Modèle statistique de l'animation expressive de la parole et du rire pour un agent conversationnel animé

Yu Ding

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2014-ENST-0050

EDITE - ED 130

Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « SIGNAL et IMAGES »

présentée et soutenue publiquement par

Yu DING

le 26 September 2014

Data-Driven Expressive Animation Model of Speech and Laughter for an Embodied Conversational Agent

Directeur de thèse : **Catherine PELACHAUD**
Co-encadrement de la thèse : **Thierry ARTIERES**

Jury

M. Pierre CHEVAILLIER, Professeur, Ecole Nationale d'ingénieurs de Brest	Rapporteur
M. Olivier PIETQUIN, Professeur, Université Lille 1	Rapporteur
M. Jean-Claude MARTIN, Professeur, Université Paris-Sud	Examinateur
M. Thierry DUTOIT, Professeur, Université Mons, Belgique	Examinateur
M. Frédéric BEVILACQUA, Chercheur, Inst. de Rech. et Coord. Acous. /Musique	Examinateur
Mme. Catherine PELACHAUD, Directeur de recherche, Télécom ParisTech	Directeur de thèse
M. Thierry ARTIERES, Professeur, Université Pierre et Marie CURIE	Directeur de thèse

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech

46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

THÈSE

Acknowledgements

Firstly, I would like to express my sincere appreciation to my supervisors Catherine Pelachaud and Thierry Artières. They always guide me with their profound knowledge, inspire me with their remarkable insight and support me with their great patience. It would have been impossible for me to complete my research works and finish this thesis smoothly without their generous assistance. Their edification will benefit me not only in my graduate stage, but also throughout my entire research career.

Thank Catherine PELACHAUD and Thierry ARTIERES to give me the chance of enjoying research work. The PhD experience will affect all my life. It gives me courage and confidence to face unknown problems in work and even in life, to analyze them and then to solve them. The PhD experience gives me an important cognitive training on how to find and solve an unknown problem.

I would also like to thank Professor Gérard Chollet for his suggestions in my mi-term PhD presentation.

I would also like to thank Dr Mathieu Radenenm for his cooperation in our collaborative works.

I would also like to thank Ms. Laurence Zelmar, Ms. Janique Regis, Ms. Florence Besnard and Mr. Dominique Roux for their help in administrative processing.

I am indebted to my colleagues who supported me in my works and helped me in my life. I would like to thank Andre-Marie Pez, Radoslaw Niewiadomski, Jing Huang, Nesrine Fourati, Ken Prepin, Sabrina Campano, Magalie Ochs, Pierre Philippe, Brice Donval, Kevin Sanlaville, Nadine Glas, Brian Ravenet, Florian Pécune, Chloé Clavel, Caroline Langlet, Mathieu Chollet, Angelo Cafaro, Sylwia Julia Hyniewska, Elisabetta Bevacqua, Quoc Anh Le and Etienne de Sevin for their great help and assistance.

Here I have the honor to be friend of some excellent people and I'd like to express my thanks to them for their kind encouragement and company. They are Dr. Hao Cai, Dr. Ruomin Wang, Dr. Yuhui Bai, Dr. Jiaji Liu, Dr. Haiyang Wang and Mr. Yuanjing Xiao. My life in Paris becomes colorful and memorable thanks to them.

Finally from my heart, I would like to express my gratitude and love to my family, my dear father, mother and wife, who always supported me whenever and wherever. If any achievement comes to me, I'd like to delicate it to them. I would like also to express thank and love to my parents-in-law who supported me in my works and my life.

Modèle Statistique de l’Animation Expressive de la Parole et du Rire pour un Agent Conversationnel Animé

RESUME : Texte

Notre objectif est de simuler des comportements multimodaux expressifs pour les agents conversationnels animés ACA. Ceux-ci sont des entités dotées de capacités affectives et communicationnelles; ils ont souvent une apparence humaine. Quand un ACA parle ou rit, il est capable de montrer de façon autonome des comportements multimodaux pour enrichir et compléter son discours prononcé et transmettre des informations qualitatives telles que ses émotions. Notre recherche utilise les modèles d’apprentissage à partir données. Un modèle de génération de comportements multimodaux pour un personnage virtuel parlant avec des émotions différentes a été proposé ainsi qu’un modèle de simulation du comportement de rire sur un ACA. Notre objectif est d'étudier et de développer des générateurs d'animation pour simuler la parole expressive et le rire d’un ACA. En partant de la relation liant prosodie de la parole et comportements multimodaux, notre générateur d'animation prend en entrée les signaux audio prononcés et fournit en sortie des comportements multimodaux.

Notre travail vise à utiliser un modèle statistique pour saisir la relation entre les signaux donnés en entrée et les signaux de sortie; puis cette relation est transformée en modèle d’animation 3D. Durant l'étape d’apprentissage, le modèle statistique est entrainé à partir de paramètres communs qui sont composés de paramètres d'entrée et de sortie. La relation entre les signaux d'entrée et de sortie peut être capturée et caractérisée par les paramètres du modèle statistique. Dans l'étape de synthèse, le modèle entrainé est utilisé pour produire des signaux de sortie (expressions faciale, mouvement de tête et du torse) à partir des signaux d'entrée (F0, énergie de la parole ou pseudo-phonème du rire). La relation apprise durant la phase d'apprentissage peut être rendue dans les signaux de sortie.

Notre module proposé est basé sur des variantes des modèles de Markov cachés (HMM), appelées HMM contextuels. Ce modèle est capable de capturer la relation entre les mouvements multimodaux et de la parole (ou rire); puis cette relation est rendue par l’animation de l’ACA.

Mots clés : Modèle de Markov caché, Agent Conversationnel Animé, Synthèse d’Animation, Animation de la Parole, Animation du Rire

Data-Driven Expressive Animation Model of Speech and Laughter for an Embodied Conversational Agent

ABSTRACT : Text

Our aim is to render expressive multimodal behaviors for Embodied conversational agents, ECAs. ECAs are entities endowed with communicative and emotional capabilities; they have human-like appearance. When an ECA is speaking or laughing, it is capable of displaying autonomously behaviors to enrich and complement the uttered speech and to convey qualitative information such as emotion. Our research lies in the data-driven approach. It focuses on generating the multimodal behaviors for a virtual character speaking with different emotions. It is also concerned with simulating laughing behavior on an ECA. Our aim is to study and to develop human-like animation generators for speaking and laughing ECA. On the basis of the relationship linking speech prosody and multimodal behaviors, our animation generator takes as input human uttered audio signals and output multimodal behaviors.

Our work focuses on using statistical framework to capture the relationship between the input and the output signals; then this relationship is rendered into synthesized animation. In the training step, the statistical framework is trained based on joint features, which are composed of input and of output features. The relation between input and output signals can be captured and characterized by the parameters of the statistical framework. In the synthesis step, the trained framework is used to produce output signals (facial expression, head and torso movements) from input signals (F0, energy for speech or pseudo-phoneme of laughter). The relation captured in the training phase can be rendered into the output signals.

Our proposed module is based on variants of Hidden Markov Model (HMM), called Contextual HMM. This model is capable of capturing the relationship between human motions and speech (or laughter); then such relationship is rendered into the synthesized animations.

Keywords : Hidden Markov Model, Embodied Conversational Agent, Animation Synthesis, Speech Animation, Laughter Animation

Résumé

Notre objectif est de simuler des comportements multimodaux expressifs pour les agents conversationnels animés ACA. Ceux-ci sont des entités dotées de capacités affectives et communicationnelles; ils ont souvent une apparence humaine. Quand un ACA parle ou rit, il est capable de montrer de façon autonome des comportements multimodaux pour enrichir et compléter son discours prononcé et transmettre des informations qualitatives telles que ses émotions. Notre recherche utilise les modèles d’apprentissage à partir données. Un modèle de génération de comportements multimodaux pour un personnage virtuel parlant avec des émotions différentes a été proposé ainsi qu’un modèle de simulation du comportement de rire sur un ACA. Notre objectif est d'étudier et de développer des générateurs d'animation pour simuler la parole expressive et le rire d’un ACA. En partant de la relation liant prosodie de la parole et comportements multimodaux, notre générateur d'animation prend en entrée les signaux audio prononcés et fournit en sortie des comportements multimodaux.

Notre travail vise à utiliser un modèle statistique pour saisir la relation entre les signaux donnés en entrée et les signaux de sortie; puis cette relation est transformée en modèle d’animation 3D. Durant l'étape d’apprentissage, le modèle statistique est entrainé à partir de paramètres communs qui sont composés de paramètres d'entrée et de sortie. La relation entre les signaux d'entrée et de sortie peut être capturée et caractérisée par les paramètres du modèle statistique. Dans l'étape de synthèse, le modèle entrainé est utilisé pour produire des signaux de sortie (expressions faciale, mouvement de tête et du torse) à partir des signaux d'entrée (F0, énergie de la parole ou pseudo-phonème du rire). La relation apprise durant la phase d'apprentissage peut être rendue dans les signaux de sortie.

Notre module proposé est basé sur des variantes des modèles de Markov cachés (HMM), appelées HMM contextuels. Ce modèle est capable de capturer

la relation entre les mouvements multimodaux et de la parole (ou rire); puis cette relation est rendue par l’animation de l’ACA.

La suite de ce document se présentera comme suit ; la première section introduira le domaine d’application du travail de thèse ; la deuxième section décrira les contributions de cette thèse; la troisième section présentera brièvement des travaux existants sur lesquels notre travail s’appuie; la quatrième section présentera les modèles développés pour la synthèse d’animation à partir de la parole; la cinquième section présentera les modèles développés pour la synthèse d’animation à partir du rire ; la cinquième section conclura ce résume.

1. Introduction

La communication humaine implique des signaux audio et visuels. Les signaux audio correspondent à la parole y compris le contenu parlé et la prosodie; les signaux visuels concernent l'expression du visage et les mouvements du corps. De plus, les humains sont très habiles à déduction et l'alignement de divers mouvements subtils pour satisfaire une intention communicative, y compris l'émotion. Ils sont également capables de lire et décoder ces comportements complexes [99]. Les signaux de communication peuvent être utilisés par l'homme pour déduire et exprimer une intention, un état émotionnel, etc.

Les Agents Conversationnels Animés (ACA) sont une forme d'interface homme-machine. Ce sont des entités dotées de capacités de communication et d'expression car ayant une apparence humaine. Un ACA est capable de communiquer avec l'homme ou un autre ACA. Il est souvent installé sur un appareil équipé d'une caméra, d'un écran, d'un haut-parleur et d'un microphone, ce qui permet de recevoir et de transmettre les signaux audio et visuels. Par conséquent, il est capable d'écouter et de parler à son interlocuteur (par exemple, un utilisateur); il est aussi capable d'observer les expressions du visage et de corps de l'interlocuteur et d'affichage leur expression à l'interlocuteur.

Depuis quelques années, l'ACA est devenu de plus en plus populaire dans plusieurs applications d'interactions homme-machine, comme l'encadrement social. Par exemple, dans le Project Européen TARDIS (http://public.tardis-project.eu/), l'ACA est utilisé en tant que recruteur virtuel. L'ACA communique avec un utilisateur (qui joue le rôle d'un candidat) lors de la formation d'entrevue d'emploi. Le recruteur virtuel peut décider de façon autonome de procéder à l'entretien; il peut choisir attitudes sociales à exprimer envers la personne interrogée. La figure 1 montre deux images d'un ACA avec des attitudes amicales et hostiles. La figure 2 montre les captures d'écran des interactions entre le recruteur virtuel et l'interviewé. Dans cette application, l’ACA est évalué et perçu par l'homme (les utilisateurs), donc la qualité d’expression est cruciale pour engager les utilisateurs dans une telle application.

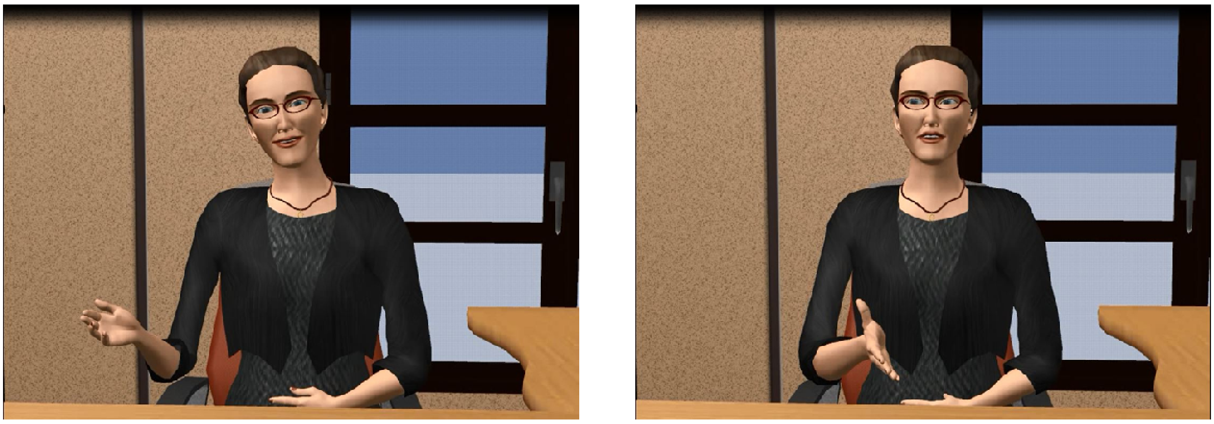


Figure 1: Exemples de l’ACA avec une attitude amicale (à gauche) et hostile (à droite) dans l’entretien d'embauche

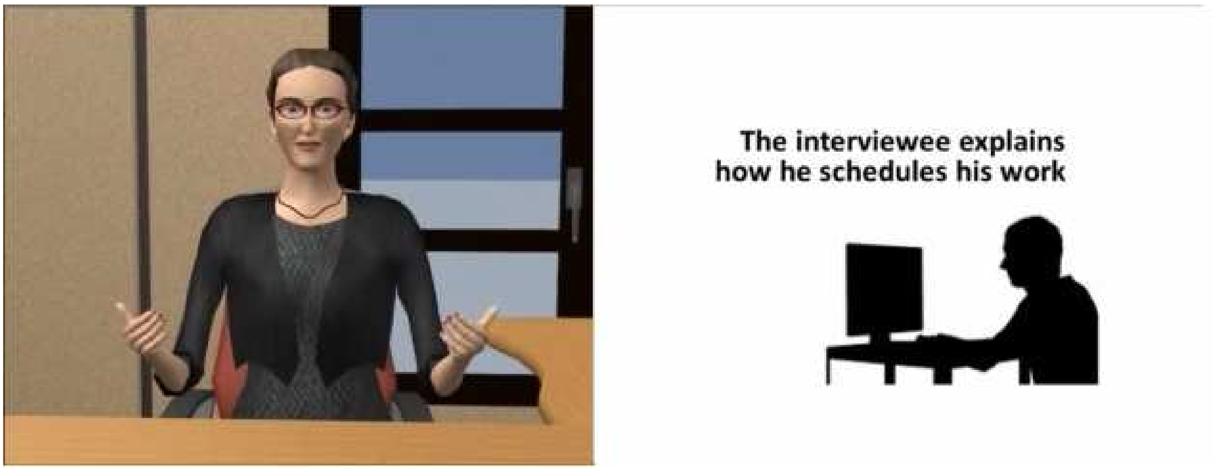


Figure 2: Captures d'écran de l'entretien d'embauche modélisé [20]. L’image de gauche montre le recruteur virtuel; l’image de droite montre la réponse de la personne interrogée.

De nombreuses études de psychologie ont été menées sur la caractérisation des expressions multimodales d'émotions. Les résultats de ces études sont exploités par des chercheurs afin de modéliser les comportements émotionnels des ACA. Ces modèles concernent non seulement la manifestation des émotions par le biais d’expressions faciales [12], mais aussi via les mouvements du corps [28].

Pour modéliser les comportements de communication d’un ACA, les modèles de procédure ont été initialement proposés par [23, 95, 22, 8]. Elles sont basées sur un ensemble de règles provenant de la littérature en communication humaine. Toutefois, il est extrêmement difficile de définir un grand ensemble de règles pour décrire toutes les possibilités des comportements humains. De plus, plusieurs règles peuvent entraîner des conflits dans la synthèse d’animation []. En outre, les animations obtenues à partir de ces modèles souffrent d'un manque de naturel et variabilité. Donc, des chercheurs ont exploré d’autres approches basées sur les données humaines, appelées modèles statistiques, qui ont été élaborées pour améliorer l’intelligibilité et la variabilité des animations synthétisées. Ces modèles reposent sur l’usage d’une grande base de données des mouvements humains expressifs. Ces modèles ont la capacité de capturer la corrélation entre le signal audio de parole et les expressions faciaux ou corporelles [14, 123, 39, 31, 19, 18, 56, 78, 70, 71, 26].

Notre recherche s’appuie sur ces modèles statistiques. Elle vise à utiliser de tels modèles pour calculer la relation entre le signal audio de parole ou de rire; puis utiliser cette relation pour synthétiser l’animation d’un ACA à partir du signal audio de parole ou de rire.

2. Contributions

Nos contributions peuvent être divisées en deux parties indépendantes. Tout d'abord, nous avons construit des modèles de synthèse d'animation faciale à partir du signal audio de parole pour l’ACA; ensuite, nous avons construit des modèles de synthèse de l'animation à partir du signal audio de rire pour l’ACA. Il s’agit de l'animation de l'expression du visage, le torse et des épaules. Ces modèles peuvent être utilisés comme générateurs d’animation de parole et de rire pour l’ACA.

Dans notre travail de synthèse de l'animation de la parole, nous introduirons l'utilisation des modèles de Markov cachés contextuels [121, 103] pour synthétiser les animations. Des travaux existant ont utilisé les HMM contextuels pour la reconnaissance. Dans les HMM contextuels, les fichiers DPG (distribution de probabilité gaussienne) sont définis en fonction de certaines variables externes (ou contextuelles); ces variables peuvent être l’état émotionnel, l'âge, etc. Dans notre travail, les variables contextuelles sont les caractéristiques du signal audio. En outre, nous avons étendu les HMM contextuels à des HMM entièrement paramétrés. Puis les HMM entièrement paramétrés sont utilisés pour synthétiser l'animation à partir des signaux de parole, où les caractéristiques de la parole sont utilisées comme variables externes.

Les modèles proposés sont évalués par le calcul de la distance entre les trajectoires du mouvement humain et de l'animation de synthèse. Les résultats montrent que nos modèles donnent de meilleurs résultats sur cette tâche que les modèles de référence (HMM standard).

Les HMM ont été utilisés pour synthétiser l'animation dans des travaux existants [19, 56, 18, 78]. Nos modèles donnent de meilleurs résultats que les travaux existants. L'étude montre que prendre en compte à la fois les sourcils et les mouvements de tête augmentent la précision de l'animation de synthèse;

ce résultat confirme aussi que les mouvements des sourcils et de la tête se sont reliés.

Dans notre travail de synthèse de l'animation du rire, nous avons construit des générateurs d'animation. Ces générateurs concernent le haut du corps, par exemple la mâchoire, les lèvres, les joues, les paupières, les sourcils, la tête, les épaules et le torse. A notre connaissances, nos modèles sont la première tentative de construire des générateurs d'animation pour la synthèse de l'animation du rire pour tout le corps.

D'abord, nous avons utilisé trois méthodes pour calculer les animations des différentes parties du corps. Des modèles de fonctions linéaires de régression sont utilisés comme générateurs de la lèvre et de la mâchoire, qui prennent en entrées les séquences de pseudo-phonèmes de rire et des caractéristiques de prosodie de rire. Les mouvements de la joue, des paupières, des sourcils et de la tête sont synthétisés par concaténation des mouvements segmentés issus de la base de données. Cette synthèse prend en entrée la séquence de pseudo-phonèmes de rire et les intensités de ces phonèmes. Les mouvements des épaules et du torse sont estimés par proportionnel-dérivé (PD) Contrôleur, qui prend en entrée les mouvements synthétisés de la tête. Ces trois approches peuvent synthétiser les animations correspondantes en temps réel. Une évaluation subjective a été réalisée pour estimer la pertinence de ces approches.

Pour étudier en outre les animations du haut du corps, nous avons construit un nouveau jeu de données de rire humain, où l'audio du rire humain, les rotations de la tête et du torse sont enregistrés. Un HMM spécifique, boucle HMM (BHMM), est utilisé pour capturer et synthétiser les types de mouvements (tremblements) de la tête et du torse. En plus, les probabilités de transition d'états paramétrées sont utilisées dans LHMM; elle est appelée TPLHMM. TPLHMM est utilisé pour capturer et rendre la relation entre le signal audio du rire et les mouvements de la tête et du torse. Enfin, nous nous sommes inspirés de HMMs couplés. Nous en construisant une variante en combinant plusieurs TPLHMMs couplés, appelés CTPLHMMs, pour capturer et rendre la relation entre les mouvements de la tête et du torse.

L'évaluation objective montre que les modèles proposés peuvent générer rire des séquences de mouvements. L'évaluation subjective montre que nos modèles sont capables de capturer le dynamisme du mouvement de rires et d'améliorer la perception humaine.

3. L’état de l’art

Synthétiser une séquence d'observations réaliste (appelée ci-après une trajectoire) à partir d'un HMM est une question essentielle. La synthèse de la séquence d'observations la plus probable étant donné une séquence d'états particulière donne une trajectoire constante par morceaux très peu probable. Bien qu'une technique d'interpolation puisse être utilisée pour lisser la séquence d’observations, cela pourrait endommager la dynamique perçue de l'animation déduite du HMM appris.

Une technique a été proposée dans [113] pour synthétiser une trajectoire réaliste à partir d'un HMM appris. Dans ce travail, un vecteur de caractéristiques comprend un ensemble de caractéristiques dites statiques accompagnées de leurs vitesses de leurs accélérations. Cette méthode vise à synthétiser la séquence d'observations (autrement dit de caractéristiques statiques) la plus probable qui satisfait certaines contraintes sur leurs évolutions. La figure 3 montre un échantillon de la trajectoire générée par un HMM [57] en utilisant cette méthode. Comme on peut le voir, la trajectoire produite (ligne pleine) est beaucoup plus lisse que la séquence la plus probable de l'observation (ligne pointillée).

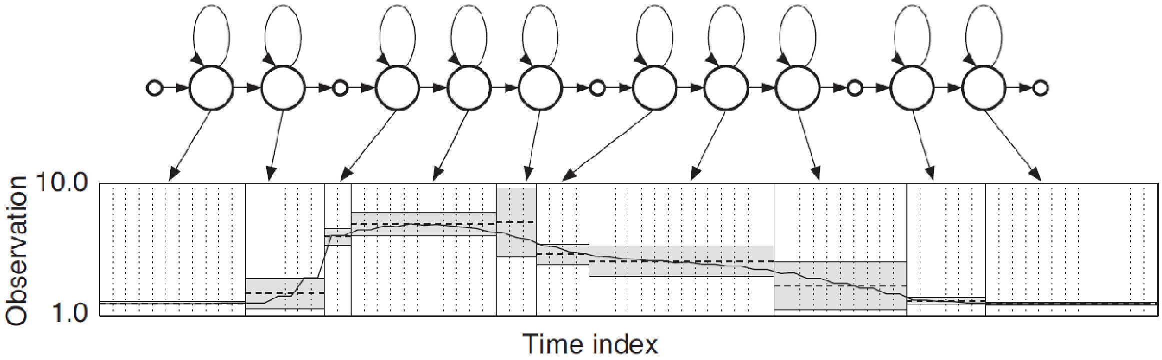


Figure 3: Un échantillon de la trajectoire générée à partir d'un HMM appris [57]. La ligne pointillée est la séquence des moyennes. La ligne continue est générée par la technique de synthèse de [113]. Il s'agit d'une trajectoire beaucoup plus

lisse que la ligne pointillée.

Dans notre travail, les chaînes de Markov sont utilisées pour modéliser les séquences de descripteurs de mouvement en considérant ces descripteurs comme les observations des états cachés. Pour dépasser les limites inhérentes aux HMMs, dans notre travail, les descripteurs de la parole ne sont pas toujours pris comme des observations émises par l'état caché, ils sont utilisés pour calculer les paramètres des HMMs dans l'étape de synthèse. Cette méthode s'inspire d'une extension des HMMs, appelée HMMs contextuels [121, 103]. Ceux-ci sont des HMMs dont les distributions des observations dépendent de variables contextuelles, aussi appelées variables externes. Les HMMs contextuels ont été utilisés pour la reconnaissance de gestes dans des travaux antérieurs, alors qu'ils sont utilisés pour la synthèse des comportements dans notre travail. Basé sur les HMMs contextuels existants, nous avons développé une nouvelle extension de ceux-ci, appelés HMMs entièrement paramétrables (FPHMM). Pour les FPHMMs, non seulement les observations, mais aussi les probabilités de transition entre les états dépendent de variables contextuelles. Dans notre travail, les variables contextuelles sont les caractéristiques de prosodie de parole.

Les Modèles de Markov cachés (HMM) sont des modèles génératifs statistiques (autrement dit, ils peuvent être utilisés pour générer/synthétiser les données) bien connu pour analyser les flux des données. Ils peuvent être appliqués à des données tel que le signal de parole ou les descripteurs du mouvement humain. Ces modèles sont appris automatiquement à partir d'un corpus de données. Cela permet de capturer les caractéristiques dynamiques du flux de données. Une fois que le modèle a été appris, il peut être exploité pour déduire de nouvelles instances de flux de données dans une phase de synthèse [19]. Ci-dessous, nous rappelons les bases de la modélisation par HMMs afin d'introduire les notations nécessaires. Une présentation détaillée peut être trouvée dans [102].

Plus précisément, un HMM est basé sur une chaîne de Markov de premier ordre. Il comprend un nombre fini d'états dont la séquence obéit aux propriétés markovienne (nous considérons les modèles de Markov d'ordre 1 seulement), ce qui signifie que l'état à l'instant t est indépendant de tous les états antérieurs à l'exception de l'état au temps. La propriété de Markov permet de restreindre le calcul des probabilités de transition d'un état à un autre.

Une fonction de densité de probabilité (pdf) est associée à chacun des états, c’est habituellement une distribution gaussienne ou un mélange de distributions gaussiennes. Un tel paramétrage des HMMs vient de la seconde hypothèse appliquée au modèle: l'observation à l'instant, t, est supposée indépendante de toutes les autres observations et de tous les états étant donné l'état à l'instant, t.

Ces deux propriétés des HMMs ont rendu ces modèles très populaires car ils permettent d’utiliser algorithmes très efficaces et simples. Pourtant, ces hypothèses ne sont jamais satisfaites dans la pratique avec des données réelles. Du point de vue de la synthèse, ces hypothèses sont beaucoup trop fortes et se traduisent par une limitation de l’expressivité pour générer des trajectoires réalistes [73]. Un certain nombre de chercheurs ont contribué à surmonter un tel inconvénient, à des fins de reconnaissance en introduisant les HMMs segmentaires ou les HMMs trajectoire qui détendent l'hypothèse d'indépendance entre les observations successives [107, 38]. Parmi ceux-ci, quelques travaux peuvent être exploités pour synthétiser des trajectoires plus réalistes [121, 103], nous nous appuierons sur ces derniers ici.

4. Synthèse d’animation de parole

De nombreux travaux ont étudié les relations entre différentes modalités d’expression. Ces modalités sont manipulées afin de produire un message unifié correspondant à l’intention du locuteur. Elles sont reliées et synchronisés avec chacune d’entre elles. Les sourcils et les mouvements de tête, en tant que prosodie visuelle, sont souvent utilisés pour compléter l’information verbale. Notre travail se focalise sur la réalisation d’un modèle statistique de génération de mouvements de têtes et de sourcils à partir du signal de parole. Ce modèle peut être entrainé sur des échantillons d’apprentissage et peut alors synthétiser des animations à partir des caractéristiques du discours. Etant donné les avantages des modèles HMM pour représenter des flux de données, nous avons choisi d’utiliser un tel modèle pour notre travail.

4.1 Modèle contextuel

Dans les HMMs contextuels, les distributions de probabilité gaussienne sont définies comme étant fonctions de variables externes (ou contextuelles). Dans le cadre de la reconnaissance de parole, ces variables peuvent être la morphologie, le genre où l’âge comme dans les travaux [121, 103]. Inspiré par ces travaux, nous proposons un nouveau modèle que nous appellerons Fully Parameterized Hidden Markov Model (FPHMM). Un FPHMM est une extension d’un HMM contextuel où en plus des moyennes et matrices de covariances, tel que dans [121, 103]. Les probabilités de transition et la distribution de l’état initial sont également paramétrées (utilisant un ensemble de variables externes) et dépendent des variables contextuelles au lieu d’avoir une valeur fixe.

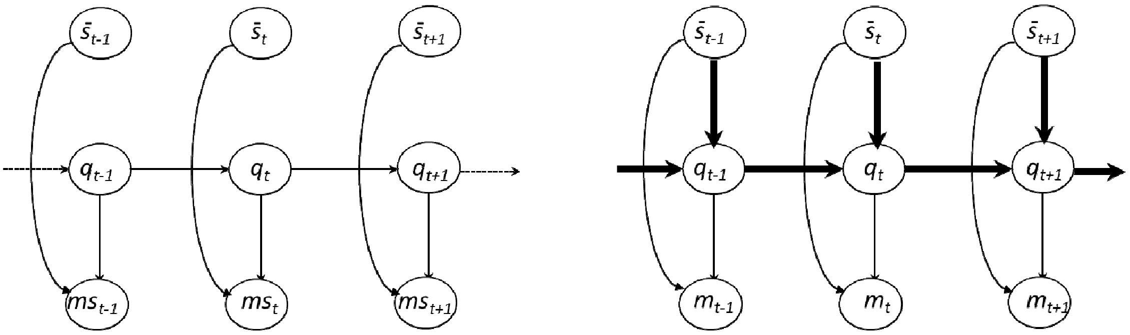


Figure 4: Représentation d'un CHMM (à gauche) et d'un FPHMM (à droite) en tant que Réseau Bayésien Dynamique pour générer une séquence de mouvements à partir d'une séquence de paramètres de parole. Alors qu'un CHMM utilise les paramètres de la parole pour modifier les paramètres, un FPHMM utilise en plus ces paramètres pour configurer les transitions entre les états.

4.2 Modèle Contextuel pour synthèse d'animation

4.2.1 CHMM

Afin de construire un système *speech-to-motion*, on peut apprendre un CHMM avec les paramètres du discours comme variables contextuelles (et dynamiques). Une fois qu'un tel modèle a été entrainé, on peut définir un CHMM pour le discours seulement en ignorant les paramètres du mouvement. On peut également utiliser le signal du discours pour définir un CHMM pour le mouvement, modifiant les paramètres de celui-ci grâce au flux du discours. En réalité, il s'agit d'un CHMM où les paramètres varient dans le temps. Lors de l'étape de synthèse, les paramètres du discours sont d'abord interprétés afin de trouver la séquence d'état la plus probable, puis nous utilisons la «*Single Method*» pour synthétiser une trajectoire à travers cette séquence d'état.

4.2.2 FPHMM

Dans notre travail, un FPHMM est utilisé pour synthétiser le flux de mouvement à partir du flux du discours comme suit. Premièrement, nous apprenons un FPHMM qui considère les paramètres du discours comme variables contextuelles et qui produit des paramètres du mouvement. Puis, lors de l'étape d’entrainement, les flux de mouvements et de discours sont tous les deux utilisés pour entrainer le FPHMM. Lors de l’étape de synthèse du mouvement (c.à.d. synthèse de l’animation), seulement le flux du discours est connu. Il est utilisé pour calculer les probabilités des transitions et les distributions de probabilités dans les états en fonction du temps. Une fois que tous les paramètres du modèle sont configurés, on peut calculer la séquence d’états la plus probable, ou on peut déduire la distribution de probabilités sur toutes les séquences d’états. Finalement, à partir de lé séquence la plus probable ou des distributions sur l’ensemble des séquences, on peut générer une trajectoire.

4.2.3 Combinaison des modèles FPHMMs et CRFs

Nous avons étudié la combinaison du modèle HMM entièrement paramétré et du modèle Conditional Random Fields (CRFs) [65]. Nous appelons ce modèle hybride FPHMM-CRF. Il est inspiré du travail réalisé par [71].

Pour l’apprentissage, nous apprenons d'abord un FPHMM. Dans cette étape, nous déterminons la séquence d'état la plus probable en utilisant uniquement des caractéristiques de parole dans le FPHMM appris. Ensuite, le CRF est appris en utilisant l'ensemble des séquences d’état comme ensemble de données d’apprentissage.

Pour la synthèse de l’animation, le CRF prend en entrée un signal de parole pour obtenir une distribution des probabilités sur toutes les séquences d'états. De plus, les caractéristiques de parole sont utilisées pour définir les distributions des probabilités de transition entre les états du FPHMM. Enfin, étant donné la séquence d'états la plus probable (la distribution sur les séquences de l'état), on peut estimer une séquence d’observation (intégration) suivant toutes les séquences d’état des séquences d’observation.

5. Synthèse d’animation à partir du signal audio du rire

Le rire apparaît souvent dans la communication humaine. Même si le rire est un signal de communication important dans l'interaction humaine, il n’a commencé à être étudié que depuis la fin des années 1990 [105]. Avant cela, très peu d'études en psychologie (par exemple, [34]) décrivaient le comportement associé au rire et ses fonctions communicatives. Le rire est fortement lié à des émotions positives et à la bonne humeur [106].

Toutefois le rire n'est pas toujours lié à la gaieté. Des études ont rapporté 23 différents types de rire [58], comme le rire hilare, hystérique embarrassé, désespéré, ou méprisant. Le rire est donc lié à divers états émotionnels. Ses expressions, au niveau acoustique et comportemental, varient en conséquence.

Dans notre travail, nous nous concentrons sur le rire hilare qui est le rire faisant suite à un événement drôle, et qui est déclenché par des stimuli amusants et positifs (par exemple une blague). Le rire hilare est un des 23 types de rire définis par [58]. La morphologie du rire comprend les expressions faciales, les mouvements du corps et des vocalisations [105].

Notre objectif est de développer un Agent Conversationnel Animé (ACA) capable de rire. Pour cela, nous devons proposer des approches pour synthétiser les signaux multimodaux du rire. Niewiadomski et Pelachaud [87] indiquent que la synchronisation entre toutes les modalités est cruciale pour la synthèse de l'animation du rire. Les humains sont très habiles dans l’observation des comportements non verbaux, et ils détectent la moindre incongruité dans les animations multimodales synthétisées. Nous avons développé des modèles statistiques pour la synthèse des comportements multimodaux de rire. Le modèle statistique est d'abord appris sur des données humaines, puis est utilisé pour synthétiser des animations multimodales. Nous présentons successivement deux approches que nous avons conçues et évaluées.

Le premier système est une preuve de concept qui combine plusieurs briques de base, traitant toutes les données d'animation en sortie de système. Le deuxième système est un système élaboré focalisé sur la synthèse des mouvements de la tête et du torse.

5.1 Premier système de synthèse d'animation à partir du signal audio du rire

Pour la synthèse d'animation du rire, plusieurs types d'entrées sont considérés pour notre système, compris dans le jeu des données disponibles (celui des bases de données). Il s'agit d'un choix raisonnable des entrées, mais aussi d'une contrainte qui vient du jeu des données à disposition.

Nous avons utilisé des caractéristiques prosodiques en entrée, comme dans les travaux précédents [14, 29, 16, 62, 48, 39, 74, 31, 66, 109, 19, 18, 56, 70, 71, 73, 78]. Nous avons aussi utilisé une autre information de plus haut niveau sur la réalisation du rire. Il s'agit de la séquence de pseudo-phonèmes du rire, qui exprime la séquence des unités élémentaires du rire. C'est l'équivalent pour le rire de la séquence de phonèmes pour la parole. Pour aucun pseudo-phonème du rire, on dispose d'informations sur son intensité et sa durée.

Le jeu de données que nous avons utilisé est une base de donnée existante, appelé AVLC [114]. Cette base de données comprend les signaux audio, les mouvements de tête et des expressions faciales. Cela nous permet d'exploiter des modèles statistiques. L'inconvénient de cette base de données est le manque de mouvements de torse et des épaules. Pour pallier cet inconvénient, nous nous sommes orientés vers une méthode déterministe.

