisfaction", (e) "Relief", (f) "Pride", (g) "Gratification", (h) "Gratitude", (i) "Admiration", (j) "Love", (k) "Pity" and (l) "Sadness".

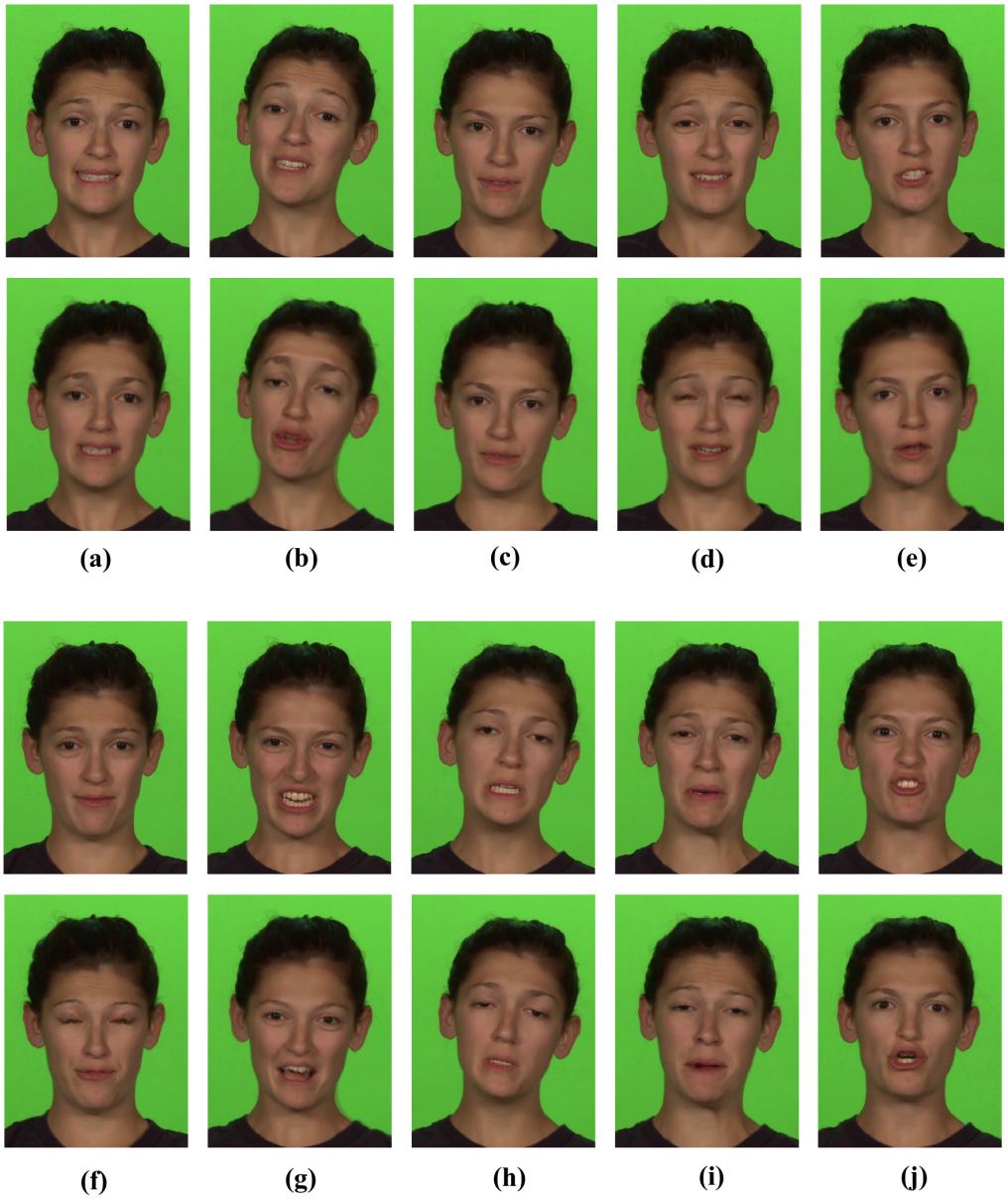


Figure 6.3 – Examples of frames extracted from real video stimuli (top row) and facial animation stimuli (bottom row). (a) "Fear", (b) "Resentment", (c) "Fears Confirmed", (d) "Shame", (e) "Reproach", (f) "Remorse", (g) "Gloating", (h) "Disapointment", (i) "Disgust", (j) "Anger".

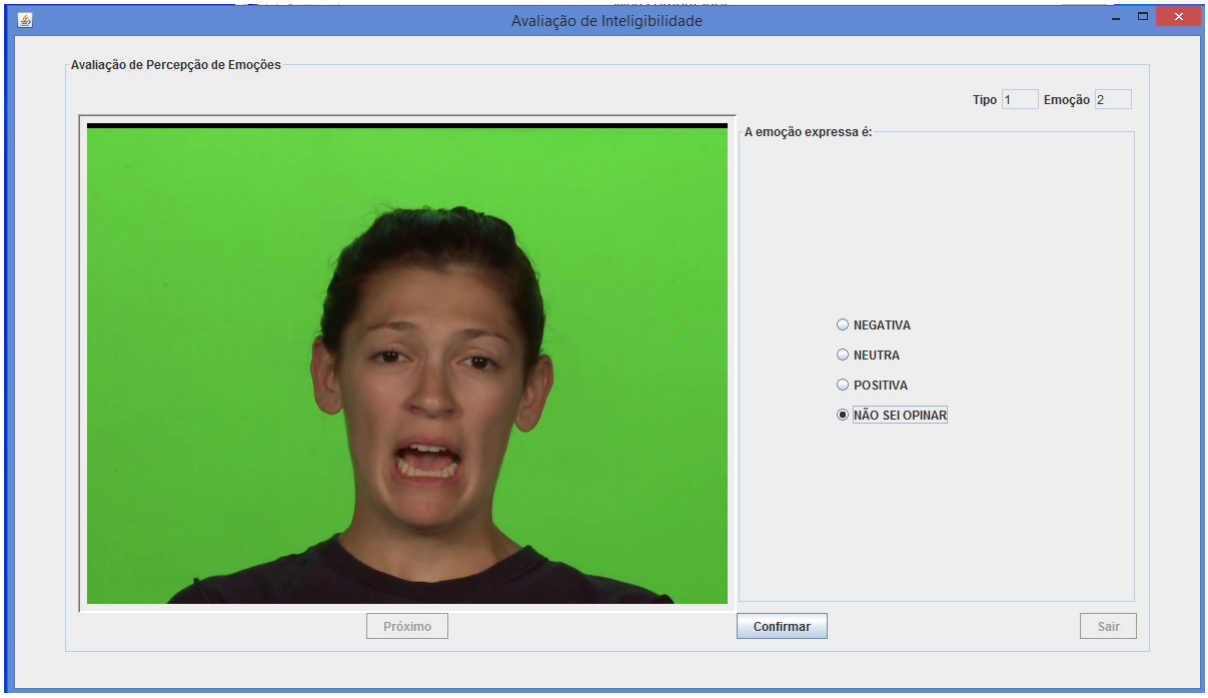


Figure 6.4 – Screenshot of the test application, with the video screen on the left and the voting options at the right. For each video, the participant is asked first, to give his/her opinion regarding the valence of the expressed emotion: negative ("NEGATIVA"), neutral ("NEUTRA"), positive ("POSITIVA") or no opinion ("NÃO SEI OPINAR").

4. A video is played on the screen, as depicted in Figure 6.4. As described in Section 6.1, real video and facial animation video clips are muted and presented without any accompanying audio. The participant cannot vote before the video clip finishes and also, he cannot advance the test without confirming his vote. The text in the screen asks the subject his opinion about the perceived valence of the emotion through the sentence "The expressed emotion is:" ("A emoção expressa é:"). The voting options are: "negative"("NEGATIVA"), "positive"("POSITIVA"), "neutral"("NEUTRAL") or "no opinion" ("NÃO SEI OPINAR").

5. The participant votes, confirms his option ("Confirmar") and advances the test ("Próximo").

6. The same video is played again in a new screen, as depicted in Figure 6.5. The text in the screen asks the participant to complete the sentence: "The informant seems to be:" ("A informante parece estar:"). The "NDA" voting option corresponds to the "None of the above" (NOTA) option. The remaining voting options correspond to the OCC emotions vocabulary adapted to complete the sentence, as detailed in Table 6.1.

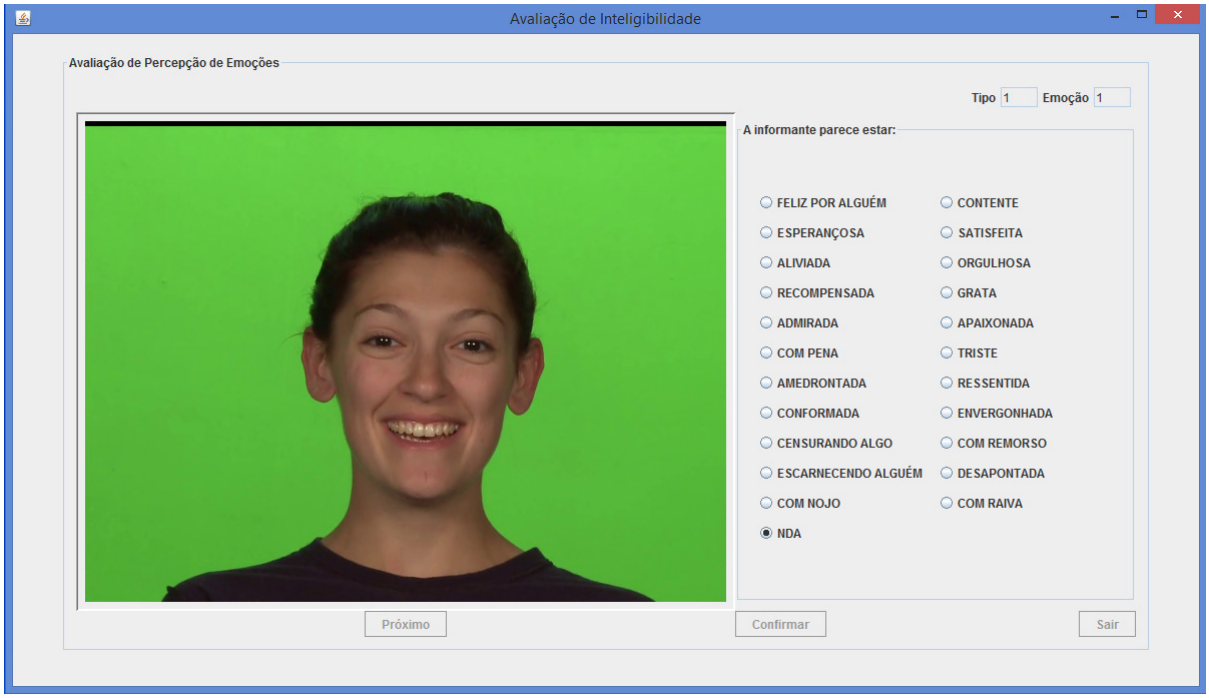


Figure 6.5 – Screenshot of the test application, showing the twenty-two voting options corresponding to the OCC emotions, in Brazilian Portuguese. The voting options correspond to the OCC emotions as mapped in Table 6.1.

7. The participant votes, confirms his option and advances the test.

8. A new video clip of the same type of stimulus (real video or facial animation) is played. If all emotions of the first type of stimulus were played, the test application proceeds to the next type. The test continues from step 2.

9. The steps 3 to 7 are repeated for both types of stimuli. The test finishes after the presentation of the last video clip. In total, the subjects are asked to vote 88 times (2 types of stimuli $\times$ 22 video clips $\times$ 2 reproductions per emotion).

After the evaluation, the subjects were interviewed again and the test supervisor asked and wrote down the answers for the following open questions:

- "Did you identify what the face was saying at any particular moment?" (If positive, the test supervisor took notes of the situations and the sentences that the participant remembered).

- "What were the visual cues that helped you to identify the emotions?"

The average duration of the tests was sixteen minutes.

OCC Emotion	Adapted term to complete the sentence: "The informant seems to be..."	Voting option in Brazilian Portuguese
Happy For	happy for someone	"FELIZ POR ALGUÉM"
Joy	cheerful	"CONTENTE"
Hope	hopeful	"ESPERANÇOSA"
Satisfaction	satisfied	"SATISFEITA"
Relief	relieved	"ALIVIADA"
Pride	proud	"ORGULHOSA"
Gratification	rewarded	"RECOMPENSADA"
Gratitude	thankful	"GRATA"
Admiration	admired	"ADMIRADA"
Love	in love	"APAIXONADA"
Pity	sorry for someone	"COM PENA"
Sadness	sad	"TRISTE"
Fear	fearful	"AMEDRONTADA"
Resentment	resentful	"RESSENTIDA"
Fears Confirmed	resigned	"CONFORMADA"
Shame	embarrassed	"ENVERGONHADA"
Reproach	reproaching something	"CENSURANDO ALGO"
Remorse	remorseful	"COM REMORSO"
Gloating	gloating	"ESCARNECENDO ALGUÉM"
Disappointment	disappointed	"DESAPONTADA"
Disgust	disgusted	"COM NOJO"
Anger	angry	"COM RAIVA"

Table 6.1 – Mapping between the OCC emotions vocabulary and the voting options of emotion recognition test.

6.3 Participant's Profile

Fifty-three subjects participated in the emotion recognition test as volunteers. This group of individuals consisted of undergraduate and graduate students and administrative personnel of the University of Campinas. They had no prior knowledge of the test purpose or any involvement with the research. The average age of the participants is 32 years old, ranging from 19 to 66 years. All the participants were Brazilian Portuguese native speakers.

The test results of three subjects were discarded due to subjects' vision problems. Two of them reported color blindness and "nystagmus", respectively; vision problems that cannot be corrected through the use of eyeglasses or eye lenses. One of the subjects forgot his eyeglasses for hypermetropia correction. The test results of other two subjects were discarded because they reported the "lipreading" of several sentences during the test. Test supervisor

checked that they identified the sentences correctly. In summary, the analysis of the results was performed considering the test results of forty-eight subjects.

6.4 Results

The following subsections present the perceptual evaluation results accompanied by the statistical analysis of the data. The mechanisms of expression and perception of emotions by humans is a wide and complex field of study in psychology and neuroscience. For this reason, it is important to highlight that all the analyses presented in this section are not focused on checking the precision of the expression of emotions nor the capability of the subjects of recognizing the emotions with accuracy. Instead, all the analyses presented in this section are focused on comparing the facial animation scores with the equivalent results obtained by the real video reference. The statistical tests and plots discussed in the following subsections were generated using R — a language and environment for statistical computing and graphics (R CORE TEAM, 2014). The R script to implement the analyses is presented in Annex F[1].

6.4.1 Perceived Valence of Emotions

Table 6.2 summarizes the results of the perceived valence of the emotions expressed by the talking face. The first column lists the emotions stimuli and the adjacent two groups of four columns report the correspondent percentage of votes that each voting option received, discriminated by the type of stimulus (real video or facial animation).

First, considering the voting options as categorical variables, the statistical analysis of the results was performed based on the contingency tables built for each emotion. In statistics, a contingency table is a two-dimensional matrix arrangement that classifies the samples from some population with respect to two or more qualitative variables (EVERITT, 1992). The entries in the matrix cells are frequencies. Tables 6.3 and 6.4 are examples of the contingency tables obtained for “Anger” and “Relief”, respectively. In this case, the samples are the participant votes and their frequencies are discriminated according the type of stimulus: real video or facial animation. In Table 6.3, for example, all the population subjects (48 subjects) perceived “Anger” as “Negative” when they observed the real video stimulus. Considering facial animation results, 46 subjects (96%) perceived “Anger” as “Negative” and 2 (4%) subjects voted for “Neutral”.

[1] The script can also be downloaded at <https://github.com/pdpcosta/myRScripts/tree/master/EmotionRecognitionEvaluation>.

	Distribution of Votes Real Video								Distribution of Votes Facial Animation								Classification		p-value Fisher's Exact Test	p-value MWM Test
	Negative		Neutral		Positive		No Opinion		Negative		Neutral		Positive		No Opinion		Real Video	Facial Animation		
Anger	48	100%	0	0%	0	0%	0	0%	46	96%	2	4%	0	0%	0	0%	Strong Negative	Strong Negative	0.4947	0.1595
Disgust	47	98%	1	2%	0	0%	0	0%	46	96%	2	4%	0	0%	0	0%			1	0.5677
Reproach	47	98%	0	0%	1	2%	0	0%	39	81%	9	19%	0	0%	0	0%			0.0026	0.0096
Disapointment	41	85%	7	15%	0	0%	0	0%	28	58%	18	38%	0	0%	2	4%		Negative	0.0072	0.0028
Fear	35	73%	13	27%	0	0%	0	0%	30	63%	15	31%	0	0%	3	6%	Negative		0.2438	0.2116
Shame	35	73%	13	27%	0	0%	0	0%	35	73%	12	25%	0	0%	1	2%			1	0.9545
Sadness	34	71%	14	29%	0	0%	0	0%	30	63%	16	33%	0	0%	2	4%			0.3872	0.3281
Fears Confirmed	30	63%	17	35%	1	2%	0	0%	32	67%	13	27%	1	2%	2	4%			0.5298	0.8097
Pity	30	63%	18	38%	0	0%	0	0%	28	58%	15	31%	0	0%	5	10%			0.0799	0.4304
Remorse	25	52%	22	46%	1	2%	0	0%	19	40%	26	54%	0	0%	3	6%		Neutral	0.1312	0.1707
Resentment	18	38%	23	48%	5	10%	2	4%	18	38%	28	58%	0	0%	2	4%	Neutral		0.1374	0.5684
Gratitude	6	13%	17	35%	25	52%	0	0%	9	19%	33	69%	5	10%	1	2%	Positive		0.0001	0.0007
Love	3	6%	13	27%	29	60%	3	6%	6	13%	9	19%	33	69%	0	0%		Positive	0.1867	0.6623
Gloating	5	10%	1	2%	40	83%	2	4%	7	15%	6	13%	33	69%	2	4%	Strong Positive		0.2119	0.1438
Relief	1	2%	3	6%	44	92%	0	0%	4	8%	14	29%	27	56%	3	6%			0.0004	0.0104
Pride	2	4%	0	0%	46	96%	0	0%	8	17%	4	8%	34	71%	2	4%			0.0041	0.0322
Gratification	0	0%	2	4%	46	96%	0	0%	2	4%	8	17%	37	77%	1	2%		Strong Positive	0.0227	0.0354
Admiration	1	2%	1	2%	46	96%	0	0%	1	2%	1	2%	46	96%	0	0%			1	1
Hope	0	0%	2	4%	46	96%	0	0%	0	0%	5	10%	43	90%	0	0%			0.4353	0.2447
Satisfaction	0	0%	2	4%	46	96%	0	0%	1	2%	0	0%	45	94%	2	4%			0.2448	0.1924
Joy	0	0%	0	0%	47	98%	1	2%	2	4%	1	2%	44	92%	1	2%			0.4893	0.1828
Happy For	0	0%	0	0%	48	100%	0	0%	0	0%	8	17%	40	83%	0	0%			0.0057	0.0034

Table 6.2 – Valence Perception Results (population size = 48 subjects)

Taken into account the small sample size of the contingency tables, the Fisher's exact test is applicable (AGRESTI, 2002). Fisher's exact test is used to determine if there are nonrandom associations between two categorical variables. The test analyzes whether the proportions of one variable are different depending on the value of the other variable. In the present study, the categorical variables considered are the perceived valence versus the type of stimulus. If the valence of the emotions expressed in real video or facial animation is similarly perceived by the subjects, it is expected to observe similar distributions of votes for both types of stimulus. In other words, the null hypothesis (H_0) to be tested is that the votes proportions are not dependent on the type of stimulus. In the present work, the null hypothesis rejection at the 5% significance level is adopted ($p < 0.05$). For "Anger", for example, the resulting Fisher's exact test p-value is 0.4947 and, consequently, the H_0 cannot be rejected (observe in Table 6.3 that the proportions of votes are very similar for both types of stimuli). On the other hand, the p-value obtained for "Relief" is 0.0004, resulting in H_0 being rejected. In this case, different distributions of votes have been associated to different types of stimulus. In fact, Table 6.4 shows that when compared to the real video, a greater proportion of subjects interpreted the facial animation of "Relief" as "neutral" instead of "positive". Among the twenty-two emotions, seven emotions resulted in H_0 being rejected at the 5% level: "Reproach", "Disappointment", "Gratitude", "Relief", "Pride", "Gratification" and "Happy For".

	Real Video **Number of Votes (% of Votes)**	**Facial Animation** **Number of Votes (% of Votes)**
Negative	48 (100%)	46 (96%)
Neutral	0 (0%)	2 (4%)
Positive	0 (0%)	0 (0%)
No Opinion	0 (0%)	0 (0%)

Table 6.3 – Frequency contingency table of the voting options versus the type of stimulus for "Anger" (population size = 48 subjects).

	Real Video **Number of Votes (% of Votes)**	**Facial Animation** **Number of Votes (% of Votes)**
Negative	1 (2%)	4 (8%)
Neutral	3 (6%)	14 (29%)
Positive	44 (92%)	27 (56%)
No Opinion	0 (0%)	3 (6%)

Table 6.4 – Frequency contingency table of the voting options versus the type of stimulus for "Relief" (population size = 48 subjects).

