HAL
open science

Generation of audio-visual prosody for expressive virtual actors

Adela Barbulescu

To cite this version:

Adela Barbulescu. Generation of audio-visual prosody for expressive virtual actors. Graphics [cs.GR]. Université Grenoble Alpes, 2015. English. NNT : 2015GREAM071 . tel-01241413v2

HAL Id: tel-01241413

https://theses.hal.science/tel-01241413v2

Submitted on 10 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques-Informatique**

Arrêté ministériel : 1 décembre 2015

Présentée par

Adela BARBULESCU

Thèse dirigée par **Rémi RONFARD**
et codirigée par **Gérard BAILLY**

préparée au sein **du LJK et du GIPSA-Lab**
et de **l'école doctorale MSTII**

Génération de la Prosodie Audio-Visuelle pour les Acteurs Virtuels Expressifs

Thèse soutenue publiquement le **23 Novembre 2015**,
devant le jury composé de :

Ronan BOULIC
MER/HDR, EPFL Lausanne, Rapporteur

Catherine PELACHAUD
DR CNRS, Telecom Paris Tech, Rapporteur

Marc SWERTS
PRF, Tilburg University, Examinateur

Slim OUNI
MCF/HDR, Université de Lorraine, Nancy , Examinateur

Marie-Christine ROUSSET
PRF, Université Grenoble Alpes, Président

Rémi RONFARD
CR/HDR INRIA, Université Grenoble Alpes, Directeur de thèse

Gérard BAILLY
DR CNRS, Université Grenoble Alpes, Co-Directeur de thèse

Table of contents

Abstract vii

Résumé ix

Acknowledgements xi

Glossary xiii

1 Introduction 1
- 1.1 Context and Motivation . . . 1
- 1.2 Expressiveness . . . 2
- 1.3 Goal and contributions . . . 4
- 1.4 Thesis organization . . . 6

2 State of the art 7
- 2.1 Virtual humans . . . 7
  - 2.1.1 Appearance, motion and behavior . . . 8
  - 2.1.2 Virtual actors in movies, games, live performances . . . 10
- 2.2 Acoustic prosody . . . 12
- 2.3 Visual prosody . . . 15
  - 2.3.1 Visual prosodic cues . . . 16
  - 2.3.2 Prosody for the expression of attitudes . . . 19
- 2.4 Visual text-to-speech . . . 20
  - 2.4.1 Rule-based approaches . . . 21
  - 2.4.2 Exemplar-based approach . . . 24
  - 2.4.3 Model-based approach . . . 25
- 2.5 Speech-driven animation . . . 27
  - 2.5.1 Exemplar-based approaches . . . 27
  - 2.5.2 Model-based approaches . . . 28
- 2.6 Expressive audiovisual conversion . . . 32
  - 2.6.1 Voice . . . 32
  - 2.6.2 Facial animation . . . 33

2.7 Summary . . . 34

3 Dataset of dramatic attitudes **37**
3.1 Expressive datasets . . . 37
3.2 Selected attitudes . . . 40
3.3 Recordings . . . 42
3.4 Segmentation . . . 48
3.5 Animation platform . . . 50
3.6 Auto-evaluation . . . 56
3.7 Limitations . . . 57
3.8 Summary . . . 58

4 Analysis of audiovisual prosody **59**
4.1 Feature characterization . . . 59
4.1.1 Data pre-processing . . . 59
4.1.2 Frame characteristics . . . 60
4.1.3 Syllable characteristics . . . 62
4.2 Segmental vs suprasegmental . . . 63
4.3 Audiovisual prosody stylization . . . 64
4.3.1 Virtual syllable . . . 65
4.4 Discrimination between attitudes . . . 65
4.5 Attitude-specific signatures . . . 68
4.5.1 Blinks . . . 72
4.6 Objective measures . . . 72
4.6.1 Objective attitude recognition . . . 72
4.6.2 Actor recognition . . . 75
4.7 Perceptual assessment . . . 76
4.7.1 First perceptual test: original video . . . 76
4.7.2 Second perceptual test: realistic animation . . . 79
4.7.3 Third perceptual test: cartoon-style animation . . . 82
4.8 Limitations . . . 84
4.9 Summary . . . 85

5 Generation of audiovisual prosody **87**
5.1 Frame-based generation . . . 88
5.1.1 Slope feature . . . 89
5.2 Exemplar-based generation . . . 92
5.2.1 Random selection of exemplars . . . 92
5.2.2 Prediction of segmental features . . . 93
5.3 Model-based generation . . . 93
5.3.1 Prosodic feature prediction . . . 94
5.3.2 Blinking prediction . . . 97

5.3.3 Prediction of segmental features . . . 97
5.4 Audiovisual synthesis . . . 98
5.4.1 Reconstruction of stylized prosodic contours . . . 99
5.4.2 Alignment of segmental features . . . 101
5.4.3 Performance synthesis . . . 102
5.5 Evaluation . . . 105
5.5.1 Objective test . . . 106
5.5.2 Subjective test . . . 108
5.6 Limitations . . . 110
5.7 Summary . . . 111

6 Perspectives **113**
6.1 Dialog generation . . . 113
6.2 Large vocabularies . . . 114
6.3 Learning by imitation . . . 114
6.4 Full-body animation . . . 115

7 Conclusion **117**
7.1 Summary . . . 117
7.2 Publications . . . 118

A **119**
A.1 Gaussian Mixture Model conversion . . . 119
A.2 Superposition of Functional Contours model . . . 121

Bibliography **125**

Abstract

The work presented in this thesis addresses the problem of generating audiovisual prosody for expressive virtual actors. A virtual actor is represented by a 3D talking head and an audiovisual performance refers to facial expressions, head movements, gaze direction and the speech signal.

While an important amount of work has been dedicated to emotions, we explore here expressive verbal behaviors that signal mental states, i.e "how speakers feel about what they say". We explore the characteristics of these so-called dramatic attitudes and the way they are encoded with speaker-specific prosodic signatures i.e. patterns of trajectories of audio-visual prosodic parameters. We analyze and model a set of 16 attitudes which encode interactive dimensions of face-to-face communication in dramatic dialogues.

We ask two semi-professional actors to perform these 16 attitudes first in isolation (exercises in style) in a series of 35 carrier sentences and secondly in a short interactive dialog extracted from the theater play "Hands around" by Arthur Schnitzler, under the guidance of a professional theater director.

The audiovisual trajectories are analyzed both at frame-level and at utterance-level. In order to synthesize expressive performances, we used both a frame-based conversion system for generating segmental features and a prosodic model for suprasegmental features. The prosodic model considers both the spatial and temporal dimension in the analysis and generation of prosody by introducing dynamic audiovisual units.

Along with the implementation of the presented system, the following topics are discussed in detail: state of the art (virtual actors, visual prosody, speech-driven animation, text-to-visual speech, expressive audiovisual conversion), the recording of an expressive corpus of dramatic attitudes, the data analysis and characterization, the generation of audiovisual prosody and evaluation of the synthesized audiovisual performances.

Résumé

Le travail presenté dans cette thèse adresse le problème de génération des performances expressives audiovisuelles pour les acteurs virtuels. Un acteur virtuel est répresenté par une tête parlante en 3D et une performance audio-visuelle contient les expressions faciales, les mouvements de la tête, la direction du regard et le signal de parole.

Si une importante partie de la littérature a été dediée aux émotions, nous explorons ici les comportements expressifs verbaux qui signalent les états mentaux, i.e. "ce que le locuteur sent par rapport à ce qu'il dit". Nous explorons les caractéristiques de ces attitudes dites dramatiques et la manière dont elles sont encodées par des signatures prosodiques spécifiques pour une personne i.e. des motifs spécifiques à l'état mental de trajectoires de paramètres prosodiques audio-visuels. Nous analysons et explorons un set de 16 attitudes qui encodent des dimensions interactives de la communication face à face dans les dialogues dramatiques.

Nous demandons à deux acteurs semi-professionnels de jouer ces 16 attitudes tout d'abord en mode isolé (exercices de style) dans une série de 35 phrases, puis dans un dialogue interactif extrait de la pièce de théâtre "La ronde" d'Arthur Schnitzler, sous la supervision d'un directeur de théâtre professionnel.

Les trajectoires audiovisuelles sont analysées à la fois au niveau des trames et au niveau des phrases. Pour synthétiser les performances expressives, nous utilisons un système de conversion basé sur les trames pour la génération des descripteurs segmentaux, et un modèle prosodique pour les descripteurs suprasegmentaux. Le modèle prosodique considère la dimension temporelle et spatiale dans l'analyse et la génération de la prosodie en utilisant des unités dynamiques audiovisuelles.

L'implémentation du système presenté, ainsi que les sujets suivants sont discutés en détail: l'état de l'art (acteurs virtuels, prosodie visuelle, animation guidée par la parole, synthétiseur de parole visuelle à partir du texte, conversion audiovisuelle expressive), l'enregistrement d'un corpus expressif d'attitudes dramatiques, l'analyse et la caractérisation des données, la génération de prosodie audiovisuelle et l'évaluation des performances audiovisuelles synthétisées.

Acknowledgements

First and foremost I would like to thank my advisors, Rémi Ronfard and Gérard Bailly, who gave me the opportunity to do this thesis. Over the course of three years, they have constantly offered their guidence and support, and spent time and energy in helping me to dive into new and complementary areas of research. I would also like to thank the jury members, whose feedback and valuable suggestions helped me to improve this thesis.

The "dramatic" dimension of this thesis was achieved with the help of Georges Gagneré, to whom I owe immense gratitude for his help and long, interesting discussions. Georges brought essential knowlege and help in building our expressive dataset and in attending Experimenta. I strongly thank Lucie Carta and Grégoire Gouby, the acting students who perfomed this difficult exercice of recording the selected attitudes.

Special thanks to Laura Paiardini and Estelle Charleroi, who did all the creative work behind the modelling of cartoon-style actors, and Romain Testylier who was always there to help with Blender related issues.

I would also like to thank fellow researchers with whom it was a great pleasure to collaborate and share ideas: Thomas Hueber, Mael Pouget, Nicolas d'Alessandro, Huseyin Cakmak. Thanks to all my colleagues and friends at INRIA and GIPSA-lab with whom i shared interesting, passionate and many times foolish discussions.

I would like to thank Jordi Gonzalez, my master thesis advisor, who played an important role in my decision to follow the path of academic research and with whom I share great memories from participating in my first international conference.

Finally, I would like to express the gratitude towards my family who has always supported me and to whom I owe the privilege of being here. I would like to thank Alex and Emeric for their friendship (and savoir-être!). And I will conclude with a huge thanks to Rogelio, my constant source of motivation, inspiration and self-improvement.

Glossary

Attitude = the voluntary display of complex, affective mental states as a socio-culturally built response; "how we feel about what we say" (Bolinger, Intonation and its uses) [Bolinger, 1989].

Audiovisual performance = audiovisual speech performed by a virtual actor.

Audiovisual signature = attitude-specific patterns of trajectories of audiovisual prosodic parameters.

Audiovisual speech = a speech performed acoustically and visually.

Audiovisual text-to-speech = technique that uses textual information as input to determine the target phoneme and viseme sequence and then generate the audiovisual speech.

Didascalia = scenic indications provided along dramaturgic text; in our case, it is represented by attitudinal labels to be applied at sentence level.

Director = user of the proposed system who can choose virtual actors, the text and didascalia to recreate a desired theatre play.

Dramatic attitude = acted attitude intended for social contexts.

Emotion = the involuntary display of primary affective states as a physiological response; "how we feel when we say" (Bolinger, Intonation and its uses) [Bolinger, 1989].

Exemplar = original stylized prosodic contour for a given syllabic length.

Exercise in style = a repetition technique in which actors perform the same line in different styles [Queneau, 1947].

Expressiveness = the ability of an audiovisual speech to display an affective style i.e. to express emotions or attitudes.

Intonation = variation of spoken pitch that is not used to distinguish words. It contrasts with tone, in which pitch variation in some languages does distinguish words, either lexically or grammatically.

Mocap actor = an actor whose movements have been captured using a motion capture system.

Modal attitude = attitude expressing modality: declarative, interrogative or exclamative.

Personality = speaker-specific patterning of affect, behavior, cognition, and desires (goals) over time and space.

Phoneme = minimal distinctive unit of spoken languages. It is a perceptive unit: phonemes are not the physical segments themselves, but are cognitive abstractions or categorizations of them.

Prosody = characteristics of sound patterns that do not individuate vowels and consonants but are properties of syllables and larger units of speech. These contribute to such linguistic functions as intonation, tone, stress, and rhythm.

Segmental feature = feature related to the phonetic segments (units of speech): voice spectra, lip motion etc.

Sentence = grammatically complete string of words expressing a complete thought.

Stylization = the extraction of several values at specific locations from the contours with the main purpose of simplifying the analysis process while maintaining the perceptual characteristics of the original contour.

Stylization keypoints = the specific locations at which contour values are extracted in the stylization process.

Stylized contour = set of values extracted from an audiovisual prosodic contour at specific locations.

Suprasegmental feature = feature whose effects are spread over several segments and have meaning in relation units above the phonetic segment: voice pitch, eyebrow motion etc.

Syllable = unit of organization for a sequence of speech sounds. A syllable is typically made up of a syllable nucleus (most often a vowel) with optional initial (onset) and final margins (typically, consonants).

Utterance = a stretch of talk preceded and followed by silence.

Virtual actor = piece of software which can be used to generate ”virtual humans” animations and which can be controlled and edited.

Virtual human = any three-dimensional human-based representation in computer graphics.

Virtual syllables = silences (with the average syllable duration) preceding and following each utterance.

Viseme = minimal distinctive unit of facial animation. A viseme is the visual equivalent of a phoneme.

Word = the smallest unit of organization of sounds that symbolizes and communicates a meaning.

List of Figures

2.1 Examples of face rendering systems and results 8

2.2 Examples of virtual humans 12

2.3 Examples of melodic clichés 14

2.4 Examples of f_0 contours and variance 14

2.5 Examples of families of f_0 contours 15

2.6 Outline of AV TTS synthesizers 20

2.7 Examples of generated facial expressions 23

2.8 SAIBA framework for multimodal generation 23

2.9 Example of HMM trajectory 25

2.10 Example of photorealistic expressions 26

2.11 Outline of speech-driven animation systems 27

2.12 Head motion synthesis overview 28

2.13 Example frames for dynamic units of visual speech 30

2.14 Examples of full-body gestures 31

2.15 Outline of expressive audiovisual conversion systems 32

3.1 Examples of the 16 attitudes interpreted by Lucie. 47

3.2 Examples of the 16 attitudes interpreted by Greg. 47

3.3 Faceshift calibration and setting for dialog recording 48

3.4 Illustration of Praat usage 49

3.5 Video and animated frames for the attitude Seductive. 50

3.6 Video and animated frames for the attitude Thinking. 51
3.7 Video and animated frames for the attitude Scandalized. 51
3.8 Realistic rendering depicting all blendshapes for Greg. 52
3.9 Realistic rendering depicting all blendshapes for Lucie. 53
3.10 Cartoon style rendering depicting all blendshapes for Greg. 54
3.11 Cartoon style rendering depicting all blendshapes for Lucie. 55
3.12 Confusion matrices for the AV auto-evaluation 56

4.1 PCA areas highlighted on the Faceshift template mesh 61
4.2 Poses corresponding to PCA components 62
4.3 Prosodic and segmental features on the generic Faceshift template mesh . 64
4.4 Position-specific LDA recognition rates for acoustic prosody 66
4.5 Cumulative LDA recognition rates for acoustic prosody 66
4.6 Position-specific LDA recognition rates for visual prosody 67
4.7 Cumulative LDA recognition rates for visual prosody 67
4.8 Examples of stylized contours 69
4.9 Attitude-specific signatures 69
4.10 Speaker individuality 70
4.11 Examples of outliers 70
4.12 Examples for left-right rotations 71
4.13 Blink periods 71
4.14 Confusion matrices for the objective recognition rate 74
4.15 Confusion matrices for the first perceptual test 78
4.16 Correlation values between the objective and the first perceptual test . . 78
4.17 Confusion matrices for the second perceptual test 80
4.18 Correlation values between the objective and the second perceptual test . 81
4.19 Confusion matrices for the third perceptual test 83
4.20 Correlations between the objective and the third perceptual test 83

5.1 Outline of the generation system of audiovisual expressive performances . 88
5.2 DTW alignment path and slope contour 90
5.3 Training with frame-based approach 91
5.4 Trajectory generation with frame-based approach 91
5.5 Original vs predicted slope 92
5.6 Prosody generation with exemplar-based technique 93
5.7 Prosody generation with model-based technique 94
5.8 Prediction with model-based approach 95
5.9 Example of families of contours 1 96
5.10 Example of families of contours 2 96
5.11 Example of families of contours 3 97
5.12 Original and predicted blinks 98
5.13 Outline of audiovisual performance synthesis 98
5.14 Reconstructed F0 contour 100
5.15 Reconstructed head rotation contour 101
5.16 Reconstructed eyebrow contour 102
5.17 Example frames for the proposed generation techniques 1 103
5.18 Example frames for the proposed generation techniques 2 104
5.19 Example frames for model-based prediction 105
5.20 Snapshot of the ranking area 108
5.21 Results of the perceptual ranking test 109

A.1 Implementation of contour generators 122
A.2 Analysis-by-synthesis loop 123

List of Tables

3.1 Datasets for affect recognition systems 39
3.2 Modal attitudes 40
3.3 Description of chosen dramatic attitudes 41
3.4 Dialog sentences with didascalia 42
3.5 Isolated sentences 45
3.6 Names of all blendshapes returned by Faceshift 49
3.7 Recognition rates obtained for the auto-evaluation test 56

4.1 Color-coded blendshapes names 61
4.2 Number of motion parameters and variance explained 62
4.3 Feature categorization in terms of segmental structure 63
4.4 Objective attitude recognition 73
4.5 Objective attitude recognition: average rates 73
4.6 Objective actor recognition: average rates 75
4.7 Recognition rates for the first perceptual test 77
4.8 Recognition rates for the second perceptual test 80
4.9 Recognition rates for the third perceptual test 82

5.1 Objective results for the proposed generation approaches 1 107
5.2 Objective results for the proposed generation approaches 2 107
5.3 Results of the perceptual ranking test per attitudes 110

Chapter 1

Introduction

1.1 Context and Motivation

Computer generated imagery represents the primary source of visual content for a wide variety of applications, for example movies and video games, and is entering broader domains such as human-machine interaction and telecommunication. The increased exposure to such content brings about a demand for higher quality animations and the main target for this demand is the representation of *virtual humans*. While animations containing emotionless objects have reached a photorealistic level, the representation of virtual humans is a more challenging task, and more so when motion is involved.

The challenges of creating believable virtual humans engage multiple research directions ranging from appearance to motion control and cognitive aspects. One very important aspect of believable behavior is the ability of virtual humans to display intelligible and realistic audiovisual speech. The automatic generation of audiovisual speech is a research topic that has received a great amount of attention for several decades (see for instance the series of Auditory-visual Speech Processing workshops).

Different types of systems have been proposed towards this goal and the main differences consist in the following characteristics: the input information, the desired output signal (modality, dimension, degree of realism, style), visual articulators representation (image-based, 3D) and technical approach to achieve the prediction of output.

The most popular techniques for audiovisual speech generation use text or a speech signal as input. The so-called audiovisual *text-to-speech* (TTS) synthesis systems use the input textual information - eventually augmented with style-related information - to determine the target phoneme sequence and then generate the audiovisual speech. *Speech-driven* animation systems use an auditory signal as input and estimate the target facial expressions based on features extracted from the input signal. Other approaches

use as input an audiovisual speech and use a *conversion* system to generate a new synthetic audiovisual speech signal displaying a different speaking style or a different speaker altogether.

Besides the lexical content, speech-based communication also displays *prosody*: the set of characteristics which define the style of speech. The totality of acoustic prosodic features include voice pitch, duration, intensity, while visual prosody relates to all the gestures which help convey functions such as prominence, affective or phrasing: head motion, eyebrow motion, gaze direction etc. *Segmental* features such as voice spectra and lip motion refer to the short-time segments i.e. phonetic units of speech. On the other side, prosody is *suprasegmental* in the sense that its effects are spread out all over the utterance and have meaning in relation to each other. The suprasegmental domain is related to syllables, words, phrases, sentences etc.

Naturalistic speech carries a wide range of emotions and complex mental states. Therefore, in order to obtain realistic speech animations, audiovisual prosody should also be taken into account. Text-to-speech and speech-driven techniques require a large amount of recorded data in order to generate audiovisual speech. In the case of *expressive* audiovisual speech synthesis, data needs to be recorded for each different target expression, such as angry, bored, ironic etc.

The following section presents emotion theory elements which should be considered in the implementation of such systems.

1.2 Expressiveness

The first scientific study of emotion appeared in 1872 when Darwin presented his observations on expression of emotion in humans and animals [Darwin, 1872]. His book provided inspiration for a large amount of research, a main part being carried by Ekman and his collaborators [Ekman and Friesen, 1971] [Ekman and Scherer, 1984] [Ekman, 1992].

Such research showed that a few discrete emotions manifest facial expressions which are universally recognized, spanning widely across cultures. Ekman also developed the Facial Action Coding System (FACS) [Ekman and Friesen, 1977], thus providing researchers with a means of precisely quantifying the activity in the facial musculature in order to express distinct emotion.

Overall, this research has lead to a greater understanding of nonverbal communication and of the human emotion system. Ekman's research helped shaping one of the dominant theories of emotion: the existence of a limited set of "basic" emotions, which are universal and correspond to biologically innate emotion response systems. These basic emotions are: happiness, fear, anger, sadness, disgust and surprise.

Another important theory, regarding the effects of emotion on voice production, was proposed by Scherer [Scherer et al., 1980] [Scherer, 1986] and argued that the expression of emotion in speech reflects two types of effects: ”push” and ”pull”. The *push effects* represent the effect of underlying - often unconscious - psychobiological mechanisms. One example is the arousal increase, which leads to the raise of fundamental frequency caused by muscle tension. These effects are expected to be largely involuntary. The *pull effects*, are triggered by conventions, norms, cultural behaviors, which pull the voice in certain directions. Such effects include accepted socio-cultural speaking styles and vocal display rules, the vocal equivalent to facial display rules introduced by Ekman [Ekman, 1973].

Thus, affective expression in speech communications happens either involuntarily (expression of emotion) or voluntarily (expression of attitude)[Scherer and Ellgring, 2007]. The socio-communicative functions trigger specific behaviors of intonation and facial expressions as Bolinger [Bolinger, 1989] notably states:

> ”Intonation [is] a nonarbitrary, sound-symbolic system with intimate ties to facial expression and bodily gesture, and conveying, underneath it all, emotions and attitudes... [it is] primarily a symptom of how we feel about what we say, or how we feel when we say”.
>
> (D. Bolinger, *Intonation and its uses: Melody in grammar and discourse*, p.1 [Bolinger, 1989])

A discussion on the expressive function of prosody is carried in [Ohala, 1996] and [Fónagy et al., 1983]. The topic of universality of emotional expression is approached by Ohala by underlining the essential role of ethology. He creates a taxonomy based on 3 expressive levels, from most to least universal in expression: survival-related signals, psychological and physiological state-related signals and attitudes towards the receiver, the utterance content or the transmitter. Fonagy, on the other side, decomposes the expressive information delivered by prosody into three categories, from least to most intentional: primary emotions, social attitudes and modal attitudes.

Another aspect of emotion theory studies the effect of personality on emotions and attitudes. An illustrative definition of personality with respect to emotion is provided by Orthony et al [Ortony et al., 2005]:

> ”Personality [is a] self-tunable system comprised of the temporal patterning of affect, motivation, cognition, and behavior. Personality traits are a reflection of the various parameter settings that govern the functioning of these different domains at all three processing levels: [...] the reactive, the routine, and the reflective levels.”
>
> (A. Orthony, *Who needs emotions*, p.173 [Ortony et al., 2005])

We consider *expressiveness* the ability of audiovisual speech to display an affective style i.e. to express emotions or attitudes. The generation of expressive audiovisual speech is a challenging problem which has been tackled either using speech-driven or text-to-speech systems, but the target expression was mostly limited to the basic six emotions. Overcoming this limitation represents one of our main motivations.

Numerous studies have been focused on the auditory expression of attitudes and proposed that speakers use global prosodic patterns to convey an attitude [Fónagy et al., 1983][Bolinger, 1989]. However, the amount of research focusing on the visual dimension of complex attitudes is considerably smaller.

Another objective of this work is to understand how attitudes are encoded via audiovisual prosodic contours.

1.3 Goal and contributions

This thesis focuses on the generation of audiovisual prosody for the expression of complex attitudes. We are interested in the "pull" effect and the exploration of characteristics of voluntary behaviors which are triggered in the expression of attitudes. The work is not intended to explore the full spectrum of personality nor the underlying mechanism that triggers emotion.

The thesis was initiated as a collaboration between different research domains, combining theoretical concepts and techniques from computer graphics, vision, speech processing and theater communities.

Goal. Our goal is to generate audiovisual prosody for the expression of attitudes such as: comforting, fascinated, thinking, scandalized etc.

For this, we propose a system for the automatic generation of expressive audiovisual speech from a neutral audiovisual speech. The system requires as input: (1) a neutral version of the audiovisual speech and (2) the label of the desired attitude.

The proposed system was initiated as a tool for the recreation of a virtual theater scene. Gagneré et al [Gagneré et al., 2012] and Ronfard [Ronfard, 2012] also proposed the development of software tools and specific notations to assist the work of a theater director. In our case, the user of the system is able to select the virtual actors, the text of a play and the *didascalia* i.e. scenic indications represented by the desired attitudes. Given the context and the nature of the target expressiveness categories, we denominate these attitudes as *dramatic*. Therefore, *didascalia* is represented by categorical dramatic attitudes at sentence-level. We will refer to audiovisual speech in the context of the proposed system as *audiovisual performance*. For easier control and editing of motion we opt for a 3D representation of audiovisual performance of an actor.

An important inspiration for our work is the Superposition of Functional Contours (SFC) model proposed by Holm et al [Bailly and Holm, 2005]. SFC is a comprehensive model of intonation which proposes a method of generating prosodic contours based on the direct link between phonetic forms and prosodic functions (for example, the attitude, the focus, the phrasing). They propose that prosodic contours present prototypical shapes which depend on the carrier speech size.

Similarly, we hypothesize the existence of visual prosodic *signatures* as manifestations of the attitudinal functions. Joint audio and visual prosodic signatures represent attitude-specific patterns of trajectories of audiovisual prosodic parameters. To this extent, we record a dataset of dramatic attitudes in French, from which we analyze the voice (pitch, duration) and upper-face movements (head and gaze motion, facial expressions) in terms of prosodic trajectories. The SFC model encodes the attitudinal function via prototypical contours. We extend the SFC model to include the visual dimension and we will show that these attitude-specific audiovisual *prototypical contours* do actually exist and efficiently encode attitudes.

While expressive control of audiovisual prosody remains the main focus, this thesis also presents work carried on: the implementation of an animation platform which incorporates different rendering styles for each actor, a crowdsourced evaluation system for perceptual experiments and comparison with objective measures.

Contributions. The main contributions presented within this work:

- We record an audiovisual expressive dataset of 16 different attitudes carried by multimodal recordings of sentences of variable lengths: video, 3D motion capture of facial expressions, head and gaze motion.
- We present an extensive analysis of our corpus, including subjective and objective evaluations of the recognizability of dramatic attitudes.
- We propose three approaches for the generation of audiovisual prosody: frame-based, exemplar-based and model-based generation.
- We train separate statistical methods for the generation of segmental and suprasegmental features.
- We extend the SFC model [Bailly and Holm, 2005] for the generation of audiovisual prosody.
- We present an extensive experimental comparison of the three approaches using subjective and objective tests.

1.4 Thesis organization

The remainder of this dissertation is organized as follows:

Chapter 2, State of the art: presents notable work related to: (1) virtual actors, (2) acoustic prosody, (3) visual prosody, (4) visual text-to-speech, (5) speech-driven animations and (6) expressive audiovisual conversion. We describe the general concepts and techniques related to these subjects and then focus on the usage of expressiveness in each topic and how they are related to our work.

Chapter 3, Dataset of dramatic attitudes: presents the motivation, design and process of recording our expressive corpus, as well as pre-processing techniques and data representation.

Chapter 4, Data analysis: presents several analysis techniques that are relevant to our goals: data processing, feature characterization, discriminant analysis, parameterization and reconstruction, observations of prosodic contours and perceptual evaluation of original data.

Chapter 5, Audiovisual performance generation: presents three approaches proposed: (1) audiovisual expressive conversion of segmental and suprasegmental features which consists of a frame-based generation method, (2) exemplar-based method which uses a random original performance and (3) model-based method which generates prototypical contours for each attitude and sentence-length.

Chapter 6, Perspectives: presents propositions for future directions of this research.

Chapter 7, Conclusion: presents a summary and comments on the contributions. The publications extracted from this work are also presented.

Chapter 2

State of the art

This chapter presents a review of theoretical elements, techniques and results related to virtual humans, acoustic and visual prosody, visual text-to-speech, speech-driven animations and expressive audiovisual conversion. We will describe the state-of-the-art related to each of these topics while maintaining a focus on the relation to expressiveness.

2.1 Virtual humans

The study of virtual humans in the context of this thesis is important because our goal is to generate audiovisual speech performed by virtual humans. This section focuses on the challenges of making virtual humans believable and underlines the topics on which we focus.

Human representations in computer graphics appeared in the early seventies. Historically, they emerged independently in several types of applications: crash simulations, motion analysis, workplace assessment, dance and movement notation, entertainment and motion understanding. Today, virtual human representations are used in an overwhelming number of applications (virtual people for simulation-based learning and training, virtual psychotherapies, virtual presenters, virtual actors, virtual characters in games etc).

Depending on the application of the system and field of usage, virtual humans have been known under numerous terms: virtual characters, virtual actors, conversational agents, digital actors, virtual humans, avatars etc. In the broadest sense, we will use the term *virtual humans* to define any three-dimensional human-based representation in computer graphics.

2.1.1 Appearance, motion and behavior

In all these applications, one stimulating task has been the integration of believable virtual humans in virtual environments. Observing and interacting with virtual humans has proven to be particularly challenging notably because every human is an expert at this process. We have unconsciously been trained to observe fine details in human motion, facial expressions and physical appearance of the human body. At the same time, this familiarity is the cause of the difficulty in accepting synthesized human appearance and behavior. This difficulty is often associated with the concept of the "uncanny valley" exposed by Mori in 1970 [Mori, 1970].

According to Thalmann et al [Magnenat-Thalmann and Thalmann, 2005b], achieving believability of virtual humans implies three important aspects:

1. Appearance: looking like real people includes realistic face and body shapes, skin textures, hair, clothes etc
2. Motion: gestures, facial expressions should be realistic, smooth and flexible
3. Behavior: includes abilities such as: language understanding and synthesis, emotion perception and expression, reactivity to the environment, memory, interaction and social skills etc

These aspects are equally important and interdependent, and engage numerous directions of research. Decades of work in these directions have brought great improvement. For example, in terms of **appearance**, face modeling and rendering have evolved from Parke's parametric model [Parke, 1974] developed in 1974 to photorealistic digital results obtained with scanning techniques such as the USC ICT's Light Stage [1] [Debevec, 2012]. Remarkable work is carried on realtime photoreal rendering in order to portray realistic reflections and wetness of the eye [Jimenez, 2012] (see Figure 2.1).

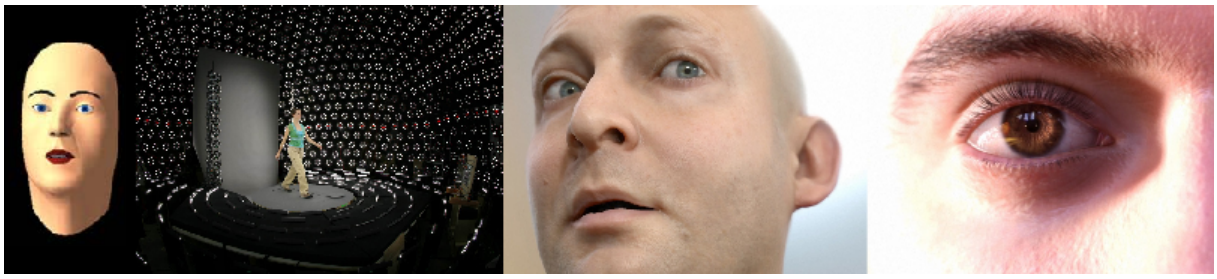

Figure 2.1: Examples of face rendering systems and results. From left to right: The 3D face model introduced by Parke [Parke, 1974], the Light Stage model 6, Digital Ira [Alexander et al., 2013] obtained with the Light Stage technology, realtime photorealistic eye rendering [Jimenez, 2012].

[1] http://ict.usc.edu/prototypes/light-stages/

The most successful results for simulating the flexibility and smoothness of human **motion** are obtained using commercial motion capture systems [2]. Work has been carried to transit towards realtime markerless facial motion capture [Weise et al., 2011] [Ichim et al., 2015], with commercial systems developed for individual-level usage using either a 3D range camera [3] or a simple RGB camera respectively. Motion depends on the underlying modeling techniques and their intrinsic capability of generating expressive shapes. This is particularly applicable to facial animation where subtle movements can cause a dramatic change in the expressed state.

Different techniques were introduced for virtual human representation presenting various advantages and disadvantages, making collaborative research difficult. Therefore, a few standards have been defined, such as FACS and MPEG-4. Facial Action Coding System (FACS) [Ekman and Friesen, 1977] which is a description of the movements of facial muscles and jaw/tongue, defining 44 action units (AUs). The effects of the action units are represented with blendshapes, whose linear weighted sum define a facial pose. Much work is currently spent on defining optimal blendshape models. Similar to FACS, the MPEG-4 [Ostermann, 1998] is another widely used standard. It is an object-based multimedia compression standard and uses facial animation parameters (FAP) that combine AUs into more synergetic degree of freedom.

The most difficult task remains that of creating believable **behavior**, which implies showing realistic interaction with (real) users and other virtual humans. Human communicative behaviors span a broad set of skills, from natural language generation and production, to coverbal gesture, to eye gaze control and facial expression. A common behavior specification framework was proposed [Vilhjálmsson et al., 2007] [Heylen et al., 2008] in order to unify international efforts for virtual agent behavior generation. For instance, the framework is used by the first embodied conversational agent (ECA), Greta [Niewiadomski et al., 2009]. Greta is able to communicate using verbal and nonverbal channels such as gaze, head and torso movements, facial expressions and gestures. Work has also been carried for the development of standards for scripting believable interaction [Perlin and Goldberg, 1996].

Another important aspect of believable motion and behavior is represented by perception and expression of emotion, implemented using emotion modeling and expressive audiovisual performance techniques. Synthesized expressive motion has obtained good results using speech-driven methods [Levine et al., 2010][Marsella et al., 2013] [Chuang and Bregler, 2005] or visual text-to-speech [Albrecht et al., 2002a] [Thiebaux et al., 2009]. An in-depth look at state-of-the-art techniques for visual text-to-speech and speech-driven animation is presented in sections 2.4 and 2.5.

[2] http://www.vicon.com/
[3] http://www.faceshift.com/

2.1.2 Virtual actors in movies, games, live performances

Movies

Beginning with the 1980s, an interest appeared in producing short films and demos involving virtual humans. Examples of the earliest computer-generated short films include "The juggler" by Richard Taylor and Gary Demos, appeared in 1981, and facial animations appeared in 1985: "Tony de Peltrie", the first animation to illustrate emotions through facial expressions and the music video for Mick Jagger's song "Hard Woman". In 1987, With the occasion of the 100th anniversary of the Engineering Institute of Canada, Nadia Magnenat Thalmann and Daniel Thalmann presented the short film "Rendez-vous in Montreal". This is the first film to create digital versions of existing actors, portraying a meeting between Marilyn Monroe and Humphrey Bogart in a cafe. With a team of 6 persons, the production of the film took over one year and the virtual characters were capable of speaking, showing emotion, and shaking hands [Magnenat-Thalmann and Thalmann, 2005a].

In 1988, Pixar's "Tin Toy" was the first entirely computer-generated movie to win an Academy Award (Best Animated Short Film). In 1989, "The Abyss", directed by James Cameron included a computer-generated face placed onto a watery pseudopod. In 1991, Terminator 2: Judgment Day, also directed by Cameron, included a mixture of synthetic actors with live animation, including computer models of Robert Patrick's face [Magnenat-Thalmann and Thalmann, 2005a].

The integration of virtual actors in movies became a well-established practice after the production of "The crow" in 1994. Brandon Lee died throughout the shooting of the film and all the remaining scenes were filmed by digitally superimposing his face over a double's body. Beginning with the 2000's, movies started to incorporate realistic 3D humans; the first fully animated movie to use only computer-generated actors was "Final Fantasy: The Spirits Within" (Hironobu Sakaguchi, Motonori Sakakibara, 2001). In 2004, "The Polar Express" (Robert Zemeckis) became the first movie to use motion capture on all of its actors. Other movies which pioneered numerous technical novelties to achieve a high degree of realism and aesthetic visual effects are: Avatar(James Cameron, 2009), Tintin (Steven Spielberg, 2011), Gravity (Alfonso Cuaron, 2013).

The subject of using virtual actors (created from scratch or by digitally scanning real actors) is addressed in the movies "S1m0ne" (Andrew Niccol, 2002) and "The congress" (Ari Folman, 2013). The movies tackle ethical and legal issues that have bred a long-term debate.

Games

A characteristic of computer games is the move from performance to agency. The non-interactive nature of movies allows character animators to produce or at least refine

memorable performance for virtual actors. On the other hand, a computer game is interactive, allowing the user to act and to control the virtual character, which is rendered in realtime, thus increasing the visual realism. The attention of a user is not only focused on performance but also on the motion and behavior of the character, and how these evolve under his/her control, as in Assassin's Creed (Ubisoft Montreal, 2008), or shifts between athletic poses, as in FIFA Soccer 11 (EA Canada, 2010).

In recent years, the performance of computer game characters has increasingly received more attention in the developmental stage. In particular, facial animation has become the new focus of developers and players alike. Facial performance-capture technology has been used to enhance character performances in games such as Grand Theft Auto IV (Rockstar North, 2008), Operation Flashpoint: Dragon Rising (Codemasters, 2009), Assassin's Creed II (Ubisoft Montreal,2009) and Napoleon: Total War (The Creative Assembly, 2010). As players get more familiar with realistic facial animation, realism becomes more demanded in computer games. Virtual character presentation and performance have reached stunningly realistic effects with the recent games: Heavy rain (Quantic Dream, 2010), L.A. Noire (Team Bondi, 2011), Beyond: Two souls (Quantic Dream, 2013), Rise of the Tomb Rider (Crystal Dynamics, 2015).

Live performances

Virtual humans have also been used in live performances. From musical groups which assume an imaginary identity (Gorillaz) to bringing to life late musicians using holograms (Michael Jackson, Tupac), the public can experience performances made by virtual characters. Hatsune Miku gives live performances to sell-out crowds of thousands which indulge the realtime music created by Yamaha's Vocaloid [4] joined by a real live band. In theater, we are not aware of the usage of virtual actors but several plays have successfully used robotic actors, such as the ones created by Hiroshi Ishiguro [5]. These virtual performers present prerecorded movements, obtained using motion capture or manually defined. On the contrary, RoboThespian is the first commercial robot designed for human interaction in a public environment [6] and is fully interactive. Its movements and voice can be prerecorded or synthesized in realtime. The Robothespian proves to be very useful for performing in theater plays because of its interactive capabilities.

Discussion. This section provided a general insight on the usage of virtual humans both in commercial systems and in research studies. Our focus is the generation of believable behavior of virtual humans, specifically for affective performances. Therefore we want to create controllable and interactive humans.

In previous work and within the entertainment industry, the term virtual actor has mostly referred to the representation of a human in the virtual world. In this thesis, the

[4] http://www.vocaloid.com/en/
[5] http://www.geminoid.jp/
[6] https://www.engineeredarts.co.uk/robothespian/

Figure 2.2: Examples of virtual humans. From left to right: portrayal of Marylin Monroe in "Rendez-vous in Montreal" (1987), digital superposition of Brandon Lee's face in "The crow" (1994), hologram of Michael Jackson surrounded by real dancers at Billboard Music Awards 2014, Hatsune Miku hologram in concert (world tour 2015), Beyond: two souls (2013), Tintin (2010), The congress (2013).

term *virtual actor* is defined as the piece of software that allows the generation of virtual human animations which can be controlled and edited. The term is used in opposition to the *mocap actor*, which represents an actor whose movements are transferred from a motion capture system.

2.2 Acoustic prosody

This section presents theoretical elements related to the linguistic aspect of an expressive speech.

Acoustic prosody refers to the characteristics of sound patterns that do not individuate vowels and consonants but are properties of syllables and larger units of speech. These contribute to such linguistic functions as intonation, tone, stress, and rhythm. Prosody reflects various features of the speaker's identity, mental and emotional state or the utterance.

While receiving a wide range of interpretations in linguistic literature, prosody can be defined as the broad set of features that do not determine *what* people are saying, but rather *how* they are saying it. Such a set of features includes pitch, voice quality, loudness, rhythm, speech rate, pauses, which are encoded in the speech signal it-

self [Johan't Hart and Cohen, 1990]. These features are called "suprasegmental" because they "comprise properties of speech that cannot be understood directly from the linear sequence of segments" [Heuven, 1994].

Distinctive phonetic features apply to simple phonetic segments (e.g. bilabial articulation for /b/, rounding for /u/) as well as to strings of several sounds, such as a syllable, a word or utterance. In the first case, the features are termed as segmental versus suprasegmental features in the last case. The study of the suprasegmental features is known as *prosody* [Lehiste, 1970] and mainly deals with the organization of the duration, intensity and pitch of sounds. Note that prosodic features can be both segmental and suprasegmental according to the scope of the analysis: voicing or quantity are segmental features that distinguish between /b/ and /p/. Both voicing and quantity are also suprasegmental features when considering voicing assimilations across segments, syllabic lenghtening or changes of voice quality due to emotions.

Prosody plays a fundamental role in message communication for spoken interaction. It provides information which can be complete or change the semantics of the discourse (such as the ironic intonation). An extensive list of prosody functions and related studies is provided by Xu [Xu, 2011]. Among these, we can mention: focus, prominence, newness and givenness, boundary marking, grouping, structuring, topic and contrastive topic, turn-taking, modality, emotion and attitude.

As speech synthesis matured, it became clear that prosody should be integrated for a natural-sounding effect for a desired expressive style. Therefore, a main goal in the study of prosody is modeling and generation of prosodic features for expressive speech synthesis.

To this extent, two approaches are generally used: rule-based or statistical models, which require an extensive training corpus and heavy annotation. The methods involved depend also on the targeted prosodic function. One approach which has influenced our work is that of Fonagy [Fónagy et al., 1983], who proposes the existence of prosodic patterns for the expression of attitudes.

Fonagy states that "our utterances are made up of large prestored pieces" that he calls *clichés mélodiques* (see figure 2.3). The shape contours of these pieces are highly dependent on the expressed attitude.

These studies have inspired the research on prosody generation based on metalinguistic information. For example, Morlec et al [Morlec et al., 1995], [Morlec et al., 2001] studied a set of social attitudes in order to reveal the existence of global prototypical contours which are attitude-specific and are deployed onto the whole utterance (see figure 2.4).