Notre premier système s'appuie sur plusieurs briques de base. Il est composé de trois modules de synthèse. Le premier module est la synthèse des mouvements des lèvres et de la mâchoire. Ce module prend en entrée les caractéristiques prosodiques et les pseudo-phonèmes du rire. Le deuxième module est la synthèse des mouvements de la tête et des autres expressions faciales, tel que les sourcils, les paupières et les joues. Ce module prend en entrée les pseudo-phonèmes du rire. Le troisième module est la synthèse des mouvements du torse et des épaules. Comme nous ne disposons pas de données humaines pour ces mouvements, ce module prend en entrée une autre information, les mouvements synthétisés de la tête. On peut trouver tous les détails sur les entrées de ces trois modules sur la figure 5. Dans la section suivante, nous présentons ces trois modules successifs.

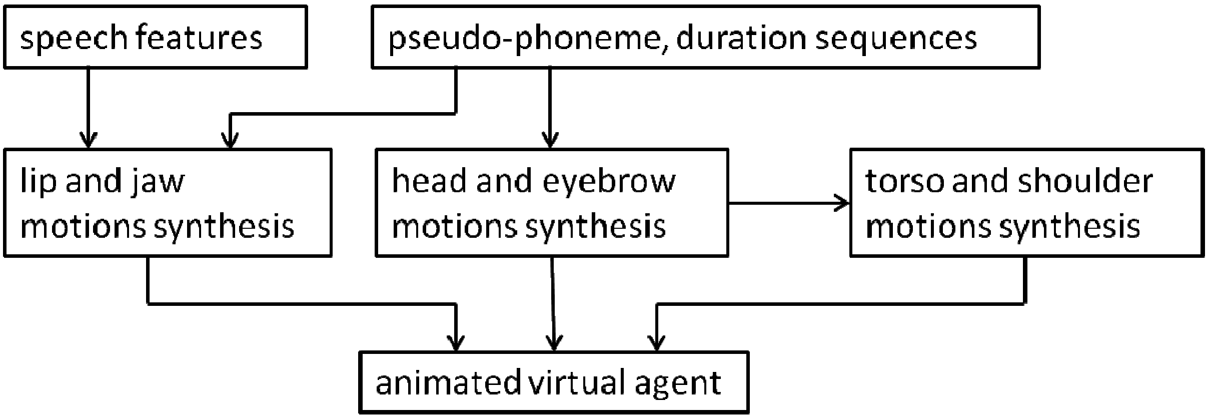

speech features
pseudo-phoneme, duration sequences
lip and jaw motions synthesis
head and eyebrow motions synthesis
torso and shoulder motions synthesis
animated virtual agent

Figure 5 Trois modules de synthèse.

5.1.1 Synthèse des mouvements des lèvres et de la mâchoire

Pour la synthèse des mouvements des lèvres et de la mâchoire, notre travail s'appuie sur l'hypothèse que ces mouvements ne dépendent que des caractéristiques prosodiques et des pseudo-phonèmes du rire. On utilise des fonctions linéaires pour modéliser la relation entre la prosodie et les mouvements. Les mouvements des lèvres et de la mâchoire sont définis par 23 paramètres. Nous utilisons donc 23 fonctions linéaires pour estimer les valeurs de ces paramètres. Les coefficients des fonctions linéaires sont appris à partir de la base de données par le critère des moindres carrés.

5.1.2 Synthèse des mouvements de la tête et des expressions faciales

Les mouvements synthétisés sont ici ceux de la tête et des expressions faciales, tels que les sourcils, les paupières et les joues. La synthèse de ces mouvements est indépendante et basée sur la même méthode.

La Figure 6 montre la procédure de la synthèse des mouvements de la tête. Pour chaque pseudo-phonème du rire, on a un ensemble des mouvements issu de la base de données d'apprentissage. On sélectionne donc un de ces mouvements pour chaque pseudo phonème de rire des entrées. La sélection de la séquence des mouvements de tête correspondant à la séquence de pseudo phonème de rire peut être vue comme le choix de la meilleure séquence. Nous expliquons ce choix de correspondance ci-après.

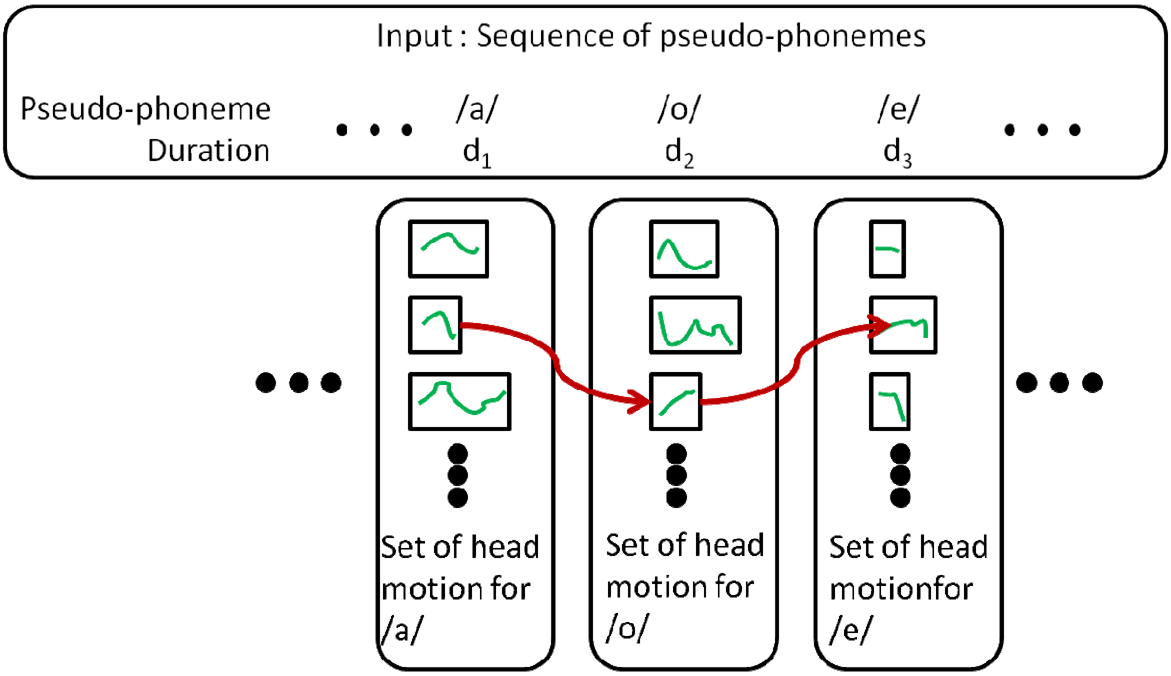

Input : Sequence of pseudo-phonemes
Pseudo-phoneme
Duration
/a/
d_1
/o/
d_2
/e/
d_3
Set of head motion for /a/
Set of head motion for /o/
Set of head motionfor /e/

Figure 6 Synthèse des mouvements de la tête

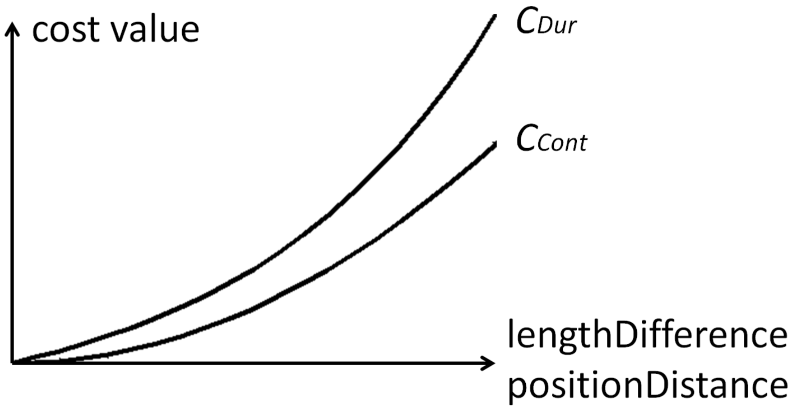


Figure 7 Deux critères pour la sélection des mouvements

La qualité de la correspondance s'appuie sur deux critères. Le premier critère est le coût de duré, qui est le mesure de l'adéquation de duré entre les mouvements sélectionnés et pseudo-phonème de rire. Le deuxième critère est la qualité d'enchaînement des mouvements des pseudos phonèmes de rire. Il est l'écart de distance entre la position à la fin du mouvement précédent et celle au début du mouvement suivant. Ces deux critères sont montrés sur la figure 7. La qualité de correspondance est calculée par la somme pondérée de ces deux critères. Le paramètre de combinaison de ces deux critères sont déterminés à la main.

5.1.3 La synthèse des mouvements du torse et des épaules

L'inconvénient de la base de données que l'on utilise est le manque de mouvements du torse et des épaules. Dans la mesure où les mouvements du torse et des épaules sont importants pour les animations du rire, on développe donc une méthode déterministe, appelé *PD controller*.

PD controller prend en entrée les mouvements synthétisés de la tête et calcule en sortie les mouvements du torse et des épaules. Cette synthèse est basée sur l'hypothèse que les mouvements du torse et des épaules suivent bien les mouvements de la tête. Par exemple, quand la tête penche vers l'avant, le torse penche vers l'avant aussi.

Pour être plus précis dans le *PD controller*, la sortie est toujours lissée ainsi que l'entrée. La sortie dépend de l'entrée courante et de la sortie précédente. La figure 8 montre un exemple d'animations produite avec *PD controller*, où la courbe bleue est l'entrée, l'animation de la tête et la rouge la sortie, l'animation du torse.

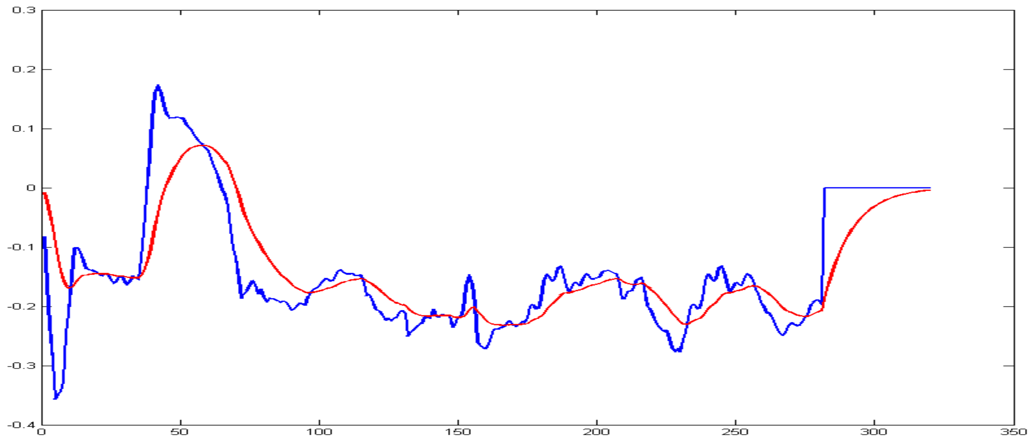


Figure 8 Exemple des animations de la tête en entrée et du torse en sortie avec le *PD controller*. La courbe bleue est la trajectoire du torse, et la rouge est celle de la tête.

5.2 Le second système de la synthèse d’animation à partir de signal audio de rire

Pour synthétiser les mouvements de la tête et du torse avec un modèle statistique, on enregistre une nouvelle base de données, qui comprend les signaux audio et les mouvements de la tête et du torse.

Les mouvements de la tête et du torse sont connus comme des mouvements de tremblements. Pour modéliser tel mouvements, nous avons exploité un modèle Markovien particulier, appelé Loop HMM (LHMM). 6 LHMMs sont développés respectivement pour les 6 dimensions caractéristiques des mouvements (3 rotations de la tête and 3 rotations du torse). La figure 9 est une représentation de la topologie de LHMM.

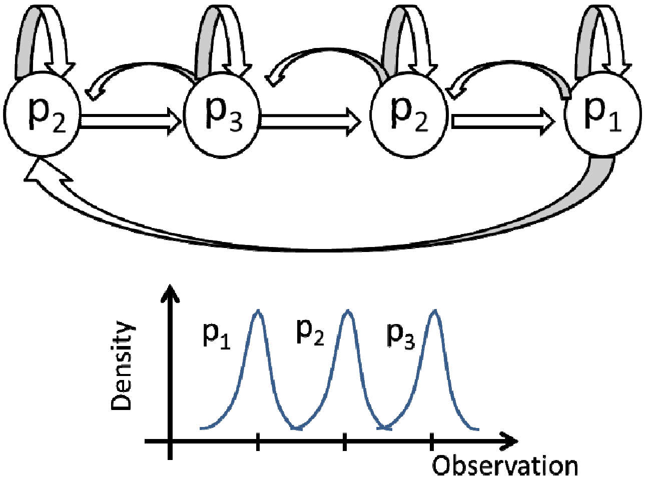


Figure 9 Représentation de la topologie de LHMM.

LHMM est un modèle markovien de gauche à droite dans lequel les retours en arrières sont possibles. Les probabilités de transition d'état sont définies de tel façon que le processus va grosso modo de la gauche vers la droite. Le cycle correspond à un mouvement d'aller-retour. La distribution des probabilités associées aux états est définie comme une distribution gaussienne dont les moyennes sont reparties sur le domaine de valeur de caractéristique modélisé. En parcourant ce modèle de gauche à droite, on décrit un mouvement en forme de vague similaire au tremblement. La figure 10 montre un exemple de l'animation synthétisée par LHMM.

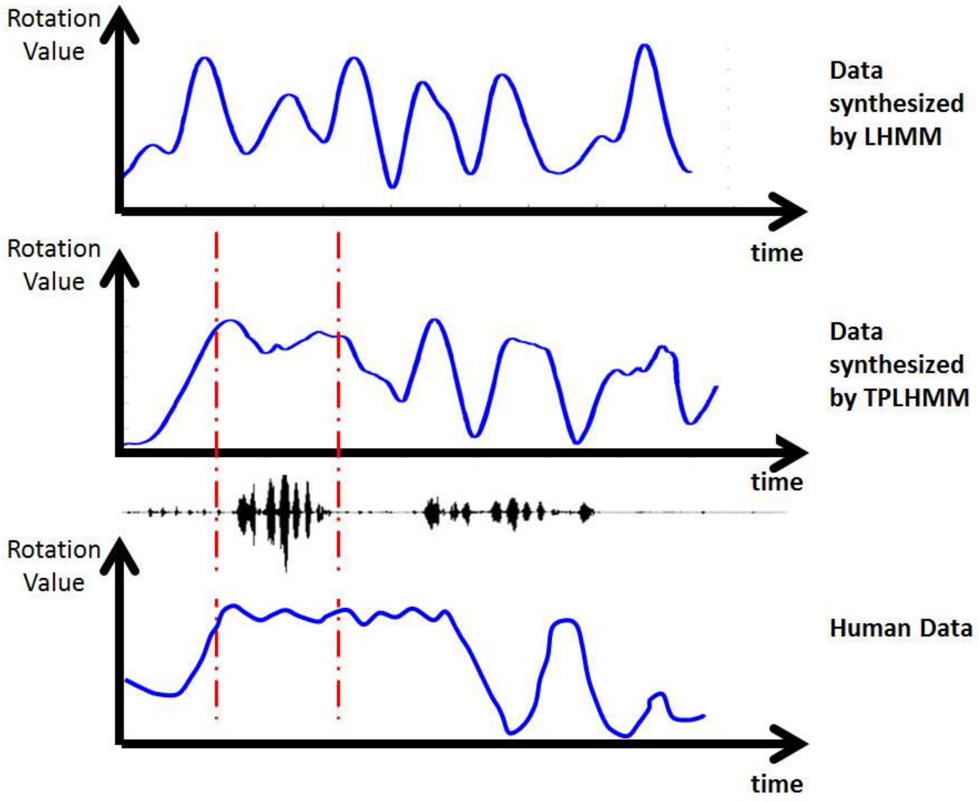


Figure 10 Comparaison des trajectoires à partir des différents modèles

L'autre idée importante est le paramétrage des probabilités de transition d'état selon les caractéristiques de prosodie. C'est-à-dire que les probabilités de transition d'état, à un instant t, dépendent de les signaux de prosodie à un instant donné, t. Cela permet d'augmenter le réalisme et la variabilité de l'animation synthétisée, mais aussi le lien entre l'audio et l'animation.

La figure 10 est un exemple de l'animation. La trajectoire du bas est un exemple humain; celle du haut est synthétisée avec LHMM; celle du milieu est la trajectoire synthétisée avec LHMM intégrant des transitions paramétrées, appelé TPLHMM. La trajectoire avec LHMM sont similaires entre elles. LHMM ne prend pas l'audio en entrée, la trajectoire se ressemble donc beaucoup. La trajectoire avec TPLHMM est assez marquée par le signal audio et dépend des fichiers audio donnés en entrée.

Comme indiqué précédemment, on cherche la synthèse des modalités de la tête et du torse. L'idée simple est de construire le modèle par modalité, l'un pour la tête et l'autre pour le torse. En fait, nous proposons un modèle joint pour modéliser les deux modalités des mouvements à la fois. Cela permet de modéliser le lien entre deux modalités des mouvements et d'augmenter la qualité de l'animation synthétisée.

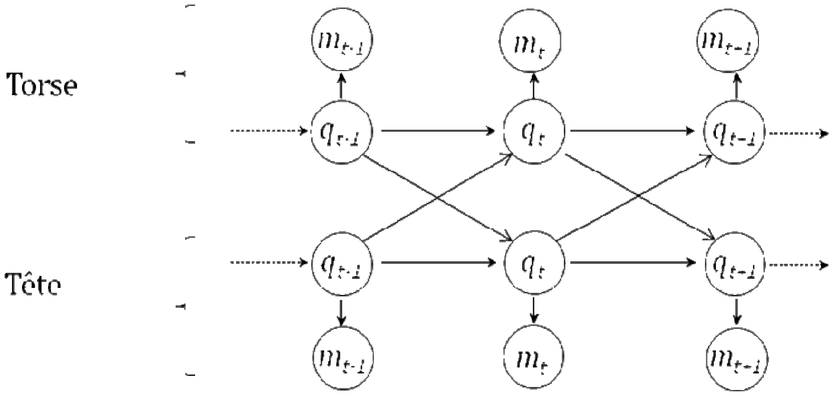


Figure 11 Topologie de LTPHMM couplé pour la synthèse des mouvements

Nous nous sommes inspirés du HMM couplé. Nous en avons une variante en combinant les modèles appelés LTPLHMM couplé (CTPLHMM). La figure 11 montre une représentation de topologie de CTPLHMM. La position courant de la tête dépend non seulement de la position précédente de la tête mais aussi la position précédente du torse.

6. Conclusion de la Synthèse d'Animation à Partir de Parole

Dans notre premier travail (synthèse d'animation de la parole), notre but est de construire des modèles pour la synthèse d'animation de la parole. Tout d'abord, nous nous sommes concentrés sur la synthèse d'animation des sourcils, puis nous avons généralisé cette synthèse d'animation des sourcils à la synthèse des animations des sourcils et de la tête en même temps.

Dans la première étape, synthèse de l'animation des sourcils, les modèles contextuels existants basés sur des HMM, appelés HMM contextuels [121, 103] dans cette thèse, sont utilisés comme générateurs d'animation. De plus, nous avons étendu ces HMM contextuels en HMM entièrement paramétrés (*Fully Parameterized HMM, FPHMM*). Puis ces FPHMM sont utilisés comme générateurs d'animation. Nous avons effectué des évaluations où nous avons comparé les résultats produits par différents types de HMM. Ces évaluations objectives montrent que les HMM contextuels et les FPHMM surpassent les HMM standards [56, 19] et que les FPHMM sont meilleurs que les HMM contextuels.

Dans la deuxième étape, synthèse d’animation des sourcils et de la tête, les FPHMM sont généralisées en générateurs d'animation des sourcils et de la tête. Dans cette étape, nous étudions si l'intégration de la relation entre les mouvements des sourcils et de la tête augmente la qualité des signaux synthétisés. Pour répondre à cette question, nous comparons les résultats des deux modèles. Les animations synthétisées sont évaluées en utilisant des méthodes objectives et subjectives. Les résultats de ces évaluations montrent que la relation apprise par notre modèle, entre les mouvements des sourcils et de la tête, peut être utilisée pour améliorer la qualité des animations synthétisées.

Dans le travail de synthèse de l'animation de la parole, nous avons proposé une approche statistique pour générer les mouvements de la tête et des sourcils pour un agent virtuel à partir de la parole. Le FPHMM est utilisé pour capturer la correspondance directe entre l'audio et l'information visuel. Le FPHMM formé permet de définir l'animation visuelle en fonction du signal de parole. L'évaluation objective montre qu'en considérant les mouvements de sourcils et de tête simultanément, la précision de l'animation résultante est améliorée. Il confirme également que les mouvements des sourcils et de la tête ne sont pas indépendants l'un de l'autre mais, au contraire, sont liés ; les signaux multimodaux renforcent le sens de la communication. L'évaluation subjective montre que notre modèle proposé améliore la perception de l'animation de l'agent virtuel au niveau de l'expressivité émotionnelle.

7. Conclusion de la Synthèse de l’Animation du Rire

Dans notre deuxième travail (synthèse de l'animation de rire), nous avons construit des générateurs d’animation du rire, qui inclut la mâchoire, les lèvres, les joues, les paupières, les sourcils, la tête, les épaules et le torse. A notre connaissance, nos générateurs d'animation du rire sont la première tentative de synthèse de l'animation du corps entier pour le rire. Des modèles de régression linéaires sont utilisés pour animer les lèvres et la mâchoire, où les mouvements des lèvres et de la mâchoire dépendent des phonèmes du rire et des caractéristiques de la prosodie. Une méthode de concaténation est proposée pour générer l’animation des joues, des paupières, des sourcils et de la tête, où l’intensité et la durée des phonèmes sont utilisées pour sélectionner des segments de mouvements humains. Le modèle *PD Control* est appliqué à la synthèse de l’animation des épaules et du torse, qui sont synchronisés les uns aux autres et dirigés par les mouvements de la tête. Nous avons évalué notre modèle pour vérifier la façon dont le rire de l'agent est perçu en racontant ou en écoutant d'une blague. Nous avons réalisé une évaluation où nous avons comparé deux conditions : un agent qui sourit et un agent qui rit. Les résultats montrent que les participants préféraient interagir avec un agent riant plutôt qu’un agent souriant.

De plus, des FPHMM couplés sont utilisées comme générateurs d’animations de la tête et du torse, où les animations synthétisées sont capables de reproduire des mouvements de secousses ou de tremblements. Dans cette méthode, les animations synthétisées sont synchronisées aux caractéristiques de la prosodie. Ils sont également fortement corrélés avec les uns les autres. Une évaluation subjective a été menée pour valider les générateurs d'animation du rire proposés. Nous avons comparé les résultats provenant des FPHMM couplés et ceux non couplés. Nous avons aussi fait une comparaison avec des mouvements humains. Les résultats montrent que les participants ont trouvé que les animations synthétisées étaient de meilleure qualité pour les FPHMM couplés.

Contents

Chapter 1

Introduction

Human communication involves audio and visual signals. Audio signals correspond to human speech, spoken language, prosody, para-linguistic features, etc; visual signals contain facial expression, body motion, arm and hand gestures, etc. Humans are very skilled at inferring and aligning various subtle motions to satisfy communicative intention including emotion. They are also capable of reading and decoding such complex behaviors [99]. Communicative signals can be used by humans to infer and to express an intention, an emotional state and personality traits. For example, when saying "hello" to start a conversation, human may first turn his/her head while smiling to the interlocutor and then say "hello" with a head nod and leaning forward.

Embodied conversational agents, ECAs, are a kind of Human-Computer Interface. They are entities endowed with communicative and expressive capabilities; they have human-like appearance. An ECA is capable of communicating with human or another ECA. It is often installed on a device endowed with camera, screen, speaker and microphone, which are used to receive and convey the audio and visual signals. Hence, it can listen to and speak to its interlocutor (e.g. a user); it can also watch interlocutor's facial expression and display gestures to the interlocutor. Reeves and Nass, in their book "The Media Equation" [104] stated that users, when interacting with human-like agent, tend to apply similar protocols than when interacting with other humans. Hence, an ECA should communicate with humans in a human-like manner. It means that an ECA should be capable of speaking and displaying natural behaviors as humans do to convey information.

In recent years, ECA has become increasingly popular in several applications of human-computer interactions, such as social coaching. For example, in the Europe project TARDIS (http://public.tardis-project.eu/) ECA is used as a virtual recruiter. The ECA interacts with a user (interviewee) during job interview training. The virtual recruiter can decide autonomously how to conduct the interview; it can select which social attitudes to express toward the interviewee. Figure 1.1 shows two images of an ECA with friendly and unfriendly attitudes. Figure 1.2 shows screenshots of interactions between the virtual recruiter and the interviewee. In such an application, ECA is evaluated and perceived by humans (users), so ECA with expressive quality is crucial for engaging users in such an application.

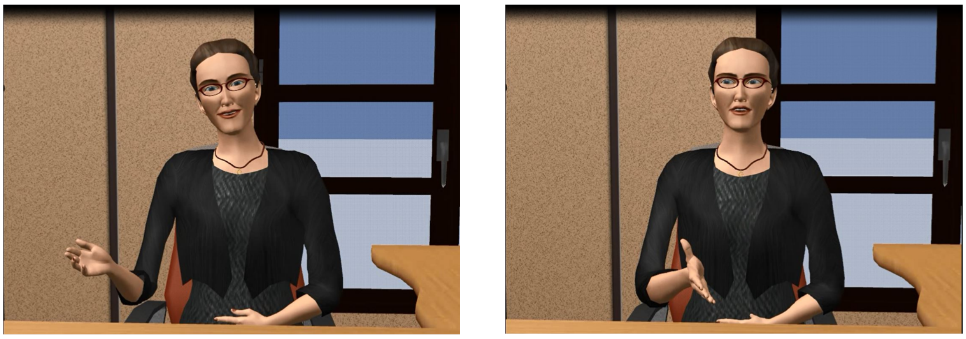


Figure 1.1: Examples of ECA with friendly (left) and unfriendly (right) attitudes in job interview [20]

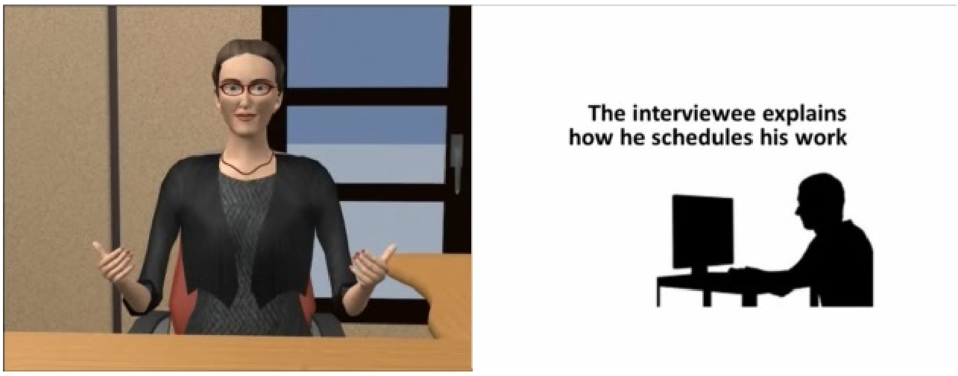


Figure 1.2: Screenshots of the simulated job interview [20]. The left is the virtual recruiter; the right is the response of the interviewee.

Many studies from the psychology literature [44, 117, 59] have been conducted on characterizing the multimodal expressions of emotions. Results from these studies are exploited by computer scientists to model emotional behaviors of ECAs [32].

These models describe not only the facial expressions of emotions [12] but also the body posture and quality of movements [28]. They also looked as rendering aspects such as wrinkles and skin coloring [36].

To simulate communicative behaviors for ECAs, procedural models have been first elaborated [23, 95, 22, 8]. They are based on a set of rules extracting from the literature in human communication or from the analysis of multimodal corpora. However, since human behaviors arise from and may be influenced by various factors, such as emotion, personality, gender, physiological state and social context [68], it is extremely difficult to define a large set of rules to fully capture the role of these factors onto human behaviors, even after decades of studies in psychology. Besides, multiple rules could result in synthesizing conflicting expressions [96]. Moreover the animations obtained from these models suffer from lack of naturalness. The virtual characters lack liveliness and dynamism, phenomena that are not easily captured by rules. Data-driven approaches have been elaborated to overcome low animation quality. These models rely on large database of expressive human motions. They learn the correlation between speech and face/body modalities [14, 123, 39, 31, 19, 18, 56, 78, 70, 71, 26]. While rule-based models allow rendering the semantic values of behaviors, data-driven models capture expressiveness of behaviors.

Our research lies in the data-driven approach. It focuses on generating the multimodal behaviors for a virtual character speaking with different emotions. It is also concerned with simulating laughing behavior on an ECA. .

1.1 Motivation

Humans communicate through verbal and nonverbal means. Their behaviors follow a very complex process. Communicative behaviors are polysemic; that is, a same behavior may convey several meanings. For example, a head nod can convey agreement, mark an emphasis, or be a back-channel signal. The meaning attached to a signal is disambiguated from the spoken context. When addressing multimodal behaviors it is important to consider that human expression is a multi-modal process and that multi-modal expressions are linked and synchronized to each other.

Multi-modal Communication: Humans communicate not only semantic content but also emotion, intention and desires, which are expressed through audio and visual means such as speech, facial expressions, gesture, etc. McGurk and MacDonald [81] emphasize the importance of both audio and visual signals in human-human communication. In social and psychological studies [33, 7], these means are classified into two separate communication channels: explicit and implicit channels, which are respectively linked to what is said and to how to say. Explicit channel contains communicative means involving linguistic and communicated content, such as spoken words, co-articulation behaviors (e.g. lip motions), emblem gestures (e.g., Ok-gesture, waving goodbye, head nod expressing Yes), etc; implicit channel contains other communicative means involving speaker's emotional state, intention and desire. Implicit channel is built by various multi modal means such as paraverbal (i.e., prosody) and non-verbal (i.e. gestures). Speech prosody is an important mean to express emotion, emphasize a word, mark an utterance as a question, etc. Many works [111, 5, 122, 67] have exploited speech prosody to recognize speaker's emotion. To build their multimodal synthesis generator, Busso et al. [18] rely on the property that head and eybrows motion are linked to speaker's communicative and emotional state.

Multi-modalities Relationship: Several works have studied the relationship between various multi-modalities of expression. The multimodal behaviors are intricately integrated to produce a unifying message according to speaker's intention. They are linked and synchronized with each other. Kendon and Key [60] and McNeill[82] indicated that a body gesture is not only strongly related to the uttered content but is also synchronized with the flow of speech. Several studies [52, 53] have shown that eyebrow and head motions are tightly coupled with speech prosody. Bolinger [11] found a strong correlation between the raise of speech pitch and of eyebrow movements. Rising fundamental frequency of speech is accompanied with rising or falling eyebrow [24], while pitch accent often co-occurs with raising eyebrows [50]. Munhall et al. [84] reported that head movement improves the auditory speech perception.

As introduced above, human communication is a multi-modal process, where several communicative modalities are related to each other. In particular, human behaviors are always synchronized to human spoken speech. Our work is to develop animation generators which allow ECA to display autonomously appropriate motions as humans do when speaking or laughing. The animation generators take audio signals of speech or laughter as input.

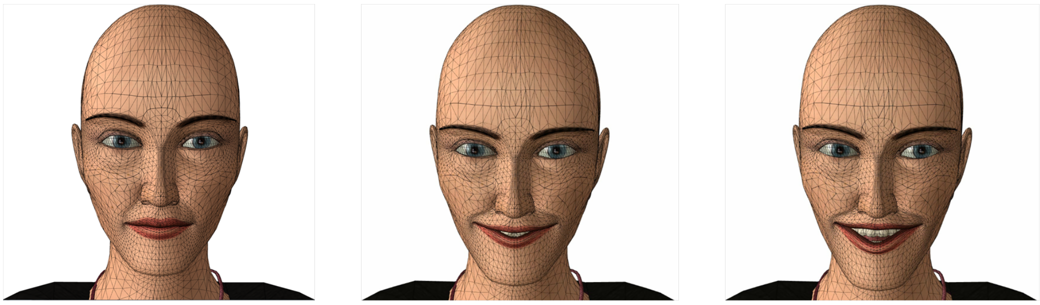


Figure 1.3: Face models: (a)Neutral; (b)Smiling; (c)Laughing.

1.2 Animation Synthesis for Virtual Characters

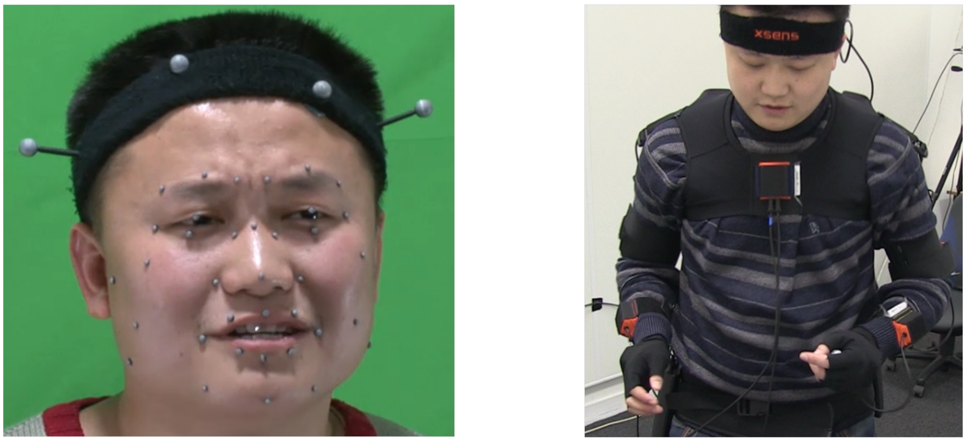


Figure 1.4: Actors equipped with motion capture technology. The left image shows the actor with motion capture sensors on his face. The right image shows the actor with motion capture sensors on his body.

Human (User) tends to take human as reference when interacting with ECA in Human-Computer Interface [104], so ECA should be able to exhibit various human-like aspects including communicative and emotional behaviors. In the last three decades, amount of works have focused on improving the perceived quality of virtual human-like characters.

The first attempt on building human-like appearance is performed by Parke [93]. He was the first to create a 3D facial model. 250 polygons involving 400 vertices are defined to shape a face surface. The face in this first model looks very sketchy. With more vertices, a face model looks more like a human face. Mesh face model can be coloured by human texture information to display human-like appearance. Figure 1.3 shows three images from a mesh face model covered by texture information [94, 9]. Similarly, body model is structured based on a skeleton model and is composed of rotation joints. An example of skeleton model is shown in the left image of Figure 1.5 [94, 9].

In face models, vertices can be controlled to deform face shape to simulate human expression. Once vertices move, texture information deforms according to the values of the related vertices. Figure 1.3 shows an example of deformation of texture information along with face deformation to simulate muscular contraction. Traditional animation systems rely on key-frames as old 2D cartoons movies did. A key-frame, also called a key-pose, consists in defining all the vertices of the face model. The animation is obtained by interpolating between the key-frames. Different interpolation techniques have been proposed. To name a few, we have linear interpolation [98], cosine interpolation [118], bilinear interpolation [92], etc. Similarly body skeleton model can be controlled at each key-frame, by defining values of joint angles. Figure 1.5 shows a key-frame of a body model that is configured by the values of joint angles. For simiplicity, we call the vertices in facial mesh structure and joints in body skeleton model, controllable points. Once the values of controllable points are known at each key-frame, one can animate a virtual character by applying interpolation techniques to infer values of the controllable points.

To animate a virtual character, one needs to specify the position of the controllable points. One of the most labor intensive methods (but which produces high quality result) is to specify the controllable points manually at each key-frame; this is commonly done by animators. Another popular method is to use motion capture systems, called MoCap method [79]. In MoCap method, actors are equipped with motion capture sensors on their face and body. The motions of the sensors are captured. Then the captured motion data is directly used to define the values of the controllable points. The MoCap method is being widely used in video games and character movies. Figure 1.4 shows two images of actors equipped with motion capture sensors on their face and on body. While MoCap method allows capturing subtleness and expressivity of human motions, it is an expensive method in time and cost. Equipment and actors can be very expensive. Capturing, filtering and segmenting the data is viewed to reproduce it on virtual characters is a time-consuming task. MoCap data needs to be recorded ahead of time. Moreover the captured data can only be used to reproduce the particular scenario; it cannot be so easily deformed to simulate any other new scenarios. So, ECA cannot take MoCap method to generate its animations. Indeed ECAs are interactive characters. As such their behaviors and the corresponding animation need to be computed in real-time on the fly. Another method, called performance-driven method, is developed to capture human expression from camera or video through various tracking techniques [13, 72]. Such animation method gives very good result when the virtual character is made to reproduced tracked motion of a human actor. Similarly to MoCap method, the tracked data cannot be used to animate autonomous entities such as, ECAs.