In Table 6.2, the rows have been organized so that the emotions classified by the

majority of the voters as "negative" come first, followed by the emotions classified as "neutral" and finally, the emotions classified as "positive", taking the real video results as sorting reference. The pair of columns named "Classification" propose a grouping of the emotions based on the perception of the majority of the subjects as follows:

- *Strong Negative*: 75% to 100% of the participants perceived the emotion stimulus as "negative".
- *Negative*: 50% to 75% of the participants perceived the emotion stimulus as "negative".
- *Neutral*: it is not possible to classify the emotion stimulus as "negative" or "positive"; the majority of the participants voted for "Neutral".
- *Positive*: 50% to 75% of the participants perceived the emotion stimulus as "positive".
- *Strong Positive*: 75% to 100% of the participants perceived the emotion stimulus as "positive".

This classification shows that it is possible, based on the test results, to organize the vocabulary of emotions as a transition from *Strong Negative* to *Strong Positive* emotions, passing through intermediate levels of valence. Those who advocate the dimensional modeling of emotions (Section 2.2.2, Chapter 2) would call this an evidence of valence as an axis of projection of emotions. From another point of view, this behavior suggests that the voting options "negative", "neutral" and "positive" could be associated to numeric ordered values and analyzed as ordinal variables. The advantage of this approach is to explore the numeric relationships of the variable distributions. In the present work, the following values were adopted to represent the valence: -1 to "negative" votes, 0 to "neutral" votes and +1 to "positive" votes. The "no opinion" results are defined as "not available" (NA) data (in R, this nomenclature triggers continuity correction algorithms). From this setup, it was possible to apply the Wilcoxon-Mann-Whitney (WMW) test, which is appropriate for small samples and ordinal numeric variables that present non-normal distributions (KLOKE; MCKEAN, 2014). The objective of the WMW test is to detect if two sample populations are centered differently. Let X and Y be continuous random variables and $f(t)$ and $g(t)$ respectively denote the probability density functions of X and Y. X and Y are said to present identical distributions with location shift Δ when $g(t) = f(t - \Delta)$. The WMW test null hypothesis is that the distributions of real video and facial animation scores present a location shift $\Delta = 0$, against the alternative hypothesis that $\Delta \neq 0$. In the cases that the H_0 is rejected, the two sample distributions are statistically distinguishable, meaning that different perception mechanisms took place. The last column of Table 6.2 presents the obtained p-values for the WMW

test for each emotion stimulus. The WMW p-values interpretation was coincident with the results obtained by the Fisher's exact test; the null hypothesis is rejected at the 5% level ($p < 0.05$) for: "Reproach", "Disappointment", "Gratitude", "Relief", "Pride", "Gratification" and "Happy For".

6.4.2 Recognition of Emotions

This section presents the results of the second part of the perceptual evaluation, in which participants are asked to recognize the emotions expressed by the talking face as described in Section 6.2.

6.4.2.1 Analysis of the Correct Answers

A first approach to analyze the results is to count a "correct answer" every time a subject votes for an emotion that is coincident with the emotion stimulus that was played in the screen. From this definition, Table 6.5 presents the summary of the statistics of correct answers given by the subjects for both real video and facial animation cases. As shown in the table, the average percentage of correct answers per subject for real video is 22%, compared to 19% for facial animation.

Figure 6.6 shows the boxplots of the distributions of the percentage of correct answers obtained by the participants for each type of stimulus. The Shapiro-Wilk test was applied to test the normality of both sample distributions (SHAPIRO *et al.*, 1968). The null hypothesis that the distribution is normal was rejected for the facial animation results with significance level $p < 0.05$ ($p = 0.0338$). In this case, as discussed in Section 6.2, the Wilcoxon-Mann-Whitney (WMW) test can be applied to compare the two sample distributions. The resulting p-value is $p = 0.0717$ and the null hypothesis that both distributions are identical and centered in the same region (location shift $\Delta = 0$) cannot be rejected at the 5% level. In other words, considering the adopted hypothesis, the real video and facial animation distributions of correct answers are not statistically distinguishable.

6.4.2.2 Analysis of the Votes Distributions per Emotion

The barplots presented in Figures 6.7 to 6.12 illustrate the voting patterns obtained for each emotion stimulus for both real video and facial animation cases. For each emotion, it is presented a pair of graphs: the top graph shows the distribution of votes obtained when the subjects judged the real video stimulus; and the bottom graph presents the distribution of votes for the facial animation stimulus. In Figure 6.7(a), for example, the top graph shows that when the subjects were asked to observe the real video of "Anger", more than 80%

Correct Answers
Population Size = 48 Subjects

	Real Video		Facial Animation	
	Score	**Percentage**	**Score**	**Percentage**
Minimum	1	4.5%	0	0.0%
1st Quartile	3.750	17.0%	3	13.6%
Median	5	22.7%	4	18.2%
Mean	4.896	22.3%	4.104	18.7%
3rd Quartile	6	27.3%	6	27.3%
Maximum	9	40.9%	8	36.4%

Table 6.5 – Summary of statistics of recognized emotions obtained for real video and facial animation. Twenty-two different emotions are presented to the subjects.

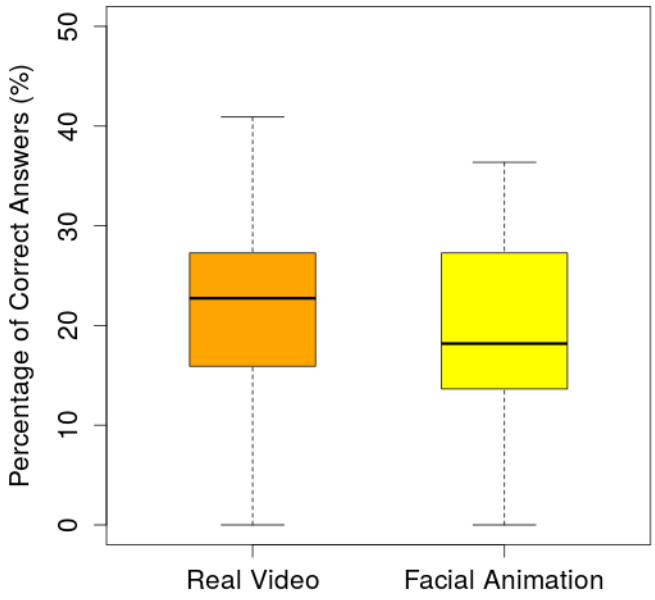


Figure 6.6 – Distribution of the percentage of correct answers obtained by the participants for real video and facial animation. The lower and upper whiskers of the boxplots represent the minimum and the maximum of the distribution, respectively. The bottom and top sides of the rectangles are the first and the third quartiles of the distribution, respectively. The line in the middle of the rectangles indicates the distributions medians.

of them perceived it as anger, but some subjects perceived it as gloating, pity, reproach or disappointment. Similarly, the bottom graph shows that more than 70% of the subjects perceived the facial animation of "Anger" as anger, but some subjects confused it with reproach, gloating and resentment. Additionally, a small percentage of the subjects voted for NOTA ("None of the above"), which means that they were not able to give an opinion about the emotion expressed by the facial animation or the emotion they identified was not

present among the voting options.

The barplots of Figures 6.7 to 6.12 can be considered the graphical representations of the contingency tables for each emotion stimulus. Table 6.6, for example, presents the contingency table for "Anger", which is illustrated by the barplots of Figure 6.7(a). The voting options are listed in the first column of the table in the same sequence they are presented in the test application screen (see Table 6.1 and Figure 6.5). The second column of the table refers to the frequency distribution of votes for the real video stimuli and the last column refers to the facial animation voting scores.

Similarly to the analysis performed in Section 6.4.1, the Fisher's exact test can be applied to test the association between the voting scores and the type of stimulus. Table 6.7 present the p-values derived from this analysis. As shown in the table, the H_0 was rejected with $p < 0.05$ for "Reproach", "Gratitude" and "Relief". For all other nineteen emotions, the similarity between real video and facial animation voting scores proportions is such that is not possible to reject the hypothesis that the voted options are not dependent of the stimulus type, i.e. real video and facial animation results are not statistically distinguishable.

6.4.3 Final Interview Results

After the evaluation, the test supervisor asked the subjects: "What were the visual cues that helped you to identify the emotions?". The subjects were free to respond with their own words while the test supervisor wrote down the answers. Frequently, the subjects mentioned more than one facial element, for example: "the mouth, the eyes and the eyebrows"; or simply, "the mouth and the eyes". The answers provided by the participants can be summarized as follows:

- The mouth region was cited 37 times. Among these, 4 subjects mentioned "the mouth and the teeth".
- The eyes were cited 27 times.
- The eyebrows were mentioned by 12 participantes.
- The head, in particular the way that it moves, was mentioned by 4 participants.
- The movements below the face region (on the neck and on the shoulders) were cited 4 times.
- The forehead frown was cited 3 times.
- The skin texture was mentioned by one participant.

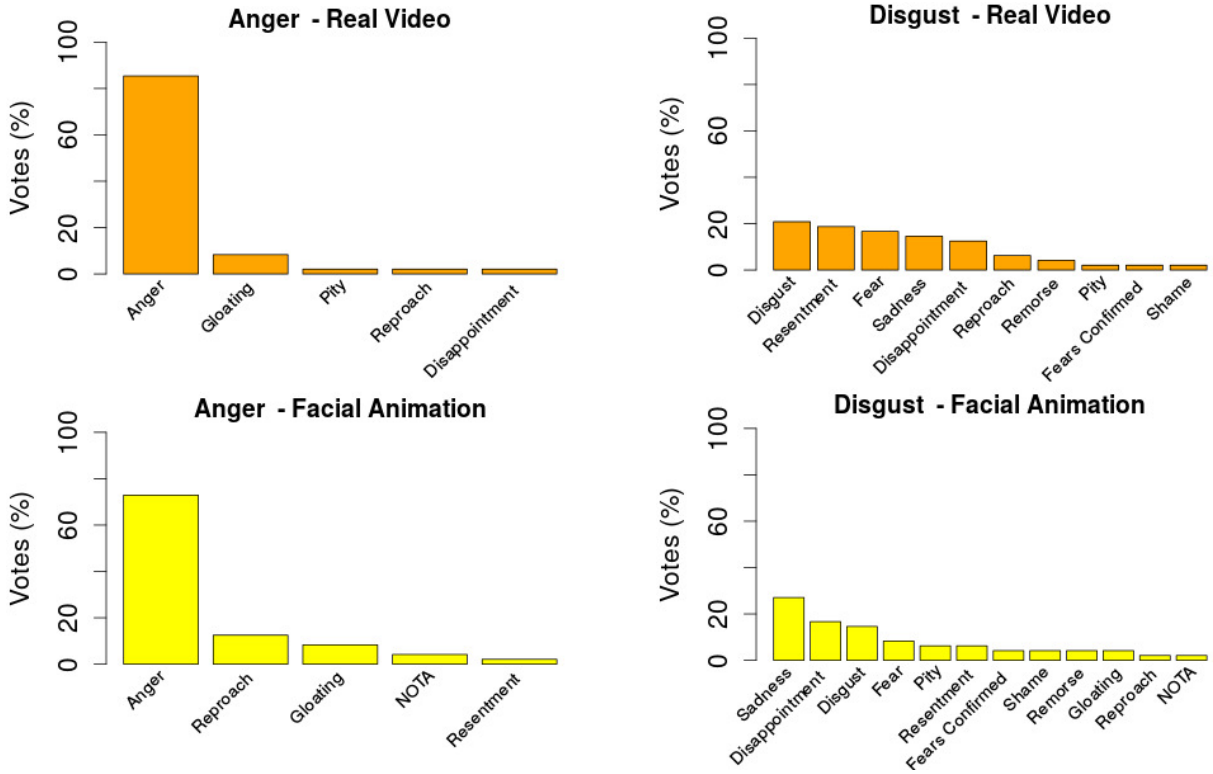
	Real Video Number of Votes (% of Votes)	Facial Animation Number of Votes (% of Votes)
Happy For	0 (0%)	0 (0%)
Joy	0 (0%)	0 (0%)
Hope	0 (0%)	0 (0%)
Satisfaction	0 (0%)	0 (0%)
Relief	0 (0%)	0 (0%)
Pride	0 (0%)	0 (0%)
Gratification	0 (0%)	0 (0%)
Gratitude	0 (0%)	0 (0%)
Admiration	0 (0%)	0 (0%)
Love	0 (0%)	0 (0%)
Pity	1 (2.1%)	0 (0%)
Sadness	0 (0%)	0 (0%)
Fear	0 (0%)	0 (0%)
Resentment	0 (0%)	1 (2%)
Fears Confirmed	0 (0%)	0 (0%)
Shame	0 (0%)	0 (0%)
Reproach	1 (2.1%)	6 (13%)
Remorse	0 (0%)	0 (0%)
Gloating	4 (8.3%)	4 (8%)
Disappointment	1 (2.1%)	0 (0%)
Disgust	0 (0%)	0 (0%)
Anger	41 (85.4%)	35 (73%)
NOTA	0 (0%)	2 (4%)

Table 6.6 – Frequency contingency table of the recognition voting options versus the type of stimulus for "Anger" (population size = 48 subjects). The emotions are listed following the same order they appear as voting options in the test application screen (Figure 6.5).

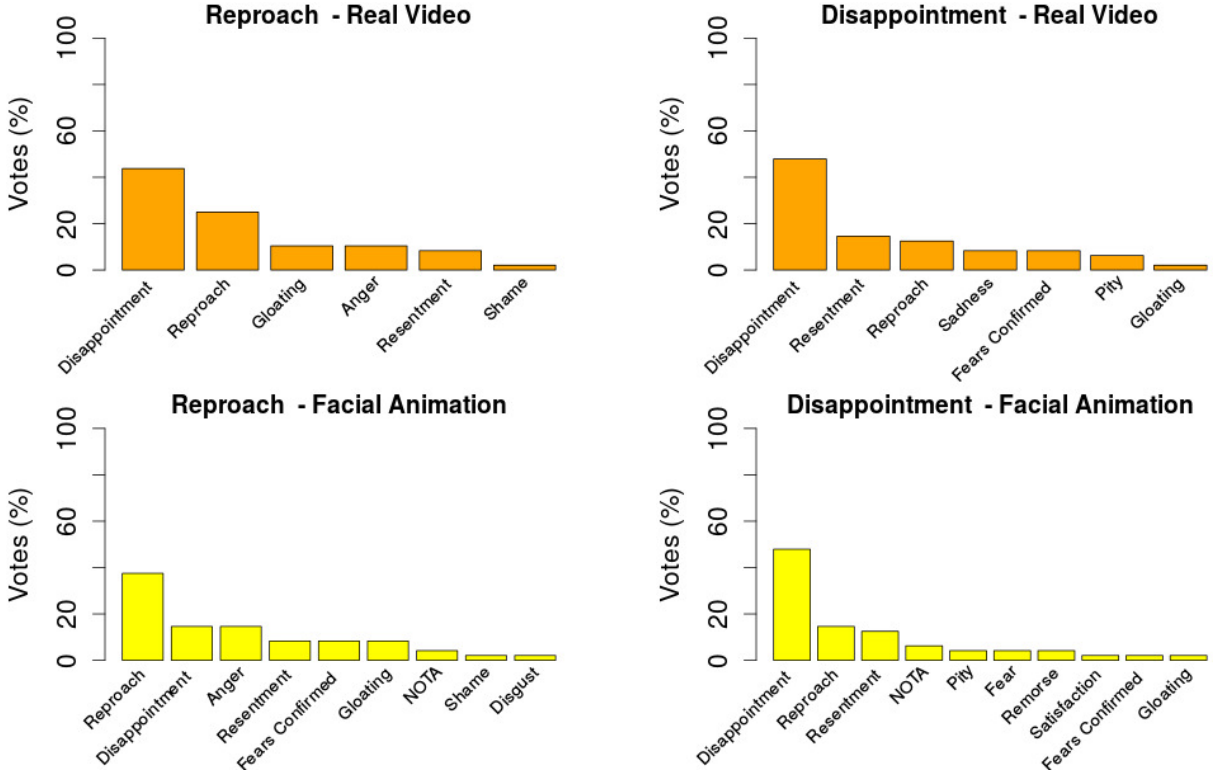
6.4.4 Discussion

Sections 6.4.1 and 6.4.2 show that in most cases facial animation and real video results present similar distributions and, according to the adopted hypotheses, they cannot be statistically distinguished. Therefore, this section focuses on discussing the cases in which this situation is not verified.

In Section 6.4.1, the perceptual valence evaluation showed that "Reproach", "Disappointment", "Gratitude", "Relief", "Pride", "Gratification" and "Happy For" emotions obtained statistically distinguishable results for real video and facial animation. For these emotions, Table 6.2 shows that the facial animation had a greater proportion of "Neutral" votes when compared to the real video stimulus. These results suggest that the facial anima-

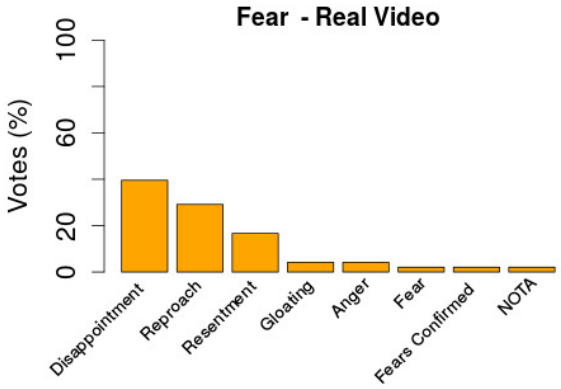


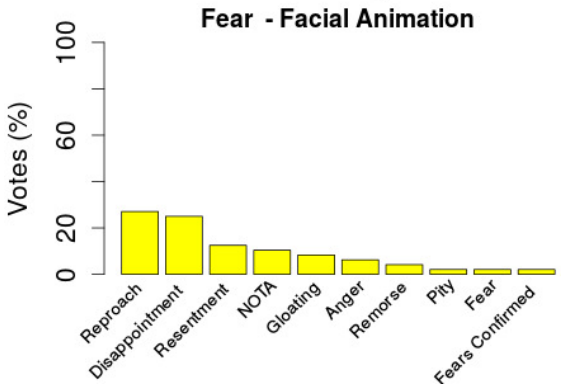
(a) Anger (b) Disgust



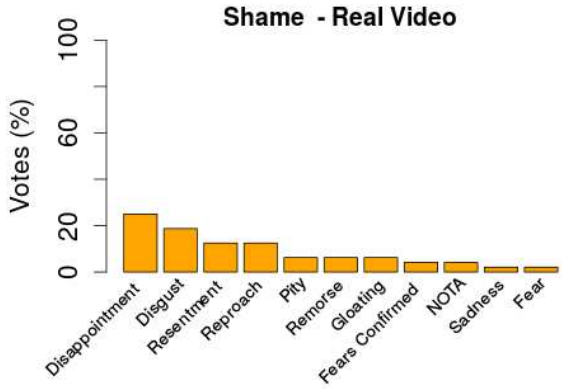
(c) Reproach (d) Disappointment

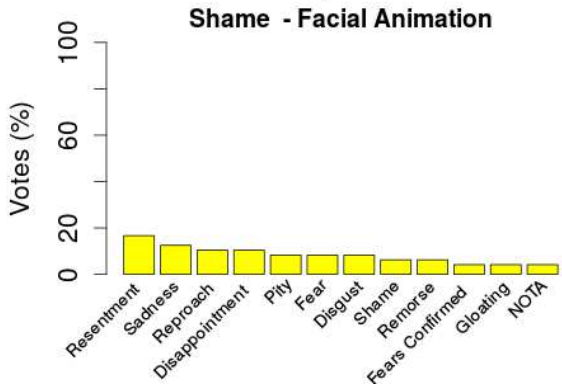
Figure 6.7 – Distribution of votes for real video and facial animation in function of the emotion stimulus.