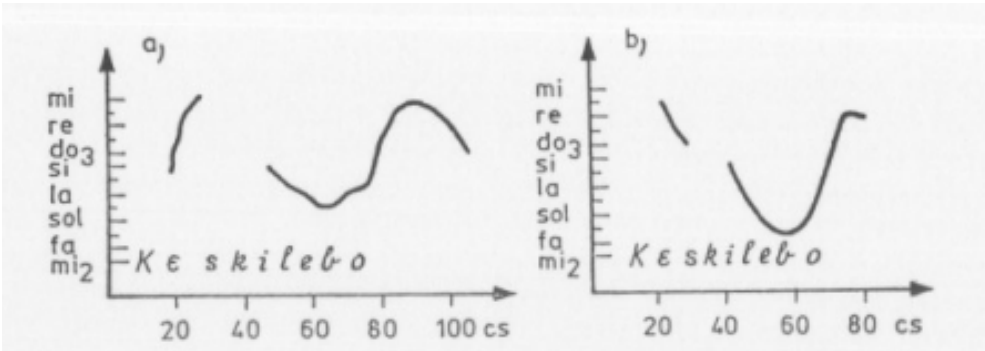

Figure 2.3: Examples of melodic clichés. F_0 contours for sentence "Qu'est-ce qu'il est beau!" ("How handsome he is!") for Fascination and Irony [Fónagy et al., 1983]. Both contours follow typical schemes for these attitudes: the contour for Fascination is more stretched and presents an initial strong rise and final fall as opposed to Irony.

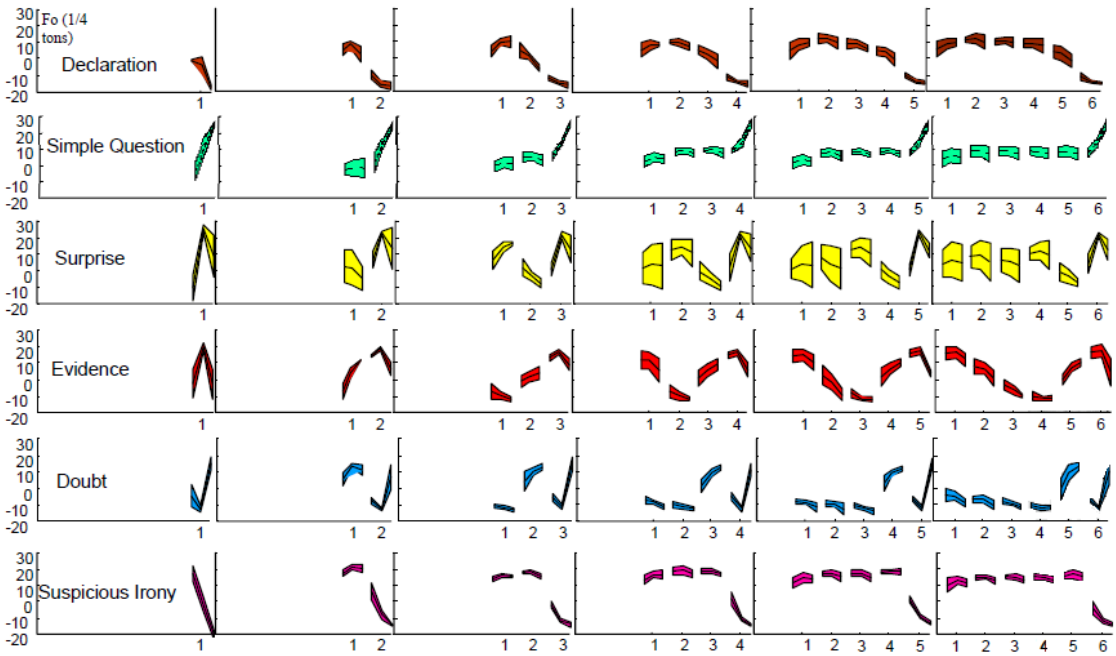

Figure 2.4: Examples of f_0 contours and variance. Complete f_0 contours and variance illustrated in [Morlec et al., 2001] for Declaration, Question, Surprise, Evidence, Doubt and Suspicious Irony, for sentences containing from 1 to 6 syllables (from left to right).

This work was continued by Holm and Bailly [Bailly and Holm, 2005] which implemented a trainable generation model of intonation: the Superposition of Functional Contours (SFC) model. This model promotes an intimate link between phonetic forms and linguistic functions [Aubergé and Bailly, 1995]: metalinguistic functions acting on different discourse units are directly implemented as global multiparametric contours. Also, the

functional contours applied to different scopes but encoding the same function constitute a morphologically coherent family of contours. Such a family of contours can be generated using a so-called contour generator which is implemented using a simple neural network that is fed only with phonotactic information. This hypothesis has been tested using three main categories of socio-communicative functions: attitudes (using declaration, question, doubt-incredulity, obviousness, exclamation and irony), hierarchy (inter-dependence between phrases within an utterance), emphasis (narrow focus on a word). Examples are illustrated in figure 2.5.

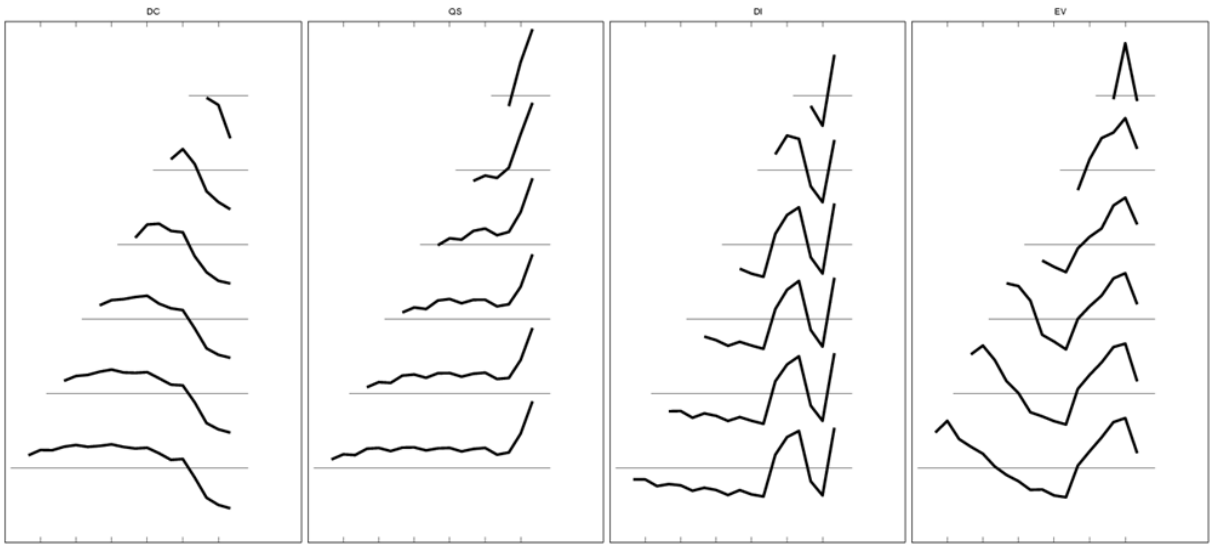

Figure 2.5: Examples of families of f_0 contours. Families of f_0 contours illustrated in [Bailly and Holm, 2005] for Declaration, Question, Doubt and Obviousness, for sentences containing from 1 to 6 syllables (from top to bottom). Note the similarity between these contours and the mean f_0 contours from figure 2.4.

2.3 Visual prosody

The concept of *visual prosody* emerged as visual information was proven to play an important role in communicative functions. Visual cues influence speech intelligibility and signal issues such as expression of emotion and attitudes, turn-taking etc. The study of this particular topic was intensified due an increasing interest in the effects of visual information on speech perception (e.g. visual contributions to speech intelligibility [McGurk and MacDonald, 1976], [Massaro and Palmer, 1998]) and audiovisual speech synthesis, which required believable visual support.

Swerts et al [Krahmer and Swerts, 2009] present a detailed overview of studies carried on audiovisual prosody. One aspect of this research relates directly to the communication functions that were traditionally attributed to auditory prosody:

prominence, focus, phrasing. Analysis on head nods, eye blinks, eyebrows movement showed that these visual cues present a high influence on word prominence [Pelachaud et al., 1996][Granström et al., 2002][Cvejic et al., 2010]. Other areas of study include emotion, attitude and modality. Other aspects presented in the overview refer to the effect of sign language on visual prosody and to tools and methods that are especially relevant for the study of audiovisual prosody.

2.3.1 Visual prosodic cues

This section presents important work carried on the topic of visual prosody, with a focus on the emotional and attitudinal functions. We perform a simple categorization based on the type of visual cues studied.

Eyebrow movements

An early study on the correlation between f_0 variations and *eyebrow movements* was carried by Cavé et al [Cavé et al., 1996]. The movements were recorded for yes/no questions, reading and dialogues. Rapid eyebrow movements were associated to 71% of the f_0 rises, proving that the two behaviors are synchronized as a consequence of communicational choices rather than automatically. 38% of the eyebrow movements were produced during silences, thus pointing the role in turn-taking or as back-channel signals.

Krahmer et al [Krahmer et al., 2002] carry an experiment with a talking head, aimed at finding out the relative contributions of pitch accents and rapid eyebrow movements for the perception of focus. The results reveal that both features have a significant effect on the perception of focus, albeit that the effect of pitch is much larger than that of eyebrows. Ding et al [Ding et al., 2013] and Roon et al [Roon et al., 2015] further explored the correlation between eyebrow movements and f_0.

Head and eyebrow movements

With the goal of creating an animated talking agent capable of displaying realistic behavior, evaluation experiments were carried by Granstrom et al [Granström and House, 2005] to test various hypotheses on audiovisual prosody. The contribution of *head nods, eyebrows and timing* is investigated for the perception of prominence, feedback [Granström et al., 2002] and modal attitudes. The results show that audiovisual cues are relevant for signaling negative or affirmative feedback in a conversation, while presenting different strengths: smile and f_0 were the most prominent, followed by eyebrow and head movement, and then eye closure and delay contributing only marginally. Tests on statement/question posing proved the dominance of the auditory cues, indicating that question intonation is less variable than visual cues for questions as final high intonation is generally very robust.

Busso et al [Busso and Narayanan, 2007] analyze an audiovisual dataset containing semantic-neutral expressive sentences. Among the correlations studied, we mention the emotional-dependent relation between speech features and *lip, head and eyebrow motions*. They show that facial activeness changes under emotional speech and that acoustic features such as pitch, energy, and spectral envelope vary as function of emotions. Another important observation is that facial gestures and speech are connected at different time scales. While the lips movements are locally tightly connected with the spoken message, the eyebrow and head motion are coupled only at the sentence level. They therefore conclude that a coarse-to-fine decomposition of acoustic and facial features may be beneficial to model the coupling between different facial areas and speech.

Swerts et al [Swerts and Krahmer, 2010] present an analysis of *head and eyebrow movement* for newspaper readers depending on content and intended auditory. The analysis shows a strong correlation between the studied visual features and pitch accents which mark emphasis.

Head movements

Munhall et al [Munhall et al., 2004] show that *head motion* conveys linguistic information. Analyzing Japanese sentences, they show that head movement correlates strongly with the pitch and amplitude of the talker's voice. They also carry a perception study in which Japanese subjects viewed realistic talking-head animations in a speech-in-noise task, where head motion could be manipulated without changing other characteristics of the visual or acoustic speech. Subjects correctly identified more syllables when natural head motion was present in the animation than when it was eliminated or distorted.

Busso et al [Busso et al., 2007] perform statistical measures on an audiovisual database, revealing characteristic patterns in *emotional head motion* sequences. The discriminating information is then used in a hidden Markov model framework to synthesize new expressive head motion.

Head movements and facial expressions

Another study on *head movements* and *facial expressions* carried by Graf et al [Graf et al., 2002] shows that these movements exhibit a wide variety of patterns which depend on speaker's personality, mood, text content (greeting or reading newspaper paragraphs). A characteristic for all speakers is the existence of simple motion patterns which are well synchronized with main prosodic events (pitch accents, phrase boundaries). Wide speaker variations appear especially in direction and strength of head movements. An interesting type of simple motion pattern is represented by *preparatory head nods*.

A study carried by Beskow et al [Beskow et al., 2006] consisted in measuring *head and facial motion* for short read sentences in Swedish. The speakers used variable focal ac-

cent for several expressive modes including: certain, confirming, questioning, uncertain, happy, angry and neutral. A novel measure, the focal motion quotient, is introduced to compute the average amount of variability introduced by a focal accent. The quotient shows that shifting from a non-focal to a focal pronunciation results on average in greater dynamics in all facial movements for all expressive modes.

The study carried by Cvejic et al [Cvejic et al., 2010] addresses visual prosody analysis by measuring the overall *head and facial movements* for several speakers and applying guided principal component analysis. The speakers were engaged in a dialog in order to elicit: narrow focused statements, broad focused statements and echoic questions. The results showed robust visual prosodic cues both in the mouth region movements and brow and head movements, an increase of rigid rotations for echoic questions and individual variation of visual cues.

Eyeblinks

Realistic eye behavior implies several types of motion related to the underlying anatomical structure: eyeball movement (saccades, vestibulo-ocular reflex, vergence), eyelid movement, combined eye-head movement. Eyelid movement can be broken into several subclasses, all presenting different dynamics: spontaneous, voluntary and reflexive [Oyekoya et al., 2010]. The spontaneous blinks are linked to cognitive state and activity. Blink occurrence is influenced by factors such as attentional processes, fatigue, lying, communication context. However, blink rate is variable and has been generated in animation systems through different methods, from identical interval generation to physiological data analysis and high-level behavior generation dependent on the cognitive state and communication functions. Nakano et al [Nakano and Kitazawa, 2010] carried experiments in which Japanese participants were asked to view videos of a speaker and compared the listener-speaker blinking behaviors. The results suggest that *blink* entrainment reflects smooth communication since participant synchrony mostly occurs at the end and during speech pauses.

Studies based on contextual *eyelid motion* behavior [Bentivoglio et al., 1997] (silent rest, short passage reading, natural induced talking) show that blink rate is influenced by cognitive state and that patterns follow a certain frequency distribution. Also, the blink pattern of a person generally has a strong correlation with the conversational role. For example, speakers exhibit a higher frequency blink than listeners [Albrecht et al., 2002b], [Bailly et al., 2010].

Gaze motion

An extensive survey of eye and gaze animation is presented by Ruhland et al [Ruhland et al., 2014]. During face-to-face conversational interactions, eyes exhibit conversational turn-taking and agent thought processes through gaze direction, saccades, and scan patterns [Lee et al., 2002] [Vertegaal et al., 2001]. Also, gaze motion is a pow-

erful method to express the emotional state [Izard, 1991] [Fukayama et al., 2002].

2.3.2 Prosody for the expression of attitudes

Several studies have been carried on the audiovisual prosody of emotions. Studies on attitudes, especially in interaction, are more seldom. Here we present a few works which focused on the audiovisual prosody of attitudes.

A perceptual analysis of audiovisual prosody for Brazilian Portuguese is carried by Moraes et al [De Moraes et al., 2010] using video data recorded by two speakers. Moraes et al studied 12 attitudes categorized as social (*arrogance, authority, contempt, irritation, politeness* and *seduction*), propositional (*doubt, irony, incredulity, obviousness* and *surprise*) and assertion (*neutral*). They conducted an attitude recognition test which showed: the difference in perception between the two speakers for certain attitudes (different strategies developed for irony and seduction, which result in statistically different recognition rates), different dominant modality (one speaker is better recognized in audio while the other in video), better overall recognition rates for audio-video among all modalities, the propositional and social attitudes show different perceptual behaviors i.e. subjects rely on both audio and visual cues for propositional expressions, while they mainly use visual cues for social attitudes.

Sendra et al [Sendra et al., 2013] investigated the interaction between facial expressions and pitch contours for simple and *incredulous question*. The analyzed data was obtained by recording two actors who were Dutch and Catalan speakers, respectively. Facial expressions were extracted from the video by experienced FACS annotators. The study illustrates the differences between the performances given by the two speakers for expressing the same attitude and the particular differences between the two modalities chosen (e.g. incredulous question uses the eye squint muscle activation).

Another work on perception of audiovisual attitudes is focused on the expression of 16 social and/or propositional attitudes in German. Honemann et al [Hönemann et al., 2015] perform a set of attitude recognition tests. While the observations are valuable, these studies focus on the perceptual results of attitude recognition tests and do not carry a complete analysis, including facial features and voice parameters.

A special mention is the recognition tests conducted by Baron-Cohen et al [Baron-Cohen et al., 1997]. They gathered a set of static photos illustrating acted postures for basic emotions and complex mental states. The stimuli include full-face, eye area and mouth area for two actors. The tests show that the eye area is almost as significant as the full-face for the recognition of complex mental states.

Discussion. The purpose of this section was to highlight the importance of visual cues in the audiovisual speech and to show how the two dimensions are correlated. The results

of the mentioned studies show that visual prosody both correlates but also compliments the auditory prosody.

To our knowledge, there are no extensive studies of the correlation between acoustic features and nonverbal gestures for the production of a large set of complex attitudes. Except for a few works related to modalities (question: [Srinivasan and Massaro, 2003],[Cvejic et al., 2010], [Sendra et al., 2013]), there is no qualitative analysis dedicated to the dynamics of visual prosodic contours of attitudes.

2.4 Visual text-to-speech

Text-to-speech synthesis [Dutoit, 1997] represents the generation of speech from an input text. Generally, this technique consists of a text analysis part, which produces a symbolic representation of a spoken sentence and a phonetic transcription of the words, followed by the speech synthesis part, which converts the symbolic representation into audible speech. Similarly, visual text-to-speech (visual TTS) implies an implicit or additional facial animation module which generates motion synchronized with audio. Prosodic functions such as emotional content are also provided using additional information (prosodic labels, expressive weights etc) or via semantics extracted from text. Figure 2.6 presents the general outline of a audiovisual TTS system.

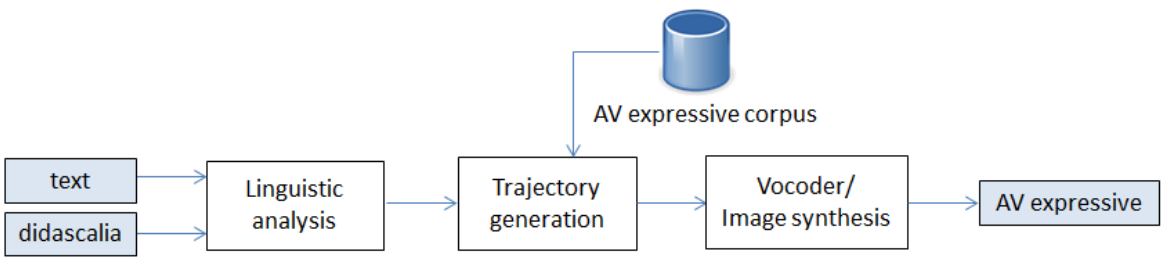

Figure 2.6: Outline of audiovisual (AV) TTS synthesizers. The system receives an input text and possibly didascalia. Visual speech parameters are generated using the analyzed text and data from an existing audiovisual dataset. The trajectory generation module usually performs a statistical mapping between symbolic information and parametric representation of the signals. The audiovisual speech is then synthesized using a vocoder and an image synthesis module.

Strictly classifying facial animation techniques is a difficult task, because exact classifications are complicated by the lack of exact boundaries between methods and the fact that recent approaches often integrate several methods to produce better results. For instance, Bailly et al [Bailly et al., 2003] classify the audiovisual speech synthesis approaches into two major groups, namely "model-based synthesis" and "image-based synthesis". On the other hand, Theobald [Theobald, 2007] uses three categories of systems: "articulatory synthesis", "rule-based synthesis" and "concatenative synthesis". In an extended survey

on audiovisual speech synthesis, Mattheyses et al [Mattheyses and Verhelst, 2015] propose several classification criteria, such as the output modality and dimensions. Output modality of visual TTS systems is split into *single-phase* (the synthetic auditory mode and the synthetic visual mode are generated at the same time, for instance by articulatory synthesis) or *two-phase* (the synthetic auditory speech is generated by an acoustic-only speech synthesizer which also provides a target phoneme sequence and its corresponding phoneme durations which are given as input to a visual speech synthesizer that generates synchronized synthetic visual speech).

Output dimension is coarsely divided into 2D-rendered and 3D-rendered visual speech. 3D-based visual speech systems use 3D polygon meshes consisting of vertices and connecting edges to represent a virtual human. Therefore, a major advantage of 3D-based facial animation is the ability to freely change camera viewpoint and rendering properties. In addition, it offers a convenient way to add visual prosody to the synthetic speech, since gestures like head movements and eyebrow raises can easily be imposed on the 3D mesh. This convenience comes at the expense of time-consuming design and heavy calculation, especially when realism is important.

2D visual speech systems aim to mimic video recordings by pursuing photorealism. An obvious disadvantage is the difficult integration of audiovisual prosody and editing possibilities. Also the applicability in virtual worlds or computer games is limited, as these are mostly rendered in 3D. On the other hand, a 2D visual speech signal can be applied in numerous other applications such as television broadcast or teaching assistants.

A few systems cannot be classified as either 2D-based or 3D-based synthesis, an example being the usage of synthetic 2D visual speech as texture map for a 3D facial polygon mesh. The resulting speech can be considered "2.5D", since there is no speech-correlated depth variation of the 3D shape.

The following sections will present notable work on TTS systems using a rough categorization for visual speech synthesis techniques based on the main types of approaches used: rule-based, exemplar-based and model-based.

2.4.1 Rule-based approaches

The success of rule-based systems depends on the complex set of rules defined to simulate the desired behavior. Rules in speech animation systems are defined according to observations and studies such as visual cues analysis for audiovisual prosody. In general, only a few frames of the output video signal are predicted directly by the predefined rules, while the rest of frames are obtained using interpolation.

One of the early visual text-to-speech systems was developed by Cassel et al [Cassell et al., 1994] using a rule-based approach. It automatically generates and ani-

mates conversations between multiple agents with appropriate and synchronized speech, intonation, facial expressions and hand gestures. A dialogue planner produces the text and generates the intonation of the utterances. Visual cues are driven from the relationship status between speakers, the text and the intonation. Coordinated arm, wrist, and hand motions are invoked to create semantically meaningful gestures. Besides being strongly correlated, speech and gestures are systematically synchronized at different scales: phoneme, word, sentence.

This work is continued by Pelachaud et al [Pelachaud et al., 1996], a system which generates facial expressions and head movements in conjunction with speech synthesis, including expressive intonation. The focus is placed on expressions conveying information correlated with the intonation of the voice: differences of timing, pitch, and emphasis. The prosodic functions studied include: focus, topic, theme and rheme, given and new information, affect and emotion. The system embodies rules that describe and coordinate the relations: intonation/information, intonation/affect, and facial expressions/affect. Then algorithms impose synchrony, create coarticulation effects, and determine expressive signals, eye and head movements.

Another rule-based approach is used by Albrecht et al [Albrecht et al., 2002a] to generate facial expressions, head motion and voluntary blinks from text input. They use the MARY text-to-speech system [Schröder and Trouvain, 2003] to obtain the speech and phoneme durations. The linguistic analysis carried provides a phonetic transcription used to generate speech-synchronized motion. Emoticons are used as a simple emotional representation for text and include six types: happy, sad, surprised, kidding, angry, disgusted. They are only applied to the visual component of the system. The lip synch algorithm is an improved version of the rule-based method proposed by Cohen and Massaro [Cohen and Massaro, 1993]. The nonverbal facial expressions are added according to well-defined rules: eye blinks are generated as punctuators during pauses to structure the sentence and then every 4.8 seconds (the average natural rate for cornea moistening), eye gaze is directed towards the user during question posing, head and eyebrows are raised at the end of a question, pitch local maxima generates head and eyebrow raising, smiles or frowns appear for the duration indicated by emoticon usage etc. Figure 2.7 presents results of the algorithms described.

The Behavior Expression Animation Toolkit (BEAT) [Cassell et al., 2004] and Spark [Vilhjálmsson, 2004] allowed animators to input typed text and to obtain as output appropriate and synchronized non-verbal behaviors and synthesized speech for human-like characters. Behavior included gaze, turn-taking, backchannel feedback in a form that can be sent to a number of different animation systems. The non-verbal behaviors rely on rules derived from extensive research on human conversational behavior [Kendon, 1994], [Dunbar et al., 1997].

To avoid replication of work, and as well as to allow for sharing modules, an international

Figure 2.7: Examples of generated facial expressions. Facial expression animation obtained by Albrecht et al [Albrecht et al., 2002a].Top: only jaw and lips movements are generated; bottom: nonverbal facial expressions are added.

effort [Kopp et al., 2006] was initiated to develop a common specification to modularize the design of conversational characters. A three stage model was proposed, called SAIBA (situation, agent, intention, behavior, animation), where the stages represent intent planning, behavior planning and behavior realization. XML-based languages were proposed to mediate between these stages: a Function Markup Language (FML) describing intent without referring to physical behavior and a Behavior Markup Language (BML) describing desired physical realization (see figure 2.8).

Figure 2.8: SAIBA framework for multimodal generation [Kopp et al., 2006]. The FML and BML languages mediate between the intent planning, behavior planning and behavior realization modules. The Behavior planning module refers to a gesticon which mainly stores exemplars of socially acceptable gestures.

Many studies have adopted this framework which is continuously being developed. Major issues underlined and new propositions towards the implementation of FML language are presented in [Lee et al., 2008][Heylen et al., 2008][Cafaro et al., 2014]. Sim-

ilar propositions are made towards introducing a perception [Scherer et al., 2012] and emotive markup language [Schröder et al., 2007].

2.4.2 Exemplar-based approach

Exemplar-based or concatenative speech synthesis has become a popular approach over the last decade. For example, in the field of auditory synthesis, concatenative synthesis approaches have almost replaced rule-based techniques in commercial products. An exemplar-based synthesis system requires a database containing original audiovisual speech recordings from a single speaker. To synthesize novel speech, a path search algorithm is applied to find for suitable segments that match the target sequence. The final signal is obtained by concatenating these segments.

The major benefit of exemplar-based synthesis is the fact that a maximal amount of original speech data is reused for generating new visual speech, thus displaying a high degree of realism and naturalness.

One of the first works on exemplar-based visual TTS was introduced by Ezzat et al [Ezzat and Poggio, 1997] for video realistic speech. A visual corpus of keywords is initially recorded, containing one keyword for each English American viseme. Coarticulation effects are ignored as a one-to-one mapping between phonemes and visemes is assumed. Next, one single image for each viseme is identified by manual search through the database. A transformation function based on two-ways optical flow computation is defined between any two visemes. Then, a TTS system is used to convert the input text to a string of phonemes. Using this information, an appropriate sequence of viseme transitions is obtained and the final visual sequence is returned by concatenating the viseme transitions.

Ouni et al [Ouni et al., 2013] present a visual TTS technique which performs simultaneous synthesis of acoustic and visual components. This is done by concatenating bimodal diphone units which consist of both acoustic and visual information. Visual synthetic speech is therefore created using original audiovisual articulations and coarticulations.

Exemplar-based techniques have also been used in hybrid synthesis approaches that are based on both statistical modeling and reusing original data have been proposed. A frequently used machine learning technique for such systems is the hidden Markov model(HMM). Govokhina et al [Govokhina et al., 2006] propose a trajectory formation system for generating speech-related facial movement. The system combines HMM-based trajectory synthesis with segment selection and concatenation.

Another work which uses an exemplar-based technique combined with a statistical model is introduced by Wang et al [Wang et al., 2010]. Wang proposes an HMM-based synthesis approach to photo-real talking head synthesis. This method is combined with a

unit-selection algorithm in the shape of a trajectory-guided real sample concatenating method for generating lip-synced articulator movements. The HMM predicted trajectory optimally guides the search for a smooth lips sample sequence from the image sample library (see figure 2.9). As the sequence is found, the corresponding lip images are then stitched back to a background head video.

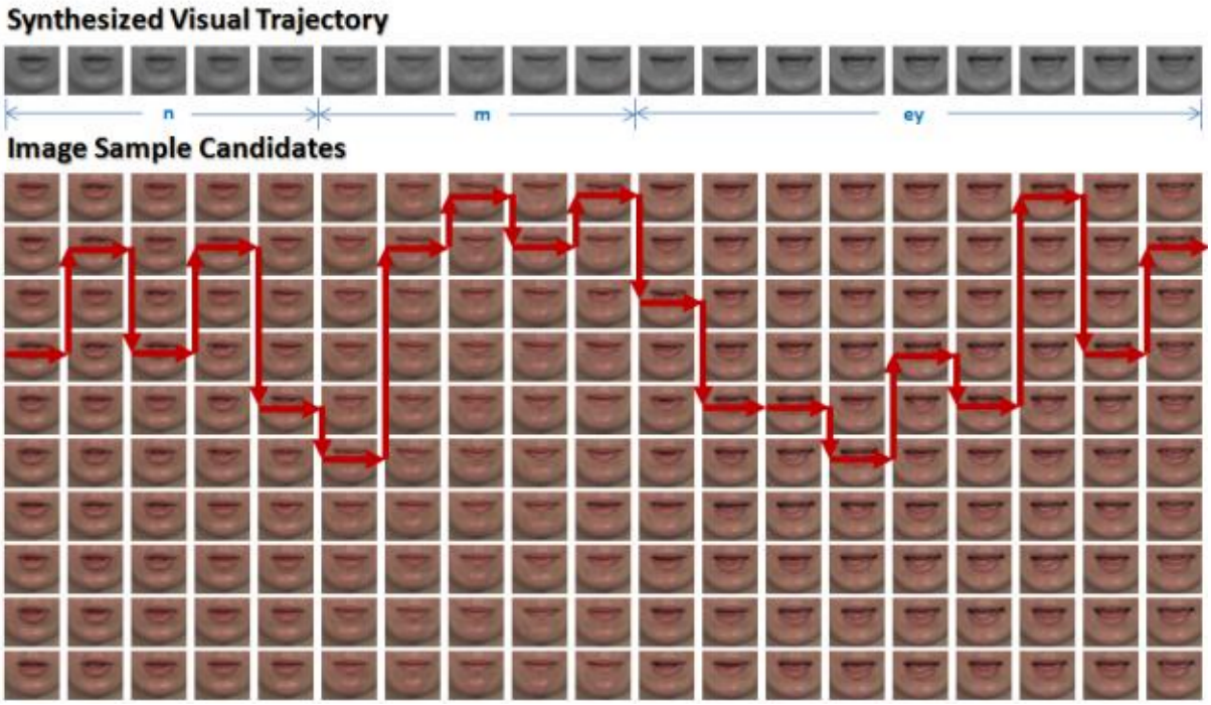

Figure 2.9: Example of HMM trajectory. The top-line lips images (gray) are the HMM predicted visual trajectory. The bottom images (colored) are real samples lips candidates where the best lips sequence (red arrow path) is selected. [Wang et al., 2010]

A similar method is used by Liu et al [Liu and Ostermann, 2011] to generate expressive talking head from a database containing video capture for neutral performances, smiling while speaking and smiling after speaking.

Since these systems are based on unit selection, the high degree of variation in appearance during expressive speech requires large databases in order to combine articulation of segments together with a large variety of facial expressions.

2.4.3 Model-based approach

Model-based approaches imply a statistical model training stage using correspondences between captured properties of original speech and the corresponding phoneme and/or viseme sequence. The model has to take into account both static (the relationship between observed features and the corresponding phoneme or viseme) and dynamic prop-

erties (the transitions between feature sets).

Model-based approaches combine the advantages of rule-based and exemplar-based synthesis: observed motion can be reused without needing to explicitly model behavior, while no original speech data needs to be stored after the training stage. A disadvantage is the fact that original audiovisual speech must be parameterized in order to train the model. The predicted speech does not reuse original data, instead it is resynthesized from predicted parameter trajectories, which can lead to a degraded signal quality.

Tamura et al [Tamura et al., 1999] use an arbitrary given text and synthesize auditory speech and lip motion. An algorithm from previous work [Masuko et al., 1996] is used to generate parameters from HMMs with dynamic features. Audio and visual features for each speech unit are modeled by a single HMM. Results show that triphone models as speech units perform better than syllable models.

Malcangi [Malcangi, 2010] proposed a TTS system that uses two artificial neural networks (ANNs): for text-to-phones and for phones-to-viseme synthesis, respectively. Afterwards, smooth trajectories are obtained using an interpolation based on fuzzy logic. The dynamic control is used to modulate the amplitude of lip-opening strength, resulting in more natural movement. The system can also dynamically control expression muscles to produce modifications such as frowning with eyebrows, wrinkling the forehead etc.

An expressive visual TTS system is proposed by Anderson et al [Anderson et al., 2013]. Given an input text and continuous expressive weights, the system generates an expressive talking head animation, where the face model is based on an extended active appearance model (AAM). The standard AAM is extended to allow the separation of models for global and local shape and for appearance deformations. The key technology behind the expressive voice and gestures method is cluster adaptive training (CAT) which extends on the HMM-based synthesis. The clusters are trained on an expressive dataset containing neutral, angry, happy, tender, sad and fearful expressions such that any combination of input expressive weights allows the exploration of this "space of emotions". This system can be considered 2.5D-based as it uses image texture maps on a 3D polygonal model. According to the authors, the combination of 2D texture maps with superimposed illumination creates a near-videorealistic effect. Figure 2.10 presents results obtained for the large range expressions.

Figure 2.10: Example of photorealistic expressions. Left to right: example synthesis for neutral, tender, happy, sad, afraid and angry [Anderson et al., 2013].

Discussion. Some of the presented works return 2D-based speech synthesis which presents a high degree of static photorealism. However, this comes at the expense of limited control of motion and rendering options. Exemplar-based approaches require large dataset sizes for each recorded speaker, especially if expressive styles are involved. Smaller datasets are required by hybrid or model-based systems. However, in terms of expressiveness, the few works that have treated this problem are limited to the set of basic cardinal emotions and do not parse finely the expressive space.

2.5 Speech-driven animation

Another popular approach for generating facial speech animation implies the use of existing speech performance to drive realistic face motion. Figure 2.11 presents the general outline of a speech-driven animation system.

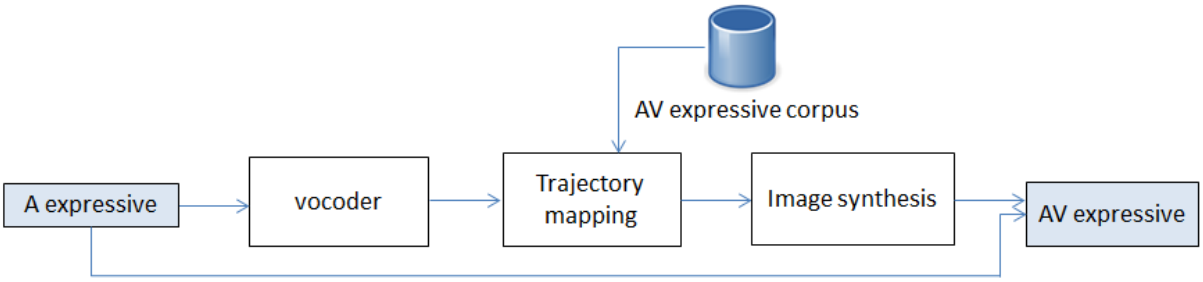

Figure 2.11: Outline of speech-driven animation systems. The system receives an audio (A) signal as input, uses a vocoder to extract the acoustic features and generates visual trajectories using existing audiovisual (AV) data. The visual speech (V) is returned by an image synthesis module.

We can generally categorize the speech-driven animation techniques as: exemplar-based and model-based.

2.5.1 Exemplar-based approaches

Video Rewrite [Bregler et al., 1997] was used on video data to derive novel motion using a triphone model. Deng et al [Deng et al., 2004] use such a method to generate head movements from an acted dataset. Based on the aligned pairs between audio features and head motion, a K-Nearest Neighbors (KNN)-based dynamic programming algorithm is used to synthesize the new head motion. However, this method has only been tested for neutral speech.

Bregler et al [Chuang and Bregler, 2005] use a example-based technique to generate expressive head motion for 3 speaking styles: happy, angry and neutral. Initially, a database

containing synchronized voice pitch and head motion is collected. For a given sentence, the pitch voiceless regions are used to segment the pitch and motion curves. For a new segmented voice input, a matching sequence of motion segments is found in two steps: finding similar pitch sequences in the database for the whole sentence (in order to keep emotional, idiosyncratic shapes which are conveyed at sentence-level) and then comparing individual pitch segments against existing ones in the database. A combined distance metric for the two steps is then used in a path search algorithm for the best matching motion segments. The algorithm is detailed in figure 2.12.

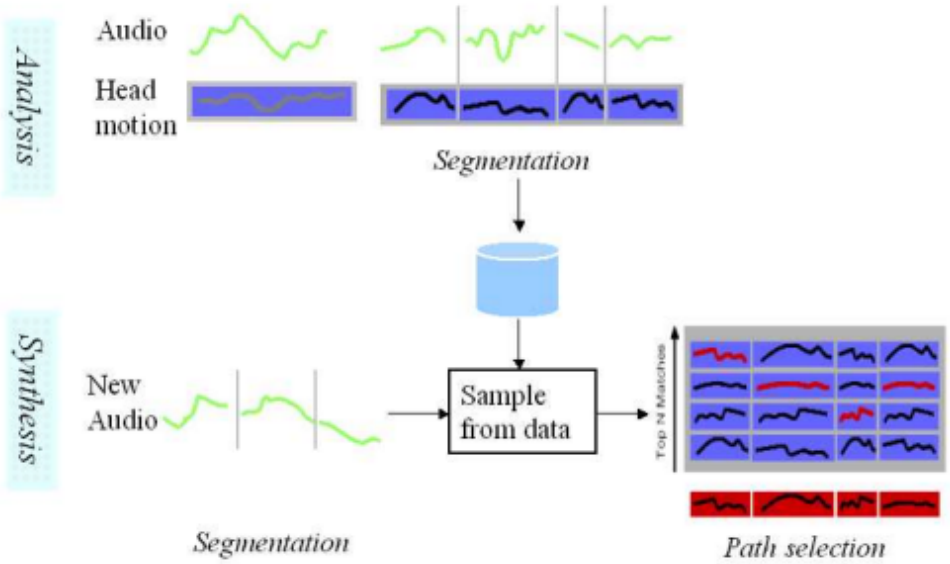

Figure 2.12: Head motion synthesis overview [Chuang and Bregler, 2005].

The quality of the speech animation for sample-based approaches depends on the amount of data available. The appearance of visemes is highly context dependent, so examples of each phoneme in as many different contexts as possible are required in order to obtain convincing correct results.

2.5.2 Model-based approaches

Some of the earliest lip-sync systems used statistical models. These models attempt to eliminate the need for large example databases by creating compact statistical models of face motion. For example, Voice puppetry [Brand, 1999] applies a hidden Markov model to audio and facial motion and predicts the most likely facial motion given an audio input.

A study involving Gaussian Mixture Models (GMMs) is carried by Costa et al [Costa et al., 2001], where audio to visual data GMM mapping is used. The data is

not annotated using emotional tags. The author mentions having obtained encouraging expressive results but no evaluation is presented.

As statistical techniques became a popular approach to lip-sync animation, the expression of emotion was largely ignored. One of the first works to address the problem of generating expressive speech animations is carried by Bregler [Chuang and Bregler, 2005]. While expressive head motion is generated using an exemplar-based approach, Bregler proposes a bilinear model for facial expressions spanning 3 emotional styles: happy, angry and neutral with the goal of editing existing facial motion.

Cao et al [Cao et al., 2005] introduce a new approach that is able to synthesize an expressive animation from an input utterance and a set of emotional tags. The approach uses database of high-fidelity recorded facial motions, which includes speech-related motions with variations across multiple emotions: Happy, Angry, Neutral, Sad, Frustrated. Independent Component Analysis (ICA) [Hyviirinen et al., 2001] is used to separate the content (speech) from the style (expressive) and for each sentence, a sequence of nodes is extracted, where each node represents a phoneme. The associated visual representation of a phoneme is called an anime and an anime graph is built from the extracted data (including the motion segment and other audio information). For an input utterance, the new motion is found by running a cost-based graph search algorithm of the anime graph. Then, expressive motion is obtained by mapping between two emotion spaces, each mapping function being trained with a Radial Basis Function (RBF) approximation [Buhmann, 2003].

Busso et al [Busso et al., 2007] use separate HMMs to synthesize new head motion for different emotions: happiness, anger, sadness and the neutral state. The problem is treated as classification of discrete representations of head poses. Realistic and smooth head motion is obtained by constraining the transition between clusters and by using quaternion representation to interpolate rotations. Several perceptual tests showed that, on average, synthesized head motion was perceived as realistic as the original head motion.

Levine et al [Levine et al., 2009] also proposed a real-time solution for the generation of appropriate body language animations from speech signals. Segment selection is driven by a HMM and uses prosody-based features extracted from speech. This study is continued with the work on gesture controllers [Levine et al., 2010], where kinematic properties of the motion are decoupled from its shape and learned separately. The controller presents a dedicated inference layer implemented by a conditional random field trained on the gesture kinematics and underlying audio data. This method consistently outperforms the previous approach which directly associates gesture segments with prosody features.

Techniques based on HMMs are also proposed by Ding et al [Ding et al., 2013] for

expressive eyebrow motion synthesis. The study shows that contextual models (able to use contextual speech information) significantly outperform the baseline method [Hofer et al., 2007], which proposes an HMM trajectory model for motion synthesis. According to the MSE criterion, best results are obtained by combining contextual HMMs with CRFs. The system performs better if the sequence of labels (anger, fear, sadness, surprise, no move) is known but this information does not affect too much the synthesized motion stream.

Taylor et al [Taylor et al., 2012] define a new dynamic unit for visual speech which represents contrastive movements of the speech articulators. The dynamic nature of the unit means that coarticulation effects are explicit in the model and the boundaries between visemes are not tied to the boundaries of the underlying phones. A large corpus of video speech data is modeled using an AAM and then segmented according to zero-crossing points in parameter space acceleration. These variable sequences are modeled as HMM super-features. Then dynamic visemes are created by clustering the super-features. A new motion is synthesized by simply stitching together the dynamic viseme cluster centers using simple spline interpolation at the unit boundaries (see figure 2.13 for results obtained with this method). Compared to static viseme synthesis, this technique receives better perceptual evaluation rates. However, this approach cannot be transferred between speakers and only neutral speech is considered.

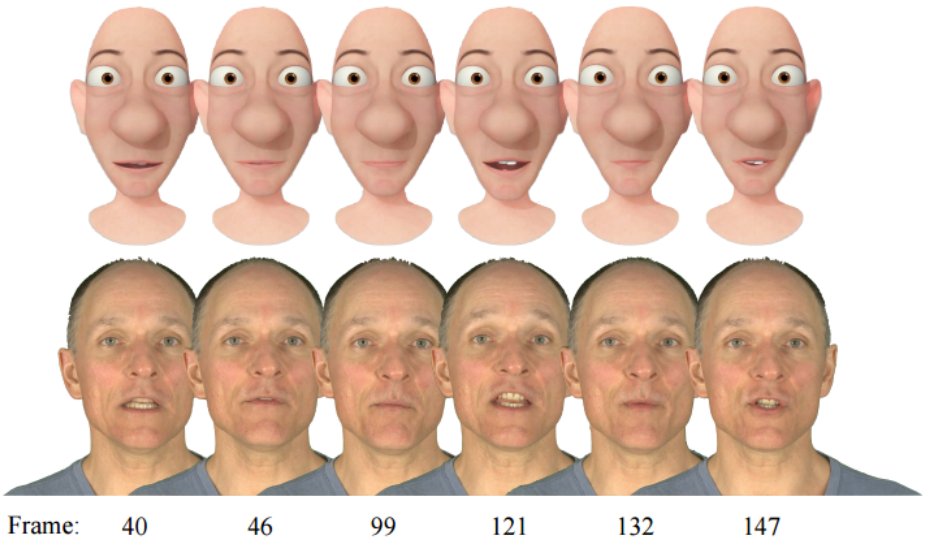

Figure 2.13: Example frames for dynamic units of visual speech. Frames from a synthesized sentence for a face model (top) and corresponding video frames (bottom) [Taylor et al., 2012].