In several ECA models [23, 116, 94, 9, 68, 35], the animation of ECAs is done by specifying the communicative intention and emotional state the ECA should display through its multimodal behaviors (ie its facial and body parameters). We call such a method an intention-driven method. In such a method, specific intention is associated to specific behaviors described by the values of the facial and body parameters. For example, the emotional state of surprise can be associated to the facial parameters defining raising eyebrows and mouth opening. A library of multimodal expression is defined, where each intention is associated to some specific facial expression and body behaviors.

Recently, speech-driven methods have been developed to animate virtual characters. This method is based on the fact that it exists a strong correlation between human speech signals and motions. Once such a correlation has been defined, one can infer animations from speech signals. For example, when spoken phoneme sequence is known or recognised from speech audio or text information, then it is used to determine lip shape associated to spoken speech.

Challenge of Animation Synthesis: As we have seen above, there exist various methods to animate virtual characters. Their behaviors can be specified by hand or can duplicate human expression using MoCap method or performance-based method. As explained, the cost of these methods can be high in time and cost. Though MoCap and performance-based methods are appropriate to animate ECA for displaying human-like animation, ECA behaviors are limited to the recorded library gathered by Mocap or through tracked samples. On the other hand, procedural methods relying on a lexicon of multimodal behavior and data-driven approach allow generating multimodal behaviors on the fly. The intention-methods are often event-driven; that is they compute a behavior only when a given communicative function is specified. Such methods capture more precisely the semantic-emotional behaviors to communicate, while the synthesized motion lacks of naturalnesses and variety.

Computer graphic community has focused on the animation synthesis of the different human communicative modalities, such as lip, eyebrow, eye gaze, head, arm, hand, body, etc. Specially, lip animation has been widely studied in the last three decades [6, 29, 16, 14, 123, 62, 39, 48, 74]. Hence, audio speech signal and uttered content can be used to drive lip animation. Other works have also attempted to study the relation between speech prosody and human behaviors, such as head motion [19, 18, 78, 56, 66], eyebrow [78, 31] and gesture [70, 71, 26, 79, 27]. The mapping between prosody and behaviors is many-to-many instead of one-to-one as between uttered content and lip shape. So, how to model this many-to-many relationship remains a challenge in the virtual agent community.

1.3 Objectives

Our aim is to study and to develop human-like animation generators for speaking and laughing ECA. When an ECA is speaking or laughing, it is capable of displaying autonomously behaviors to enrich and complement the uttered speech and to convey qualitative information such as emotion. On the basis of the relationship linking speech prosody and multimodal behaviors, our animation generator takes as input human uttered audio signals and output multimodal behaviors.

We considered two types of audio signals as input: speech and laughter. These two signals are exploited separately in our work. Our aim is not to synthesize audio signals of speech and laughter. In our work, we focused on speech and laughter animation synthesis, which mainly involve visual prosody; other behaviors particularly linked to semantic value such as emblem and pointing behaviors are not considered.

Our work focuses on using statistical framework to capture the relationship between the input and the output signals; then this relationship is rendered into synthesized animation. In the training step, the statistical framework is trained based on joint features, which are composed of input and of output features. The relation between input and output signals can be captured and characterized by the parameters of the statistical framework. In the synthesis step, the trained framework is used to produce output signals from input signals. The relation captured in the training phase can be rendered into the output signals.

In our work, we proposed two animation synthesis generators:

1. *speaking ECA: speech signal is used to synthesize head and eyebrow motions. Pitch and energy are extracted from speech signal and are used as input features; 3 head rotations and 4 eyebrow parameters are used as output features.*
2. *laughing ECA: laughter audio signal is used to synthesize head, eyebrow, eyelid, cheek, lip, jaw, torso and shoulder motions. Pitch and energy are extracted from laughter audio signal and are used as input features; furthermore laughter phonetic transcription is extracted using the method proposed by Urbain et al. [115] as well as the laughter phoneme intensity and duration, which are all taken as input features.*

1.4 Contributions

Our contributions can be divided into two independent parts. First, we built models of facial animation synthesis for speaking ECA; secondly, we built models of animation synthesis for laughing ECA, including the animation of facial expression and upper torso. In these models, acoustic signals (speech or laughter sound) are taken as input. To achieve our goals, we exploited data-driven models to capture the relationship between input and output signals, which are based on human dataset; then these models are used as speech or laughter animation generators.

Speech Animation Synthesis In our work of speech animation synthesis, we introduce the use of the contextual Hidden Markov Models (HMMs) [121, 103] to synthesize animations. Previous works have used contextual HMMs only as recognition models. In contextual HMMs the *pdfs* (Gaussian probability distribution) are defined as a function of some external (or contextual) variables; these variables can be the morphology, the gender, the age, etc. In our work, the contextual features are the speech features. Furthermore, we have extended contextual HMMs to fully parameterized HMMs (FPHMMs). For FPHMMs, not only *pdfs* but also state transition probabilities are defined as a function of some external (or contextual) variables. Then FPHMMs are used to synthesize animation from speech signals, where speech features are used as external variables to define *pdfs* and state transition probabilities.

The proposed models are evaluated by calculating the distance between the trajectories of human motion and synthesized animation. The results show that our models are better than the reference models (standard HMMs). The standard HMMs have been used to synthesize animation in the previous works [19, 56, 18, 78]. Our model outperforms existing works. The objective evaluation study shows that considering simultaneously eyebrow and head motions increases the precision of the synthesized animation; this result confirms also that eyebrow and head motions are related to each other. On the other hand, the subjective evaluation shows that our proposed models enhance the human perception, although they cannot overcome the human data.

Laughter Animation Synthesis In our work of laughter animation synthesis, we built laughter animation generators. These generators involve many upper body features, namely: jaw, lip, cheek, eyelid, eyebrow, head, shoulder and torso. To the best of our knowledge, our models are the first attempt on building laughter animation generators for the full body laughter animation synthesis.

As the first step, we used three approaches to generate various animations based on an existing laughter human dataset, called AudioVisual LaughterCycle (AVLC) [114]. Linear regression models are used as lip and jaw animation generators, which take as input sequences of phonemes of laughter and prosody features of laughter sound. Cheek, eyelid, eyebrow and head animations are synthesized by concatenating human segmented motions, which takes as input laughter phoneme sequences and laughter sound intensity. Shoulder and torso are inferred by Proportional-Derivative (PD) Controller [1], which takes as input the synthesized head movements. These three approaches can synthesize their corresponding animations in real time. A subjective evaluation was conducted to evaluate these approaches. To furthermore investigate laughing upper body animations, we built a new laughter human dataset, where human laughter audio, head rotation and torso motions are recorded. A specific HMM, Loop HMM (LHMM), is used to capture

[1] Proportional-Derivative (PD) Controller is widely used in graphics simulation domain [85], which is a simple version of proportional-integral-derivative controller (PID) in classical mechanics

and synthesize the motion patterns (shaking/trembling) of head and torso. Then the parameterized state Transition Probabilities are used in LHMM; it is called TPLHMM. TPLHMM is used to capture and render the relation between the dynamical characteristics of motions and prosody features. Finally, we combine the TPLHMMs and the Coupled HMMs [15], called Coupled TPLHMMs (CTPLHMMs) to capture and to render the relation between the movements of head and of torso. The objective evaluation shows that the proposed models can generate laughing motion patterns. The subjective evaluation shows that our models are able to capture the dynamism of laughter movement and to enhance the human perception, although they cannot overcome animation directly copied from human data.

1.5 Chapter Overview

The remainder of this thesis is organized as follows:

Chapter 2 reviews previous research works related to our work. We first review the data-driven approaches for speech-to-animation synthesis in section 2.1; then we review the existing approaches for laughing animation synthesis in section 2.2; finally we introduce in section 2.3 the Greta System which is the system of Embodied Conversational Agent that we used in our work.

Chapter 3 introduces our work on synthesizing eyebrow and head motions from prosody features. Section 3.1 introduces contextual models based on HMMs. Section 3.2 describes three proposed approaches of using contextual models to infer the motion signals from the speech signals. Section 3.3 introduces the dataset used to build motion generators. For simplicity, the three proposed approaches are first applied to only one modality motion (eyebrow) synthesis. The experiments are introduced in section 3.4 as well as their results. According to the experiment results, one out of the approaches is selected and applied to multiple modalities of motions of eyebrow and head. The experiments and results are presented in section 3.5. This work is concluded in section 3.6

Chapter 4 presents our work on laughter animation synthesis. The first work is detailed in section 4.1. It involves synthesis of facial expression (lip, jaw, eyebrow, eyelid and cheek) and extrapolated movements of head, shoulders and torso. The second work is detailed in section 4.2. It focuses on investigating more in depth, head and torso animation synthesis. The work on laughter animation synthesis is concluded in section 4.3.

Chapter 5 summarizes our work on speech and laughter animation synthesis.

Chapter 6 presents future work which merits to be further studied and investigated.

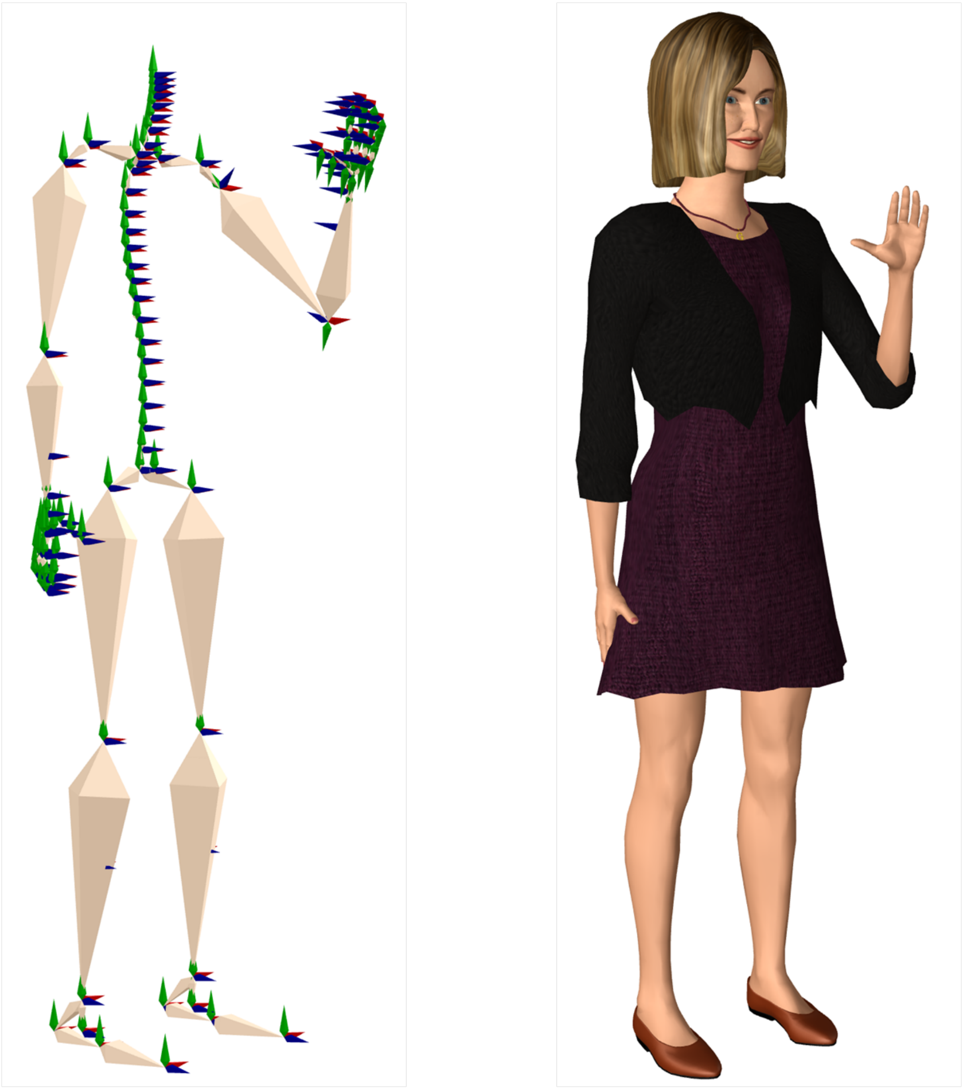


Figure 1.5: Full Body Model: (a)Skeleton Model; (b)Fully Body Model with texture. The right image is configured by the values of joint angles in the left skeleton model.

Chapter 2

Related Works

Virtual humans ought to be capable of displaying high quality behaviors while speaking or laughing. Speech animation synthesis has been a challenge that has been addressed since many years. But, few works have focused on laughter animation synthesis. Most existing models of speech-to-animation can be clustered into two main groups. In one group, models are based on theoretical models taken from domains such as psychology, emotion studies, linguistic. Such models apply often a rule-based approach. On the other hand, statistical models, called data-driven models, have been applied to capture the correlation between speech and multimodal behaviors from human datasets. Then, they render the captured correlation into outputted animation in the synthesis step.

In the 1990s, rule-based approaches have been proposed to generate various nonverbal communicative features, including head motion, facial expression and gesture. Such examples include works by [95, 22, 8]. For example, Pelachaud et al. [95] infer lip, head and eye motions from spoken text, which has been annotated according to predefined rules. In particular, blink occurs on accent; rapid movements of head take place on pitch accent and emphasis while slow movement of head at the end of an utterance; lip suck occurs with phonemes of f and v. However, since human behaviors arise from and may be influenced by various factors, such as emotion, personality, gender, physiological state and social context [68], it is extremely difficult to define a large set of rules to fully capture the role of these factors onto human behaviors, even after decades of studies in psychology. Besides, multiple rules could result in synthesizing conflicting expressions [96]. For example, each performative [1] and each qualifier [2] have both a corresponding lexical expression. A frown is included in the facial expression for the performative "I criticize"; while a raising eyebrows occurs for a qualifier of uncertainty. One may criticize another one while being a bit doubtful (uncertain) on one's criticism. One cannot frown (criticizing) and raise one's eyebrows (be uncertain) at the same time without creating an expression with a third meaning (oblique eyebrows of sadness) [96]. Later, researches turned to other approaches for animation synthesis.

In the late 1990s and early 2000s, data-driven approaches have been applied to speech-to-animation synthesis including gesture and facial expression. Data-driven approaches rely on human dataset containing both speech and facial expression data. First, the proposed models (data-driven approaches) are trained on the human dataset for capturing the link between human speech and motion. Then, in the step of synthesis, the captured link is rendered when computing the synthesized animation given a new speech input. [14, 29, 16, 62, 48, 39, 74] are examples of models following such an approach.

In these works, speech signals that are being considered can be categorized into two groups: spoken content and speech prosody. Spoken content can be directly obtained from speech text. It can also be recognized from speech features, such as LPCC and MFCC, which are strictly related to speech text (e.g. phoneme and syllable). Speech prosody is related to context information, such as speaker's emotion and attitude. It is usually featured by speech pitch and energy. Busso et al. [19] indicate that spoken content is related to what has been said and that speech prosody is related to how the spoken content is uttered. Therefore, spoken content is usually used to synthesize co-articulation animation, such as lip movement, while speech prosody is usually used to synthesize non-verbal animation such as eyebrow, head movements and gestures.

In section 2.1, we first review the data-driven approaches for speech-to-animation synthesis. Then, in section 2.2, we present the existing approaches for laughing animation synthesis. Finally, in section 2.3, we introduce the Greta System which is a system of Embodied Conversational Agent that we used in our work.

[1] performative: *the social attitude in which we put ourselves towards the hearer, and the reason why we are talking to him* [96].

[2] qualifier: propositional content [96].

2.1 Speech Animation

In the last 15 years, data-driven models have been proposed to build body and facial animation generators for embodied conversational agents (ECAs). These models take speech signals as input including speech content or speech prosody; they output various bodily animations involving lip, facial expression, head, or gesture. [16, 6, 29, 14, 123, 62, 39, 74] are examples of synthesizing lip animation; [14, 123] are examples of synthesizing facial expression (including upper face and lower face); [19, 56, 18] are examples of synthesizing head motion; and [70, 71, 26, 27, 79] are examples of synthesizing communicative gestures and body postures.

Deng et al. [39] contributed to model co-articulation transition between lip shapes of successive phonemes. In their work, lip shape at time t is defined as the sum of weighted visemes (visual counterpart to phonemes) of adjacent phonemes. The used weights vary with time; they are modeled by third degree polynomial curves. Yehia et al. [124] used linear mapping to model facial motion and speech features. In synthesis, facial motion is linearly conditioned on speech features. Later, Kuratate et al. [64] built a non-linear mapping between facial motion and speech features using a neural network working on curent speech frame and few past motion frames. Most of previous works used graphical models, such as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Conditional Random Field (CRF) and Restricted Boltzmann Machines (RBM). We will now review these works.

Most of the existing graphical models proposed to use Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). Section 2.1.2 reviews a synthesis technique which allows to synthesize directly a smoothing observation sequence given a learnt HMM. This synthesis technique is reported in Tokuda et al. [113]. Section 2.1.3 reviews a classic methodology of speech-to-animation generators based on GMM or HMM. Section 2.1.4 reviews the other works.

2.1.1 Notations

In the following we will note vectors in lowercase (e.g. x) and matrices in uppercase (e.g. Σ). A sequence will be noted in bold (e.g. $\boldsymbol{x}$) and elements of a sequence will be noted in standard font. If $\boldsymbol{x}$ is a sequence of T elements, then $\boldsymbol{x} = (x_1, ..., x_T)$, where x_t might be scalar or vector.

2.1.2 Using a HMM for synthesis

Synthesizing a realistic sequence of observations (called a trajectory hereafter) from a HMM is a key issue. Of course, synthesizing the most likely observation sequence given a particular state sequence yields a very unlikely piecewise constant trajectory (state mean sequence). Although interpolation technique can be used to smooth animation trajectory, it could damage perceived dynamic of animation inferred from the trained HMM.

A key technique has been proposed in [113] to synthesize more realistic and smooth trajectories from a trained HMM with Gaussian probability density functions. In their work, a feature vector includes a set of features (called static features) together with their velocity and acceleration features where velocity and acceleration features are calculated as follows:

$$\Delta o_t = -0.5 o_{t-1} + 0.5 o_{t+1} \tag{2.1.1}$$

$$\Delta\Delta o_t = o_{t-1} - 2 o_t + o_{t+1} \tag{2.1.2}$$

where o_t stands for the static features at time t; and Δo_t and $\Delta\Delta o_t$ respectively stand for velocity and acceleration features, these are called dynamic features. Actually a frame at time t, x_t, is defined as $x_t = [o, \Delta o, \Delta\Delta o]$ and may be computed as a function of static features o_t with a particularly shaped matrix W according to $x_t = W \times o_t$.

In some way, Eq. (2.1.1) and (2.1.2) may be viewed as a set of constraints on the evolution of static features in a sequence. The work by [113] showed that such explicit constraints may be taken into account with some benefit in the synthesis step. More precisely, consider one is given a state sequence $\boldsymbol{q}$ and wants to synthesize an observation sequence from this state sequence. A naive synthesis approach would first look for $\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} P(x|\boldsymbol{q}, \lambda)$ (with $\boldsymbol{x} = [\boldsymbol{o}, \Delta\boldsymbol{o}, \Delta\Delta\boldsymbol{o}]$) and then synthesize the static features sequence extracted from $\hat{\boldsymbol{x}}$. Alternatively the technique by Tokuda et al. directly look for $\hat{\boldsymbol{o}}$ such that $\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} P(W \times \boldsymbol{o}|\boldsymbol{q}, \lambda)$ where W is the matrix defined above.

This procedure yields synthesizing an observation sequence which is different from the most likely trajectory which would be, assuming a known path (i.e. a state sequence), the sequence of the means of the Gaussian distribution in the successive states of the path. In some way their method aims at synthesizing the most likely

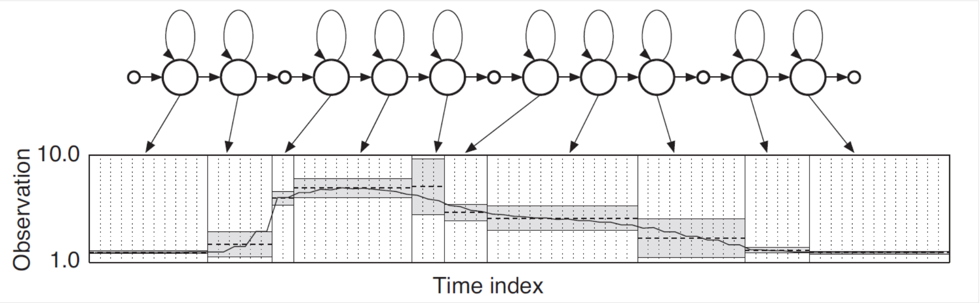


Figure 2.1: A sample of the trajectory generated from a learnt HMM [57]. The dotted line is the sequence of state means. The solid line is generated by the synthesis technique in [113]. It is a much smoother trajectory than the dotted line because successive observations have been made dependent on each others.

observation sequence (of static features) which satisfies some constraints on their evolution. Figure 2.1 shows a sample of the trajectory generated from a HMM [57] using this method. As can be seen, the generated trajectory (solid line) is much smoother than the most likely observation sequence (dotted line).

Tokuda et al. [113] derived algorithms for solving the following problems:

Case 1. For a given HMM (noted λ) and for a given state sequence, noted $\boldsymbol{q}$, determine the observation sequence $\boldsymbol{o}$ that maximizes $P(\boldsymbol{o}|\boldsymbol{q},\lambda)$ under the conditions given by Eq. (2.1.1) and Eq. (2.1.2).

Case 2. For a given HMM λ, determine the observation sequence $\boldsymbol{o}$ and the state sequence $\boldsymbol{q}$ that maximize $P(\boldsymbol{o},\boldsymbol{q}|\lambda)$ under the conditions given by Eq. (2.1.1) and Eq. (2.1.2).

Case 3. For a given HMM λ, determine the observation sequence $\boldsymbol{o}$ that maximize $P(\boldsymbol{o}|\lambda)$ with respect to $\boldsymbol{o}$ under the conditions 2.1.1, 2.1.2.

where λ stands for set of parameters of a given continuous mixture HMM; $\boldsymbol{q}$ stands for a state sequence; and $\boldsymbol{o}$ stands for a sequence of static features. In $Case1$, $P(\boldsymbol{o}|\boldsymbol{q},\lambda)$ is calculated for a given $\boldsymbol{q}$; in $Case3$, although $\boldsymbol{q}$ is unknown, $P(\boldsymbol{o}|\lambda)$ can be viewed as the integration of $P(\boldsymbol{o},\boldsymbol{q}|\lambda)$ over all the possible $\boldsymbol{q}$ as follows:

$$P(\boldsymbol{o}|\lambda) = \sum_{all\ \boldsymbol{q}} P(\boldsymbol{o}|\boldsymbol{q},\lambda)P(\boldsymbol{q}) \tag{2.1.3}$$

$$\tag{2.1.4}$$

In the work of Tokuda et al. [113], the synthesis algorithms mentioned above have been used to generate speech features.

2.1.3 Synthesizing Motion Animation from Speech with Standard HMM

The synthesis method described above has been used for synthesing a sequence of observations from a HMM learnt on such sequence of observations. For instance [113] aimed at syntheising speeh signals from a HMM that had been trained on speech utterances.

Yet the same synthesis method has been also used in a different design in the past, for synthesizing an information stream from another information stream, e.g. for synthesing head motion (of an animated agent) from speech signal (that the agent is supposed to utter). One may cite here the work by Hofer et al. [56] who relied on the case 1 above.

The key idea, that was followed by a number of researchers, has been to use Gaussian distributions on feature vectors including speech and motion features [31, 66, 109, 19, 18, 56] hence introducing some link between the observation of particular motion features and of particular speech features. This has been investigated for Gaussian mixture models and for HMMs, we will focus our presentation on these latter works.

Training HMMs operating on observations built from the concatenation of two sets of features (e.g. speech and motion) is probably the most popular approach for synthesizing one set of features from another one (e.g. generating motion from speech), we will use this approach as a baseline in our experiments.

The key idea consists in designing and learning a Gaussian *joint* HMM, named λ hereafter, working on concatenated observation vectors of the two streams (i.e. a frame at time t is $x_t = [x_t^1 x_t^2]$ where x_t^i stands for the feature vector at time t for stream i). A key point is that one can build from the *joint* HMM a Gaussian HMM for every stream, named λ_1 and λ_2, by keeping only parameters related to the stream under consideration. Note that these models λ_i have the same architecture and share transition probabilities. Note also that the training of the HMM makes that states naturally focus on a particular combination of the observation in both streams, it undirectly links the two information streams.

Once such a joint HMM is trained one can synthesize a trajectory for the second stream from the observation sequence of the first stream as follows. Using λ_1 one determines the most likely state sequence according to the information stream 1. Then using λ_2 one can determine an observation sequence for stream 2 using either a naive approach (the most likely sequence which is the sequence of the means of the Gaussian distributions in the successive states) or a more efficient approach using *case 1* procedure that we evoked in previous section. We name this synthesis approach the *single* method, it has been used for instance in [19, 18].

Alternatively, one may use λ_1 to compute the probability distribution over all state sequences given the stream 1 observation sequence. Then using λ_2 one can determine an observation sequence using the *case 3* algorithm proposed in Tokuda et al., this has been investigated e.g. in [73].

Usually whatever the approach used, the synthesis ends with an interpolation technique (e.g. quaternion unit sphere [3] used in [19, 18]) which is applied to smooth the observation sequence. The two approaches above can be detailed by the following equations:

$$\boldsymbol{o}^m = \underset{\boldsymbol{o}^m}{\arg\max} P(\boldsymbol{o}^m|\boldsymbol{o}^s,\lambda) \tag{2.1.5}$$

$$\boldsymbol{o}^m = \underset{\boldsymbol{o}^m}{\arg\max} \sum_{all\ \boldsymbol{q}} P(\boldsymbol{o}^m|\boldsymbol{q},\boldsymbol{o}^s,\lambda)P(\boldsymbol{q}|\boldsymbol{o}^s,\lambda) \tag{2.1.6}$$

integrated method :

$$\boldsymbol{o}^m = \underset{\boldsymbol{o}^m}{\arg\max} \sum_{all\ \boldsymbol{q}} P(\boldsymbol{o}^m|\boldsymbol{q},\lambda)P(\boldsymbol{q}|\boldsymbol{o}^s,\lambda) \tag{2.1.7}$$

single method :

$$\boldsymbol{q}_{max} = \underset{\boldsymbol{q}}{\arg\max} P(\boldsymbol{q}|\boldsymbol{o}^s,\lambda) \tag{2.1.8}$$

$$\boldsymbol{o}^m \approx \underset{\boldsymbol{o}^m}{\arg\max} P(\boldsymbol{o}^m|\boldsymbol{q}_{max},\lambda) \tag{2.1.9}$$

where $\boldsymbol{o}^m$ is motion feature vector sequence; $\boldsymbol{o}^s$ is speech feature vector sequence; λ is set of parameters used in HMMs; $\boldsymbol{q}$ is state sequence.

In some way the *single* method results, focusing on a single path, is a kind of approximation while the *integrated* method keeps all the possible information in the synthesized animation. Li and Shum [73] indicated that integrating over all state sequences the corresponding piecewise constant trajectories gives a better result in speech-to-facial expression. Similar approaches have been investigated with

GMM models [66] where the outputs from all the states of the GMM are integrated as in the *integrated* method.

Note finally that the idea of sharing a common set states between the two stream HMMs has been used in other situations. For instance [71] used the combination of a Conditional Random Field (CRF) and of a HMM, sharing the same topology, where the CRF operates on a first information stream and is learnt so as to infer the state sequence for the HMM which operates on the second stream. More details related to their work will be reviewed in section SynthesizingAnimationGraphicalModel.

The approach we just described allows building a HMM model able to synthesize one information stream from another one. The topology of the HMM may be chosen to be ergodic or according to any information one has at his disposal. For instance we used a dataset with speech and motion as the two streams of information for with the labeling information concerned the motion only. This means that we get training sequences of both speech and motion features together with a segmentation of the sequence into motion classes. In that case one may train independently one left-right HMM for every motion class and then group all these models in a global model to perform inference and synthesis. This was first proposed in [56]. In that case the single approach for synthesis method somehow resumes to first infer (from the speech signal) the most likely path in the global HMM, i.e. the most likely sequence of motion classes and their duration, and second to use this path in the model to synthesize the animation trajectory, i.e. the sequence of motion features.

2.1.4 Other related approaches

Few other approaches have been proposed for synthesizing one information stream (e.g. motion) from another information stream (speech). We briefly review these now.

Levine et al. [70] used a large set of full segments of motion as finite observation space with discrete HMMs. They proposed a cost function to first select the successive motion segments given a state sequenceand speech features. Then the selected segment are concatenated as synthesized motion stream.

In a later work Levine et al. [71] used a CRF to learn the correct state sequence in a motion HMM based on speech features. In their work, motion features are modeled by a HMM. This HMM and the CRF share the qame topology, i.e. set of states and authorized transitions. In the synthesis phase, the state sequence probability distribution is determined by CRF using the speech features, then the *single* method is used to synthesize the animation stream using the HMM and the best state sequence.

Chiu and Marsella [26] built gesture generator based on Restricted Boltzmann Machines (RBMs). The major philosophy behind their work is to identify motion features which better represent the temporal pattern of human motion. Such motion features are identified by unsupervised learning on motion data.

Finally Li and Shum [73] proposed to use input-output HMM. In their work HMM's parameters including state transition probabilities and emission probabilities are conditional on speech features. The mappings between HMM's parameters and speech features are modeled by a neural network. In the synthesis phase, HMM's parameters are computed by the neural net for the given speech features then the motion stream is synthesized with this HMM using the *integrated* method.

We will make use and combine few of these ideas in our contributions that we present in the next two chapters.

2.1.5 Discussion on Speech Animation Works

As introduced above, in section 2.1.3, HMMs have been popularly used to learn the relationship between human motions and speech features and then used to synthesize animation from speech signals.

A HMM is based on a first-order Markov finite state machine. It includes a finite number of states whose sequence obeys the Markov property (we consider order 1 Markov models only), meaning that the state at time t is independent of all the previous states given the state at time $t-1$. The state at time t is noted q_t. This Markov property allows parameterizing the models with a very limited of transition probabilities from a state to another state. These transition probabilities are noted $a_{ij} = P(q_t = j|q_{t-1} = i)$.

A probability density function (*pdf*) is associated to each state.It is usually a Gaussian distribution or a mixture of Gaussian distributions. These *pdfs* are defined on an observation space noted X and the observation at time t is noted x_t. The *pdf* in state j is denoted b_j and its value for observation x_t is noted $b_j(x_t)$. Such a parameterization of HMMs comes from the second main hypothesis underlying HMMs (the first one being the Markov property of the underlying Markov chain): the observation at time t, x_t, is assumed independent of all other observations and all states given the state at time t, x_t.

These two assumptions underlying the HMM formalism have made these models very popular since they lead to very efficient algorithms and easy to use models. Yet, these assumptions are never satisfied in practice with real data. And from a synthesis perspective these assumptions are much too strong and result in a limited expressive power for generating realistic trajectories [73].

2.2 Laughter Animation

In this section, we first review previous studies on observable periodicity of head and body motion during laughter. Secondly, we review few reported works involving laughter animation synthesis. Finally, we review the representative works on data-driven animation models for speaking character.

2.2.1 Periodicity Analysis on Head and Body Laughter Motion

Darwin reported "During excessive laughter the whole body is often thrown backward and shakes, or is almost convulsed" [34]. Ruch and Ekman [105] described laughter movements as "rhythmic patterns", "rock violently sideways, or more often back and forth", "nervous tremor ... over the body", "twitch or tremble convulsively". Melo et al. [37] built a virtual character which "convulses the chest with each chuckle". It means that periodic motions of head and body are very well known during laughing. The periodicity of body motion was used to distinguish between different videos of laughter in [77]. Ruch and Ekman [105] reported that rhythmical patterns during laughter were usually characterized by frequency around 5 Hz. Mancini et al. [77] observed 8 videos which show actors laughing while watching funny images. Laughing actors produce rhythmic body movements with frequencies in the range of [1.27Hz 3.66Hz]. Ma et al. [76] showed that head motion frequency influences directly human perception.

2.2.2 Laughter Motion Synthesis

Few works have been focused on laughter motion synthesis. This section overviews the works involving laughter motion synthesis.

DiLorenzo et al. [40] proposed a physics-based model of human torso deformation during laughter. This model is anatomically inspired and synthesizes torso muscle movements activated by audio wave signal. Yet, the animation cannot be synthesized in real-time and the model cannot be easily extended to facial motion (e.g. eyebrow) synthesis. In their work, torso model consists of rigid body and deformable components. The rigid body contains spine and clavicle which are controllable by a set of joint angles. The deformable components contain the respiratory muscles, the abdomen, and the muscles attached to the clavicles which are respectively modeled by a Hill-type muscle model [17, 86, 125]. This muscle model can simulate muscle stretching and contraction which are controlled by a parameter in reference to force. Certain parameters controlling the rigid body and the deformable components are related to each other for maintaining the torso stability. Additionally, all the parameters are classified into 4 groups. The parameters in the same group are controlled by one mid-signal. These mid-signals are computed from the air flow, based on the previous existing pressure models [83, 120, 126]. Furthermore, the air flow is estimated from audio wave signal based on the work of [75].

Niewiadomski and Pelachaud [87] consider how laughter intensity modulates facial motion. A specific threshold is defined for each key point. Each key point moves linearly according to the intensity if it is higher than the corresponding threshold. So, if the intensity is high, the facial key points concerning laughter move more. In this model, facial motion position depends only on laughter intensity. It lacks of variability. Moreover, all facial key points move always synchronously, while human laughter expressions do not. For example, for the same intensity, one human subject can move both eyebrows, another one only one eyebrow. In their perceptive study, each laughter episode is specified with a single value of intensity. It leads to only one invariable facial expression during this laughter episode.

Later on, Niewiadomski et al. [88] propose an extension of their previous model. Recorded facial motions sequence is selected by taking into account two factors: laughter intensity and laughter duration. In this model, coarticulation of lip shapes is not considered which may lead to non-synchronization between lip shape and audio information (e.g. closed lip and strong intensity audible laughter information). Moreover, the roles of intensity and duration are not attentively distinguished when selecting recorded motion sequence. As a side effect, the selected motion may last differently (e.g. too short) than the desired duration.

Urbain et al. [114] proposed to compare the similarity of new and recorded laughter audio information and then to select the corresponding facial expressions sequence. The computation of the similarity is based on the mean and standard deviation of each audio feature during the laughter audio sequence. It means that the audio sequence is specified by only two variables: mean and standard deviation. This is not enough to characterize long audio sequence.

Cosker and Edge [30] used HMM to synthesize laughter facial motion from audio features (MFCC). The authors built several HMMs to model laughter motion, one HMM per subject. To compute the laughter animation of new subject, the first step is to select one HMM from the set of trained HMMs by comparing the audio similarity between the new subject and the subjects involved in the training dataset. Then the selected HMM is used to produce the laughter animation. The authors do not precise how many HMMs should be built to cover various audio patterns of different subjects. The use of the classification operation as well as of the Viterbi algorithm make impossible to obtain animation synthesis in real time. In the states sequence computed by the Viterbi algorithm, one single state may last very long. It leads to unchanged motion position during such a state which produces unnatural animations.

Hüseyin et al. [25] observed that the visual laughter on the face appears mostly before audible laughter and disappears after the end of laughter sound. They segmented the laughter sequence of facial expression into 3 segments. One segment begins and ends at the same time as the laughter sound. It is labeled by *Audible Laugh* (L). The segments before and after *Audible Laugh* are labeled by *Neutral* (N) or *Smile* (B). They collected 3 subsets of segmented motions labeled by the 3 labels. Each subset is used to train a HMM. In the synthesis phase, they used as input a visual label sequence ($[N,L,N]$ or $[B,L,B]$) and the duration of each label. The duration information ensures that facial expression labeled by *Audible Laugh* begins and ends at the same time with laughter sound.