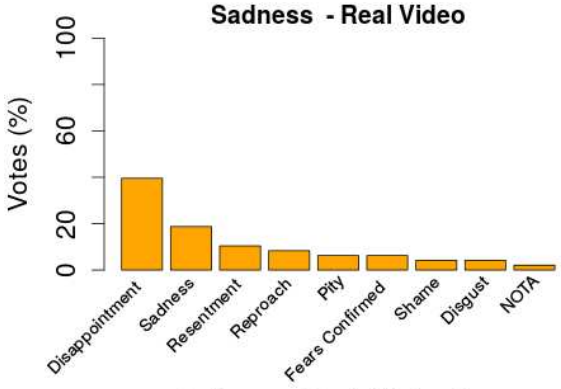


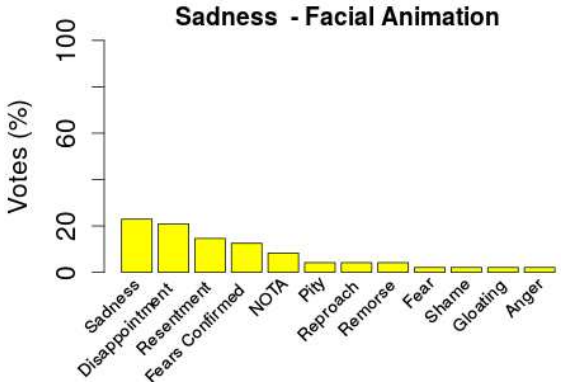
(a) Fear



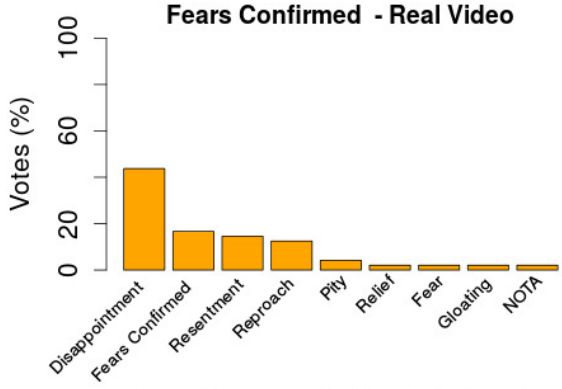


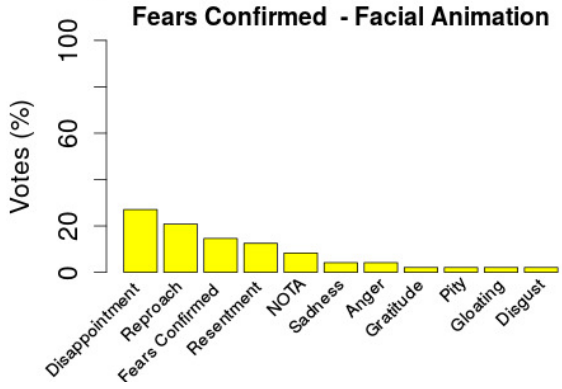
(b) Shame





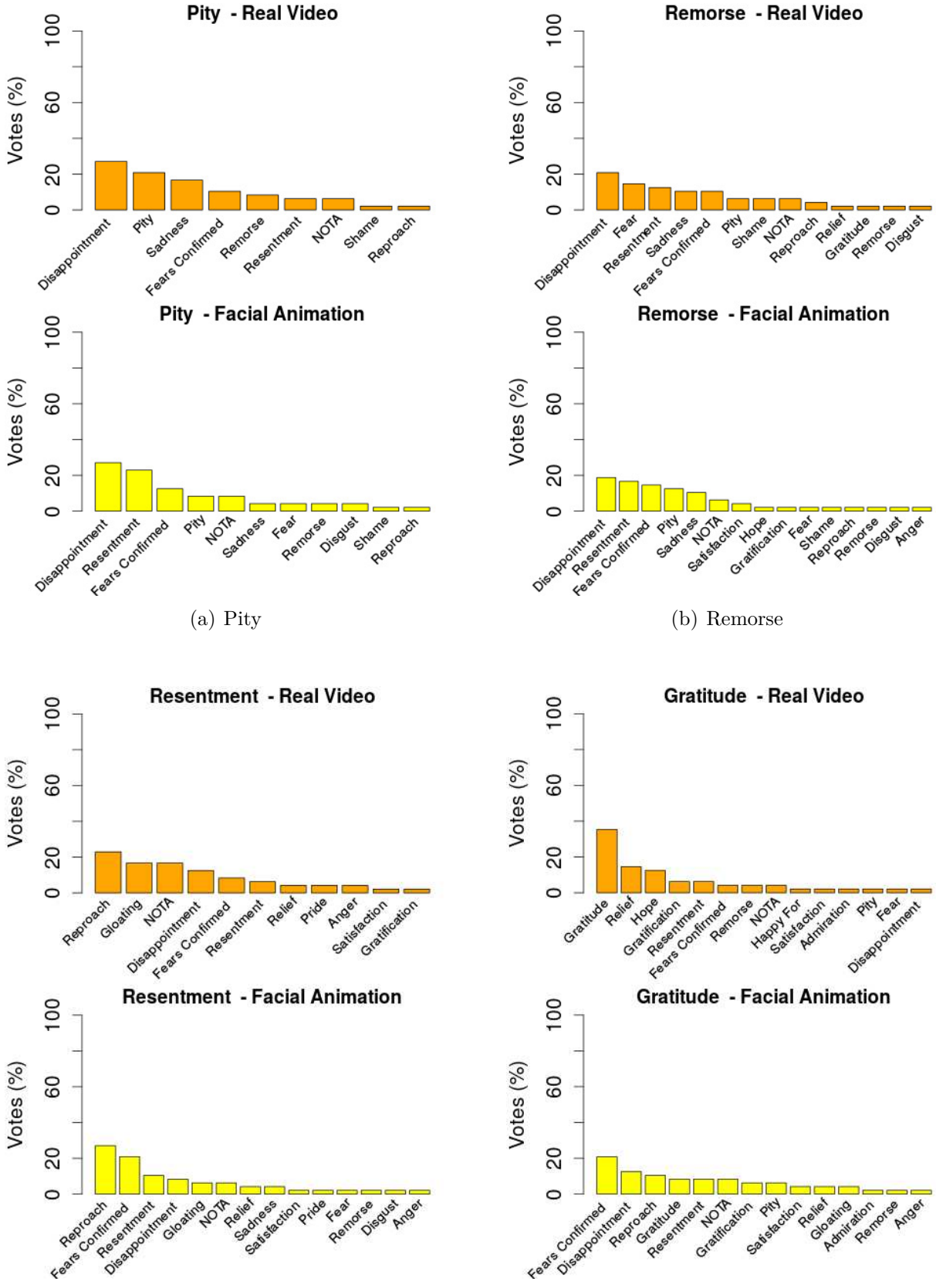
(c) Sadness





(d) Fears Confirmed

Figure 6.8 – Distribution of votes for real video and facial animation in function of the emotion stimulus.



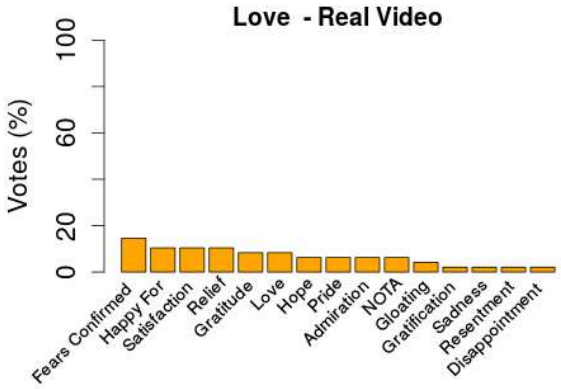
(a) Pity

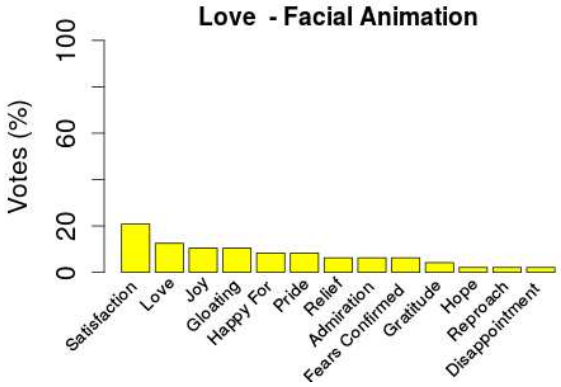
(b) Remorse

(c) Resentment

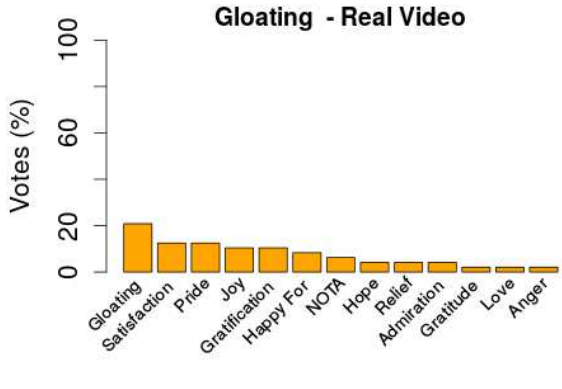
(d) Gratitude

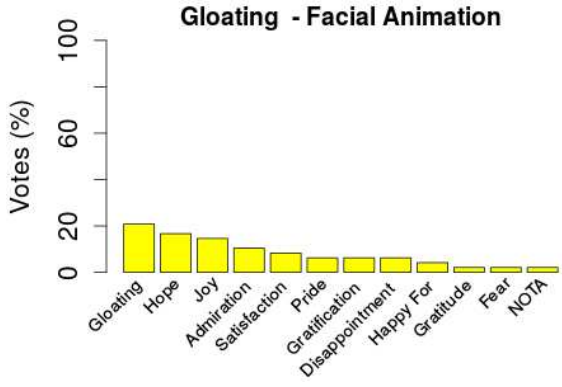
Figure 6.9 – Distribution of votes for real video and facial animation in function of the emotion stimulus.



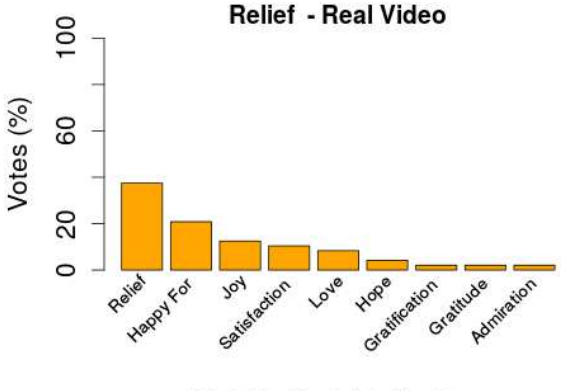


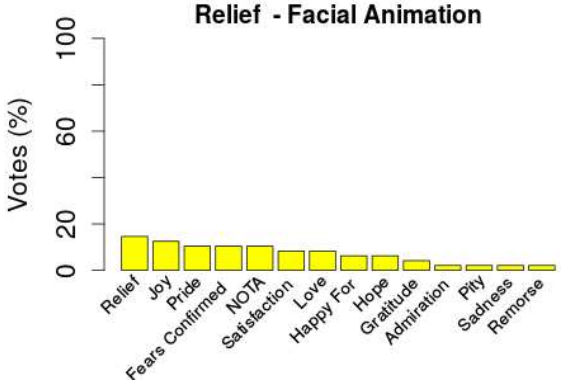
(a) Love



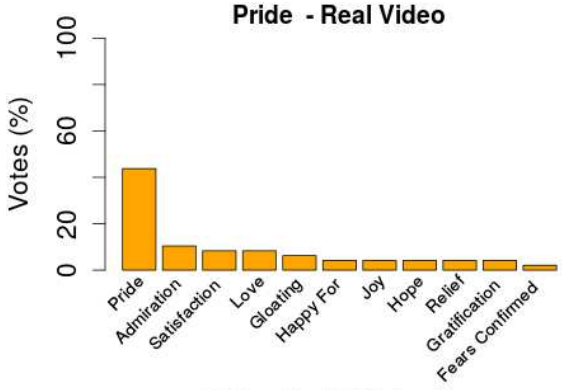


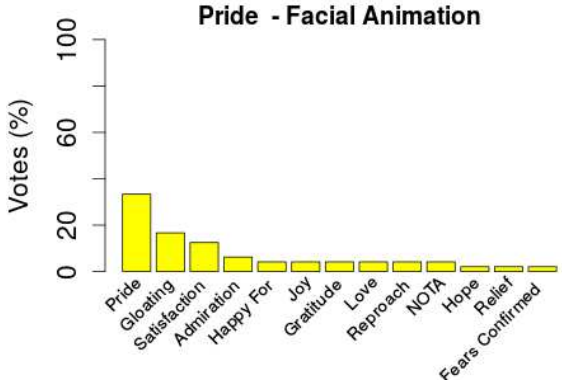
(b) Gloating





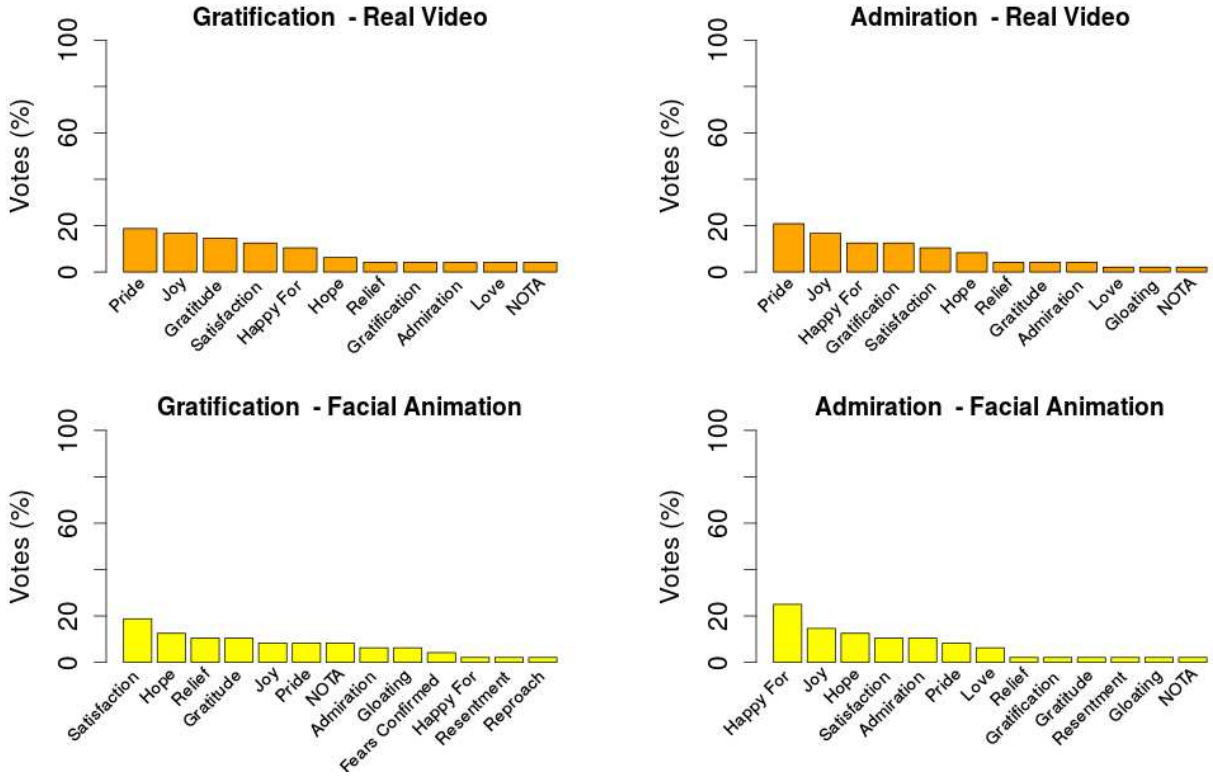
(c) Relief



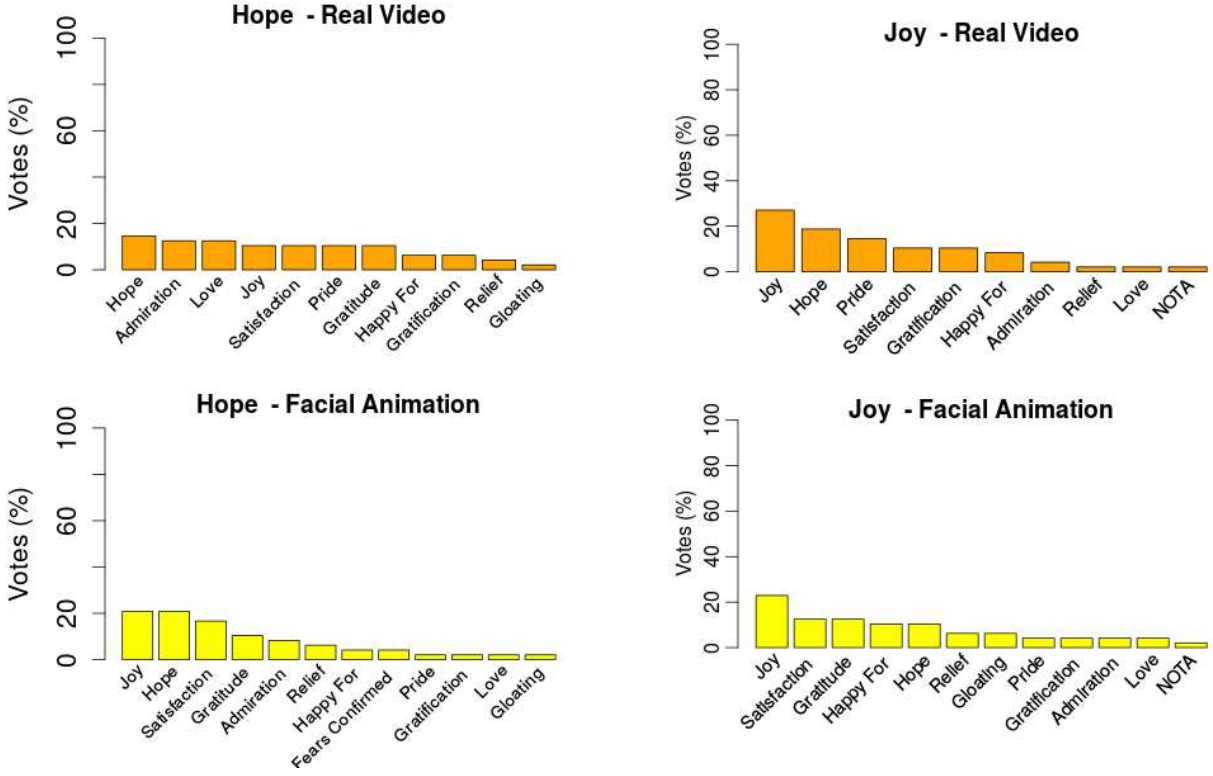


(d) Pride

Figure 6.10 – Distribution of votes for real video and facial animation in function of the emotion stimulus.



(a) Gratification (b) Admiration



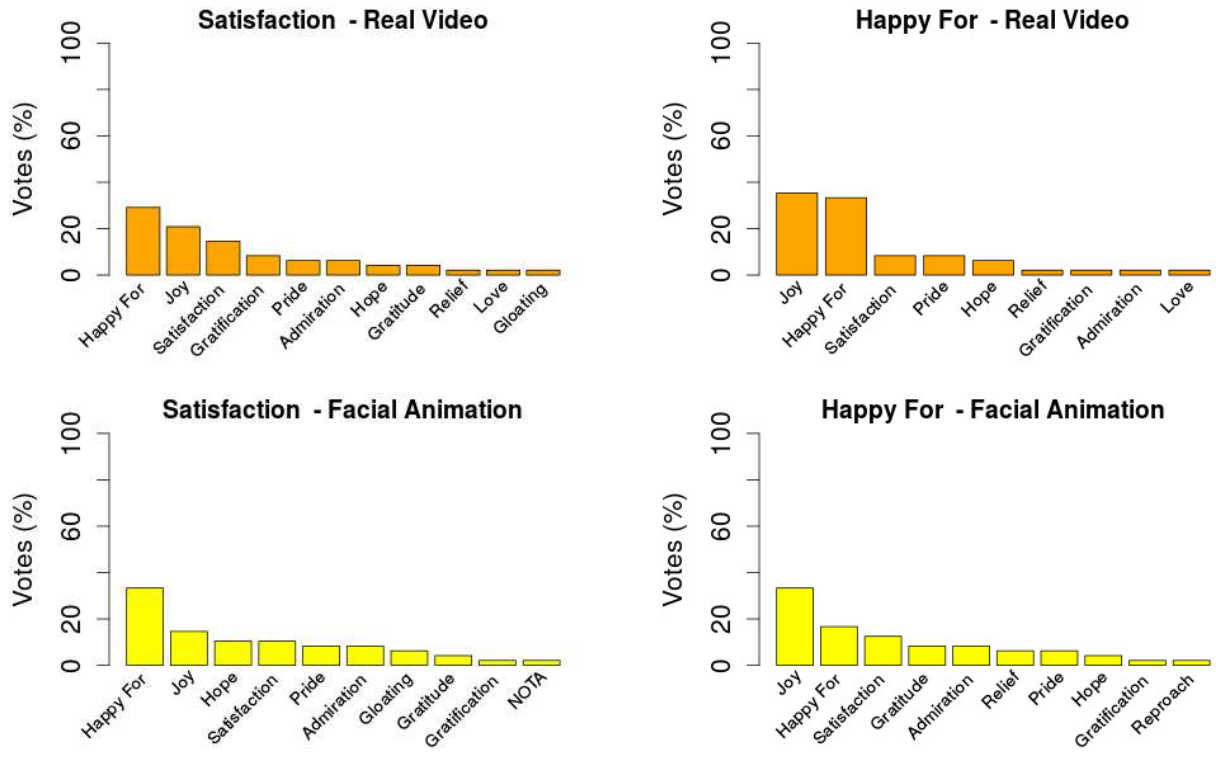
(c) Hope (d) Joy

Figure 6.11 – Distribution of votes for real video and facial animation in function of the emotion stimulus.

Emotion	p-value Fisher's Exact Test
Anger	0.1087
Disgust	0.3222
Reproach	0.0246 (H_0 rejected)
Disappointment	0.1601
Fear	0.4938
Shame	0.2413
Sadness	0.2857
Fears Confirmed	0.3571
Pity	0.0784
Remorse	0.5841
Resentment	0.3595
Gratitude	$<$0.0001 (H_0 rejected)
Love	0.2183
Gloating	0.2651
Relief	0.0023 (H_0 rejected)
Pride	0.5082
Gratification	0.1029
Admiration	0.3553
Hope	0.3054
Joy	0.1121
Satisfaction	0.7584
Happy For	0.2389

Table 6.7 – P-values obtained for the Fisher's exact test for each emotion stimulus. The emotions are listed following the same order presented in Table 6.2.

tion had an "attenuation effect" on the perceived valence of the towards neutral valence. In practice, the facial expressions and the head dynamics became less stereotyped. "Reproach" and "Disappointment" real video stimuli, for example, occupy the last two positions of the "Strong Negative" rank of emotions proposed in Section 6.4.1(see Table 6.2). This is an evidence that they present stereotypical dynamics and facial expressions that are commonly associated to the "negative" valence. However, for facial animation, "Reproach" and "Disappointment" results present a greater confusion between "negative" and "neutral" votes. For facial animation, "Disappointment" becomes part of the less stereotypical "Negative" group of emotions. The same "attenuation effect" was observed for "Gratitude" (from "Positive" in real video to "Neutral" in facial animation); and "Relief" and "Pride" (from "Strong Positive" to "Positive"). "Gratification" and "Happy For" also suffer a greater confusion between "positive" and "neutral" votes, but they are still perceived as positive by the great majority of subjects (77% and 83%, respectively).