Le et al [Le et al., 2012] propose a novel, fully automated framework to generate real-time head, gaze and eyelid motion, based on live or pre-recorded speech input. Each component is modeled using separate but inter-related statistical models for each component from a pre-recorded facial motion dataset. Specifically, GMMs and gradient descent op-

timization algorithms are used to generate head motion from speech features. Then eye gaze is generated from head motion and speech features from Nonlinear Dynamic Canonical Correlation Analysis and eyeblinks are predicted using Non-negative linear regression (voluntary blinks) and log-normal distribution (involuntary blinks). Comparative user studies show that the proposed framework outperforms state-of-the-art facial gesture generation algorithms ([Ma and Deng, 2009] [Lee et al., 2002] [Levine et al., 2009] [Busso et al., 2005] [Chuang and Bregler, 2005]).

Marsella et al [Marsella et al., 2013] uses a hybrid system which combines a speech-driven model-based technique with a rule-based visual text-to-speech synthesizer in order to generate expressive performances for a 3D virtual character using prosody and sentence semantics. The system is able to generate head motion, eye saccades, eye blinks, gazes and gestures using a complex set of rules derived through a study of video corpora of human behaviors. The generation process involves 6 steps: (1) acoustic processing: returns information related to stressed words and overall speaker agitation (interpreted as sad, neutral or agitated); (2) syntactic analysis: returns syntactic structure; (3) function derivation: derives the meaning relevant to nonverbal behavior; it uses forward-chaining inference rules to build a structured analysis and is implemented by function derivation classes (example: mental state); (4) behavior mapping: nonverbal behavior rules map from communicative functions to behavior classes; (5) animation specification: behavior classes are mapped to specific behaviors while supporting elements such as personality, culture, body type etc (6) animation synthesis: the scheduled behaviors are processed and performance synthesis is carried with the option of introducing extra constraints. Over an online evaluation test, results are compared to animations showing random gestures and prosody-based beats. The results obtained using this approach are perceived as considerably more realistic.

Figure 2.14: Examples of full-body gestures. Frames illustrating generated facial expressions and gestures for two virtual characters [Marsella et al., 2013].

Discussion. Among the disadvantages of statistical methods, we consider the lack of flexibility, as the underlying mapping functions are highly speaker-specific. Moreover, not all of the facial motion during speech can be determined from the acoustics. One important criteria is ensuring that the chosen features and mapping function capture the perceptually significant variance, and that any error in the mapped facial motion is not in perceptually important speech regions.

A notable paradigm shift was caused by the emergence of cheap and efficient technology for recording both voice and 3D motion, notably the Kinect camera. Therefore the usefulness of speech-driven animation is diminished while more research is directed at visual TTS techniques.

2.6 Expressive audiovisual conversion

Audiovisual conversion consists in changing one aspect of speech such as: speaker conversion (making the converted speech appear as if uttered by another speaker) or expressive conversion (changing the expressive style of speech). We are interested in the expressive conversion of audiovisual speech in the context of adding expressiveness to neutral speech. Figure 2.15 presents the general outline of an expressive audiovisual conversion system.

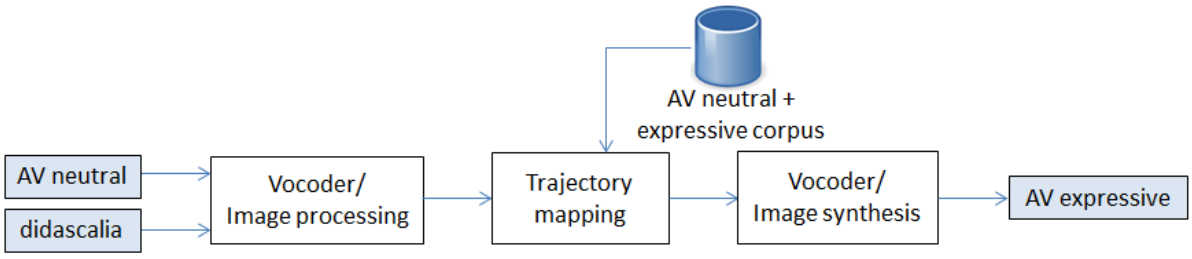

Figure 2.15: Outline of expressive audiovisual conversion systems. The system receive as input the neutral version of audiovisual speech and didascalia, extracts the audiovisual parameters which are converted to expressive trajectories using an audiovisual corpus. The expressive audiovisual speech is generated using a vocoder and an image synthesis module.

The topic of expressive speech conversion has mostly been studied for separate modalities i.e. audio or visual-only. For this reason we present the current techniques separately for voice and facial animation.

2.6.1 Voice

The earliest attempts of synthesizing expressive speech used several approaches [Cahn, 1989] [Murray and Arnott, 1995] [Bulut et al., 2002], mostly based on a concatenative speech synthesis framework. This framework provides a high quality of expressive speech because the new speech is synthesized by concatenating units of natural speech at different levels. However this method is limited by the availability of a large amount of data of speaker-specific expressions. For example, Schroeder [Schröder and Breuer, 2004] and Eide [Eide et al., 2004] generated an expressive text-to-speech engine which can be

controlled, via an extended speech synthesis markup language, to use a variety of expressive styles from about 10 h of expressive readings of sentences with neutral linguistic content.

An alternative to concatenative expressive speech synthesis consists in storing only neutral speech and synthesizing the required expressive speech by modifying the excitation source and modulating spectral information. Mori et al [Mori et al., 2006] analyzed face variations of natural speech and proposed a synthesis method for the fundamental voice frequency f_0 by identifying expressive subspaces in raw speech, such as: anger, boredom, depression etc. Wo et al [Wu et al., 2010] proposed a hierarchical prosody conversion where structure consisted of sentence, prosodic word, and subsyllable levels.

Gaussian mixture models are widely used in spectrum conversion to modify non linguistic information such as voice characteristics while keeping linguistic information unchanged. A joint density GMM is trained on aligned source and target audiovisual features using a minimum mean square error based solution [Stylianou et al., 1995] and the maximum likelihood trajectory generation solution [Toda et al., 2007]. Veaux et al [Veaux and Rodet, 2011] proposed a GMM-based f_0 conversion system for joy, fear, sadness and anger, in which the converted f_0 contour is generated under dynamic feature constraints. Aihara et al [Aihara et al., 2012] used GMMs for converting both spectrum and f_0 for anger, sadness and joy. In a comparative study for prosody conversion, Tao et al [Tao et al., 2006] test different models: a linear modification model (LMM), a GMM and a classification and regression tree (CART) model. According to the objective and perceptual tests, the best results for a small training dataset are obtained by using GMMs to convert spectrum and prosody, while CART models present better emotional speech output if trained with a large context-balanced corpus. Note however that the resulting mapping is sensitive to phonetic cues but relatively insensitive to positional cues within higher structural levels such as the syllable, phrase or sentence.

2.6.2 Facial animation

Early approaches in expressive facial animation used geometric meshes with embedded pseudo-muscles activations or more complex physically-based models [Waters, 1987][Lee et al., 1995]. These approaches are very intuitive but they lack dynamic expressiveness while being computationally expensive.

More recently, statistical approaches have been applied to expressive video sequences [Chuang and Bregler, 2002] or to motion capture data [Cao et al., 2003] [Vlasic et al., 2005] [Cao et al., 2004] in an attempt to synthesize speech and facial expression. Such approaches include bilinear [Chuang and Bregler, 2002] [Chuang and Bregler, 2005] and trilinear models [Vlasic et al., 2005], both of which factorize expressive speech into separate components so that parameterized neutral speech

sequences can be modulated with expression parameters. Ma et al [Ma and Deng, 2009] use Gaussian processes to generate a mapping function between any two emotion styles from the set: neutral, angry, happy.

Discussion. To our knowledge, most techniques used in expressive conversion of audio-visual speech follow frame-based approaches. While presenting expressive contours, the synthesized performances have the same rhythm as the neutral source stimuli. Another limitation of these works is represented by the set of target emotions, usually represented by a limited number of basic emotions.

2.7 Summary

This chapter covered a brief description of state-of-the-art techniques and theoretical concepts used in several topics which are strongly related to our work: virtual actors, auditory and visual prosody, speech-driven animation and expressive text-to-speech animation.

Section 2.1 discussed the challenges implied in the creation of believable virtual humans. Virtual humans present high-quality appearance which can now be achieved at cheap costs. Many systems have been proposed for audiovisual speech generation, using complex rules, exemplars and/or parametric models. However, the expression of emotion and attitude remains a difficult problem.

Automatically generating expressive speech for virtual actors implies studying the prosody of real-life data. Section 2.2 presented this topic and was continued with visual prosody in Section 2.3. Many valuable observations have been drawn from a large amount of studies. Typically, tight correlations have been found between vocal tract features and mouth area motion, high correlation between prosody and lip movements and weaker correlation between prosody and features such as head motion, eyebrow motion, gaze, gestures, blinks. Another important conclusion is that contour shapes are also emotion-dependent and that correlation values depend on the scale: word, phrase, sentence.

These observations are important for the process of automatic behavior generation as it enables the creation of joint audiovisual models which correctly describe audiovisual patterns to create perceptually realistic motion. The relationship between speech and motion varies from emotion to emotion, indicating that specific emotional models are needed for expressive facial animations.

State of the art techniques for the generation of audiovisual speech were presented in Sections 2.4, 2.5 and 2.6. Speech-driven animations, visual text-to-speech approaches and expressive audiovisual conversion systems have treated a limited set of basic emotions.

However, improvements are required and open questions remain related to the dynamics of visual prosody contours, complex attitudes etc. We decided to focus on a larger set of *discrete attitudes* which would be more appropriate for dramatic performances. As described in this chapter, the analysis and modeling of expressive speech imply well-defined steps and are based on expressive audiovisual datasets. In order to create our virtual actors, we design and record an expressive corpus.

In the presented studies, audiovisual prosody is not modeled explicitly, so the global dynamics of prosodic contours are not explored. However, some studies have shown that speech prosody for attitudes is encoded via specific signatures and we want to see if such *signatures* exist also in the visual prosodic contours.

Rhythm can be considered both audio and visual. In the case of speech-driven techniques, the rhythm is provided by the speech. In the case of expressive conversion, most techniques take the rhythm of the neutral performance. For visual text-to-speech systems, the rhythm is predicted, in most cases by a TTS engine. In our approach the *rhythm* will be treated as an additional parameter.

Finally, most of the presented approaches are frame-based in the sense that they use frame-by-frame methods to generate trajectories from an input. A few others use also the syllable as unit, especially for processing pitch segments. We will compare two types of methods: one which uses blind *frame-based* conversion, with a local speech rate prediction, and one that uses *sentence-based* conversion, in which the prosody, including the rhythm, are sensible to the position within the sentence.

Chapter 3

Dataset of dramatic attitudes

With the goal of synthesizing audiovisual prosody for realistic social contexts, our work includes the analysis and modeling of these specific prosodic signatures. The extraction of prosodic shapes requires sufficient statistical coverage of the metalinguistic functions at varied positions and sentence sizes. For this reason, we designed and recorded an expressive dataset of dramatic attitudes. The chapter includes a short review of expressive datasets and describes the process of selection and recording of attitudes. We also detail our pre-processing techniques and evaluation platforms for synthetic data representation.

3.1 Expressive datasets

Although recent years have brought a substantial progress in the field of affective computing [Zeng et al., 2009], the development of emotion-modeling systems strongly depends on available affective corpora. As training and evaluation of algorithms require a great amount of data which is hard to collect, publicly available datasets represent a bottleneck for research in this field. Moreover, the majority of available datasets are limited to the six basic emotion categories proposed by Ekman [Ekman and Friesen, 1971] and only include happiness, sadness, fear, anger, disgust, and/or surprise.

Databases containing affective data can be categorized under several criteria: data types used (2D or 3D visual data, speech), spontaneity (naturalistic, artificially induced or posed by professional actors or not), affective state categorization (emotion, attitudes etc). Audiovisual recording is obviously more expensive and time-consuming than audio-only recording. This is proven by the comparative amounts of publicly available audio and audiovisual datasets. For instance, the Interspeech Computational Paralinguistic Challenge [1] provides audio data from a high diversity of speakers and different languages

[1] http://compare.openaudio.eu/

((Non-native) English, Spanish, and German).

A comprehensive overview of the existing audiovisual corpora can be obtained from [Cowie et al., 2005][Zeng et al., 2009]. Table 3.1 presents a set of expressive datasets which are most relevant to our work.

The works listed in the table present publicly available data that are used in several research topics: analysis, affective recognition, expressive performance generation, audiovisual conversion etc. IEmbarrassedOCAP [Busso et al., 2008] and CAM3D [Mahmoud et al., 2011] contain motion capture data of the face and upper-body posture from spontaneous performances. Large data variability is presented by Bosphorud [Savran et al., 2008] and CK+ [Lucey et al., 2010] as they include more than 100 subjects posing over 20 expressions each, in the shape of action units and combinations. Other special mentions are the MPI facial expressions dataset [Kaulard et al., 2012] which includes 55 states (basic emotions and attitudes) and the Mind Reading dataset [Baron-Cohen, 2003] which includes video recordings of 412 expressive states classified under 24 main categories. While these serve as valuable references for the expressive taxonomy, these datasets often do not contain audio data.

To our knowledge, the only publicly available affective datasets that include 3D data and speech are the BIWI, CAM 3D and 4D Cardiff corpuses. Although the expressive categories contained extend the set of basic emotions, only a few present conversational potential (thinking, confused, frustrated, confidence). Most importantly, the sentences used in the datasets do not present a high variability or a systematic variation of syllable lengths.

The expression of attitudes is highly dependent on the studied language. The following works focus on the study of attitudes and/or generation of prosody (intonation): [Morlec et al., 2001] (French), [Mac et al., 2012] (Vietnamese), [De Moraes et al., 2010] (Brazilian Portuguese), [Hönemann et al., 2015](German). However, the datasets recorded for these studies are not publicly available and only few [Mac et al., 2009] [Hönemann et al., 2015] feature audio-visual data.

We designed and recorded an expressive corpus consisting of attitudes performed by three French speakers. The corpus is designed to include sentences of varied sizes. The data gathered consists both in audio, video and 3D motion capture of the recorded performances. The following sections describe the recording process.

Table 3.1: Datasets for affect recognition systems.

	# subjects	# samples	Data type	Speech & # sentences	Categories	Spontaneity
BIWI [Fanelli et al., 2010]	14	1109	3D face & video	Yes - 40	11 affective labels	acted
4D Cardiff [Vandeventer et al., 2015]	4	N/A	3D face & video	Yes - N/A	10 expressions (basic emotions and attitudes)	spontaneous
MPI facial expressions [Kaulard et al., 2012]	6	N/A	3D face	N/A	55 expressions (basic emotions and attitudes)	acted
Bosphorus [Savran et al., 2008]	105	4652	3D face (static)	N/A	37 expressions (action units and basic emotions)	acted
CK+ [Lucey et al., 2010]	123	593	video	N/A	23 expressions (action units and combinations)	posed and non-posed
CAM3D [Mahmoud et al., 2011]	7	108	3D face & upper body	Yes - N/A	12 mental states	spontaneous
Mind Reading [Baron-Cohen, 2003]	6	N/A	video	Yes - N/A	412 emotions in 24 categories	acted
IEmbarrassedOCAP [Busso et al., 2008]	10	N/A	3D face & upper body	N/A	8 emotions	acted and spontaneous

3.2 Selected attitudes

Our corpus consists of "pure" attitudes, i.e. isolated sentences carrying only one attitude over the entire utterance.

Two cognitively different categories of attitudes can be distinguished: social and propositional, whether they refer to the interpersonal relationship of the speakers (seductive, polite etc) or the propositional content of the sentence (doubt, irony etc) respectively. As we refer to acted expressions of attitudes with the goal of recreating dramatic dialogues, we coin the term of *dramatic attitude*. This is partly due to the fact that we want to create virtual actors that are capable of acting a chosen set of attitudes, but also because the actual expressive dataset from which we will analyze prosodic trajectories, is an acted corpus. A similar denomination, *dramatic expression* [Johnson, 2003] is used in the context of dramatic portrayal in opera and application to the development of embodied conversational agents.

A starting point for the possible attitudes was the Baron-Cohen's Mind Reading project [Baron-Cohen, 2003]. The taxonomy proposed in this work gathers a total of 412 emotions grouped under 24 main categories, each comprising several layers of subexpressions. Due to the complexity of the taxonomy, we choose a limited number of attitudes which are possibly expressed in one selected act from a play by Arthur Schnitzler, translated in French as "La ronde" by Maurice Rémon and Wilhelm Bauer [Schnitzler, 1912] and in English as "Hands around" by Marya Mannes [Schnitzler, 1920]. This play nicely consists in a series of seduction pitches during dyadic face-to-face conversation. The characters involved in the scene are the Count and the Actress. The set of short turns exhibit a large variety of emotions and attitudes, intimately associated with the verbal content.

Table 3.3 contains the list of attitudes we decided to study, to which the three basic modalities have been added: assertion (declarative), exclamation and full question (interrogative)(see table 3.2). The correspondence to the Mind Reading emotional set [Baron-Cohen, 2003] is indicated in the *Category-Subgroup* and *Definition* fields. The *Name* and *Abbreviation* fields indicate the notations used throughout the manuscript.

Table 3.2: Modal attitudes.

Name	Abbr.	Definition
Declarative	**DC**	Making a declaration; neutral
Exclamative	**EX**	Making an exclamation
Interrogative	**QS**	Making an interrogation; question

Table 3.3: Description of chosen dramatic attitudes. Information in Category-Subgroup and Definition fields is extracted from Mind Reading, while Name and Abbr. are the notation used in the manuscript.

#	Category-Subgroup	Name	Abbr.	Definition
1	Kind-Comforting	**Comforting**	**CF**	Making people feel less worried, unhappy or insecure
2	Fond-Liking	**Tender**	**TE**	Finding something or someone appealing and pleasant; being fond of something or someone
3	Romantic-Seductive	**Seductive**	**SE**	Physically attractive (an attractive quality that tempts in some way)
4	Interested-Fascinated	**Fascinated**	**FA**	Very curious about and interested in something that you find attractive or impressive
5	Wanting-Jealous	**Jealous**	**JE**	Feeling resentment against someone because of that person's rivalry, success, or advantages
6	Thinking-Thoughtful	**Thinking**	**TH**	Thinking deeply or seriously about something
7	Disbelieving-Incredulous	**Doubtful**	**DI**	Unwilling or unable to believe something
8	Unfriendly-Sarcastic	**Ironic**	**IR**	Using words to convey a meaning that is the opposite of its literal meaning
9	Surprised-Scandalized	**Scandalized**	**SS**	Shocked or offended by someone else's improper behavior
10	Surprised-Dazed	**Dazed**	**SD**	So shocked and stunned that you can't think clearly
11	Sorry-Responsible	**Responsible**	**RE**	Feeling that you are the cause of something that has happened and must therefore take the blame for its affects
12	Hurt-Confronted	**Confronted**	**HC**	Approached in a critical or threatening way
13	Sorry-Embarrassed	**Embarrassed**	**EM**	Worried about what other people will think of you

3.3 Recordings

The process of data recording was carried in collaboration with theater director Georges Gagneré of Université Paris 8. Two semi-professional actors performed recordings under the active supervision of the director. In the following sections the actors are named Greg and Lucie, and the director is named Georges.

- Isolated phrases: The director and each actor recorded 35 sentences uttered with each of the selected 16 attitudes (see Figure 3.2). We call "exercices in style" the technique in which the actor separates the text from the semantics and plays the same text in different styles. The name is inspired from Queneau [Queneau, 1947] who uses the method of retelling the same story in 99 different styles. These sentences were selected from a target scene in order to span the distribution of lengths of sentences in the final text (from one syllable up to 21-syllable sentences) and have a large variability of positions and lengths of constitutive grammar groups. Table 3.5 presents the sentences and the number of syllables. Note that due to syllabification and notably the mute "e", the number of syllable may vary for a given sentence.

- Dialogs: The two actors played the selected scene in a neutral and an expressive manner. Greg (male) performs 38 phrases with lengths ranging between 1 and 18 syllables, while Lucie (female) performs 20 phrases with lengths ranging between 1 and 18 syllables. The director (male) also performs the entire dialogue scene in a neutral style. Table 3.4 presents all sentences included in the dialog as performed in French, the translations in English and the attitude associated to each sentence in the expressive version.

Table 3.4: Dialog sentences with didascalia.

Actress	Count	Sentence	Sentence translation
Tender		C'est vous, comte ?	It's you, Count?
	Embarrassed	Madame votre mère m'a autorisé... autrement je ne me serais pas...	Your good mother gave me permission, or of course I wouldn't
Tender		Asseyez-vous, mon cher comte.	Please come right in.
	Comforting	Madame votre mère m'a dit que vous étiez souffrante... mais j'espère que ce ne sera rien.	Your mother said you weren't very well, Fräulein. Nothing too serious, I hope?

Scandalized		Rien ? J'ai été à la mort !	Nothing serious? I was dying!
	Doubtful	Ce n'est pas possible ?	Not really?
Seductive		Vous êtes trop aimable d'être venu.	In any case it's very kind of you to . . . trouble to call.
	Doubtful	À la mort ! Et hier encore vous avez joué comme un archange	Dying! And only last night you played like a goddess!
Fascinated		Oui, ç'a été un vrai triomphe.	It was a great triumph, I believe.
	Fascinated	Un triomphe inouï ! Toute la salle était emballée. Je ne parle pas de moi.	Colossal! People were absolutely knocked out. As for myself, well
Seductive		Merci de vos jolies fleurs.	Thanks for the lovely flowers.
	Embarrassed	Oh ! Je vous en prie...	Not at all, Fräulein.
Seductive		Les voilà.	There they are!
	Jealous	Vous avez été comblée de fleurs, hier.	Last night you were positively strewn with flowers and garlands!
Comforting		Elles sont restées dans ma loge. Je n'ai fait apporter que les vôtre.	I left them all in my dressing room. Your basket was the only thing I brought home.
	Embarrassed Dazed	Vous êtes trop gentille. Mais... mademoiselle...	You're very kind. Fräulein!
Comforting		Ne craignez rien, mon cher comte, cela ne vous engage à rien.	Don't be afraid, Count. It commits you to nothing!
	Fascinated	Vous êtes une personne extraordinaire... Je dirais même énigmatique.	You're a strange creature . . . a puzzle, one might almost say.
Ironic		Mademoiselle Birken vous paraît moins mystérieuse ?	Fräulein Birken is . . . easier to solve?
	Confronted	évidemment, la petite Birken n'est pas compliquée, quoique... Je la connais très peu, vous savez.	Oh, little Birken is no puzzle. Though . . . I know her only superficially.
Ironic		Ah !	Indeed?
	Confronted	C'est vrai.	Oh, believe me.

	Fascinated	Mais vous, vous êtes l'énigme faite femme et le mystère m'a toujours attiré.	But you are a problem. And I've always longed for one.
	Thinking	Ah ! Que de temps, que de plaisir perdu quand je songe que c'est hier... que je vous ai vue jouer pour la première fois.	As a matter of fact, last night I realized what a great pleasure I'd been missing. You see, it was the first time I've seen you act.
Scandalized		Hier seulement ?	Is that true?
	Responsible	Que voulez-vous, mademoiselle, ce n'est pas facile d'aller au théâtre. Je suis habitué à dîner tard... alors, quand on arrive, le plus intéressant est passé. Vous comprenez ?	You see, Fräulein, it's a big problem with the theater. I'm used to dining late. By the time I get there, the best part of the play is over, isn't it?
Jealous		Eh bien, désormais, vous dînerez plus tôt.	You'll have to dine earlier from now on.
	Responsible	J'y avais déjà pensé. Je pourrais aussi ne pas dîner. Dîner n'est pas un plaisir.	I'd thought of that. Or of not dining at all. There's not much pleasure in it, is there-dining?
Seductive		Quels sont exactement les plaisirs que vous goûtez encore, jeune vieillard ?	What do you still find pleasure in, young fogey?
	Thinking	Je me le demande quelquefois. Mais je ne suis pas un vieillard. Il doit y avoir un autre motif.	I sometimes ask myself. But I'm no fogey. There must be another reason.
Seductive		Croyez-vous ?	You think so?
	Thinking	Oui. Mon ami, le comte Lulu, me dit souvent que je suis un philosophe. Vous savez, mademoiselle, il veut dire par là que je réfléchis trop.	Yes. For instance, Lulu always says I'm a philosopher. What he means is: I think too much.
Seductive		Oui... réfléchir, c'est le grand malheur.	He's right . . . it is a misfortune, all that thinking.

Table 3.5: Isolated sentences and the corresponding number of syllables.

No	Text	Text(English)	# of syllables
1	Oui	Yes	1
2	Non	No	1
3	Ah	Oh	1
4	Bonjour	Goodday	2
5	Là-bas	There	2
6	Eh bien	Well	2
7	Mademoiselle	Miss	3
8	Vous savez	You know	3
9	J'ai été	I was	3
10	C'est vrai	It is true	2
11	À la mort	Dying	3
12	Mon cher comte	My dear count	3
13	Un triomphe ïnoui	A colossal triumph	5
14	Un jeune vieillard	A young fogey	4
15	Ne craignez rien	Do not fear	4
16	Je la connais très peu	I know her very little	6
17	Ce n'est pas possible	It is not possible	5
18	J'y avais déjà pensé	I have already considered it	7
19	Votre mère m'a autorisé	Your mother allowed me in	7
20	Je dirais même énigmatique	I might even say enigmatic	8
21	Merci de vos jolies fleurs	Thank you for the pretty flowers	7
22	Il doit y avoir un autre motif	There must be another reason	9
23	Toute la salle était emballée	The audience was excited	8
24	Elles sont restées dans ma loge	They are left in my dressing room	7
25	La petite Birken n'est pas compliquée	Miss Birken is not complicated	10
26	Désormais, vous dînerez plus tôt	From now on, you will dine earlier	9
27	Je pensais qu'à Vienne tout cela changerait	I thought that everything would change in Vienna	11

28	Il veut dire par là que je réfléchis trop	He means that i think too much	11
29	Ce n'est pas facile d'aller au théâtre	It is not easy to go to the theater	10
30	Et hier encore vous avez joué comme un archange	And yesterday you played like an archangel	12
31	Quand on arrive, le plus intéressant est passé	When we arrive everything interesting is over	13
32	Mon ami me dit souvent que je suis un philosophe	My friend tells me often that I am a philosopher	14
33	Ce sont exactement les plaisirs que vous goûtez encore	These are the pleasures you still enjoy	15
34	Mais vous êtes l'énigme faite femme et le mystère m'a toujours attiré	But you are an enigma in the shape of a woman and the mistery has always attracted me	17
35	En Hongrie j'ai fait du service dans les petits trous où j'ai été en garnison	In Hungary I did service in the small towns where I was stationed	21

The recording session began with an intensive training of the actors, who received scenic indications from the theater director. The training consisted in fully understanding the interpreted attitudes and developing the ability to dissociate the affective state imposed by each attitude from the meaning of the text and to maintain a constant voice modulation, specific for each attitude, throughout uttering the 35 sentences. The actors did not receive any instruction related to head movements. The same set of sentences are then recorded by the director himself.

The synchronized recording of voice signals and motion are performed with the commercial system Faceshift [2], a short-range Kinect camera and a Lavalier microphone. Faceshift enables the creation of a customized user profile consisting of a 3D face mesh and an expression model characterized by a set of predefined blendshapes that correspond to facial expressions (smile, eye blink, brows up, jaw open etc). The sampling rate for audio is 44.1 kHz. Recordings are done in front of the camera, while seated, without additional markers such that the acting is not constrained. The use of Faceshift requires soft, non-saturated light conditions. Therefore, the recordings are done in a sound-proof, uniformly illuminated studio. Due to the nature of the desired attitudes and to the eye gaze tracker, the actors perform the isolated sentences as if they are addressing to a person standing in front of them, at the same height.

[2] http://www.faceshift.com/

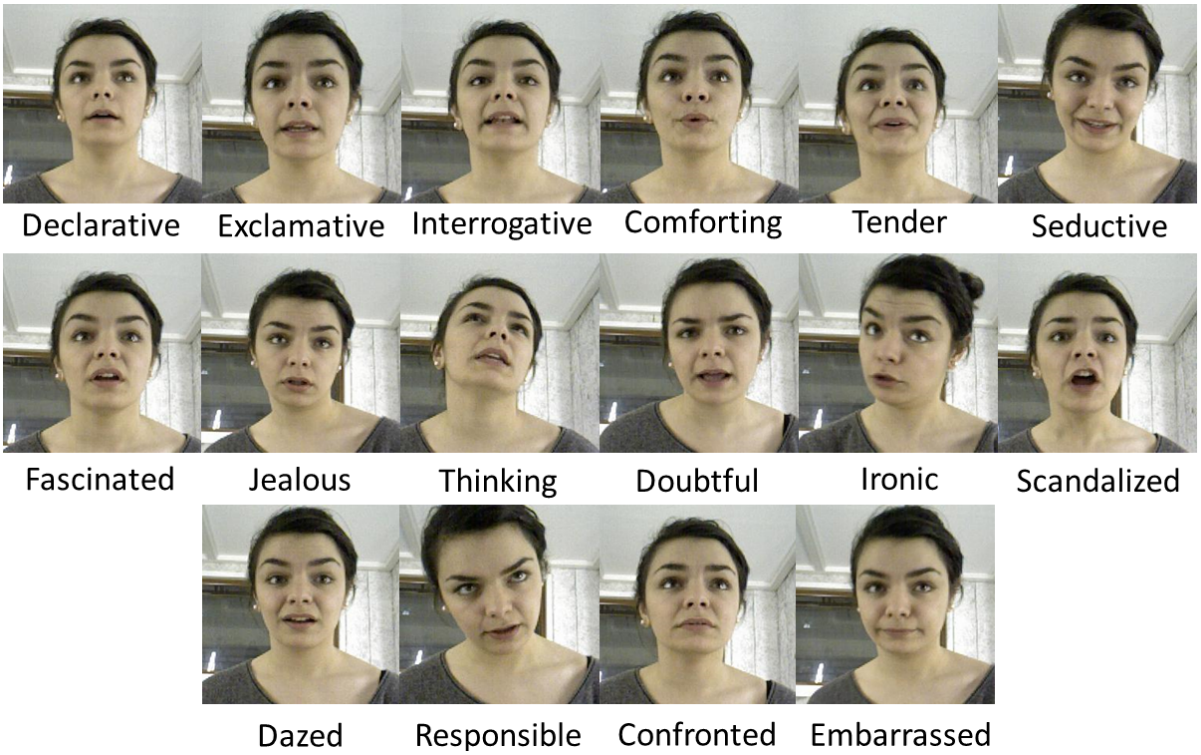

Figure 3.1: Examples of the 16 attitudes interpreted by Lucie.

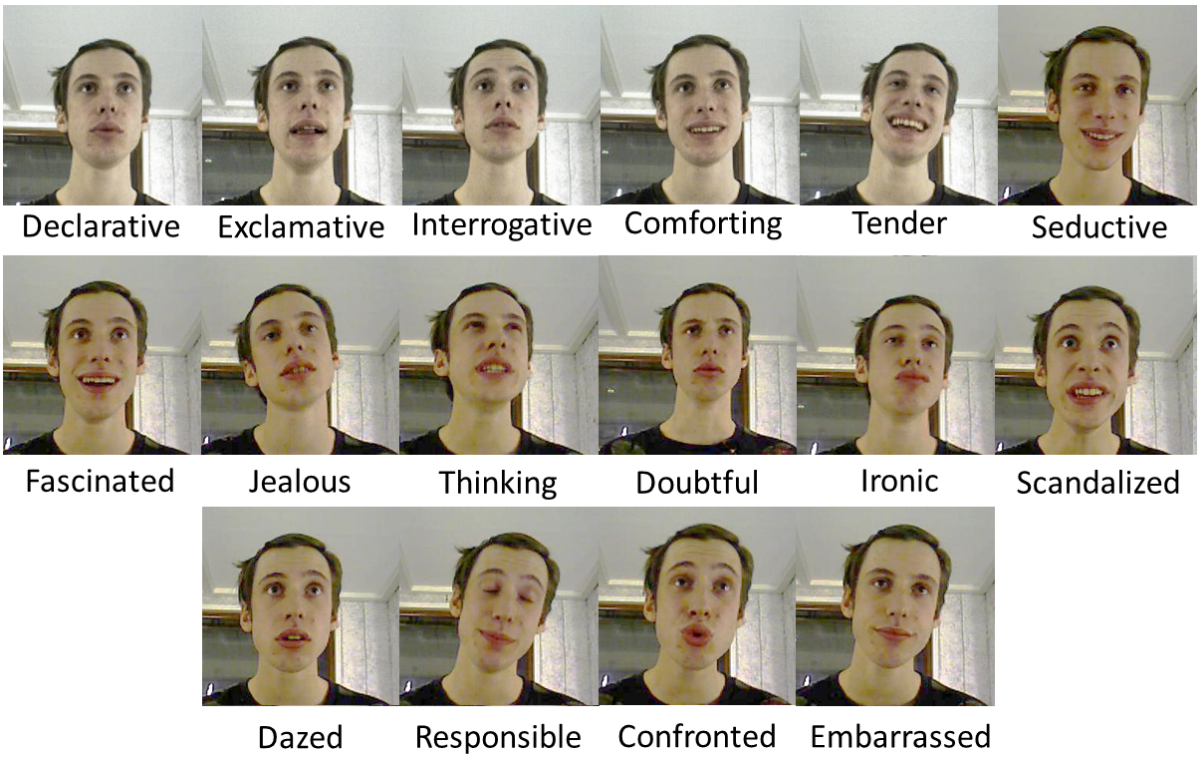

Figure 3.2: Examples of the 16 attitudes interpreted by Greg.

For the dialogs, the actors sat in front of each other across a table, where two kinect cameras were disposed between them (see Figure 3.3). All lines were learned by heart so that the performances were not affected by need to read text. This is particularly important for the recording of eye gaze direction.

Figure 3.3: Faceshift calibration and setting for dialog recording.

3.4 Segmentation

All utterances were automatically aligned with their phonetic transcription obtained by an automatic text-to-speech phonetizer [Bailly et al., 1991]. The linguistic analysis (part-of-speech tagging, syllabation), the phonetic annotation and the automatic estimation of melody were further checked and corrected by hand using the Praat speech analysis software [Boersma, 2002]. The manual verification of melodic contours represented an extensive effort due to the large amount of data recorded (a total of 3 hours of speech). Figure 3.4 presents a snapshot of Praat and corresponding video frames.

The Faceshift software exports the facial expressions data as a collection of 48 blendshapes per frame. Each blendshape corresponds to an action unit and is represented by a value bounded between 0 and 1, where 0 corresponds to the neutral expression and 1 to the highest expression intensity. Table 3.6 presents the original 48 blendshapes and the intuitive meaning of the underlying motion.

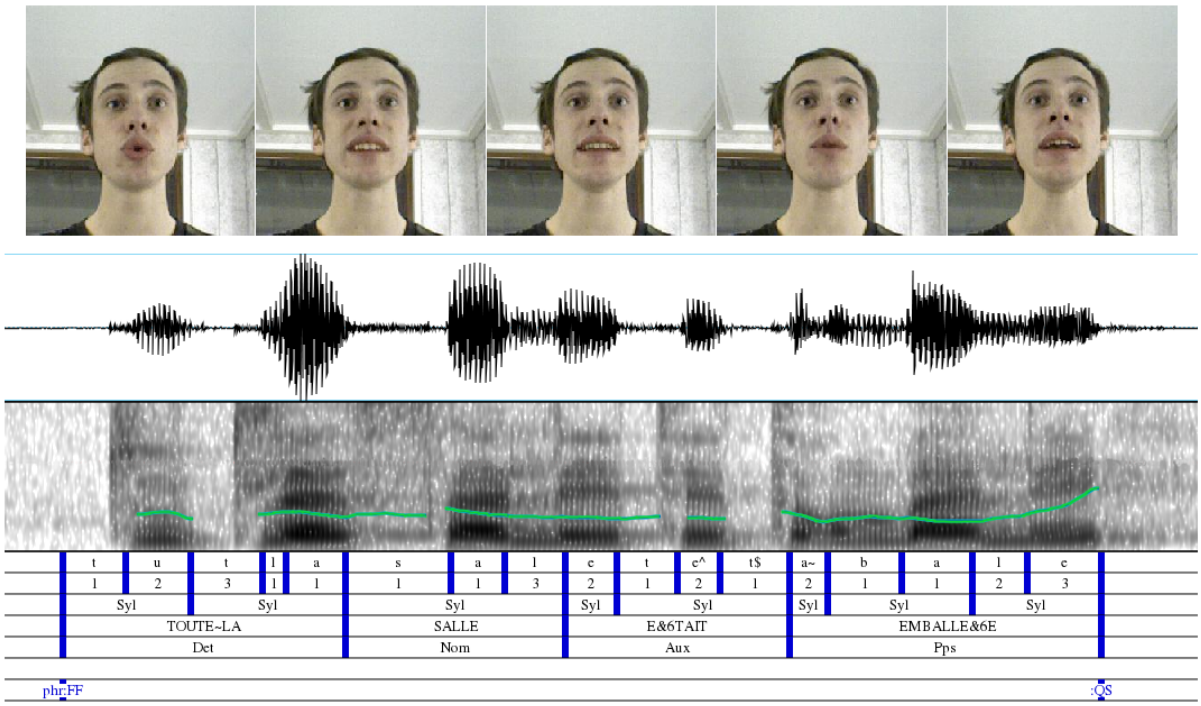

Figure 3.4: Illustration of Praat usage. Usage of Praat for the phonetic annotation of the phrase: "Toute la salle était emballée." uttered by Greg with the attitude Question. The contour highlighted over the spectogram represents the F_0 trajectory. The rising of the contour towards the end of phrase is characteristic for interrogative sentences. Every 8 video frames are selected from the performance and they illustrate the production of sounds: /t/, /a/, /e/, /a∼/ and /e/ respectively.

Table 3.6: Names of all blendshapes returned by Faceshift. The numbers associated to each blendshape are used in the following figures fo the corresponding rendered poses.

# Name	# Name	# Name	# Name
1 Eye Blink Left	13 Eye Up Left	25 Mouth Left	37 Lips Upper Up
2 Eye Blink Right	14 Eye Up Right	26 Mouth Right	38 Lips Lower Down
3 Eye Squint Left	15 Brows Down Left	27 Mouth Frown Left	39 Lips Upper Open
4 Eye Squint Right	16 Brows Down Right	28 Mouth Frown Right	40 Lips Lower Open
5 Eye Down Left	17 Brows Up Center	29 Mouth Smile Left	41 Lips Funnel
6 Eye Down Right	18 Brows Up Left	30 Mouth Smile Right	42 Lips Pucker
7 Eye In Left	19 Brows Up Right	31 Mouth Dimple Left	43 Chin Lower Raise
8 Eye In Right	20 Jaw Forward	32 Mouth Dimple Right	44 Chin Upper Raise
9 Eye Open Left	21 Jaw Left	33 Lips Stretch Left	45 Sneer
10 Eye Open Right	22 Jaw Open	34 Lips Stretch Right	46 Puff
11 Eye Out Left	23 Jaw Chew	35 Lips Upper Close	47 Cheek Squint Left
12 Eye Out Right	24 Jaw Right	36 Lips Lower Close	48 Cheek Squint Right

3.5 Animation platform

Given the nature of the original data (speech and 3D motion), a natural choice is to create a 3D animation platform to allow the visualization of a synthesized performance, where a virtual actor is represented by a blendshape model. Both the actors and the director are recorded, but we are only interested in obtaining animations of the actors.

We propose two animation platforms implemented in Blender. The first one uses the 3D model and texture generated by Faceshift, seeking for a realistic rendering style. The second uses a 3D blendshape model which was generated by artists as a cartoon style mesh and texture starting from original photos of the actors. Each cartoon style actor is created by modeling 48 meshes (one for each blendshape) along with new textures. Figures 3.8, 3.9, 3.10 and 3.11 illustrate individual blendshape values at maximum intensity for the two actors and animation styles. The numbers joining each pose correspond to the blendshape names presented in table 3.6. The following figures present snapshots of Seductive, Thinking and Scandalized poses for the two actors: video, realistic and cartoon style animated frames. In the realistic animation we use a floating head, as generated by Faceshift, while for the cartoon style animation a torso is added. Translations are applied to the torso and rotations only to the head.

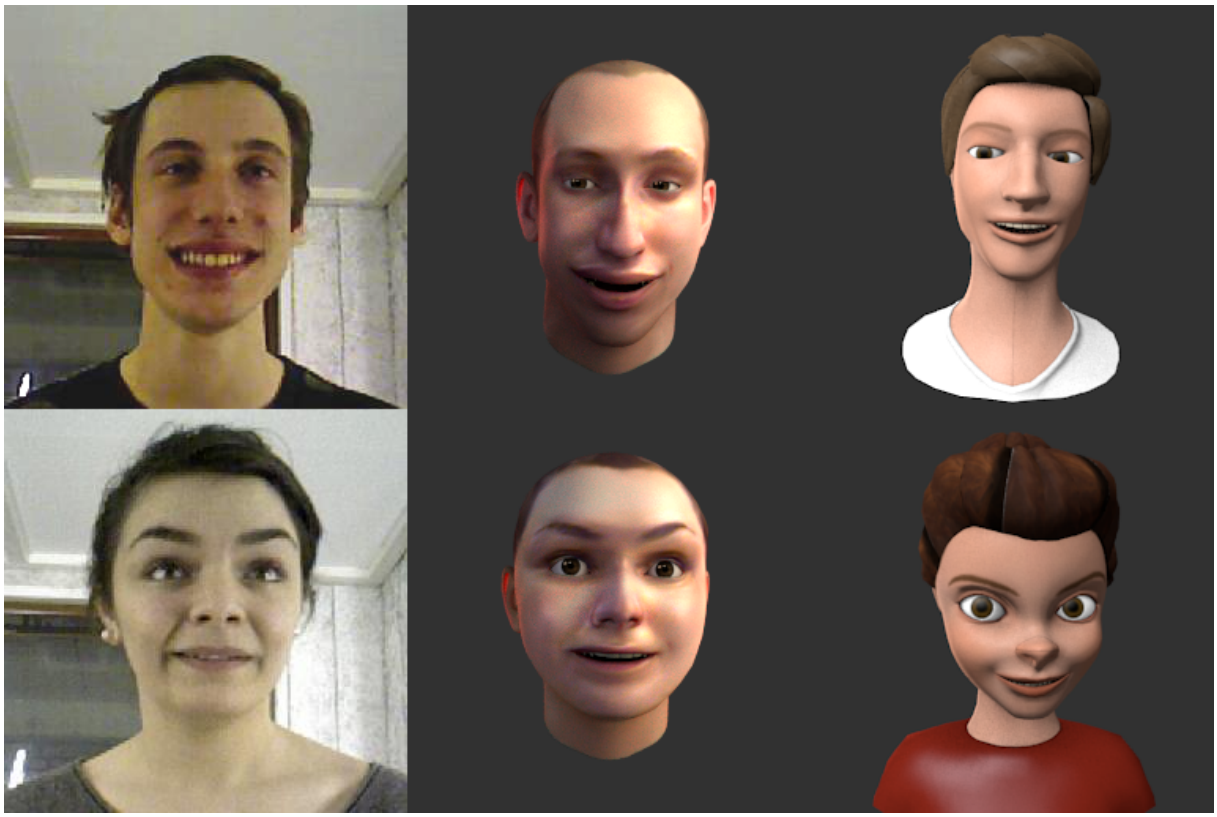

Figure 3.5: Video and animated frames for the attitude Seductive.