2.2.3 Discussion on Laughter Animation Works

As introduced above, in Section 2.2.2, we are not aware of existing work on modeling full body laughter animation synthesis. Our aim is to build such a synthesis model that includes facial expressions and upper body behaviors.

DiLorenzo et al. [40] proposed the muscle-based model of torso animation synthesis. Their approach cannot be generalised to head and facial expressions animation. Moreover, in the Greta system [94, 9], the virtual character model consists of rigid joint articulations (details can be seen in section 2.3); hence, muscle-based model is not adapted to the Greta system. Other pioneers [87, 88, 114, 30, 25] have attempted to build models of laughter animation synthesis, but these models are too simple and cannot generate human-like laughter animation.

2.3 Greta System

The Greta system controls autonomous software characters that have human-like appearances (see Figure 2.2) and endowed with communicative and expressive capabilities [94, 9]. They are capable of using speech and multi-modal behaviors to convey intentions and to express emotions as humans do. Figure 2.3 shows examples of facial expression displaying two emotions: joy and fear. Speech synthesis models (OpenMary TTS [110], Cereproc [2], Acapela [1]) have been integrated into the Greta system. They take as input the text to be said and output speech audio signal (wav files). They also provide the list of phonemes and their duration that are used to compute the lip shapes. The text to be said is augmented with information regarding communicative acts and emotional states [99]. This information is embedded into tags following the Function Markup Language FML [55], called Affective Presentation Markup Language APML [21]. The behavior engine instantiates and synchronizes multimodal behaviors (facial expressions, gestures, body movements) from the APML-FML tags.

The body of ECA in Greta system is a full skeleton of 186 articulations as defined by the MPEG-4 standard [91]. The skeleton is animated through 186 parameters called Body Animation Parameters (BAPs) [91], which control the angles of rotation of body joints. Figure 2.4 shows the topology structure of the leg and the spine. Motions of human body are defined by changing the values of BAPs. The left side of Figure 2.4 shows the joint articulation of the leg; the right side of Figure 2.4 shows those of the spine.

The facial model has a mesh structure with the shape of human face (see Figure 2.5). The mesh structure consists of vertices connected by edges. Vertices can be controlled to deform the shape of the face to simulate human expression. The

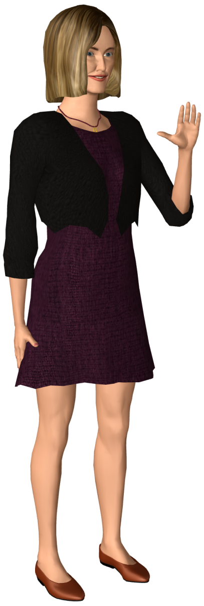
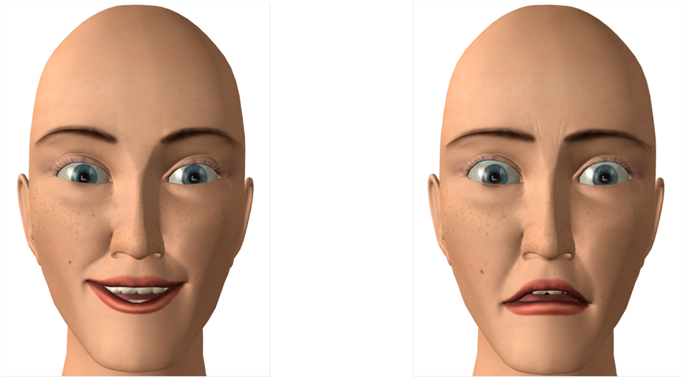
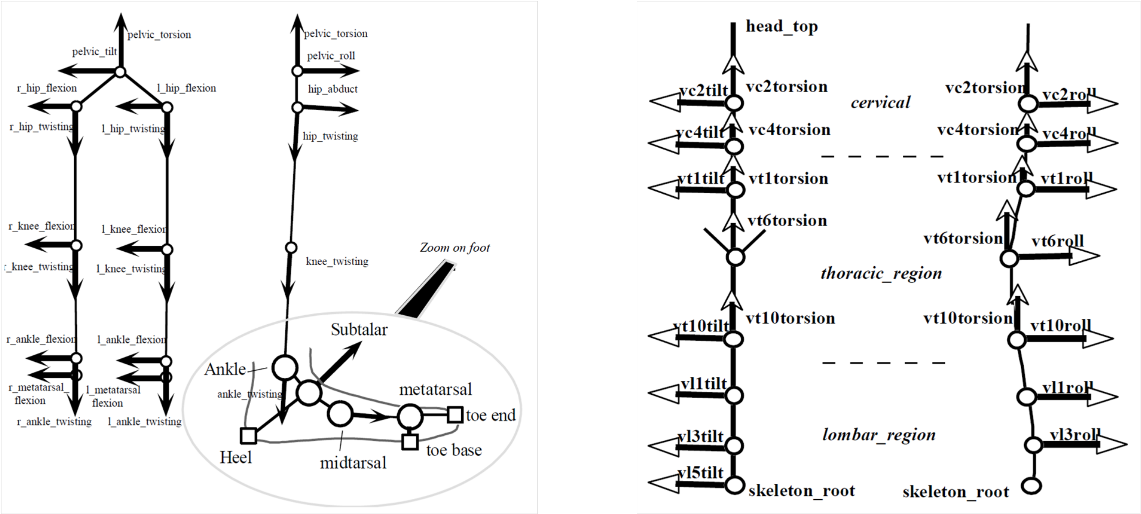


Figure 2.2: The appearance of Greta full body



Left: facial expression of joy Right: facial expression of fear

Figure 2.3: the examples of facial expressions



Left: Front and side views of the mobilities of the leg

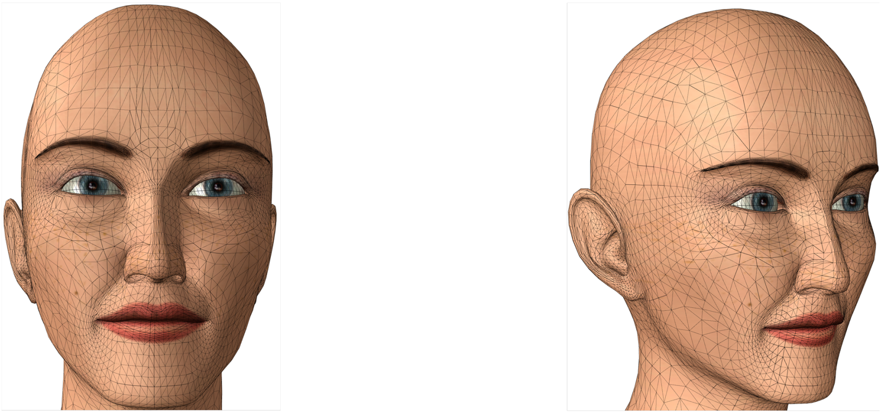
Right: Front and side views of the mobilities of the simple spine

Figure 2.4: Topology structure of the Body Animation Parameters (BAPs)[91]

texture information applied on the face moves with the vertices. A simple and common animation approach is to define all the vertices of the facial mesh at all times. They are viewed as key poses or key frames. The intermediate frames between key frames can be calculated by interpolation techniques such as linear interpolation [98], cosine interpolation [118], bilinear interpolation [92], etc.

Key points are specified on the mesh of the face. Their displacements control the displacements of the surrounding vertices. MPEG-4 refers to these vertices as feature points (FPs). Figure 2.6 shows all the 44 FPs considered in the MPEG-4 standard [91]. Their displacements are determined by 66 Facial Animation Parameters (FAPs). A FAP specifies the displacement of a FP along one dimension. So 66 parameters can be used to define the displacement of all the vertices of the face mesh model.

In the Greta system, an area of influence is defined for each FAP. The area of influence has an ellipsoid shape [94]. Each FP is the centroid of ellipsoid shape. Figure 2.7 shows an example of area of influence. The displacement of a vertex in the area is proportional to value of FAP and its position related to the center of the ellipsoid. Proportional weight is based on the distance between a vertex and the FP; it is computed by a cosine function shown in Figure 2.8. Figure 2.9 shows an example of deformation within an area of influence.



Left: facial mesh structure (front view) Right: facial mesh structure (side view)

Figure 2.5: Facial mesh structure

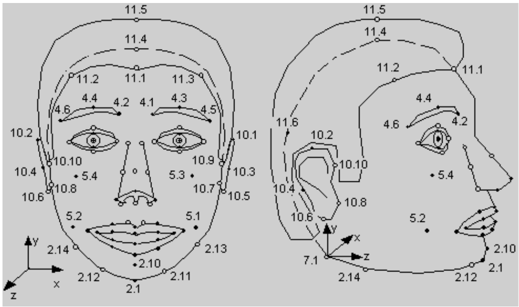


Figure 2.6: The feature points (FP) [91]

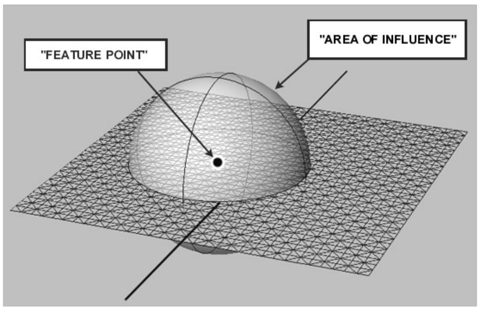


Figure 2.7: Area of Influence [94]

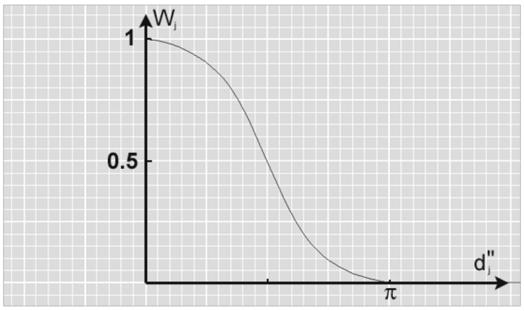
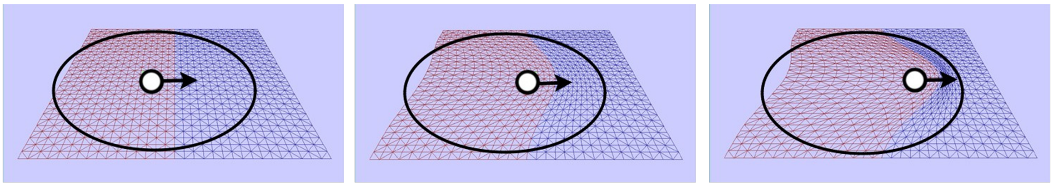


Figure 2.8: The function of proportional weight. In this figure, W_j stands for proportional weight, $d_j^{''}$ stands for the distance between the vertex and the feature point.[94]



Left: null intensity

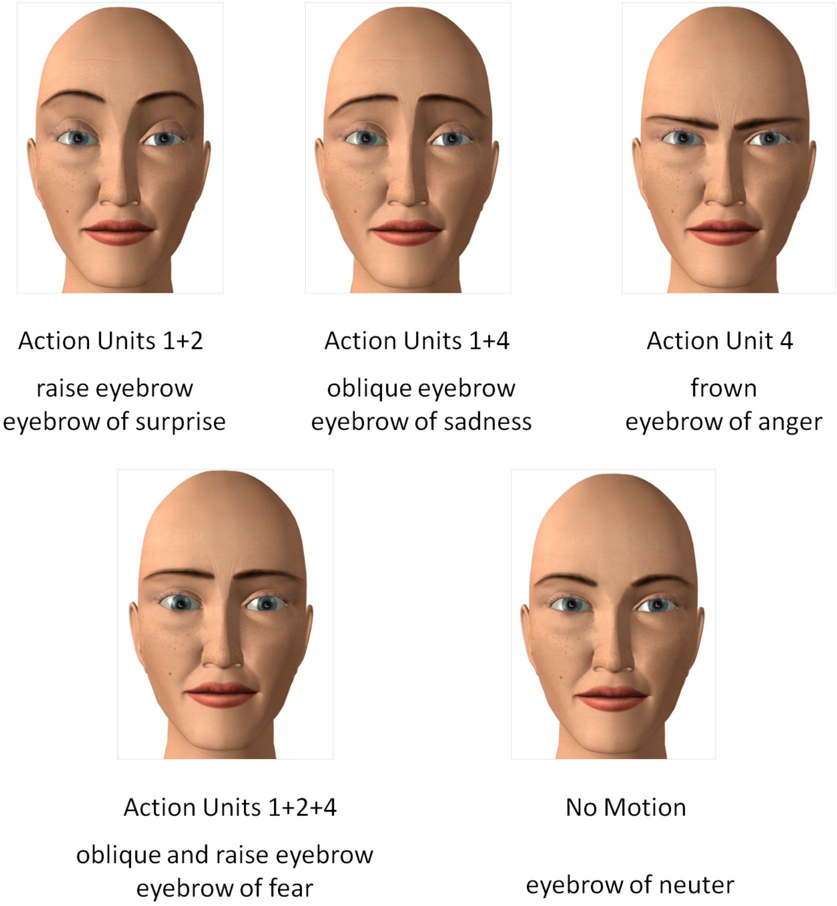
Middle: medium intensity

Right: maximum intensity

Figure 2.9: Deformation within an area of influence [94]

Another approach defined in the Greta system for deforming the face mesh is based on 44 basic Action Units (AU), which are defined in the Facial Action Coding System (FACS) [45]. An AU corresponds to a minimal visible muscular contraction. Examples are AUs are Inner Brow Raiser (AU1), Outer Brow Raiser (AU2), Nose Wrinkler (AU9), Lid Corner Puller (AU12), etc. Facial expressions are described as a combination of AUs. Figure 2.10 shows five examples of combination of AUs in the eyebrow regions. he movement of all the AUs can be defined independently. Within the Greta system, a mapping between AU and FAP has been designed. So the movement of an AU is computed by moving the corresponding FAPs.

Greta system follows an international common multi-modal behavior generation framework, called SAIBA architecture [61]. SAIBA architecture is shown in Figure 2.11. The SAIBA architecture consists of three modules: *Intent Planner*, *Behavior Planner* and *Behavior Realizer*. These three modules represent three levels of abstraction from the selection of ECA's communicative intention to behavior selection and realization. The information exchanged between the modules is realized through the Functional Markup Language (FML) [55] and the Behavior Markup Language (BML) [61].



Action Units 1+2

raise eyebrow
eyebrow of surprise

Action Units 1+4

oblique eyebrow
eyebrow of sadness

Action Unit 4

frown
eyebrow of anger

Action Units 1+2+4

oblique and raise eyebrow
eyebrow of fear

No Motion

eyebrow of neuter

Figure 2.10: Eyebrow Action Units (AUs)

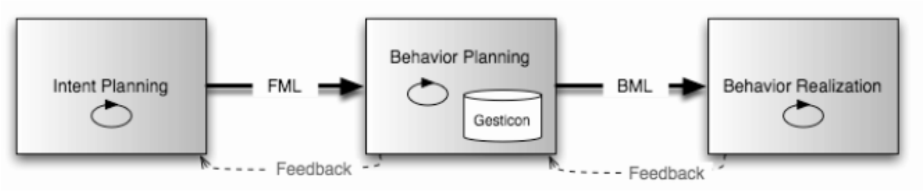


Figure 2.11: SAIBA framework for multimodal generation [61]

The *Intent Planner* module defines the ECA's goals, emotional state and beliefs which are encoded into FML tags. Given the FML tags, the *Behavior Planner* module selects and schedules multimodal signals such as facial expressions, gestures, gaze and then encodes them into BML tags. The *Behavior Realizer* module takes as input BML tags and then generates the corresponding animation.

Chapter 3

Speech Animation Synthesis

Embodied conversational agents, ECAs, are autonomous software characters that often have a human-like appearance and are endowed with communicative and expressive capabilities. They are capable of using speech and multimodal behaviors to convey intentions and to express emotions as humans do. Natural and lively behaviors are crucial for engaging users in human-computer interactions. Humans are very sensitive to subtle behaviors. They have very good skills to interpret the behaviors they perceive in their interlocutors. ECAs ought to be capable of displaying such high quality behaviors. In fact, previous works have proposed various behavior generators to augment the quality of displayed behaviors. Their effort has made ECAs more and more realistic, natural and believable in human-machine interaction, which motivated the prevalence of ECAs in interactive systems such as online web applications.

In video games or cinematographic applications the animation of virtual characters are reproduced from large motion capture datasets of an actor's performance. This approach induces two non-negligible disadvantages: data acquisition expenses and restriction to movement reproduction of the recorded scenarios [19], hence it cannot be applied to the autonomous ECAs. In fact, amount of works have contributed to develop models of behavior synthesis for autonomous ECAs. Most existing models of ECAs' behaviors can be clustered into two main groups. In one group, models are based on experimental data and theoretical models taken from domains such as psychology, emotion, linguistic. Examples of such models are works by [95, 22, 8, 9]. On the other hand, statistical techniques have been applied to learn from data the correlation between speech and multimodal behaviors.

These models make use of the tight relationship between acoustic and visual behaviors, [14, 29, 16, 62, 48, 39, 74, 31, 66, 109, 19, 18, 56, 70, 71, 73, 78] belong to this cluster. These models make use of the tight relationship between acoustic and visual behaviors.

Humans communicate not only semantic content but also emotion, intention and desires, which are expressed through acoustic and visual means such as speech, facial expressions, gesture, etc. McGurk and MacDonald [81] emphasizes the importance of both audible and visual signals in human-human communication. Acoustic and visual signals are sophistically modulated to express and emphasize spearker's emotion, intention, etc [22]. Humans are sensitive to subtle expressions during face-to-face conversation. For example, they are skilled in inferring their interlocutor's affective and mental states from their accompanied facial expression or body gestures.

Amounts of works have studied the relationship between various multi-modalities of expression. They are intricately manipulated to produce a unifying message according to speaker's intention. They are linked and synchronized with each other. Eyebrow and head motions are often used to complement the verbal information as visual prosody. Eyebrow movement can be used as important factor to segment a uttered speech sequence; Munhall et al. [84] reported that natural head motion significantly facilitates auditory speech perception. Graf et al. [52] indicated that eyebrow and head motions are strictly linked to the speech prosodic. Rising fundamental frequency of speech is accompanied with rising or falling eyebrow [24]. It exists a strong correlation ($r = 0.8$) of head motions and acoustic prosodic features [63].

Our work is focused on investigating a statistical model to infer eyebrow and head motions from speech signals. This model can be parameterized from training samples and can then synthesize natural animation motion from speech features. Considering the advantages of HMM to model data stream, we chose to base our models on HMM. HMM has been used to synthesize animations in previous works (more details can be seen in section 2.1.3). As discussed in section 2.1.5, previous works suffer limitations of independence assumption from standard HMM. To overcome such limitations and to take advantages of the HMM to model data sequence, we developed a variant of standard HMM, whose parameters are defined during the synthesis phase instead of during the training phase.

In our work, to take advantage of HMM, Markov chain is always used to model sequences of motion features by taking motion features as state observation, as it has been done in previous works. To overcome HMM's limitations, in our work, speech features are not always taken as state observation; they are used to calculate the parameters of HMM in the step of synthesis. This choice is inspired from an extension of HMM, called Contextual HMMs [121, 103]. Contextual HMMs are HMMs whose state observation parameters including state mean and covariance depend on contextual variables, also called external variables. Contextual HMMs have been used for gesture recognition in previous works, while they are used for synthesizing behaviors in our work. Inspired by existing Contextual HMMs, we developed a new extension of HMM, called Fully Parameterized HMM (FPHMM). In FPHMM, not only state observation parameters but also state transition probabilities depend on contextual variables. In our work, contextual variables are prosodic features.

This chapter introduces our work on synthesizing eyebrow and head motions from prosody features. Section 3.1 introduces contextual models based on HMMs. Section 3.2 describes three approaches of using contextual models to infer the motion signals from the speech signals. Section 3.3 introduces the dataset used to build motion generators. For simplicity, three proposed approaches are first applied to only one modality motion (eyebrow) synthesis. The experiments are introduced in section 3.4 as well as the results. From to the experiment results, the approach that obtained the best results is selected and applied to eyebrow and head motions. Head and eyebrow motions are not separately synthesized from speech signals; rather the relationships between these two-modal motions is taken into account. The experiments and results are presented in section 3.5.

3.1 Contextual Models

Hidden Markov Models (HMMs) are statistical *generative* models (meaning they may be used to generate/synthesize data) that are well-known for dealing with data streams. It can be applied to data such as speech signal and human motion features. These are automatically trained from a set of training samples of data stream to capture dynamic characteristics of the data stream. Once the model has been trained, it may be exploited to infer new instances of the data stream in a synthesis step [19]. Below, we only recall basic elements on HMMs in order to introduce the necessary notations; a detailed presentation may be found in [102].

In more details,a HMM is based on a first-order Markov finite state machine. It includes a finite number of states whose sequence obeys the Markov property (We consider order 1 Markov models only), meaning that the state at time t is independent of all the previous states given the state at time $t-1$. The state at time t is noted q_t. This Markov property allows parameterizing the models with a very limited of transition probabilities from a state to another state. These transition probabilities are noted $a_{ij} = P(q_t = j|q_{t-1} = i)$.

A probability density function (*pdf*) is associated to each of the states, it is usually a Gaussian distribution or a mixture of Gaussian distributions. These *pdfs* are defined on an observation space noted X and the observation at time t is noted x_t. The pdf in state j is denoted $b_j(x_t)$ and its value for observation x_t is noted $b_j(x_t)$. Such a parameterization of HMMs comes from the second main hypothesis underlying HMMs (the first one being the Markov property of the underlying Markov chain): the observation at time t, x_t, is assumed independent of all other observations and all states given the state at time t, x_t.

These two assumptions underlying the HMM formalism have made these models very popular since they lead to very efficient algorithms and easy to use models. Yet, these assumptions are never satisfied in practice with real data. And from a synthesis perspective these assumptions are much too strong and result in a limited expressive power for generating realistic trajectories [73]. A number of researchers have contributed to overcome such a drawback for recognition purpose by introducing segmental HMMs or trajectory HMMs that relax the independence hypothesis between successive observations [107, 38]. Among these, few works may be exploited to synthesize more realistic trajectories [121, 103], we will build on these latter works here.

3.1.1 Contextual HMMs

In contextual HMMs the *pdfs* (Gaussian probability distribution) are defined as being a function of some external (or contextual) variables such as the morphology, the gender or the age in speech recognition [121, 103]. We recall briefly the existing contextual HMM models [121, 103].

We note θ (a vector of dimension c) as the vector of the values of the contextual variables for an observation sequence $\boldsymbol{x} = (x_1, ..., x_T)$ where x_t are observations (d-dimensional feature vectors). Considering for simplicity a contextual HMM model with a single Gaussian distribution as *pdf*, the mean of the Gaussian distribution in state j, $\hat{\mu}_j$ (d- dimensional vector), and its covariance matrix, $\hat{\Sigma}_j$($d \times d$ matrix), are defined according to:

$$\hat{\mu}_j(\theta) = W_j^{\mu}\theta + \bar{\mu}_j \tag{3.1.1}$$

$$\begin{aligned} &\hat{\Sigma}_j(\theta) = D_j(\theta) \times \bar{\Sigma}_j \times D_j(\theta) \\ &\text{with } D_j(\theta) = diag(exp(W_j^{\Sigma}\theta + \widetilde{\Sigma}_j)) \end{aligned} \tag{3.1.2}$$

with W_j^{μ} and W_j^{σ} two $d \times c$ matrices, and $\bar{\mu}_j$ and $\widetilde{\Sigma}_j$ their *offsets* vectors. The first equation says that the mean in a state depends on the contextual variables θ, allowing us to model some of the variability in the model. For instance in the case of gesture modeling and recognition it is intuitive to think that an older person will make smaller amplitude gestures and will perform them slower than a young person. The parameters of a contextual HMM gesture recognizer could take into account such variability by including the age of the performer as external variables. The same idea holds for the covariance matrix, whose parameterization allows us to slightly change the shape of the distribution around its mean. The exponential function ensures the elements included in $D_j(\theta)$ to be strictly positive. Note finally that θ varies with time without big changes since all above equations may be rewritten using θ_t instead of θ. In that case the mean of the Gaussian distribution in a state of the HMM will depend on this time varying external variable and will also vary with time.

When only the mean is parameterized according to Eq. (3.1.1) one gets a Parametric HMM which were proposed in Wilson and Bobick [121]. When both Eq. (3.1.1) and Eq. (3.1.2) are used one gets a Contextual HMM as proposed in Radenen and Artières [103].

The training phase consists in two steps. First, a Parametric HMM is learnt, it is performed with an adaptation of standard re-estimation formula used for HMMs. Next, all other parameters are kept fixed to estimate the W_j^{σ} and $\widetilde{\Sigma}_j$. This latter step is performed via the Generalized Expectation Maximization algorithm. The interested reader may look at [103] for more details.

3.1.2 Fully Parameterized HMMs

Inspired by the works [121, 103] introduced above, we proposed a new model that we named hereafter fully parameterized Hidden Markov model (FPHMM)[1]. A FPHMM is an extension of a Contextual HMM where in addition to means and covariance matrices, the transition probabilities and the initial state distribution are also parameterized and depend on θ instead of being fixed at particular values. In a FPHMM means and covariance matrices are defined as described above for contextual HMMs [121, 103] while the transition probabilities a_{ij} are now defined as (using a dynamic set of external variables):

$$\begin{aligned} a_{ij}(\theta_t) &= P(q_t = j | q_{t-1} = i, \theta_t) \\ &= \frac{\exp(log\bar{a_{ij}} + W_{ij}\theta_t)}{\sum_{j'} \exp(log\bar{a_{ij'}} + W_{ij'}\theta_t)} \end{aligned} \tag{3.1.3}$$

where θ_t is the c-dimensional vector of contextual features at time t; $W^{tr}_{i,j}$ is a c-dimensional vector; $\bar{a_{ij}}$ may be viewed as an offset value that would correspond to transition probabilities in case $\theta_t = 0$. According to this modeling schema the transition probabilities may change at every time step according to the contextual variables. Using such a modeling framework it may happen that a transition might be more likely to occur or not according to the values of the contextual variables. It means also that a state sequence may be sampled form the external variables only. Figure 3.1 illustrates the difference between CHMM and FPHMMs.

As previously defined contextual models a FPHMM is trained via likelihood maximization with a Generalized Expectation-Maximization algorithm. To ease learning it is initialized with a trained Contextual HMM, and by setting A_{ij} to the learnt transition probabilities.

Estimation of $W^{tr}_{i,j}$ is then performed via the Generalized Expectation-Maximization algorithm. First, we define an auxiliary function $Q(\lambda, \lambda')$ as the one used for learning HMMs, where λ' stands for the current value of the FPHMM parameters and

[1] This is a joint work with Mathieu Radenen while he was Ph.D. student at UPMC, France

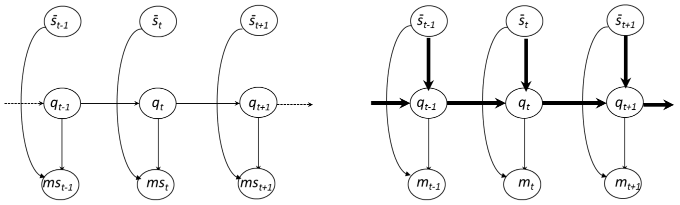


Figure 3.1: Representation of a CHMM (left) and of a FPHMM (right) as Dynamic Bayesian Networks (DBNs) for inferring a sequence of motion features (ms_t) from a sequence of speech features ($\bar{s}_t$). While a CHMM uses speech features to modify the *pdfs*, a FPHMM in addition takes into account the speech features to determine the state transitions (State at time t is noted q_t)

.

λ denotes the new values of the FPHMM parameters that we are looking for. Secondly we compute the derivative of the auxiliary function Q with respect to $W^{tr}_{i,j}$. Thirdly, we do a gradient ascent. It is a similar strategy than for the estimation of the covariance matrices in Contextual HMMs.

More precisely $Q(\lambda,\lambda')$ is defined as follows:

$$Q(\lambda,\lambda') = \sum_{all\ k,\boldsymbol{q}} P(\boldsymbol{q}|\boldsymbol{o}^k,\ \lambda) logP(\boldsymbol{q},\boldsymbol{o}^k|\lambda') \tag{3.1.4}$$

where $\boldsymbol{k}$ stands for the training sequence number and $\boldsymbol{q}$ stand for a state sequence path. This auxiliary function may be rewritten as:

$$\begin{aligned}
Q(\lambda,\lambda') &= \sum_{all\ k,\boldsymbol{q}} P(\boldsymbol{q}|\boldsymbol{o}^k,\ \lambda) logP(\boldsymbol{q},\boldsymbol{o}^k|\lambda') \\
&= \sum_{all\ k,\boldsymbol{q}} P(\boldsymbol{q}|\boldsymbol{o}^k,\ \lambda) \sum_{t=1}^{T} (loga_{q_{t-1}q_t}(\theta_t^k)(\theta_t^k) + logP(O_t^k|q_t,\lambda')) \\
&= \sum_{all\ k} \sum_{i=1}^{N} P(q_{t-1} = i|\boldsymbol{o}^k,\lambda) log\pi_i(\theta_{t=1}^k) \\
&+ \sum_{all\ k} \sum_{i,j=1}^{N} \sum_{t=2}^{T^k} P(q_{t-1} = i, q_t = j|\boldsymbol{o}^k,\lambda) loga_{ij}(\theta_t^k) \\
&+ \sum_{all\ k} \sum_{i}^{N} \sum_{t=1}^{T^k} P(q_t = i|\boldsymbol{o}^k,\lambda) logP(o_t^k|q_t,\lambda')
\end{aligned} \tag{3.1.5}$$

Now we detail how to compute the derivative of the auxiliary function Q with respect to $W^{tr}_{i,j}$:

$$\frac{\partial Q}{\partial W_{ij}} = \frac{\partial Q}{\partial a_{ij}(\theta^k_t)} \frac{\partial a_{ij}(\theta^k_t)}{\partial W_{ij}} + \sum_{j' \neq j} \frac{\partial Q}{\partial a_{ij'}(\theta^k_t)} \frac{\partial a_{ij'}(\theta^k_t)}{\partial W_{ij}} \tag{3.1.6}$$

as can be seen, 4 elements should be calculated to get the derivative of the auxiliary function Q with respect to $W^{tr}_{i,j}$. They are $\frac{\partial Q}{\partial a_{ij}(\theta^k_t)}$, $\frac{\partial a_{ij}(\theta^k_t)}{\partial W_{ij}}$, $\frac{\partial Q}{\partial a_{ij'}(\theta^k_t)}$ and $\frac{\partial a_{ij'}(\theta^k_t)}{\partial W_{ij}}$. Now we show how to calculate them.

We denote occupancy probability as follows:

$$\gamma_{k,t,i,j} = P(q_{t-1} = i, q_t = j | \boldsymbol{o}^k, \lambda) \tag{3.1.7}$$

Calculation of $\frac{\partial Q}{\partial a_{ij}(\theta^k_t)}$:

$$\begin{aligned}
\frac{\partial Q}{\partial a_{ij}(\theta^k_t)} &= \frac{\partial}{\partial a_{ij}(\theta^k_t)} \sum_{all\ k} \sum_{t=2}^{T^k} P(q_{t-1} = i, q_t = j | \boldsymbol{o}^k, \lambda) log a_{ij}(\theta^k_t) \\
&= \sum_{all\ k} \sum_{t=2}^{T^k} \frac{\partial}{\partial a_{ij}(\theta^k_t)} P(q_{t-1} = i, q_t = j | \boldsymbol{o}^k, \lambda) log a_{ij}(\theta^k_t) \\
&= \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j} \frac{1}{a_{ij}(\theta^k_t)}
\end{aligned} \tag{3.1.8}$$

Calculation of $\frac{\partial a_{ij}(\theta^k_t)}{\partial W_{ij}}$:

$$\begin{aligned}
\frac{\partial a_{ij}(\theta^k_t)}{\partial W_{ij}} &= exp(log a_{ij} + W_{ij}\theta^k_t)(\theta^k_t)^T \frac{1}{\sum_{j'} \exp(log a_{ij'} + W_{ij'}\theta^k_t)} \\
&\quad - \frac{exp(log a_{ij} + W_{ij}\theta^k_t)}{(\sum_{j' \neq j} \exp(log a_{ij'} + W_{ij'}\theta^k_t))^2} exp(log a_{ij} + W_{ij}\theta^k_t)(\theta^k_t)^T \\
&= \frac{exp(log a_{ij} + W_{ij}\theta^k_t)}{\sum_{j'} \exp(log a_{ij'} + W_{ij'}\theta^k_t)}(\theta^k_t)^T - \frac{(exp(log a_{ij} + W_{ij}\theta^k_t))^2}{(\sum_{j' \neq j} \exp(log a_{ij'} + W_{ij'}\theta^k_t))^2}(\theta^k_t)^T \\
&= a_{ij}(\theta^k_t)(\theta^k_t)^T - (a_{ij}(\theta^k_t))^2(\theta^k_t)^T \\
&= a_{ij}(\theta^k_t)(1 - a_{ij}(\theta^k_t))(\theta^k_t)^T
\end{aligned} \tag{3.1.9}$$

Calculation of $\frac{\partial Q}{\partial a_{ij'}(\theta_t^k)}$:

$$\frac{\partial Q}{\partial a_{ij'}(\theta_t^k)} = \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j'} \frac{1}{a_{ij'}(\theta_t^k)} \tag{3.1.10}$$

Calculation of $\frac{\partial a_{ij'}(\theta_t^k)}{\partial W_{ij}}$:

$$\begin{aligned}\frac{\partial a_{ij'}(\theta_t^k)}{\partial W_{ij}} &= -\frac{\exp(loga_{ij'} + W_{ij'}\theta_t^k)}{(\sum_{j'} \exp(loga_{ij'} + W_{ij'}\theta_t^k))^2} \exp(loga_{ij} + W_{ij}\theta_t^k)(\theta_t^k)^T \\ &= -\frac{\exp(loga_{ij'} + W_{ij'}\theta_t^k)}{\sum_{j'} \exp(loga_{ij'} + W_{ij'}\theta_t^k)} \frac{\exp(loga_{ij} + W_{ij}\theta_t^k)}{\sum_{j'} \exp(loga_{ij'} + W_{ij'}\theta_t^k)} (\theta_t^k)^T \\ &= -a_{ij'}(\theta_t^k) a_{ij}(\theta_t^k)(\theta_t^k)^T\end{aligned} \tag{3.1.11}$$

To calculate $\frac{\partial Q}{\partial W_{ij}}$, we should calculate $\frac{\partial Q}{\partial a_{ij}(\theta_t^k)} \frac{\partial a_{ij}(\theta_t^k)}{\partial W_{ij}}$ and $\sum_{j' \neq j} \frac{\partial Q}{\partial a_{ij'}(\theta_t^k)} \frac{\partial a_{ij'}(\theta_t^k)}{\partial W_{ij}}$.