(a) Satisfaction (b) Happy For

Figure 6.12 – Distribution of votes for real video and facial animation in function of the emotion stimulus.

In Section 6.4.2, the analysis of the emotion recognition test also resulted that "Reproach", "Relief" and "Gratitude" presented statistically distinguishable results between real video and facial animation. A general hypothesis is that, since the facial animation is a simplified model of visual expressive speech, the present model was not able to reproduce subtle details that would help human observers to distinguish these emotions through the visual channel alone. Considering the results of the perceptual valence evaluation, a possible explanation is that for these emotions, the visual dynamic and static cues provided by the facial animation were "attenuated" and became less efficient to help the subject to identify the emotions.

Among these emotions, "Gratitude" in particular, presents a high level of confusion among the voted options for both real video and facial animation, as illustrated in the barplots of Figure 6.9(d). It is also interesting to note that "Gratitude" was identified by some participants as a positive emotion (like "Happy For" and "Satisfaction") and by others, as a negative emotion (like "Disappointment" and "Fear"). Besides, Table 6.2 shows that "Gratitude" was voted "neutral" by a great number of subjects. The high level of confusion for

real video may be interpreted as an indicator that the performance of the actress for "Gratitude" is not typical or stereotypical and, for this reason, it is not easily recognized by the subjects. Since the expressive context-dependent visemes for "Gratitude" are extracted from performances of the same actress, this would also affect the perception of facial animation. Another hypothesis is that a "neutral" valence emotion is harder to be recognized without the knowledge of the context provided by the speech audio. "Remorse" and "'Resentment" for example, were also voted "Neutral" by almost 50% of the subjects and they also present a high level of confusion among the voted options for both real video and facial animation(see the barplots shown in Figures 6.9(b) and 6.9(c)).

6.5 Concluding Remarks

This chapter presented the results of a perceptual evaluation that included the judgement of valence and the recognition of emotions by volunteer subjects that were asked to observe and identify different emotion stimuli, presented as real video and facial animation muted video clips.

Together with the speech intelligibility test results reported in (COSTA; DE MARTINO, 2013), the results of the present evaluation validates the expressive speech animation synthesis methodology presented in Chapters 4 and 5.

From the results analyses, it is possible to conclude:

- The proposed synthesis methodology is capable of synthesizing expressive speech animation at a videorealism level that enable human observers to recognize the valence of the emotions.

- In the OCC model, the emotions are presented as complementary pairs of positive and negative valence emotions. "Pride" and "Shame", for example, is a complementary pair of emotions. "Pride" represents the positive reaction to a self agent action. "Shame" represents the complementary negative reaction. The test results showed that the perceived valence of emotions in real video and facial animation are consistent with the OCC model.

- The average emotion recognition rate achieved by the synthetic talking head cannot be statistically distinguished from the rate achieved by the real video.

- When comparing real video and facial animation confusion patterns, most emotions — in particular those that can be directly mapped to the Ekman's emotions (anger,

disgust, sadness, joy and fear) — , present distributions that cannot be statistically distinguished from each other.

The present work required the design of an emotion recognition evaluation protocol that comprehends the assessment of the twenty-two emotions of the OCC model. Similar efforts to present the results of emotion recognition assessments were made in (BESKOW; NORDENBERG, 2005) and (ANDERSON *et al.*, 2013). Beskow e Nordenberg (2005) reported the results of an emotion recognition test in which the expressive speech animation was presented with synchronized neutral speech audio. The subjects were asked to classify the expression observed in the audiovisual stimuli into one of four categories: happy, angry, sad or neutral. Ten subjects participated in the evaluation and the average recognition rate was: 73% for happy, 60% for angry and 40% for sad facial expressions. The authors do not discuss influence of speech audio on the identification of the categories. Anderson *et al.* (2013) presented the results of an emotion recognition evaluation in which the participants were presented either video or audio clips of a sentence and were asked to identify the emotion expressed by the speaker, selecting it from a list of six emotions: neutral, tender, angry, afraid, happy and sad. The synthetic stimuli were generated following the visual-text-to-speech synthesis approach proposed by the work. The results obtained by the synthetic talking face were compared with versions of synthetic video only (no audio), synthetic audio only and real video. Ten sentences in each of the six emotions were evaluated by twenty subjects. The average recognition rates were 73% for real video, 77% for facial animation, 52% for the muted facial animation and 68% for the synthetic audio only. The authors argue that the stylization of the expression in the synthesis may explain the higher recognition rate obtained by facial animation. They also report that tender and neutral expressions are most easily confused in all cases (consistent with the results discussed in Section 6.4.4). Additionally, they highlight that some emotions are better recognized from audio only, but that recognition rate is higher when using both audiovisual cues.

The analyses performed in this chapter provided feedback information that helps to point the directions for future work. First, the trajectory comparison of facial feature points of real video and facial animation is encouraged. The objective is to analyze if the statistical dissimilarities found between real video and facial animation for some emotions are correlated to different representations of expressive speech dynamics; and if the "attenuation effect" of facial animation discussed in Section 6.4.4 can be objectively verified. Second, the evaluation should be repeated for different face models in order to investigate their impact on the user perception of emotions and the robustness of the synthesis methodology. Finally, the results of the interview presented in Section 6.4.3 confirm that any effort to improve the face model should consider the enhancement of the visual quality of animation in the mouth and eye

region a priority.

7 Conclusions

The present work proposed a 2D expressive speech animation synthesis methodology, focusing on:

- the realistic synthesis of the facial appearance;
- the realistic reproduction of the speech articulatory movements;
- the synthesis of facial expressions compatible with everyday conversational interactions;
- the expression of emotions.

Chapter 2 provided a review of speech synchronized facial animation approaches and showed that the problem of modeling the speech accompanied by the expression of emotions is not trivial. First, the emotions, how they are elicited, how the humans express them, and how many emotions exist are open questions discussed for centuries by philosophers, artists, psychologists, biologists, neuroscientists and, more recently, by computer scientists and engineers. Despite the lack of a consensus around a computational model of emotions, Chapter 2 showed that many initiatives to model the expression of emotions by the face were influenced by the work of the psychologist Paul Ekman, regarding the universality of human facial expressions. The "big six" emotions of Ekman (anger, happiness, sadness, surprise, fear and disgust) are references of emotions for numerous computational systems in different research fields. Chapter 2 also discussed the limitation of such a small vocabulary of stereotypical facial expressions for the reproduction of the everyday dialogue episodes. In this sense, the present work supports the thesis that appraisal model of emotions provides a clear correlation between stimuli and emotional reaction that is compatible with the interaction model of virtual Embodied Conversational Agents (ECAs), as illustrated in Figure 2.6.

Considering this, as highlighted in Table 2.1, a first contribution of the present work is the implementation of a 2D expressive speech synthesis methodology that is based on an appraisal model of emotions. In particular, the proposed expressive speech face modeling is based on the Ortony, Clore and Collins (OCC) model of emotions. The OCC model embraces a more complex but still concise vocabulary of twenty-two emotions that arise as consequence of the appraisal analysis regarding the consequences of events, the actions of agents or the attractiveness of things. This approach enables the integration of the present solution with an intelligent system capable of simulating the appraisal process proposed by the OCC model.

For the study of the dynamics of visual expressive speech, great effort was applied to the construction of an expressive speech corpus for Brazilian Portuguese that integrates samples of high quality audio speech recordings, motion capture data and recorded video material, as described in Chapter 3. Such corpus is considered one of the contributions of the present work since it enables numerous studies regarding the expression of emotions in the speech, in the face and it also includes different personality trait performances. Among the difficulties encountered to process the corpus were the need of accurate segmentation of hours of speech audio material and the precise detection of feature points in the facial images. Considering this, just part of the recorded video material was processed and analyzed to construct the *CH-Unicamp* expressive viseme image database: an annotated database of high quality images of a female face uttering the different phones of Brazilian Portuguese accompanied by the expression of the twenty-two OCC emotions.

The *CH-Unicamp* database characterized the samples dataset used to build the 2D expressive speech face model, described in details in Chapter 4. A key aspect of the proposed model is the representation of the visemes of a language as "shape-independent" appearance coefficients. Together with the appearance model parameters derived from the samples database, such representation carries relevant visual information to discriminate both, visemes and emotions; without loosing the flexibility of being shape modulated. The proposed model is scalable, i.e. new samples can be added to the expressive speech face model database, making possible to extend and enrich the diversity of represented phonetic contexts.

In Chapter 5, the elements of the expressive speech face model were combined with other processing modules to compose a 2D expressive speech animation synthesis methodology. The proposed methodology provides a robust modeling of visible speech, including the effects of coarticulation, and it also enables the expression of the twenty-two OCC emotions. The underlying concepts of the methodology can be applied to any language and any vocabulary of emotions. The synthesis framework characterizes a rule-based synthesis approach. The rules are based on the analysis of the timed phonetic transcription of the speech to be animated and, consequently, different "voices" can be animated without any additional training or modeling steps.

In order to validate the proposed synthesis methodology, Chapter 6 presented the results of an emotion recognition perceptual evaluation. The results showed that in numerous situations, the identification of valence and the rate of recognition of emotions obtained by the synthesized animations are comparable to those obtained by the recorded videos of a real face.

In summary, the main contributions of the present work are:

- The proposal and implementation of a 2D expressive speech animation synthesis methodology that is based on an appraisal model of emotions (see Table 2.1).

- The proposal of a rule-based animation synthesis methodology compatible with the OCC model of emotions (see Table 2.1).

- The "shape-independent" modeling of expressive context-dependent visemes as a set of appearance coefficients associated to an appearance model (Chapter 4, Section 4.2).

- The creation of the *CH-Unicamp* expressive viseme image database (Chapter 3, Section 3.2).

- The building of a multimodal expressive speech corpus for Brazilian Portuguese (Chapter 3).

- The presentation of an evaluation protocol and the results of the emotion recognition perceptual evaluation presented in Chapter 6.

Additionally, the appendices to the current text present specific evaluations that were performed to assess the validity of assumptions that guided the project. Appendix B shows the results of an objective evaluation of different formulations of the active appearance model and Appendix C investigates the sensitiveness of the user perception to the dimensionality reduction of the appearance model. Such evaluations can also be considered contributions to the area since they embrace evaluation methodologies and results that throw light over very particular aspects of the studied techniques.

At the current stage of development, the synthesis methodology presented in the current work present limitations including:

- the 2D head model, constructed from frames extracted from a video corpus, only supports limited excursions of the head;

- the animation synthesis is not performed in real time;

- there is no mechanism to predict the visual prosody or to control the expressiveness intensity to be reflected in the trajectory of the head and the shape of facial elements during the animation of speech accompanied by the expression of emotions.

Among the possible approaches to overcome such limitations, the following section points out future work directions identified as able to provide promising results.

7.1 Future Work

The results of the perceptual evaluation described in Chapter 6 provided relevant feedback for the improvement of the proposed synthesis methodology. First, the results show the potential of the methodology to synthesize videorealistic expressive speech animations. Thus, the full implementation of the "shape modulator" entity described in Chapter 5 is strongly encouraged. Ideas to be explored are:

- Tagged scripts can input the shape processing module providing predefined key-shapes at particular instants of time. The key-shapes may define different levels of shape information, such as:
  - *Changes in facial elements that are independent of the mouth*: a script could be used to set the instants that the eyes should blink or the eyebrows should be frowned, for example. Such predefined actions can be implemented in the system following a rule-based approach.
  - *Changes in head orientation*: a script could be used to determine the trajectory desired for the head over time. To implement this feature it would be necessary to project the head orientation into final shapes.
- An intelligent system may be developed through the application of machine learning algorithms on the motion capture data of the same female actress available in the expressive speech corpus described in (Chapter 3). Ideally, the system would be capable of predicting the shapes trajectory based on the likely transitions learned from the speech articulation patterns recorded by the mocap system.

Further objective evaluations, like the trajectory comparison of facial feature points of real video and facial animation are also recommended. The objective is to analyze if the statistical dissimilarities found between real video and facial animation for some emotions are correlated to different representations of expressive speech dynamics and if the "attenuation effect" of facial animation discussed in Chapter 6, Section 6.4.4, can be objectively verified.

Of importance is also the study of alternative techniques to model the appearance variability inside the mouth region. All analyses performed thorough this work show that this region is subject to the worst reconstruction errors and, at the same time, it is the region cited more frequently as important to judge the expressiveness of the face (Chapter 6, Section 6.4.3).

Finally, finding strategies to make the expressive speech face model database as compact as possible while delivering high quality animations is an important aspect when considering the use of the technology in real world applications.

Bibliography

AGRESTI, A. Inference for Contingency Tables. In: *Categorical data analysis*. Second edition. [S.l.]: John Wiley & Sons, Inc., 2002. cap. 3, p. 91–101. Cited in page 105.

ALBRECHT, I.; HABER, J.; KAHLER, K.; SCHRODER, M.; SEIDEL, H.-P. " May i talk to you?:-)"-Facial Animation from Text. In: IEEE. *Computer Graphics and Applications, 2002. Proceedings. 10th Pacific Conference on*. [S.l.], 2002. p. 77–86. Cited 2 times in pages 26 and 38.

ALBRECHT, I.; HABER, J.; SEIDEL, H.-P. Speech synchronization for physics-based facial animation. In: *Proceedings of the tenth international conference in central Europe on computer graphics, visualization and computer vision*. [S.l.: s.n.], 2002. p. 9–16. Cited in page 25.

ALBRECHT, I.; SCHRÖDER, M.; HABER, J.; SEIDEL, H.-P. Mixed feelings: expression of non-basic emotions in a muscle-based talking head. *Virtual Reality*, Springer, v. 8, n. 4, p. 201–212, 2005. Cited 2 times in pages 26 and 93.

ALEXANDER, O.; ROGERS, M.; LAMBETH, W.; CHIANG, J.-Y.; MA, W.-C.; WANG, C.-C.; DEBEVEC, P. The digital emily project: Achieving a photorealistic digital actor. *Computer Graphics and Applications, IEEE*, IEEE, v. 30, n. 4, p. 20–31, 2010. Cited in page 3.

ANDERSON, R.; STENGER, B.; WAN, V.; CIPOLLA, R. Expressive Visual Text-to-Speech Using Active Appearance Models. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2013. p. 3382–3389. ISSN 1063-6919. Cited 7 times in pages 36, 38, 59, 93, 94, 119, and 151.

ARELLANO, D.; VARONA, J.; PERALES, F. J. Generation and visualization of emotional states in virtual characters. *Computer Animation and Virtual Worlds*, Wiley Online Library, v. 19, n. 3-4, p. 259–270, 2008. Cited in page 20.

ARNOLD, M. B. Emotion and personality. Columbia University Press, 1960. Cited 2 times in pages 14 and 18.

BARTNECK, C. Integrating the occ model of emotions in embodied characters. In: CITESEER. *Workshop on Virtual Conversational Characters*. [S.l.], 2002. Cited in page 45.

BELL, C. *Essays on the Anatomy and Philosophy of Expression*. [S.l.]: J. Murray, 1824. Cited in page 13.

BESKOW, J. Rule-based visual speech synthesis. In: *Proceedings of of the Fourth European Conference on Speech Communication and Technology*. [S.l.: s.n.], 1995. p. 299–302. Cited in page 25.

BESKOW, J. Trainable articulatory control models for visual speech synthesis. *International Journal of Speech Technology*, Springer, v. 7, n. 4, p. 335–349, 2004. Cited 2 times in pages 93 and 94.

BESKOW, J.; NORDENBERG, M. Data-driven synthesis of expressive visual speech using an MPEG-4 talking head. In: *INTERSPEECH*. [S.l.: s.n.], 2005. p. 793–796. Cited 5 times in pages 25, 28, 38, 94, and 119.

BEVACQUA, E.; MANCINI, M.; NIEWIADOMSKI, R.; PELACHAUD, C. An expressive ECA showing complex emotions. In: *Proceedings of the AISB annual convention, Newcastle, UK*. [S.l.: s.n.], 2007. p. 208–216. Cited 4 times in pages 4, 27, 38, and 93.

BINSTED, K.; LUKE, S. Character design for soccer commentary. In: *RoboCup-98: Robot Soccer World Cup II*. [S.l.]: Springer, 1999. p. 22–33. Cited 2 times in pages 26 and 38.

BOERSMA, P.; WEENINK, D. Praat, a system for doing phonetics by computer. *Glot International*, v. 5, n. 9/10, p. 341–347, 2001. Cited 2 times in pages 54 and 96.

BOS, E.-J. Princess Elizabeth of Bohemia and Descartes' letters (1650–1665). *Historia Mathematica*, Elsevier, v. 37, n. 3, p. 485–502, 2010. Cited in page 11.

BOSSELER, A.; MASSARO, D. W. Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. *Journal of autism and developmental disorders*, Springer, v. 33, n. 6, p. 653–672, 2003. Cited in page 2.

BOULOGNE, G.-B. D. de; CUTHBERTSON, R. A. *The mechanism of human facial expression*. [S.l.]: Cambridge university press, 1990. Cited in page 13.

BRADSKI, G.; KAEHLER, A. *Learning OpenCV: Computer vision with the OpenCV library*. [S.l.]: " O'Reilly Media, Inc.", 2008. Cited in page 54.

BRAND, M. Voice puppetry. In: ACM PRESS/ADDISON-WESLEY PUBLISHING CO. *Proceedings of the 26th annual conference on computer graphics and interactive techniques*. [S.l.], 1999. p. 21–28. ISBN 0201485605. Cited 2 times in pages 35 and 93.

BREGLER, C.; COVELL, M.; SLANEY, M. Video Rewrite: driving visual speech with audio. In: *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. [S.l.: s.n.], 1997. p. 353–360. Cited 2 times in pages 32 and 93.

BURKHARDT, F.; SCHRÖDER, M. *Emotion Markup Language (EmotionML) 1.0*. [S.l.], 2014. Http://www.w3.org/TR/2014/REC-emotionml-20140522/. Cited in page 21.

CANNON, W. B. The James-Lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology*, JSTOR, p. 567–586, 1987. Cited in page 14.

CAO, Y.; TIEN, W. C.; FALOUTSOS, P.; PIGHIN, F. H. Expressive speech-driven facial animation. *ACM Transactions on Graphics*, v. 24, n. 4, p. 1283–1302, 2005. Cited 4 times in pages 4, 33, 34, and 38.

COHEN, M. M.; MASSARO, D. W. Modeling coarticulation in synthetic visual speech. In: *Models and techniques in computer animation*. [S.l.]: Springer, 1993. p. 139–156. Cited in page 25.

COHN, J. F. Advances in behavioral science using automated facial image analysis and synthesis [social sciences]. *Signal Processing Magazine, IEEE*, IEEE, v. 27, n. 6, p. 128–133, 2010. Cited in page 1.

COOTES, T. F.; EDWARDS, G. J.; TAYLOR, C. J. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 23, n. 6, p. 681–685, 2001. ISSN 01628828. Cited 4 times in pages 59, 61, 66, and 145.

COOTES, T. F.; TAYLOR, C. J.; COOPER, D. H.; GRAHAM, J. Active shape models-their training and application. *Computer vision and image understanding*, Elsevier, v. 61, n. 1, p. 38–59, 1995. Cited 3 times in pages 62, 63, and 64.

COSATTO, E.; GRAF, H. P. Photo-realistic talking-heads from image samples. *IEEE Transactions on Multimedia*, IEEE, v. 2, n. 3, p. 152–163, 2000. ISSN 1520-9210. Cited 2 times in pages 33 and 93.

COSATTO, E.; OSTERMANN, J.; GRAF, H. P.; SCHROETER, J. Lifelike talking faces for interactive services. *Proceedings of the IEEE*, IEEE, v. 91, n. 9, p. 1406–1429, 2003. Cited 2 times in pages 1 and 94.

COSI, P.; FUSARO, A.; GRIGOLETTO, D.; TISATO, G. Data-Driven Tools for Designing Talking Heads Exploiting Emotional Attitudes. In: *Affective Dialogue Systems*. [S.l.]: Springer, 2004. p. 101–112. Cited 2 times in pages 27 and 38.

COSKER, D.; MARSHALL, D. Speech driven facial animation using a hidden Markov coarticulation model. In: *Proceedings of the 17th International Conference on Pattern Recognition*. [S.l.: s.n.], 2004. v. 1, p. 4–7. Cited 4 times in pages 35, 59, 147, and 151.

COSTA, P. D. P. *Animação facial 2D sincronizada com a fala baseada em imagens de visemas dependentes do contexto fonetico*. Master's Thesis — School of Electrical and Computer Engineering, 2009. Cited 6 times in pages 4, 45, 81, 84, 85, and 89.

COSTA, P. D. P.; DE MARTINO, J. M. Context dependent visemes: a new approach to obtain realistic 2D facial animation from a reduced image database. In: *23rd SIBGRAPI Conference on Graphics, Patterns and Images*. [S.l.: s.n.], 2010. Cited in page 6.

COSTA, P. D. P.; DE MARTINO, J. M. Compact 2D Facial Animation Based on Context-dependent Visemes. In: *Proceedings of the SSPNET 2Nd International Symposium on Facial Analysis and Animation*. New York, NY, USA: ACM, 2010. (FAA '10), p. 20–20. ISBN 978-1-4503-0388-0. Cited in page 6.

COSTA, P. D. P.; DE MARTINO, J. M. Captura de Movimento Aplicada à Pesquisa de Agentes Conversacionais Expressivos. In: *Latin Display 2012*. [S.l.: s.n.], 2012. Cited in page 51.

COSTA, P. D. P.; DE MARTINO, J. M. Assessing the Visual Speech Perception of Sampled-Based Talking Heads. In: *Proc. of the 12th International Conference on Auditory-Visual Speech Processing*. [S.l.: s.n.], 2013. Cited 5 times in pages 6, 25, 93, 94, and 118.