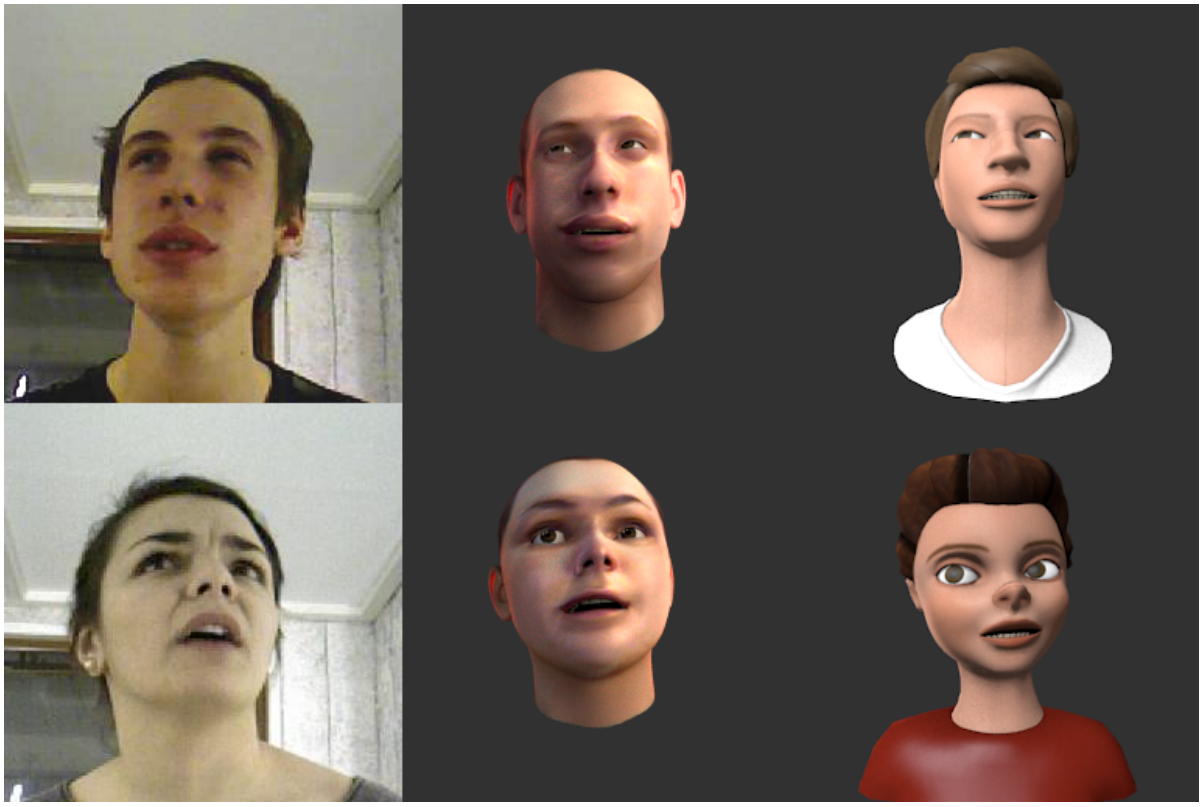

Figure 3.6: Video and animated frames for the attitude Thinking.

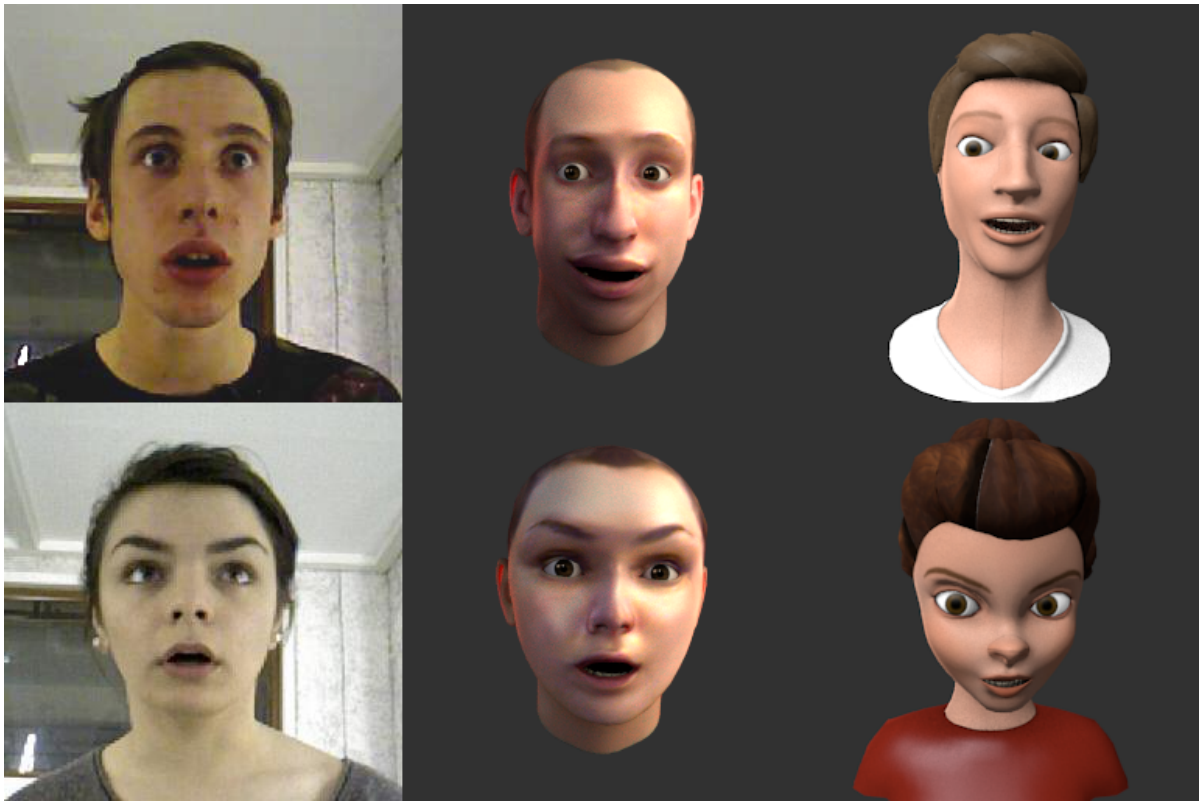

Figure 3.7: Video and animated frames for the attitude Scandalized.

Figure 3.8: Realistic rendering depicting all blendshapes for Greg.

Figure 3.9: Realistic rendering depicting all blendshapes for Lucie.

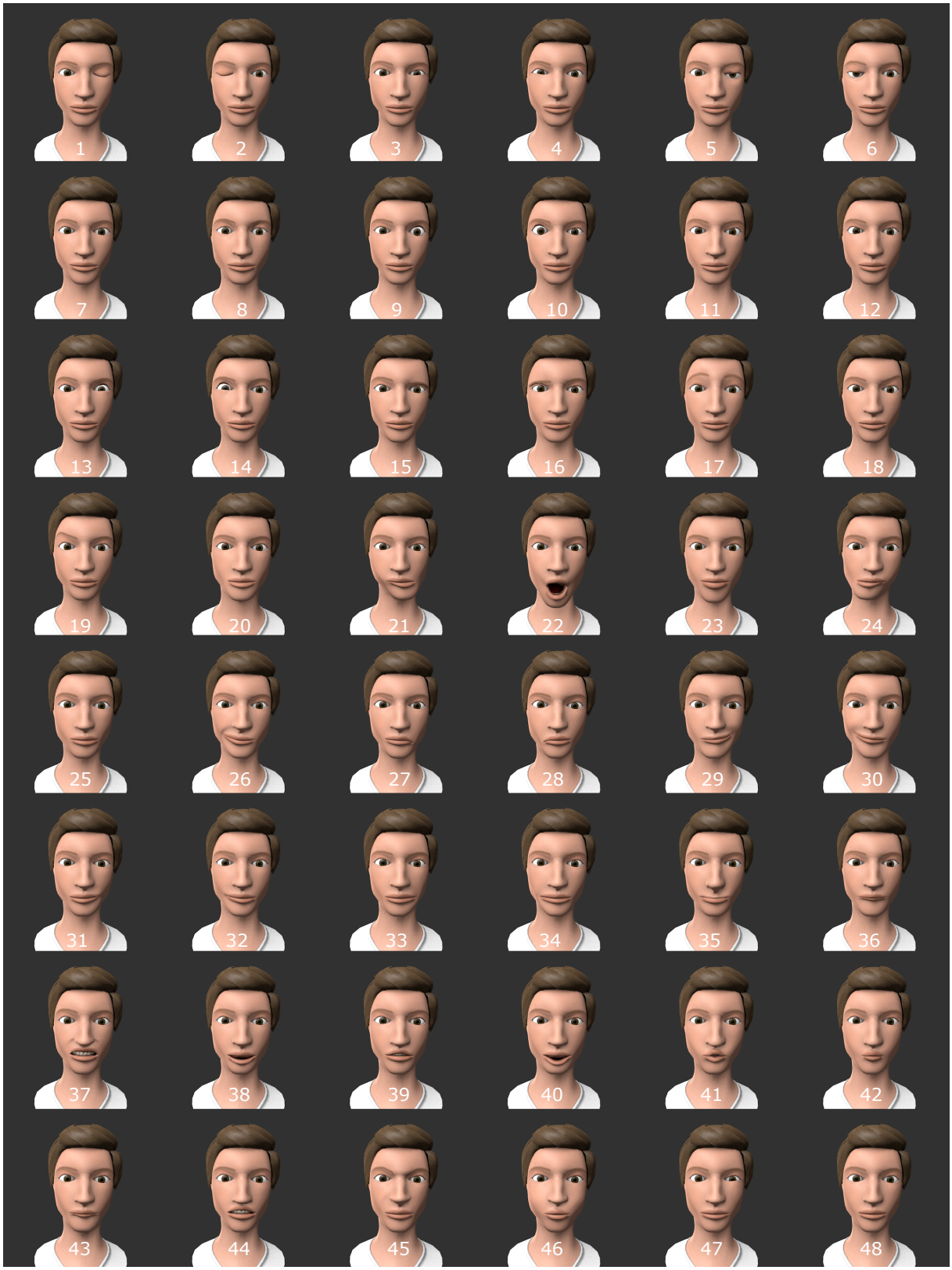

Figure 3.10: Cartoon style rendering depicting all blendshapes for Greg.

Figure 3.11: Cartoon style rendering depicting all blendshapes for Lucie.

3.6 Auto-evaluation

The two actors and director performed an auto-evaluation test in which they were asked to label samples from their own performances, a few days after having been recorded. They are presented with random sets of 32 performances for each modality (audio, video and audio-video), such that all attitudes are represented by 2 performances in each test. The performances can be played several times. No explanation is given for the labels. Table 3.7 shows the accuracy rates for all performers on all types of material used:

Table 3.7: Recognition rates obtained for all modalities and performers for the auto-evaluation test.

Performer	Gender	Audio	Video	Audio-Video
Georges	male	43.75%	68.75%	68.75%
Greg	male	65.62%	71.87%	75.00%
Lucie	female	**78.12%**	**81.25%**	**78.12%**

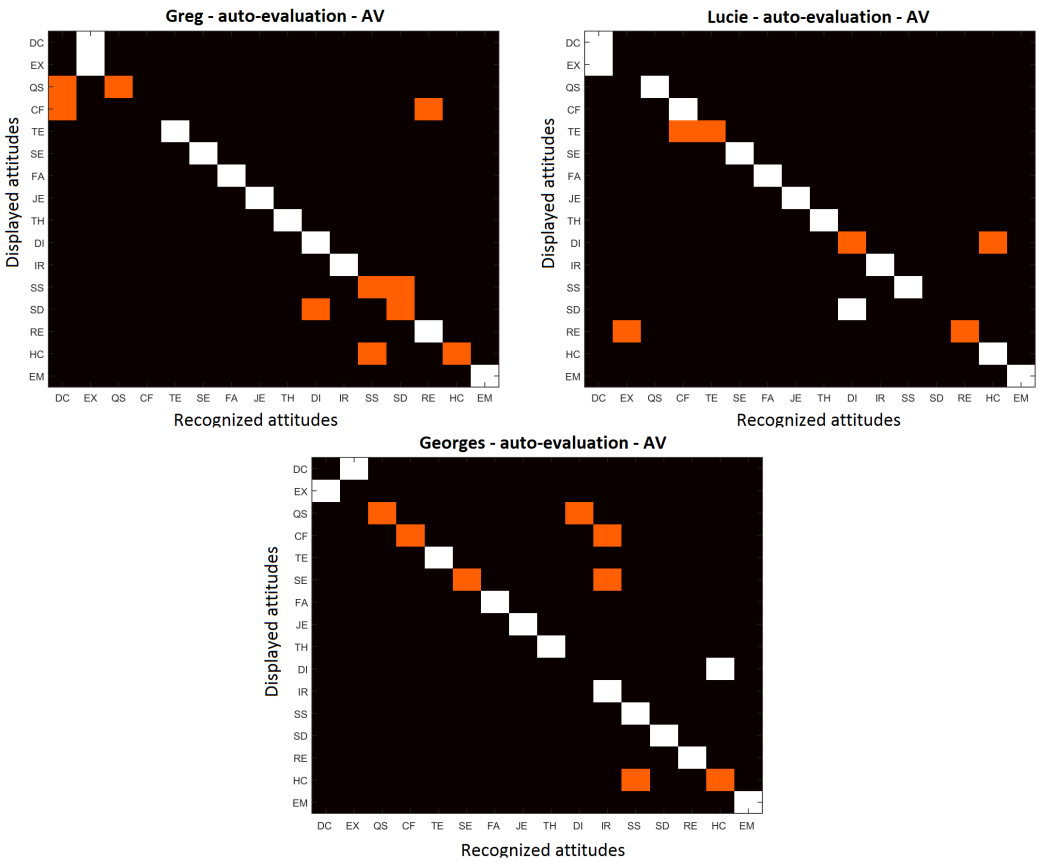

Figure 3.12: Confusion matrices for the AV auto-evaluation. Heatmaps representation of the confusion matrices obtained for the auto-evaluation test using audio-video material. The results are shown for all speakers and for all attitudes.

The audio channel seems to encode slightly less discriminant information than the video channel. Overall, the recognition rate is high above the chance level (6.25%) and in the case of the actors, it is generally higher than 60%. Lucie presents the highest consistency.

Figure 3.12 illustrates the confusion matrices for the audio-video modality. Since only two performances are shown for each attitude, the recognition rate per attitude can be 0, 50 or 100%. The attitudes with 100% recognition rate for all speakers are: Fascinated, Jealous, Thinking, Ironic, Embarrassed. The attitudes which are not recognized at all by at least one speaker are: Declarative, Exclamative, Comforting, Doubtful, Dazed. Generally, there is a high confusion between Declarative and Exclamative, between Doubtful and Dazed, and between Scandalized and Confronted.

We retain these results for the improvement of perceptual evaluation tests. We consider that the average participant should not spend more than 20 minutes for an attitude recognition test (such that the results remain unaffected by fatigue etc). Given the multimodal nature of the data, the number of speakers and attitudes, in the next evaluation tests we will mostly use performances recorded by the actor with the highest general recognition rates (Lucie) and the best recognized attitudes.

3.7 Limitations

Dataset limitations may refer to data quality and quantity. In terms of quality, all recordings were carried under director's supervision and the video data was validated initially through the auto-evaluation tests.

Faceshift presents high-level motion capture capabilities and has successfully been used for facial expression analysis [Davis et al., 2015]. However, we do not hold a reliable method to assess the results delivered i.e. to compare with ground-truth 3D data. Moreover, we expect a decrease in expressiveness perception with the usage of an animated platform. For this reason, we propose two types of meshes and textures for the actors: realistic and cartoon style. The transposition of original performances through Faceshift and the animated platforms is evaluated in section 4.7.

In terms of quantity, our dataset is limited both in terms of number of attitudes (16) and number examples per attitude (35). This is not a problem for our purpose, which is to extract "signatures" of individual attitudes at a suprasegmental level. The choice for the number of utterances per attitude is also related to the capabilities of one actor to record "pure" attitudes over an extended period of time.

On the other hand, our corpus does not make it possible to train a full text-to-speech synthesizer for the three speakers. The dialogue part of the corpus is provided for testing purposes only.

3.8 Summary

This chapter described the requirements, design and implementation of a dataset of dramatic attitudes. In Section 3.1 we presented a brief overview of available expressive datasets.

Cognitive studies on vocal and facial expression and previous work on the generation of auditory prosody indicate that attitudes are encoded via prosodic signatures which are dependent on the utterance size and are attitude-specific. As the study of audiovisual expression of attitudes requires a specific amount and type of data, we designed and recorded a dataset of 13 dramatic attitudes and 3 modal attitudes. Section 3.2 presented the set of selected attitudes and Section 3.3 described the dataset design and the process of recording the actors.

In Section 3.4 we described the process of data segmentation. Next, Section 3.5 presented how the animation platform is used to synthesize audiovisual performances using a realistic style and a cartoon style. Section 3.6 presented the result of an auto-evaluation test carried by the two actors and the theater director after recording the dataset.

Chapter 4

Analysis of audiovisual prosody

This chapter presents an analysis of the data described in the previous chapter, with the goal of better understanding how the different attitudes are encoded in the audio-visual signals. We therefore propose inter-class distance measures. Both the spoken performance and the pre-phonatory and post-phonatory movements for each actor are analyzed. We compare the results of the objective measures to perceptual test results on the original video and animated performances. The following sections describe: feature characterization, stylization, discriminant analysis and objective together with perceptual test results.

4.1 Feature characterization

We first describe the pre-processing step and then detail the frame and syllable characteristics.

4.1.1 Data pre-processing

Faceshift automatically aligns audio and visual data. Visual data is resampled at 30 Hz and pre-processed using built-in cubic smoothing presets. Resampling is necessary in order to solve the problem of audio-video asynchrony which may sometimes appear due to camera framedrops. Once exported, data is prepared for the analysis step by carrying a few pre-processing techniques:

- head translation is normalized such that all performances start at the same location
- head movements are restricted (eg. location was restricted in case the actor made a sudden move and exceeded the recording area)

- eye-area blendshapes (blinks and squints) that are partly occluded simply copy their symmetric counterparts

4.1.2 Frame characteristics

Spectrum

The audio features are extracted using the STRAIGHT vocoder [Kawahara, 2006] which returns the voice spectra, aperiodicities and the fundamental frequency f_0. Specifically, we work with 25 mel-cepstral coefficients [Bogert et al.,] and 5 Bark-scale aperiodicities.

Voice pitch

As mentioned in 3.3, all utterances are automatically aligned with their phonetic transcription and the automatic estimation of melody is further checked and corrected by hand using the Praat [Boersma, 2002]. Therefore, we obtained reliable f_0 contours which are further normalized and then converted to tones according to the equations:

$$f_0 ref = \begin{cases} 210Hz, & \text{if female} \\ 110Hz, & \text{if male} \end{cases} \tag{4.1.1}$$

$$\tag{4.1.2}$$

$$f_0[tone] = \frac{240}{\log 2} \log \frac{f_0[Hz]}{f_0\ ref} \tag{4.1.3}$$

where $f_0\ ref$ represents the speaker's register. The resulting f_0 contours are comparable across speakers.

Motion

Head, gaze motion and facial expressions are processed using the principal component analysis (PCA). The visual features are split into several groups to which PCA is applied. The analyzed data is composed of the audiovisual speech segments and the visual information gathered from a fixed amount of segments before and after each spoken sentence. The fixed interval is equivalent to 30 video frames (1 second as the video framerate is 30 HZ). PCA is applied on all data gathered from all speakers.

We keep only the components which explain more than 90% of the information contained thus leading to: 5 components for head movements, 2 components for gaze, 3 components for eyebrow movements, 6 components for eye-area movements and 8 components for the lower part of the face (see figure 4.1). The whole face is parameterized with 24 PCA components. The high percentage is particularly important for the blendshape

parameters as we require high accuracy for the reconstruction of facial expressions. Figure 4.2 illustrates a set of poses corresponding to the resulted PCA components. Table 4.1 presents the original 48 blendshapes where the background color defines the facial regions of interest for applying PCA. Table 4.2 presents the variance explanation for the face areas considered:

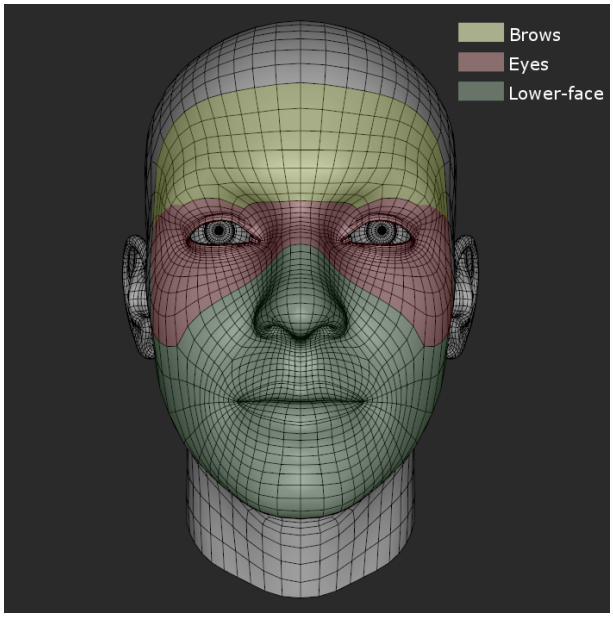

Figure 4.1: PCA areas highlighted on the Faceshift template mesh. Green covers the lower-face area, red covers the eye-area and yellow covers the brows-area.

Table 4.1: Color-coded blendshapes names returned by Faceshift. The color associated to each blendshape corresponds to the three face regions defined in figure 4.1.

# Name	# Name	# Name	# Name
1 Eye Blink Left	13 Eye Up Left	25 Mouth Left	37 Lips Upper Up
2 Eye Blink Right	14 Eye Up Right	26 Mouth Right	38 Lips Lower Down
3 Eye Squint Left	15 Brows Down Left	27 Mouth Frown Left	39 Lips Upper Open
4 Eye Squint Right	16 Brows Down Right	28 Mouth Frown Right	40 Lips Lower Open
5 Eye Down Left	17 Brows Up Center	29 Mouth Smile Left	41 Lips Funnel
6 Eye Down Right	18 Brows Up Left	30 Mouth Smile Right	42 Lips Pucker
7 Eye In Left	19 Brows Up Right	31 Mouth Dimple Left	43 Chin Lower Raise
8 Eye In Right	20 Jaw Forward	32 Mouth Dimple Right	44 Chin Upper Raise
9 Eye Open Left	21 Jaw Left	33 Lips Stretch Left	45 Sneer
10 Eye Open Right	22 Jaw Open	34 Lips Stretch Right	46 Puff
11 Eye Out Left	23 Jaw Chew	35 Lips Upper Close	47 Cheek Squint Left
12 Eye Out Right	24 Jaw Right	36 Lips Lower Close	48 Cheek Squint Right

Table 4.2: Number of motion parameters and variance explained over the groups considered: head motion, gaze direction, eyebrows, eye area and lower-face area. The columns represent: the original number of parameters returned from Faceshift, the number of PCA components considered and the variance explained by each group.

	Head	Gaze	Eyebrows	Eye	Lower
No orig	6	4	5	14	29
No PCA	5	2	3	6	8
Variance	97.11%	99.95%	93.31%	95.3%	92.88%

Figure 4.2: Poses corresponding to PCA components. From left to right: Eyebrows 1st component, Eyebrows 2nd component, Eyes 1st component, Eyes 2nd component, Lower-face 1st component, Lower-face 2nd component. Semantically, these poses correspond more or less to: eyebrows up, eyebrows down, eyes closed, eyes open, mouth open and lips spreading, respectively.

4.1.3 Syllable characteristics

Rhythm

For rhythm we use a duration model [Campbell, 1992] [Bailly and Holm, 2005] where syllable lengthening/shortening is characterized with a unique z-score model applied to log-durations of all constitutive segments of the syllable. We consider the syllable as rhythmical unit (RU) and compute a coefficient of lenghtening/shortening C equal to the deviation of the RU duration relative to an expected duration Δ':

$$\Delta' = (1 - r) \cdot \sum_i \bar{d_{p_i}} + r \cdot D \tag{4.1.4}$$

where i is the phoneme index within the RU, $\bar{d_{p_i}}$ is the average duration of phoneme i, D is the average duration for a RU and r is a weighting factor. For a RU with a measured duration Δ, the shortening/lenghtening coefficient is:

$$C = \frac{\Delta - \Delta'}{\Delta'} \tag{4.1.5}$$

We note C as the rhythm coefficient which is computed for every syllable of all sentences in the corpus.

4.2 Segmental vs suprasegmental

An important step in the analysis and modeling of attitudes is the characterization of features relative to structural properties. Segmental features concern individual sounds or phonemes, working at the segmental level since each phoneme is usually assumed to be one segment of speech. Looking at larger chunks of speech that span a number of segments, such as whole words or phrases, implies dealing with features on the suprasegmental (prosodic) level.

While some models such as the SFC [Bailly and Holm, 2005] propose to combine segmental and suprasegmental information onto the signals via overlap and add techniques, most prosodic models promote a strict separation between segmental and suprasegmenta features. We consider the separate contribution of segmental and suprasegmental features via specific audiovisual streams of parameters.

Among the features we extract and analyze, we consider the ones associated to the segmental level: voice spectra and lower-face movements (such as jaw open, funnel etc). Power (intensity) is implicitly processed with the mel-cepstra coefficients and will be treated as a segmental feature. On the other side, we consider voice pitch, rhythm, head motion, gaze and upper-face facial expressions (such as: eyebrows up, squint etc) as suprasegmental features and will be characterized at the level of the syllable (see figure 5.19). Table 4.3 presents a summary of this characterization.

Due to the nature of these features, the following sections refer to the prosodic dimension.

Table 4.3: Feature categorization in terms of segmental structure.

	Audio	Visual
Segmental	mel-cepstra aperiodicities	lower-face blendshapes
Suprasegmental	f_0 rhythm	upper-face blendshapes head movements eye gaze rhythm

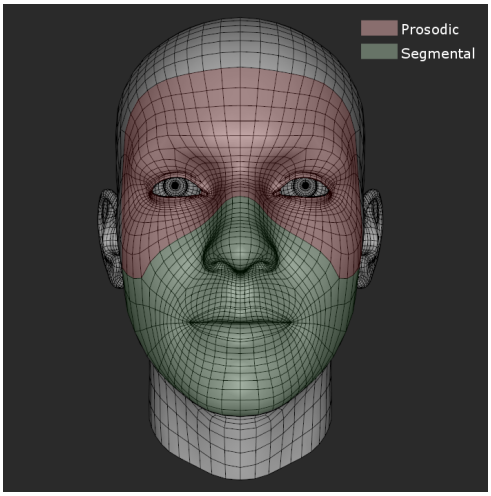

Figure 4.3: Representation of prosodic (red area) and segmental (green area) feature regions on the Faceshift template mesh.

4.3 Audiovisual prosody stylization

By *stylization* we mean the extraction of several values at specific locations from the prosodic continuum with the main purpose of simplifying the analysis process while maintaining the original contour characteristics. A stylized contour is composed of stylization keypoints which represent the contour locations from which values are extracted. Stylization solves the problem of trajectory comparison when different phonetic content is carrying the prosodic contour.

In speech, stylization has been mainly used for pitch contours: Black and Hunt [Black and Hunt, 1996] notably used only 3 values per voiced part of the syllable and interpolation between these; Tournemire [De Tournemire, 1994] tested different sampling instants per syllable and stated that the evolution of f_0 inside consonants does not require further modeling and that interpolation between three values collected within vowels is sufficient.

We propose the following stylization methods for the audiovisual prosodic features: f_0, rhythm and visual contours (head and gaze motion, upper-face expressions) by extracting 3 values per syllable.

- f_0: the log-pitch contour is stylized by extracting 3 values: at at 20%, 50% and 80% of the vocalic nucleus of each syllable. The sampling is performed after a polynomial interpolation of the voiced part of the syllable.
- *Motion*: all PCA components obtained are also stylized by extracting contour values at 20%, 50% and 80% of the length of each syllable.

- *Rhythm*: the rhythm is represented by one parameter per syllable, the lenghtening/shortening coefficient. Therefore, the stylized rhythm contour for one sentence is identical to the initial sequence of coefficients.

4.3.1 Virtual syllable

Our study focuses also on the motion occurring in pre and postphonatory silences. As previously observed, preparatory movements are discriminant to a certain degree to specific emotion categories. We therefore introduce the *virtual syllables* which are silences with a duration of 250 ms (approximately the average syllable duration) preceding and following each utterance. We add these "virtual syllables" to account for the pre- and post-phonatory movements for all visual components. Therefore, stylization of motion is also done for the virtual syllables, by extracting motion contour values at 20%, 50% and 80% of the length of each virtual syllable.

4.4 Discrimination between attitudes

We verify the quality of the data gathered from our corpus by analyzing the stylized motion contours extracted from virtual syllables and from the stylized f_0 and rhythm contours extracted from the first and last syllables of speech.

Discriminant analysis is performed using Fisher classification over the motion performed within virtual syllables, first and last syllables of the utterance and also on the auditory prosodic features over the first and last syllables. Fisher classification was implemented with 10-folds cross validation. We perform position-specific classification when we take into account only the stylized contour at a certain syllable or cumulative classification when we gradually classify accumulated data from the first up to the currently analyzed syllable. Only data gathered from Lucie is considered.

We first perform the classification using acoustic prosody only. Results are illustrated in figures 4.4 and 4.5.

On average, the recognition rate for auditory prosody (including melody and rhythm) for first and last syllable is 20.12% and 40.92% respectively, where chance level is 6.25%. Best results are obtained for Scandalized, Question, Thinking and Confronted, with Scandalized yielding 66.67% and 75% for the first and last syllable respectively. These attitudes reportedly show high pitch contours, especially for the last syllable. The least recognized attitudes are Fascinated, Comforting and Embarrassed, with some obtaining 0% recognition for the first syllable. These results show that the last syllable contains more discriminant pitch values than the first syllable, which is consistent with observations

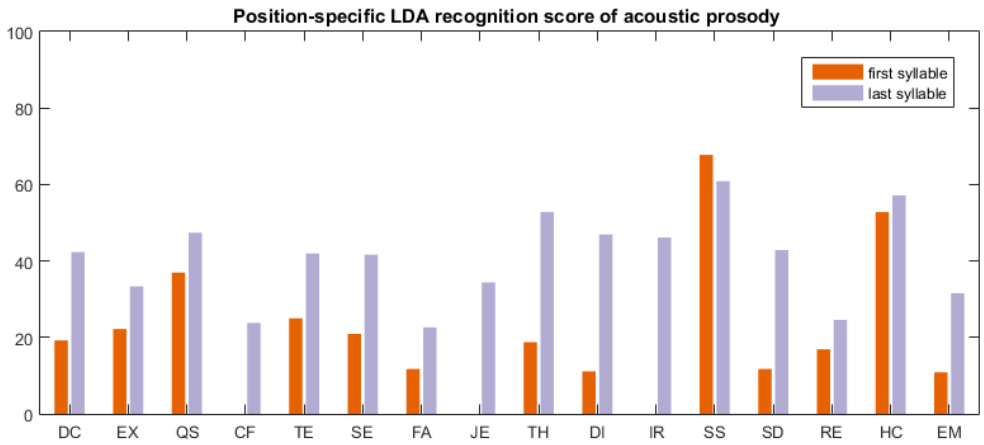

Figure 4.4: Position-specific auditory prosody recognition rates for first and last syllables per attitude.

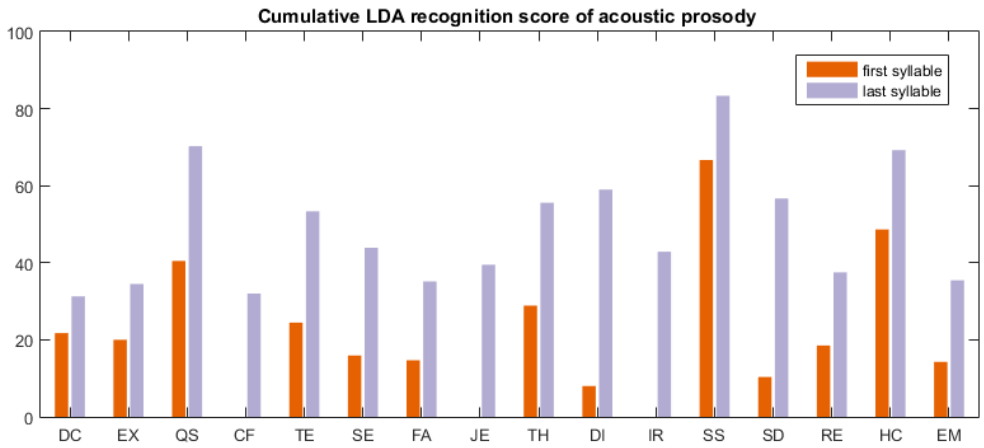

Figure 4.5: Cumulative auditory prosody recognition rates for first and last syllables per attitude.

extracted from studies on melodic signatures [Morlec et al., 1995] [Haan et al., 1997] [Bailly and Holm, 2005].

The cumulative recognition scores are on average 20.70 % and 48.63 %, showing slightly higher rates for the last syllable. Therefore, considering prosodic information from both syllables increases the discrimination rate between attitudes.

Next, we perform Fisher classification for the visual prosodic features only, extracted from the first and last syllables and from the virtual syllables also. Results are illustrated in figures 4.6 and 4.7.

Visual prosody shows high recognition rates both for virtual syllables and marginal syllables within utterances. The average obtained is 75.16 %, 74.76%, 82.14%, 85.54% for pre-phonatory, first, last and post-phonatory syllables respectively. Best recognition

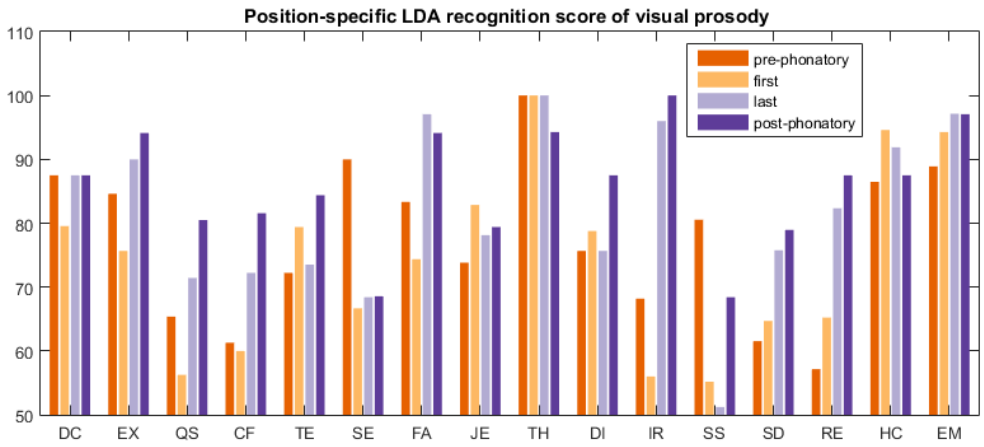

Figure 4.6: Position-specififc visual prosody recognition rates for virtual syllables, first and last syllables per attitude.

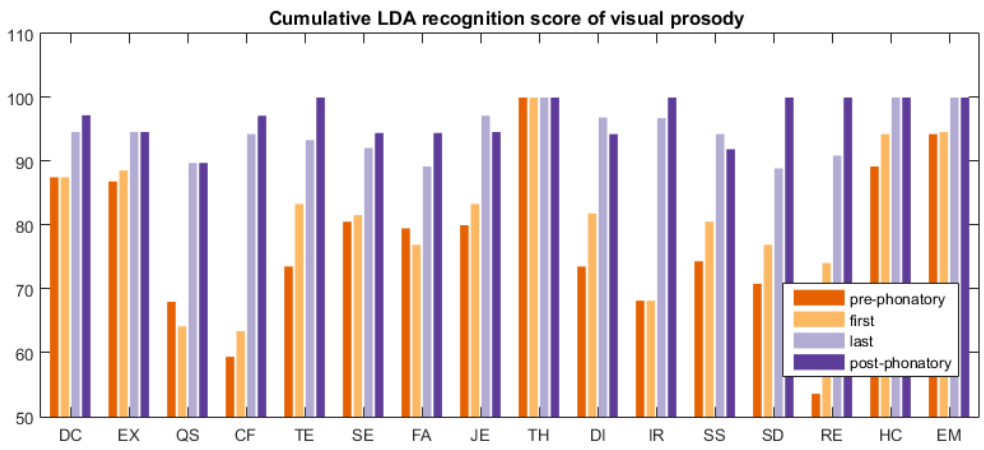

Figure 4.7: Cumulative visual prosody recognition rates for virtual syllables, first and last syllables per attitude.

is obtained for Thinking and Embarrassed and lowest for Comforting and Scandalized. Generally, the post-phonatory syllable (with a minimum of 61.54% for Scandalized) obtained better rates than the pre-phonatory (with a minimum of 60.71% for Responsible), while being above the chance level.

The average cumulative classification rates in the temporal order of the syllables are 76.91%, 80.89%, 95.59% and 95%. These results show that the information contained in the virtual syllables is discriminant for the expression of specific attitude patterns. Therefore, the visual information contained within these syllables will be used in the modeling and synthesis of the visual prosodic signatures.

4.5 Attitude-specific signatures

As pointed out by Morlec [Morlec et al., 2001] and Holm [Bailly and Holm, 2005], coherent behaviors for prosodic contours are observed for various-length sentences for specific attitudes. This section illustrates audiovisual prosodic stylized contours extracted from our corpus, along with observations for each prosodic feature.

The following set of figures represent audiovisual prosodic stylized contours extracted from original performances. The dotted lines represent a common reference value for all actors, equating the average value of the represented component. Each set of contours for one parameter are displayed for all sentences containing 1, 3, 5, 7, 9 and 11 syllables. Contours associated with an equal number of syllables are overlapping. Notice that different syllable lengths collect different numbers of sentences. For easier visualization, the overlapped contours are colored differently.

Below we describe several observations on the behavior of the stylized contours:

1. Sentence-scope signatures: there is coherence both in contours belonging to one attitude for a given number of syllables and across all sentence-lengths (see figure 4.8).

2. Attitude-specific signatures: f_0 tends to follow the same patterns for one attitude across speakers. These are similar to patterns signaled by work mentioned in the previous chapter (see figure 2.5, 4.9).

3. Speaker individuality: for some attitudes there is coherence across sentences for each speaker, but not across speakers (see figures 4.10).

4. Outliers: there are exemplars which break the seeming patterns above-mentioned. These can be considered as outliers or prototypical alternative variations (see figure 4.11).

5. Left-right rotations: the first head component, which corresponds roughly to left-right head rotation, presents for same-length exemplars which are symmetric relative to the reference contours. This indicates that this rotation can be implemented in opposite directions across same-scope sentences (see figure 4.12)

These observations represent the base for the implementation of an audiovisual prosodic model.

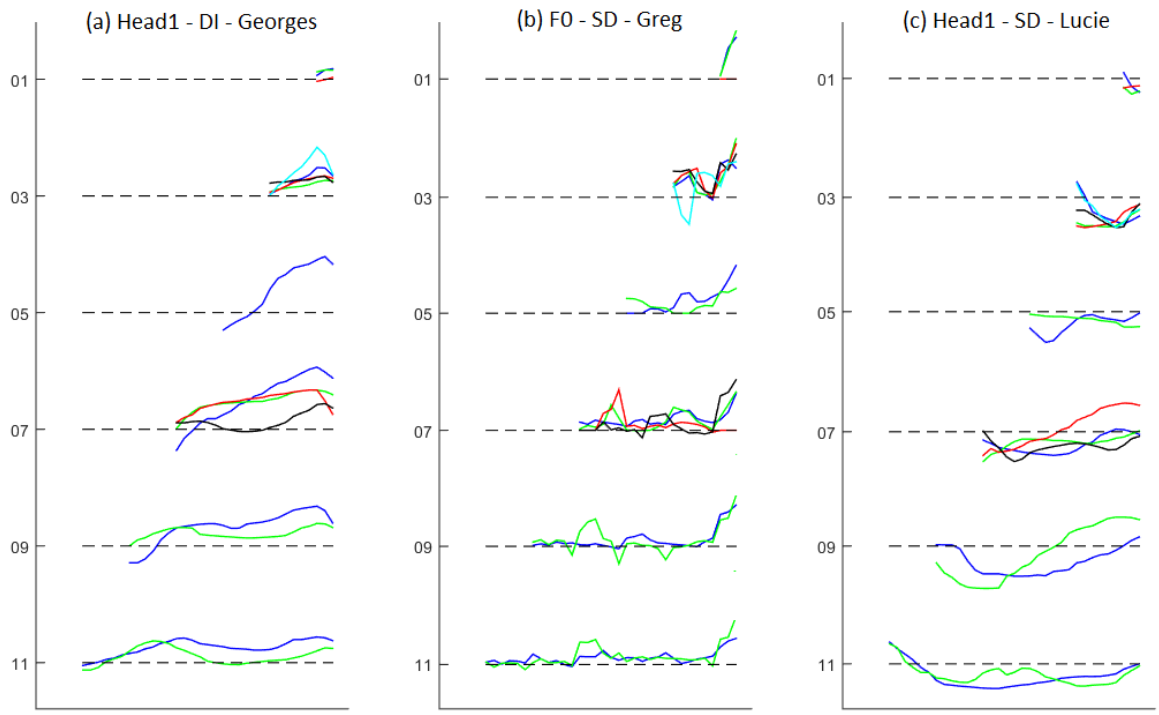

Figure 4.8: Examples of stylized contours. From left to right, the figures represent: (a) component head 1 for attitude Doubtful for Georges, (b) f_0 for attitude Dazed for Greg (c) component head 1 for attitude Dazed for Lucie.