Given the calculations of $\frac{\partial Q}{\partial a_{ij}(\theta_t^k)}$ and $\frac{\partial a_{ij}(\theta_t^k)}{\partial W_{ij}}$ by equation 3.1.8 and equation 3.1.9, we can calculate $\frac{\partial Q}{\partial a_{ij}(\theta_t^k)} \frac{\partial a_{ij}(\theta_t^k)}{\partial W_{ij}}$ as follows:

$$\begin{aligned}\frac{\partial Q}{\partial a_{ij}(\theta_t^k)} \frac{\partial a_{ij}(\theta_t^k)}{\partial W_{ij}} &= \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j} \frac{1}{a_{ij}(\theta_t^k)} \times a_{ij}(\theta_t^k)(1 - a_{ij}(\theta_t^k))(\theta_t^k)^T \\ &= \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j}(1 - a_{ij}(\theta_t^k))(\theta_t^k)^T\end{aligned} \tag{3.1.12}$$

Given the calculations of $\frac{\partial Q}{\partial a_{ij'}(\theta_t^k)}$ and $\frac{\partial a_{ij'}(\theta_t^k)}{\partial W_{ij}}$ by equation 3.1.10 and equation 3.1.11, we can calculate $\sum_{j' \neq j} \frac{\partial Q}{\partial a_{ij'}(\theta_t^k)} \frac{\partial a_{ij'}(\theta_t^k)}{\partial W_{ij}}$ as follows:

$$\begin{aligned}\sum_{j' \neq j} \frac{\partial Q}{\partial a_{ij'}(\theta_t^k)} \frac{\partial a_{ij'}(\theta_t^k)}{\partial W_{ij}} &= \sum_{j' \neq j} \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j'} \frac{1}{a_{ij'}(\theta_t^k)} (-1) a_{ij'}(\theta_t^k) a_{ij}(\theta_t^k)(\theta_t^k)^T \\ &= -\sum_{j' \neq j} \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j'} a_{ij}(\theta_t^k)(\theta_t^k)^T\end{aligned} \tag{3.1.13}$$

Given the calculations of $\frac{\partial Q}{\partial a_{ij}(\theta_t^k)} \frac{\partial a_{ij}(\theta_t^k)}{\partial W_{ij}}$ and $\sum_{j' \neq j} \frac{\partial Q}{\partial a_{ij'}(\theta_t^k)} \frac{\partial a_{ij'}(\theta_t^k)}{\partial W_{ij}}$ by equation 3.1.12 and equation 3.1.13, we can calculate $\frac{\partial Q}{\partial W_{ij}}$ as follows:

$$
\begin{aligned}
\frac{\partial Q}{\partial W_{ij}} &= \frac{\partial Q}{\partial a_{ij}(\theta_t^k)} \frac{\partial a_{ij}(\theta_t^k)}{\partial W_{ij}} + \sum_{j' \neq j} \frac{\partial Q}{\partial a_{ij'}(\theta_t^k)} \frac{\partial a_{ij'}(\theta_t^k)}{\partial W_{ij}} \\
&= \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j}(1 - a_{ij}(\theta_t^k))(\theta_t^k)^T - \sum_{j' \neq j} \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j'} a_{ij}(\theta_t^k)(\theta_t^k)^T \\
&= \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j}(\theta_t^k)^T - \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j} a_{ij}(\theta_t^k)(\theta_t^k)^T - \sum_{all\ k} \sum_{j' \neq j} \sum_{t=2}^{T^k} \gamma_{k,t,i,j'} a_{ij}(\theta_t^k)(\theta_t^k)^T \\
&= \sum_{all\ k} \sum_{t=2}^{T^k} \gamma_{k,t,i,j}(\theta_t^k)^T - \sum_{all\ k} \sum_{j'} \sum_{t=2}^{T^k} \gamma_{k,t,i,j'} a_{ij}(\theta_t^k)(\theta_t^k)^T \\
&= \sum_{all\ k} \sum_{t=2}^{T^k} [\gamma_{k,t,i,j}(\theta_t^k)^T - \sum_{j'} \gamma_{k,t,i,j'} a_{ij}(\theta_t^k)(\theta_t^k)^T] \\
&= \sum_{all\ k} \sum_{t=2}^{T^k} \{[\gamma_{k,t,i,j} - \sum_{j'} \gamma_{k,t,i,j'} a_{ij}(\theta_t^k)](\theta_t^k)^T\} \\
&= \sum_{all\ k} \sum_{t=2}^{T^k} \{[\gamma_{k,t,i,j'} - a_{ij}(\theta_t^k) \sum_{j'} \gamma_{k,t,i,j'}](\theta_t^k)^T\} \\
&= \sum_{all\ k} \sum_{t=2}^{T^k} [\gamma_{k,t,i,j} - \frac{\exp(logA_{ij} + W_{ij}\theta_t^k)}{\sum_{j'} \exp(logA_{ij'} + W_{ij'}\theta_t^k)} \sum_{j'} \gamma_{k,t,i,j'}](\theta_t^k)^T
\end{aligned}
\tag{3.1.14}
$$

So, we can calculate the derivative of the auxiliary function Q with respect to $W_{i,j}^{tr}$ as follows:

$$
\frac{\partial Q}{\partial W_{ij}} = \sum_{all\ k} \sum_{t=2}^{T^k} [\gamma_{k,t,i,j} - \frac{\exp(logA_{ij} + W_{ij}\theta_t^k)}{\sum_{j'} \exp(logA_{ij'} + W_{ij'}\theta_t^k)} \sum_{j'} \gamma_{k,t,i,j'}](\theta_t^k)^T \tag{3.1.15}
$$

where

$$
\gamma_{k,t,i,j} = P(q_{t-1} = i, q_t = j' | \boldsymbol{o}^k, \lambda) \tag{3.1.16}
$$

3.2 Contextual Models for Animation Synthesis

We present now three approaches that we investigated for synthesizing a motion stream from a speech stream. These approaches are based on the various contextual models that we just described in the previous sections.

We start with a system based on contextual HMMs which generalizes the method in [56]. Then we present two new approaches that improve upon this baseline. The first approach is based on Fully Parameterized HMM (FPHMM), which we introduced in previous section 3.2.2. The second approach is a combination of a FPHMM and of a Conditional Random Field (CRF) [65], which has been introduced in Section 3.2.3. In the following we consider a training dataset where every observation sequence is a sequence of frames x_t's that are composed of motion features m_t and of speech feature s_t, i.e. $x_t = [m_t; s_t]$.

3.2.1 Contextual HMMs

To design a speech-to-motion system one may learn one CHMM with speech features as (dynamic) contextual variables (i.e. pdfs are conditioned on speech features) and for observations either motion features or both motion and speech features (as in [56]). Note that when used as contextual variables a good choice is to use short term means of the speech frames computed on a sliding window of length 10 (we note these features $\bar{s}$) while standard speech features are used in observation vectors in case observation include both motion and speech features.

Once such a model is trained one can determine a CHMM on speech only, that we note λ_s, by ignoring *pdf* parameters on motion features. Also one can use the speech signal to define a CHMM on motion whose parameters are modified by the speech stream, we note this model $\lambda_{m/s}$. Actually it is a CHMM with time varying parameters (e.g. the mean of a Gaussian changes with time).

At the synthesis step, speech features are first processed with λ_s to find the most likely state sequence, then we use the *single method* of [113] (cf. section 2.1.3) to synthesize a trajectory along this state sequence with $\lambda_{m/s}$. This approach is somehow close to [56]. However we use contextual HMMs instead of HMMs which allows capturing more complex dependencies between speech and motion, yielding improved synthesis as we will demonstrate.

3.2.2 Fully Parameterized HMMs

In our work, a FPHMM is used to synthesize the motion stream from the speech stream as follows. We first learn a FPHMM that takes speech features as contextual variables and that produces motion features observation. [2] Then, during the training phase, motion and speech streams are both used to learn a FPHMM. During the motion synthesis phase (i.e. the animation generation phase), only the speech stream θ is known. It is used to compute the time dependent transition probabilities and the time dependent altered emission probability distributions in states (cf. e.g. Eq. 3.1.3. Once all the parameters of the model are set, one can compute the most likely state sequence, or one can infer the probability distribution over all state sequences. Finally, from this single state sequence or from the distribution over state sequences, one can synthesize a trajectory using techniques such as in [113].

3.2.3 Combining FPHMMs and CRFs

At last, we have investigated the combination of Fully Parameterized HMMs and of Conditional Random Fields (CRFs) [65], we name this hybrid model FPHMM-CRF. It is inspired from the work in [71]. The FPHMM has the same architecture as in previous cases. It takes speech features as external variables and motion features as observation. The CRF has the same architecture as the FPHMM (same number of states and topology, i.e. same authorized set of transitions...). It takes speech features as input and is used to output the most likely state sequence or a probability distribution over state sequences.

For training, we first learn a FPHMM as described previously in section 3.2.2. Then for every training sequence $\boldsymbol{x}^i$, we determine the most likely state sequence $\boldsymbol{h}_{s^i}$ using only speech features in the motion FPHMM built from λ, $\lambda_{m/s}$. Then the CRF is trained using the set of $(\boldsymbol{s}^i, \boldsymbol{h}_{s^i})$ as a training dataset. That is, the CRF operating on speech is learnt to output the right state sequence for the motion stream in the FPHMM model. The rationale behind this choice is that a CRF, being learnt in a discriminative way, is probably more accurate to predict the right state sequence than the non discriminatively learned FPHMM.

[2] Note that one could have included speech features in observations as we have investigated for contextual HMMs. But it did not yield significant improvements.

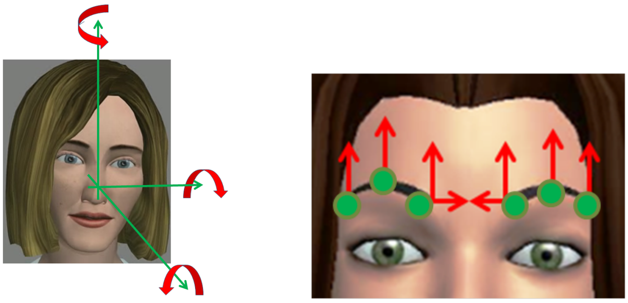


Figure 3.2: Facial motion features - Left: 3 head rotations. Right: Eyebrow animation parameters (arrows illustrate displacements).

To perform synthesis, a speech signal s is first input to the CRF to get a probability distribution over hidden state sequences in $\lambda_{m/s}$. Then speech features are also used to define the parameters of $\lambda_{m/s}$ from the Fully Parameterized HMM. Finally, given the most likely state sequence or the distribution on state sequences output by the CRF, one may synthesize using $\lambda_{m/s}$ a smooth trajectory using either the *single* method or the *integrated* method from [113]. This main benefit of this approach is to take advantage of the discriminative training of the CRF to infer an accurate probability distribution over all hidden state sequences.

3.3 Speech Database on Facial Expression

A human dataset is used to train our animation generators. We use the dataset called Biwi 3D Audiovisual Corpus of Affective Communication database (B3D / AC) [49]. This dataset can be obtained by contacting the authors [49]. To build this dataset, 14 subjects were invited to speak 80 short English sentences. In total, this corpus includes 1109 episodes of spoken speech, each lasting 4.67s long on average (actually we should have 80×14 episodes but few have been disgarded because of bad recording settings). Speech audio signal, 3D head rotations and facial expressions are recorded synchronously. Audio signal is recorded by a studio condenser microphone; 3D head rotations and facial expressions are captured by the real-time 3-D scanner [119] at 25 fps. Facial expression is represented by 3D face geometry made of a total of 23370 facial points. 3D face geometry is shown in Figure 3.3. We used a part of this corpus that corresponds to 240 sentences from three subjects.

Our work is based on the Greta system [94]. As described in section 2.3, the

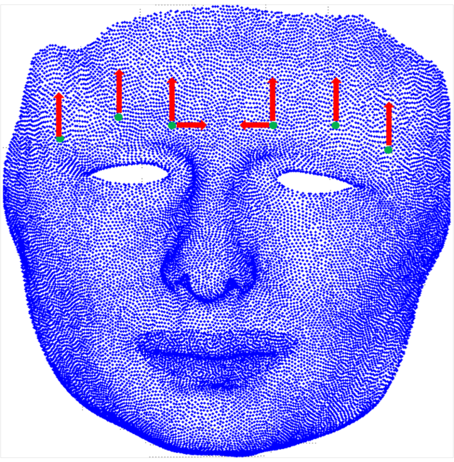


Figure 3.3: Eyebrow animation parameters (Eyebrow FAPs) (arrows illustrate displacements) and 3D face geometry. The 3D face geometry is used in Biwi 3D Audiovisual Corpus of Affective Communication database (B3D / AC) [49]. The 3D face geometry is featured by 23370 facial points.

facial expression in the Greta system can be specified by the Facial Animation Parameters (FAPs) [91]. In speech animation synthesis, we focus on motions of head and eyebrow, hence we extracted a subset of facial motion features that correspond to 3 head rotations and 8 eyebrow features (see Figure 3.2). These features coincide with the Facial Animation Parameters as defined by the norm MPEG-4 [91]. To obtain motion features, the recorded head data is directly used as the 3 FAPs corresponding to the 3 head rotations. The norm MPEG-4 [91] has defined 6 feature points on the eyebrow regions, 3 on each eyebrow. The feature points on the inner side of the eyebrows can move along 2 dimensions (horizontal and vertical) while the other feature points (central and outer side of the eyebrows) can move along 1 dimension. Thus 6 feature points and 8 FAPs are defined for the eyebrow regions. From the face geometries of the B3D / AC database, we select manually 6 face points. In the recorded database, these 6 face points are specified by 18 parameters (each face point is located by 3 coordinates). 8 of 18 parameters correspond to eyebrow FAPs' definition; they are used as eyebrow FAPs. Details can be shown in Figure 3.3.

For the sake of simplicity, we assume both eyebrows move identically and we take the mean of the right and the left eyebrows as the eyebrow motion features. At the end head and eyebrow motion signals are respectively transformed in a sequence of 3-dimensional head feature vector and a sequence of 4-dimensional eyebrow feature vector at a rate of 25 frames per second (fps).

Concerning the speech features we consider 2 prosodic features (pitch and RMS energy), which were extracted with PRAAT software [10] at the same sample rate as for motion feature extraction (25 fps). Both features are related to speech prosody that characterizes speech expressiveness.

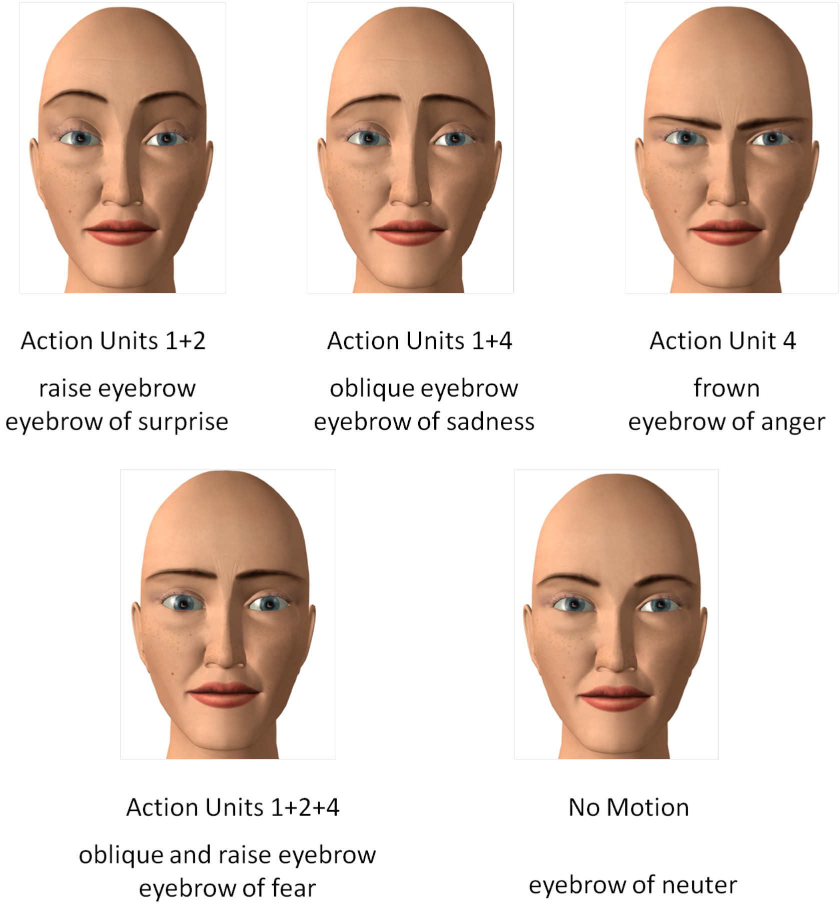
We used augmented feature vectors both for motion and for speech streams by adding first and second order derivatives of static features (i.e. velocity and acceleration). Hence we got 6−dimensional feature vector for speech, 12−dimensional feature vector for eyebrow motion and 9−dimensional feature vector for head motion. In contextual models, the speech feature $\bar{s}$ used as contextual variables are short term means of the speech frames computed on a sliding window of length 10 (found by trials and errors to give the best results).

3.4 Single Modality Motion Synthesis

In this section, we focus on evaluating objectively our approaches. We first detail the experimental settings of the experiments then we compare the results obtained by our approaches and by state of the art methods.

Experimental settings: For the sake of simplicity the evaluation is conducted on a single modality, we investigate eyebrow motion synthesis.

We used 240 videos from three subjects in the dataset and annotated them with respect to five labels $\{c_1,...,c_5\}$ that consist in a combination of Action Units. An Action Unit AU as defined by [45, 47] is a minimal visible muscular contraction (e.g. raise eyebrow). Each AU is precisely defined by changes happening on the face, in term of feature displacement, apparition of bulges, furrows or wrinkles, as well as changes of the skin color [47]. Facial expressions are described as a combination of AUs. We considered the combination of Action Units 1+2+4 (oblique and raise eyebrow, that often corresponds to the eyebrow of fear), the combination of Action Units 1+2 (raise eyebrow, eyebrow of surprise), Action Unit 4 (frown, eyebrow of anger), and the combination of Action Units 1+4 (oblique eyebrow, eyebrow of sadness), plus a fifth additional *no motion*class. Such eyebrow images are shown in Figure 3.4. We annotated videos of the data set using these 5 eyebrow expressions. The annotation was done manually as we did not have access to a facial analysis



Action Units 1+2

raise eyebrow
eyebrow of surprise

Action Units 1+4

oblique eyebrow
eyebrow of sadness

Action Unit 4

frown
eyebrow of anger

Action Units 1+2+4

oblique and raise eyebrow
eyebrow of fear

No Motion

eyebrow of neuter

Figure 3.4: Eyebrow Motion

tool that was precise enough to perform the annotation along these 5 labels. A sequence of observation is then labeled as a sequence of labels (a specific combination of action units) together with their boundaries, just like a speech signal is annotated in phones. Our work presented in this section is inspired by Hofer et al. [56]. These authors have developed head animation generator using HMM. Their method is based on five labels of head motion: nod, shake, silence, shift and pause [56].

Every training sequence consists then in a triple $(\boldsymbol{s},\boldsymbol{m},\boldsymbol{y})$ of a sequence of speech feature vectors (of length T), a sequence of motion feature vectors (of length T) and a sequence of labels $\boldsymbol{y}$ (of length T, with $\forall t, y_t \in \{AU1+2+4, AU1+2, AU4, AU1+4, no\ motion\}$). As introduced in section 3.3, speech and motion feature vectors are respectively 6−dimensional and 12−dimensional at each frame.

3.4.1 Objective Evaluation

We performed experiments with our approaches and with the method in [56] that exploits HMMs. We considered as many models as there are eyebrow motion classes (namely 5). We used an ergodic model for the *no motion* class and left-to-right models for the other classes. We trained the models with a dataset including speech and motion features for each sentence. We first trained independently class models (whatever the models used, HMM, CHMM, PFHMM and PFHMM-CRF) using corresponding segments of training sequences. Then we combined these submodels into a global model which is finally re-estimated on all the sentences.

For the test we used the sequence of speech features only. We primarily evaluated our methods with respect to a reconstruction error, i.e. the mean squared error between the synthesized motion signal (from the speech signal) and the real motion signal (MSE criterion). To gain more insight on the behavior of the methods we also evaluated the methods with respect to their labeling quality, i.e. the recognition of the sequence of labels. We computed the recognition accuracy with respect to the Hamming distance (H criterion) and to the edit distance (E criterion) between recognized and manually annotated sequences of labels. Reported results are averaged results over 20 random splits of the dataset into 80% for training and 20% for testing, together with standard deviation.

Table 3.1 reports the performance, on the test set, of the four methods with respect to the three evaluation criteria and for a number of states per class model ranging from 3 to 7. As can be seen in Table 3.1 our three novel approaches (CHMM, PFHMM and PFHMM-CRF) perform better than conventional HMM used by [56]; the performance with PFHMM-CRF is the best. Table 3.2 reports similar results in a slightly different setting. We computed the same performance criterion as in table 3.1 but in that case the sequence of labels was assumed to be known for every test sequence (but not the time boundaries between labels). Of course the H and MSE obtained here show significant improvements compared to table 3.1 but the gap is not so big. This means that even if the system does not always recognize the correct labels, it does not affect too much the synthesized motion stream.

Our results clearly show that contextual models are significantly better than the benchmark method in the field. In particular, the new approaches, e.g. fully parameterized HMMs, perform significantly better than standard HMMs and than contextual HMMs. In the following section, we will use fully parameterized HMM

Model	#states	MSE	Acc (H)	Acc (E)
HMM [56]	3	0.67 (0.052)	37% (4.7)	45% (4.2)
	5	0.59 (0.042)	43% (4.7)	49% (4.4)
	7	0.56 (0.056)	53% (5.7)	51% (4.3)
Contextual HMM	3	0.51 (0.055)	55% (4.8)	49% (4.4)
	5	0.49 (0.064)	58% (5.7)	50% (4.9)
	7	0.47 (0.056)	59% (4.5)	50% (3.4)
FPHMM	3	0.55 (0.042)	60% (5.3)	57% (4.7)
	5	0.46 (0.051)	61% (5.1)	61% (3.8)
	7	0.45 (0.037)	63% (3.0)	62% (3.7)
FPHMM-CRF	3	0.47 (0.054)	58% (4.2)	60% (3.7)
	5	0.44 (0.061)	61% (4.0)	65% (3.8)
	7	0.39 (0.051)	66% (4.1)	64% (3.7)

Table 3.1: Performance of the models with respect to the synthesis quality (MSE) and to labelling accuracy where accuracy is computed by evaluating Hamming distance (H) and edit distance (E). Performances are averaged results gained on 20 experiments (standard deviations are given in brackets). HMM model is a reference model used by [56]; Contextual HMM model is introduced in section 3.2.1; FPHMM (Fully Parametrized HMM) is introduced in section 3.2.2; FPHMM-CRF model (the combination of Fully Parametrized HMM and Conditional Random Field) is introduced in Section 3.2.3

Model	#states	MSE	Acc (H)
HMM [56]	3	0.43 (0.055)	73% (4.7)
	5	0.39 (0.051)	75% (4.4)
	7	0.36 (0.063)	78% (4.7)
Contextual HMM	3	0.37 (0.057)	77% (5.0)
	5	0.31 (0.061)	81% (4.7)
	7	0.30 (0.061)	82% (5.0)
FPHMM	3	0.33 (0.043)	80% (4.1)
	5	0.28 (0.048)	83% (5.3)
	7	0.25 (0.052)	84% (4.9)
FPHMM-CRF	3	0.31 (0.044)	81% (5.8)
	5	0.26 (0.040)	84% (5.5)
	7	0.23 (0.038)	84% (5.4)

Table 3.2: Similar results as in Table 3.1 but where we assume the sequence of labels of each test observation sequence is known (but not the time boundaries). HMM model is a reference model used by [56]; Contextual HMM model is introduced in Section 3.2.1; FPHMM (Fully Parametrized HMM) is introduced in Section 3.2.2; FPHMM-CRF model (the combination of Fully Parametrized HMM and Conditional Random Field) is introduced in Section 3.2.3

to build synthesis of both eyebrow and head motions from speech. They perform slightly lower than FPHMM-CRF but they are much simpler models and offer a good trade-off between simplicity and accuracy.

3.5 Multiple Modality Motions Synthesis

In this section, we build an eyebrow and head animation model for embodied conversational agent (ECA) using the FPHMM approach which we evaluate through both objective and subjective evaluations.

Contrary to our approach presented in the previous section, here we did not used the strategy of learning as many models as there are motion classes (as in [56]) corresponding here to eyebrow and head motion. Although it was found accurate and successful (see previous section) such a strategy requires an important annotation effort that we wanted to avoid here. Alike in [78] we explored here another strategy which doesn't require any prior manual annotation of the training dataset.

Every training sequence consists then in a pair $(\boldsymbol{s},\boldsymbol{m})$ of two sequences of equal length (T), the first being a sequence of $6-$dimensional speech feature vectors, the

Table 3.3: Performance of the models with respect to the synthesis quality (MSE). Performances are averaged results gained on 50 experiments (standard deviations are given in brackets).

Model	10 states	20 states	30 states	50 states
HMM [78]	0.57 (0.054)	0.53 (0.049)	0.49 (0.061)	0.46 (0.053)
separate PFHMM	0.45 (0.069)	0.41 (0.066)	0.39 (0.059)	0.38 (0.051)
joint PFHMM	0.39 (0.071)	0.34 (0.059)	0.30 (0.045)	0.30 (0.036)

second one being a sequence of 21−dimensional motion feature vectors. Each motion feature vector is composed by 9−dimensional head feature vector and 12−dimensional eyebrow feature vector. More details concerning features of prosody and motions may be found in Section 3.3.

3.5.1 Objective evaluation

We first performed experiments for modeling head and eyebrow features separately with 2 PFHMMs, one for each motion stream. Then experiments were performed to jointly model head and eyebrow features with a single PFHMM, where the joint features of head and eyebrow were considered as FHMM's observation features. We compare our results with the baseline approach proposed in [78] that we have implemented and that we have tuned for the task at hand.

These two sets of experiments were evaluated by computing the reconstruction error defined as the mean square error between the synthesized motion signal (from the speech signal) and the real motion signal (MSE criterion). Table 3.3 reports the experimental performances with different numbers of states based on the averaged results and the standard deviation over 50 random splits of the dataset into 80% for training and 20% for testing. The experiments are configured using full covariance matrix and ergodic topology. As can be seen in Table 3.3, the joint model outperforms the baseline [78] as well as the other two using separate models.

In the joint model, the relationship between eyebrow and head features can be learned through the use of full covariance matrices. In the synthesis phase (trajectory computation), generated eyebrow and head motions are defined not only from the speech features but also from their mutual influence. This influence is captured during the training phase through the combination of covariance matrices. On the other hand, when using two separate models, there is an underlying hypothesis implying that head and eyebrow movements are independent from each other that is carried out during the training phase and that produces poorer results.

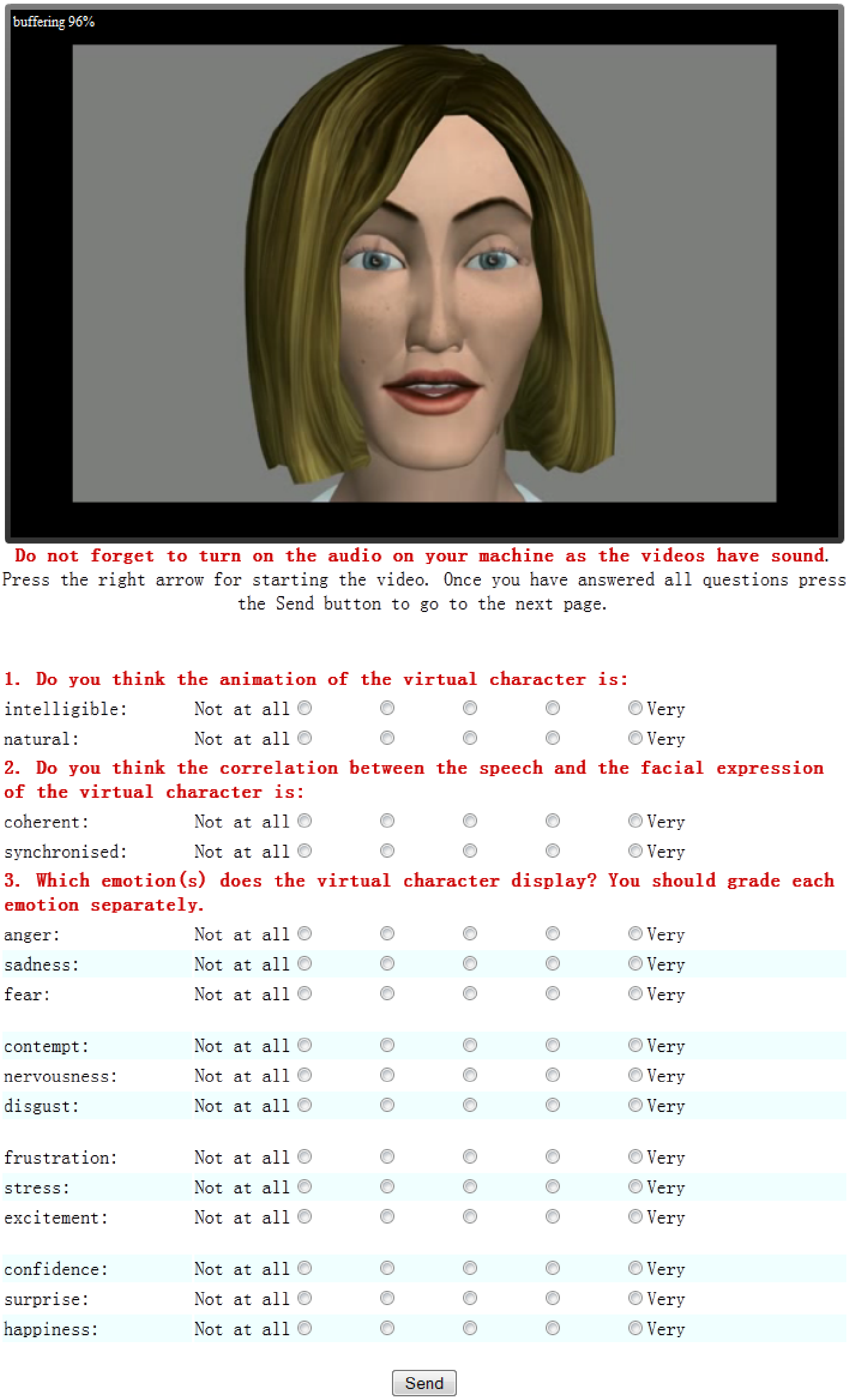
As can be seen in Table 3.3, the joint model with 30 states achieves the best result. We used this model to compute the animation of the virtual agent which we evaluate through subjective evaluation (see next section). It is a necessary complement of the objective measure we just presented since objective measures do not allow to actually measure how well synthesized motions are perceived by human eyes [69].

3.5.2 Subjective evaluation

In this section, we detail the subjective evaluation we conducted with human participants to evaluate the qualitative aspects of FPHMM as generator of head and eyebrow motion from speech signals. The evaluation was done through an online web application. An example of web page is shown in Figure 3.5

Hypothesis The subjective evaluation was conducted to investigate two hypotheses: 1) the perception of the virtual agent displaying head and eyebrow motions synthesized by FPHMM is similar to the perception of the virtual agent animated directly by human data; 2) the two-modal motions (head and eyebrow) outperform single model motion (either head or eyebrow) at a perceptual level. Through the first hypothesis we aim to measure if FPHMM is capable of capturing and of rendering the sophisticated relationship between motion and speech streams and that the animation of a virtual character offers similar results when driven by FPHMM and by real human data. The second one is to verify that multimodal motions facilitate human perception over monomodal motions. To verify the first hypothesis, we compare the perceptions resulted from real and generated motions. To answer the second hypothesis, we compare how animations of the virtual agent driven either by one of the modal motions (either head or eyebrow motions) or by two modal motions (head and eyebrow motions) are perceived.

Our work focuses on nonverbal communication and not on the appearance of the virtual agent. Moreover to eliminate changes of too many variables, we did not use the original data as such. Rather, motions from both, FPHMM types and human data, are displayed through the same identical virtual agent.



Do not forget to turn on the audio on your machine as the videos have sound.
Press the right arrow for starting the video. Once you have answered all questions press the Send button to go to the next page.

1. Do you think the animation of the virtual character is:

intelligible:	Not at all ○	○	○	○	○ Very
natural:	Not at all ○	○	○	○	○ Very

2. Do you think the correlation between the speech and the facial expression of the virtual character is:

coherent:	Not at all ○	○	○	○	○ Very
synchronised:	Not at all ○	○	○	○	○ Very

3. Which emotion(s) does the virtual character display? You should grade each emotion separately.

anger:	Not at all ○	○	○	○	○ Very
sadness:	Not at all ○	○	○	○	○ Very
fear:	Not at all ○	○	○	○	○ Very
contempt:	Not at all ○	○	○	○	○ Very
nervousness:	Not at all ○	○	○	○	○ Very
disgust:	Not at all ○	○	○	○	○ Very
frustration:	Not at all ○	○	○	○	○ Very
stress:	Not at all ○	○	○	○	○ Very
excitement:	Not at all ○	○	○	○	○ Very
confidence:	Not at all ○	○	○	○	○ Very
surprise:	Not at all ○	○	○	○	○ Very
happiness:	Not at all ○	○	○	○	○ Very

Figure 3.5: An example of web page to evaluate the performances of FPHMM or human data. After watching each video clip, each participant was invited to answer questions.

Protocol The participants went on a web page where after answering few questions about themselves, they have to view videos of the virtual agents. Their task consisted in answering few questions. We provide elements of the protocol we follow for the perceptive evaluation study:

(1) Participants: in total, there were 280 participants consisting of 136 males and 144 females with age ranging from 18 to 65 (M=32.89 years, SD=7.99 years).

(2) Stimuli: 7 spoken utterances were randomly selected from the testing database. They were given as input to the trained FPHMM. Then the synthesized motions (sequences of MPEG-4 FAPs frames) and the corresponding WAV file of the spoken utterances are used to drive a virtual agent. In all the animations the lip shapes and body movements of the virtual agents are reproduced from real data. The final animations including eyebrow and head motion as well as lip shape and body movements are stored as video clips.

To study both hypotheses, 7 versions (conditions) of the virtual agent animations were created for each selected sentence. In all conditions the lip shapes and body movements remain constant and are duplicated from real human data:

1^{st} condition (cond1): No eyebrow and head motion;

2^{nd} condition (cond2): Only human eyebrow motion (no head motion);

3^{rd} condition (cond3): Only synthesized eyebrow motion (no head motion);

4^{th} condition (cond4): Only human head motion (no eyebrow motion);

5^{th} condition (cond5): Only synthesized head motion (no eyebrow motion);

6^{th} condition (cond6): Human eyebrow and head motion;

7^{th} condition (cond7): Synthesized eyebrow and head motion;

Therefore, there are a total of 49 video clips (7 sentences × 7 conditions), each of which lasts about 4.5s.

(3) Design and Procedure: Subjective evaluations were conducted online. At first, each participant fills out a demographic questionnaire concerning their age, gender, education level, occupation and country in which participant spent the majority of his/her life. Then, the participant watches 7 randomly selected video clips out of 49. The 7 video clips watched by any participant are comprised of the 7 sentences and of the 7 conditions. After watching each video clip, each participant is invited to answer the following questions using a 5 point Likert scale:

1. Do you think the animation of the virtual character is intelligible?

2. Do you think the animation of the virtual character is natural?

Table 3.4: Results of F-statistic from repeated measures ANOVA

	intelligible	natural	coherent	synchronized
only eyebrow	F=20.6, p<.001	F=1.79, p>.05	F=2.02, p>.05	F=2.57, p>.05
only head	F=1.7, p>.05	F=0.39, p>.05	F=0.02, p>.05	F=2.57, p>.05
eyebrow&head	F=3.51, p>.05	F=0.93, p>.05	F=1.9, p>.05	F=0.01, p>.05

3. Do you think the correlation between the speech and the facial expression of the virtual character is coherent?
4. Do you think the correlation between the speech and the facial expression of the virtual character is synchronized?
5. Which emotion(s) does the virtual character display? You should grade each emotion separately.

The same 12 emotional states as used in the BIWI experiments are considered [49]: anger, sadness, fear contempt, nervousness, disgust, frustration, stress, excitement, confidence, surprise and happiness. Each video clip has been evaluated 40 times (i.e., by 40 participants).

Result To investigate the differences of participants perception from human and from synthesized motions, we conducted three pairwise comparisons in term of intelligibility, naturalness, coherency and synchronization: only eyebrow motion (2nd and 3rd conditions), only head motion (4th and 5th conditions) and both (6th and 7th conditions). The comparison results based on repeated measures ANOVA are shown in Table 3.4. The results show no significant differences between human and generated motions in almost all pairwise conditions except for the only eyebrow condition in term of intelligibility. In this latter case, the mean scores of this pairwise condition are 2.67 and 2.27 for only human and synthesized motions, respectively. While the difference is significant, they are still not too highly different.