COSTA, P. D. P.; DE MARTINO, J. M. Expressive Talking Head for Interactive Conversational Systems. In: *9th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*. [S.l.: s.n.], 2014. Cited 2 times in pages 6 and 39.

COSTA, P. D. P.; DE MARTINO, J. M. 2D Expressive Speech Talking Head Based on the OCC Model of Emotions. In: *27th Conference on Computer Animation and Social Agents, CASA 2014*. [S.l.: s.n.], 2014. Cited in page 6.

COSTA, P. D. P.; DE MARTINO, J. M.; GOUVEIA, M. de Fátima de. Towards Interactive Conversational Talking Heads. In: *Proceedings of the 3rd Symposium on Facial Analysis and Animation*. New York, NY, USA: ACM, 2012. (FAA '12), p. 12:1–12:1. ISBN 978-1-4503-1793-1. Cited in page 39.

COURGEON, M.; MARTIN, J.-C.; JACQUEMIN, C. Marc: a multimodal affective and reactive character. In: *Proceedings of the 1st Workshop on AFFective Interaction in Natural Environments*. [S.l.: s.n.], 2008. Cited in page 20.

DARWIN, C. *The expression of the emotions in man and animals*. [S.l.]: Oxford University Press, 1998. Cited in page 14.

DE MARTINO, J. M.; MAGALHÃES, L. P.; VIOLARO, F. Facial animation based on context-dependent visemes. *Computers & Graphics*, Elsevier, v. 30, n. 6, p. 971–980, 2006. Cited 7 times in pages 6, 25, 43, 44, 45, 83, and 84.

DENG, Z.; LEWIS, J. P.; NEUMANN, U. Synthesizing speech animation by learning compact speech co-articulation models. In: IEEE. *Proceedings of Computer Graphics International 2005*. [S.l.], 2005. p. 19–25. ISBN 0780393309. ISSN 1530-1052. Cited in page 32.

DENG, Z.; NEUMANN, U.; LEWIS, J.; KIM, T.-Y.; BULUT, M.; NARAYANAN, S. Expressive Facial Animation Synthesis by Learning Speech Coarticulation and Expression Spaces. *IEEE Transactions on Visualization and Computer Graphics*, v. 12, n. 6, p. 1523–1534, 2006. ISSN 1077-2626. Cited 3 times in pages 29, 38, and 93.

DENG, Z.; NOH, J. Computer facial animation: A survey. *Data-Driven 3D Facial Animation*, 2007. Cited in page 22.

DESCARTES, R. The Passions of the Soul, 1649. *The Philosophical Writings of Descartes*, v. 1, 1989. Cited 2 times in pages 11 and 16.

DEY, P.; MADDOCK, S.; NICOLSON, R. A talking head for speech tutoring. In: ACM. *Proceedings of the SSPNET 2nd International Symposium on Facial Analysis and Animation*. [S.l.], 2010. p. 14–14. Cited in page 2.

DING, C.; XIE, L.; ZHU, P. Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications*, Springer, p. 1–18, 2014. Cited in page 36.

EDGE, J. D.; LORENZO, M. A. S.; MADDOCK, S. Reusing motion data to animate visual speech. In: *Symposium on Language, Speech and Gesture for Expressive Characters, Bath, UK*. [S.l.: s.n.], 2004. Cited in page 32.

EGGES, A.; KSHIRSAGAR, S.; MAGNENAT-THALMANN, N. Generic personality and emotion simulation for conversational agents. *Computer Animation and Virtual Worlds*, Wiley Online Library, v. 15, n. 1, p. 1–13, 2004. Cited in page 45.

EKMAN, P. Universals and cultural differences in facial expressions of emotion. In: UNIVERSITY OF NEBRASKA PRESS. *Nebraska symposium on motivation*. [S.l.], 1971. Cited 2 times in pages 14 and 20.

EKMAN, P. Basic emotions. In: DALGLEISH, T.; POWER, M. J. (Ed.). *Handbook of Cognition and Emotion*. [S.l.]: John Wiley & Sons, Ltd., 1999. Cited in page 16.

EKMAN, P.; FRIESEN, W. V. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, American Psychological Association, v. 17, n. 2, p. 124, 1971. Cited in page 16.

EKMAN, P.; FRIESEN, W. V. *Unmasking the face: A guide to recognizing emotions from facial cues*. [S.l.]: Englewood Cliffs, NJ: Prentice Hall, 1975. Cited in page 16.

EKMAN, P.; FRIESEN, W. V. *Manual for the facial action coding system*. [S.l.]: Consulting Psychologists Press, 1978. Cited 3 times in pages 16, 20, and 26.

EKMAN, P.; FRIESEN, W. V. A new pan-cultural facial expression of emotion. *Motivation and emotion*, Springer, v. 10, n. 2, p. 159–168, 1986. Cited in page 16.

EKMAN, P.; ROSENBERG, E. L. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. [S.l.]: Oxford University Press, 1997. Cited in page 16.

EVERITT, B. S. The analysis of contingency tables. In: . [S.l.]: CRC Press, 1992. cap. 1, p. 2–3. Cited in page 103.

EZZAT, T.; GEIGER, G.; POGGIO, T. Trainable videorealistic speech animation. In: *Proceedings of the 29th annual conference on computer graphics and interactive techniques*. [S.l.]: ACM, 2002. v. 21, n. 3, p. 388–398. Cited 3 times in pages 4, 35, and 93.

EZZAT, T.; POGGIO, T. Miketalk: A talking facial display based on morphing visemes. In: IEEE. *Proceedings of Computer Animation 98*. [S.l.], 1998. p. 96–102. Cited 2 times in pages 25 and 37.

FIELD, T. M.; WOODSON, R.; GREENBERG, R.; COHEN, D. Discrimination and imitation of facial expression by neonates. *Science*, American Association for the Advancement of Science, v. 218, n. 4568, p. 179–181, 1982. Cited in page 2.

FOLEY, J. D. *Computer graphics: principles and practice*. [S.l.]: Addison-Wesley Professional, 1990. Cited in page 90.

GEISSLER, L. R. Three Experimental Studies in Psychology. In: FISHER, S. W. (Ed.). *The Supremacy of Mind: A Lecture*. [S.l.]: Munsell and Tanner, printers, 1845, (Eleventh Annual Course of Lectures Before the Young Men's Association of the City of Albany). Cited in page 13.

GENDRON, M.; BARRETT, L. F. Reconstructing the past: A century of ideas about emotion in psychology. *Emotion Review*, Sage Publications, v. 1, n. 4, p. 316–339, 2009. Cited in page 14.

GLASBEY, C. A.; MARDIA, K. V. A review of image-warping methods. *Journal of applied statistics*, Taylor & Francis, v. 25, n. 2, p. 155–171, 1998. Cited 2 times in pages 65 and 90.

GOVOKHINA, O.; BAILLY, G.; BRETON, G.; BAGSHAW, P. *et al.* TDA: A new trainable trajectory formation system for facial animation. In: *Interspeech*. [S.l.: s.n.], 2006. p. 2474–2477. Cited in page 35.

GOYAL, U. K.; KAPOOR, A.; KALRA, P. Text-to-audiovisual speech synthesizer. In: SPRINGER. *Virtual Worlds*. [S.l.], 2000. p. 256–269. Cited in page 25.

HILL, D. R.; PEARCE, A.; WYVILL, B. Animating speech: an automated approach using speech synthesised by rules. *The visual computer*, Springer, v. 3, n. 5, p. 277–289, 1988. Cited in page 25.

INTERNATIONAL PHONETIC ASSOCIATION. *Handbook of the international phonetic association-a guide to the use of the international phonetic alphabet*. [S.l.]: Cambridge University Press, 1999. Cited 4 times in pages 44, 45, 83, and 84.

ITU-T RECOMMENDATION, P. Subjective video quality assessment methods for multimedia applications. 1999. Cited 2 times in pages 153 and 154.

IZARD, C. E. *Human emotions*. [S.l.]: Boom Koninklijke Uitgevers, 1977. Cited in page 14.

IZARD, C. E. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, SAGE Publications, v. 2, n. 3, p. 260–280, 2007. Cited in page 15.

JACKSON, J. E. *A User's Guide to Principal Components*. [S.l.]: John Wiley & Sons, Inc., 2003. Cited 2 times in pages 66 and 147.

JIA, J.; ZHANG, S.; MENG, F.; WANG, Y.; CAI, L. Emotional Audio-Visual Speech Synthesis Based on PAD. *IEEE Transactions on Audio, Speech & Language Processing*, v. 19, n. 3, p. 570–582, 2011. Cited 3 times in pages 29, 38, and 93.

JIANG, D.; RAVYSE, I.; SAHLI, H.; VERHELST, W. Speech driven realistic mouth animation based on multi-modal unit selection. *Journal on Multimodal User Interfaces*, Springer, v. 2, n. 3-4, p. 157–169, 2008. Cited in page 32.

JIANG, D.; ZHAO, Y.; SAHLI, H.; ZHANG, Y. Speech driven photo realistic facial animation based on an articulatory DBN model and AAM features. *Multimedia Tools and Applications*, Springer US, p. 1–19, jul. 2013. ISSN 1380-7501. Cited in page 59.

JOLLIFFE, I. *Principal component analysis*. [S.l.]: Wiley Online Library, 2002. 21–27 p. Cited in page 69.

JONES, E.; OLIPHANT, T.; PETERSON, P. *et al. SciPy: Open source scientific tools for Python*. 2001. Cited 2 times in pages 66 and 90.

KING, S. A.; PARENT, R. E. Creating speech-synchronized animation. *IEEE Transactions on visualization and computer graphics*, Published by the IEEE Computer Society, p. 341–352, 2005. Cited in page 25.

KLOKE, J.; MCKEAN, J. W. Two-Sample Problems. In: *Nonparametric Statistical Methods Using R*. [S.l.]: CRC Press, 2014, (The R Series). cap. 3. Cited in page 106.

KSHIRSAGAR, S. A multilayer personality model. In: ACM. *Proceedings of the 2nd international symposium on Smart graphics*. [S.l.], 2002. p. 107–115. Cited 5 times in pages 21, 26, 27, 38, and 45.

KSHIRSAGAR, S.; MAGNENAT-THALMANN, N. Visyllable based speech animation. *Computer Graphics Forum*, v. 22, n. 3, p. 631–639, 2003. ISSN 1467-8659. Cited in page 32.

LAZARUS, R. S. Psychological stress and the coping process. McGraw-Hill, 1966. Cited 2 times in pages 14 and 18.

LE BRUN, C. *Methode pour apprendre à dessiner les passions, proposée dans une conference sur l'expression générale, et particuliere*. [S.l.]: Chez François van-der Plaats, 1702. Cited in page 13.

LE GOFF, B.; BENOIT, C. A text-to-audiovisual-speech synthesizer for French. In: *Proceedings of Fourth International Conference on Spoken Language*. [S.l.: s.n.], 1996. v. 4, p. 2163–2164. Cited in page 25.

LEONE, G. R.; PACI, G.; COSI, P. LUCIA: An Open Source 3D Expressive Avatar for Multimodal hmi. In: *Intelligent Technologies for Interactive Entertainment*. [S.l.]: Springer, 2012. p. 193–202. Cited in page 27.

LIU, J.; YOU, M.; CHEN, C.; SONG, M. Real-time speech-driven animation of expressive talking faces. *International Journal of General Systems*, v. 40, n. 4, p. 439–455, maio 2011. ISSN 0308-1079. Cited 3 times in pages 30, 38, and 93.

LIU, K.; OSTERMANN, J. Optimization of an Image-Based Talking Head System. *EURASIP Journal on Audio, Speech, and Music Processing*, v. 2009, p. 1–13, 2009. ISSN 1687-4714. Cited 2 times in pages 33 and 59.

LIU, K.; OSTERMANN, J. Realistic facial expression synthesis for an image-based talking head. In: IEEE. *Multimedia and Expo (ICME), 2011 IEEE International Conference on*. [S.l.], 2011. p. 1–6. Cited 2 times in pages 4 and 33.

LIU, K.; OSTERMANN, J. Evaluation of an image-based talking head with realistic facial expression and head motion. *Journal on Multimodal User Interfaces*, Springer, v. 5, n. 1-2, p. 37–44, 2012. Cited in page 93.

LYONS, M. J.; AKAMATSU, S.; KAMACHI, M.; GYOBA, J.; BUDYNEK, J. *The Japanese female facial expression (JAFFE) database*. 1998. Cited in page 29.

MAGNENAT THALMANN, N.; THALMANN, D. Digital actors for interactive television. *Proceedings of the IEEE*, IEEE, v. 83, n. 7, p. 1022–1031, 1995. Cited in page 1.

MARSELLA, S.; GRATCH, J.; PETTA, P. Computational models of emotion. In: *A Blueprint for Affective Computing - A sourcebook and manual* . [S.l.: s.n.], 2010. p. 21–46. Cited 2 times in pages 15 and 21.

MASSARO, D. W.; LIGHT, J. Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of Speech, Language, and Hearing Research*, ASHA, v. 47, n. 2, p. 304–320, 2004. Cited in page 2.

MATTHEYSES, W.; LATACZ, L.; VERHELST, W. Active appearance models for photorealistic visual speech synthesis. In: *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech)*. [S.l.: s.n.], 2010. p. 1113–1116. Cited 2 times in pages 33 and 59.

MATTHEYSES, W.; VERHELST, W. Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, v. 66, n. 0, p. 182–217, 2015. ISSN 0167-6393. Cited 4 times in pages 3, 22, 23, and 35.

MCGURK, H.; MACDONALD, J. Hearing lips and seeing voices. Nature Publishing Group, 1976. Cited in page 2.

MEHRABIAN, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, Springer, v. 14, n. 4, p. 261–292, 1996. Cited in page 18.

MEHRABIAN, A.; RUSSELL, J. A. *An approach to environmental psychology.* [S.l.]: the MIT Press, 1974. Cited in page 14.

MENDI, E.; BAYRAK, C. Text-to-Audiovisual Speech Synthesizer for Children with Learning Disabilities. *Telemedicine and e-Health*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 19, n. 1, p. 31–35, 2013. Cited in page 2.

MORI, M.; MACDORMAN, K. F.; KAGEKI, N. The uncanny valley [from the field]. *Robotics & Automation Magazine, IEEE*, IEEE, v. 19, n. 2, p. 98–100, 2012. Cited 2 times in pages 2 and 3.

NOH, J.-y.; NEUMANN, U. *A survey of facial modeling and animation techniques*. [S.l.], 1998. Cited in page 22.

OCHS, M.; NIEWIADOMSKI, R.; PELACHAUD, C.; SADEK, D. Intelligent expressions of emotions. In: *Affective computing and intelligent interaction*. [S.l.]: Springer, 2005. p. 707–714. Cited in page 28.

ÖHMAN, S. E. G. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, v. 41, n. 2, p. 310–320, 1967. Cited in page 25.

ORTONY, A.; CLORE, G.; COLLINS, A. *Cognitive Structure of Emotions*. [S.l.]: Cambridge University Press, 1988. Cited 6 times in pages 6, 14, 18, 19, 46, and 47.

ORTONY, A.; TURNER, T. J. What's basic about basic emotions? *Psychological review*, American Psychological Association, v. 97, n. 3, p. 315, 1990. Cited 2 times in pages 15 and 20.

OUNI, S.; COHEN, M. M.; ISHAK, H.; MASSARO, D. W. Visual contribution to speech perception: Measuring the intelligibility of animated talking heads. *EURASIP Journal on Audio, Speech, and Music Processing*, Hindawi Publishing Corp., v. 2007, 2007. Cited in page 93.

PANDZIC, I. S. Facial animation framework for the web and mobile platforms. In: ACM. *Proceedings of the seventh international conference on 3D Web technology*. [S.l.], 2002. p. 27–34. Cited in page 1.

PANDZIC, I. S.; FORCHHEIMER, R. *MPEG-4 facial animation*. [S.l.]: Wiley Online Library, 2002. Cited 2 times in pages 26 and 54.

PARKE, F. I. Computer generated animation of faces. In: ACM. *Proceedings of the ACM annual conference - Volume 1*. [S.l.], 1972. p. 451–457. Cited 3 times in pages 1, 4, and 21.

PARKE, F. I.; WATERS, K. *Computer facial animation*. [S.l.]: AK Peters Wellesley, 1996. Cited in page 2.

PELACHAUD, C.; BADLER, N. I.; STEEDMAN, M. Generating facial expressions for speech. *Cognitive Science*, Elsevier, v. 20, n. 1, p. 1–46, 1996. ISSN 0364-0213. Cited 2 times in pages 25 and 38.

PELACHAUD, C.; BILVI, M. Computational model of believable conversational agents. In: *Communication in multiagent systems*. [S.l.]: Springer, 2003. p. 300–317. Cited 2 times in pages 27 and 38.

PELTOLA, M. J.; LEPPÄNEN, J. M.; MÄKI, S.; HIETANEN, J. K. Emergence of enhanced attention to fearful faces between 5 and 7 months of age. *Social Cognitive and Affective Neuroscience*, Oxford University Press, p. nsn046, 2009. Cited in page 2.

PICARD, R. W. Affective computing. Citeseer, 1995. Cited in page 14.

PIGHIN, F.; HECKER, J.; LISCHINSKI, D.; SZELISKI, R.; SALESIN, D. H. Synthesizing realistic facial expressions from photographs. In: ACM. *ACM SIGGRAPH 2006 Courses*. [S.l.], 2006. p. 19. Cited in page 24.

PLUTCHIK, R. *Emotions: A general psychoevolutionary theory*. [S.l.: s.n.], 1984. 197–219 p. Cited in page 17.

PLUTCHIK, R. *The emotions*. [S.l.]: University Press of America, 1991. Cited in page 14.

PLUTCHIK, R. The nature of emotions. *American Scientist*, v. 89, n. 4, p. 344–350, 2001. Cited in page 17.

QUEIROZ, R. B.; COHEN, M.; MUSSE, S. R. An extensible framework for interactive facial animation with facial expressions, lip synchronization and eye behavior. *Computers in Entertainment (CIE)*, ACM, v. 7, n. 4, p. 58, 2009. Cited 2 times in pages 28 and 38.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2014. Cited in page 103.

REVÉRET, L.; BAILLY, G.; BADIN, P. MOTHER: A new generation of talking heads providing a flexible articulatory control for video-realistic speech animation. In: *Proceedings of Sixth International Conference on Spoken Language Processing*. [S.l.: s.n.], 2000. v. 2, p. 755–758. Cited in page 25.

RUSSELL, J. A.; MEHRABIAN, A. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, Elsevier, v. 11, n. 3, p. 273–294, 1977. Cited in page 18.

SCHERER, K. R. Appraisal theory. In: *Handbook of cognition and emotion*. [S.l.: s.n.], 1999. p. 637–663. Cited 4 times in pages 14, 15, 18, and 45.

SCHERER, K. R. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, v. 92, p. 120, 2001. Cited in page 19.

SCHLOSBERG, H. Three dimensions of emotion. *Psychological review*, American Psychological Association, v. 61, n. 2, p. 81, 1954. Cited in page 18.

SCOTT, K. C.; KAGELS, D. S.; WATSON, S. H.; ROM, H.; WRIGHT, J. R.; LEE, M.; HUSSEY, K. J. Synthesis of speaker facial movement to match selected speech sequences. In: *Proceedings of the fifth australian conference on speech science and technology*. [S.l.: s.n.], 1994. v. 2, p. 620–625. Cited 2 times in pages 25 and 37.

SHAPIRO, S. S.; WILK, M. B.; CHEN, H. J. A comparative study of various tests for normality. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 63, n. 324, p. 1343–1372, 1968. Cited in page 107.

SHEIKH, H. R.; BOVIK, A. Image information and visual quality. In: *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*. [S.l.: s.n.], 2004. v. 3, p. iii–709–12 vol.3. ISSN 1520-6149. Cited in page 149.

STEGMANN, M. B. *Active appearance models: Theory, extensions and cases*. 262 p. Master's Thesis, 2000. Cited 5 times in pages 62, 65, 66, 146, and 151.

STEUNEBRINK, B. R.; DASTANI, M.; MEYER, J.-J. C. The OCC model revisited. In: *Proceedings of the 4th Workshop on Emotion and Computing*. [S.l.: s.n.], 2009. v. 65, p. 2047–2056. Cited in page 45.

SUMBY, W.; POLLACK, I. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, Acoustical Society of American, v. 26, p. 212–215, 1954. Cited in page 93.

TAO, J.; XIN, L.; YIN, P. Realistic visual speech synthesis based on hybrid concatenation method. *Audio, Speech, and Language Processing, IEEE Transactions on*, IEEE, v. 17, n. 3, p. 469–477, 2009. Cited 3 times in pages 35, 38, and 93.

THEOBALD, B.-J.; BANGHAM, J. A.; MATTHEWS, I. A.; CAWLEY, G. C. Near-videorealistic synthetic talking faces: Implementation and evaluation. *Speech Communication*, Elsevier, v. 44, n. 1, p. 127–140, 2004. Cited in page 32.

THEOBALD, B.-J.; MATTHEWS, I.; MANGINI, M.; SPIES, J. R.; BRICK, T. R.; COHN, J. F.; BOKER, S. M. Mapping and manipulating facial expression. *Language and speech*, SAGE Publications, v. 52, n. 2-3, p. 369–386, 2009. Cited 2 times in pages 24 and 76.

TINWELL, A.; GRIMSHAW, M.; NABI, D. A.; WILLIAMS, A. Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human Behavior*, Elsevier, v. 27, n. 2, p. 741–749, 2011. Cited in page 2.