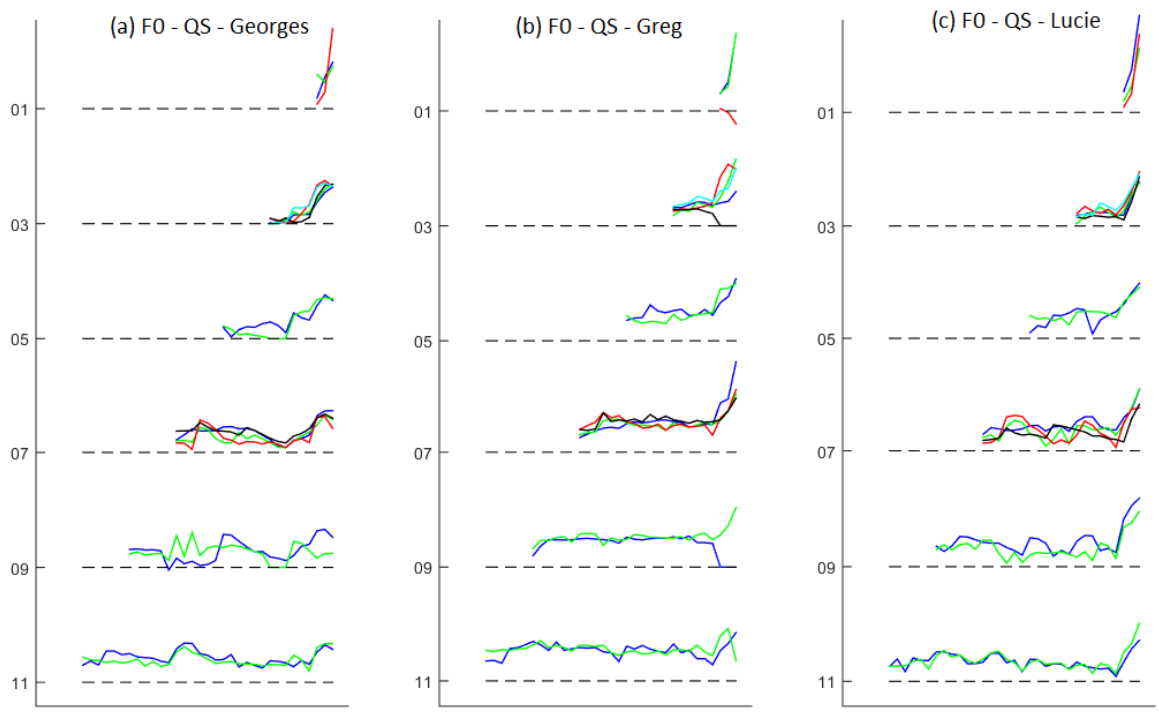

Figure 4.9: Attitude-specific signatures. f_0 contours for attitude Question for Georges, Greg and Lucie, respectively.

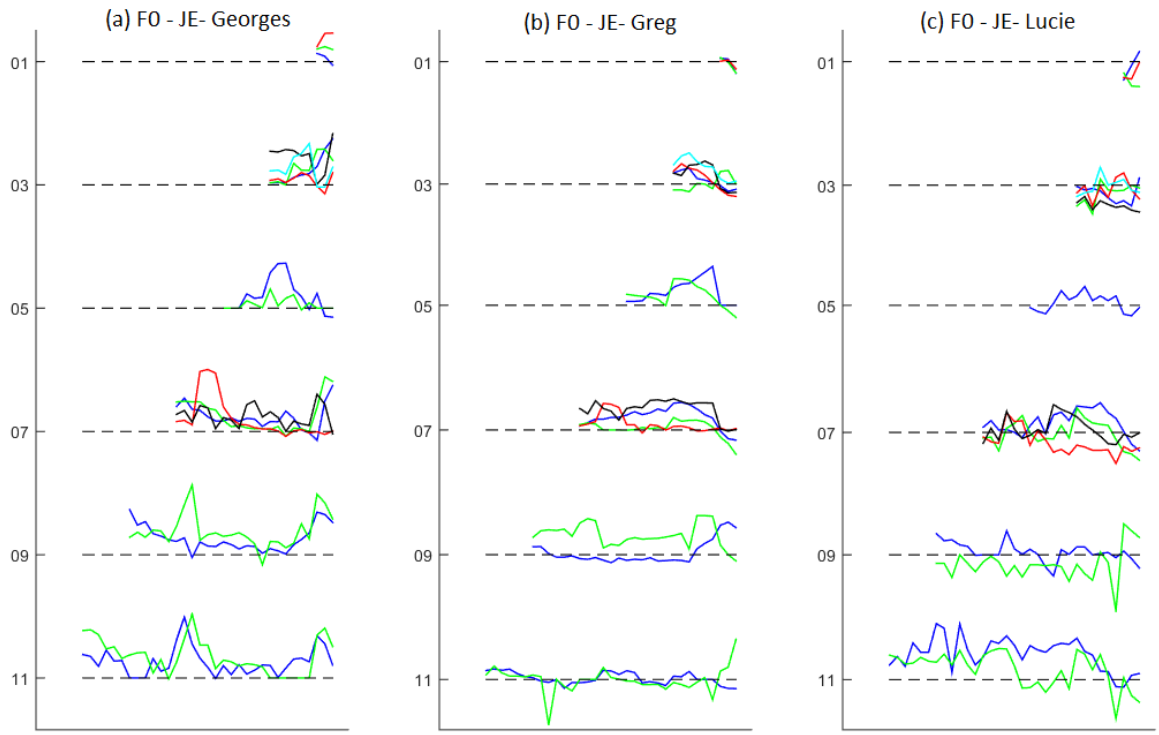

Figure 4.10: Speaker individuality. f_0 contours for attitude Jealous for Georges, Greg and Lucie, respectively.

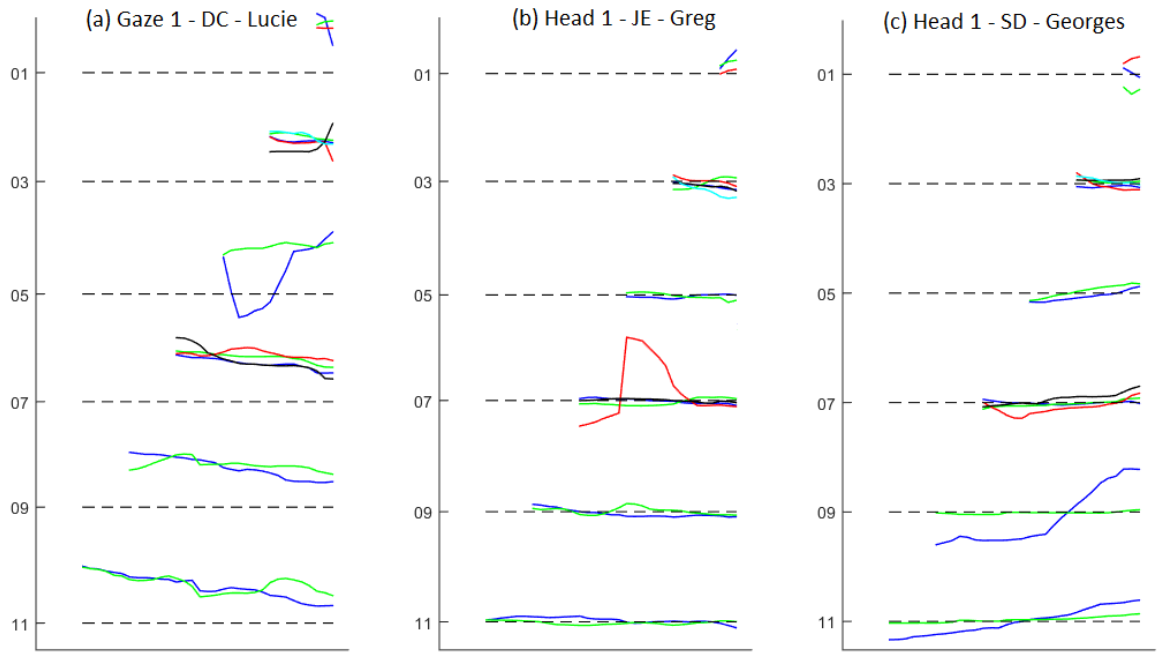

Figure 4.11: Examples of outliers. From left to right: exemplars for Lucie, first gaze component for Declarative, Greg first head component for Jealous and Georges, first head component for Dazed. Outliers are visible for a sentence of 5, 7 and 9 syllables, respectively.

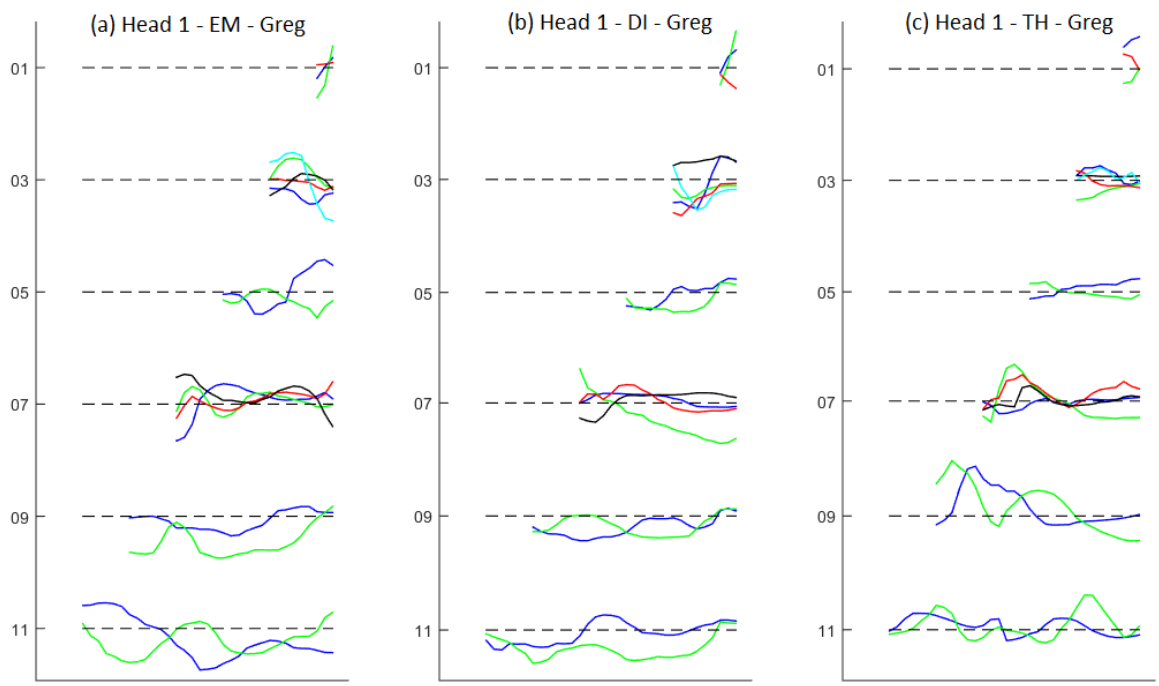

Figure 4.12: Examples for left-right rotations. From left to right: exemplar for Greg, first head component for Embarassed, Doubtful and Thinking, respectively. Symmetric contours are visible for sentences of 5 and 9 syllables.

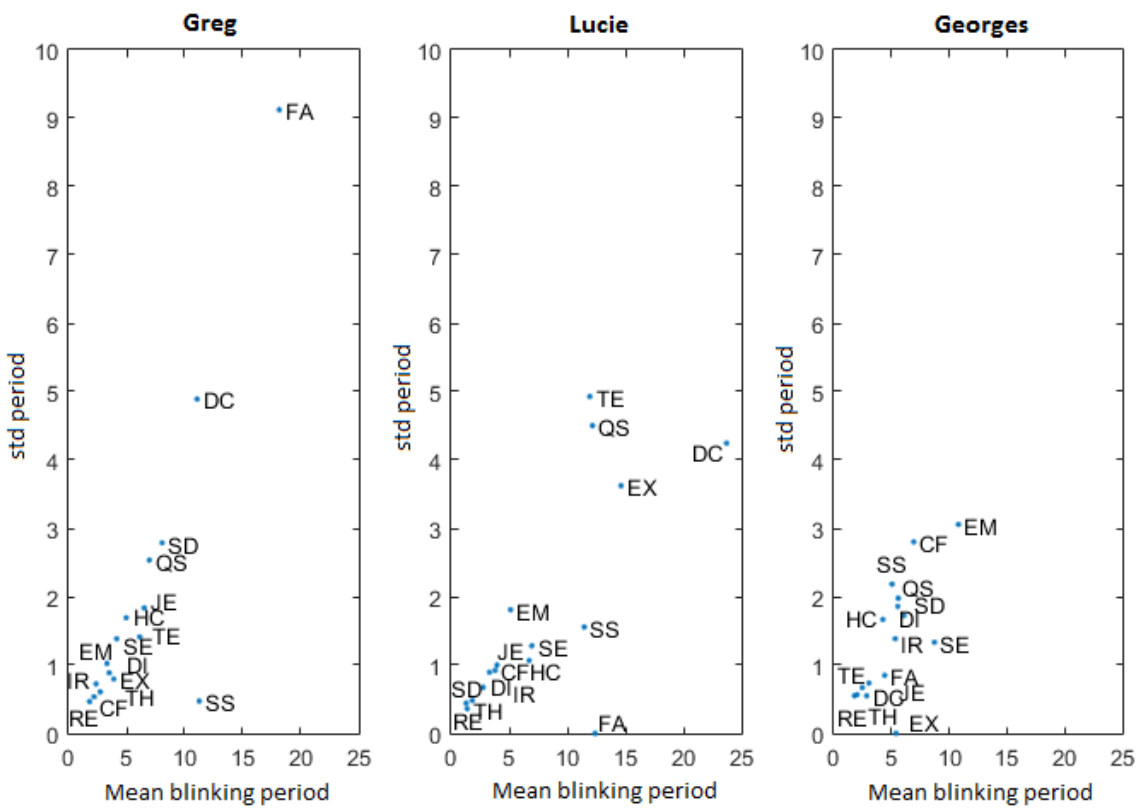

Figure 4.13: Blink periods (seconds) per attitudes and speakers.

4.5.1 Blinks

Eye blinks can be generated using various methods (see section 2.3.1). For our dataset, the first eyes PCA component corresponds to blinking. We extract blinks by thresholding the first eyes PCA component and we compute the blinking rate (defined as the time elapsed between consecutive blinks). Figure 4.13 illustrates the mean and standard deviation values for blink period per attitude.

Blinks show an irregular behavior relative to the position within a sentence. We observe that blinking rate depends on the speaker and attitude. For example, attitudes such as Responsible, Comforting and Tender show similar rates across speakers, as opposed to Declarative, Fascinated and Exclamative: while Greg performs few blinks, Georges blinks very frequently.

4.6 Objective measures

The previous section illustrated examples of attitudes which display different behavior expressed through more or less consistent dynamic shapes. We encounter the problem of discriminating attitude- or actor-specific data. Therefore, we propose an objective measure for inter-class distances for the original recorded data.

We define the "objective recognition rate" of a prosodic feature as a inter-class distance measure. The algorithm for computing the proposed measure is based on the k-nearest neighbor (k-NN). The following section describe how the algorithm is applied to compute objective recognition rates of prosodic features with respect to attitudes and actors.

4.6.1 Objective attitude recognition

For each sentence and each attitude, we find the closest exemplar with the same number of syllables, within or outside the attitude considered. The distance measure used to define closeness in this case is the Root mean square error (RMSE) computed for the stylized contour parameters which define the sentences. Based on these distances, a confusion matrix is computed for all attitudes. The recognition rate of one attitude is represented by the recall of that attitude. An attitude confusion matrix is computed for each actor. A prosodic feature obtains a high attitude recognition rate if the attitude is very discriminated from the others i.e. it is encoded differently via the analyzed prosodic feature.

Syllable lengths associated to only one exemplar for an attitude are excluded from the algorithm. Table 4.4 presents the precision and recall values obtained per speaker for all

attitudes and overall features considered and table 4.5 presents the average recognition rate per actor and feature.

Table 4.4: Objective attitude recognition. Precision and recall values per speaker for all features. Values larger than 0.7 are highlighted in bold.

	Greg		Lucie		Georges	
	Recall	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
Declarative	0.64	**0.96**	0.57	**1.00**	**1.00**	**1.00**
Exclamative	0.45	**0.75**	0.48	**0.71**	**0.71**	**0.96**
Interrogative	**0.82**	**0.96**	0.42	0.64	**0.78**	**0.89**
Comforting	0.43	0.43	0.38	0.46	0.65	**0.74**
Tender	**0.74**	**0.71**	**0.71**	0.36	**0.79**	**0.82**
Seductive	0.47	0.50	0.44	0.39	0.43	0.54
Fascinated	0.64	**0.96**	0.52	0.54	0.62	**1.00**
Jealous	0.18	0.07	0.62	0.46	0.53	**0.71**
Thinking	**1.00**	**0.96**	**0.87**	0.46	0.65	0.46
Doubtful	**0.88**	**0.79**	0.65	0.39	**0.88**	**0.85**
Ironic	**1.00**	**0.82**	**0.77**	0.36	0.24	0.14
Scandalized	0.61	0.50	0.32	0.21	**0.85**	0.39
Dazed	0.66	**0.75**	0.44	0.39	0.58	0.64
Responsible	**0.70**	0.50	**0.82**	0.32	**0.80**	0.43
Confronted	**0.87**	0.48	0.45	**0.89**	**0.79**	0.39
Embarrassed	**1.00**	**0.79**	0.67	**0.86**	0.62	**0.75**

Table 4.5: Average recognition rates per speaker for all sets of features.

	Greg	Lucie	Georges
f_0	0.34	0.30	0.25
Head	0.51	0.44	0.49
Brows	0.45	0.30	0.19
Eyes	0.48	0.41	0.46
Gaze	0.48	0.30	0.38
All	0.68	0.51	0.66

The confusion matrices are illustrated in figure 4.14.

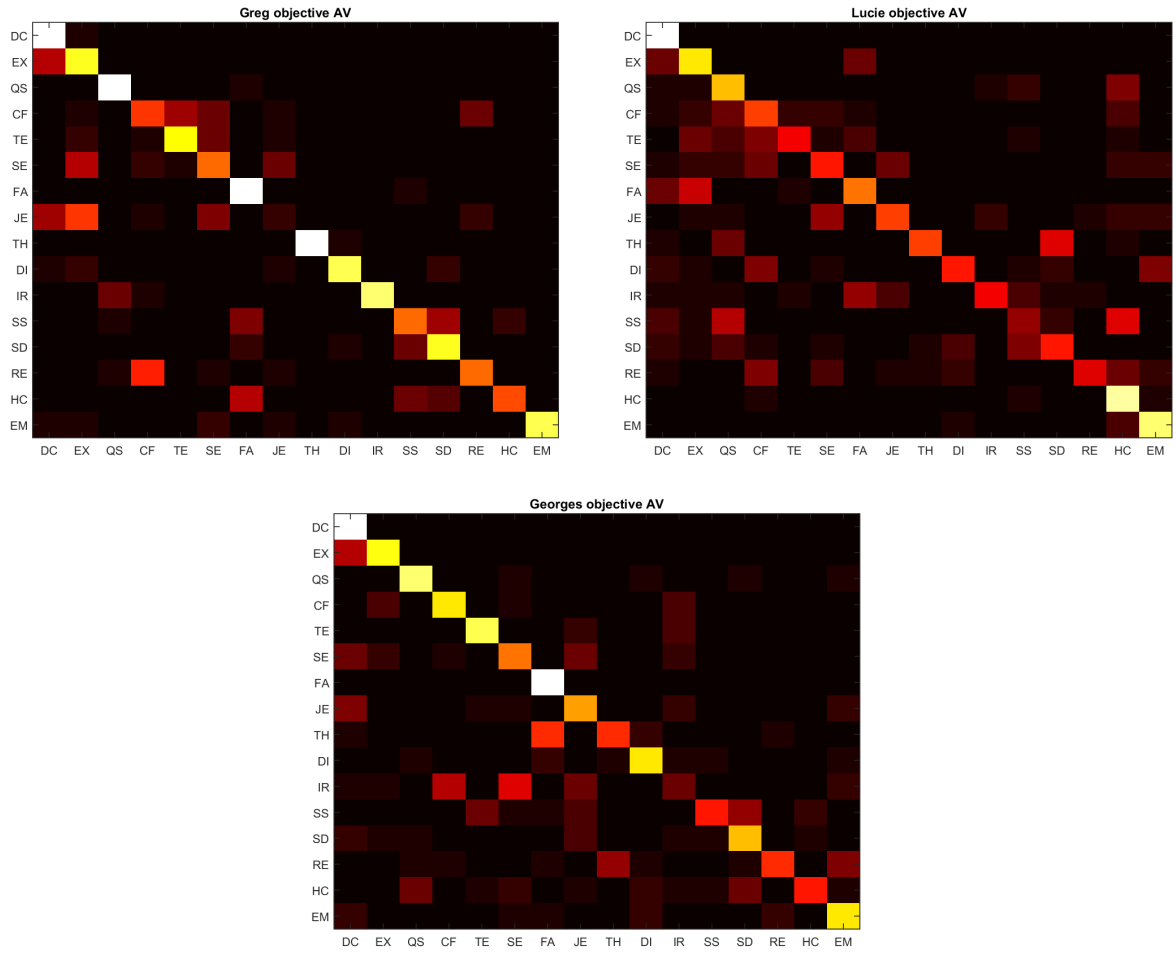

Figure 4.14: Heatmaps representation of the objective confusion matrices per speaker for all features.

Overall, the actor with best recognition rates for all features considered is Greg (68.23%), followed by the Georges (66.82%) and then by Lucie (51.79%). Note that this is not in accordance with the auto-evaluation results (see table 3.12). The recognition rate for all parameter types range between 19 % and 68 %. Visual parameter scores are again higher than the acoustic parameter scores, implying that specific patterns exist at the level of each visual region: head, brows, eyes and gaze. The feature for which the highest recognition scores were obtained is the head motion, which means that head motion is the most discriminant feature between attitudes.

Declarative has a very high recognition rate for all actors implying that patterns in neutral speech are completely different from the patterns displayed under the other expressive states. High rates are also obtained across speakers for the other modal attitudes and for Embarrassed.

However, most attitudes present different recognition rates across speakers. For example, Tender and Doubtful are well recognized by Greg and Georges, while for Lucie the rates are around 50% smaller. Scandalized has a 50% higher recognition rate for Greg than for the other two speakers. Thinking has a 96% rate for Greg while it obtains 46% for both Lucie and Georges being confused with Dazed and Fascinated, respectively.

An important observation is that each speaker presents a different least discriminated attitude: Jealous for Greg (very close to chance level: 7% as it is confused with Exclamative), Scandalized for Lucie (21%, confused with Confronted) and Ironic for Georges (14%, confused with Seductive).

4.6.2 Actor recognition

A similar algorithm is applied to compute the actor recognition rate per attitude. Actor recognition rate computation takes into account all utterances in the training dataset.

Table 4.6: Average recognition rates per attitude for all sets of features.

	f0	**Head**	**Brows**	**Eyes**	**Gaze**	**All**
Declarative	0.69	1	0.98	1	0.954	1
Exclamative	0.51	0.91	0.95	0.98	0.91	0.98
Interrogative	0.72	0.80	0.83	0.94	0.84	0.95
Comforting	0.31	0.91	0.65	0.80	0.77	0.91
Tender	0.38	0.83	0.75	0.77	0.73	0.94
Seductive	0.55	0.79	0.78	0.61	0.65	0.84
Fascinated	0.71	0.82	0.60	1	1	0.97
Jealous	0.61	0.80	0.79	0.46	0.52	0.79
Thinking	0.58	0.64	0.60	0.61	0.58	0.76
Doubtful	0.80	0.75	0.78	0.90	0.83	0.91
Ironic	0.42	0.76	0.77	0.64	0.59	0.84
Scandalized	0.38	0.75	0.71	0.69	0.72	0.90
Dazed	0.36	0.78	0.51	0.73	0.76	0.78
Responsible	0.47	0.69	0.73	0.67	0.77	0.85
Confronted	0.35	0.71	0.80	0.72	0.79	0.89
Embarrassed	0.41	0.77	0.82	0.89	0.65	0.95

The attitudes which are most actor-specific are: Declarative, Exclamative, Fascinated, Embarrassed, then followed by Tender, Doubtful and Scandalized. The lowest rates are obtained for Jealous, Thinking and Dazed, showing that these attitudes are expressed similarly across speakers. We observe that the type of feature accounts differently to the average recognition rate per attitude. For example, Jealous has a high audio recognition

rate but low rate contributions of eyes and gaze, while Dazed has a low audio rate and higher contributions of eyes and eyes. The most discriminant feature remains head motion while f_0 receives the lowest rates, especially for Comforting, Confronted, Tender and Dazed.

For evaluation purposes, the series of objective measures is followed by several perceptual tests. Our goal is to compute the correlation between perceptual and objective confusion matrices and to tentatively assess how the difference in contour patterns explains the perceptual judgments.

4.7 Perceptual assessment

The perceptual tests are carried on the original video recordings (original voice and video), and also on animations performed using original motion capture data. The purpose of these tests is to: (a) compare auto-evaluation with third parties, (b) measure the contribution of discriminant prosodic shapes to perception test results and (c) obtain a perceptual validation of original performances and the proposed animation platform.

All perceptual evaluations consist in forced-choice identification tests. Subjects are asked to label heard and/or viewed performances with one of the 16 attitudes. The perceptive tests were carried on a crowd-sourced online platform. Only data from the native French participants are considered here.

4.7.1 First perceptual test: original video

A first perceptual test is carried on video data recorded by Lucie for all attitudes and the three modalities: audio, visual and audio-visual. 80 anonymous participants accessed the online evaluation test [1] which was carried using a normal web browser. Before starting the test, user information is collected (e.g. age, sex, native language), of which the most important is whether French is the native language. Participants were given the definitions of all the attitudes and asked to identify the attitudes performed by the chosen actor using audio files, video-only files and audio-video files. For each modality, the participants have to label 32 random performances. Each set of performances is obtained by randomly selecting 2 performances from each attitude with the condition of not retrieving identical consecutive sentences or attitudes. Performances last up to 5 seconds and can be played several times until they are labeled. Only the answers provided by native speakers are taken into account. After test completion, users have the option of leaving commentaries related to the experience. Evaluation tests took on average 25 minutes to be completed.

[1] The test is no longer available online.

This test used original audio and video files obtained from sequences of 2D images which are synchronized with sound. Confusion matrices are built based on the choices made in the perceptual tests for each modality. For representation of attitudes, we use the abbreviations indicated in Tables 3.3 and 3.2. The recognition rates for all modalities are presented in Table 4.7. The confusion matrices are represented in figure 4.15.

Table 4.7: Recognition rates for the online test containing original audio and/or video performances recorded by Lucie.

	Audio	Video	Audio-Video
Declarative	0.148	0.251	0.170
Exclamative	0.048	0.014	0.066
Question	**0.725**	0.412	**0.788**
Comforting	0.130	**0.563**	0.290
Tender	0.481	0.340	0.368
Seductive	**0.506**	**0.731**	**0.761**
Fascinated	0.192	0.289	0.391
Jealous	0.000	0.250	0.167
Thinking	**0.597**	**0.722**	**0.814**
Doubtful	0.180	0.083	0.214
Ironic	**0.672**	0.383	**0.640**
Scandalized	0.435	0.250	0.264
Dazed	0.325	0.294	0.357
Responsible	0.314	0.462	0.273
Confronted	0.216	0.361	0.150
Embarrassed	0.263	0.464	**0.617**

The best recognized attitudes are: Question, Seductive, Thinking and Ironic, while the lowest result values are obtained for: Exclamative, Jealous, Doubtful and Confronted. Except for Exclamative, most attitudes have a recognition score high above chance level (0.0625%). Jealous was not recognized at all in the audio modality. Declarative also presents low scores as it was confused with Exclamative and Responsible in all modalities. According to the comments left by users, the meaning of the Confronted attitude was not well understood, thus explaining the low score obtained. A factor which may explain why for some attitudes (Declarative, Jealous, Responsible, Confronted) the scores obtained for video-only were higher than the ones including audio cues is the difficulty of dissociating

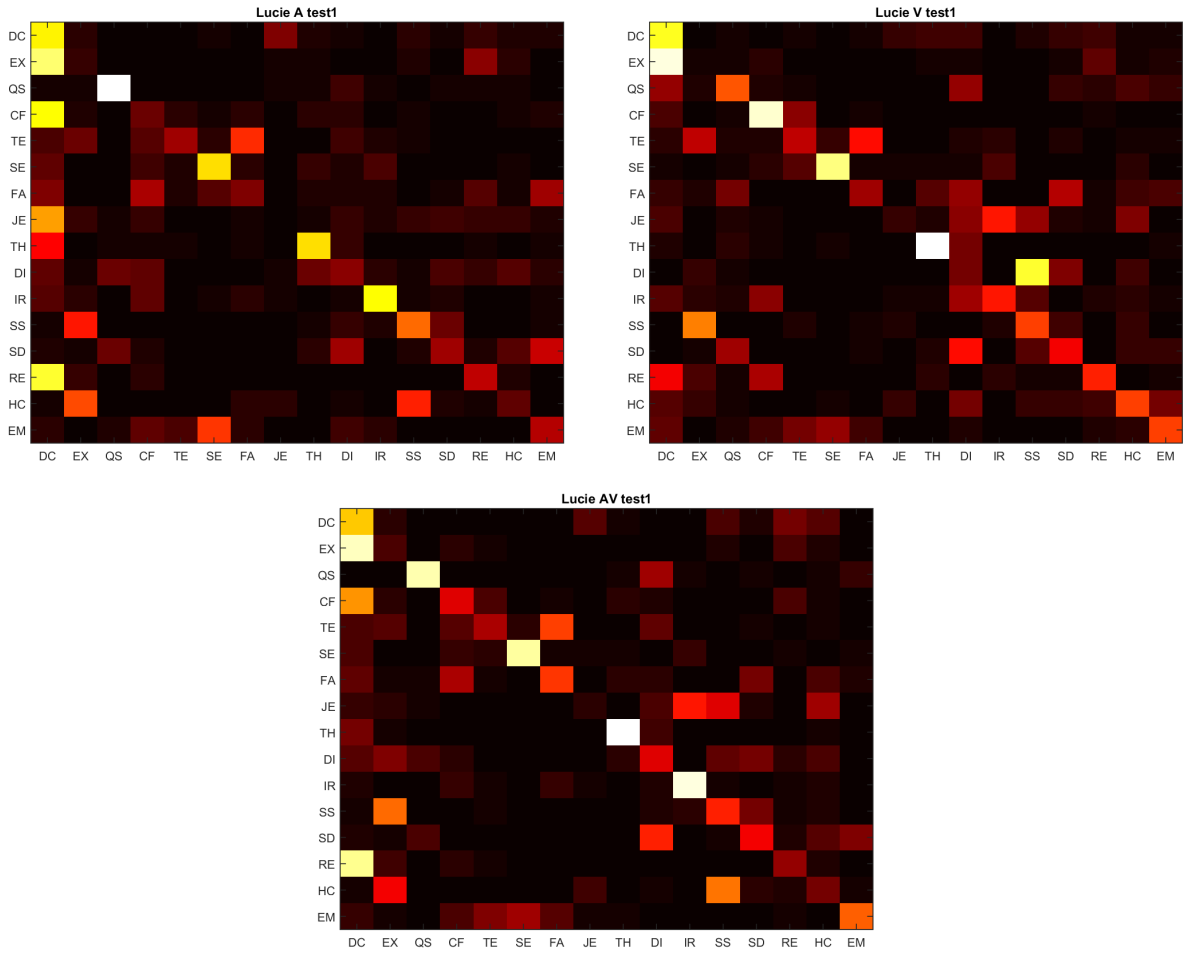

Figure 4.15: Heatmap representation of the confusion matrices for recognition rates obtained for audio, video and audio-video performances of Lucie.

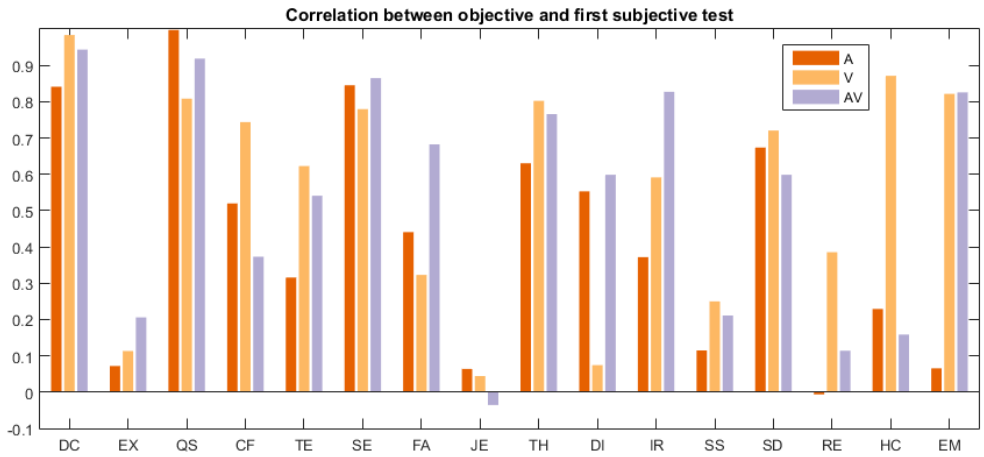

Figure 4.16: Correlation values between the recognition rates obtained for the objective and the first perceptual test. The values are computed per attitude and per modality.

the attitude from the semantic content of some carrier sentences such as the sentence: "Ce n'est pas possible" ("This is not possible"). As the test was considered too long, we decided to keep only the audiovisual information in the following perceptual tests.

Next, we compute the correlations between the recognition rates obtained in this test and the corresponding objective recognition rates (see figure 4.16).

The attitudes which present high correlation values are Declarative, Question, Seductive, Thinking, Dazed and Embarrassed. These attitudes are highly correlated for all modalities, except for the attitude Embarrassed, which presents no correlation for the audio modality. The attitudes whose recognition rates are not correlated are: Exclamative, Jealous, Scandalized, Responsible and Confronted. This indicates that the audiovisual prosodic features used in the objective test are not sufficient to explain the perceptual results obtained in this perceptual test. The correlation values are small because the discrimination between attitudes behaved differently. For instance, in the objective test Scandalized was mostly confused with Confronted and Jealous obtained 46% recognition rate, being mostly confused with Seductive (see figure 4.14). In this perceptual test, Scandalized was mostly confused with Confronted and Jealous obtained a lower recognition rate 17 % while being mostly confused with Ironic.

4.7.2 Second perceptual test: realistic animation

A second test was performed using both video and animated original performances with the goal of assessing the realistic animation system for facial animation. We conducted a perceptual test to compare the recognition performance of participants watching original video data vs. the recreated audiovisual animations. Participants were instructed to label 16 original *video* stimuli and then 16 synthetic *animated* stimuli, with the appropriate attitude. Stimuli were randomly selected from a subset of 7 sentences out of the 35 training sentences. The online test can be found at [2].

A total of 77 French native speakers participated in this experience. Table 4.8 presents the precision and recall obtained for each attitude for the original videos and the animations.

For the video test, the best recognized attitudes are: Declarative, Question, Comforting, Seductive, Thinking, Ironic, Scandalized and Embarrassed. High recognition rates are also obtained for Declarative, Question and Scandalized for the animation test. However, the low precision values for Declarative and Exclamative show that these attitudes have a higher chance of being chosen when another attitude is played. The least recognized attitudes are: Exclamative, Doubtful, Responsible, Confronted. These attitudes are subtle in audiovisual changes (as opposed to Scandalized for example) and have meanings

[2] http://www.gipsa-lab.fr/∼ adela.barbulescu/test1/

Table 4.8: Recognition rates for the second perceptual test. Precision and recall obtained for videos and animations of original data recorded by Lucie for all attitudes. Values above 0.3 are outlined in bold.

	Video		Animation	
	Recall	Precision	Recall	Precision
Declarative	0.25	**0.61**	0.18	**0.66**
Exclamative	0.07	0.09	0.07	0.14
Question	**0.50**	**0.60**	**0.40**	**0.31**
Comforting	0.26	**0.45**	0.16	0.23
Tender	**0.51**	0.29	0.29	0.15
Seductive	**0.60**	**0.45**	**0.56**	0.19
Fascinated	**0.39**	0.21	**0.43**	0.20
Jealous	**0.50**	0.10	0.10	0.04
Thinking	**0.50**	**0.39**	**0.34**	0.25
Doubtful	0.13	0.18	0.09	0.08
Ironic	**0.32**	**0.43**	0.14	0.16
Scandalized	**0.66**	**0.58**	**0.38**	**0.35**
Dazed	0.18	0.14	0.09	0.06
Responsible	0.11	0.09	0.00	0.00
Confronted	0.23	0.13	0.15	0.10
Embarrassed	**0.47**	**0.36**	0.25	0.18

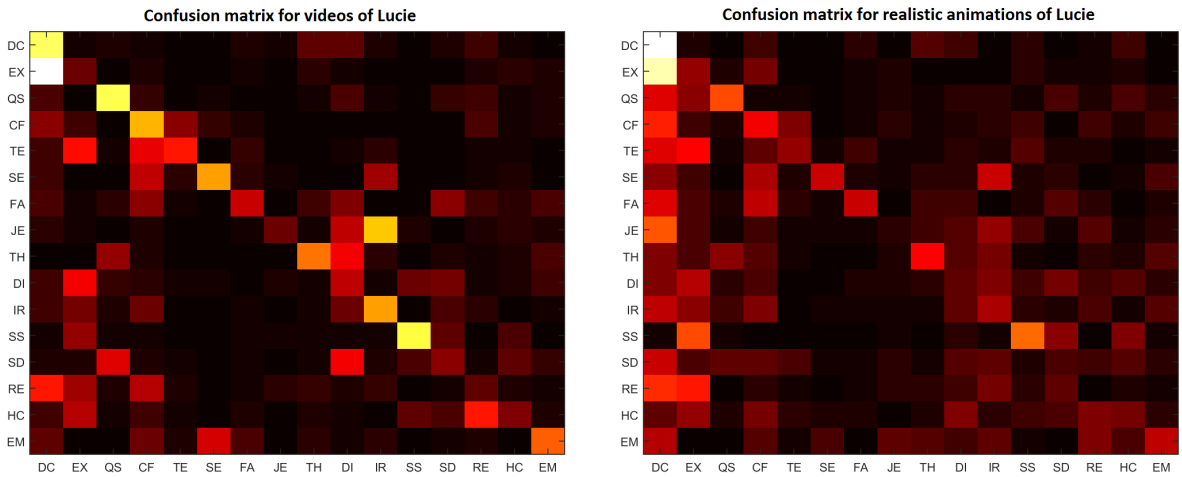

Figure 4.17: Confusion matrices for the second perceptual test. Heatmap representation of the confusion matrices for recognition rates obtained for video and animation representation of Lucie. The two matrices have a correlation r=0.81 (p<0.01).

that are more difficult to grasp and dissociate from the meaning of the sentence. All these attitudes are generally confused with Declarative in both tests. This may be explained, especially in the case of animation, by the fact that users tend to select the Declarative or Exclamative style in the lack of a more suitable expressive category.

Some notable results: Exclamative is generally confused with Declarative, Doubtful with Exclamative, Dazed with Declarative and Doubtful, Responsible with Exclamative. As expected, all recognition rates decrease as the animation version is used and it is most apparent for attitudes such as: Doubtful, Ironic, Dazed, Responsible and Embarrassed. However, the attitude recognition rates for video and animated material are highly correlated (Pearson's r=0.81, $p < 0.01$).

Figure presents the correlation values computed between the recognition rates obtained for the objective test and the second perceptual test. The values are computed separately for the scores obtained for video and animated performances.

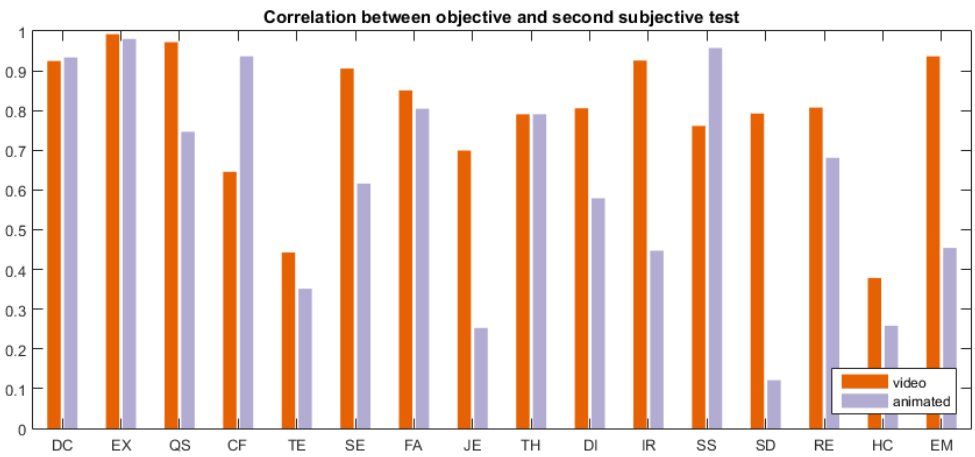

Figure 4.18: Correlation values between the objective and the second perceptual test recognition rates. The values are computed per attitude and modality (audiovisual).

Overall, most attitudes present a high correlation and the animated performances present smaller correlation values compared to the video version. The attitudes which present high correlation values are Declarative, Exclamative, Question, Comforting, Seductive, Fascinated, Thinking, Scandalized and Responsible. Particularly low correlations are obtained for the animated version for Tender, Jealous, Dazed and Confronted. An interesting note is that for the video version, Jealous presents now a high correlation as it is mostly confused with Ironic, similar to the result obtained with the objective test. This behavior is different from the one observed in the previous test.

For the following studies on data analysis and generation of dramatic attitudes, we decided to focus on a subset of attitudes in order to simplify the users' selection and the evaluation process. We will consider among these the attitudes which received the

highest recognition rates.

4.7.3 Third perceptual test: cartoon-style animation

A third perceptual test was carried on the cartoon style animation system. We focused on a subset of 8 attitudes (Tender, Seductive, Fascinated, Jealous, Thinking, Ironic, Scandalized and Embarrassed) and used both actors in the test. Each subject is asked to recognize the attitudes of a set of 32 animations which present original animated performances of the two actors alternating (such that no two consecutive performances contain identical attitudes, sentences or actors). For one test, random sentences are chosen such that each attitude appears twice for each actor. Also, we only used a subset of 6 sentences, all containing more than 5 syllables because raters are more comfortable evaluating longer sentences. The online test can be found at [3].

A total of 53 French native speakers participated in this experience. Table 4.9 presents the precision and recall obtained for each attitude for the animations.

Table 4.9: Recognition rates for the third perceptual test. Precision and recall obtained for cartoon style animations of the two actors for 8 attitudes. Values above 0.5 are outlined in bold.

	Greg		**Lucie**	
	Recall	**Precision**	**Recall**	**Precision**
Tender	**0.53**	**0.51**	0.40	0.38
Seductive	**0.58**	**0.51**	**0.55**	**0.50**
Fascinated	**0.54**	**0.72**	0.38	0.33
Jealous	0.48	0.35	0.36	0.19
Thinking	**0.63**	**0.75**	0.42	**0.52**
Ironic	0.35	0.48	0.45	0.36
Scandalized	**0.82**	**0.68**	**0.71**	**0.84**
Embarrassed	**0.63**	0.45	0.40	**0.62**

All recognition rates are obtained above chance level (12.5%), with the best rates being obtained for Seductive, Thinking and Scandalized, and lowest for Jealous for both actors. Generally, Greg was better recognized, this observation being supported also by the results obtained in the objective attitude recognition tests. Notable observations are: for Greg, Jealous and Ironic are interchangeably confused, while for Lucie, Jealous is confused with Embarrassed. These attitudes are expressed using similar intensity and rhythm and also more subtle facial expressions.