Then, to test our second hypothesis (the two-modal motions outperforms single model motion), we compare the 4 different conditions involving synthesized motions with each other, namely: no head and eyebrow motions (1st condition), only synthesized eyebrow motion (3rd condition), only synthesized head motion (5th condition)

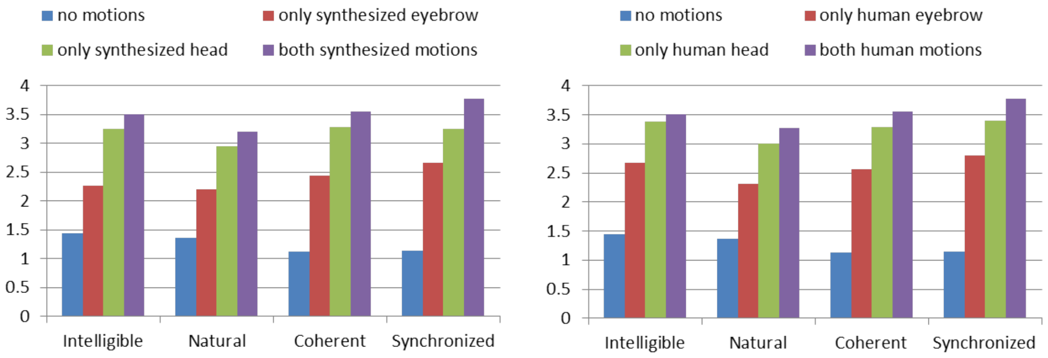


Figure 3.6: Comparison results between animations from synthesized data (left) and human data (right).

and both synthesized (head and eyebrow) motions (7th condition). The comparison results presented in Figure 3.6 show that when eyebrow and head motions are modeled together, the human perception improves in the term of intelligibility, naturalness, coherency and synchronization. The results based on repeated measures ANOVA show significant differences between any two different types of motions among the 4 cases (synthesized eyebrow or head, both synthesized eyebrow and head, no motion of eyebrow and head). The same conclusions are supported by the similar pairwise for the animations from human data (see details in Figure 3.6).

At last, we investigated how synthesized motion conveyed emotional information. To do this, we extracted the scores in term of 12 emotions from both synthesized head and eyebrow motions (7th condition) and from both human motions (6th condition), respectively. Moreover, the BIWI corpus provides the emotion recognition rate for the videos of real humans for these 12 emotions using a 5 point Likert scale. We consider these results as reference when evaluating the virtual agent's performance. Figure 3.7 reports the recognition rate of participants for the virtual agent when driven from human data and from synthesized model, as well as for the videos of real humans. The results are average over the recognition rate for 5 of the 7 sentences spoken by the virtual agent. Two of the sentences have to be disregarded for this test as no result is provided for them in the case of the evaluation of the real human videos. The number of participants differs between the study made with videos of real humans and the study made with videos of virtual agents. In case of the BIWI study, the first set of videos was evaluated when the BIWI corpus was built. There is a very low number of participants that evaluated

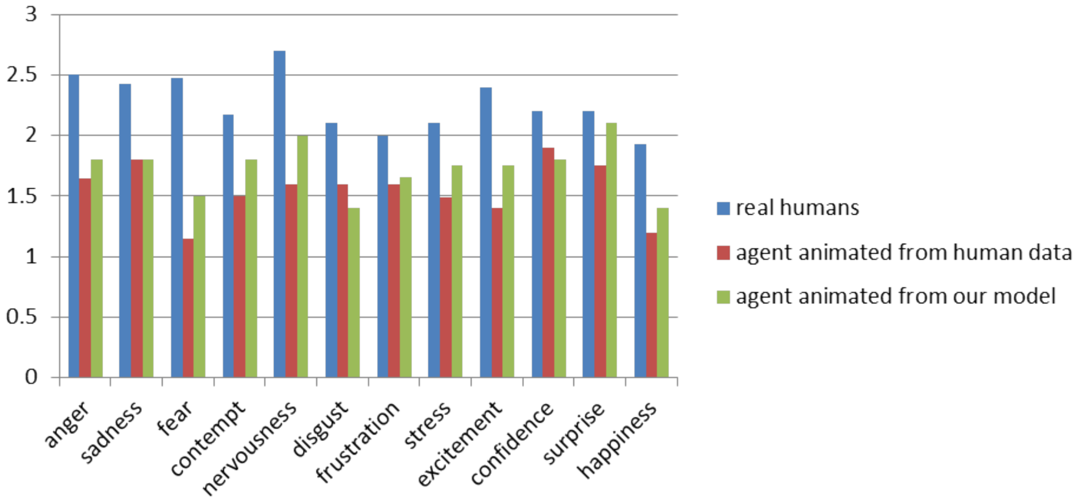


Figure 3.7: Perceived emotions: average values over 5 sentences.

each video (often around 4 participants per sentence) while in our study we have 40 participants evaluating all the sentences in average. As can be seen in Figure 3.7, in all three cases, the faces, be of a real human or of a virtual agent, are able to convey some emotional communication. However, in term of the perception of emotional expressiveness, the videos of the real humans outperform the videos of the virtual character driven from our statistical model; in turn, the videos of the virtual character driven from our statistical model outperform the videos of the virtual character driven from human data.

Due to the too big difference between participants number, we only compare results between the conditions of the virtual agent driven from our model and from human data. The animations with synthesized motions are perceived as showing more emotions in a statistically significant way for the emotion: fear, contempt, nervousness, stress, excitement, surprise, happiness. There is no significant difference for the emotions: anger, sadness, frustration. For the emotions disgust and confidence, the animations from human data are ranked higher than the animations from synthesized data.

Discussion The post-hoc pairwise comparisons between identical modal motions show no significant difference between the human and synthesized motions. In the training phase, the mapping between human audio-visual signals is captured and recorded in FPHMM, and thus rendered in the output synthesized animations.

Therefore the perception of the animation of the virtual agent with synthesized motions is similar to the perception of animation with human motions.

The post-hoc pairwise comparisons of conditions show that there are significant differences among the monomodal and multimodal conditions. The animations created in the multimodal conditions are perceived as more intelligible, natural, coherent and synchronized than the animations created with one modality only. Humans are very skilled in reading nonverbal signals. So the lack of either eyebrow or head motions are negatively perceived along these 4 qualitative dimensions. The comparisons of the perceptions between only eyebrow and only head motions reveals that head motion plays a more important role than eyebrow motion at the perception level. Similar results have also been reported in [78].

Rather than examining the recognition rate of emotions from the videos in different cases, we look at the level of emotional expressiveness. Indeed even in the reference case, namely the videos of real humans speaking the various utterances in emotionally-colored fashion, we remark that the recognition rate level of emotion is rather low and that there is a lot of confusion. That is human participants did not show a strong agreement in their perception of which emotion the human actor aimed to convey when s/he spoke the utterances. Such a result is reproduced with the virtual agent. In both cases, animation of the virtual character from real human data and from our model, a lot of confusion can be noticed. However we can remark that the virtual agent driven by our model is perceived as speaking with emotions with a higher level than when the virtual agent is driven by human data. This can be interpreted as the virtual agent driven by our model is able to exhibit more expressive behaviors; that is, it can communicate in a more emotionally colored manner. Our statistical model relying on FPHMM is capable of capturing the speech/motion relationship. It is also able to render the quality of emotional behaviors: not only its types of movements (i.e. which head movements and eyebrow shapes) but also the dynamism of the movements. Our model computes the types of the visual cues but also their trajectory that carries out dynamics characteristics.

The results of both evaluation studies, the objective and subjective studies, show that our model is able to capture pertinent information that are conveyed through the nonverbal behavior animation of the virtual agent. Considering the link between prosody and both head and eyebrow motions together allows seizing their tight coupling. This is in link with results found in studies from psychology domain

[82, 43].

We can conclude that the machine learning approach (FPHMM) can capture the link between speech prosody and facial movements and can reproduce the movement dynamism in the output synthesized animations. Thus our animation model based on FPHMM is able to gather these both aspects that are important to compute when animating a virtual agent.

3.6 Conclusion on Speech Animation Synthesis

In this section, we have investigated three data-driven approaches to generate head and eyebrow motions for a virtual agent from speech prosody. All of three approaches are variants of HMM: contextual HMMs, fully parameterized HMMs, the combination of Conditional Random Fields and fully parameterized HMMs. They are used to capture the direct mapping between audio and visual information. First, they are objectively evaluated based on (single modality) eyebrow animation synthesis. The results show that they perform better than the standard HMM. Furthermore, fully parameterized HMM is used to synthesize the motion of two modalities, head and eyebrow. We investigate two cases: separate FPHMM and joint FPHMM. In separate FPHMM, motions of head and eyebrow are synthesized independently from one another; in joint FPHMM, motions of head and eyebrow are related to each other in synthesis. The objective evaluation study shows that considering simultaneously eyebrow and head motions increases the precision of the resulting animation. It also confirms that eyebrow and head motions are not independent from each other but rather are connected; the multimodal signals reinforce the communicative meaning. Moreover, the subjective evaluation shows that our proposed model enhances the perception of the virtual agent animation at the level of emotional expressiveness.

Chapter 4

Laughter Animation Synthesis

Laughter is frequently used in human communication. Even-though laughter is an important communicative signal in human interaction, it is only since late 1990 that it started to be studied [105]. Before that, very few studies (e.g., [34]) existed in psychology, neither regarding its behavior description, nor its communicative functions. Laughter is strongly linked to positive emotions and even more to cheerful mood [106]. In particular, the frequency of laughter and its intensity are best predicted for people whose personality trait is to be cheerful and who are currently in cheerful mood [105]. Humans laugh at humorous stimuli or to mark their pleasure when receiving praised statements[101]; they also laugh to mask embarrassment[58] or to be cynical. Laughter can also act as social indicator of in-group belonging [4]; it can work as speech regulator during conversation [100]; it can also be used to elicit laughter in interlocutors as it is very contagious [101].

However laughter is not always linked to cheerfulness. Studies have reported up to 23 different types of laughter [58] ranging from hilarious, hysterical to embarrassed, desperate, contemptuous laughter. That is laughter is linked to various emotional states. Its expressions, at the acoustic and behavioral levels, vary accordingly. The facial expressions and body movements for the different laughter types have hints of the underlying emotions and these variations are perceivable by humans [105][54].

In our work, we focus on hilarious laughter that is laughter following a funny event and triggered by amusing and positive stimuli (e.g., a joke). Hilarious laughter is one among the 23 types of laughter defined by [58]. Laughter morphology involves facial expressions, body movements and vocalizations [105]. For hilarious laughter, muscular activities include mainly the zygomatic major, mouth opening and jaw movement. Orbicularis oculi is squinted as for Duchenne smile [46]. Eyebrows may be raised or even frown in very intense laughter [105]. Saccadic movements affect the whole body. Torso may bend back and forth and shoulder may shake. Changes in respiration patterns are also prominent. Inhalation and exhalation phases are very noticeable. All these movements are done very rhythmically. They are also highly correlated. Indeed they arise from the same physiological processes [105].

Embodied conversational agents ECAs are autonomous virtual agents able to converse with human interactants. Our aim is to develop an embodied conversational agent able to laugh. To achieve our aim to simulate laughing agent, we ought to propose approaches to reproduce the multimodal signals of laughter. Niewiadomski and Pelachaud [87] indicate that synchronization among all the modalities is crucial for laughter animation synthesis. Humans are very skilled in reading nonverbal behaviors and in detecting even small incongruences in synthesized multimodal animations. We have developed statistical models for laughing multimodal behaviors synthesis. The statistical model is first trained on human data and then is used to synthesize multimodal animations. We present successively two approaches that we designed and that we evaluated. In both cases the input of the system is the audio signal together with the corresponding sequence of laughter phonemes [1] and of their duration.

The first work is detailed in section 4.1. It involves synthesis of facial expression (lip, jaw, eyebrow, eyelid and cheek) and extrapolated movements of head, shoulders and torso. The animation generators are built based on a human dataset, called AudioVisual LaughterCycle (AVLC) database [114]. This dataset contains human laughter episodes. Each episode corresponds to trajectory data of the head movements and facial expression as well as a laughter audio file, while the torso motion has unfortunately not been recorded, hence such motion cannot be learned as other motion signals. Indeed the torso animation is deterministically infered from the synthesized head motion, this strategy is based on the assumption that there exists a relation between head and torso movements.

The second work is detailed in section 4.2. It focuses on head and torso animation synthesis only which was not actually investigated in the first approach by

[1] [115] has categorized audible information from laughter into 14 laughter phonemes, small sound units, according to human hearing perception.

lack of training data. To build such a system we recorded a new human laughter dataset and gathered the trajectory of torso and head motions as well as the audio signal. Using this new dataset we developed a new animation generator for torso and head and evaluate it in a global animation system by reusing subparts of the first system.

4.1 A first Facial Expression and Upper Body Animation

This section describes our first approach for building animation generator of hilarious laughter. The animation generator can synthesize multimodal signals that include facial expression (eyebrow, eyelid, cheek, lip and jaw), movements of head, torso and shoulder; it takes audio signal of laughter as input. The proposed generator learned first the relationship between the acoustic features of laughter and the face signals based on human data; then the generator is used to produce automatically synthesis animation; that is, given laughter audio as input, it generates the correlated laughter facial expression in real time. Lip and jaw motion synthesis are based on linear regression models; head, eyebrow, eyelid and cheek motion synthesis is based on selecting and concatenating segments of motion capture data of human laughter. In our first model we rely on a database that contains motion data of only the head and the face. To be able to compute the body (torso and shoulder) motion, we made the (strong) hypothesis that head and body motions are correlated; and more particularly that body animation can be inferred from the synthesized head animation.

In the remaining of this subsection we first describe the used dataset and introduces the used features in subsection 4.1.1. Then we present motion synthesis module in section 4.1.2, which contains lip animation synthesis in section 4.1.2.1, head movement and upper face expression in section 4.1.2.2 and shoulder and torso animation synthesis in section 4.1.2.3. Finally, we end this chapter by presenting the evaluation of our animation generator we have conducted (see section 4.1.3).

4.1.1 Laughter Database on Facial Expression

Our work makes use of the AVLaughterCycle database [114]. This database contains more than 1000 audiovisual spontaneous laughter episodes produced by 24 subjects. 66 facial landmarks coordinates were detected by an open-source face tracking tool - FaceTracker [108]. Among these 66 landmarks, 22 of them correspond to the Facial Animation Parameters FAPs of MPEG-4 [91] for the lips and 8 landmarks for the FAPs for the eyebrows.

In this database, subjects are seated in front of a PC and a set of 6 cameras. They watch funny movies for about 15mn. Their facial expressions, head movements and laughter are then analyzed using FaceTracker. However body behaviors (e.g. torso and shoulders behaviors) are not recorded in this database. 24 subjects were recorded but only 4 subjects had their head motion tracked. Therefore, a sub dataset of 4 subjects with head motion data is used in our work.

This database includes acoustic data of laughter. In addition it contains the segmentation of laughter into small sound units. [115] has categorized audible information from laughter into 14 laughter phonemes according to human hearing perception. These 14 laughter phonemes correspond to (number of occurrences of these laughter phonemes are specified in parentheses): silence(729), ne(105), click(27), nasal(126), plosive(45), fricative(514), ic(162), e(87), o(15), grunt(24), cackle(10), a(144), glotstop(9) and vowel(0) [2]. So laughter is segmented into sequences of laughter phonemes and their durations. Laughter prosodic features (such as energy and pitch) have been extracted using PRAAT [10] and are provided with the database.

In our model we focus on face and head motion synthesis from laugher phoneme sequence (e.g. [a, silence, nasal]) and their duration (e.g. [0.2s, 0.5s, 0.32s]). We take prosodic features as additional inputs for lip and jaw motion synthesis. Section 4.1.2.1 and Section 4.1.2.2 provide details on our model. Since the AVLaughterCycle database does not contain any data on torso movement, we base our torso animation model on the observations that head and torso movements are correlated. We build a PD controller that extrapolates torso movement from head motion as explained in Section 4.1.2.3.

[2] In the sub-dataset of the 4 subjects, no vowel laughter phoneme was found. However it does appear in other data.

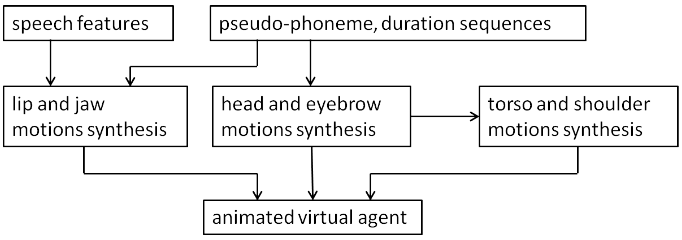

speech features
pseudo-phoneme, duration sequences
lip and jaw motions synthesis
head and eyebrow motions synthesis
torso and shoulder motions synthesis
animated virtual agent

Figure 4.1: Overall architecture of multimodal behavior synthesis.

4.1.2 Motion synthesis

Figure 4.1 illustrates the overall architecture of our multimodal behavior synthesis. Our aim is to build a generator of multiple outputs (lip, jaw, eyelid, cheek, head, eyebrow, torso and shoulder motions) from an input sequence of laughter phonemes together with their duration and from speech prosodic features (i.e. pitch and energy). We briefly motivate our choices for developing three different synthesis models adapted to different modalities; then we present in details these three models.

Lip articulator: First to accurately synthesize lip and jaw motions, which play an important role in articulation, we exploit all our inputs, namely the speech features and the laughter phoneme sequence, in a new statistical model that we describe in section 4.1.2.1. Using speech features as input yields an accurate synthesized motion that is well synchronized with speech, which is required for high quality synthesis.

Face and head: Secondly, although it has been demonstrated in the past that speech features allow accurate prediction of head motion and upper face expression for normal speech [19, 78, 42, 41, 73, 14], the relationship between speech features and facial motion of laughter is unknown. Moreover exploring our laughter dataset we found that some segments have significant head motion and upper face expression while they are labeled as unvoiced segments. We then turned to exploit a more standard *synthesis by concatenation* method that we simplify to allow real time animation. Our method is described in section 4.1.2.2.

Upper body: At last, body (torso and shoulder) motion, which is an important feature of laughter [105], is determined in a rather simple way from the synthesized head motion output by the algorithm in section 4.1.2.2. The main reason for doing so is that there is no torso and shoulder motion information gathered in the AVLaughterCycle dataset so that none of the two synthesis methods above may be used here. Moreover we noticed in our dataset a strong correlation between head move on the one hand and torso and shoulders moves on the other hand. We then decided to hypothesize a simple relationship between the two motions that we modeled with a proportional-derivative (PD) controller. We present such a model in section 4.1.2.3.

4.1.2.1 Lip and Jaw Synthesis Module

To design the lip and jaw motion synthesis system, we used a linear regression model. Basically the underlying idea of a linear regression model is to estimate a desired quantity x (the lip and jaw motion) as a linear function of an observed quantity θ (the speech features). Such a linear regression model is defined as follows:

$$\hat{\mu}(\theta) = W^{\mu}\theta + \bar{\mu}_j \tag{4.1.1}$$

$$\tag{4.1.2}$$

where W^{μ} is a $d \times c$ matrix, and $\bar{\mu}$ is an offset vector. θ stands for the value of the contextual variables.

We use one such linear regression model for each of the 14 laughter phonemes so that we get a set of 14 linear regression models. Somehow, each linear regression model is learned to model the dependencies between the lip/jaw motion and the speech features from a collection of training pairs of speech features and of lip and jaw motion.

The linear regression model of a laughter phoneme is learned through the method of least squares. For compact notation, we first define the matrix $Z^{\mu} = [W^{\mu} \;\; \bar{\mu}]$ and the column vector $\Omega_t = [\theta_t \;\; 1]^T$. Equation 4.1.1 can then be rewritten as $\hat{\mu}(\theta_t) = Z^{\mu} \times \Omega_t$. The sum of the squares of the errors, Q, is defined as follows:

$$Q = \sum_{all\ k,t} (\mu(\theta_t^k) - o_t^k)^T (\mu(\theta_t^k) - o_t^k) \tag{4.1.3}$$

which can be carried out as follows:

$$\begin{aligned}
Q &= \sum_{all\ k,t} (Z\Omega_t^k - o_t^k)^T (Z\Omega_t^k - o_t^k) \\
&= \sum_{all\ k,t} ((Z\Omega_t^k)^T - (o_t^k)^T)(Z\Omega_t^k - o_t^k) \\
&= \sum_{all\ k,t} ((Z\Omega_t^k)^T Z\Omega_t - (o_t^k)^T Z\Omega_t - (Z\Omega_t^k)^T o_t^k + (o_t^k)^T o_t^k) \\
&= \sum_{all\ k,t} ((Z\Omega_t^k)^T Z\Omega_t - (o_t^k)^T Z\Omega_t - ((Z\Omega_t^k)^T o_t^k)^T + (o_t^k)^T o_t^k) \\
&= \sum_{all\ k,t} ((Z\Omega_t^k)^T Z\Omega_t - (o_t^k)^T Z\Omega_t - (o_t^k)^T Z\Omega_t^k + (o_t^k)^T o_t^k) \\
&= \sum_{all\ k,t} ((Z\Omega_t^k)^T Z\Omega_t - 2(o_t^k)^T Z\Omega_t + (o_t^k)^T o_t^k)
\end{aligned} \tag{4.1.4}$$

Now, we may calculate $\frac{\partial Q}{\partial Z}$ as follows:

$$\frac{\partial Q}{\partial Z} = 2 \sum_{all\ k,t} Z\Omega_t^k (\Omega_t^k)^T - 2 \sum_{all\ k,t} o_t^k (\Omega_t)^T \tag{4.1.5}$$

Setting $\frac{\partial Q}{\partial Z}$ to 0 we get:

$$Z = \left[\sum_{all\ k,t} o_t^k (\Omega_t^k)^T \right] \left[\sum_{all\ k,t} \Omega_t^k (\Omega_t^k)^T \right]^{-1} \tag{4.1.6}$$

where we considered a single training sequence case and the sum ranges over all indices in the sequence.

At synthesis time one has as inputs a series of speech features and a sequence of laughter phonemes together with their duration. The synthesis of the lip and jaw motion is performed independently for every segment corresponding to a laughter phoneme of the sequence then the obtained signal is smoothed at articulation between successive laughter phonemes. One can adopt few techniques to synthesize the lip and jaw motion segment given a laughter phoneme (with a known duration) and speech features.

A first technique consists in relying on a synthesis method that has been pro-

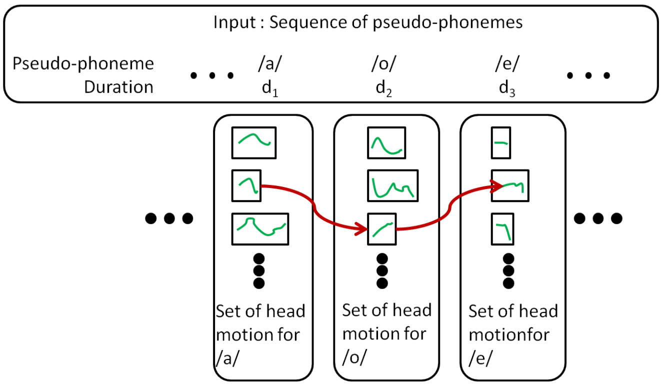

Input : Sequence of pseudo-phonemes
Pseudo-phoneme
Duration
/a/
d1
/o/
d2
/e/
d3
Set of head motion for /a/
Set of head motion for /o/
Set of head motionfor /e/

Figure 4.2: Head and upper face synthesis is performed by the concatenation of motion segments, gathered from real data, corresponding to a given laughter phoneme sequence and their duration. Green curve are samples of motion segments while the red arrow indicates the sequence of selected motion segments. The chosen motion segment is the one that minimizes a cost function of fit with the sequence of laughter phonemes.

posed for Hidden Markov Models by [113] which yields smooth trajectories. Alternatively, a simpler approach consists in using the speech features θ_t at time t to compute the most likely lip and jaw motion using a regression model, i.e. $\mu(\theta_t)$. This is the approach we used in our implementation to ensure real time synthesis.

4.1.2.2 Head and Upper Face Synthesis Module

Our approach to head and upper face synthesis system is a synthesis by concatenation approach. The motivation for using a different strategy than for animation synthesis from speech comes from the fact that preliminary results using the same approach as thos presented in previous chapter did not work at all, one reason being that often we laugh without making any noise.

The synthesis by concatenation approach consists in selecting and concatenating motions from original data corresponding to the input laughter phonemes sequence. This may be done provided one has a large enough collection of real motion segments corresponding to every laughter phoneme. Such data are available from the AVLaughterCycle database [114] which includes head and upper face motion data and which has been manually labeled into laughter phoneme segments. Actually for each of the 14 laughter phoneme labels, pp_i, we have a number N_i of head and upper face real moves that we note $S_i = \left\{m^i_j, j = 1..N_i\right\}$.

For a given laughter phoneme sequence of length K, $(p_1,...p_K)$ (with $\forall k \in 1..K, p_k \in \{pp_1,...,pp_{14}\}$), noting $d(p_k)$ the duration of the k^{th} laughter phoneme in the sequence, the *synthesis by concatenation* method aims at finding the best sequence of segments $(s_1,s_2,...,s_K)$ belonging to $S_{p_1} \times S_{p_2} \times ... \times S_{p_K}$ (with $d(s_k)$ the duration of the segment s_k) such that a cost function (that represents the quality of fit between the segment sequence and the laughter phonemes sequence) is minimized. Figure 4.2 illustrates our head and upper face synthesis framework. In our case the cost function is defined as:

$$C[(s_1,s_2,...,s_K),(p_1,p_2,...,p_K)] \tag{4.1.7}$$

$$= \gamma \sum_{u=1..K} C_{Dur}(d(s_u),d(p_u)) \tag{4.1.8}$$

$$+ (1-\gamma) \sum_{u=2..K} C_{Cont}(s_{u-1},s_u) \tag{4.1.9}$$

where C_{Dur} is a *duration* cost function that increases with the difference between the length of a segment and the length of the corresponding laughter phoneme, and where C_{Cont} is a *continuity* cost function that increases with the distance between the last position of a segment and the first position of the following segment, and where γ is a manually tuned parameter (between 0 and 1) that allows weighting the importance of continuity and duration costs.

The two elementary cost functions are defined as follows, there are illustrated in Figure 4.3:

$$C_{Dur}(d,d') = e^{|d-d'|} - 1 \tag{4.1.10}$$

and:

$$C_{Cont}(s,s') = \left\| last(s) - first(s') \right\|^2 \tag{4.1.11}$$

where $first(s)$ and $last(s)$ stand for the first and the last positions in segment s.

Once a sequence of segments $(s_1,s_2,...,s_K)$ has been determined the synthesis of head and eyebrow motion corresponding to the laughter phoneme sequence requires some processing. Indeed the selected segments' duration may not be exactly the same as the laughter phonemes' duration. Selected segments are then linearly

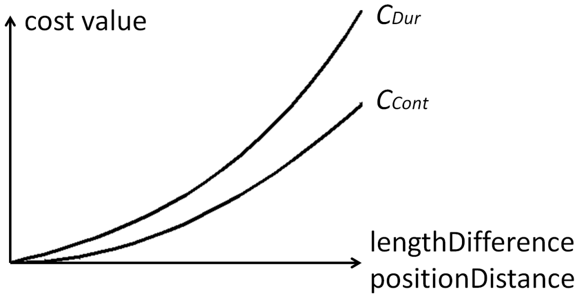


Figure 4.3: Shape of the duration cost function $C_{Dur} = f(v) = e^{v} - 1$ and of the continuity cost function $C_{Cont} = g(v) = v^2$ as a function of their argument v.

stretched or shrank to obtain the required duration. Note that it is assumed that stretching and shrinking of segment motion have no effect on human perception as long as segment duration has minimal variation. Also it may happen that there is a significant distance between the last frame of a segment and the first frame of the next segment which would yield discontinuous moves. To avoid this we perform a local smoothing by linear interpolation at the articulation between two successive segments.

Note that to allow real-time animation, we use a simplified version of the *synthesis by concatenation* method by selecting iteratively the first segment, then the second, then the third according to a *local* cost function focused on the current segment s, $\gamma C_{Dur}(d(s),d(p))+(1-\gamma)C_{Cont}(s',s)$ where p stands for the current laughter phoneme, whose duration is $d(p)$, and s' stands for the previous segment. The obtained sequence of segments may then not be the one that minimizes the cost in Eq. (4.1.7), it is an approximation of it.

Note finally that the duration cost increases much quicker than the continuity cost (see Figure 4.3), which is wanted since as we said previously stretching and shrinking are tolerable only for small factors, while smoothing the end of a segment and the beginning of the following segment is always desirable to avoid discontinuous animation. Defining the cost functions as in equations (4.1.10) and (4.1.11) strongly discourages high stretching and shrinking factors.

4.1.2.3 Torso and Shoulder Synthesis Module

As we explained before torso and shoulder motion is synthesized from the synthesized head motion which is output by the algorithm described in the previous section. Although Ruch and Ekman [105] reported torso and shoulders motions are important components of laughter, there is no such motion data in the AVlaughtercycle corpus. Thus the synthesis method used for lip and jaw or for head and facial expressions, since it requires training data, cannot be exploited here. Yet, through careful observation of the AVlaughtercycle dataset we notice a strong correlation between torso and head movements. For instance we did not find any case where torso and head are going in opposite direction. Thus we hypothesize that torso and shoulder motion follows head motion and that a simple prediction module may already perform well for natural-looking animation.

Based on these observations, torso and shoulder movements of the virtual agent are synthesized from head movements. In more details, we define a desired intensity (or amplitude) of each torso and shoulder movement which is decided by the head movement. This desired intensity is the desired value in a Proportional-Derivative (PD) controller. We choose to use a PD controller (illustrated in Fig 4.4) since it is widely used in graphics simulation domain [85], which is a simple version of Proportional-Integral-Derivative controller (PID) in classical mechanics. The PD controller ensures smooth transitions between different motion sequences and removes discontinuity artifacts.

The PD controller is defined as:

$$\tau = k_p(\alpha_{current} - \alpha) - k_d\overline{\alpha}$$

where τ is the torque value, k_p is the proportional parameter, $\alpha_{current}$ is the current value of the head pitch rotation (ie vertical rotation as in head nod), α is the previous head pitch rotation, k_d is the derivative parameter, $\overline{\alpha}$ is the joint angle velocity. At the moment, we defined manually, by trial and error, the parameters of the PD controller.

We defined two controllers, one for torso joints (vt3, vt6, vt10, vl2) and one for shoulders joints (acromioclavicular, sternoclavicular) which are defined in MPEG4 H-ANIM skeleton [91]. The other torso joints are extrapolated from these 4 torso joints. To avoid any "freezing" effect we added a Perlin noise [97] on the 3 dimensions of the predicted torso joints.

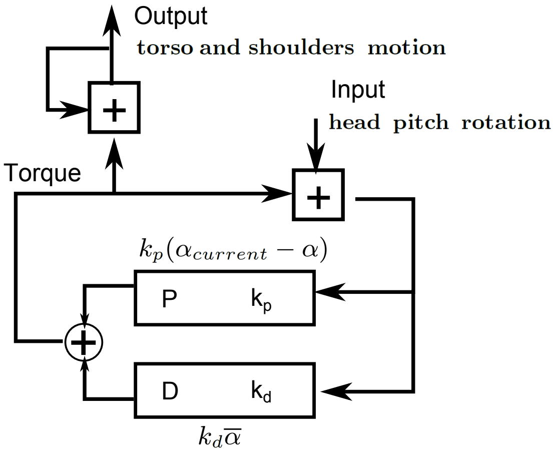


Figure 4.4: PD controller is used to compute torso and shoulders motion for each frame. Input: current head pitch rotation; Output: torso and shoulders joints

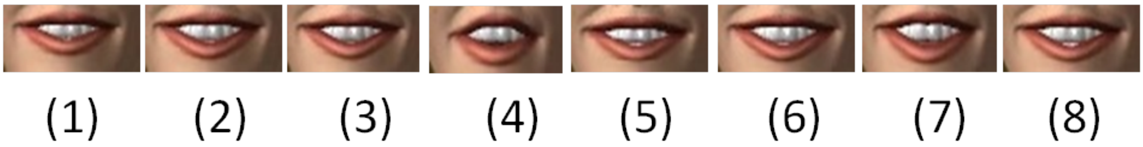


Figure 4.5: Synthesized lip, front view

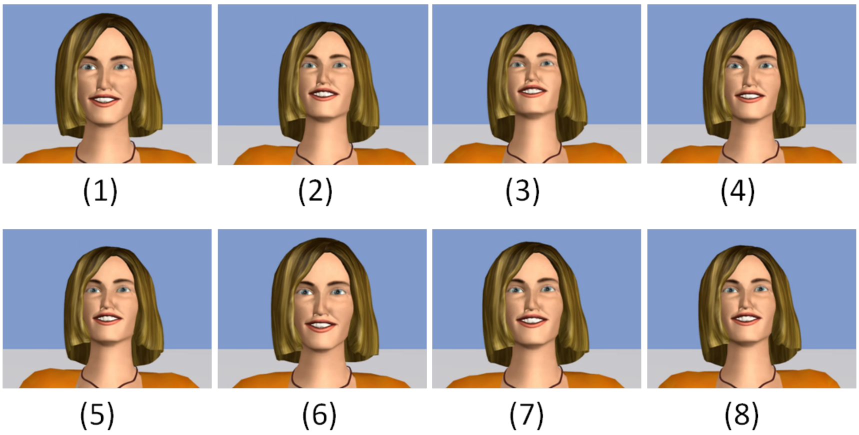


Figure 4.6: Synthesized data, front view

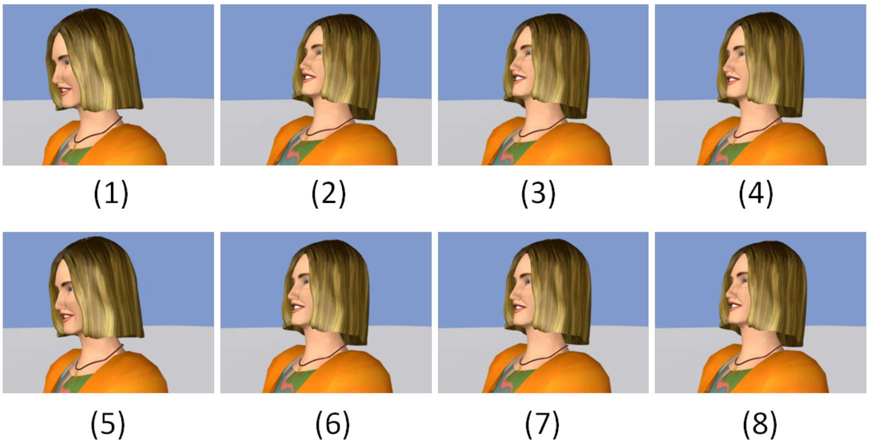


Figure 4.7: Synthesized data, side view

4.1.3 Experiments

In this section we describe examples of laughter animations. We also present an evaluation study where the agent and human participants exchange riddles. The input to our motion synthesis model includes laughter phonemes sequence, each phoneme duration and audio features (pitch and energy) sequence. Our motion synthesis model generates facial expression synchronized with laughter audio in real time. Figure 4.5, Figure 4.6 and Figure 4.7 present several frames of the animation synthesized by our approach.

Our next step is to measure the effect of these laughs on partners of an interaction with a laughing agent. For this purpose, we have conducted a study to test how users perceive laughing virtual character when interacting with it. This study has been thought as a step further of Ochs and Pelachaud's study on smiling behavior [90] (see below for a short description): the smiling behaviors used in [90] are used as the control condition; that is the virtual character smiles instead of laughing.

Considering the type of behavior that we want to test, i.e. laugh, the experimental design of [90] is particularly appropriate. Indeed, in order to explore the effect of amusement smiling behaviors on users' perception of virtual agents, the authors chose positive situations to match the types of smile: in their experiment, the agent asks a riddle to the users, make a pause and give the answer. We use the four jokes and the description of polite and amused smiles of [90]'s evaluation study.

We have conducted an experimental study to evaluate how users perceive how a virtual character laughs or smiles when, either telling a riddle, or listening to a riddle. We consider the following conditions: when the virtual character tells the riddle and laughs or smiles, and when the human user tells the riddle and the virtual character laughs or smiles. Thus, we have two "test conditions" which are the laughing conditions, when speaking or listening, and two "control conditions" which are the smiling conditions, when speaking or listening.