TSAPATSOULIS, N.; RAOUZAIOU, A.; KOLLIAS, S.; COWIE, R.; DOUGLAS-COWIE, E. Emotion recognition and synthesis based on MPEG-4 FAPs. In: *MPEG-4 Facial Animation*. [S.l.: s.n.], 2002. p. 141–167. Cited 2 times in pages 26 and 38.

TURING, A. M. Computing machinery and intelligence. *Mind*, JSTOR, p. 433–460, 1950. Cited in page 93.

TURK, M.; PENTLAND, A. Eigenfaces for recognition. *Journal of cognitive neuroscience*, MIT Press, v. 3, n. 1, p. 71–86, 1991. Cited 2 times in pages 59 and 70.

WANG, L.; SOONG, F. K. HMM trajectory-guided sample selection for photo-realistic talking head. *Multimedia Tools and Applications*, Springer, p. 1–21, 2014. Cited in page 35.

WANG, Z.; BOVIK, A.; SHEIKH, H.; SIMONCELLI, E. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, v. 13, n. 4, p. 600–612, April 2004. ISSN 1057-7149. Cited in page 149.

WEHRLE, T.; KAISER, S. Emotion and facial expression. In: *Affective interactions*. [S.l.]: Springer, 2000. p. 49–63. Cited 2 times in pages 20 and 21.

WEISER, M. The Computer for the 21st Century. *Scientific American*, v. 265, p. 94–104, 1991. Cited in page 1.

WHISSELL, C. The dictionary of affect in language. *Emotion: Theory, research, and experience*, New York, NY: Academic Press, v. 4, n. 113-131, p. 94, 1989. Cited in page 18.

WINN, M.; RHONE, A.; CHATTERJEE, M.; IDSARDI, W. The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Frontiers in Psychology*, v. 4, n. 824, 2013. ISSN 1664-1078. Cited in page 23.

WOLBERG, G. Image morphing: a survey. *The Visual Computer*, v. 14, n. 8/9, p. 360–372, 1998. Cited in page 88.

ZHANG, S.; WU, Z.; MENG, H. M.; CAI, L. Facial expression synthesis using PAD emotional parameters for a Chinese expressive avatar. In: *Affective Computing and Intelligent Interaction*. [S.l.]: Springer, 2007. p. 24–35. Cited in page 20.

ZHANG, S.; WU, Z.; MENG, H. M.; CAI, L. Facial Expression Synthesis Based on Emotion Dimensions for Affective Talking Avatar. In: *Modeling Machine Emotions for Realizing Intelligence*. [S.l.]: Springer, 2010. p. 109–132. Cited 3 times in pages 29, 31, and 38.

ZHANG, Y.; JI, Q.; ZHU, Z.; YI, B. Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters. *Circuits and Systems for Video Technology, IEEE Transactions on*, IEEE, v. 18, n. 10, p. 1383–1396, 2008. Cited in page 24.

ZHAO, W.; CHELLAPPA, R.; PHILLIPS, P. J.; ROSENFELD, A. Face Recognition: A Literature Survey. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 35, n. 4, p. 399–458, dez. 2003. ISSN 0360-0300. Cited in page 59.

ZHOU, C.; LIN, X. Facial expressional image synthesis controlled by emotional parameters. *Pattern Recognition Letters*, Elsevier, v. 26, n. 16, p. 2611–2627, 2005. Cited in page 24.

APPENDIX A – Brazilian Portuguese Texts used during the OCC Emotion Experiment

Joy

Lucas! Tulha! Puxa! Mas que convite maravilhoso... Preciso contar imediatamente para o Lilo! Então quer dizer que a Luciana será minha afilhada? Esteja convicto que serei um(a) ótimo(a) padrinho (madrinha)! Sempre que ela precisar, poderá me chamar. Estarei disponível vinte e quatro horas para ajudá-la em qualquer caso. E quando ela quiser dar uma fugida dos pais, ela vai dormir na nossa casa. Terei sempre um quarto muito lindo esperando por ela. Juro que cumprirei meu papel com muito carinho e dedicação!

Sadness

Lucas! Tulha! Puxa...estou arrasada! Não sei nem como dar a notícia para o Lilo e a Luciana. Juro que jamais imaginei que acabaria assim. Sinceramente não consigo entender porque eles não nos deixaram adotá-la como filha... Nós até já tínhamos preparado um quarto lindo para ela... Cuidaríamos dela com muito carinho e dedicação. Estaríamos disponíveis vinte e quatro horas do dia, prontos para ajudá-la em qualquer situação. E ela nunca mais precisaria fugir de casa, pois ela finalmente teria um lar.

Happy For

Lucas! Tulha! Puxa! Que notícia excelente! Há meses eu vejo Rafaela e suas filhas preocupadas e assustadas, correndo alucinadas para diferentes lugares! Sempre em busca do melhor tratamento para curar o tumor do Lilo. A ajuda do Xulhu finalmente fez esta situação chata acabar. Não é à toa que ele é um oncologista tão importante. Agora finalmente eles poderão arrumar as malas para voltar para casa. E certamente andarão na rua com orgulho por terem vencido esta batalha.

Pity

Luku! Tulha! Puxa, que fim triste... Rafaela e suas filhas devem estar arrasadas. Nem a ajuda de um ilustre médico, como Xulhu, foi suficiente para encontrar um que tratamento fizesse o tumor da Lila regredir. Jamais imaginei que esta situação acabaria desta maneira chata. Agora elas terão que arrumar as malas e voltar tristes para casa. Talvez nunca mais as veremos andando pela rua felizes e com orgulho.

Gloating

Lucas! Tulha! O pior já passou! Finalmente eu venci! Nem a ilustre ajuda do Xulhu acabou com esta situação! O caso foi fechado e nunca mais será apurado ou passado a limpo. Elas estão afundadas em acusações. quela chata da Rafaela e suas filhas nunca mais andarão com orgulho pela rua. Elas estarão sempre assustadas arrumando as malas. Fugindo alucinadamente... Cada dia em um novo lugar.

Resentment

Lucas! Tulha! O pior aconteceu e agora acabaram as chances delas serem presas. Com a ajuda daquele ilustre advogadozinho Xulhu, o caso foi supostamente apurado e passado a limpo. Todas as acusações foram retiradas. Agora, aquela charlatã da Rafaela e suas filhas andarão na rua com orgulho, esnobando a liberdade na nossa cara. Eu preferia vê-las de malas na mão, assustadas e alucinadas, enquanto fugiam daqui. Puxa! Bem que eu desconfiava que o rapaz não ia cumprir o combinado e conseguir um chalé desocupado na data que eu pedi... Com esta previsão de chuva, nosso final de semana está acabado!

Hope

Lucas, nestes números pus a esperança de mudar radicalmente a minha vida! Se eu ganhar este prêmio, juro que tudo que já sofri batalhando para sobreviver fará parte do passado. Finalmente vou dar a festa de aniversário que o Lilo sempre me pediu com tanto carinho! Vou ter uma linda casa com um quintal enorme para ele brincar. Vou poder cuidar da saúde do meu pai e ajudar meus irmãos. Nunca mais vou precisar fugir das visitas, com vergonha do barraco entulhado de lixo onde moro. Vou comprar um carro e nunca mais andar de ônibus. E então, vou realizar meu maior sonho: abrir a melhor oficina de carros antigos da cidade.

Fear

Lucas... Tulha... Estou muito preocupado... Se não conseguirmos este contrato, tudo que realizei e pelo qual batalhei nesta vida pode ser arrasado. Sem este contrato ficarei sem dinheiro para pagar o que devo para o Lilo e o Juliano. Eles me tomarão a casa e o carro. Nunca mais poderei olhar com orgulho para minha família. E o pior é que já passei por dificuldades no passado e sei que nessas horas, muitos dos que se dizem meus amigos, simplesmente sumirão... Sei que estarei sozinho e não terei para quem pedir ajuda.

Satisfaction

Que ótima notícia! Com esta previsão de chuva, se não tivéssemos um lugar bom para ficar, nosso final de semana ficaria arruinado! Pessoal, vamos para a praia! O rapaz cumpriu o combinado e conseguiu um chalé desocupado! Lilo arrume suas malas! Lucas, coloque comida no aquário do Pipo! Não esqueçam o guarda-sol! Juro que desta vez não vou ficar todo queimado!

Fears Confirmed

Puxa! Bem que eu desconfiava que o rapaz não ia cumprir o combinado e desocupar o chalé na data que eu pedi... Com esta previsão de chuva, nosso final de semana está acabado. Não temos outra solução a não ser desfazer as malas. Lilo, me ajude a tirar o guarda-sol do carro. Vamos entulhar esta tralha tudo de novo no quartinho. Lucas, coloque o aquário do Pipo de volta no quintal. Vou ligar rápido para seu pai antes que ele saia mais cedo do trabalho. Juro que ele vai surtar com a notícia.

Relief

Lucas! Quer dizer que o transplante da Tulha já acabou? Ela está acordada?! Ufa! Você não sabe o alívio que isto me dá! Nunca fiquei tão alucinado esperando uma notícia! Ai que notícia boa! Eu preciso contar imediatamente prá minha filha! Há dias que ela não come de tão assustada e preocupada que a pobre estava. Isto é incrível! Não é à toa que o Xulhu é um médico tão ilustre. Agora, em breve, finalmente poderemos arrumar as malas e voltar PARA casa. Andaremos na rua com muito orgulho por termos vencido esta batalha.

Disappointment

Lucas, você tem certeza? Puxa... Não acredito que o resultado do exame da Tulha ainda não saiu. Vou ligar agora mesmo para o Xulhu. É impressionante que um médico tão ilustre nos deixe tantos dias neste suspense para saber se o tumor é maligno ou não. Minha filha já está preocupada e assustada com esta situação chata. Precisamos ter uma resposta rápida para iniciar o tratamento o quanto antes se for necessário. Não vemos a hora deste pesadelo acabar. Só queria arrumar as malas e voltar para casa.

Pride

Luku, modéstia à parte, o meu pudim de limão é um dos mais elogiados da família. O meu filho Lilo simplesmente adora! Apesar de ter uma calda açucarada ele deixa no fundo da língua um leve tom cítrico que causa um imenso prazer degustativo. Juro que você vai se arrepender de experimentar... Nunca mais você vai dizer que não gosta deste tipo de sobremesa. E se eu fosse você, já punha logo um pedaço numa vasilha para levar para casa para a Alice experimentar.

Shame

Luku! Puxa! Sei que as quartas-feiras são reservadas para sua turma, mas eu ouvi a Alice comentar que você estaria viajando até junho e então assumi que não haveria problema se eu viesse arrumar esta minha tralha que está entulhada há mais de um ano neste armário. Eu deveria ter seguido o procedimento e agendado primeiro com o Lilo por telefone. Juro que vou tirar imediatamente minhas coisas daqui e em dez minutos a bancada estará arrumada para seus alunos. Sinto muito ter atrasado sua aula. Isto nunca mais se repetirá. Peço apenas que no final do dia eu possa ficar com a chave para acabar o que comecei.

Admiration

Luku, realmente fiquei admirada com a conquista da Alice. Puxa filho, juro que você precisava ter visto! O Lilo é um ótimo professor! Esta audição de Paulínia nunca será esquecida. No palco estava ela com um lindo tutu de tule azul. No fundo da platéia toda sua família. Não se ouvia um grunhido sequer. Quando a música começou, parecia que um anjo tinha descido dos céus para dançar. Quando acabou, vi o Juliano coberto de lágrimas. Ninguém duvidou que era ela quem deveria ter vencido.

Reproach

Luku, não acredito que você estacionou o carro na vaga para deficientes! É um abuso! Eu supunha que a multa restringiria esse tipo de comportamento, mas percebi que de nada adianta. Juro que eu prefiriria ter colocado meu carro lá na ladeira, ou em qualquer outro buraco e ficar tranquilo com minha consciência. Você já imaginou se seu filho andasse de cadeira de rodas e você precisasse levá-lo urgente para uma consulta? Você não tem idéia de como a vida destas pessoas fica difícil no seu dia-a-dia. Se todos tivessem consciência de seu dever como cidadão as coisas seriam muito mais simples.

Gratification

Luku! Meu filho, a Tulha passou no vestibular! Grite na rua! Liga prá toda a família! Puxa, nunca me senti tão realizada... Só Deus sabe quantos anos de luta e sacrifício para que ela conseguisse estudar. Quantas noites mal dormidas... Acordando de madrugada PARA caminhar vendendo churros de porta em porta. Mas agora tudo isso faz sentido! Minha filha será uma ilustre e linda médica!

Remorse

Lucas! Puxa vida, você não sabe o remorso que estou sentindo pelas consequências desastrosas do meu ato... Eu estava na fila do ônibus na hora do rush. O barulho era uma loucura. e repente, alguém quis furar a fila e eu inventei de reclamar. Meus amigos Danilo e Ciro, tentaram me acalmar. Como a pessoa nem ligou para minhas reclamações comecei a falar em alta voz com agressões verbais. Foi a fagulha que faltava. O dito cujo avançou sobre mim, na base do chute e com uma faca na mão. Tive que me defender como pude. Na luta, levei chutes e murros. Por isso, depois de tanto apanhar, só me restou como alternativa a fuga. O pior é saber que fui eu quem provocou esta situação chata.

Gratitude

Lula! Puxa, obrigada por vir me ajudar! Sem você eu não conseguiria lavar toda essa sujeira do xixi da gata no tapete. O filho do Luku já está quase chegando com o carro. Meu cabelo cheira a alho, minhas unhas estão cheias de terra, tenho um quilo de roupas para passar e agora o chuveiro está queimado. Meu único consolo é ter uma amiga incrível como você ao meu lado.

Anger

Xuxa, você fez xixi na sala? Gata dos infernos! Bem que você poderia fugir para a rua e nunca mais voltar! Ai que ódio! Lula, olha isso! Meu cabelo cheira a alho, minhas unhas estão cheias de terra e tenho dez quilos de roupas para passar! E o pior, agora vou ter que lavar toda essa sujeira no tapete. Eu simplesmente não acredito que o filho do Luku já está chegando com o carro e eles vão me ver neste estado.

Love

Filha, quando olho para você e vejo essa fagulha de vida que brilha nos seus olhos, juro que me sinto agraciada por um milagre. Quando vejo como pula com as perninhas no ar, sem nunca sair do lugar, fico com vontade de te imitar. Nada me faz mais criança que você. Seu riso divertido é som que invade a casa e deixa tudo colorido. Quando você chora, é chuva lá fora. Luto para te entender, pois sei que o choro é fala. Nem em meus sonhos eu poderia imaginar que você seria tão linda e nem que eu me apaixonaria tão perdidamente. Olha bem para mim, sou eu quem vai te amar infinitamente e incondicionalmente pelo resto da sua vida.

Hate

Luku, credo! Puxa, que gosto horrível! Ainda bem que eu não chamei meu sobrinho para almoçar em casa hoje. Juro que essa foi a última vez que preparo dobradinha. Além de eu odiar o aspecto da carne, o futum na cozinha ficou insuportável durante a manhã toda. Coloquei tudo que você pode imaginar para deixar qualquer carne deliciosa. E mesmo temperando com alho, salsa, cebolinha e colocando quase um quilo de batata e tomate no molho, o gosto continua péssimo. Vou mesmo é comer saladinha de chuchu com rúcula e rabanete para ver se tira este gosto ruim da boca.

APPENDIX B – Comparative Study of AAMs Applied to the Synthesis of Facial Images

B.1 Introduction

The results of a systematic comparison of different formulations of AAMs (Active Appearance Model), with focus on facial images, are presented. The study was conducted to investigate how the AAM modeling could be adapted to increase the resolution and to improve the blurred aspect typically observed inside the mouth region of facial images synthesized by AAM. The study was performed as part of the development process of the facial modeling methodology presented in Chapter 4. Further references of AAM are found in the work of Cootes *et al.* (2001) and in Chapter 4.

B.2 AAM Background

Three different AAM formulations are compared: the independent AAM ($iAAM$), the combined AAM ($cAAM$) and the combined AAM based on the correlation matrix of the feature sets ($corrAAM$).

B.2.1 Independent AAM (*iAAM*)

The independent AAM is called this way because the correlation between the shape and appearance parameters is not taken into consideration. While the face shape is defined by a set of landmarks in the face, the face appearance, sometimes referred as the face texture, is represented by the intensities of the pixels that lie inside the facial region.

The modeling process can be summarized as follows:

- the coordinates of k feature points are obtained from an image i of the training set to define the shape vector $\mathbf{s} = [x_{i1}, y_{i1}, ..., x_{ik}, y_{ik}]^T$;

- the shapes are aligned (see Chapter 4, Section 4.1.1) and a reference mean shape $\bar{\mathbf{s}}$ is computed;

- PCA (Principal Components Analysis) is used to derive a compact model that is represented by the linear combination of a mean shape $\bar{\mathbf{s}}$ and n eigenvectors $\mathbf{s_i}$:

$$\mathbf{s} = \bar{\mathbf{s}} + \sum_{i=1}^{n} u_i \mathbf{s_i} \tag{B.1}$$

  where u_i are the shape parameters that allow the reproduction of shape variations such as changes in the head pose or the reproduction of different facial expressions.

- the training images are warped to the reference shape $\bar{\mathbf{s}}$ and the intensities of the normalized shape image form the appearance vector $\mathbf{a}$.

- another PCA is performed to obtain the appearance model $\mathbf{a}$ that is the linear combination of the mean facial appearance $\bar{\mathbf{a}}$ and m eigenvectors $\mathbf{a_i}$:

$$\mathbf{a} = \bar{\mathbf{a}} + \sum_{i=1}^{m} v_i \mathbf{a_i} \tag{B.2}$$

  where v_i are the appearance parameters.

B.2.2 Combined AAM (*cAAM*)

Taking into consideration the correlation between shape and appearance, a third PCA can be performed on the concatenated shape and appearance parameters vector, $\mathbf{b}$, of the training set:

$$\mathbf{b} = \begin{bmatrix} \mathbf{W_s u} \\ \mathbf{v} \end{bmatrix} \tag{B.3}$$

The matrix $\mathbf{W_s}$ is a diagonal matrix of weights for each shape parameter, coping with the difference in units between the shape and appearance models.

The resulting model has the form:

$$\mathbf{b} = \sum_{i=1}^{r} w_i \mathbf{f_i} \tag{B.4}$$

where w_i are the combined model weights and $\mathbf{f_i}$ are the r eigenvectors resulting from the PCA analysis. Since the shape and appearance parameters have zero mean, there is no mean offset in the model. More details about the typical implementation of the combined model formulation and how to compute the matrix $\mathbf{W_s}$ can be found in the work of Stegmann (2000).

B.2.3 Combined AAM Based on the Correlation Matrix (*corrAAM*)

While not frequently explored, an alternative method to obtain a combined model is to concatenate both shape and appearance variables into one single vector and perform the PCA on the correlation matrix of these observations (that is equivalent to perform the PCA on the standardized data). In this case, for each database image the combined vector $\mathbf{c_i}$, is built:

$$\mathbf{c_i} = \begin{bmatrix} \mathbf{s_i} \\ \mathbf{a_i} \end{bmatrix} \tag{B.5}$$

A detailed description of the application of PCA with correlation matrices can be found in (JACKSON, 2003).

B.3 Training Database and Reference Images

The *CH-Unicamp* database was adopted as the training dataset to build the AAM models (Chapter 3, Section 3.2).

As discussed in Chapter 3, the video material of the female actress performance used to build *CH-Unicamp* database was divided in two parts. The first part was used to extract the visemes to build *CH-Unicamp* database. The complementary part was reserved for tests, characterizing a test dataset. Considering this, four video excerpts were selected from the test dataset, corresponding to four different emotional states: “Happy For”, “Anger”, “Sadness” and “Disgust”. Each video sequence is approximately two seconds, or 60 frames, long . The frames from the video excerpts are considered the reference frames for comparison purposes. The shape and appearance parameters for each frame of the reference videos were used as inputs to the different synthesis models to be compared. In order to obtain the shape parameters, the images were semi-automatically annotated using the same procedure applied to build the *CH-Unicamp* database.

B.4 Full Face versus Piecewise Modeling

This section compares the impact of the full face modeling versus a piecewise approach modeling. The full face modeling consists of considering the appearance of the face as whole, as shown in Figure B.1(a). In the piecewise approach, the face image is divided into four regions: forehead; eyes and eyebrows; cheeks and nose; and mouth (Figure B.1(b)). An individual AAM is built for each region. This division was inspired, for example, by Cosker e Marshall (2004).

(a) (b)

Figure B.1 – Full face versus piecewise modeling.

In order to assess the quality of the synthesized facial appearance the shape-free appearance ($\bar{\mathbf{s}}$) synthesized by the AAM was analyzed, before the warping step to the final synthesized shape. This approach enables to isolate the errors that are due to the deficient modeling of the appearance from the errors that can be attributed to the shape modeling or the final warping process. In the present study, all the eigenvectors resulting from the PCA analysis are kept.

The metric adopted to assess the results of different models is the reconstruction error. In this case, the reconstruction error refers to how close the synthesized face is to the reference facial image (both warped to the reference shape $\bar{\mathbf{s}}$).