The rates obtained for Lucie are compared to the ones obtained in the previous test,

[3] http://www.gipsa-lab.fr/∼adela.barbulescu/test_recognition/

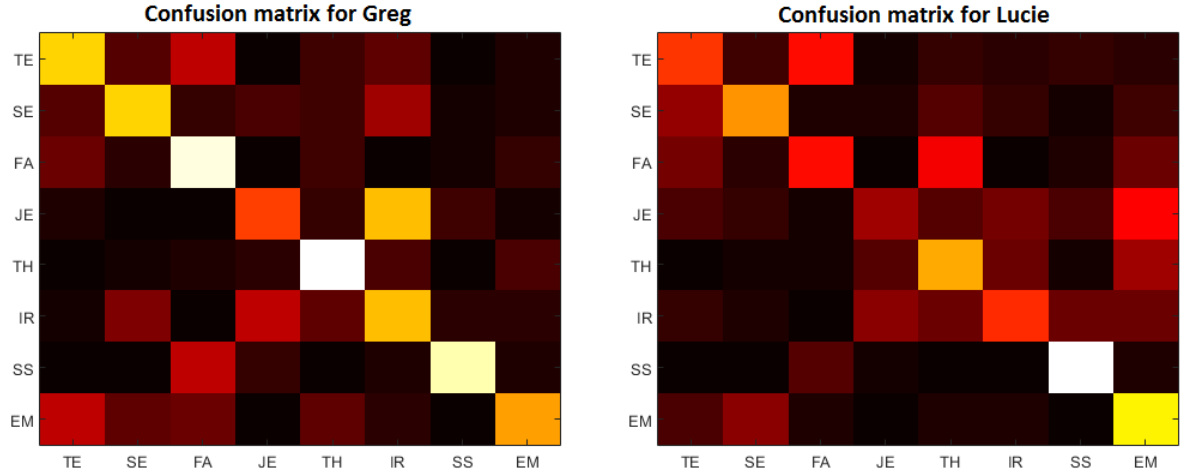

Figure 4.19: Confusion matrices for the third perceptual test. Heatmap representation of confusion matrices for recognition rates of cartoon-style animations of Greg and Lucie.

showing very strong correlations between the two types of animation platforms for Lucie (Pearson r=0.88 for p<0.05) and a slight smaller correlation for the video material (Pearson r=0.84 for p<0.05). The recognition rates for the cartoon style animations are generally more than 50% increased than in the case of realistic animations, especially in the case of Scandalized and Embarrassed, which is expected since we use a smaller number of attitudes.

We also compute the correlations between the recognition rates obtained for the objective test and the third perceptual test. Correlation values are presented in figure 4.20.

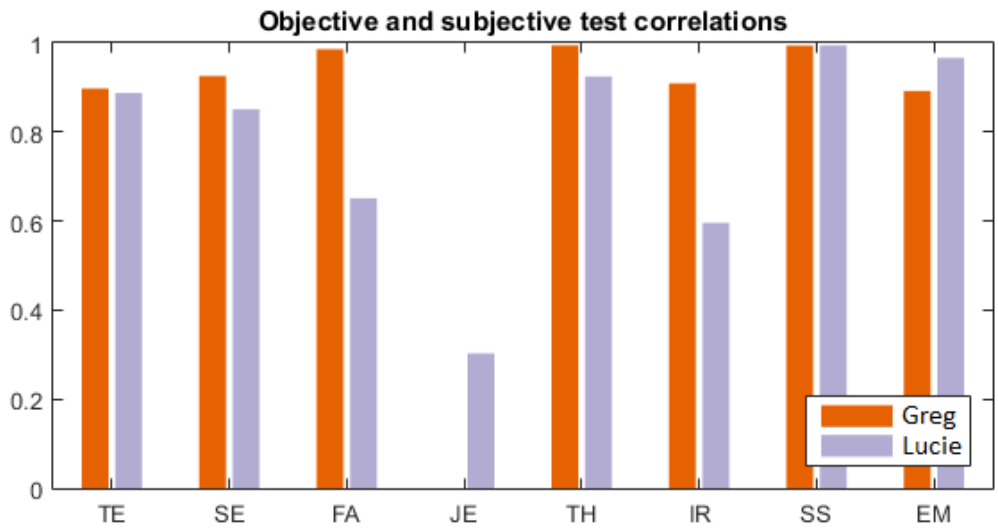

Figure 4.20: Correlations obtained between objective and the third perceptual test results per attitude.

Overall, the attitudes present very strong correlations and higher values for Greg. The

exception is represented by the attitude Jealous, which shows low rates in both objective and perceptual tests but either there is no correlation (Greg) or it remains weak (Lucie). This happens because in the objective test Jealous was mostly confused with Seductive for both Greg and Lucie (we take into consideration only the attitudes studied in this perceptual test), while in this test Jealous was confused with Ironic and Embarrassed, respectively.

Discussion. The series of objective and perceptual studies were carried in order to provide a validation of our expressive corpus. The results show that for the different attitudes the actors use expression techniques which can be more or less easily discriminated depending on the modality used.

The correlations between recognition rates obtained in the objective and the perceptual tests for the audiovisual modality are very strong for attitudes such as Declarative, Question, Seductive, Fascinated, Thinking and Embarrassed. This means that the audiovisual prosodic contours do explain the results of perceptual tests. This remains valid for performances rendered using the proposed animated systems. Attitudes such as Exclamative, Jealous, Confronted and Responsible generally show weak or no correlations, because they are generally confused with different attitudes across tests.

In the last perceptual test we study a subset of attitudes and observe similar behaviors for the two actors: Tender is confused with Fascinated and Ironic is confused with Jealous. The attitude Jealous was initially chosen because it was very well recognized during the auto-evaluation tests. However, it obtains very low recognition rates both in the objective and the online perceptual tests.

4.8 Limitations

In this chapter we defined an objective measure which can be used to compute the discrimination between attitudes for a given actor. This measure is based on the distances between exemplars (stylized prosodic contours) containing the same syllable number. However, we don't have a similarity measure for prosodic contours which contain different numbers of syllables. For this reason, we make observations on contour shapes with the goal of improving a statistical model which would be able to learn and generate the prosodic contour patterns.

A specific problem is encountered at the level of eye blink motion as the prosodic contours show irregular behavior within attitudes. While blinking frequency is influenced by physiological and expressive functions, blink motion cannot be considered a segmental feature. We must therefore define a separate statistical model to generate this motion.

Another interesting problem discussed in this chapter is the perceptual evaluation of

original data. In order to measure to which extent perception of expressiveness is influenced by the representation of the performance, we should carry comparative tests including sentences represented with video data and the two animation styles.

The generally high recognition rates obtained in the auto-evaluation test indicate that participants of online tests should first be accustomed to the animated performances. This could be done using a short training process (viewing random animations for a few minutes).

4.9 Summary

This chapter described the analysis techniques and evaluation tests carried on the original recorded data from our dataset and discussed the results obtained.

Section 4.1 presented the set of prosodic and segmental features extracted from the recorded audiovisual data. In Section 4.2, we characterized these features according to their structural properties (segmental or suprasegmental level).

Section 4.3 presented the stylization of audiovisual prosodic contours and introduced the virtual syllables as pre- and post-phonatory silences with the duration of an average syllable. In Section 4.4, we presented a discriminative analysis of audiovisual prosody extracted from the marginal and virtual syllables, which shows that the prosodic information is attitude-specific. Section 4.5 presented observations of attitude-specific behaviors of the stylized audiovisual contours.

Stylized contours were then used in Section 4.6 to compute an objective measure for the inter-class differences with respect to attitudes or actors. The results obtained by this algorithm were then compared with the perceptual evaluation recognition scores in Section 4.7.

Chapter 5

Generation of audiovisual prosody

This chapter describes the techniques proposed for the generation of audiovisual prosody. Our goal is to model and generate audiovisual prosodic contours for a discrete set of dramatic attitudes. Audiovisual data analysis has shown that auditory and visual prosodic contours exhibit specific patterns for the expression of attitudes. This information is used to train separate statistical models for segmental vs suprasegmental contributions.

In order to generate a complete audiovisual performance, we also have to generate the expressive segmental features. However, the size and phoneme occurrence coverage of our dataset is not sufficient to contain all the realizations of phonemes in context (Le Maguer [Le Maguer, 2013] has shown that left and right contexts are sufficient to faithfully reproduce spectral variability of speech sounds). Our corpus is insufficient for a visual text-to-speech synthesizer. Therefore, segmental generation is done using voice conversion techniques i.e. we add expressiveness to pre-recorded declarative performances (considered as neutral). Figure 5.1 describes the pipeline for audiovisual expressive performance generation.

The source neutral audiovisual speech is phonetically transcribed and segmented in the same manner as all the utterances existing in the training corpus. The prosodic contours and segmental features are extracted using a vocoder and a motion analysis technique which also returns stylized prosodic contours. We propose three techniques for expressive trajectory generation: (1) frame-based approach, (2) exemplar-based approach and (3) a model-based approach which uses a sentence-level dynamic unit for prosody modeling. The trajectory generation module is followed by an audiovisual synthesis module in which the expressive speech and the facial animation are synthesized. The chapter presents also objective and subjective evaluations.

The following sections describe the approaches used for trajectory generation, focusing on the prosody generation.

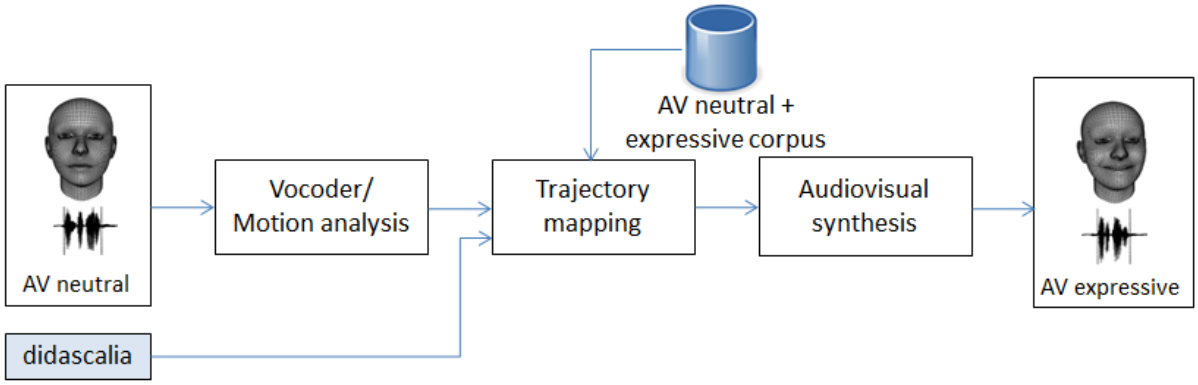

Figure 5.1: Outline of the generation system of audiovisual expressive performances. The system receives as input the neutral version of audiovisual speech and didascalia and returns a synthesized audiovisual expressive performance. The system consists in three steps: (1) audiovisual feature extraction from the neutral audiovisual speech, (2) trajectory generation (prosody and segmental features) and (3) audiovisual performance synthesis.

The organization of the chapter is as follows. In Section 5.1, we describe our frame-based approach for trajectory generation. In Section 5.2, we describe our exemplar-based approach for the same problem. In Section 5.3, we describe our model-based approach, where we generalize the SFC method to the case of audiovisual utterances, based on the findings of Chapter 4. In Section 5.4, we give details of the audiovisual reconstruction step, which is common to all methods. In Section 5.5, we compare the three methods experimentally. In Section 5.6, we explain the limitations of our work and propose directions to overcome them.

5.1 Frame-based generation

The first method proposed for expressive audiovisual generation is a frame-based approach that consists in standard Gaussian mixture model regression. This method was initially implemented by the authors in the context of speaker identity conversion [Barbulescu et al., 2013]. Speaker identity conversion refers to the problem of converting multimodal features between different speakers such that the converted performance of a source speaker can be perceived as belonging to the target speaker. Representative features for speaker individuality are the speaker's voice (represented by timbre and pitch) and articulation (represented by PCA components of the facial movements). In the identity conversion context, a mapping was created between audiovisual speech data collected from one speaker (source) and the audiovisual speech data collected from another speaker (target). The training speech data was neutral and no expressive labels were used.

Similarly, we use the GMM regression method for creating a mapping function between a

neutral performance (source) and the expressive version of the performance for a given attitude (target). Detailed information about the GMM regression [Stylianou et al., 1995] is presented in the Annex A.1.

The most common technique in predicting the pitch of the speaker is that of normalizing the mean and variance of the source speaker's (log-) f_0 distribution to the target speaker's mean and variance. This can be implemented by a straightforward linear transformation at the frame-level, such that given the input source pitch f_0_s, the predicted pitch f_0_t is obtained:

$$f_0_t = \frac{\sigma_s}{\sigma_t}(f_0_s - \mu_s) + \mu_t \tag{5.1.1}$$

where μ_s and μ_s represent the mean log pitch values for source and target speaker and σ_s and σ_s represent the variance log pitch values for source and target speaker respectively.

5.1.1 Slope feature

A mapping function between source and target audiovisual speech signals can be built if the two signals are aligned according to a chosen criteria. In our case, an alignment path between corresponding frames is retrieved using a Dynamic Time Warping (DTW) algorithm [Berndt and Clifford, 1994]. For parallel data, DTW aligns the source and the target feature sequences by minimizing the spectral distance with temporal structure constraint.

A correlation is observed between the local slope of the path and the instantaneous rhythm of speech. By computing the local slope on a window of frames (which is centered on the current frame), we extract a new prosodic feature which characterizes the change of local speech rate. At frame t, for the path alignment between two speakers, the slope parameter is given by:

$$slope_t = \frac{\Delta p_{attitude}}{\Delta p_{neutral}} \tag{5.1.2}$$

where $p_{attitude}$ and $p_{neutral}$ represent the attitude-specific alignment path vectors.

Figure 5.2 presents the DTW alignment path and the corresponding contour of the slope feature.

The conventional MMSE approach [Stylianou et al., 1995] can be applied by concatenating the slope parameter to the spectral feature vector in order to estimate the speaker-specific rhythm of speech for a new utterance [Barbulescu et al., 2013].

Speaking style features are extracted by computing the local alignment path slope by sliding a 7-dimensional window (3 frames before and after the current frame) for each

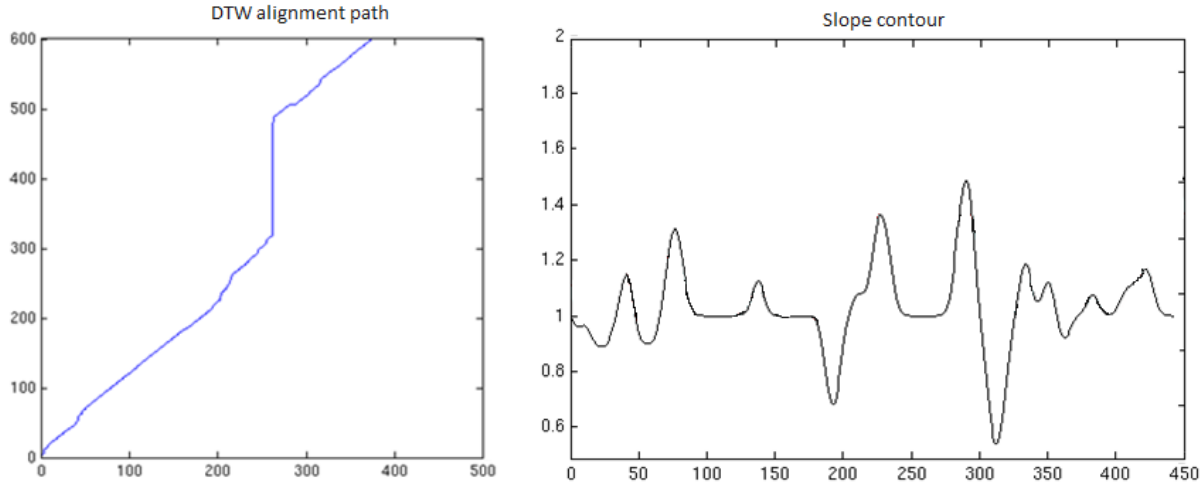

Figure 5.2: DTW alignment path and slope contour. Left: DTW alignment path for features extracted from the neutral and expressive (Fascinated) version of the sentence "Et hier encore vous avez jouée comme un archange" performed by Lucie. Right: the corresponding slope contour obtained.

frame. This slope parameter is concatenated to the spectral feature vector. Therefore we call our method "GMM conversion with slope".

GMM training. In the expressive conversion context, the source is represented by neutral audiovisual speech (declarative attitude) and the target is an expressive audiovisual speech (any other attitude). For a given attitude we train a GMM with audio and visual features. For each GMM, a joint neutral (source) and expressive (target) vector is created by concatenating 25 mel-cepstra coefficients and 5 bark-scale aperiodicities, 24 visual parameters (as presented in 4.1.2), for both neutral and expressive style, and then we add the local path slope coefficient. The GMMs are trained using 16 Gaussian components. Figure 5.3 illustrates the GMM training process.

GMM regression. GMM regression returns the predicted mel-cepstra, aperiodicities, PCA motion components and the slope contour. The slope contour is then used to build a timeline for both the audio and visual predictions. The expressive pitch contour is obtained by scaling the source contour to the target register. In order to compare our results with original target data, we test the GMM models on the utterances extracted from the training database (35 utterances per attitude) using a leave-one-out training framework. Throughout this chapter all testing will be carried on utterances extracted from the training dataset. Figure 5.4 illustrates how prosodic and segmental features are generated using this frame-based approach.

Figure 5.5 illustrates an example of expressive target (original) and predicted slope parameter contours.

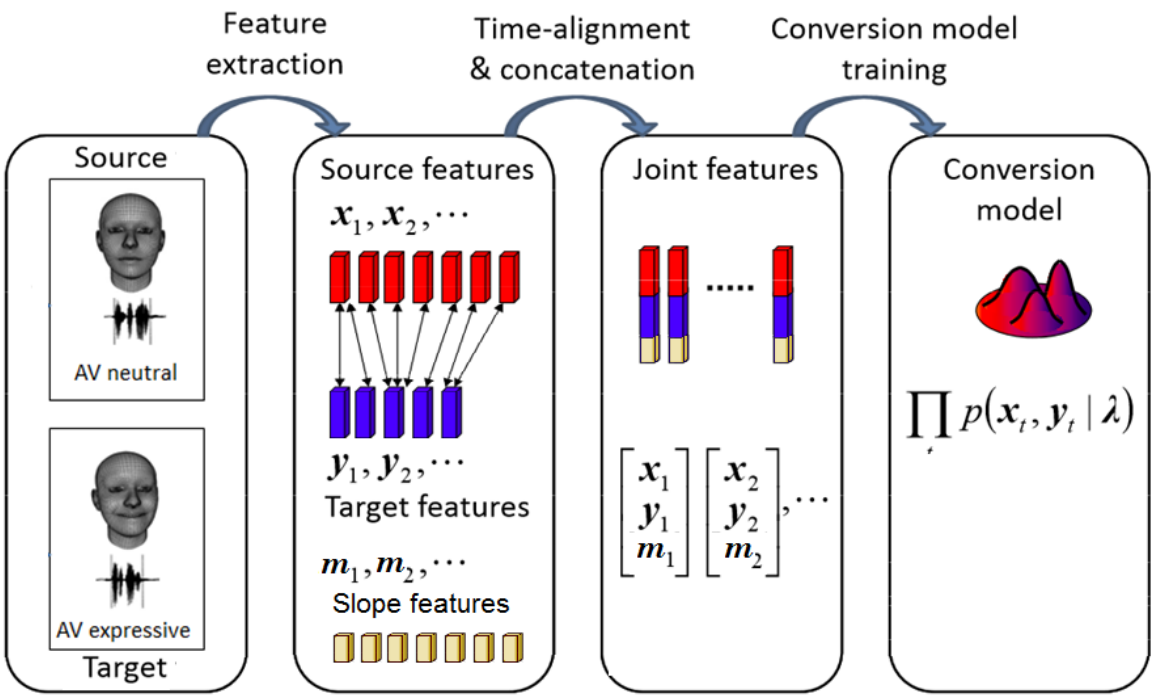

Figure 5.3: Training with the frame-based approach. Features are extracted from source (neutral) and target (expressive). Then they are aligned using DTW and the slope features are computed. GMMs are trained using the aligned and concatenated features.

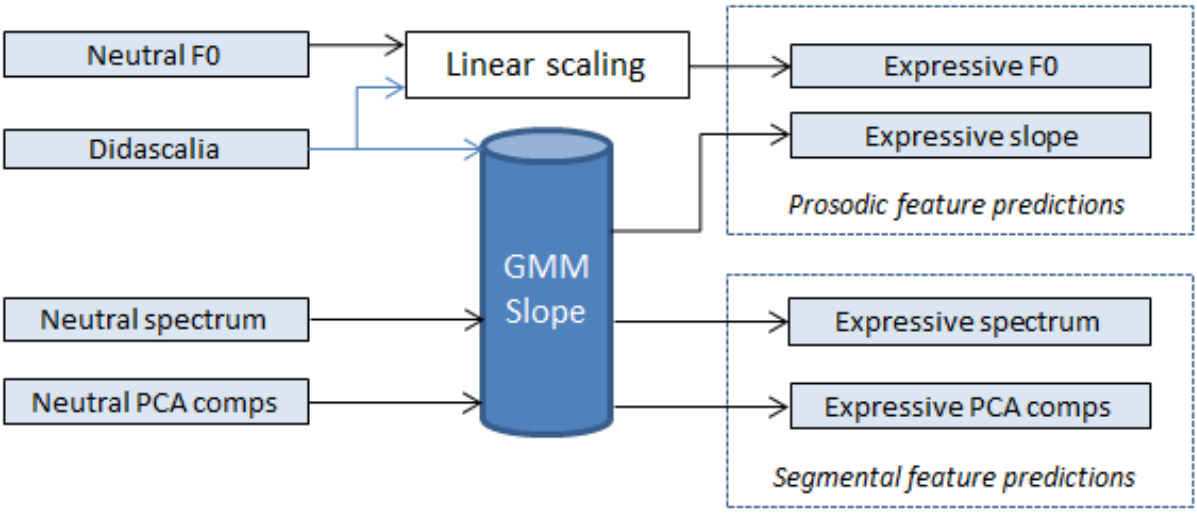

Figure 5.4: Trajectory generation with the frame-based approach. Prosody generation: F_0 is obtained by scaling the neutral pitch contour and the segmental timeline (rhythm) is determined by the predicted slope. Note that we consider all the visual PCA components as segmental because they were obtained using a frame-based technique, just as the voice spectrum.

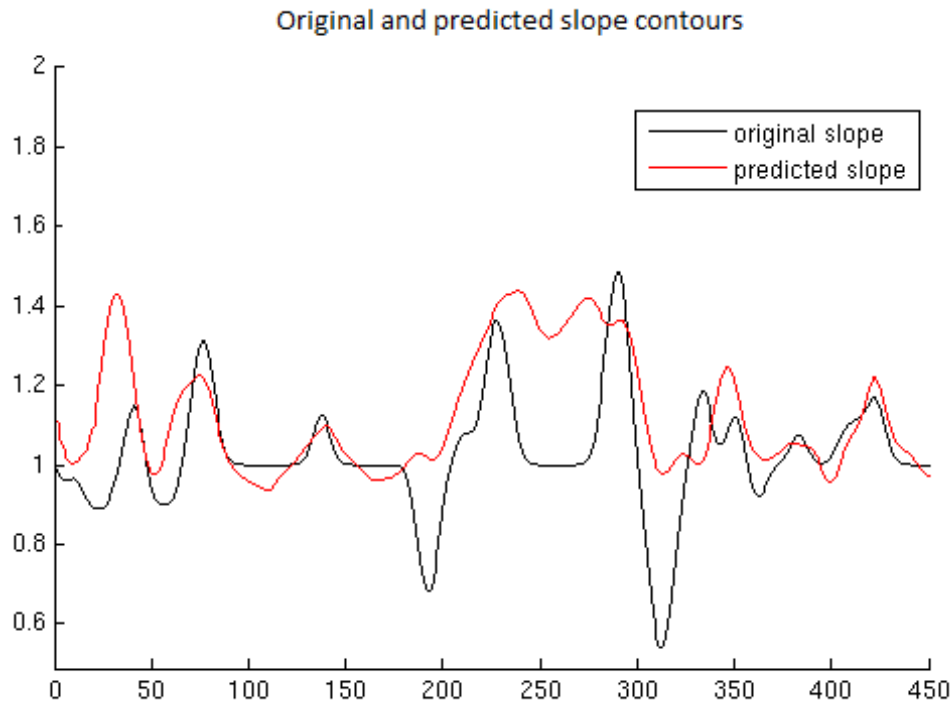

Figure 5.5: Original vs predicted slope. Example of target and predicted slope for attitude Fascinated, sentence: "Et hier encore vous avez jouée comme un archange" performed by Lucie. Note that the source contour would always be equal to 1.

5.2 Exemplar-based generation

In order to generate expressive performances, we separately use a frame-based approach for converting segmental features and an exemplar-based approach to generate suprasegmental features.

5.2.1 Random selection of exemplars

As we work with sentence-level contours, we propose to interchange the original prosodic contours between sentences of same length as a method to generate audiovisual speech. This method is exemplar-based since the data used in generation is extracted from the original database. Exemplars are characterized by the audiovisual stylized contours. They are indexed by the attitude and number of syllables.

Similarly to the approach used in [Levine et al., 2009], we select a random exemplar and substitute the audio-visual prosody to the one of the source utterance. Therefore, given an attitude and a new sentence, a random sentence with identical syllabic number is selected from our corpus (thud irrespective to the verbal content). The stylized audiovisual prosodic contours (melody, rhythm, upper-face and head motion) are then retrieved

and used for the reconstruction of prosodic contours.

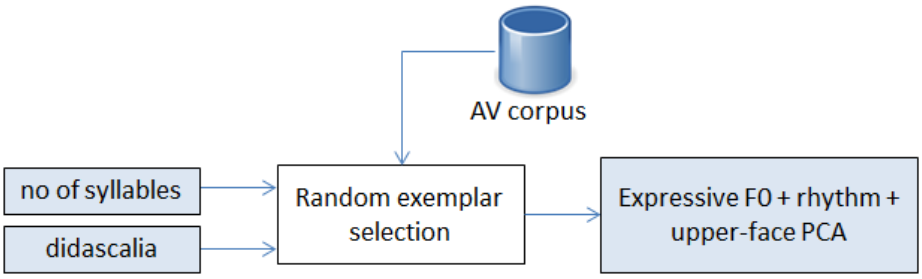

Figure 5.6: Prosody generation with frame-based technique. Exemplar-based method for the generation of audiovisual speech.

5.2.2 Prediction of segmental features

Features such as lip movements and spectra depend on the underlying phoneme pronounced at a certain position within the speech. For this reason, the generation of expressive segmental features is performed from neutral performances using a frame-based conversion technique as described in Section 5.1. The conversion with slope is not used in this case because rhythm is obtained from the selected exemplar.

GMM training. For a given attitude we train a GMM with audio and visual features. For each GMM, a joint neutral (source) and expressive (target) vector is created by concatenating 25 mel-cepstra coefficients and 5 bark-scale aperiodicities and 8 visual parameters (the lower-face PCA components as presented in 4.1.2), for both neutral and expressive style. The GMMs are trained using 16 Gaussian components.

GMM regression. GMM regression returns the predicted mel-cepstra, aperiodicities and lower-face PCA motion components. The predicted expressive signal presents the time structure of the neutral performance.

The exemplar selection and the segmental feature prediction methods are combined by stretching the segmental features to the rhythm imposed by the exemplar.

5.3 Model-based generation

The last proposed method is similar to the exemplar-based as the segmental and suprasegmental generation approaches are separate. In order to generate expressive performances, we separately use a frame-based approach for converting segmental features and a parametric prosodic model to explicitly generate suprasegmental features. Figure 5.7 presents the outline of the proposed system:

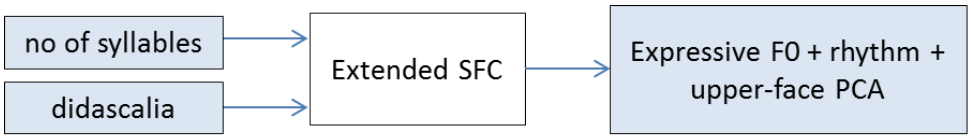

Figure 5.7: Prosody generation with model-based technique. The SFC model uses phonotactic information and didascalia in order to generate prosodic stylized contours.

5.3.1 Prosodic feature prediction

We test hypotheses described in chapter 4 with the trainable model SFC for generating speech prosody i.e. obtaining voice frequency contours and syllabic durations. Research within the field of audiovisual prosody has revealed strong correlations between voice frequency and gestures, head movements, facial expressions etc. The description of stylized prosodic contours for the recorded attitudes and the objective measures show that expressive performances are encoded via attitude-specific and sometimes actor-specific patterns. Therefore we test the hypothesis that communicative functions are implemented using prototypical audiovisual prosodic contours.

To test this hypothesis, we propose a framework for generating the audiovisual prosody of attitudinal functions by extending the SFC model to motion components. Along with f_0 and syllable durations, we also include visual data: head and eye movements and facial expressions of the upper part of the face. The stylization of audio-visual parameters i.e. three keypoints per syllable and the rhythmic component remain identical to the ones described in [Bailly and Holm, 2005]. Detailed information about the SFC learning and prediction is presented in the Annex A.2.

In the general framework, the process of overlap and add of contours associated to each communicative function is also applied to motion components (as simple addition) in order to obtain the final prosodic contours. The rhythmical unit is the syllable and the carrier unit of all functions is the entire utterance. Audiovisual prototypical contours are here predicted given an attitudinal function and *scope* of the carrier utterance i.e. we do not consider other communicative functions such as phrasing or focus.

Prosodic contours of different scopes belonging to the same function constitute a morphologically coherent *family of contours*. Such a family of contours can be generated using a so-called *contour generator* which is implemented using a simple neural network that is trained in an analysis-by-synthesis loop. Specifically, the neural networks are chosen because we need a non-linear mapping between phonotactic information (length of sentence and position within the sentence) and the stylized contour values. We use simple feed-forward networks with a hidden layer of 17 neurons which use the logistic activation function. The implementation and learning (using the backpropagation algorithm) is

done using the SNNS library [1].

SFC prediction. As explained in A.2, the contour generator receive as input didascalia and phonotactic information extracted from the number of syllables. The output is represented by the stylized prosodic contours which consist in: 3 values for melody, 1 value for rhythm and 48 for motion (3 times 5 PCA components for head motion, 2 PCA components for gaze rotation, 3 components for eyebrow motion and 6 components for eye-area expressions). Figure 5.8 illustrates how contour generators are used for the prediction of stylized contours.

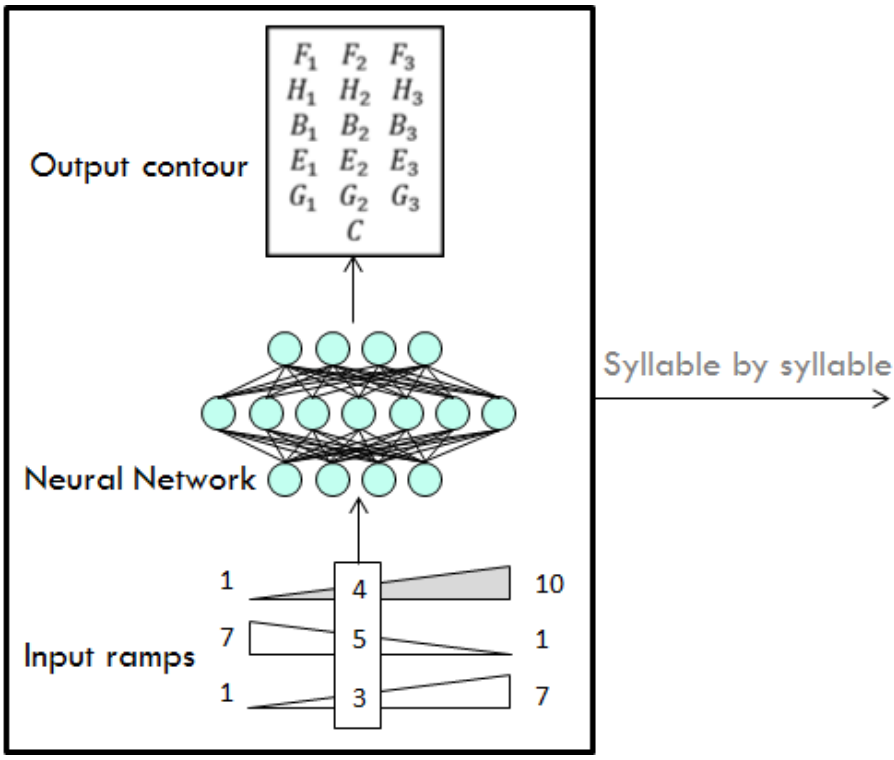

Figure 5.8: Prediction with the model-based approach. For a sentence with 7 syllables we observe the input ramps for a given position within the syllable and the output stylized contour: F_0,H (head), B (brows-area), E (eyes-area), G (gaze-area) and C (rhythm). The prediction is done syllable by syllable.

The following figures illustrate families of contours consisting in the predicted prototypical contours (blue contours), along with all exemplars (green contours) and the average exemplar contour per syllable-length (red contours). The figures show how the predicted contours capture the mean profiles of the original contours. The smoother predictions create families of contours which present overall behavior coherence. The contours are shown only for utterances of 3, 5, 7, 9 and 11 syllables in length (but the SFC is able to predict the prosodic contour of any syllabic length even if data for that length is missing or underrepresented).

[1] http://www.ra.cs.uni-tuebingen.de/SNNS/

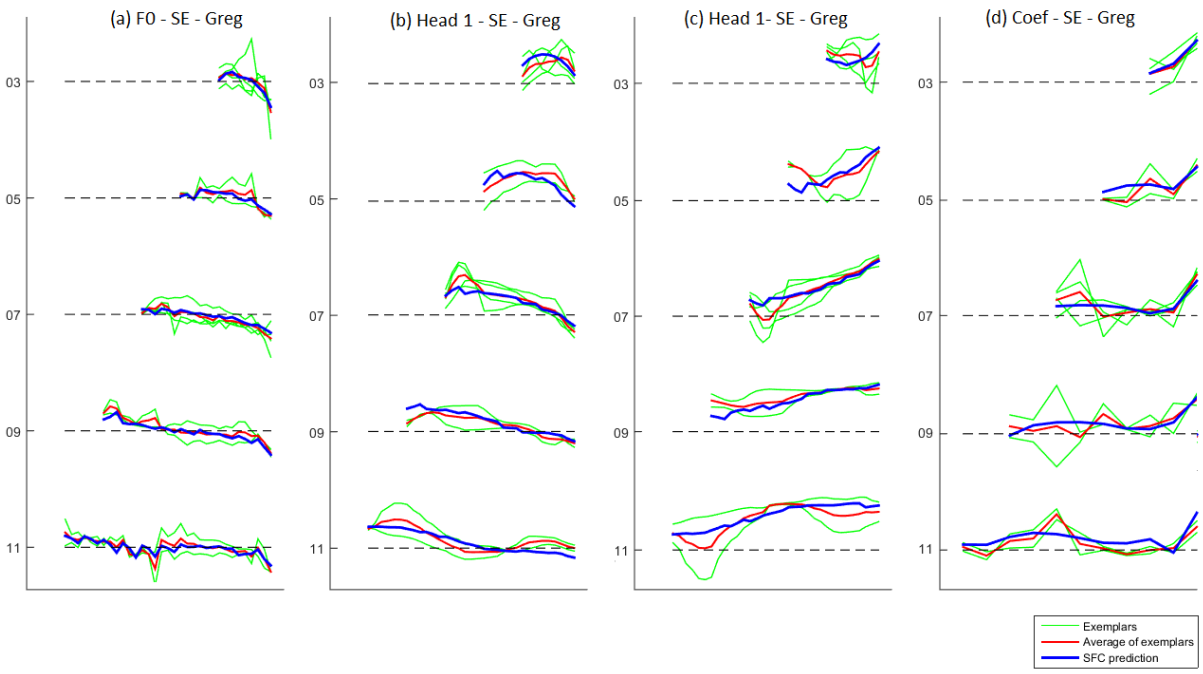

Figure 5.9: Example of families of contours 1. The figures represent: (a) f_0, (b) head 1, (c) head 2 and (d) rhythm parameter for attitude Seductive for Greg. We use the following color coding: the green contours are exemplars, the red contour is the average of exemplars and the blue contour is the SFC prediction.

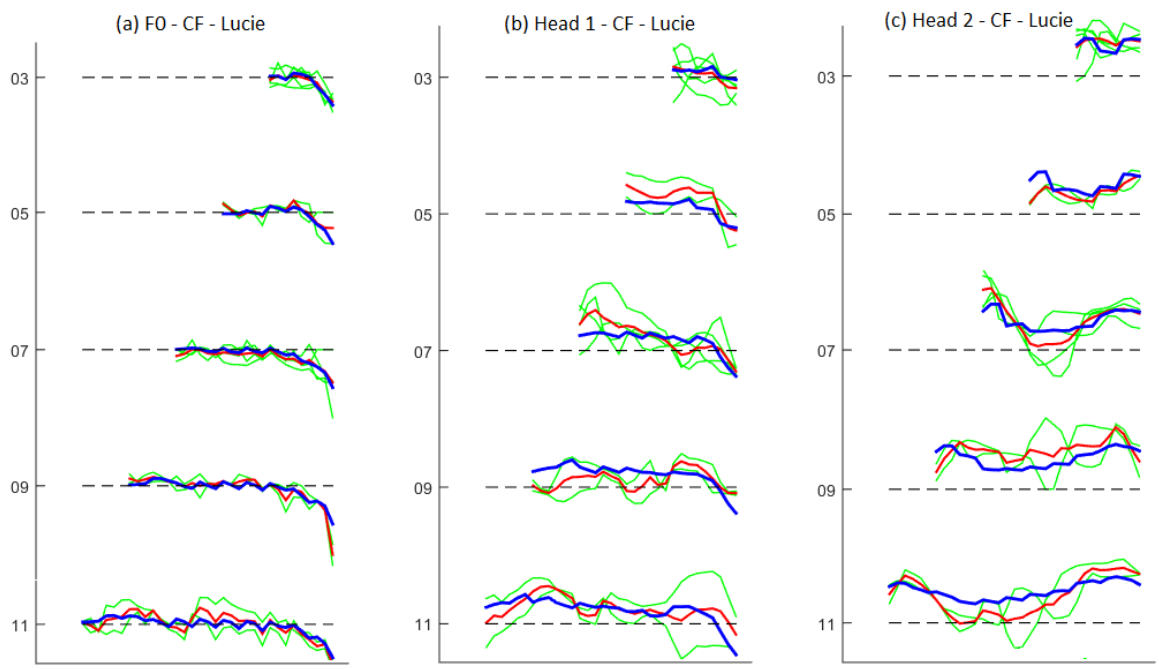

Figure 5.10: Example of families of contours 2. The figures represent: (a) f_0, (b) head 1, (c) head 2 for attitude Comforting for Lucie. The contours are colored as in figure 5.9.

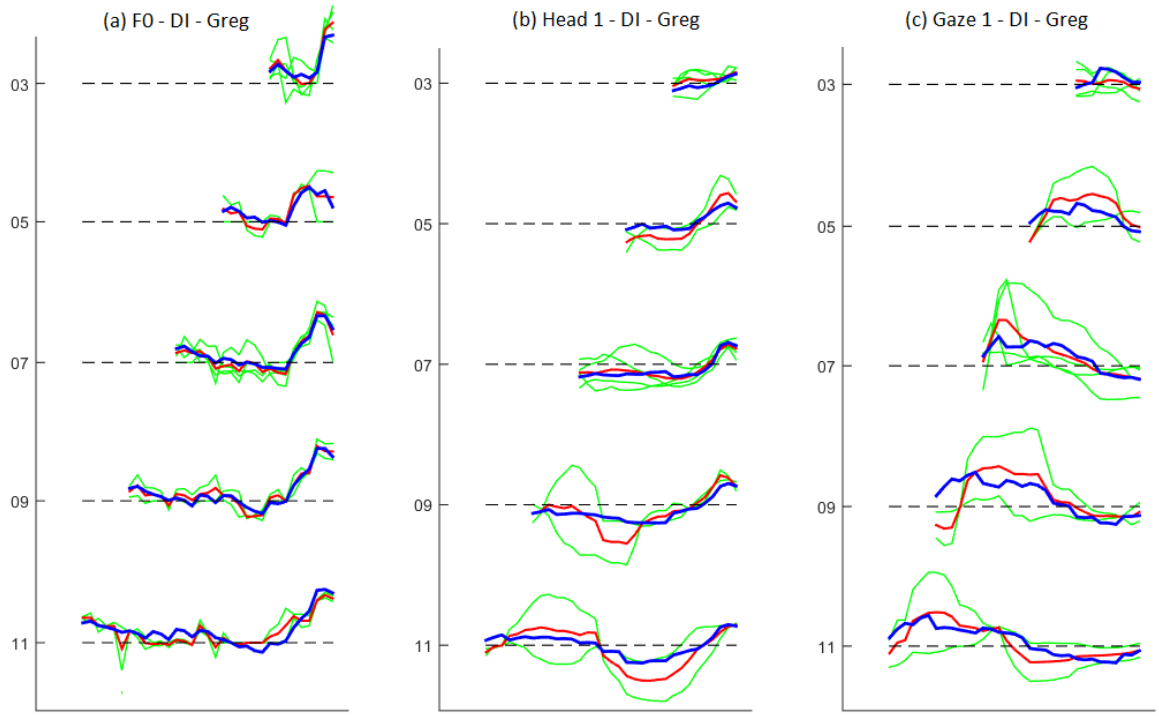

Figure 5.11: Example of families of contours 3. The figures represent: (a) f_0, (b) head 1, (c) Gaze 1 for attitude Doubtful for Greg. The contours are colored as in figure 5.9.

5.3.2 Blinking prediction

As mentioned in section 4.5.1, blinking presents an irregular behavior. SFC predictions cannot directly incorporate the variations present at different positions within the sentence. One solution to this problem would be to compute a local blinking rate and incorporate this as an additional parameter in the SFC model.

Given the overall quantity of observed blinks, we preferred to create a separate model for blinks using an attitude-specific Gaussian distribution of the blinking rate. Then we iteratively sample this distribution starting from the beginning of the sentence.

The following figure presents predicted contours using the normal distribution model and the original contour over concatenated sentences for a given attitude:

5.3.3 Prediction of segmental features

The expressive segmental features are predicted in the same manner as in the exemplar-based approach (see section 5.2.2). Both the predicted segmental features and predicted blinking are further stretched according to the rhythm predicted by SFC.

Figure 5.12: Original and predicted blinks for Greg and Lucie for concatenated sentences for attitude Confronted. The mean and standard deviation for blinking period for the attitude Confronted are: 3.32 ± 1.02 and 5.13 ± 1.81 for Greg and Lucie respectively.

5.4 Audiovisual synthesis

The audiovisual synthesis module involves three steps: prosodic contour reconstruction, segmental feature alignment and the actual audiovisual performance synthesis, which involves a vocoder for the audio signal and the proposed animation platform for the visual output (see figure 5.13).