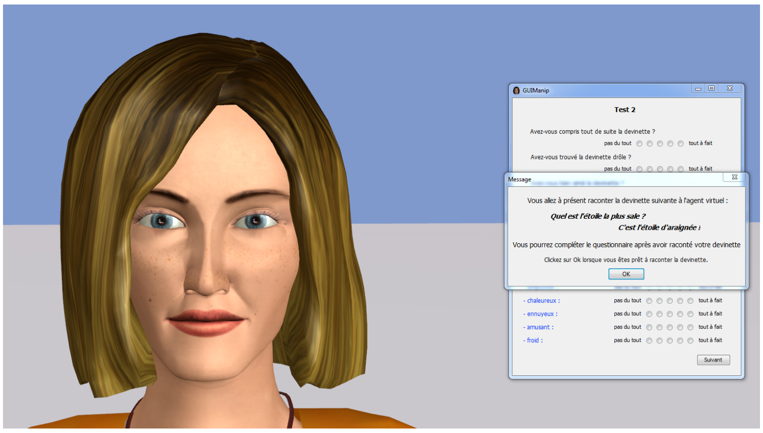
Hypotheses Our hypotheses are:

- the evaluation of the agent's attitude: we expect that the agent which laughs when the human user tells a riddle will be perceived as warmer, more amused, more positive than the agent which only smiles.
- the evaluation of the riddle: we expect that when the agent laughs to the user's riddle, the user will evaluate "his/her" riddle as funnier.

4.1.3.1 Setup

The main constraint for our evaluation is to have real time reaction of the agent to the human user's behavior. This constraint exists mainly in the listening agent condition in which the user tells the riddle and the agent has to react at appropriate time, i.e. at the end of the riddle. As a consequence for the design of our study, we cannot use pre-recorded videos of the agent's behavior and thus, we cannot perform the evaluation using a web application as in [90]. We performed the evaluation in our lab.

Participants sit on a chair in front of computer screen. They wear headphones and microphone and have to use the mouse to start each phase of the test and to fill in the associated questionnaires (see Figure 4.8).



Test 2

Avez-vous compris tout de suite la devinette ?
pas du tout ○ ○ ○ ○ ○ tout à fait

Avez-vous trouvé la devinette drôle ?
pas du tout ○ ○ ○ ○ ○ tout à fait

Vous allez à présent raconter la devinette suivante à l'agent virtuel :

Quel est l'étoile la plus sale ?
C'est l'étoile d'araignée !

Vous pourrez compléter le questionnaire après avoir raconté votre devinette

Clickez sur Ok lorsque vous êtes prêt à raconter la devinette.

- chaleureux : pas du tout ○ ○ ○ ○ ○ tout à fait
- ennuyeux : pas du tout ○ ○ ○ ○ ○ tout à fait
- amusant : pas du tout ○ ○ ○ ○ ○ tout à fait
- froid : pas du tout ○ ○ ○ ○ ○ tout à fait

Figure 4.8: Screenshot of experiment interface.

Each participant saw four riddles in the four conditions, alternating speaking and listening conditions. Here is an example of the sequence of conditions that a participant can have: Agent speaks and smiles, Agent listens and laughs, Agent speaks and laughs, Agent listens and smiles. These sequences of condition are counter balanced for each participant to avoid any effect of their order.

Questionnaires To evaluate how is the act of telling a riddle perceived when the agent listens to the user's riddle and when the agent tells a riddle to the user, we used a questionnaire similar to [90]. After watching each condition, the user has to rate two sets of factors on five degrees Likert scales:

- 3 questions: Did the participant find the riddle funny? How well s/he understood the riddle? Did s/he like the riddle?
- 6 questions related to the perceived stance of the virtual character. Stance is defined in Scherer [28] as the "affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, colouring the interpersonal exchange in that situation (e.g. being polite, distant, cold, warm, supportive, contemptuous)". We used positive qualifiers for the stance of the virtual agent:
  - Is the speaker-agent: spontaneous, warm, amusing?
  - Is the listener-agent: spontaneous, warm, amused?

  We used negative qualifiers:

  - Is the speaker-agent: stiff, boring, cold?
  - Is the listener-agent: stiff, bored, cold?

  For the stance, the questions are of the form: Do you think the agent is stiff/-cold...?

Speaking agent condition A message pop-up on the screen explaining that the agent will tell a small riddle and that the questionnaire will need to be filled just afterwards. When the user clicks on the "ok" button, the agent tells the riddle (and then smiles or laughs depending on the condition). Then the user fills in the questionnaire.

Listening agent condition A message pop-up on the screen, with a short riddle (two lines) and explaining that the user has to tell this story to the agent and that the questionnaire can be filled just afterwards. When the user clicks on the "ok" button, the text of the joke disappears, the user tells the story to the agent; the agent either smiles or laughs at the joke, depending on the condition. In the listening agent condition, the speech and pauses of the human participants are detected to automatically trigger the smiles and laughs of the agent at appropriate time, ie at the end of the riddle. After having told the riddle, the user fills in the questionnaire.

4.1.3.2 Virtual agent's behavior and conditions

To evaluate the impact of agent's laugh on user's perception of the agent and of the riddle, we have considered four conditions.

- Two "test conditions" which are the laughing conditions:
  - the virtual character asks the riddle and laughs when it gives the answer of the riddle.
  - the virtual character listens to the riddle and laughs when the participant gives the answer.
- Two "control conditions" which are the smiling conditions:
  - the virtual character asks the riddle and smiles when it gives the answer of the riddle.
  - the virtual character listens to the riddle and smiles when the participant gives the answer.

Riddles Both the virtual character and the human user tell their riddle in French. When translated into English the joke is something like: “What is the future of I yawn? (speech pause) I sleep!”. In this riddle, the answer of the riddle is “I sleep”. We consider four riddles chosen from [90]. According to [90] the selected four riddles are rated equivalently on their level of funniness.

Smiles The smiles synthesized here correspond to the smiles validated in [90]. We used a polite smile for the question part of the riddle and an amused smile at the end of the answer.

Laughs The laughs that are used in the experiment are the two laughs that were described at the beginning of section 4.1.3.

4.1.3.3 Participants

Seventeen individuals participated in this study (10 female) with a mean age of 29 (SD = 5.9). They were recruited among the students and professors of our University. The participants have all spent the majority of the last five years in France and were mainly native from France (N=15). Each participant took all the four conditions. In the next section, we present in details the results of this test.

4.1.3.4 Results

To measure the effects of laughs on the user’s perception, we have performed repeated measures ANOVA (each participant saw the four conditions) and the post hoc Tukey’s test to evaluate the significant differences of rating between the different conditions (agent Speaks and Smiles (SS), agent speaks and laughs (SL), agent listens and smiles (LS), agent listens and laughs (LL)).

No significant differences were found between conditions for *Understanding* and *Finding funny* the riddle. No significant differences were found between conditions for the agent’s *Spontaneous* and *Stiff*. Significant differences between conditions were found for the other variables: How much the agent finds the riddle funny ($F = 1.3, p < .001$), How much the agent is stiff ($F = 3.8$, $p < 0.05$), warm ($F = 6.58$, $p < .001$), boring/bored ($F = 6.23$, $p < .001$), enjoyable/amused ($F = 6.31$, $p < .001$) and cold ($F = 5.46$, $p < .001$).

The post-hoc analysis on the significant results are presented in Table 4.1. For each conditions pair we report results to items of the questionnaire that were given to the participants for which significant differences were found. Thus we do not report results for the various conditions presented just above (e.g. Understanding, Finding Funny the riddle) for which no significant differences between conditions were found. We also report only the results for the qualifier *Stiff* as no significant difference has been found between *Stiff* and *Spontaneous*. In the Table 4.1, the first column indicates which conditions are compared (agent Speaks and Smiles (SS), agent speaks and laughs (SL), agent listens and smiles (LS), agent listens and laughs (LL)) and the first line indicates the concerned variables. The other columns are the positive and negative qualifiers for speaker-agent and listener-agent (e.g., bored /boring). The second column indicates results regarding if the agent liked the riddle (either told by the participant or by itself, depending on the condition). The inside elements of the table correspond to the condition in which the variable is significantly higher (n.s. means non significant, *: p < .05, **: p < .01, ***: p < .001). If in a comparison, no significant differences are found, we mark n.S.; while if there are significant differences, we indicate the condition with a higher result followed by the number of stars that gives the confidence level of the results. For instance, in Table 4.1, the notation LL*** at the intersection of the line LL-LS and the column Warm means that, the agent when it Listens and Laughs is perceived significantly warmer (with p < .001) than when it Listens and Smiles.

4.1.4 Discussion

Listening conditions The results of the second line of Table 4.1 (LL-LS) tend to show that a listening agent which laughs at the joke of the user is perceived significantly more positive (warmer, more amused, less bored and less cold) than if it only smiles. When it listens, smiling agent appears to be negatively perceived (agent is considered as bored and cold).

Consistently with this result, participants expressed disappointment when the agent did not laugh at their joke (i.e. condition user tells a joke) and satisfaction when the agent did laugh to their joke.

Speaking conditions By contrast, the results of the first line of Table 4.1 (SL-SS) tend to show that there is not much effect of smiling vs laughing when the agent

Conditions	Agent riddle liking	Stiff /Stiff	Warm /Warm	Boring /Bored	Enjoyable /Amused	Cold
SL-SS	SL**	n.s.	n.s.	n.s.	n.s.	n.s.
LL-LS	LL***	n.s.	LL***	LS***	LL***	LS***
SS-LL	LL**	n.s.	n.s.	n.s.	LL*	n.s.
SS-LS	SS**	n.s.	n.s.	LS*	n.s.	LS*
SL-LS	SL***	LS*	SL***	LS***	SL**	LS**
SL-LL	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Table 4.1: Results of ANOVA tests when comparing the pairs of conditions described in column 1 (SL vs. SS, LL vs. LS, etc). Results indicate no significant difference (n.s.), or significant difference at various levels (indicated by the number of stars). See Section 4.1.3.4 for more explanation.

speaks: only the agent's liking of its riddle is perceived significantly higher when the agent laughs.

Smiling condition The results of the fourth line of Table 4.1 (SS-LS) show that an agent which speaks and smiles is better perceived than an agent who listens and smiles. Again the negative perception of listener-agent who "just smiles" to the user's jokes seems to explain the result.

Laughing condition The laughing conditions (last line of Table 4.1 (SL-LL)), when the agent speaks and when the agent listens, show no significant differences.

These results give a hierarchy of conditions in the context of telling a riddle:

To listen and "just smile" is the most negatively perceived attitude: the agent seems to like significantly less the joke but among others to be significantly more *bored* and *cold* than in any other conditions, and to be significantly less *warm* and *amused* than in laughing conditions.

To "just smile" is perceived less negatively when the agent speaks: compared to the laughing speaking agent, only the liking of the riddle is lower.

Laughing does not appear to change the perception when the agent speaks or listens whereas smiling does: "just" smiling when listening is perceived negatively.

The laugh synthesized animation clearly enriched the agent with fine interaction capacities, and our study points out that laugh contrasts with smile through two facets: (1) when laugh is triggered in reaction to the partner's talk, it appears as a reward and a very interactive behavior; (2) when laugh is triggered by the speaker itself, it appears as more self-centred behavior, an epistemic stance.

4.1.5 Comments

We have presented a laughter motion synthesis model that takes as input laughter phonemes and their duration as well as speech features to compute a synchronized multimodal animation. We have evaluated our model to check how laughing agent is perceived when telling / listening to a joke.

Contrasting with one of our expectations, we did not found any effect of agent's laugh on human user's liking of the joke. This may be explained by the fact that human had to read the joke before telling it to the agent: thus they had already evaluated the joke while reading it for themselves before telling it to the agent and seeing its reaction.

However, our data shows that laugh induces a significant positive effect in the context of telling a riddle, when the agent is listening and reacting to the user. The effect is less clear when the agent is speaking, certainly due to this very context of telling a riddle: laughing at its own joke is more an epistemic stance (concerning what the speaker thinks of what it says) than a social stance (i.e. a social attitude directed toward the partner).

4.2 Upper Body Animation Based on HMM

We now present another approach that focuses on upper body laughing motion where we use a motion capture database we gathered for this particular purpose. We first present the dataset in section 4.2.1 then we detail our animation system in section 4.2.2, which is based on Fully Parameterized HMM (FPHMM) and we report in section 4.2.3 experimental results gained through objective as well as subjective evaluation.

4.2.1 Laughter Dataset of Body Motion

Three human subjects participated in the collection of laughter data. During the recording session, the subjects watched funny movies for about 40 minutes. Since laughter occurs mainly during social interactions [80], [89], we propose an interactive setup where two subjects watch funny videos together. Only the movement of one person was gathered. Three-dimensional torso and head movements and audio signal are recorded by a motion capture system xSens[3] at 125 frames per second (fps) and a microphone at 44100 Hz, which were synchronized using the approach described in [51]. During data processing, all laughter episodes were manually extracted. In total, we obtain 259 laughter episodes; each one lasts from 1 to 37 seconds. Then the phonetic transcription of each laughter is extracted according to Urbain et al [115], in which 14 laughter laughter phonemes are defined in reference to speech phoneme. Phonetic transcription contains laughter phoneme (text signal) and its duration. An intensity value is also provided for each laughter phoneme.

[3] http://www.xsens.com

Finally, PRAAT [115] is used to extract acoustic signals at 125 fps including pitch and energy.

4.2.2 Head and Torso Motion Synthesis

The system we propose produces head and torso motions featured by 3D rotation angles (hence a 6 dimensional signal) from a number of input signals which are: the laughter phoneme sequence together with their duration and their intensity (low or high), and audio features (pitch and energy).

Animation Generator We chose to build one model for generating animation for every $(phoneme, intensity)$ pair, we name a model for each pair an Animation Generator (AG). Since *silence* laughter phoneme has only one distinguishing between low or high for the remaining 13 laughter phonemes have to we build 23 AGs. Each of these 23 AGs is learned independently from the training corpus with corresponding (input, output) pairs where the input stands for all the above input features and the output stands for a sequence of animation motion for the 6 data streams we want to learn to synthesize (the 6 dimensions of the animation signal). Our modeling framework is based on three ideas that we detail now.

- Modeling one dimensional shaking-like movement with what we call *Loop HMM*.
- Introducing speech influence on motion through transition probability parameterization, yielding what we call Transition Parameterized Loop HMM (TPLHMM).
- Taking into account the dependencies between the 6 dimensions of the animation movements with coupled HMMs, yielding Coupled TPLHMM (CTPLHMM).

Modeling Shaking Motion with a Loop HMM. We propose a specific HMM that we call a Loop HMM (LHMM) to model (and synthesize) a one-dimensional shaking-like (and/or trembling) signal (Figure 4.9). It has an approximate left-to-right chain structure where transitions are allowed from one state to itself, to the previous and to the next state. Yet it is intended that the transition probability from one state to the previous state be very small so that a likely state sequence will depict the entire chain form the first state to the last state with some *hesitation* corresponding to few back transitions.

The HMM is designed so that an observation sequence produced along such a state sequence will correspond to one shake pattern (with some trembling effect coming from back transitions). There is one Gaussian distribution associated to each state of the chain, which are set by hand rather than learned, as follows. We first divide the range of the signal value in N intervals and define N Gaussian Probability Density Function (PDF), one for each interval. The mean of the Gaussian distribution for a given interval is the mean of this interval and its variance is defined according to the width of this interval. Then we assign one of the PDF to every state of the left-right HMM so that going from the first state to the last state corresponds to a trajectory of a shaking movement. For instance in Figure 4.9, the first state has PDF p_2 which outputs intermediate values in the observation space, the second state has PDF p_3 which outputs higher values, it is followed by a state with PDF p_2, then by a state with PDF p_1 which outputs lower values. If a signal is produced by this HMM along a state sequence that goes from the first (left) to the last (right) state it will correspond to a shaking-like motion.

Finally, there is a loop from the last state to the first state to enable the repetition of such a shaking and trembling pattern. Figure 4.9 (top) shows one example of a synthesized motion stream by a LHMM. As can be seen, the animation inferred by a LHMM shows the repetition of a pattern.

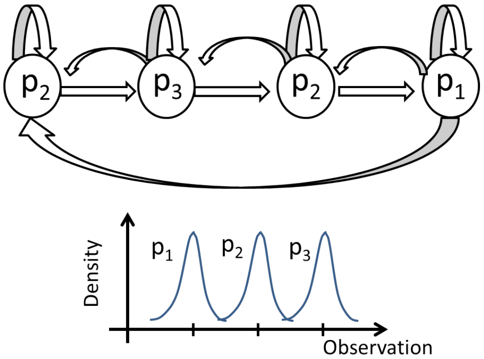


Figure 4.9: A Loop HMM whose manual design allows to model shaking and trembling one dimensional moves.

Taking into account the dependency with speech through parameterized transitions Some evidence about the motion pattern may be gained from taking into account the dependencies between audio signal and motion during laughter [40]. Audio signal (we use pitch and energy) may then be used to shape the synthesized animation stream. In addition to introducing some variability in the inferred animation such a strategy makes animation look more realistic because of an increased consistency with the audio signal. One could have thought at exploiting contextual models where emission probability function also depend on the audio signal, as in previoous chapter. Yet in Loop HMMs the strong constraint on emission probability functions makes observation sequence almost determined by the state sequence. In some way parameterizing the transition probabilities in a Loop HMM plays a similar role as parameterizing the Gaussian distribution in a more conventional contextual HMM.

To exploit such a correlation between speech and movements we developed an extension of our LHMM, whose state transition probabilities depend on acoustic features. We call these models Transition Parameterized Loop HMM (TPLHMM). They may be used to model and synthesize one dimensional shaking movements that are linked in some way with speech. We consider that transition probabilities from state i to state j at time t are defined according to:

$$a_{i,j}(t) = \frac{e^{W_{i,j}\theta_t}}{\sum_{j'} e^{W_{i,j'}\theta_t}} \tag{4.2.1}$$

where θ_t and W are c-dimensional vectors. The parameters of a TPLHMM (the W's) are learned via likelihood maximization with a Generalized EM algorithm. To ease learning it is initialized with a trained LHMM.

Isolated and joint modeling of the 6 dimensional animation signal We investigated few possibilities for synthesizing the 6 dimensional animation signal.

- A first possibility to model and synthesize the 6 dimensional animation signal is to assume the 6 signals are independent from each other and to learn independently one LHMM per dimension.
- Following our discussion above one may follow a very similar strategy but use one TPLHMM per dimension instead of one LHMM.
- Finally one could consider jointly modeling head and torso motions. For example, Ruch and Ekman [105] reported that the backward tilt of the head

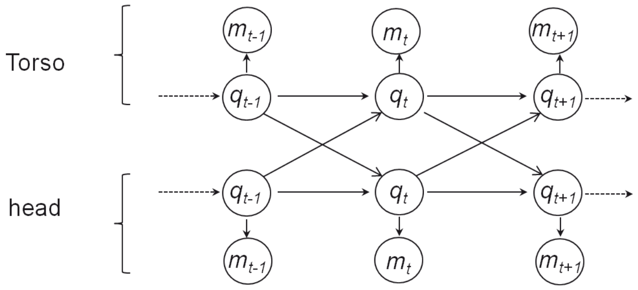


Figure 4.10: Coupled Transition Parameterized Loop HMM (CTPLHMM).

facilitates the forced exhalations, while exhalation directly influences torso motion as being done in DiLorenzo et al. [40]. Therefore, the relationship between head and torso motions should be modeled jointly for improving naturalness of synthesized animations. In our work, we used Coupled HMMs [15] which have been designed to model multiple smoothly interdependent streams of observations. In a Coupled HMMs with K streams of observations, there is one HMM per stream and transition probabilities account for transiting from K-tuple of states (one state in each stream's HMM) to another K-tuple of states. In our experiments we use 6 trained TPLHMMs to initialize one Coupled HMMs, we then get a six branch Coupled TPLHMM, whose transitions are parameterized with speech features. After initialization it is retrained through maximum likelihood estimation.

Figure 4.10 shows a CTPLHMM which is composed of two TPLHMMs. One is used to model the torso motion; the other one the head motion. The relation between the torso and head motion is learnt by Coupled models (Coupled TPLHMMs). In synthesis, the inferred head position at time t, depends not only on the head position at time $t-1$, but also on the torso position at time $t-1$.

Animation Synthesis Given a sequence of phonemes of length T, together with their intensity and duration, we successively synthesize T segments of appropriate duration with the corresponding model of the (laughter phoneme, intensity) pair, which is either a set of 6 LHMMs, or a set of 6 TPLHMMs, or a CTPLHMM with 6 streams. In case TPLHMMs or CTPLHMM are used the acoustic features are exploited to alter the transition probabilities. At the end all the T synthesized sequences are concataned.

Whatever the models used, the synthesis is performed in two steps. First we

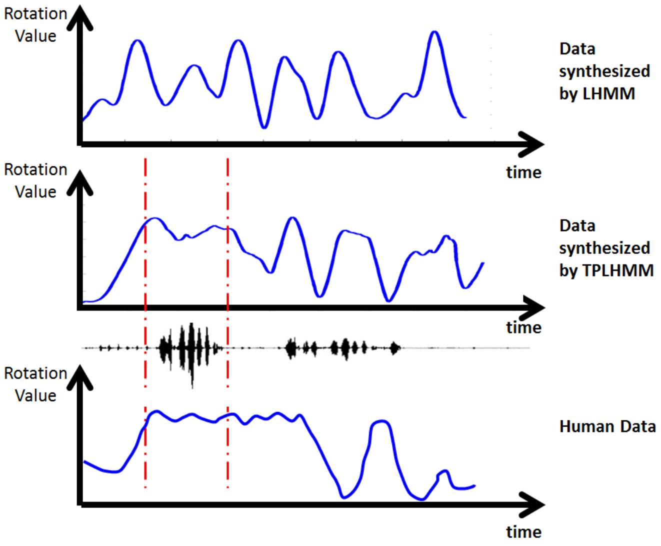


Figure 4.11: Examples of head pitch trajectories.

randomly generate a state sequence according to the transition probability distribution. Second, we computed the synthesized sequence as the most likely observation sequence given the state sequence, it consists in the sequence of the means of the Gaussian distribution of the states in the sequence.

Figure 4.11 shows 3 examples of head pitch trajectories. The top one is synthesized by LHMM; the middle one is by TPLHMM; the bottom one reproduces human data from our dataset. The top and the middle are inferred from audio signals. The three examples are synchronized with the same audio signals. As we can see, the LHMM generates the animation by repeating quite similar head movement patterns. As LHMM does not make use of the audio signals, its output does capture variations in line with the audio signals. On the other hand, TPLHMM is able to generate head patterns showing high similarity with human data.

4.2.3 Experiments

Animation synthesis model is built from human data of 2 subjects. The data contains 205 laughter sequences in total. Human data from a third subject is used for validation through subjective and objective evaluation studies. It contains 54 laughter sequences. Objective and subjective evaluations are conducted to validate the proposed animation synthesis model.

4.2.3.1 Objective Evaluation

As described in Section 4.2.2, LHMM and TPLHMM treat separately each dimension motion of head and torso, while the coupled model is intended to capture the relationships between them. We first investigate whether such a coupling is relevant; then we compare the animations synthesized by LHMM, TPLHMM and CTPLHMM with respect to few quantitative criteria.

Investigating Relation between Head and Body Motions To investigate the relevance of joint modeling of the 6 dimensions animation we tested the probabilistic independency between the 6 random variables corresponding to the states that are occupied at the same time in the 6 streams' LHMMs. For each pair of streams we built a contingency table for the two random variables of being in a state in the HMM for stream 1 while being in a state in the HMM for stream 2, then we computed a χ^2 test to evaluate the independency between the two random variables. We found that whatever the two streams are and whatever the model is, i.e whatever the pair (laughter phoneme, intensity) is, the two random variables were found statistically dependent at a p-value lower than 0.001. This means jointly modeling the multiple streams is actually relevant and should lead to improved animation.

Furthermore to quantify the degree of dependency between the multiple streams we computed relative mutual information. The mutual information between two random variables X and Y, $I(X,Y)$, equals the difference between the entropy of X, $H(X)$ and the conditional entropy of X given Y, $H(X|Y)$. If X and Y are independent, Y does not bring any information about X and $I(X,Y)=0$. Alternatively, if Y includes some information about X, the uncertainty on X is reduced when knowing Y so that the conditional entropy $H(X|Y)$ is lower than $H(X)$ and $I(X,Y)>0$. Furthermore one can measure the amount of information Y brings on X by computing a normalized mutual information $\hat{I}(X,Y)=I(X,Y)/H(X)$ where $H(X)$ is the entropy of X. The normalized mutual information belongs to the range $[0,1]$. It equals 0 if X and Y are fully independent, while it equals 1 if X may be deterministically predicted from Y. In all the tests we performed we obtained normalized mutual information between 17% and 22% which shows that some uncertainty exists between the 6 dimensions of the animation but that it is not fully random either.

As a conclusion, the 6 dimensions of the animation are not independent. Hence, independent modeling of the 6 streams would be suboptimal, and these are not de-

Table 4.2: Performance of the models with respect to the synthesis quality (frequency, amplitude and energy errors). Performances are averaged results gained on 54 test sequences (standard deviations are given in brackets).

Model	frequency	amplitude	energy
LHMM	0.21 (0.074)	0.24 (0.100)	0.41 (0.071)
TPLHMM	0.17 (0.063)	0.19 (0.066)	0.34 (0.057)
CTPLHMM	0.17 (0.061)	0.20 (0.059)	0.31 (0.052)

terministically linked, meaning that a pure synchronous modeling of the 6 streams in a single LHMM or a single TPLHMM would not be a good option either. Finally these results justify our choice of modeling the 6 dimensional animation signal within a coupled HMM that enables modeling dependency (even a weak dependency) between the streams.

Similarity between synthesized and real animations We compared our models by computing 3 criteria which allow evaluating the similarity between a synthesized signal and a real signal. Basically we consider the quality of the synthesized signal with respect to three features: the main frequency of the signal, as extracted by the Periodicity Algorithm [112], the amplitude of this main frequency, and the energy of this frequency. These criteria allow investigating if the main features of a shaking-like movement are well modeled by the synthesis system.

For each of the three features we computed a normalized error (e.g. $\left|\frac{f^s-f^h}{f^h}\right|$ for the frequency feature, where f^s and f^h stand for the frequency of the synthesized and of the human animation signals averaged over all laughter phonemes realizations. The lower such a measure is the closer the synthesized signal is from the original one. The frequency, amplitude and energy errors obtained for our various models are reported in Table 4.2. According to these measures, TPLHMM and CTPLHMM do perform much better than LHMM while the difference of performance between TPLHMM and CTPLHMM is less clear.

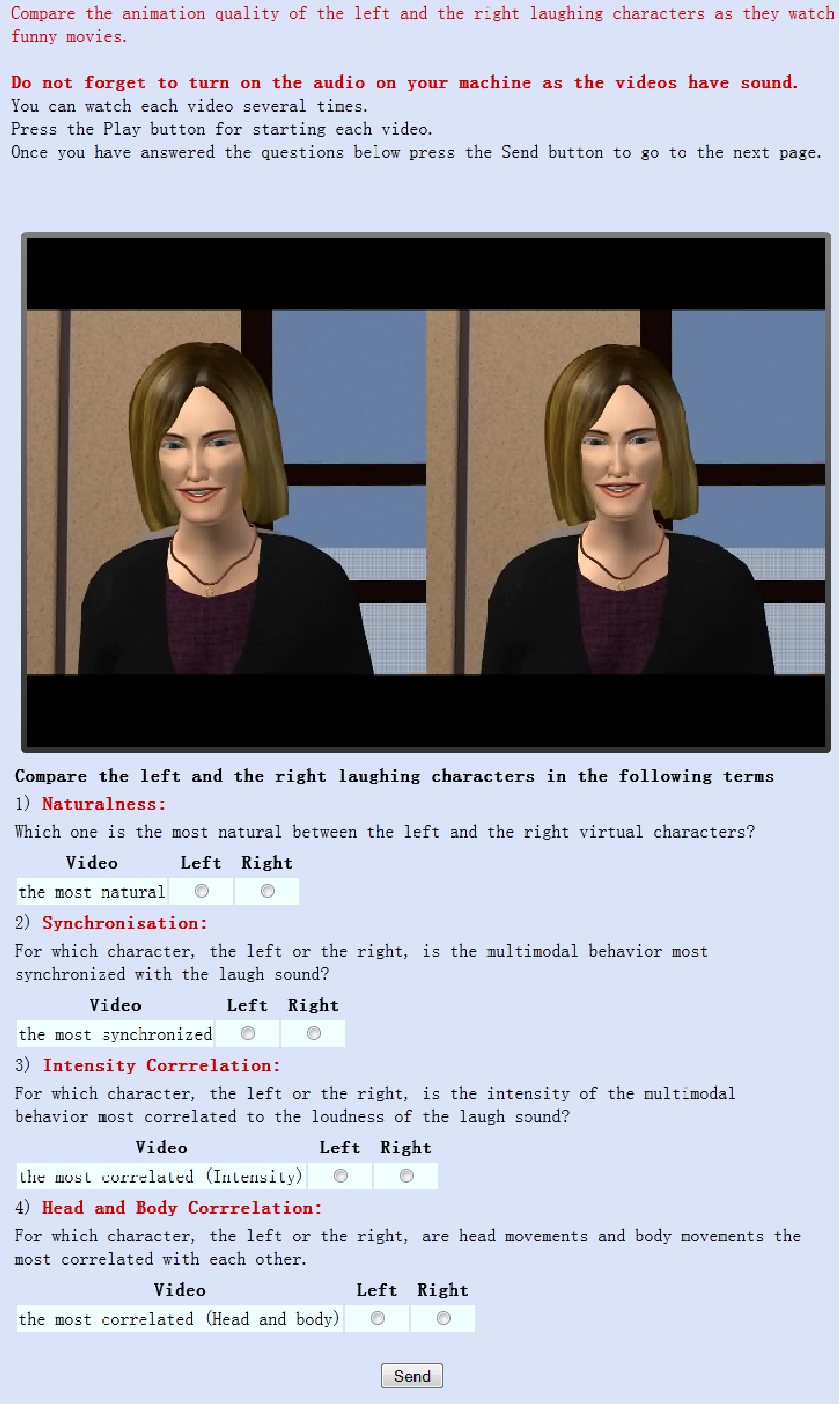
4.2.3.2 Subjective Evaluation

Two subjective evaluations were conducted through an online web application. First, we compare the animations synthesized by TPLHMM and CTPLHMM; then the best one is compared to human data. The participants were invited to watch

5 videos of laughing virtual character and to answer few questions for each video. They could control when to start the videos and could watch them as many times as they wish. Our aim is to evaluate the behaviors animation and not the appearance of the virtual agent. We used the same virtual agent to display motion data for both subjective evaluations. Motion data displayed with the virtual character consists of head and torso movements (motion capture or generated data) and facial expression. Facial expression of laughter was computed using the approaches described in Section 4.1.2.1 and Section 4.1.2.2. The 5 videos used in both subjective studies last respectively 9s, 10s, 18s, 26s and 27s.

TPLHMM and CTPLHMM Comparison: To compare TPLHMM and CTPLHMM, both trained models were applied to the 5 test samples. For each test sample, a pair of videos was recorded in which the virtual agent's head and torso motions were driven respectively by these models. Each pair of video clips was displayed on the same web page and randomly arranged on the right or on the left. After watching each pair of video clips, participants were invited to select the best animation along four dimensions: naturalness of the animation, synchronization of head and torso movements with laughter sound, correlation of laughter intensity and torso movements, inter-correlation of head and torso movement. An example of web page is shown in Figure 4.12. This evaluation study involved 120 participants, 67 males and 53 females with age ranging from 18 to 65 years old (Mean=33.5 years, SD=9.6 years). We computed 95% confidence intervals that show that CTPLHMM is significantly better than TPLHMMs with respect to the 4 questions: we obtained a confidence interval equal to [66% 77%] for CTPLHMM being better than TPLHMMS with respect to Naturalness, [60% 72%] for Synchronisation, [58% 70%] for Intensity correlation and [63% 74%] for Head and Torso inter-correlation.

Synthesized and Human Data Comparison: With respect to the results above, CTPLHMM is perceived as the best animation synthesis framework; so we use the animations obtained with CTPLHMM in the comparison test with human data. This subjective evaluation was conducted to investigate how similar is the perception of the virtual agent displaying head and torso motions synthesized by CTPLHMM to the perception of the virtual agent displaying head and torso motions synthesized by CTPLHMM is similar to the perception of the virtual agent animated directly by human data (ie from the motion capture data). As the previous



Compare the animation quality of the left and the right laughing characters as they watch funny movies.

Do not forget to turn on the audio on your machine as the videos have sound.
You can watch each video several times.
Press the Play button for starting each video.
Once you have answered the questions below press the Send button to go to the next page.

Compare the left and the right laughing characters in the following terms

1) **Naturalness:**

Which one is the most natural between the left and the right virtual characters?

Video	Left	Right
the most natural	○	○

2) **Synchronisation:**

For which character, the left or the right, is the multimodal behavior most synchronized with the laugh sound?

Video	Left	Right
the most synchronized	○	○

3) **Intensity Corrrelation:**

For which character, the left or the right, is the intensity of the multimodal behavior most correlated to the loudness of the laugh sound?

Video	Left	Right
the most correlated (Intensity)	○	○

4) **Head and Body Corrrelation:**

For which character, the left or the right, are head movements and body movements the most correlated with each other.

Video	Left	Right
the most correlated (Head and body)	○	○

Figure 4.12: An example of web page to compare the performances of TPLHMM and CTPLHMM. After watching each pair of video clips, participants were invited to select the best animation along four dimensions.

study, a comparison test was conducted. An example of web page is shown in Figure 4.13.

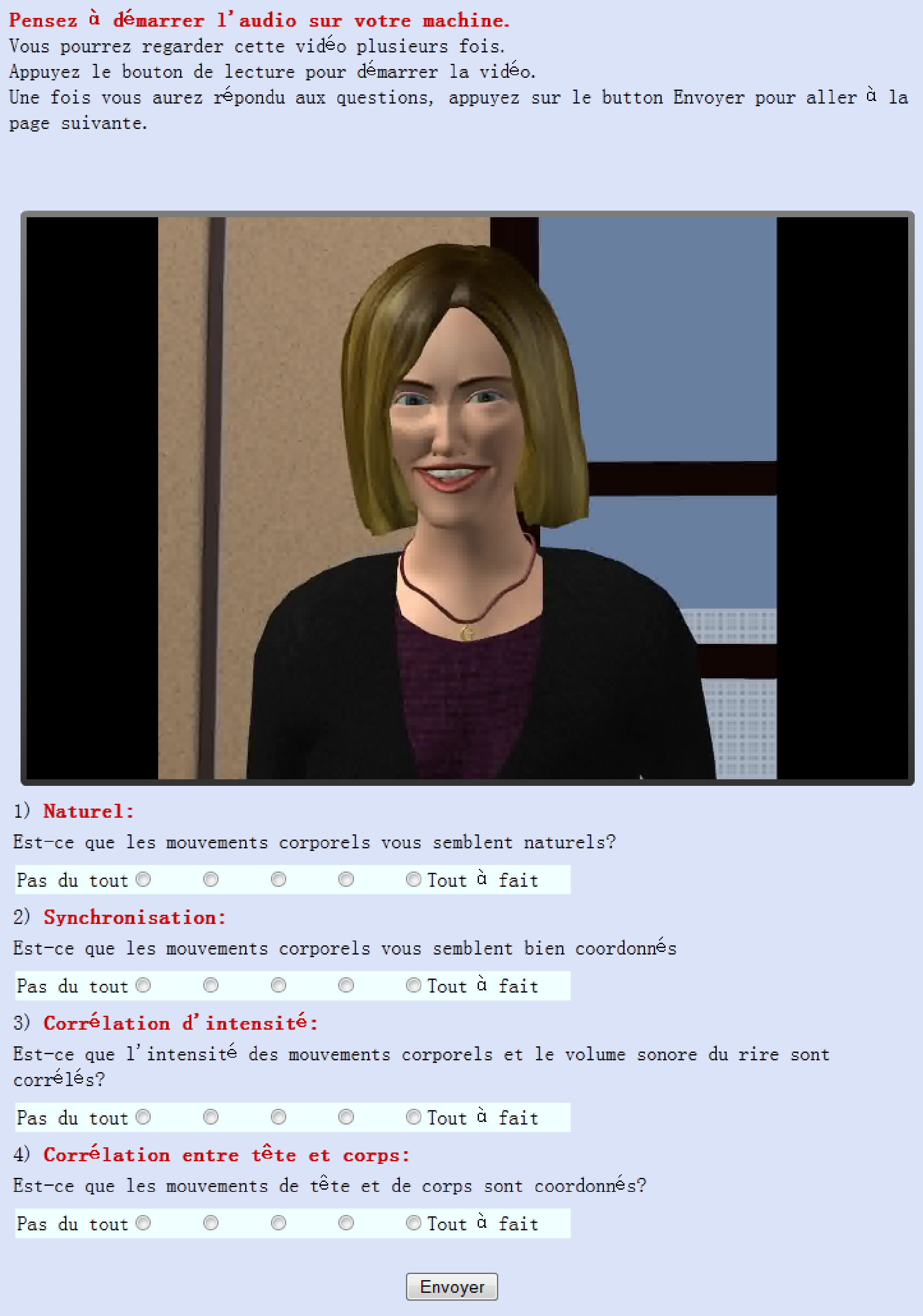
In total, there were 80 participants consisting of 46 males and 34 females with age ranging from 12 to 78 (M=40.65 years, SD=17.91 years). To verify the hypothesis, 2 versions (conditions) of the virtual agent animations were created for each selected test sample. They are human and synthesized motions. There are a total of 10 video clips (5 input samples × 2 conditions). Each participant watched 5 video clips, each of which is randomly selected from the 2 conditions. Each video clip has been evaluated 40 times (i.e., by 40 participants). After watching each video clip, each participant was invited to answer the same four questions as in the first evaluation study, but this time the participant answered using a 5 point Likert scale.