In this section we present the reconstruction error as the mean square error (MSE) between a reference appearance $\mathbf{A}$ and a reconstructed appearance $\hat{\mathbf{A}}$:

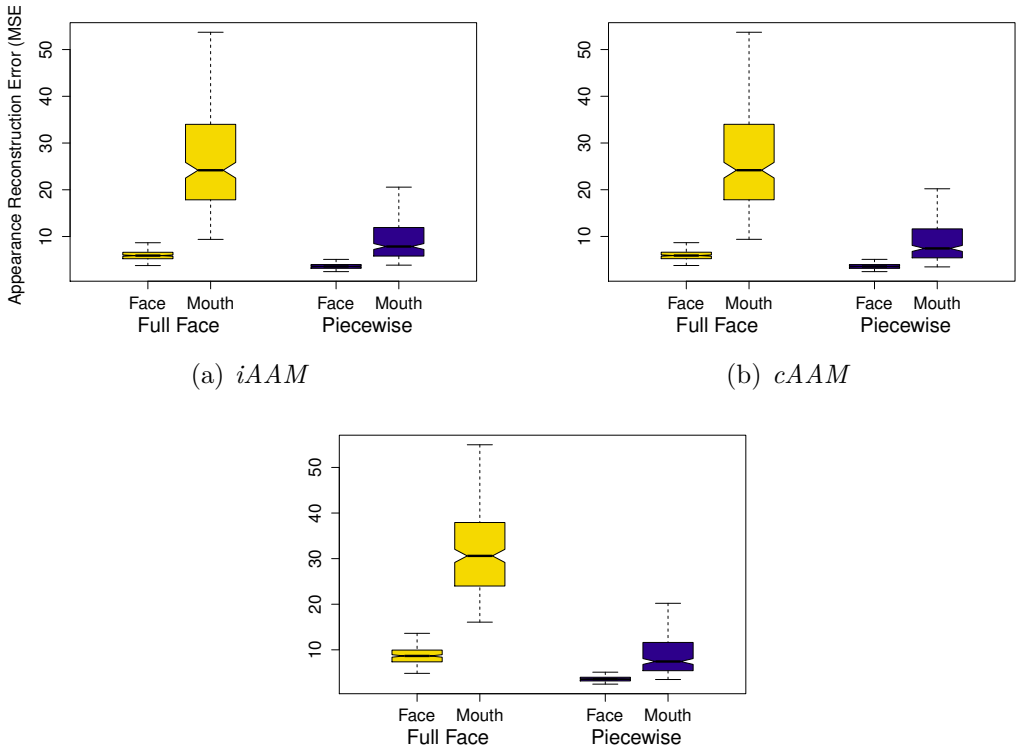
$$MSE = \frac{1}{M} \sum_{\mathbf{p} \in \bar{\mathbf{s}}} [A(\mathbf{p}) - \hat{A}(\mathbf{p})]^2 \tag{B.6}$$

where $\mathbf{p} \in \bar{\mathbf{s}}$ denotes the set of pixels $\mathbf{p} = (x, y)$ inside the reference shape to which the images are warped to build the AAM model and M is the total number of pixels inside this region.

The boxplots of Figure B.2 show the reconstruction error for the full face and the piecewise modeling approaches for the *iAAM*, *cAAM* and *corrAAM* building methods. For each case, the boxplots also highlights the MSE measured in the mouth region.

Considering the full face modeling, all the tested cases presented a mean reconstruction error in the mouth region approximately 50% greater than the average error observed in the face as whole, coherently with the visual perception that the AAM synthesis lacks resolution in the region inside the mouth (Figure B.3).

The two boxplots on the right of graphs shown in the Figure B.2(a), Figure B.2(b)



(a) *iAAM*

(b) *cAAM*

(c) *corrAAM*

Figure B.2 – The piecewise modeling contributes to reduce ($p < 0.001$) the reconstruction error when the *iAAM* and the *corrAAM* building methods are applied. The same effect is not observed when using the *cAAM* method.

and Figure B.2(c), show that the piecewise modeling contributes to reduce the reconstruction error of the synthesized face appearance. In fact, when adopting the piecewise approach, the reconstruction error not only dropped significantly but also the MSE difference between the mouth region and the whole face becomes comparable (t-paired test: $p < 0.001$).

B.5 Comparing Different AAM Formulations

Focusing on the piecewise modeling approach, Table B.1 presents the mean reconstruction error computed between the original frame images extracted from the corpus and the final synthesized faces. Besides MSE, Table B.1 also shows the mean reconstruction error using the structural similarity index (SSIM) (WANG *et al.*, 2004) and the visual information fidelity (VIF) visual quality metric (SHEIKH; BOVIK, 2004). The metrics were computed using the Python package PyMetrikz.

Independently of the metric, all three formulations are similar in terms of the obtained reconstruction error.

Figure B.3 – Synthesized faces using the *iAAM* formulation. The face on the right was synthesized using the full face modeling approach. It is possible to observe the blurred aspect inside the mouth region. The face on the left was synthesized through the piecewise modeling approach.

	iAAM	*cAAM*	*corrAAM*
MSE	3.589	3.570	4.778
SSIM	0.9972	0.9972	0.9965
VIF	0.8393	0.8396	0.8142

Table B.1 – Reconstruction errors for piecewise modeling.

	Time (min)
iAAM	29.8
cAAM	70.7
corrAAM	3.9

Table B.2 – Average time to compute the model.

On the other hand, Table B.2 shows the average time needed to compute each model taking a Dell Vostro 3550 Notebook platform as reference. We observe that the *corrAAM* is more than 15 times faster to compute than the *cAAM*. This difference can be easily understood from the fact that the *cAAM* involves multiple concatenated PCAs and the intermediate step of computing the weights of the training database samples.

B.6 Conclusion

The study explored different methods of building AAMs and observed their impact on the synthesized facial images taking the reconstruction error as an objective metric for comparison purposes.

The typical implementation of the combined AAM (*cAAM*), as adopted in the works

of Cosker e Marshall (2004) and Anderson *et al.* (2013)), is partly influenced by the development history of the AAM as an evolution of the active shape models (ASM) technique (STEGMANN, 2000). It was showed that alternative formulations deliver similar results in terms of the visual quality of synthesized faces. On the other hand, the *corrAAM* modeling can be considered an advantageous implementation since its computation is significantly faster than the *cAAM*.

Additionally, this study contributes with objective results showing that the resolution of the whole face and inside the mouth region can be improved if the piecewise modeling strategy is applied. This strategy was explored, for example, by Cosker e Marshall (2004).

Finally, while the different objective metrics applied lead to the same conclusions, it is understood that the present study can be complemented and enriched by subjective evaluation results.

APPENDIX C – Evaluating the Impact of Dimensionality Reduction on the Perceived Image Quality

C.1 Introduction

The dimensionality reduction mechanism offers advantages for the implementation of the expressive speech face model.

Consider, for example, a full face ROI containing approximately 120k pixels (approx. 300 x 400 pixels). A unique full face PC vector is then composed of $3 \times 120 \times 10^3$ elements (considering three color planes RGB), occupying circa 2.9Mb of memory (24 bits/pixel). Thus, the dimensionality reduction from $m = 782$ to, for example, $l = 300$ principal components, signifies a memory usage reduction of more than 1Gb. Additionally, since the appearance reconstruction is a linear combination of PCs (Equation 4.19), the reduction of dimensionality also improves the speed of reconstruction.

This section presents the results of a subjective evaluation conducted to investigate the impact of appearance model dimensionality reduction, from the user perception point of view.

The designed evaluation was inspired on the simultaneous presentation of sequence pairs assessment methodology described on the International Telecommunication Union (ITU) recommendation "P.910: Subjective video quality assessment methods for multimedia applications" (ITU-T RECOMMENDATION, 1999). The methodology is based on the presentation, side by side, of two synchronized videos: a reference video and a version of the same video to be evaluated(Figure C.1). The observers are asked to evaluate the image quality of the second video compared to the reference.

In the present evaluation, the reference video is characterized by video clips generated from a sequence of reconstructed frames using all the principal components of the appearance model, i.e. the full quality reconstructed version. Deteriorated versions of the same stimulus, corresponding to different levels of dimensionality reduction, are presented to the observers as the video clips to be evaluated. The following subsections describe the process applied to generate the test stimuli, the profile of the participants, the test protocol and the evaluation

results.

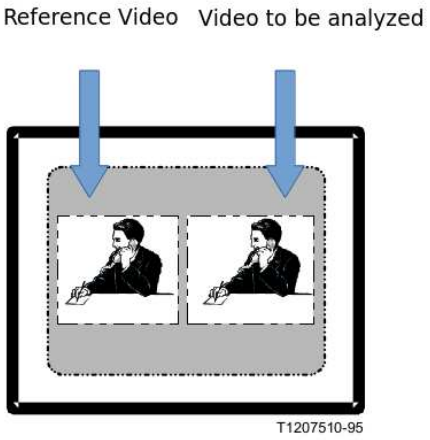


Figure C.1 – Illustration of the simultaneous presentation of sequence video pairs. Source: Adapted from (ITU-T RECOMMENDATION, 1999).

C.2 Test Stimuli

In order to generate the test stimuli, four video excerpts were selected from the test dataset video material associated to the *CH-Unicamp* database, as described in Chapter 3, Section 3.2. Each video excerpt corresponds to a different emotional state: "Happy For", "Anger", "Sadness" and "Disgust". The video sequences are approximately two seconds, or 60 frames, long. In order to obtain the shape information, the frames from the video excerpts were semi-automatically marked following the same procedure applied to build the *CH-Unicamp* database. All video frames were warped to the mean shape $\bar{\mathbf{s}}$ (Section 4.1.1) and projected into the axes defined by the principal components (PCs) of the appearance models computed for the full face; the cheeks+lips region; and the lip region, depicted in Figure 4.5.

After the projection, the same models were used to reconstruct the original frames following the hierarchical approach in which the full face is reconstructed first, followed by the cheeks+lips region; and the lips are superimposed last. The reconstruction process was repeated multiple times, each time with a different number of PCs being used to recover the original frames. For each modeled region, 10 different sets of PCs were chosen in order to represent different levels of cumulative explained variance ratio (EVR). Taking into consideration the different profiles of EVR for each region (Figure 4.7), Table C.1 presents the EVR levels considered for each of them, and their corresponding number of PCs. For each test stimulus the whole face was presented, but only one of the three regions specified in

the columns of Table C.1 (Full Face, Cheeks+Lips or Lips) was subject to dimensionality reduction. For example, to assess the effect of dimensionality reduction on the lip region, the reconstruction of the other regions was performed with full quality (including all PCs), while only the lips region was synthesized with a reduced number of principal components. The reconstructed video excerpts for each emotional state were concatenated and mixed together in such a way that each emotion was played twice. The final video clips are muted (no audio track) and they have approximately 10 seconds of duration. A total of 30 test stimuli video clips were generated for the evaluation (3 regions $\times$ 10 different levels of dimensionality reduction).

C.3 Test Protocol

The assessment was performed presenting two synchronized videos side by side on the screen. The video presented on the left is the reconstructed video with full quality, i.e. synthesized with no dimensionality reduction ($m = 782$) and characterizing the reference video. The video presented on the right is the deteriorated version of the video, being reconstructed with some level of dimensionality reduction in one of the regions of interest.

For each video clip the participants were asked to observe both faces on the screen, trying to detect differences between the two videos. Figure C.2 shows the test application screen with both videos being played synchronously on the left; and the voting options on the right side of the screen. After the video is played, the subjects are asked to vote if the differences they perceived were: imperceptible ("Imperceptível"), perceptible but not annoying ("Perceptível mas não incômoda"), slightly annoying ("Ligeiramente incômoda") or very annoying ("Muito incômoda").

Before the evaluation starts, the participant is oriented with the instructions provided in Annex D. Additionally, the users interacted with the test application a few times before the beginning of the evaluation. The test application randomly selects the order of presentation of the 30 test stimuli for each participant and guides the evaluation session up to its end.

C.4 Population

Fifty-three subjects, students and professionals of the University of Campinas, aged between 19 and 66 years, volunteered to participate in the test. The test results of three subjects were discarded because they reported vision problems that are not corrected with the use of eyeglasses or contact lenses. The test result of one subject was discarded due

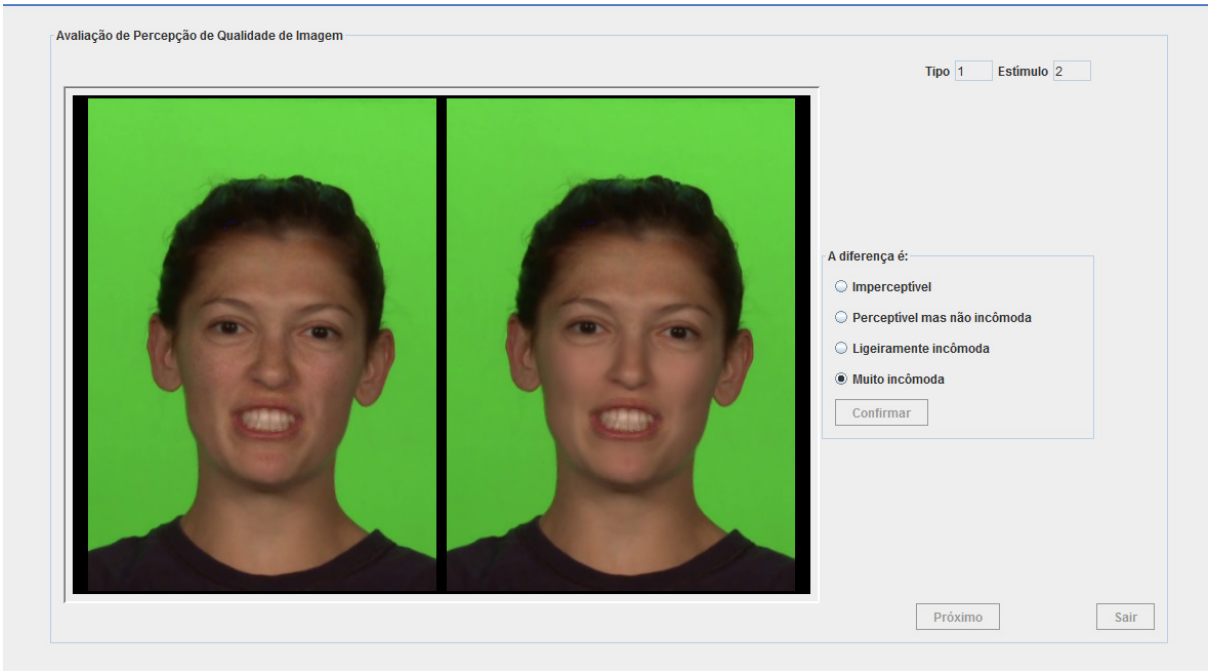
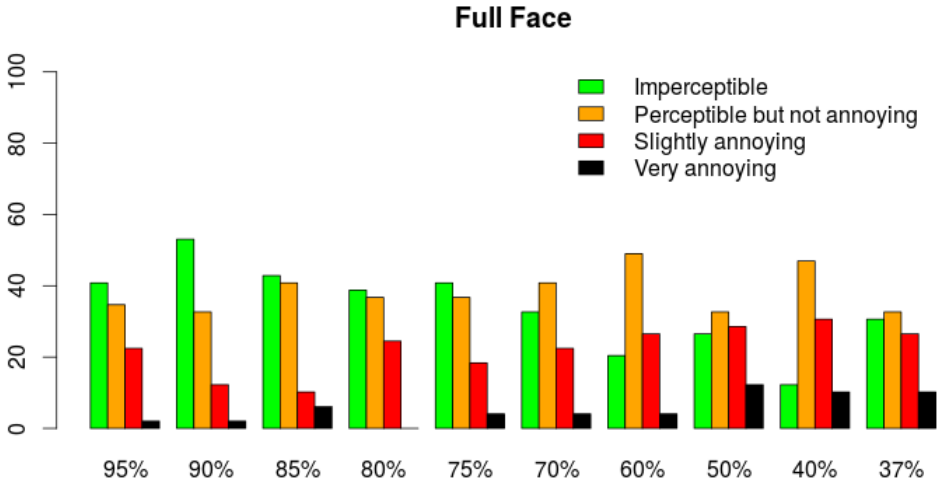


Figure C.2 – Screenshot of the test application. The video presented on the left is the video reference, reconstructed with no dimensionality reduction of the appearance model. On the right, it is the face reconstructed with a lower number of principal components in one of the regions of interest.

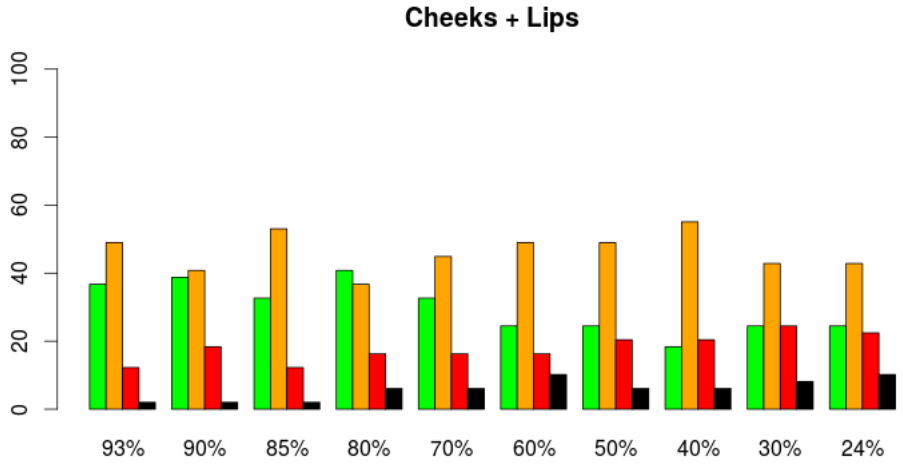
Full Face		**Cheeks + Lips**		**Lips**	
EVR (%)	**Number of PCs**	**EVR (%)**	**Number of PCs**	**EVR (%)**	**Number of PCs**
93	315	93	288	95	188
90	271	90	237	90	90
85	151	85	124	85	48
80	84	80	71	80	29
75	52	70	29	75	19
70	34	60	14	70	13
60	16	50	8	60	7
50	9	40	4	50	4
40	5	30	2	40	2
23	1	24	1	37	1

Table C.1 – Number of principal components (PCs) used to generate test stimuli.

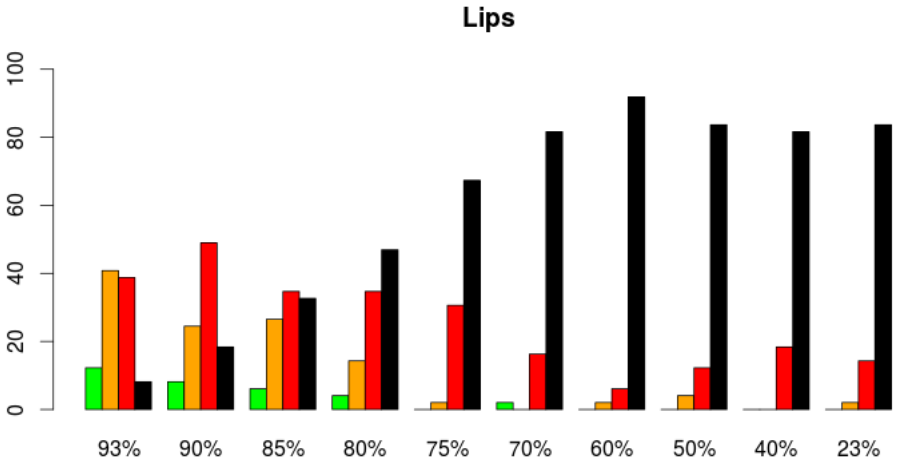
to a technical problem experienced during the test session. All participants are Brazilian Portuguese native speakers.



(a)



(b)



(c)

Figure C.3 – Percentage of votes for each region and for different levels of dimensionality reduction.

C.5 Evaluation Results

The barplots of Figures C.3(a) to C.3(c) present the results obtained for the fullface, cheeks+lips region and lip region, respectively. Each barplot shows the percentage of votes obtained for each level of dimensionality reduction, as specified in Table C.1. The green and yellow bars represent the votes for "imperceptible" and "perceptible but not annoying"; and the red and black bars represent the votes for "slightly annoying" and "very annoying", respectively.

C.6 Discussion

In order to identify configurations of dimensionality reduction that could be applied to the synthesis of facial images without significant loss of visual quality or excessive distortion, a possible strategy is to focus on the EVR configurations in which the green and yellow bars represent the great majority of votes. Following this criteria, for full face for example, the dimensionality reduction at the level of 90% ($l = 271$) or 93% ($l = 315$) seems to be promising. For the cheeks+lips region, a dimensionality reduction up to 85% also seems to be acceptable.

However, it is important to note that in some cases, the dimensionality reduction may not be desirable. The barplot of the lip region for example, show that discarding PCs should not be a priority for this region. The model built for the lip region is highly specialized on expressing the many modes of variation of this region. In this case, discarding PCs brings the risk of affecting the visual quality of the synthesized images.

APPENDIX D – Brazilian Portuguese Instructions for the Evaluation of Video Image Quality

D.1 Avaliação de Síntese de Fala Expressiva

D.1.1 Objetivo

Avaliar como diferentes parâmetros de síntese da animação facial influenciam a qualidade de imagem percebida pelo usuário.