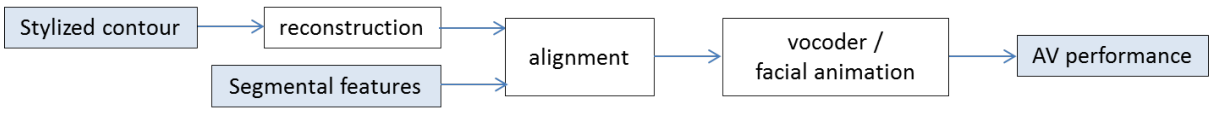

Figure 5.13: Outline of audiovisual performance synthesis. Note that in the case of approaches using a stylized contour (exemplar-based and model-based), a prosodic feature reconstruction step is required.

The first step in synthesizing the audiovisual performances is computing the speech rate. This is done differently according to how the local speech rate was generated. In the case of the frame-based approach, the local speech rate is computed at **frame level** and is represented by the predicted slope parameters. In the case of the exemplar- and model-based approaches, the local speech rate is computed at **syllable level** and is represented by the predicted rhythm coefficient C. While the slope parameters allow the direct computation of the alignment path for segmental features, in the case of stylized prosodic contours, the phoneme durations must be computed first. For this reason, we first discuss phoneme duration and prosodic contour reconstruction for the exemplar- and model-based approaches.

5.4.1 Reconstruction of stylized prosodic contours

The stylized prosodic contours are represented by 1 value per syllable for the rhythm components and 3 values per syllable by the melody and motion components. The following sections only refer to the approaches that use stylized contours: exemplar-based and model-based.

5.4.1.1 Reconstruction of phoneme durations

The stylized contour of rhythm is represented by one elongation coefficient C per syllable. During synthesis, the value of C is first used in computing the syllable durations. Then, each syllable duration is distributed among its segmental constituents: phonemes and an optional pause. The way of obtaining these durations is to suppose that phoneme elasticity is proportional to the standard deviation of its duration (measured in an independent corpus) and that within one syllable all the phonemes are compressed or elongated in the same way [Campbell, 1992]. Then we can compute the z-score k in a straightforward manner:

$$\Delta = \sum_i d_{p_i} = \sum_i e^{\mu_i + k\sigma_i} \tag{5.4.1}$$

where μ_i and σ_i are the average and standard deviation of the distribution of the logarythm of phoneme durations d_{p_i}. The average and standard deviations of phoneme durations are computed from an independent corpus, as well as r and D which are fixed to $r = 0.6$ and $D = 190ms$ (see section 4.1.3). We used a longer, independent corpus containing a female and a male native French speakers to obtain reliable phoneme duration statistics.

Pauses appear as an emergence phenomena: they are supposed to result from an excessive lengthening of the syllable. The previously computed k is virtual: if this value is greater than a fixed threshold, then a pause will be inserted at the end of the syllable. According to Barbosa [Barbosa, 1994], the real z-score, which accounts for the duration of a pause can be computed as:

$$k_r(k_v) = \alpha \cdot \arctan(k_v/\alpha) + \beta \tag{5.4.2}$$

where α and β are fixed parameters, computed from the reference corpus. For a detailed description of the algorithm refer to [Barbosa and Bailly, 1997]. Note that α and β are speaker-specific.

Given the syllabic and phoneme durations, the predicted stylized contours can be recon-

structed into full contours and reconverted to the required domain of usage (pitch must be reconverted to Hz, PCA components to blendshape values and rotations etc). We will discuss each parameter in the following sections.

5.4.1.2 Reconstruction of f_0

The f_0 contours are reconstructed using cubic interpolation between all keypoint values and converted to Hz. Given the duration of each constituent phoneme of the syllable, the keypoints are aligned according to this timeline (such that each set of 3 keypoints is placed at 20, 50 and 80 % of the duration of the vocalic nucleus). The full contour is interpolated from these aligned keypoints. The portions of the contour corresponding to unvoiced sounds and silences are then set to 0 Hz. Figure 5.14 illustrates a reconstructed f_0 contour from the SFC prediction:

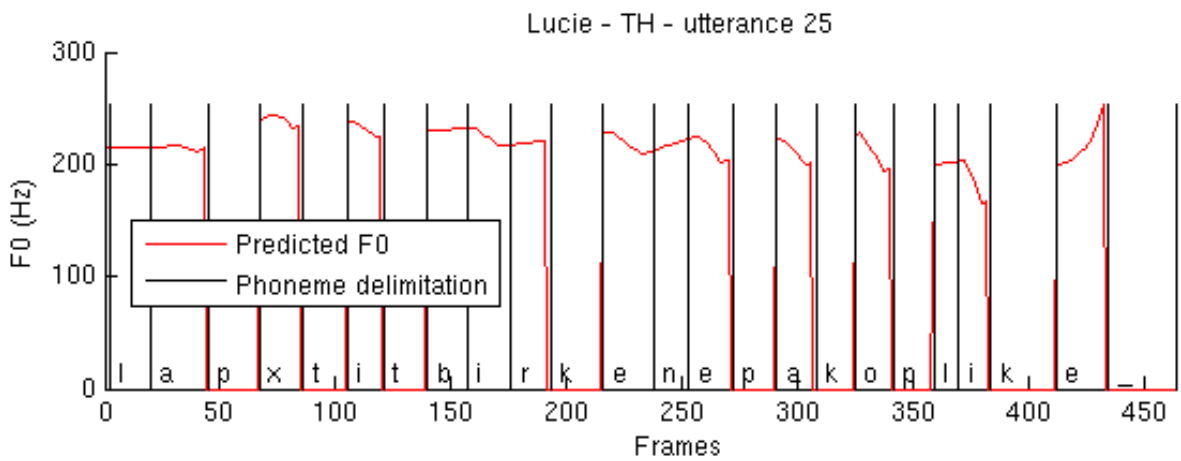

Figure 5.14: Reconstructed F0 contour. The stylized contour is predicted using the model-based approach for the attitude Thinking, for a sentence of 10 syllables. The contour is aligned according to the phoneme durations computed from the stylized rhythm contour.

5.4.1.3 Reconstruction of upper-face motion

All visual parameters are reconstructed similarly using cubic spline interpolation over all stylization keypoints, including those extracted from the virtual syllables. Similarly to the reconstruction method for f_0 contours, we first align the keypoints according to the predicted syllable durations (such that each set of 3 keypoints is placed at 20, 50 and 80 % of the duration of the syllable). then, the full contour is interpolated from these aligned keypoints.

The following figures present several examples of contours: original, stylized and reconstructed from stylized. Vertical lines represent syllable delimitation where the marginal

syllables are virtual and the rest constitute the spoken utterance. A stylized contour is represented in each figure by the sequence of 3 values per syllable and the reconstruction by cubic interpolation follows the original contour accordingly. Notice the virtual syllables which are notated in the phonetic transposition as "E", considered as a syllable with only one vocalic nucleus and with constant duration of 250 ms.

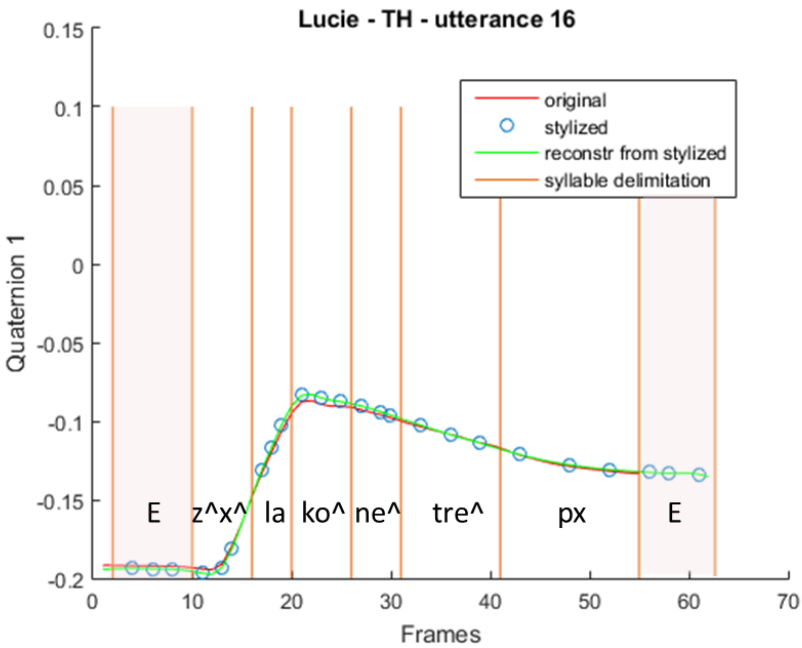

Figure 5.15: Reconstructed head rotation contour. The original, stylized and reconstructed from stylized contours for head rotation are superposed. They represent the quaternion i contour for head rotation for attitude Thinking, phrase 16: "Je la connais très peu". The virtual syllables are transcribed as "E".

The second step is the realignment of the GMM predicted segmental features to the predicted speech rate.

5.4.2 Alignment of segmental features

Frame-based approach. The slope GMMs predict expressive spectra, visual parameters and slope parameters. As in the case of Dynamic Time Warping, the string of slope parameters indicate a warping path for the spectra and visual parameters, thus dynamically changing their sampling frequency.

Exemplar-based and model-based approach. Alignment of segmental features is done by resampling the trajectories at phoneme level. For each phoneme, the neutral trajectories are linearly stretched so that they match the predicted duration of the phoneme. Mismatches in the phoneme sets may occur if new silences are encountered in the predicted phonemic string. In this case, the contours values at the previous frame are simply

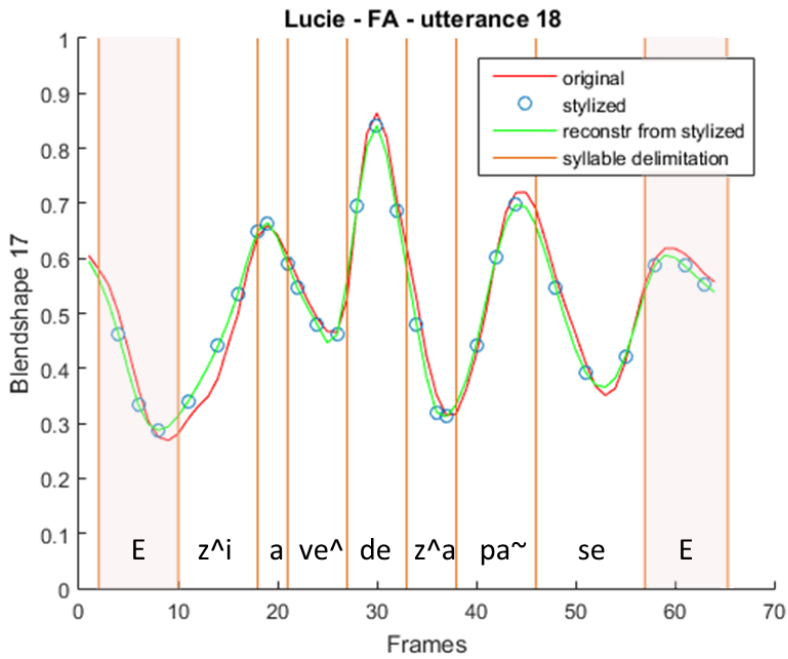

Figure 5.16: Reconstructed eyebrow contour. The original, stylized and reconstructed from stylized contours for eyebrow motion are superposed. They represent the blendshape 17 (Brows Up Center) contour for attitude Fascinated, phrase 18: "J'y avait déjà pensé". The contours are notated as in 5.15.

repeated for the duration of the silence.

5.4.3 Performance synthesis

With the aligned audio and visual parameters, a new performance can be synthesized. This is done separately for voice and the visual features.

Voice is synthesized using the STRAIGHT vocoder which receives: the cepstral features with the time-aligned expressive mel-cepstra and aperiodicities that were generated using a GMM and the reconstructed f_0 contour. As visual features are reconstructed, they are used to control a blendshape model implemented in Blender. The keyframes containing poses are rendered, thus creating an *animation*.

The following figures show rendered examples of the synthesized audiovisual speech in cartoon style. Each figure presents rendered frames from the synthesized speech using the methods described. The four rows correspond to: original reconstructed performance, frame-based prediction, exemplar-based prediction and model-based prediction respectively.

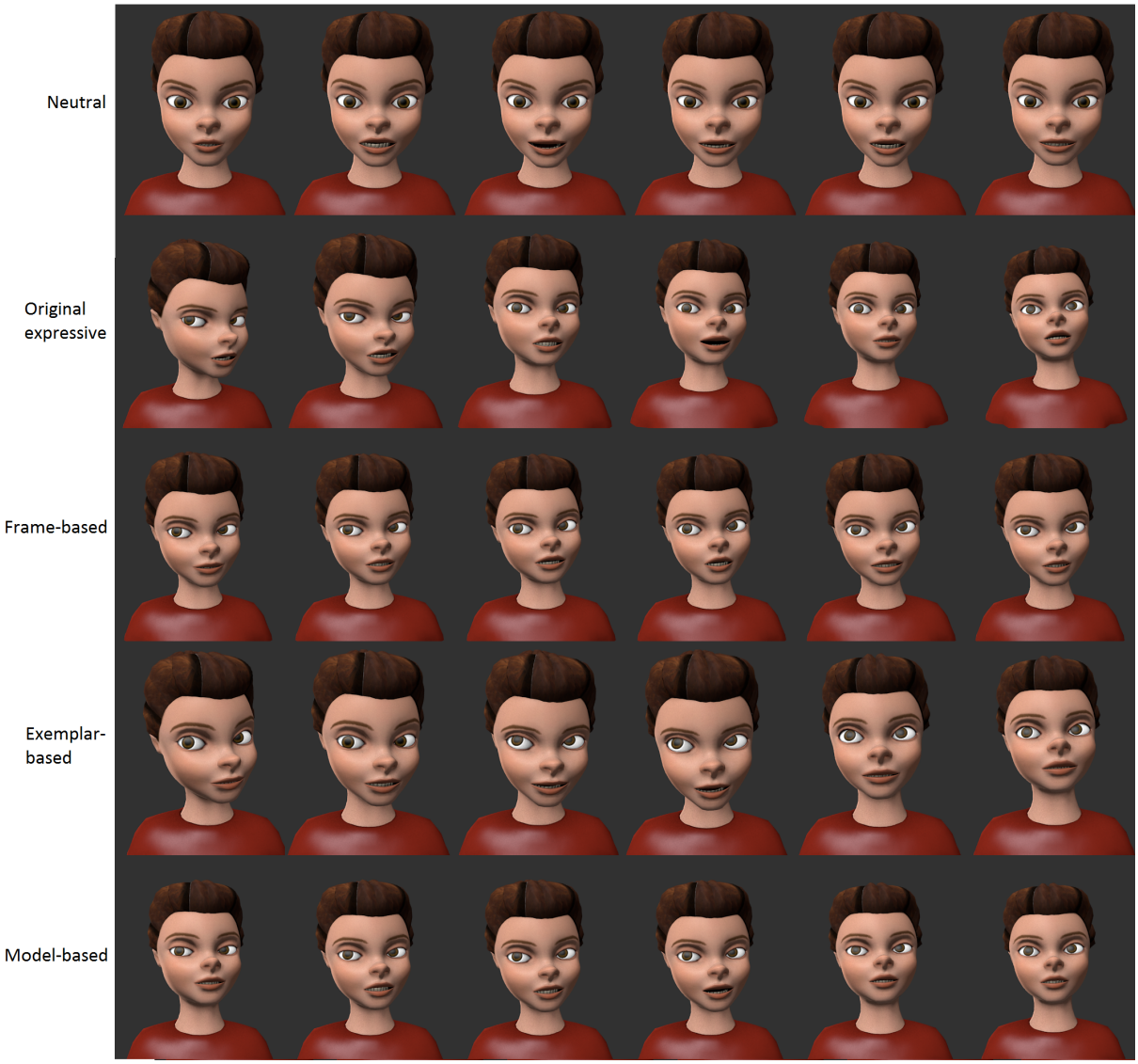

Figure 5.17: Example frames for the proposed generation techniques. Sentence: "Je dirais même énigmatique" with the attitude Thinking. The four rows correspond to: neutral performance (source), original expressive performance (target), frame-based prediction, exemplar-based prediction and model-based prediction respectively. The poses for each row correspond to the onsets of the last 6 vowels. Note the evolution of head rotation for original expressive, exemplar-based and model-based.

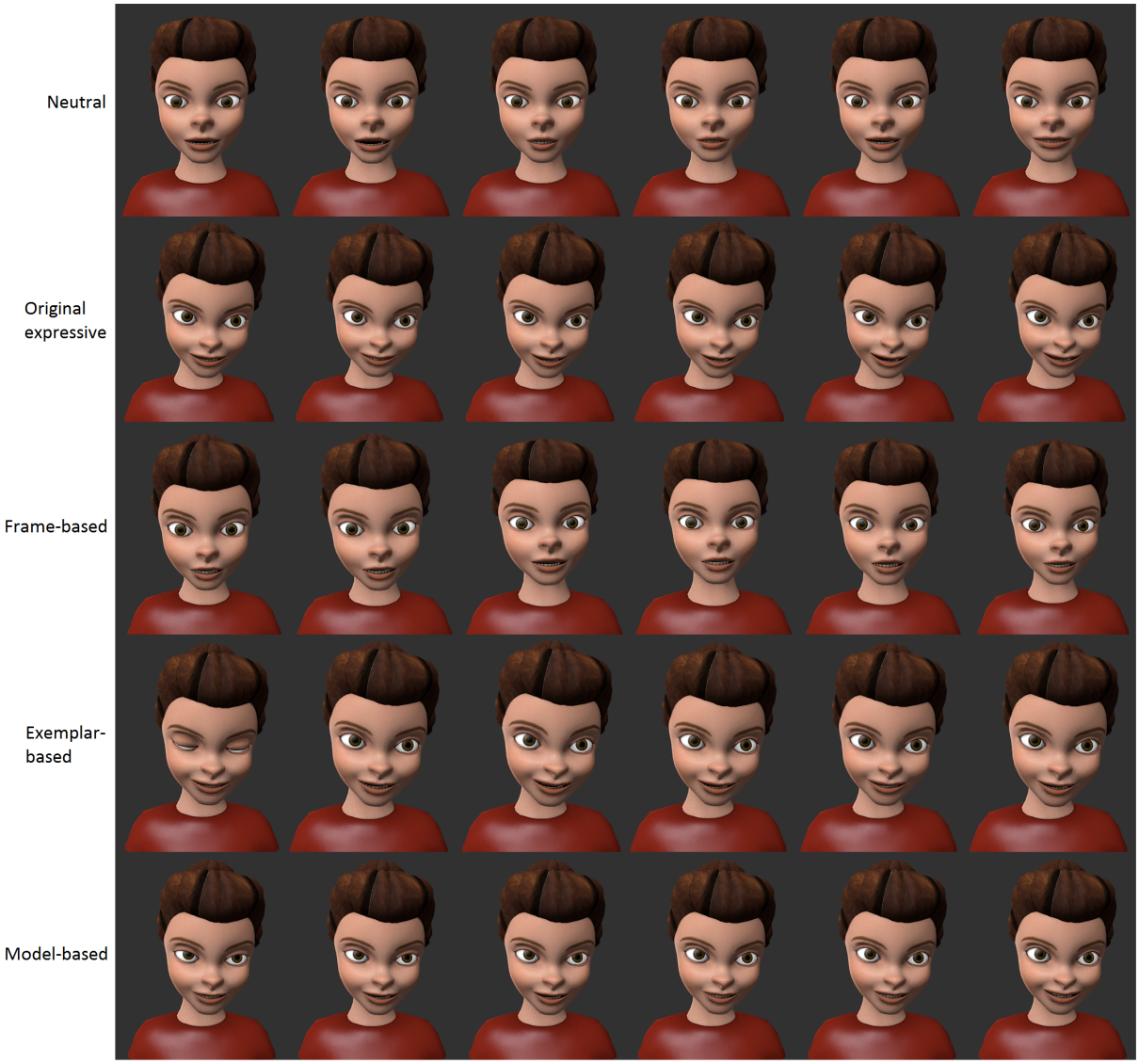

Figure 5.18: Example frames for the proposed generation techniques. Sentence:"Votre mêre m'a authorisé" with the attitude Embarrassed. The four rows correspond to: original performance, frame-based prediction, exemplar-based prediction and model-based prediction respectively. The poses for each row correspond to the onsets of the last 6 vowels. Note how the head movements are correctly generated by the exemplar-based and model-based methods while the frame-based method retrieves poses which are more similar to the neutral version. Blinking occurs in the exemplar-based result but not in the original expressive version.

Figure 5.19 illustrates results for the model-based method obtained for one sentence using the realistic animation style:

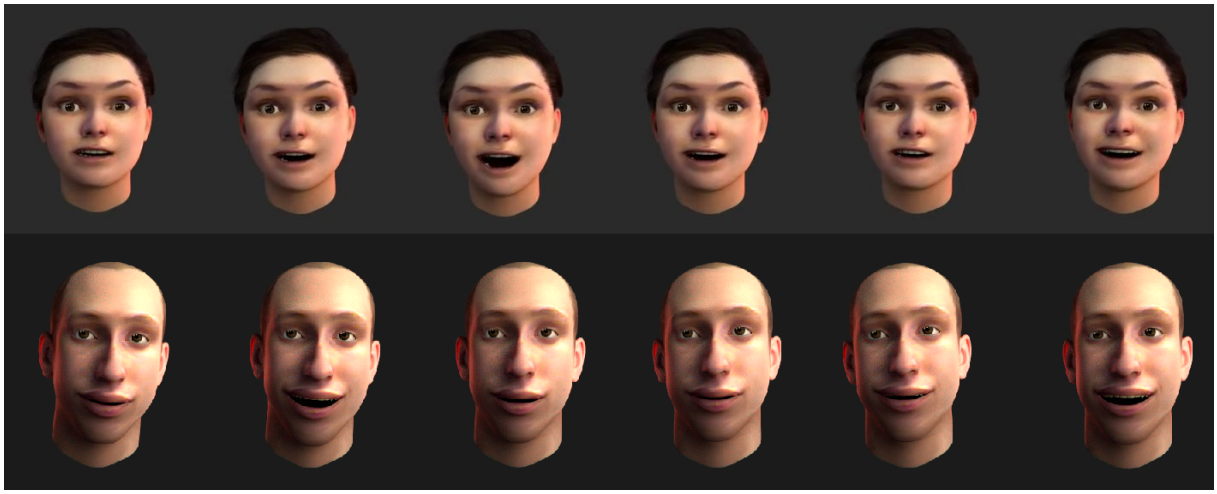

Figure 5.19: Example frames for model-based prediction. The two actors perform the sentence: "Votre mêre m'a authorisé" with the attitude Seductive. The poses correspond to the onsets of the first 6 vowels. Note how the two actors present different head motion behaviors.

5.5 Evaluation

The performance generation methods are evaluated using objective and subjective tests. The following list presents notations that will be used for evaluation purposes. They describe the type of data used in the audiovisual performance synthesis. Given an attitude and a sentence, the versions proposed are:

- *original reconstructed*: animated representation using the resynthesized voice from original spectra and pitch, and then animation from the reconstructed original stylized prosodic contours and original lower-face motion
- *frame-based*: animated representation using the synthesis as described for the frame-based generation
- *exemplar-based*: animated representation using the synthesis as described for exemplar-based generation (random exemplar as stylized contour)
- *model-based*: animated representation using the synthesis as described for the model-based generation (SFC prediction as stylized contour)

Since the generation methods use the stylized contours as original training data we consider the *original reconstructed* data as ground truth.

5.5.1 Objective test

Our approach is evaluated using a set of objective measures. From our dataset of 35 utterances, we use 7 different utterances for testing, containing 3, 4, 5, 7 and 9 syllables respectively. We use leave-one-out GMM training in order to generate frame-based and segmental feature predictions for the 7 test utterances.

Considering the original reconstruction version as ground truth we compute the errors obtained for each proposed generation method using the RMSE criteria:

$$RMSE_y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{5.5.1}$$

The values are computed for all stylization keypoints of the testing utterances. We compute the distances between ground truth data and the three other types of data. We focus on a set of attitudes that is also used in a perceptual test in the following section.

For the selected utterances and attitudes, we synthesize all combinations of performances which are obtained using the four techniques mentioned: *original reconstructed*, *frame-based*, *exemplar-based* and *model-based*. Head rotation and gaze errors are computed directly as euler angle RMSE ($\circ$).

For facial expressions, we setup three distinct sets of 3 virtual markers and extract their coordinates from aligned performances. The sets include markers selected from the eye area and the brow area, respectively. We can therefore compute errors using the RMSE criteria.

We focus on the visual prosodic features. The errors obtained on 3D virtual marker displacements are separately on computed on the eye and brow area and presented in tabls 5.1. Head and gaze rotation errors are presented in 5.2.

The model-based approach generally presents the smallest error-rate. Also we observe that the attitudes exhibiting the smallest error rates are Doubtful and Thinking. The eye area presents bigger errors than the brows area because of the natural occurrence of eyeblinks. These are unique events which cause high virtual marker displacements. We note that the largest eye area error is present

However, these measures only provide information related to the similarity to a given performance. This does not represent a measure of expressiveness. In the context of generating audiovisual performances, a perceptual test is more suitable to evaluate our results.

Table 5.1: Objective results for the proposed generation approaches. For each attitude and feature type, we compute RMSE distances between ground truth data (original reconstructed) and: (1) frame-based, (2) exemplar-based, (3) model-based for: eye-area and brow-area facial expressions respectively.

	Eye-area expressions(cm)			**Brow-area expressions(cm)**		
	(1)	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
QS	1.565	1.646	**1.239**	0.505	0.672	**0.374**
CF	0.912	1.678	**0.568**	0.378	0.415	**0.164**
SE	0.646	1.179	**0.422**	0.505	0.264	**0.216**
TH	0.707	0.977	**0.337**	0.458	0.497	**0.215**
DI	0.409	0.800	**0.386**	**0.120**	0.638	0.223
IR	0.793	0.786	**0.484**	0.407	0.445	**0.350**
EM	**0.292**	0.803	0.334	0.332	0.397	**0.163**

Table 5.2: Objective results for the proposed generation approaches. For each attitude and feature type, we compute RMSE distances between ground truth data (original reconstruction) and: (1) frame-based, (2) exemplar-based, (3) model-based for: head and gaze rotation respectively.

	Head rotation (∘)			**Gaze (∘)**		
	(1)	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
QS	0.878	1.085	**0.741**	**1.015**	1.436	1.051
CF	1.276	**0.982**	1.216	1.101	2.236	**0.876**
SE	**1.468**	1.773	1.775	1.852	1.692	**0.948**
TH	0.659	0.788	**0.577**	1.158	1.976	**1.118**
DI	0.699	1.736	**0.571**	1.970	1.687	**0.481**
IR	1.142	**0.868**	1.123	1.915	1.747	**1.744**
EM	2.652	**1.587**	2.255	2.035	9.132	**1.811**

5.5.2 Subjective test

Evaluation of the proposed conversion methods was carried using a *ranking test* paradigm similar to [Bailly and Gorisch, 2006]. The animated stimuli is obtained using the following methods: *frame-based*, *exemplar-based*, *model-based* and *original reconstructed*. The test was carried on a crowdsourced online platform and can be found at [2].

Participants to the test are asked to rate 4 animations per trial and then to specify the level of their expressiveness relative to an indicated attitude. Instead of choosing from a limited list of possible scores to describe the perceived expressiveness of one animation, the participant is able to retrieve relative and absolute perceptual information by placing symbolic icons in a ranking rectangular-shaped grid. The horizontal sides of the grid effectively represent quality ratings, from $Very\ bad$ to $Very\ good$. The vertical axis has no dimension and just eases the layout of icons by avoiding messy superpositions. Subjects can play animations on demand by clicking on the four icons representing the tested systems and can then move the icons anywhere within the grid, considering that verticals represent identical quality options. Ranking an animation is equivalent to associating it with a numerical value which ranges continuously from 1 to 5 such that $Very\ bad = 1$, $Average = 3$ and $Very\ good = 5$ (see figure 5.20).

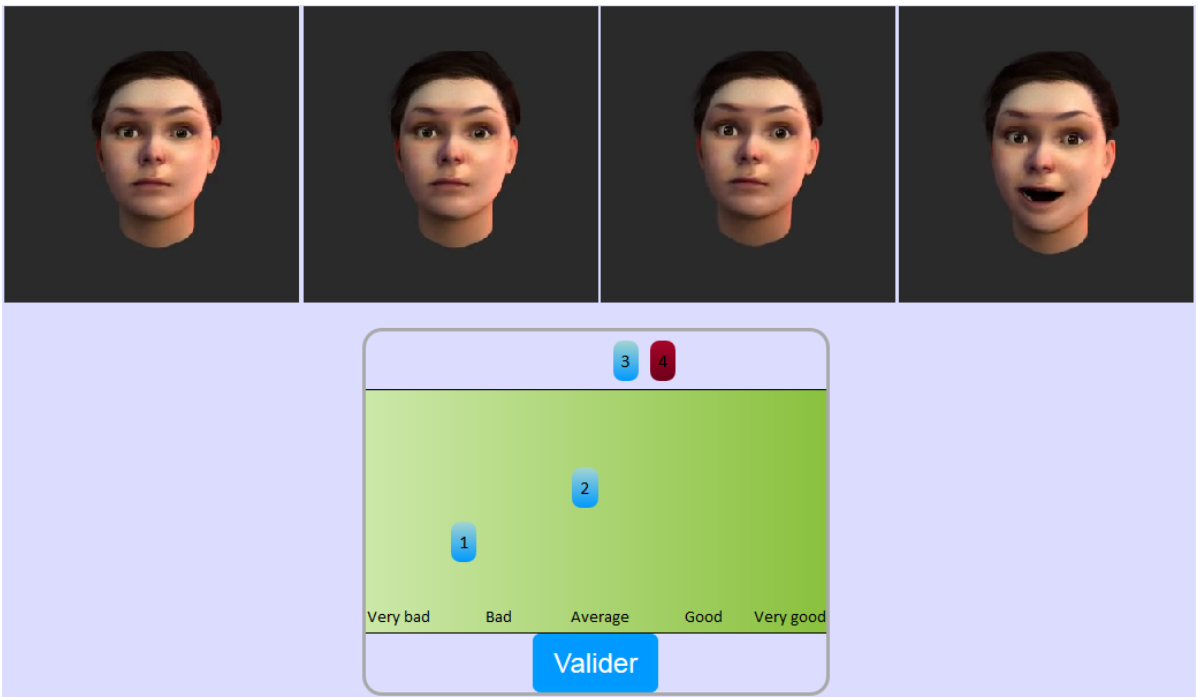

Figure 5.20: Snapshot of the ranking area. The 4 animation stimuli are placed above the ranking grid and symbolic icons 1 to 4 are randomly associated performances obtained using the 4 methods. In this snapshot, the first two animations are ranked, the last two are not ranked and the 4th is being played.

[2] http://www.gipsa-lab.fr/~adela.barbulescu/test2/

As this test requires a longer time for completing one validation, only a limited subset of attitudes is chosen. The choice is based both on the results of the previous test and the level of discrimination existing in the attitude subset. The attitudes chosen are: Question, Comforting, Seductive, Thinking, Doubtful, Ironic and Embarrassed. Each of 7 trials presents 4 animated versions for a given attitude of a random sentence (out of the 7 test sentences). The associations between icons and versions is randomized within each trial. We performed a leave-one-out GMM training per attitude for audio and visual parameters separately and 7 test sentences. The SFC predictions do not include virtual syllables, nor the blinking generation model.

A total of 41 native French subjects participated in the experiment and each evaluated the 28 animations. The statistical significance was assessed using an ANOVA test which considers the X *position* of the ranking as a continuous variable and *version* as factor. The main effect of the factor *version* was significant ($p < 0.005$), thus allowing for further analysis of the results obtained for each version. The average results per version are shown in Figure 5.21.

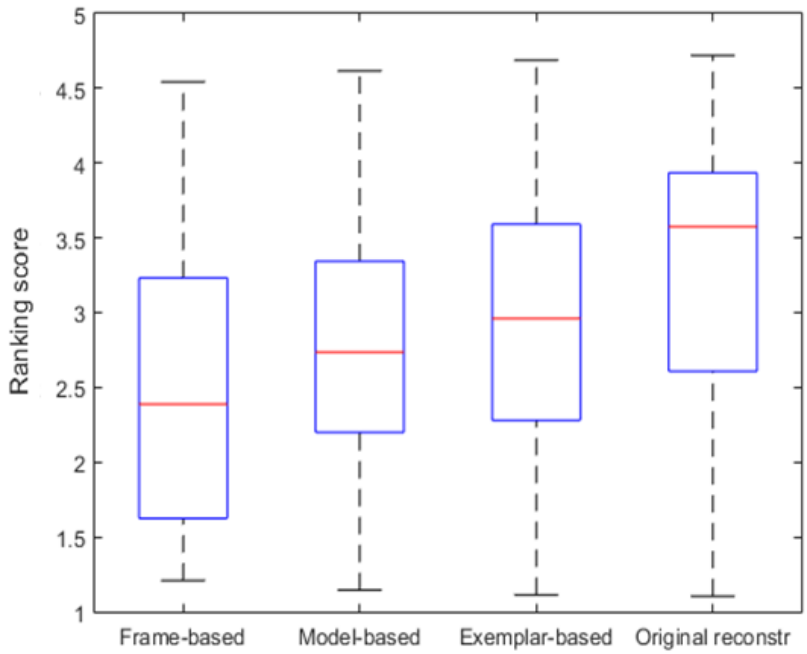

Figure 5.21: Results of the perceptual ranking test. Bars represent the rank value per version averaged across subjects and attitudes.

All version averages are situated between *Bad* and *Good*. As expected, the best results were obtained for the original reconstructed version with an average and standard deviation of 3.35 ± 0.22, while the worst were obtained for the model-based approach with 2.5 ± 0.38. A paired t-test for the model-based (with 2.76 ± 0.28) vs. exemplar-based (with 2.86 ± 0.26) versions showed that there is no significant statistical difference between the observations. This can be explained by the intrinsic quality of our speech

resynthesis system. Table 5.3 presents the average values obtained for each attitude and version used in this test:

Table 5.3: Average ranking values of the 4 methods per attitude.

	Frame-based	Model-based	Exemplar-based	Original reconstr
Question	1.87	3.30	3.11	3.59
Comforting	3.02	2.62	2.45	3.04
Seductive	2.84	2.59	2.54	3.34
Thinking	2.28	2.96	3.11	3.11
Doubtful	2.69	2.67	2.87	3.29
Ironic	2.35	2.46	3.04	3.58
Embarrassed	2.45	2.74	2.91	3.51

The ranking test results show that on average the frame-based method is outperformed by the model-based and exemplar-based methods while the original reconstructed method presents the highest test scores. The lowest scores obtained for the frame-based version are observed for attitudes in which f_0 and speech rhythm have a big impact on perception: Question, Thinking, Ironic. According to comments retrieved by subjects, a few animations presented unnatural audio which lead to a lower ranking. These samples can be both attributed to the model-based or exemplar-based methods, due to the phoneme duration generation algorithm, thus explaining the bigger scores obtained by the frame-based method in the case of Comforting and Seductive.

All version averages are situated between *Bad* and *Good*. The general lower scores obtained in the case of Comforting and Doubtful can also be explained by the fact that these attitudes are more difficult to be recognized as shown by the previous assessment results (see section 3.7).

5.6 Limitations

The proposed methods for generating expressive audiovisual speech present specific limitations. For example, at the moment, GMM training is time consuming. Since we have a small database, each sentence is trained in a leave-one-out paradigm and thus each test sentence requires a new training session. A larger database would allow the distinct separation of the database into train and test data. Although we have a neutral dialog at our disposal for testing purposes, it is not available in all expressive versions. The

tests carried until this point required the comparison to an existing expressive version of the studied text.

The exemplar-based method uses random selection of exemplars. Optimal selection of exemplars could be implemented if performance suitability is defined as a cost. One idea is to take into account the similarity of the phonological structure of the analyzed sentences.

The model-based method uses the SFC model which captures coherent mean prosodic profiles but does not propose for the moment ways to model the structure of the variability around these mean profiles. For example, blinking presents an incoherent behavior at the level of attitude, and these variations are smoothed leading to the prediction of very few blinks. In this case, a separate model is considered. However, at the rim of undertaking the perceptual ranking test, we did not use a blinking model, nor the information extracted from virtual syllables.

The general low recognition rates for the comparative perceptual tests, including for ground truth data, show that the perceptual tests need improvements; one such direction would be to design pre-test training with a series of original animations such that participants get accustomed to the animated actors or rating several sentences at once for a given attitude. Also the separation of audio/visual modalities would allow for a clear evaluation of the contributions brought by each of these modalities. Lastly, perceptual evaluations may benefit from the usage of cartoon style animations, whose appearance influence the perception of expressiveness.

5.7 Summary

This chapter presented our approaches for the generation of audiovisual prosody from neutral stimuli and the evaluation of the proposed techniques. Section 5.1 presented the *frame-based approach*, which is a variation on the conventional method for speech conversion [Stylianou et al., 1995], to which we propose a novel prosodic feature (slope parameter) for the computation of speech rate. The slope parameters are predicted using GMM regression along with the spectrum and PCA motion components.

The two other methods are based on the idea of combining separate generation approaches for the segmental and suprasegmental features. The *exemplar-based*, presented in Section 5.2 method selects random exemplars (original stylized prosodic contours) to generate prosody and the conventional GMM regression to predict segmental features.

The *model-based* approach, presented in Section refgen3 consists in extending the Superposition of Functional Contours model [Bailly and Holm, 2005] in order to generate audiovisual prototypical contours. Our contribution consists in training attitude-specific

SFC models with a motion component. This component also includes virtual syllables to account for pre- and post-phonatory movements. A separate Gaussian model is used for blinking generation. The segmental features are predicted using the conventional GMM regression.

Section 5.4 described how the predicted prosodic contours are reconstructed and aligned with predicted segmental features in order to synthesize new audiovisual performances. The audiovisual prosody generation methods are evaluated using objective measures and a perceptual ranking test, described in Section 5.5.

Experimental results confirm that methods based on suprasegmental features outperform the frame-based conversion method in subjective tests. The exemplar-based and the model-based approaches present no significant statistical differences. Experimental results also demonstrate that the task of recognizing attitudes in isolated sentences is difficult, with recognition rates ranging between *Bad* and *Good* even for ground truth sentences.

Our best results are obtained with the exemplar-based method which demonstrates that stylized prosodic contours can be interchanged between same-sized sentences. Future work is needed to further improve the model-based approach and to design better subjective evaluation experiments allowing to measure progress in the difficult task of generating recognizable dramatic attitudes.

Chapter 6

Perspectives

This chapter presents ideas for improvements of our generation and evaluation methods. The following sections describe future directions of the presented work.

6.1 Dialog generation

The first future direction we envision is the application of the presented techniques to generate a dialogued play. Given a neutral performance we can generate all the expressive versions corresponding to the 16 attitudes. A system for dialog generation would require an algorithm to combine all the predicted lines from the dialog in a desired order. If both virtual actors are present in the camera viewpoint at all times, a model for backchannel generation would also be needed.

We presented a first trial for the generation of expressive dialogues in Experimenta 2015 [1]. We proposed an application in which participants can direct a dialogued scene by choosing the attitude to be performed in each line by our two virtual actors. The performances were obtained by combining the original expressive voice with the SFC predictions for movements. More than 1000 participants tested the application thus providing valuable ad-hoc feedback.

The application can be further used as an experimental tool for the attitude recognition tests because it allows a more contextual perceptual evaluation of the attitude. We believe that the evaluation process becomes easier for the participant if more performances are played instead of isolated ones. Moreover, the idea of using dialogues for testing is suitable for carrying a gating experiment similar to the one proposed by [Morlec et al., 2001]. The application is available online [2].

[1] http://www.atelier-arts-sciences.eu/EXPERIMENTA-2015-208
[2] http://gipsa-lab.fr/∼adela.barbulescu/exp/

6.2 Large vocabularies

More exemplars. On the one side, we could focus on the best recognized 8 attitudes and record another set of sentences with the goal of increasing the number of exemplars per sentence-length. This would allow a larger pool of data from which we can develop an algorithm for best exemplar selection and outlier removal according to associated costs for each exemplar. Exemplar selection could be optimized by introducing a phonological similarity cost. A similar approach is used by Taylor et al [Taylor and Black, 1999] who synthesize speech with a unit selection based on phonological tree matching.

A larger database would also allow a deeper investigation of the contours that display multiclass signatures per attitude or head rotation symmetry related issues. Another subject we could look into is the symmetry of facial expressions. During pre-processing, for example, we make symmetric the eye opening and eye blinking blendshapes. More investigation should be carried in the effect of facial asymmetries on attitudes as previous work has demonstrated that these play an important role in expressiveness [Ahn et al., 2013].

With a sufficient coverage of phonetic context (see [Le Maguer, 2013]), we could use a TTS system to directly generate audiovisual performances from text. This would avoid the need of input utterances interpreted in neutral style.

More attitudes. On the other side, an interesting future direction is that of exploring a large amount of attitudinal functions for various languages. Our initial inspiration, the Mind reading taxonomy [Baron-Cohen, 2003] contains a total of 412 expressive labels. Our recording pipeline could be used to record a large number of acted or spontaneous attitudes for a few sentences. The existence of such a database would also represent a huge benefit for other researchers in the domain.

6.3 Learning by imitation

Another direction for our system would be considering an incremental model for the generation of audiovisual speech. In the envisioned context, the director would be allowed to demonstrate the performance first. His performance would be used as input instead of neutral performance. In a first step, our system could be changed simply by allowing the frame-based function to map between the director's performance to actors' performances, for instance using a speaker conversion approach. Towards this goal, we implemented a realtime Max [3] application for voice and motion conversion during the eNTERFACE 2013 workshop [4].

[3] https://cycling74.com/
[4] http://eventos.fct.unl.pt/enterface13

In a second step, the director would be allowed to iteratively repeat his performance in order to modify the initial prediction. This approach is similar to learning by imitation or demonstration. An example of an incremental model for learning by imitation was proposed by Calinon et al [Calinon and Billard, 2007].

6.4 Full-body animation

In this thesis, we have chosen to focus exclusively on facial animation and head movements. It would be useful to explore how the proposed methods can be generalized to other aspects of body language, including hand gestures, which are an important part of expressive actor performances.

Previous work by Kipp [Kipp, 2003] has investigated this problem using expressive datasets drawn from television talk shows. In future work, it would be interesting to apply our methodology of choosing a smaller vocabulary of dramatic attitudes and analyzing how professional actors use their body language, together with facial expressions and intonation, to express them.

Chapter 7

Conclusion

7.1 Summary

This section presents a summary of the contributions and results presented in this thesis. As mentioned in the introduction chapter, this document revolved around the topic of generating audiovisual prosody, in the context of expressive conversion of audiovisual speech.

Chapter 3 presents the process of selection and recording of attitudes which was carried with the help of a theater director. Two semi-professional actors were selected and trained to record an expressive corpus of 16 attitudes. As opposed to existing affective databases, our expressive corpus contains 3D speech performances of a set of sentences acted with complex attitudes. The corpus is designed to contain variable-sized sentences with multiple samples per sentence length (number of syllables).