D.1.2 Passo-a-Passo da Avaliação

1. Na tela da avaliação, clique no botão "Próximo" (Figura D.1) para iniciar o teste.

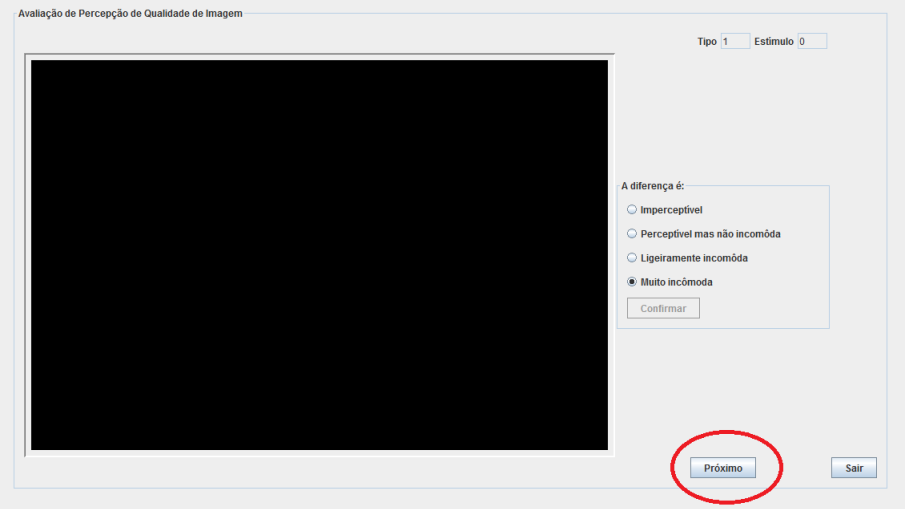


Figure D.1 – Clique em "Próximo" para iniciar o teste.

2. No quadro do lado esquerdo da tela, será reproduzido um vídeo sem áudio. Neste vídeo duas faces da mesma pessoa estarão em movimento lado a lado (veja Figura D.2). A face do lado esquerdo é a FACE DE REFERÊNCIA e face do lado direito é a FACE A SER AVALIADA (Figura D.2). A face do lado direito é uma versão deteriorada da face de referência.

3. A sua tarefa é assistir ao vídeo e, após a sua reprodução completa, classificar se qualquer diferença observada na face a ser avaliada, em relação à face de referência é: *Imperceptível*, *Perceptível mas não incômoda*, *Ligeiramente incômoda* ou *Muito incômoda*.

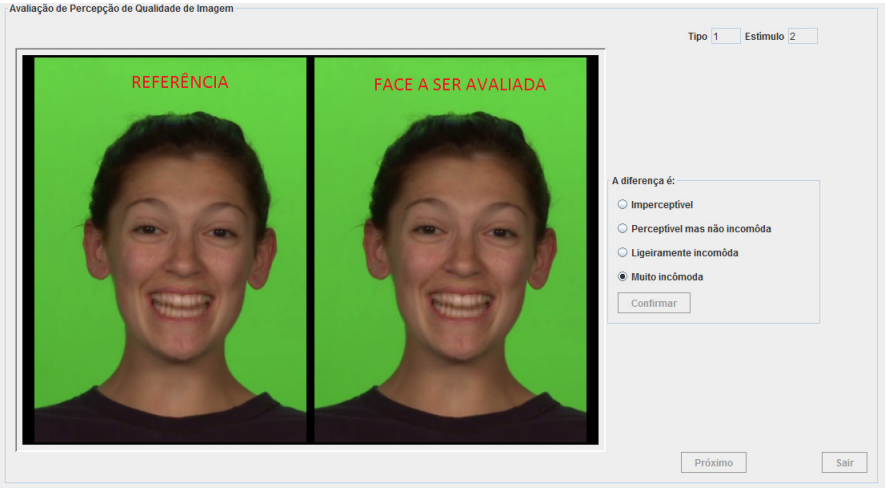


Figure D.2 – Do lado esquerdo da tela é reproduzido um trecho de vídeo de uma face feminina. Escolha uma das opções de votação.

4. A sua opinião deve ser registrada no painel de votação à direita da tela da avaliação.

5. Após registrar sua opinião você deve clicar no botão “Confirmar” e, em seguida, no botão “Próximo”.

6. Repita os passos anteriores, prosseguindo com o teste até o final, quando o botão “Próximo” ficará . Clique em “Sair” para sair da avaliação.

D.2 Recomendações Importantes

- Procure observar a face como um todo, prestando atenção também nos olhos, movimento das sobrancelhas e movimentação da cabeça.

- Evite “pensar muito” para dar uma resposta. Siga seu primeiro impulso.

APPENDIX E – Brazilian Portuguese Emotion Recognition Evaluation Instructions to Participants

E.1 Avaliação de Síntese de Fala Expressiva

E.1.1 Objetivo

Comparar a percepção de emoções a partir do vídeo de uma face real versus uma animação facial.

E.1.2 Passo-a-Passo da Avaliação

1. Na tela da avaliação, clique no botão "Próximo" (Figura E.1) para iniciar o teste.

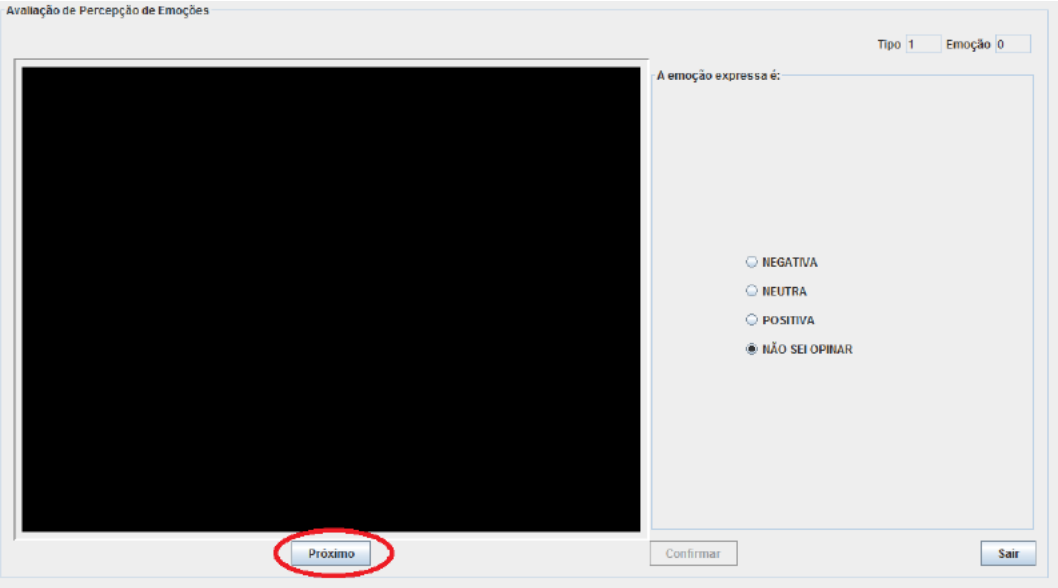


Figure E.1 – Clique em "Próximo" para iniciar o teste.

2. No quadro do lado esquerdo da tela, será reproduzido um trecho de vídeo de uma face feminina expressando uma emoção (o vídeo estará sem áudio). Veja Figura E.2.

3. Sua primeira tarefa será assinalar no painel de votação à direita do vídeo se a emoção expressa é, na sua opinião, uma emoção “NEGATIVA”, “NEUTRA” ou “POSITIVA”. Recomendamos que você opte por uma dessas três opções porém, caso você não se sinta capaz de opinar, selecione a opção “NÃO SEI OPINAR”. (Veja opções na Figura E.2).

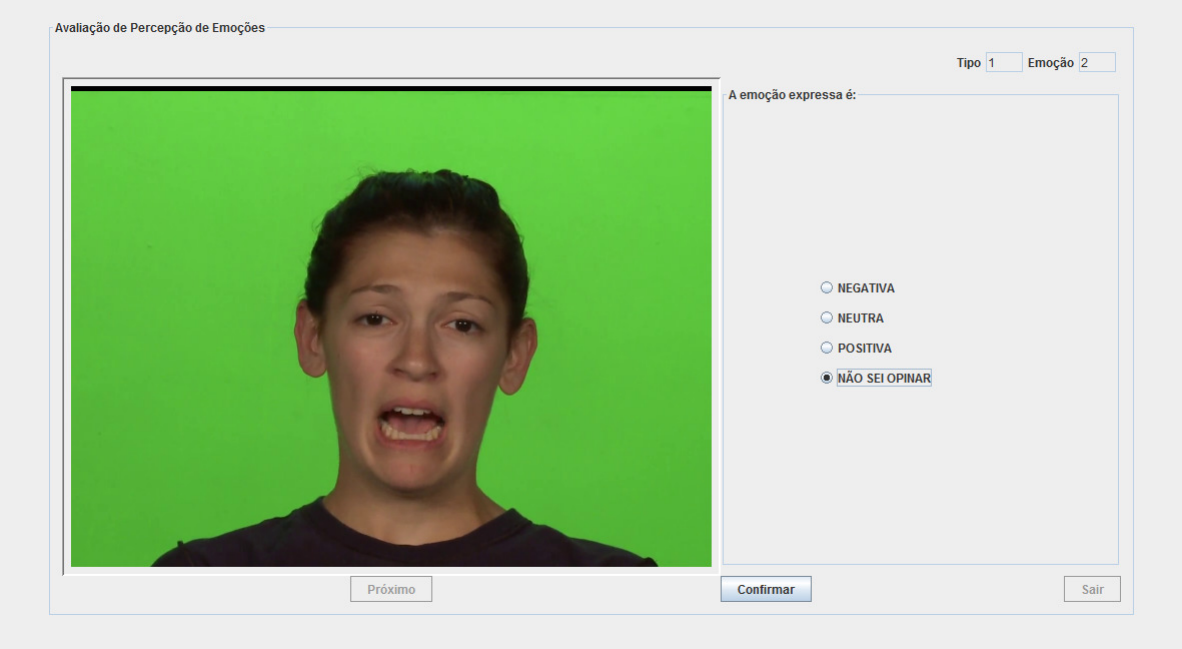


Figure E.2 – Do lado esquerdo da tela é reproduzido um trecho de vídeo de uma face feminina. Escolha uma das opções de votação.

4. Confirme sua opção clicando no botão “Confirmar” e, em seguida, clique no botão “Próximo”.

5. O mesmo vídeo será reproduzido novamente mas, desta vez, sua tarefa será a de informar sua percepção sobre qual foi a emoção expressa pela face no trecho de vídeo que acabou de ser reproduzido. Para isso, você poderá optar entre 22 termos (veja Figura E.3):

   - FELIZ POR ALGUÉM
   - CONTENTE
   - ESPERANÇOSA
   - SATISFEITA
   - ALIVIADA
   - ORGULHOSA
   - RECOMPENSADA

- GRATA
- ADMIRADA
- APAIXONADA
- COM PENA
- TRISTE
- AMEDRONTADA
- RESSENTIDA
- CONFORMADA
- ENVERGONHADA
- CENSURANDO ALGO
- COM REMORSO
- ESCARNECENDO ALGUÉM
- DESAPONTADA
- COM NOJO
- COM RAIVA

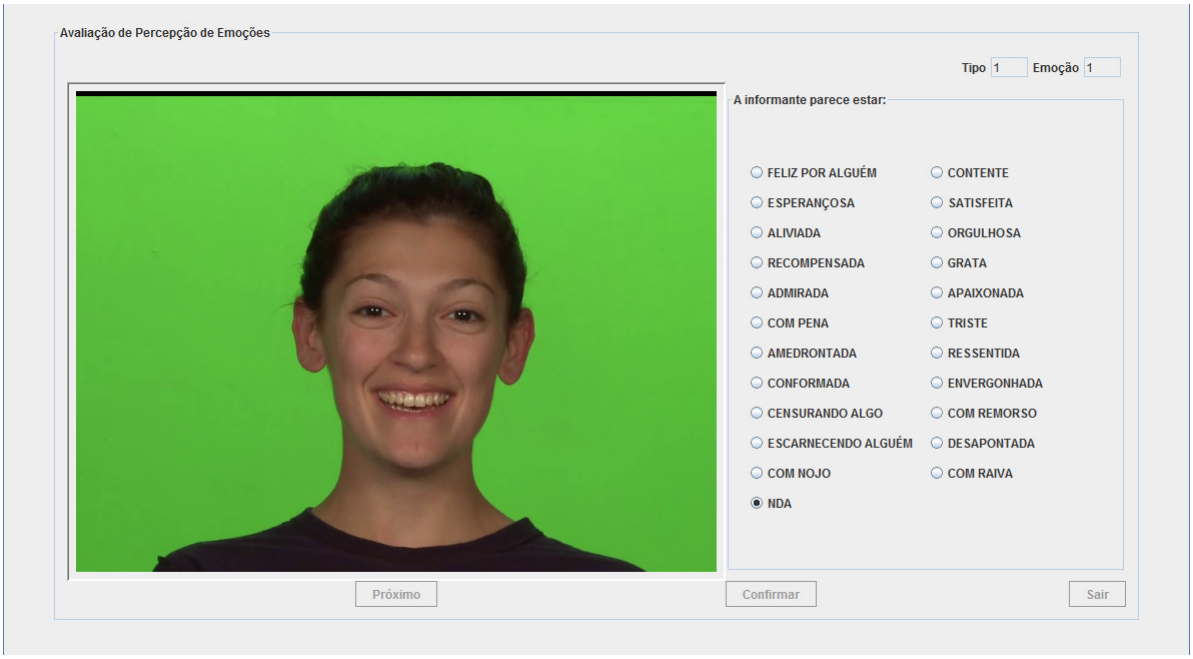


Figure E.3 – Assinale qual emoção foi expressa pela face, escolhendo uma das alternativas, ou NDA (Nenhuma das Anteriores).

6. Mais uma vez, recomendamos que você opte por um desses termos, mas caso julgue que a emoção expressa é diferente das listadas ou não se sinta capaz de opinar, você pode assinalar a alternativa NDA (Nenhuma das Anteriores).

7. Após registrar sua opinião você deve clicar no botão “Confirmar” e, em seguida, no botão “Próximo”.

8. Ao longo do teste, você notará mudanças nos aspecto visual da face. Isso é normal.

9. Prossiga com o teste repetindo os passos anteriores até o botão “Próximo” ficar DESABILITADO. Clique em “Sair” para sair da avaliação.

E.2 Recomendações Importantes

- Nas duas etapas desta avaliação a informação da fala não é relevante. Evite direcionar sua atenção para o exercício de leitura labial, tentando adivinhar o que a face está falando.

- Procure observar a face como um todo, prestando atenção também nos olhos, movimento das sobrancelhas e movimentação da cabeça.

- Evite “pensar muito” para dar uma resposta. Siga seu primeiro impulso.

APPENDIX F – Emotion Recognition Evaluation Analysis: R Script

```
#=============================================================================
#                                  EM_EVAL_ANALYSIS
#=============================================================================
# Script name: em_eval_analysis.r
# Created on: 15/12/2014
# Author: Paula D. Paro Costa
# Purpose: Script to process and proceed with statistical analysis
#          of emotions recognition evaluation results of expressive speech
#          animation system.
#
# Notice:
# Copyright (C) 2014  Paula D. Paro Costa
#=============================================================================

data<-read.csv2('datafile_clean.csv',header=TRUE,sep=',')
data$result<-ifelse(data$Stimulus.==data$VoteVP2.,1,0)
old_par<-par()
emnames=c("Happy For",
          "Joy",
          "Hope",
          "Satisfaction",
          "Relief",
          "Pride",
          "Gratification",
          "Gratitude",
          "Admiration",
          "Love",
          "Pity",
          "Sadness",
          "Fear",
```

```
          "Resentment",
          "Fears Confirmed",
          "Shame",
          "Reproach",
          "Remorse",
          "Gloating",
          "Disappointment",
          "Disgust",
          "Anger",
          "NOTA")

#############################################################################
# Analysis - Perceptual Evaluation of the Valence of Emotions
#############################################################################
valence_video<-table(data[data$Type_Name=='video',]$Stimulus_Name,
                     data[data$Type_Name=='video',]$VoteVP1_Name)
valence_animation<- table(data[data$Type_Name=='animation',]$Stimulus_Name,
                    data[data$Type_Name=='animation',]$VoteVP1_Name)

write.table(valence_video,"valence_video.csv",sep=",")
write.table(valence_animation,"valence_animation.csv",sep=",")

#
# Fisher's exact test  - Valence Voting Option versus Type of Stimulus
#
valence_pvalue<-c()
for (i in levels(data$Stimulus_Name)){
  # Creates contigency table
  vtable<-table(data[data$Stimulus_Name==i,]$VoteVP1.,
                data[data$Stimulus_Name==i,]$Type_Name)
  # Computes the p-value of Fisher's exact test
  valence_pvalue<-rbind(valence_pvalue,fisher.test(vtable)$p.value)
}
valence<-data.frame(levels(data$Stimulus_Name),round(valence_pvalue,digits=4))
write.table(valence,"valence_fisher_pvalues.csv",sep=",",row.names=FALSE)

#
```

```
# Print example of contigency table for anger emotion
#
i='anger'
vtable<-table(data[data$Stimulus_Name==i,]$VoteVP1.,
              data[data$Stimulus_Name==i,]$Type_Name)
print("Example of Contigency Table - Anger")
print (vtable)

#
# Print example of contigency table for relief emotion
#
i='relief'
vtable<-table(data[data$Stimulus_Name==i,]$VoteVP1.,
              data[data$Stimulus_Name==i,]$Type_Name)
print("Example of Contigency Table - Relief")
print (vtable)

#
# Wilcoxon-Mann-Whitney test - Considering the valence of Emotions an
# ordinal variable
#

valence_pvalue_wilcox<-c()
for (i in levels(data$Stimulus_Name)){
  valence_pvalue_wilcox<-rbind(valence_pvalue_wilcox,
     wilcox.test(
      data[data$Stimulus_Name==i & data$Type_Name=="video",]$VoteVP1.,
      data[data$Stimulus_Name==i & data$Type_Name=="animation",]$VoteVP1.)$p.value)
}
valencewilcox<-data.frame(levels(data$Stimulus_Name),
               round(valence_pvalue_wilcox,digits=4))
write.table(valencewilcox,"valence_wilcox_pvalues.csv",sep=",",row.names=FALSE)

#############################################################################
# Analysis - Emotions Recognition Test
#############################################################################
correctvideo<-c()
```

```
correctanimation<-c()
s<-c()

for (i in levels(data$Subject)){
  correctvideo<-
     rbind(correctvideo,
     sum(data$result[data$Subject==i & data$Type_Name=="video"]))
  correctanimation<-
     rbind(correctanimation,
     sum(data$result[data$Subject==i & data$Type_Name=="animation"]))
  s<-rbind(s,i)
}

#
# Generate boxplot of correct answers
#
correctanswers<-data.frame(correctvideo,correctanimation,row.names=levels(data$Subject))
old_par<-par()
par(mar=c(4.5, 5, 2, 1))
par(cex.axis=1.3)
par(cex.lab=1.3)
boxplot((correctanswers/22)*100,
        ylab="Percentage of Correct Answers (%)",
        ylim=c(0,50),
        names=c("Real Video","Facial Animation"),
        col=c("orange","yellow"),
        boxwex=0.5)

#
# Print the statistics summary for both real video and facial animation
#
print(summary(correctvideo))
print(summary(correctanimation))

#
# Print the Shapiro-Test for normality p-values
#
```

```
print(round(shapiro.test(correctvideo)$p.value,digits=4))
print(round(shapiro.test(correctanimation)$p.value,digits=4))

#
# Print the resulting p-value of WMW test
#
print(round(wilcox.test(correctvideo,correctanimation)$p.value,digits=4))

#
# Contingency tables of Emotion Recognition ------------------------------
#

recognition_pvalue<-c()

#
# Exact Fisher Test Recognition Rates per Emotion: Animation x Video
#

for (i in levels(data$Stimulus_Name)){
  vtable<-table(data[data$Stimulus_Name==i,]$VoteVP2_Name,
                data[data$Stimulus_Name==i,]$Type_Name)
  recognition_pvalue<-rbind(recognition_pvalue,fisher.test(vtable)$p.value)
}
recognition<-data.frame(levels(data$Stimulus_Name),round(recognition_pvalue,digits=4))
write.table(recognition,"recognition.csv",sep=",",row.names=FALSE)

#
# Voted Emotions Barplots   ----------------------------------------
#

t<-table(data$VoteVP2.[data$Type_Name=="video"],
         data$Stimulus.[data$Type_Name=="video"])
t<-as.data.frame.array(t,row.names=emnames)
names(t)<-emnames[1:22]

a<-table(data$VoteVP2.[data$Type_Name=="animation"],
         data$Stimulus.[data$Type_Name=="animation"])
```

```
a<-as.data.frame.array(a,row.names=emnames)
names(a)<-emnames[1:22]

for (i in names(a)){
  par(mfrow=c(2,1)) # 2 rows, 1 column

  s<-sort(t[[i]],decreasing=TRUE,index.return=TRUE)

  aux=0
  j=0
  while(aux<sum(t[[i]])){
    j=j+1
    aux=aux+s$x[j]
  }
  s$x<-s$x/sum(t[[i]])*100
  bp<-barplot(s$x[1:j],
              main=paste(i," - Real Video"),
              axes=FALSE,
              axisnames=FALSE,
              ylab="Votes (%)",
              col=rep("orange",length(s$x[1:j])),
              ylim=c(0,100))

  text(bp, par("usr")[3], labels = emnames[s$ix[1:j]],
       srt = 45, adj = c(1.1,1.1), xpd = TRUE, cex=.9)
  axis(2)


  s<-sort(a[[i]],decreasing=TRUE,index.return=TRUE)
  #s$x<-round((s$x/sum(a[[i]]))*100,digits=1)
  aux=0
  j=0

  while(aux<sum(a[[i]])){
    j=j+1
    aux=aux+s$x[j]
    print (aux)
```

```
  }
  s$x<-s$x/sum(a[[i]])*100
  bp<-barplot(s$x[1:j],
              main=paste(i," - Facial Animation"),
              axes=FALSE,
              axisnames=FALSE,
              ylab="Votes (%)",
              col=rep("yellow",length(s$x[1:j])),
              ylim=c(0,100))

  text(bp, par("usr")[3], labels = emnames[s$ix[1:j]],
       srt = 45, adj = c(1.1,1.1), xpd = TRUE, cex=.9)
  axis(2)

}
```