Chapter 4 describes prosody analysis techniques and validation tests for the recorded database. Analysis is carried on the stylized prosodic contours. We introduce the virtual syllables to account for pre- and post-phonatory movements and we show that these movements are attitude-specific. We also propose an objective measure for inter-class differences with respect to attitudes or actors for specific prosodic features. A part of this chapter is dedicated to comparing objective inter-class distances to perceptual results of attitude recognition tests. The objective measures and stylized contour descriptions show that contour shapes are discriminant and that most prosodic features show specific behaviors, especially at the beginning or end of speech.

Chapter 5 presents our proposed approaches for the generation of audiovisual prosody. These approaches are based on the idea of combining separate techniques for the generation of segmental and prosodic contours. These include a frame-based approach, an example-based approach and a model-based approach. Another important contribution

of this work is the extension of the SFC model to include a motion component. This allows us to test the hypothesis that attitudes are encoded via audiovisual prosodic signatures. The chapter presents experimental results illustrating how attitude-specific behaviors are replicated by the SFC models and can also be observed when replacing the prosodic contour of one sentence with a random exemplar.

As mentioned in the introductory chapter, this manuscript contains the details for the main contributions introduced during the thesis:

- an audiovisual dataset of dramatic attitudes is recorded by 3 speakers and validated using various objective and perceptual tests
- the data extracted from this dataset is used in the implementation of 3 methods for the generation of audiovisual prosody
- among the techniques approached, we highlight the separation of segmental and suprasegmental models and the extension of the SFC model to include a motion component

7.2 Publications

Part of the work carried during this thesis was presented in the following publications:

1. **Audio-Visual Speaker Conversion using Prosody Features**, *oral presentation*, Adela Barbulescu, Thomas Hueber, Gérard Bailly, Rémi Ronfard, proceedings AVSP 2013.
2. **Reactive Statistical Mapping: Towards the Sketching of Performative Control with Data**, Nicolas d'Alessandro, Joelle Tilmanne, Maria Astrinaki, Thomas Hueber, Rasmus Dall , Thierry Ravet, Alexis Moinet, Huseyin Cakmak, Onur Babacan, Adela Barbulescu, Valentin Parfait, Victor Huguenin, Emine Sumeyye Kalaycı and Qiong Hu, Springer, 2013.
3. **Beyond Basic Emotions: Expressive Virtual Actors with Social Attitudes**, *oral presentation*, Adela Barbulescu, Rémi Ronfard, Gérard Bailly, Georges Gagneré, Huseyin Cakmak, proceedings MIG 2014.
4. **Directing virtual actors by interaction and mutual imitation**, *oral presentation*, Adela Barbulescu, doctoral symposium IEEE VR 2015.
5. **Audiovisual Generation of Social Attitudes from Neutral Stimuli**, *oral presentation*, Adela Barbulescu, Gérard Bailly, Rémi Ronfard, Maël Pouget, proceedings FAAVSP 2015.

Appendix A

A.1 Gaussian Mixture Model conversion

Most voice conversion systems use statistical approaches to create the feature conversion function from experimental data. The state of the art performs spectral conversion using Gaussian mixture model and a similar approach is used in our paper for joint audio-video features.

The basic conversion approach implies modeling data using GMMs and building a conversion function as a weighted sum of local regression functions, thus providing a soft classification between mixture components [Stylianou et al., 1995] [Toda et al., 2007]. In the following example, the joint probability density of source and target feature data represented by parameter vectors $x_t = [x_t(1), x_t(2), ..., x_t(D_x),]^T$ and $y_t = [y_t(1), y_t(2), ..., y_t(D_y)]^T$ at frame t, are modeled by the GMM:

$$P(z_t|\lambda^{(z)}) = \sum_{m=1}^{M} \alpha_M \mathcal{N}(z_t; \mu_m^{(z)}, \Sigma_m^{(z)}), \tag{A.1.1}$$

where z_t is the joint vector $z_t = [x_t^T, y_t^T]^T$, m is the mixture component index corresponding to the weight α_m. $\mathcal{N}(\cdot; \mu, \Sigma)$ represents a normal distribution with mean μ and covariance Σ, defined as follows:

$$\mu_m^{(z)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \Sigma_m^{(z)} = \begin{bmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{bmatrix}, \tag{A.1.2}$$

Given a source vector x and m being the GMM mixture component index, the conditional probability density of the converted vector y is also modeled as a GMM with the mean and covariance:

$$E_{m,t}^{(y)} = \mu_m^{(y)} + \Sigma_m^{(yx)} {\Sigma_m^{(xx)}}^{-1} (x_t - \mu_m^{(x)}), \tag{A.1.3}$$

$$D_m^{(y)} = \Sigma_m^{(yy)} - \Sigma_m^{(yx)} {\Sigma_m^{(xx)}}^{-1} \Sigma_m^{(xy)}. \tag{A.1.4}$$

The minimum mean square estimation of the converted vector is:

$$\hat{y}_t = E[y_t|x_t] = \sum_{m=1}^{M} w_m E_{m,t}^{(y)}, \tag{A.1.5}$$

where the weight w_m represents posterior probability of x for the mth component:

$$w_m = P(m|x_t, \lambda^{(z)}) = \frac{\alpha_m \mathcal{N}(x_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(x_t; \mu_x^{(x)}, \Sigma_n^{(xx)})}, \tag{A.1.6}$$

Despite the popularity and good mapping functionality of the conventional method, the results can be improved by solving the problems created by time-independent mapping and oversmoothing. Time-independency assumes that each vector is converted on a frame-basis, disregarding the information contained in other frames. Therefore, the converted vector suffers from discontinuities and a solution is represented by introducing the dynamic features of vector parameters.

Toda [Toda et al., 2004] introduces the maximum likelihood estimation of parameter trajectory in which feature vectors are converted simultaneously over a time sequence. The source and target features at frame t consist of static and dynamic features: $X_t = [x_t^T, \Delta x_t^T]$ and $Y_t = [y_t^T, \Delta y_t^T]$ and the parameter vectors over an utterance are regarded as a single time-sequence vector: $X = [X_1^T, X_2^T, ..., X_T^T]$ and $Y = [Y_1^T, Y_2^T, ..., Y_T^T]$. The target time-sequence vector including dynamic features is computed from the initial static target vector: $Y = Wy$, where W is a $[2D_yT] - by - [D_yT]$ matrix of predefined weights.

Given a source vector X and the parameter set $\lambda^{(Z)}$ of the GMM trained on the joint vector $Z = [X^T, Y^T]$, the MLE-based mapping determines the converted target vector as follows:

$$\hat{y}_t = arg\,max\; P(Y|X, \lambda_Z) \tag{A.1.7}$$

The maximization in (A.1.7) is done by maximizing an auxiliary function such that the converted vector sequence is given for the suboptimal approximation $\hat{m} = arg\,max\; P(m|X, \lambda_Z)$ by:

$$\hat{y} = \left(W^T {D_{\hat{m}}^{(Y)}}^{-1} W\right)^{-1} W^T {D_{\hat{m}}^{(Y)}}^{-1} E_{\hat{m}}^{(Y)}, \tag{A.1.8}$$

where

$$E_{\hat{m}}^{(Y)} = [E_{\hat{m}_1,1}^{(Y)}, E_{\hat{m}_2,2}^{(Y)}, ..., E_{\hat{m}_T,T}^{(Y)}], \tag{A.1.9}$$

$${D_{\hat{m}}^{(Y)}}^{-1} = diag[{D_{\hat{m}_1}^{(Y)}}^{-1}, {D_{\hat{m}_2}^{(Y)}}^{-1}, ..., {D_{\hat{m}_T}^{(Y)}}^{-1}]. \tag{A.1.10}$$

Unlike in the MMSE method, the converted target vector is computed as a weighted sum of the mean vectors where covariance matrices are used as weights. The covariance matrices can be regarded as a confidence measure of conditional mean vectors from individual mixture components.

A.2 Superposition of Functional Contours model

The SFC model represents a comprehensive model of intonation which proposes a method of generating prosodic contours based on the link between phonetic forms and metalinguistic functions. The intonation has the ability to demarcate phonological units and to convey information about the propositional and interactional functions of these units. The hypotheses launched are that these functions are directly implemented using prototypical prosodic contours and that prosody is obtained by overlapping and adding the contours.

The hypothesis proposed by the SFC model is that the prosodic manifestation of linguistic and paralinguistic functions is represented by multiparametric contours coextensive to the carrier speech chunk. Also, the functional contours of different scopes belonging to the same function constitute a morphologically coherent family of contours. Such a family of contours can be generated using a so-called contour generator which is implemented using a simple neural network. This hypothesis has been tested using three main categories of socio-communicative functions: attitudes (using declaration, question, doubt-incredulity, obviousness, exclamation and irony), hierarchy (inter-dependence between phrases within an utterance), emphasis (narrow focus on a word). The different contours are superposed to form a global prosodic contour. Each metalinguistic function is encoded by a specific prototypical contour anchored to the scope of that function (i.e. the extent of the units to which the function applies) by a few landmarks, i.e. the beginning and end of the unit(s) concerned with this function. As the metalinguistic function can be applied to different scopes, it is characterized by a family of contours.

Contour generators are in charge of generating a coherent family of contours given only their scopes. These contour generators are implemented as simple feedforward neural networks which:

- receive as input linear ramps giving the absolute and relative distance of the current syllable from the closest landmarks of the scope
- deliver as output prosodic characteristics for the current syllable

The structure of such a neural network is depicted in Figure A.1[Bailly and Holm, 2005]:

The ill-posed problem of extracting elementary multiparametric contours from an observation represented by the sum of several contours is solved by an analysis-by-synthesis loop in which SFC generators are trained iteratively so that their overlapped and added prosodic contours best predict the observed realizations. At iteration n:

1. Generators predict the functional contours for all units of the corpus using the

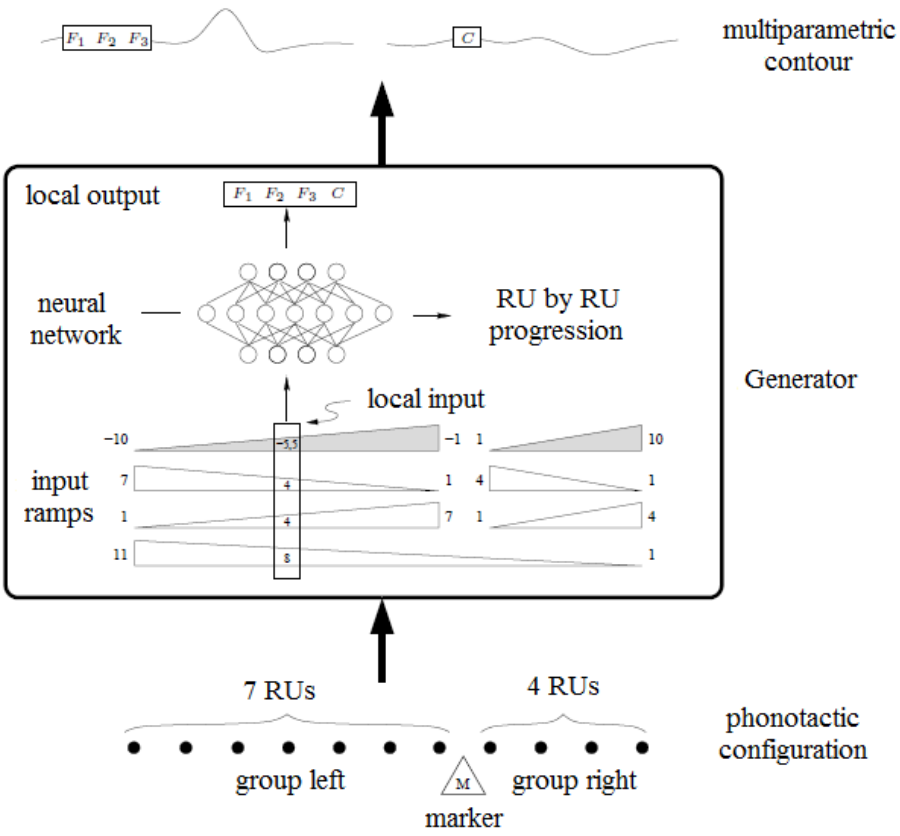

Figure A.1: Implementation of contour generators. Neural networks that receive input ramps generate 4 output values for each Rhythmic Unit (RU): F1, F2, F3 and the lengthening/shortening coefficient. For each RU, the input values represent the absolute (white ramps) and relative (grey ramps) position of that RU within the scope. Note that in the context on this thesis, only the attitudinal function is modeled, therefore the entire utterance is represented by the group left and that there is no group right.

parameters obtained from the previous iteration. If $n = 1$, the output is null.

2. For each utterance, the prosodic contours are computed by overlapping and adding the functional contours associated to all units.

3. For each utterance, the prediction error is computed and distributed among the contributing functional contours in accordance to scope variability. The prediction error is then added to each predicted functional contour to form new target functional contours.

4. The target functional contours are collected for all utterances and sorted according to the discourse function they implement. They will be considered new target patterns during a classical learning technique for neural networks.

The loop stops when errors are significantly reduced. The mechanism is explained in figure A.2:

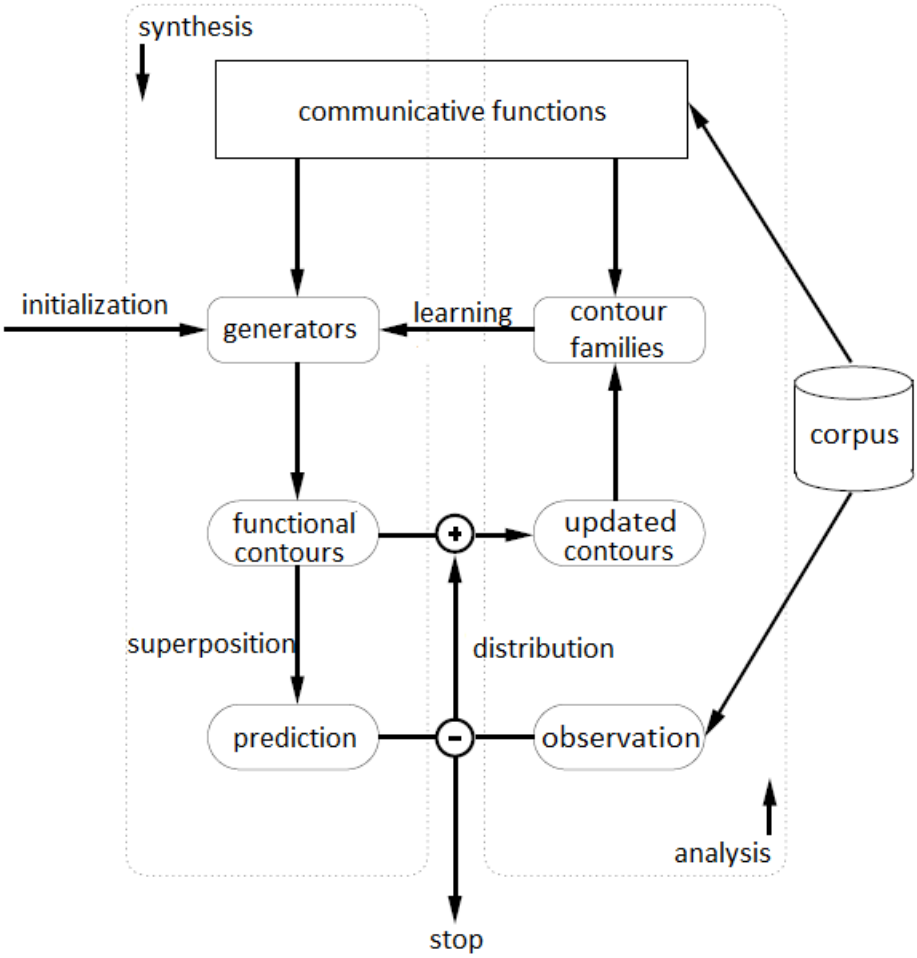

Figure A.2: Analysis-by-synthesis loop.

Bibliography

[Ahn et al., 2013] Ahn, J., Gobron, S., Thalmann, D., and Boulic, R. (2013). Asymmetric facial expressions: revealing richer emotions for embodied conversational agents. *Computer Animation and Virtual Worlds*, 24(6):539–551.

[Aihara et al., 2012] Aihara, R., Takashima, R., Takiguchi, T., and Ariki, Y. (2012). Gmm-based emotional voice conversion using spectrum and prosody features. *American Journal of Signal Processing*, 2(5):134–138.

[Albrecht et al., 2002a] Albrecht, I., Haber, J., Kähler, K., Schroder, M., and Seidel, H.-P. (2002a). " may i talk to you?:-)"-facial animation from text. In *Computer Graphics and Applications, 2002. Proceedings. 10th Pacific Conference on*, pages 77–86. IEEE.

[Albrecht et al., 2002b] Albrecht, I., Haber, J., and Seidel, H.-P. (2002b). Automatic generation of non-verbal facial expressions from speech. In *Advances in Modelling, Animation and Rendering*, pages 283–293. Springer.

[Alexander et al., 2013] Alexander, O., Fyffe, G., Busch, J., Yu, X., Ichikari, R., Jones, A., Debevec, P., Jimenez, J., Danvoye, E., Antionazzi, B., et al. (2013). Digital ira: Creating a real-time photoreal digital actor. In *ACM SIGGRAPH 2013 Posters*, page 1. ACM.

[Anderson et al., 2013] Anderson, R., Stenger, B., Wan, V., and Cipolla, R. (2013). Expressive visual text-to-speech using active appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3382–3389. IEEE.

[Aubergé and Bailly, 1995] Aubergé, V. and Bailly, G. (1995). Generation of intonation: a global approach. In *EUROSPEECH*.

[Bailly et al., 1991] Bailly, G., Barbe, T., and Wang, H.-D. (1991). Automatic labeling of large prosodic databases: Tools, methodology and links with a text-to-speech system. In *The ESCA Workshop on Speech Synthesis*, pages 77–86.

[Bailly et al., 2003] Bailly, G., Bérar, M., Elisei, F., and Odisio, M. (2003). Audiovisual speech synthesis. *International Journal of Speech Technology*, 6(4):331–346.

[Bailly and Gorisch, 2006] Bailly, G. and Gorisch, I. (2006). Generating german intonation with a trainable prosodic model. In *Interspeech*, pages 2366–2369.

[Bailly and Holm, 2005] Bailly, G. and Holm, B. (2005). Sfc: a trainable prosodic model. *Speech Communication*, 46(3):348–364.

[Bailly et al., 2010] Bailly, G., Raidt, S., and Elisei, F. (2010). Gaze, conversational agents and face-to-face communication. *Speech Communication*, 52(6):598–612.

[Barbosa, 1994] Barbosa, P. (1994). *Caractérisation et génération automatique de la structuration rythmique du français*. PhD thesis.

[Barbosa and Bailly, 1997] Barbosa, P. A. and Bailly, G. (1997). Generation of pauses within the z-score model. In *Progress in Speech Synthesis*, pages 365–381. Springer.

[Barbulescu et al., 2013] Barbulescu, A., Hueber, T., Bailly, G., Ronfard, R., et al. (2013). Audio-visual speaker conversion using prosody features. In *International Conference on Auditory-Visual Speech Processing*.

[Baron-Cohen, 2003] Baron-Cohen, S. (2003). *Mind reading: the interactive guide to emotions*. Jessica Kingsley Publishers.

[Baron-Cohen et al., 1997] Baron-Cohen, S., Wheelwright, S., Jolliffe, and Therese (1997). Is there a" language of the eyes"? evidence from normal adults, and adults with autism or asperger syndrome. *Visual Cognition*, 4(3):311–331.

[Bentivoglio et al., 1997] Bentivoglio, A. R., Bressman, S. B., Cassetta, E., Carretta, D., Tonali, P., and Albanese, A. (1997). Analysis of blink rate patterns in normal subjects. *Movement Disorders*, 12(6):1028–1034.

[Berndt and Clifford, 1994] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

[Beskow et al., 2006] Beskow, J., Granström, B., and House, D. (2006). Visual correlates to prominence in several expressive modes. In *INTERSPEECH*. Citeseer.

[Black and Hunt, 1996] Black, A. W. and Hunt, A. J. (1996). Generating f 0 contours from tobi labels using linear regression. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1385–1388. IEEE.

[Boersma, 2002] Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

[Bogert et al.,] Bogert, B., Healy, M., and Tukey, J. The frequency analysis of time series for echoes: cepstrum, pseudo-auto covariance, cross-cepstrum, and shaft cracking. In *Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed)*, pages 209–243.

[Bolinger, 1989] Bolinger, D. (1989). *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press.

[Brand, 1999] Brand, M. (1999). Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28. ACM Press/Addison-Wesley Publishing Co.

[Bregler et al., 1997] Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360. ACM Press/Addison-Wesley Publishing Co.

[Buhmann, 2003] Buhmann, M. D. (2003). *Radial basis functions: theory and implementations*, volume 12. Cambridge university press.

[Bulut et al., 2002] Bulut, M., Narayanan, S. S., and Syrdal, A. K. (2002). Expressive speech synthesis using a concatenative synthesizer. In *INTERSPEECH*.

[Busso et al., 2008] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

[Busso et al., 2007] Busso, C., Deng, Z., Grimm, M., Neumann, U., and Narayanan, S. (2007). Rigid head motion in expressive speech animation: Analysis and synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1075–1086.

[Busso et al., 2005] Busso, C., Deng, Z., Neumann, U., and Narayanan, S. (2005). Natural head motion synthesis driven by acoustic prosodic features. *Journal of Visualization and Computer Animation*, 16(3-4):283–290.

[Busso and Narayanan, 2007] Busso, C. and Narayanan, S. S. (2007). Interrelation between speech and facial gestures in emotional utterances: a single subject study. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(8):2331–2347.

[Cafaro et al., 2014] Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., and Pelachaud, C. (2014). Representing communicative functions in saiba with a unified function markup language. In *Intelligent Virtual Agents*, pages 81–94. Springer.

[Cahn, 1989] Cahn, J. E. (1989). Generating expression in synthesized speech. Master's thesis.

[Calinon and Billard, 2007] Calinon, S. and Billard, A. (2007). Incremental learning of gestures by imitation in a humanoid robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 255–262. ACM.

[Campbell, 1992] Campbell, W. N. (1992). Syllable-based segmental duration. *Talking machines: Theories, models, and designs*, pages 211–224.

[Cao et al., 2004] Cao, Y., Faloutsos, P., Kohler, E., and Pighin, F. (2004). Real-time speech motion synthesis from recorded motions. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 345–353. Eurographics Association.

[Cao et al., 2003] Cao, Y., Faloutsos, P., and Pighin, F. (2003). Unsupervised learning for speech motion editing. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 225–231. Eurographics Association.

[Cao et al., 2005] Cao, Y., Tien, W. C., Faloutsos, P., and Pighin, F. (2005). Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24(4):1283–1302.

[Cassell et al., 1994] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420. ACM.

[Cassell et al., 2004] Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2004). Beat: the behavior expression animation toolkit. In *Life-Like Characters*, pages 163–185. Springer.

[Cavé et al., 1996] Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. (1996). About the relationship between eyebrow movements and fo variations. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2175–2178. IEEE.

[Chuang and Bregler, 2002] Chuang, E. and Bregler, C. (2002). Performance driven facial animation using blendshape interpolation. *Computer Science Technical Report, Stanford University*, 2(2):3.

[Chuang and Bregler, 2005] Chuang, E. and Bregler, C. (2005). Mood swings: expressive speech animation. *ACM Transactions on Graphics (TOG)*, 24(2):331–347.

[Cohen and Massaro, 1993] Cohen, M. M. and Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. In *Models and techniques in computer animation*, pages 139–156. Springer.

[Costa et al., 2001] Costa, M., Chen, T., and Lavagetto, F. (2001). Visual prosody analysis for realistic motion synthesis of 3d head models. *Proc. of ICAV3D*, pages 343–346.

[Cowie et al., 2005] Cowie, R., Douglas-Cowie, E., and Cox, C. (2005). Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural networks*, 18(4):371–388.

[Cvejic et al., 2010] Cvejic, E., Kim, J., Davis, C., and Gibert, G. (2010). Prosody for the eyes: quantifying visual prosody using guided principal component analysis. In *INTERSPEECH*, pages 1433–1436.

[Darwin, 1872] Darwin, C. (1872). *Expression of emotion in man and animals*. Philosophical library.

[Davis et al., 2015] Davis, C., Kim, J., Aubanel, V., Zelic, G., and Mahajan, Y. (2015). The stability of mouth movements for multiple talkers over multiple sessions. In *Proceedings of the 2015 FAAVSP*.

[De Moraes et al., 2010] De Moraes, J. A., Rilliard, A., de Oliveira Mota, B. A., and Shochi, T. (2010). Multimodal perception and production of attitudinal meaning in brazilian portuguese. *Proc. Speech Prosody, paper*, 340.

[De Tournemire, 1994] De Tournemire, S. (1994). Recherche d'une stylisation extrême des contours de f0 en vue de leur apprentissage automatique. *Journées d'Etudes sur la Parole*, pages 75–80.

[Debevec, 2012] Debevec, P. (2012). The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia Technical Briefs*.

[Deng et al., 2004] Deng, Z., Narayanan, S., Busso, C., and Neumann, U. (2004). Audio-based head motion synthesis for avatar-based telepresence systems. In *Proceedings of the 2004 ACM SIGMM workshop on Effective telepresence*, pages 24–30. ACM.

[Ding et al., 2013] Ding, Y., Radenen, M., Artieres, T., and Pelachaud, C. (2013). Speech-driven eyebrow motion synthesis with contextual markovian models. In *ICASSP*, pages 3756–3760.

[Dunbar et al., 1997] Dunbar, R. I., Marriott, A., and Duncan, N. D. (1997). Human conversational behavior. *Human Nature*, 8(3):231–246.

[Dutoit, 1997] Dutoit, T. (1997). *An introduction to text-to-speech synthesis*, volume 3. Springer Science & Business Media.

[Eide et al., 2004] Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., and Pitrelli, J. (2004). A corpus-based approach to expressive speech synthesis. In *Fifth ISCA Workshop on Speech Synthesis*.

[Ekman, 1973] Ekman, P. (1973). Cross-cultural studies of facial expressions. *Darwin and facial expression: A century of research in review*, pages 169–229.

[Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

[Ekman and Friesen, 1971] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

[Ekman and Friesen, 1977] Ekman, P. and Friesen, W. V. (1977). Facial action coding system.

[Ekman and Scherer, 1984] Ekman, P. and Scherer, K. (1984). Expression and the nature of emotion. *Approaches to emotion*, 3:19–344.

[Ezzat and Poggio, 1997] Ezzat, T. and Poggio, T. (1997). Videorealistic talking faces: A morphing approach. In *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*, pages 141–144.

[Fanelli et al., 2010] Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., and Van Gool, L. (2010). A 3-d audio-visual corpus of affective communication. *Multimedia, IEEE Transactions on*, 12(6):591–598.

[Fónagy et al., 1983] Fónagy, I., Bérard, E., and Fónagy, J. (1983). Clichés mélodiques. *Folia linguistica*, 17(1-4):153–186.

[Fukayama et al., 2002] Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., and Hagita, N. (2002). Messages embedded in gaze of interface agents—impression management with agent's gaze. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 41–48. ACM.

[Gagneré et al., 2012] Gagneré, G., Ronfard, R., and Desainte-Catherine, M. (2012). La simulation du travail théâtral et sa" notation" informatique. *Notation du travail théâtral, du manuscript au numérique*.

[Govokhina et al., 2006] Govokhina, O., Bailly, G., Breton, G., and Bagshaw, P. (2006). Tda: A new trainable trajectory formation system for facial animation. In *Interspeech*, pages 2474–2477.

[Graf et al., 2002] Graf, H. P., Cosatto, E., Strom, V., and Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 396–401. IEEE.

[Granström and House, 2005] Granström, B. and House, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication*, 46(3):473–484.

[Granström et al., 2002] Granström, B., House, D., and Swerts, M. (2002). Multimodal feedback cues in human-machine interactions. In *Speech Prosody 2002, International Conference*.

[Haan et al., 1997] Haan, J., van Heuven, V. J., Pacilly, J., and van Bezooijen, R. (1997). An anatomy of dutch question intonation. *Linguistics in the Netherlands*, 14(1):97–108.

[Heuven, 1994] Heuven, V. v. (1994). Introducing prosodic phonetics. *Experimental studies of indonesian prosody*, 9:1–26.

[Heylen et al., 2008] Heylen, D., Kopp, S., Marsella, S. C., Pelachaud, C., and Vilhjálmsson, H. (2008). The next step towards a function markup language. In *Intelligent Virtual Agents*, pages 270–280. Springer.

[Hönemann et al., 2015] Hönemann, A., Mixdorff, H., and Rilliard, A. (2015). Classification of auditory-visual attitudes in german. In *International Conference on Auditory-Visual Speech Processing*.

[Hofer et al., 2007] Hofer, G., Shimodaira, H., and Yamagishi, J. (2007). Speech-driven head motion synthesis based on a trajectory model.

[Hyviirinen et al., 2001] Hyviirinen, A., Karhunen, J., and Oja, E. (2001). Independent component analysis. *Wiley and Sons*.

[Ichim et al., 2015] Ichim, A. E., Bouaziz, S., and Pauly, M. (2015). Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (TOG)*, 34(4):45.

[Izard, 1991] Izard, C. E. (1991). *The psychology of emotions*. Springer Science & Business Media.

[Jimenez, 2012] Jimenez, J. (2012). Separable subsurface scattering and photorealistic eyes rendering. Presented at Advances in reatime rendering in games course at ACM Siggraph 2012.

[Johan't Hart and Cohen, 1990] Johan't Hart, R. C. and Cohen, A. (1990). A perceptual study of intonation.

[Johnson, 2003] Johnson, W. L. (2003). Dramatic expression in opera, and its implications for conversational agents. Technical report, DTIC Document.

[Kaulard et al., 2012] Kaulard, K., Cunningham, D. W., Bülthoff, H. H., and Wallraven, C. (2012). The mpi facial expression database a validated database of emotional and conversational facial expressions. *PloS one*, 7(3):e32321.

[Kawahara, 2006] Kawahara, H. (2006). Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353.

[Kendon, 1994] Kendon, A. (1994). Do gestures communicate? a review. *Research on language and social interaction*, 27(3):175–200.

[Kipp, 2003] Kipp, M. (2003). *Gesture generation by imitation : from human behavior to computer character animation*. PhD thesis, Universite des Saarlandes, Postfach 151141, 66041 Saarbrucken.

[Kopp et al., 2006] Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., Thórisson, K. R., and Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent virtual agents*, pages 205–217. Springer.

[Krahmer et al., 2002] Krahmer, E., Ruttkay, Z., Swerts, M., and Wesselink, W. (2002). Pitch, eyebrows and the perception of focus. In *Speech Prosody 2002, International Conference*.

[Krahmer and Swerts, 2009] Krahmer, E. and Swerts, M. (2009). Audiovisual prosody — introduction to the special issue. *Language and speech*, 52(2-3):129–133.

[Le et al., 2012] Le, B. H., Ma, X., and Deng, Z. (2012). Live speech driven head-and-eye motion generators. *Visualization and Computer Graphics, IEEE Transactions on*, 18(11):1902–1914.

[Le Maguer, 2013] Le Maguer, S. (2013). *Évaluation expérimentale d'un système statistique de synthèse de la parole, HTS, pour la langue française*. PhD thesis, Université Rennes 1.

[Lee et al., 2008] Lee, J., DeVault, D., Marsella, S., and Traum, D. (2008). Thoughts on fml: Behavior generation in the virtual human communication architecture. *Proceedings of FML*, pages 83–95.

[Lee et al., 2002] Lee, S. P., Badler, J. B., and Badler, N. I. (2002). Eyes alive. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 637–644. ACM.

[Lee et al., 1995] Lee, Y., Terzopoulos, D., and Waters, K. (1995). Realistic modeling for facial animation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 55–62. ACM.

[Lehiste, 1970] Lehiste, I. (1970). *Suprasegmentals*. MIT Press Cambridge.

[Levine et al., 2010] Levine, S., Krähenbühl, P., Thrun, S., and Koltun, V. (2010). Gesture controllers. In *ACM Transactions on Graphics (TOG)*, volume 29, page 124. ACM.

[Levine et al., 2009] Levine, S., Theobalt, C., and Koltun, V. (2009). Real-time prosody-driven synthesis of body language. *ACM Transactions on Graphics (TOG)*, 28(5):172.

[Liu and Ostermann, 2011] Liu, K. and Ostermann, J. (2011). Realistic facial expression synthesis for an image-based talking head. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6. IEEE.

[Lucey et al., 2010] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE.

[Ma and Deng, 2009] Ma, X. and Deng, Z. (2009). Natural eye motion synthesis by modeling gaze-head coupling. In *Virtual reality conference, 2009. vr 2009. ieee*, pages 143–150. IEEE.

[Mac et al., 2009] Mac, D.-K., Aubergé, V., Rilliard, A., and Castelli, E. (2009). Audio-visual prosody of social attitudes in vietnamese: building and evaluating a tones balanced corpus. In *Tenth Annual Conference of the International Speech Communication Association*, pages 2263–2266.

[Mac et al., 2012] Mac, D.-K., Castelli, E., and Aubergé, V. (2012). Modeling the prosody of vietnamese attitudes for expressive speech synthesis. In *SLTU*, pages 114–118.

[Magnenat-Thalmann and Thalmann, 2005a] Magnenat-Thalmann, N. and Thalmann, D. (2005a). *Handbook of virtual humans*. John Wiley & Sons.

[Magnenat-Thalmann and Thalmann, 2005b] Magnenat-Thalmann, N. and Thalmann, D. (2005b). Virtual humans: thirty years of research, what next? *The Visual Computer*, 21(12):997–1015.

[Mahmoud et al., 2011] Mahmoud, M., Baltrušaitis, T., Robinson, P., and Riek, L. D. (2011). 3d corpus of spontaneous complex mental states. In *Affective computing and intelligent interaction*, pages 205–214. Springer.

[Malcangi, 2010] Malcangi, M. (2010). Text-driven avatars based on artificial neural networks and fuzzy logic. *International journal of computers*, 4(2):61–69.

[Marsella et al., 2013] Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., and Shapiro, A. (2013). Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 25–35. ACM.

[Massaro and Palmer, 1998] Massaro, D. W. and Palmer, S. E. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*, volume 1. MIT Press Cambridge, MA.

[Masuko et al., 1996] Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. (1996). Speech synthesis using hmms with dynamic features. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, pages 389–392. IEEE.

[Mattheyses and Verhelst, 2015] Mattheyses, W. and Verhelst, W. (2015). Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217.

[McGurk and MacDonald, 1976] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. Nature Publishing Group.

[Mori, 1970] Mori, M. (1970). The uncanny valley. volume 7, pages 33–5. Energy.

[Mori et al., 2006] Mori, S., Moriyama, T., and Ozawa, S. (2006). Emotional speech synthesis using subspace constraints in prosody. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1093–1096. IEEE.

[Morlec et al., 1995] Morlec, Y., Bailly, G., and Aubergé, V. (1995). Synthesis and evaluation of intonation with a superposition model. In *EUROSPEECH*, volume 3, pages 2043–2046.

[Morlec et al., 2001] Morlec, Y., Bailly, G., and Aubergé, V. (2001). Generating prosodic attitudes in french: data, model and evaluation. *Speech Communication*, 33(4):357–371.

[Munhall et al., 2004] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility head movement improves auditory speech perception. *Psychological science*, 15(2):133–137.

[Murray and Arnott, 1995] Murray, I. R. and Arnott, J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16(4):369–390.

[Nakano and Kitazawa, 2010] Nakano, T. and Kitazawa, S. (2010). Eyeblink entrainment at breakpoints of speech. *Experimental brain research*, 205(4):577–581.

[Niewiadomski et al., 2009] Niewiadomski, R., Bevacqua, E., Mancini, M., and Pelachaud, C. (2009). Greta: an interactive expressive eca system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 1399–1400. International Foundation for Autonomous Agents and Multiagent Systems.

[Ohala, 1996] Ohala, J. J. (1996). Ethological theory and the expression of emotion in the voice. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1812–1815. IEEE.

[Ortony et al., 2005] Ortony, A., Norman, D., and Revelle, W. (2005). Effective functioning: A three level model of affect, motivation, cognition, and behavior. *Who needs emotions*, pages 173–202.

[Ostermann, 1998] Ostermann, J. (1998). Animation of synthetic faces in mpeg-4. In *Computer Animation 98. Proceedings*, pages 49–55. IEEE.

[Ouni et al., 2013] Ouni, S., Colotte, V., Musti, U., Toutios, A., Wrobel-Dautcourt, B., Berger, M.-O., and Lavecchia, C. (2013). Acoustic-visual synthesis technique using bimodal unit-selection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):1–13.

[Oyekoya et al., 2010] Oyekoya, O., Steed, A., and Steptoe, W. (2010). Eyelid kinematics for virtual characters. *Computer animation and virtual worlds*, 21(3-4):161–171.

[Parke, 1974] Parke, F. I. (1974). A parametric model for human faces. Technical report, DTIC Document.

[Pelachaud et al., 1996] Pelachaud, C., Badler, N. I., and Steedman, M. (1996). Generating facial expressions for speech. *Cognitive science*, 20(1):1–46.

[Perlin and Goldberg, 1996] Perlin, K. and Goldberg, A. (1996). Improv: A system for scripting interactive actors in virtual worlds. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 205–216. ACM.

[Queneau, 1947] Queneau, R. (1947). *Exercises in style*. Editions Gallimard.

[Ronfard, 2012] Ronfard, R. (2012). Notation et reconnaissance des actions scéniques par ordinateur. *Notation du travail théâtral, du manuscript au numérique*.

[Roon et al., 2015] Roon, K. D., Tiede, M. K., Dawson, K. M., and Whalen, D. (2015). Coordination of eyebrow movement with speech acoustics and movement.

[Ruhland et al., 2014] Ruhland, K., Andrist, S., Badler, J., Peters, C., Badler, N., Gleicher, M., Mutlu, B., and Mcdonnell, R. (2014). Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems. In *Eurographics State-of-the-Art Report*, pages 69–91.

[Savran et al., 2008] Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., and Akarun, L. (2008). Bosphorus database for 3d face analysis. In *Biometrics and Identity Management*, pages 47–56. Springer.

[Scherer, 1986] Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):143.

[Scherer and Ellgring, 2007] Scherer, K. R. and Ellgring, H. (2007). Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion*, 7(1):158.

[Scherer et al., 2012] Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, F., Egan, A., Morency, L.-P., et al. (2012). Perception markup language: towards a standardized representation of perceived nonverbal behaviors. In *Intelligent virtual agents*, pages 455–463. Springer.

[Scherer et al., 1980] Scherer, U., Helfrich, H., and Scherer, K. (1980). Paralinguistio behaviour: Internal push or external pull? In *Language: social psychological perspectives: selected papers from the first International Conference on Social Psychology and Language held at the University of Bristol, England, July 1979*, page 279. Pergamon.

[Schnitzler, 1912] Schnitzler, A. (1912). *La ronde*. P.V. Stock, Paris.

[Schnitzler, 1920] Schnitzler, A. (1920). *Hands around*. Privately Printed for Subscribers, Ny.

[Schröder and Breuer, 2004] Schröder, M. and Breuer, S. (2004). Xml representation languages as a way of interconnecting tts modules. In *INTERSPEECH*.

[Schröder et al., 2007] Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., and Wilson, I. (2007). What should a generic emotion markup language be able to represent? In *Affective Computing and Intelligent Interaction*, pages 440–451. Springer.

[Schröder and Trouvain, 2003] Schröder, M. and Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

[Sendra et al., 2013] Sendra, V. C., Kaland, C., Swerts, M., and Prieto, P. (2013). Perceiving incredulity: The role of intonation and facial gestures. *Journal of Pragmatics*, 47(1):1–13.

[Srinivasan and Massaro, 2003] Srinivasan, R. J. and Massaro, D. W. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in english. *Language and Speech*, 46(1):1–22.

[Stylianou et al., 1995] Stylianou, Y., Cappe, O., and Moulines, E. (1995). Statistical methods for voice quality transformation. In *EUROSPEECH*.

[Swerts and Krahmer, 2010] Swerts, M. and Krahmer, E. (2010). Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics*, 38(2):197–206.

[Tamura et al., 1999] Tamura, M., Kondo, S., Masuko, T., and Kobayashi, T. (1999). Text-to-audio-visual speech synthesis based on parameter generation from hmm. In *EUROSPEECH*, pages 3745–3748 vol.6.

[Tao et al., 2006] Tao, J., Kang, Y., and Li, A. (2006). Prosody conversion from neutral speech to emotional speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1145–1154.

[Taylor and Black, 1999] Taylor, P. and Black, A. W. (1999). Speech synthesis by phonological structure matching. pages 623—-626.

[Taylor et al., 2012] Taylor, S. L., Mahler, M., Theobald, B.-J., and Matthews, I. (2012). Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284. Eurographics Association.

[Theobald, 2007] Theobald, B. (2007). Audiovisual speech synthesis. In *International Congress on Phonetic Sciences*, pages 285–290.

[Thiebaux et al., 2009] Thiebaux, M., Lance, B., and Marsella, S. (2009). Real-time expressive gaze animation for virtual humans. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 321–328. International Foundation for Autonomous Agents and Multiagent Systems.

[Toda et al., 2004] Toda, T., Black, A. W., and Tokuda, K. (2004). Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis. In *Fifth ISCA Workshop on Speech Synthesis*.

[Toda et al., 2007] Toda, T., Black, A. W., and Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(8):2222–2235.

[Vandeventer et al., 2015] Vandeventer, J., Aubrey, A. J., Rosin, P. L., and Marshall, D. (2015). 4d cardiff conversation database (4d ccdb): A 4d database of natural, dyadic conversations. In *International Conference on Auditory-Visual Speech Processing*.

[Veaux and Rodet, 2011] Veaux, C. and Rodet, X. (2011). Intonation conversion from neutral to expressive speech. In *INTERSPEECH*, pages 2765–2768.

[Vertegaal et al., 2001] Vertegaal, R., Slagter, R., Van der Veer, G., and Nijholt, A. (2001). Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 301–308. ACM.

[Vilhjálmsson et al., 2007] Vilhjálmsson, H., Cantelmo, N., Cassell, J., Chafai, N. E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A. N., Pelachaud, C., et al. (2007). The behavior markup language: Recent developments and challenges. In *Intelligent virtual agents*, pages 99–111. Springer.

[Vilhjálmsson, 2004] Vilhjálmsson, H. H. (2004). Animating conversation in online games. In *Entertainment Computing–ICEC 2004*, pages 139–150. Springer.

[Vlasic et al., 2005] Vlasic, D., Brand, M., Pfister, H., and Popović, J. (2005). Face transfer with multilinear models. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 426–433. ACM.

[Wang et al., 2010] Wang, L., Qian, X., Han, W., and Soong, F. K. (2010). Synthesizing photo-real talking head via trajectory-guided sample selection. In *INTERSPEECH*, volume 10, pages 446–449.

[Waters, 1987] Waters, K. (1987). A muscle model for animation three-dimensional facial expression. In *ACM SIGGRAPH Computer Graphics*, volume 21, pages 17–24. ACM.

[Weise et al., 2011] Weise, T., Bouaziz, S., Li, H., and Pauly, M. (2011). Realtime performance-based facial animation. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2011)*, 30(4):77:1–77:10.

[Wu et al., 2010] Wu, C.-H., Hsia, C.-C., Lee, C.-H., and Lin, M.-C. (2010). Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1394–1405.

[Xu, 2011] Xu, Y. (2011). Speech prosody: A methodological review. *Journal of Speech Sciences*, 1(1):85–115.

[Zeng et al., 2009] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58.