The results are shown in Figure 4.14. As can be seen, synthesized motion obtains score less than human motions along the four dimensions: naturalness, synchronization, correlation of laughter intensity and torso movements, inter-correlation of head and torso movement. T-test shows that there are significant differences in all terms between human and synthesized data.

4.2.3.3 Discussion

The objective evaluation for comparing LHMM, TPLHMM and CTPLHMM shows that TPLHMM and CTPLHMM perform better than LHMM. It highlights that acoustic features and motions are linked. Thus acoustic features can be used to capture and synthesize improved motion trajectories. In LHMM and TPLHMM models, head and torso motions are modeled separately. In other words, they are considered as being independent. However, through the objective evaluation investigating the relation between head and torso, we found that head and torso motions are dependent on each others. We proposed coupled models for learning such relationships between head and torso movements.

The subjective evaluation compared TPLHMM and CTPLHMM. CTPLHMM obtains higher score than TPLHMM. In the subjective evaluation on comparing synthesized and human motions, human data is perceived significantly better than synthesized data in terms of naturalness, synchronization, intensity and correlation of head and torso movements. However the difference in perception is not so severe (less than 1 on a 5 likert scale). This suggests that the proposed CTPLHMM is somehow capable of synthesizing human-like head and body motions.



Pensez à démarrer l'audio sur votre machine.
Vous pourrez regarder cette vidéo plusieurs fois.
Appuyez le bouton de lecture pour démarrer la vidéo.
Une fois vous aurez répondu aux questions, appuyez sur le button Envoyer pour aller à la page suivante.

1) **Naturel:**
Est-ce que les mouvements corporels vous semblent naturels?
Pas du tout ○ ○ ○ ○ ○ Tout à fait

2) **Synchronisation:**
Est-ce que les mouvements corporels vous semblent bien coordonnés
Pas du tout ○ ○ ○ ○ ○ Tout à fait

3) **Corrélation d'intensité:**
Est-ce que l'intensité des mouvements corporels et le volume sonore du rire sont corrélés?
Pas du tout ○ ○ ○ ○ ○ Tout à fait

4) **Corrélation entre tête et corps:**
Est-ce que les mouvements de tête et de corps sont coordonnés?
Pas du tout ○ ○ ○ ○ ○ Tout à fait

Figure 4.13: An example of web page to evaluate the performances of CTPLHMM or human data. After watching each video clip, each participant was invited to answer questions along four dimensions.

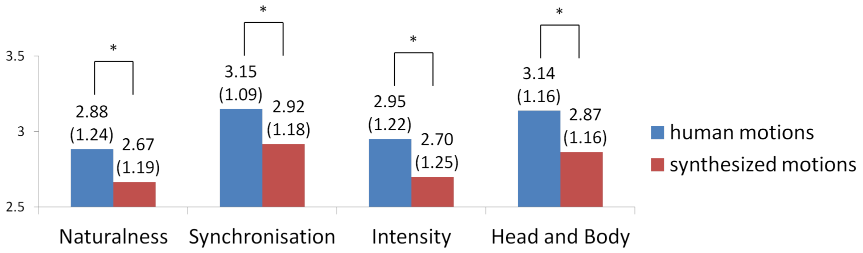


Figure 4.14: Averaged values of virtual agent animated by animations from human and synthesized. Significant differences are identified by $\star$ ($P < .05$). The averaged values are shown with an histogram and the standard deviation is specified in parenthesis.

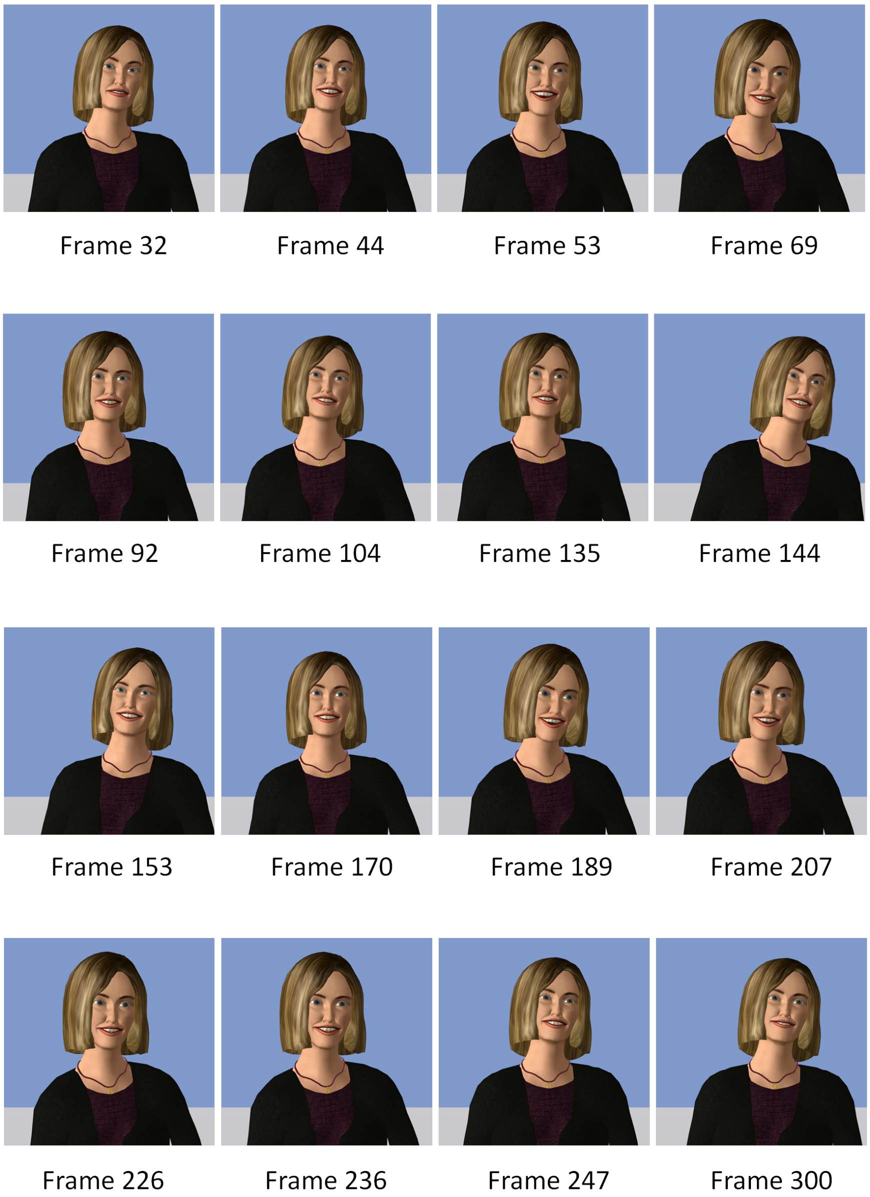
Figure 4.15 shows examples of laughing images, which are extracted from an ECA video. In this video, the movements of head and torso are computed by CT-PLHMM and the facial expression are inferred by the approaches presented in Section 4.1.2.1 and Section 4.1.2.2.

4.2.4 Comments

In this section we have presented an approach to model laughter head and torso movements, which are very rhythmic and show saccadic patterns. To capture laughter motion characteristics, we have developed a statistical approach to reproduce frequency movements, such as shaking and trembling. Our statistical model takes as input laughter phoneme sequences and acoustic features of laughter sound. Then it outputs the head and torso animations of the virtual agent. In the training model, not only the relation between input and output features is modeled, but also the relation between head and torso movements is captured. Experiments show that our model is able to capture the dynamism of laughter movement, but it does not actually reach the quality of animation directly copied from human data.

4.3 Conclusion on Laughter Animation Synthesis

In this chapter, we have presented approaches to synthesize laughing animation for embodied conversational agents (ECAs). The models are capable of generating the laughing movements of lip, jaw, eyebrow, head, shoulder and torso. These movements are driven by the transcription and prosody features of laughter sound. The



Frame 32 Frame 44 Frame 53 Frame 69

Frame 92 Frame 104 Frame 135 Frame 144

Frame 153 Frame 170 Frame 189 Frame 207

Frame 226 Frame 236 Frame 247 Frame 300

Figure 4.15: Examples of laughing images. In this figure, laughing images are extracted from an ECA video. In this video, the movements of head and torso are computed by CTPLHMM and the facial expression are inferred by the approaches presented in Section 4.1.2.1 and Section 4.1.2.2

works are composed of two parts, which have been respectively described in Section 4.1 and Section 4.2. The work described in Section 4.1 is our first attempt, where the synthesized movements involve lip, jaw, eyebrow, head, shoulder and torso. The work described in Section 4.2 concerns a new approach of generating head and torso movements.

The work described in Section 4.1 used an existing database, called AudioVisual LaughterCycle (AVLC) [114]. It contains signals of human laughter audio, facial expression and head rotation motion. In this work, lip and jaw movements are computed by linear regression method, which takes as input phoneme sequences and prosody features of laughter sound; eyebrow and head movements are synthesized by concatenating human segmented motions (motion concatenation method), which takes as input laughter phoneme sequences and intensity; shoulder and torso are inferred by PD controller, which takes as input the synthesized head movements. Linear regression method and motion concatenation method are statistical framework based on human data. PD controller is a determinate model, where the used parameters are defined by hand. Notice that shoulder and torso motions are not recorded in the used database, the AudioVisual LaughterCycle database. We evaluated such models to check how laughing agent is perceived when telling / listening to a riddle to a user. The experiments show that laugh induces a significant positive effect in the context of telling a riddle, when the agent is listening and reacting to the user. The effect is less clear when the agent is telling the riddle.

To furthermore investigate laughing upper body animation, we recorded a new database, which contains signals of human laughter audio, body motions (including head rotation and torso motions). Using this new database, we presented a new approach to model laughter head and torso movements, which are very rhythmic and show saccadic patterns. To capture laughter motion characteristics, we developed a statistical approach to reproduce frequency movements, such as shaking and trembling. Our statistical model takes as input laughter phoneme sequences and acoustic features of laughter sound. Then it outputs the head and torso animations of the ECA. In the training model, not only the relation between input and output features is modeled, but also the relation between head and torso movements is captured.

We used specific HMMs, called Loop HMMs (LHMMs) in our work, to learn shaking / trembling motion patterns. Then we used Transition Parameterized LH-

MMs (TPLHMMs) to model the relation between prosody features and transition probabilities of states. With TPLHMMs, we can capture the relation between the dynamical characteristics of motions and prosody features. Finally, we used Coupled TPLHMMs (CTPLHMMs) to capture the relation between the movements of head and torso. Experiments show that our models are able to capture the dynamism of laughter movement, but do not overcome animation directly copied from human data.

Chapter 5

Conclusions

Human-like animations are crucial for engaging users in various applications of ECA. In this thesis, we have described our approaches of generating appropriate animations respectively for speaking ECA and laughing ECA. For speaking ECA, the animation generators take as input speech prosody features (pitch and energy) and output the animations of head and eyebrow. For laughing ECA, the animation generators take as input laughter sound transcription and prosody features (pitch and energy) and output the various animations including lip shapes, facial expressions and upper body motions. The proposed approaches are statistical frameworks. The frameworks are first trained on human datasets to capture the relationship between signals of input and output. Then they are used to render the captured relationship to the synthesized signals as animation generators.

5.1 Conclusion on Speech Animation Synthesis

In our first work (speech animation synthesis), our aim is to build models for speech animation synthesis. First, we focused on eyebrow animation synthesis; then we generalised the eyebrow animation synthesis to both eyebrow and head animations synthesis.

In the first step of eyebrow animation synthesis, the existing contextual models based on HMM, called contextual HMMs [121, 103] in this thesis, are used as animation generators. Furthermore, we extended contextual HMM as fully parameterized HMMs (FPHMMs); then FPHMMs are used as animation generators. We conducted some evaluations where we compared the results produced by different types of HMM. These objective evaluations show that contextual HMMs and FPHMMs outperform the standard HMM [56, 19] and that FPHMMs are better than contextual HMMs.

In the second step of both eyebrow and head animations synthesis, FPHMMs are generalised as eyebrow and head animation generators. In this step, we investigate whether embedding in our statistical framework the relationship between the movements of eyebrow and of head augments the quality of the synthesized signals. To answer this question, we compare the results of both models. We build an *independent model* and a *joint model*. In the independent model, eyebrow and head animation generators are independently built and used. It means that the movements of the eyebrows and of the head are assumed to be independent. In the joint model, eyebrow and head animation generators are joined and acted as one common generator. The relationship between the movements of the eyebrows and of the head is recorded and captured in the step of learning. In the step of synthesis, the movements of the eyebrows and of the head are related to each other. The synthesized motions are evaluated using objective and subjective methods. The results of these evaluations show that the relationship learnt by our model between the movements of the eyebrows and of the head can be used to improve the quality of synthesized animations.

In the work of speech animation synthesis, we have proposed a data-driven approach to generate head and eyebrow motions for a virtual agent from speech prosody. The FPHMM is used to capture the direct mapping between audio and visual information. The trained PFHMM allows defining the visual animation as a function of the speech signal. The objective evaluation shows that considering simultaneously eyebrow and head motions increases the precision of the resulting animation. It also confirms that eyebrow and head motions are not independent from each other but, rather, are connected; the multimodal signals reinforce the communicative meaning. The subjective evaluation shows that our proposed model enhances the perception of the virtual agent animation at the level of emotional expressiveness.

5.2 Conclusion on Laughter Animation Synthesis

In our second work (laughter animation synthesis), we build laughter animations generators of the full body, which includes jaw, lip, cheek, eyelid, eyebrow, head, shoulder and torso. To our best acknowledge, our laughter animation generators are the first attempt of synthesizing the full body laughter animation. Linear regression models are used as lip and jaw animations generators, where the lip and jaw movements depend on laughter phoneme and prosody features. Concatenation method is proposed to build animations generators of cheek, eyelid, eyebrow and head, where phonemes intensity and duration are used to select the segmented human motions. Proportional-Derivative Control is applied to synthesize shoulder and torso animations, which are synchronised to each other and driven by the synthesized head movement. We evaluated our model to check how laughing agent is perceived when telling / listening to a riddle. We performed an evaluation study where we compared two conditions, a smiling agent and a laughter agent. The results show that participants preferred interacting with a laughing agent than a smiling agent.

Furthermore, Coupled FPHMMs are used as head and torso animations generators, where the synthesized animations are capable to render shaking or trembling motions. In such a method, the synthesized animations are synchronized to prosody features. They are also highly correlated with each other. A subjective evaluation is conducted to validate the proposed laughter animation generators. We compare results coming from coupled FPHMMs and non coupled ones. We also did a comparison with human motions. The results show that participants found the synthesized animations better for the coupled FPHMMs but they still do not overcome the naturalness of the animations directly copied by human data.

Chapter 6

Perspectives

6.1 Speech Animation

Most existing models of virtual agents' behaviors can be clustered into two main groups: the rule-based models (ex. the existing Greta system) and the data-driven models. Our works described in this thesis belong to the latter. Both of these models types have pros and cons. While data-driven models are more prone to produce natural looking animation, rule-based models capture more precisely the semantic-emotional behaviors to communicate. These latter ones are often event-driven; that is they compute a behavior only when a given communicative function is specified. Data-driven models produce animation continuously that captures the communicative colour of the message to convey but they have difficulty to compute behaviors which have specific meaning. As a result, virtual agents driven by cognitive-like system are able to convey more precise displays while those driven by statistical models look more natural and lively [68]. Hence, speech-based animation could be improved by combining the output of both the Greta system and the statistical system. By combining both approaches, that is, embedding in one model the semantically driven approach and the statistical one, would allow us to have an agent able to convey semantic-emotional messages naturally. Hybrid schemes to combine the output of both the Greta system and the statistical system will have to be designed.

6.2 Laughter Animation

In this thesis, we have described our approaches of modeling periodical motions for head and torso, where motion trajectory position is used as motion features. A specific HMM, Loop HMM, is used to model periodical motion signals. Such approaches could result in unnatural animations. In the future, we could attempt to use frequency signals of motion as motion features. This idea is based on that a laughter periodical motion can be decomposed into several periodical signals. Such signals are defined by two features: periodicity and energy. Once we determine periodicity and energy of these compositions, the smoothing animation trajectory can be directly synthesized.

Additionally, the works described in this thesis have focused on the animation synthesis of rigid torso. It is observed that torso is deformable during laughing and synchronized by breathing. We have not attempted to model laughing deformable torso. In the future, we aim to investigate how to model deformable torso; then we can investigate the relationship between the rigid movement of torso and the deformation of torso to augment the quality of the synthesized animation of torso. Also, we are interested in investigating the relationship between the movement of head/arms and the deformation of torso.

Laughter can also act as social indicator of in-group belonging [4]; it can work as speech regulator during conversation [100]; it can also be used to elicit laughter in interlocutors as it is very contagious [101]. Such factors can be taken into account for synthesizing laughter animation.

Chapter 7

Publications

Best Paper Reward —- 2013 International Conference on Intelligent Virtual Agent

French Conference

1. Y. Ding, M. Radenen, T. Artières, C. Pelachaud.
   Eyebrow Motion Synthesis Driven by Speech
   WACAI 2012 *Workshop Affect, Compagnon Artificiel, Interaction.*

2. M. Ochs, Y. Ding, N. Fourati, M. Chollet, B. Ravenet, F. Pecune, N. Glas, K. Prépin, C. Clavel et C. Pelachaud.
   Vers des Agents Conversationnels Animés Socio-Affectifs
   IHM 2013 *Interaction Humain-Machine.*

3. Y. Ding, T. Artières, C. Pelachaud.
   Laughing Body
   WACAI 2014 *Workshop Affect, Compagnon Artificiel, Interaction.*

International Conference

1. Y. Ding, M. Radenen, T. Artières, C. Pelachaud.
   Speech-Driven Eyebrow Motion Synthesis With Contextual Markovian Models
   ICASSP 2013 *International Conference on Acoustics, Speech and Signal Processing, pp. 3756-3760.*

2. Y. Ding, C. Pelachaud, T. Artières.
   Modeling Multimodal Behaviors from Speech Prosody
   IVA 2013 *International Conference on Intelligent Virtual Agent, pp. 217-228.*
3. Y. Ding, K. Prepin, J. HUANG, C. Pelachaud, T. Artières.
   Laughter animation synthesis
   AAMAS 2014 *International Conference on Autonomous Agents and Multiagent Systems, pp. 773-780.*
4. Y. Ding, J. Huang, N. Fourati, T. Artières, C. Pelachaud.
   Upper Body Animation Synthesis for a Laughing Character
   IVA 2014 *International Conference on Intelligent Virtual Agent, August 28th (to be published)*

Bibliography

[1] URL http://www.acapela-group.com/?lang=fr.

[2] URL https://www.cereproc.com.

[3] *3D Game Engine Design: A Practical Approach to Real-time Computer Graphics*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000. ISBN 1-55860-593-2.

[4] V Adelsward. Laughter and dialogue: The social significance of laughter in institutional discourse. *Nordic Journal of Linguistics*, 102(12):107–136, 1989.

[5] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572 – 587, 2011.

[6] G. Bailly. Audiovisual speech synthesis. *International Journal of Speech Technology*, 6:6–331, 2001.

[7] R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *J Pers Soc Psychol*, 70(3):614–636, 1996.

[8] J. Beskow. Rule-based visual speech synthesis. In *ESCA - EUROSPEECH '95. 4th European Conference on Speech Communication and Technology*, Madrid, September 1995.

[9] E. Bevacqua, K. Prepin, R. Niewiadomski, E. de Sevin, and C. Pelachaud. GRETA: Towards an Interactive Conversational Virtual Companion. In *Artificial Companions in Society: perspectives on the Present and Future*, pages 1–17. 2010.

[10] P. Boersma and D. Weeninck. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

[11] D.L.M. Bolinger. *Intonation and Its Uses: Melody in Grammar and Discourse*. University Press, 1989.

[12] H. Boukricha, I. Wachsmuth, A. Hofstätter, and K. Grammer. Pleasure-Arousal-Dominance driven facial expression simulation. In *3rd International Conference on Affective Computing and Intelligent Interaction, ACII 2009*, pages 119–125, Amsterdam, 2009. IEEE.

[13] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. High resolution passive facial performance capture. *ACM Trans. on Graphics (Proc. SIGGRAPH)*, 29(3), 2010.

[14] M. Brand. Voice puppetry. In *Proceedings of conference on Computer graphics and interactive techniques*, pages 21–28, 1999.

[15] Matthew Brand. Coupled hidden markov models for modeling interacting processes. Technical report, 1997.

[16] Christoph Bregler, Michele Covell, and Malcolm Slaney. Video rewrite: driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '97, pages 353–360, 1997. ISBN 0-89791-896-7.

[17] Thomas S. Buchanan, David G. Lloyd, Kurt Manal, and Thor F. Besier. Neuromusculoskeletal modeling: estimation of muscle forces and joint moments and movements from measurements of neural command. *Journal of Applied Biomechanics*, 20(4):367–395, 2004.

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

[19] Carlos Busso, Zhigang Deng, Ulrich Neumann, and Shrikanth Narayanan. Natural head motion synthesis driven by acoustic prosodic features. *Journal of Visualization and Computer Animation*, 16(3-4):283–290, 2005.

[20] Zoraida Callejas, Brian Ravenet, Magalie Ochs, and Catherine Pelachaud. A computational model of social attitudes for a virtual recruiter. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pages 93–100, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-2738-1.

[21] B. De Carolis, C. Pelachaud, I. Poggi, and M. Steedman. Apml, a markup language for believable behavior generation. In H. Prendinger and

M. Ishizuka, editors, *Lifelike Characters. Tools, Affective Functions and Applications*. Springer, 2004.

[22] J. Cassell, C. Pelachaud, N.I. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Computer Graphics Proceedings, Annual Conference Series*, pages 413–420. ACM SIGGRAPH, 1994.

[23] J. Cassell, H. Vilhjálmsson, and T. Bickmore. BEAT : the Behavior Expression Animation Toolkit. In *Computer Graphics Proceedings, Annual Conference Series*. ACM SIGGRAPH, 2001.

[24] C. Cave, I Guaitella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser. About the relationship between eyebrow movements and fo variations. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, pages 2175–2179, 1996.

[25] Hüseyin Çakmak, Jérôme Urbain, Joëlle Tilmanne, and Thierry Dutoit. Evaluation of hmm-based visual laughter synthesis. In *IEEE International Conference on Audio Speech and Signal Processing*, 2014.

[26] Chung-Cheng Chiu and Stacy Marsella. How to train your avatar: A data driven approach to gesture generation. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, pages 127–140, 2011.

[27] Chung-Cheng Chiu and Stacy Marsella. Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, pages 781–788. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[28] Céline Clavel, Justine Plessier, Jean-Claude Martin, Laurent Ach, and Benoit Morel. Combining facial and postural expressions of emotions in a virtual character. In *Intelligent Virtual Agents*, volume 5773 of *Lecture Notes in Computer Science*, pages 287–300. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-04379-6.

[29] Michael M. Cohen and Dominic W. Massaro. Modeling coarticulation in synthetic visual speech. In *Models and Techniques in Computer Animation*, pages 139–156. Springer-Verlag, 1993.

[30] Darren Cosker and James Edge. Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations. *Proceedings of Computer Animation and Social Agents*, pages 21–24, 2009.

[31] Maurizio Costa, Tsuhan Chen, and Fabio Lavagetto. Visual prosody analysis for realistic motion synthesis of 3d head models. In *In: Proc. of ICAV3Dï£¡01 - International Conference on Augmented, Virtual Environments and 3D Imaging*, pages 343–346, 2001.

[32] Matthieu Courgeon, Céline Clavel, and Jean-Claude Martin. Appraising emotional events during a real-time interactive game. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, AFFINE '09, pages 7:1–7:5, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-692-2.

[33] Roddy Cowie and Ellen Douglas-Cowie. *Speakers and hearers are people: reflections on speech deterioration as a consequence of acquired deafness*, pages 510–527. Whurr, 1995.

[34] Charles Darwin. *The expression of the emotions in man and animals*. London: John Murray, 1872.

[35] Iwan de Kok and Dirk Heylen. When do we smile? analysis and modeling of the nonverbal context of listener smiles in conversation. In *Affective Computing and Intelligent Interaction*, volume 6974 of *Lecture Notes in Computer Science*, pages 477–486. Springer Berlin Heidelberg, 2011.

[36] Celso M. de Melo and Jonathan Gratch. Expression of emotions using wrinkles, blushing, sweating and tears. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents (IVA)*, 2009.

[37] Celso M. de Melo, Patrick G. Kenny, and Jonathan Gratch. Real-time expression of affect through respiration. *JVCA*, 21(3-4):225–234, 2010.

[38] Li Deng. A generalized hidden markov model with state-conditioned trend functions of time for the speech signal. *Signal Processing*, 27(1):65 – 78, 1992. ISSN 0165-1684.

[39] Zhigang Deng, J.P. Lewis, and U. Neumann. Synthesizing speech animation by learning compact speech co-articulation models. In *Computer Graphics International 2005*, pages 19–25, 2005.

[40] Paul C. DiLorenzo, Victor B. Zordan, and Benjamin L. Sanders. Laughing out loud: control for modeling anatomically inspired laughter using audio. *ACM Trans. Graph.*, 27(5):125, 2008.

[41] Yu Ding, Catherine Pelachaud, and Thierry Artières. Modeling multimodal behaviors from speech prosody. In *IVA*, pages 217–228. 2013.

[42] Yu Ding, Mathieu Radenen, Thierry Artières, and Catherine Pelachaud. Speech-driven eyebrow motion synthesis with contextual markovian models. In *ICASSP*, pages 3756–3760, 2013.

[43] P. Ekman. About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human ethology: Claims and limits of a new discipline: contributions to the Colloquium*, pages 169–248. Cambridge University Press, Cambridge, England; New-York, 1979.

[44] P. Ekman. *Emotions Revealed*. New York: Times Books (US). London: Weidenfeld & Nicolson (world), 2003.

[45] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[46] P. Ekman and W. Friesen. Felt, false, miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–251, 1982.

[47] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System (FACS): Manual*. A Human Face, Salt Lake City (USA), 2002.

[48] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 57–64, 2004.

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

[50] María L. Flecha-García. Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in english. *Speech Communication*, 52(6):542 – 554, 2010.

[51] Nesrine Fourati and Catherine Pelachaud. Emilya: Emotional body expression in daily actions database. In *LREC 2014, the 9th International Conference on Language Resources and Evaluation*, pages 3486–3493, Reykjavik, Iceland, 2014.

[52] Hans Peter Graf, Eric Cosatto, Volker Strom, and Fu Jie Huang. Visual prosody: Facial movements accompanying speech. In *In Proceedings of AFGR 2002*, pages 381–386, 2002.

[53] Bjorn Granstrom and David House. Audiovisual representation of prosody in expressive speech communication. pages 393–400, 2004.

[54] Harry J. Griffin, Min S.H. Aung, Bernardino Romera-Paredes, Ciaran McLoughlin, Gary McKeown, William Curran, and Nadia Bianchi-Berthouze. Laughter type recognition from whole body motion. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 349–355, 2013.

[55] D. Heylen, E. Bevacqua, M. Tellier, and C; Pelachaud. Searching for prototypical facial feedback signals. In *IVA07*, pages 147–153, Paris, France, 2007.

[56] Gregor Hofer and Hiroshi Shimodaira. Automatic head motion prediction from speech data. In *Proc. Interspeech 2007*, Antwerp, Belgium, August 2007.

[57] Gregor Hofer, Junichi Yamagishi, and Hiroshi Shimodaira. Speech-driven lip motion generation with a trajectory hmm. In *in Proc. Interspeech 2008*, pages 2314–2317, 2008.

[58] Tania Huber and Willibald Ruch. Laughter as a uniform category? A historic analysis of different types of laughter. In *10th Congress of the Swiss Society of Psychology, September 13-14 2007*, University of Zurich, Switzerland., 2007.

[59] Dacher Keltner. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68(3):441–454, 1995.

[60] Adam Kendon and H. R. Key. Gesticulation and Speech: Two Aspects of the Process of Utterance. In *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. Mouton and Co, 1980.

[61] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. ThÃşrisson, and Hannes VilhjÃąlmsson. Towards a common framework for multimodal generation: The behavior markup language. In *INTERNATIONAL CONFERENCE ON INTELLIGENT VIRTUAL AGENTS*, pages 21–23, 2006.

[62] Sumedha Kshirsagar and Nadia Magnenat-Thalmann. Visyllable based speech animation. *Comput. Graph. Forum*, 22(3):632–640, 2003.

[63] Takaaki Kuratate, Kevin G. Munhall, Philip Rubin, Eric Vatikiotis-Bateson, and Hani Yehia. Audio-visual synthesis of talking faces from speech production correlates. In *EUROSPEECH*, pages 1279–1282, 1999.

[64] Takaaki Kuratate, Kevin G. Munhall, Philip Rubin, Eric Vatikiotis-Bateson, and Hani Yehia. Audio-visual synthesis of talking faces from speech production correlates. In *EUROSPEECH*, 1999.

[65] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

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

[67] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9Ű10):1162 – 1171, 2011.

[68] J. Lee and S. Marsella. Modeling speaker behavior: A comparison of two approaches. In *International Conference on Intelligent Virtual Agents*, pages 161–174, 2012.

[69] Jina Lee and S.C. Marsella. Predicting speaker head nods and the effects of affective information. *Multimedia, IEEE Transactions on*, 12(6):552–562, Oct 2010.

[70] Sergey Levine, Christian Theobalt, and Vladlen Koltun. Real-time prosody-driven synthesis of body language. In *ACM Transactions on Graphics (TOG)*, volume 28, page 172. ACM, 2009.

[71] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. *ACM Transactions on Graphics (TOG)*, 29(4):124, 2010.

[72] Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. Realtime facial animation with on-the-fly correctives. *ACM Transactions on Graphics*, 32(4), July 2013.

[73] Yan Li and Heung yeung Shum. Learning dynamic audio-visual mapping with inputoutput hidden markov models. *IEEE Trans. on Multimedia*, pages 542–549, 2006.

[74] Wentao Liu, Baocai Yin, Xibin Jia, and Dehui Kong. Audio to visual signal mappings with hmm. In *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[75] Erich S. Luschei, Lorraine O. Ramig, Eileen M. Finnegan, Kristen K. Baker, and Marshall E. Smith. Patterns of laryngeal electromyography and the activity of the respiratory system during spontaneous laughter. *Journal of Neurophysiology*, 96(1):442–450, 2006.

[76] Xiaohan Ma, Binh Huy Le, and Zhigang Deng. Perceptual analysis of talking avatar head movements: A quantitative perspective. In *CHI*, pages 2699–2702, 2011.

[77] Maurizio Mancini, Giovanna Varni, Donald Glowinski, and Gualtiero Volpe. Computing and evaluating the body laughter index. In *Proceedings of HBU*, pages 90–98, 2012.

[78] S. Mariooryad and C. Busso. Generating human-like behaviors using joint, speech-driven models for conversational agents. *IEEE Trans. on Audio, Speech & Language Processing*, 20(8):2329–2340, 2012.

[79] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew W. Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Symposium on Computer Animation*, pages 25–35. ACM, 2013.

[80] Gary McKeown, William Curran, Ciaran McLoughlin, Harry J. Griffin, and Nadia Bianchi-Berthouze. Laughter induction techniques suitable for generating motion capture data of laughter associated body movements. *FG*, pages 1–5, 2013.

[81] D. McNeill. Hearing lips and seeing voices. *Nature*, 264(246-248), 1976.

[82] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.

[83] A. H. Mines. Respiratory physiology. 1993.

[84] K. G. Munhall, Jeffery A. Jones, Daniel E. Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15:133–137, 2004.

[85] Michael Neff and Eugene Fiume. Modeling tension and relaxation for computer animation. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, SCA '02.

[86] Victor Ng-Thow-Hing. *Anatomically-based Models for Physical and Geometric Reconstruction of Humans and Other Animals*. PhD thesis, Toronto, Ont., Canada, Canada, 2001.

[87] Radoslaw Niewiadomski and Catherine Pelachaud. Towards multimodal expression of laughter. In *IVA*, pages 231–244, 2012.

[88] Radoslaw Niewiadomski, Jennifer Hofmann, Jérome Urbain, Tracey Platt, Johannes Wagner, Bilal PIOT, Hüseyin Çakmak, Sathish Pammi, Tobias Baur, Stéphane Dupont, Matthieu Geist, Florian Lingenfelser, Gary McKeown, Olivier Pietquin, and Willibald Ruch. Laugh-aware virtual agent and its impact on user amusement. In *AAMAS*, pages 619–626, 2013.

[89] Radoslaw Niewiadomski, Maurizio Mancini, and Tobias Baur. MMLI: Multimodal multiperson corpus of laughter in interaction. *In proceeding of: 4th international workshop on Human Behavior Understanding*, 8212:pp 184–195, 2013.

[90] Magalie Ochs and Catherine Pelachaud. Model of the perception of smiling virtual character. In *AAMAS*, pages 87–94, 2012.

[91] I.S. Pandzic and R. Forcheimer. *MPEG4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, 2002.

[92] Frederic Ira Parke. *A Parametric Model for Human Faces.* PhD thesis, 1974. AAI7508697.

[93] Frederick I. Parke. Computer generated animation of faces. In *in ACM National Conference*, pages 451–457, 1972.

[94] Stefano Pasquariello and Catherine Pelachaud. Greta: A simple facial animation engine. In *Soft Computing and Industry*, pages 511–525. Springer London, 2002.

[95] C. Pelachaud, N.I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, January-March 1996.

[96] Catherine Pelachaud and Isabella Poggi. Subtleties of facial expressions in embodied agents. *The Journal of Visualization and Computer Animation*, 13 (5):301–312, 2002. ISSN 1099-1778.

[97] Ken Perlin. Improving noise. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '02, pages 681–682.

[98] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin. Synthesizing realistic facial expressions from photographs. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, pages 75–84, New York, NY, USA, 1998. ACM. ISBN 0-89791-999-8.

[99] I. Poggi. *Mind, hands, face and body. A goal and belief view of multimodal communication*. Weidler, Berlin, 2007.

[100] Robert Provine. Laughter. *American Scientist*, 84(1):38–47, 1996.

[101] Robert R Provine. *Laughter: A scientific investigation*. Penguin books edition, 2001.

[102] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[103] M. Radenen and T. Artières. Contextual hidden markov models. In *ICASSP*, pages 2113–2116, 2012.

[104] Byron Reeves and Clifford Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, New York, NY, USA, 1996. ISBN 1-57586-052-X.

[105] W. Ruch and P. Ekman. The Expressive Pattern of Laughter. *Emotion qualia, and consciousness*, pages 426–443, 2001.

[106] Willibald Ruch, Gabriele Kohler, and Christoph Van Thriel. Assessing the 'humorous temperament': Construction of the facet and standard trait forms of the state-trait-cheerfulness- inventory - stci. *Humor: International Journal of Humor Research*, 9:303–339, 1996.

[107] Martin Russell. A segmental hmm for speech pattern modelling. In *Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing: Speech Processing - Volume II*, ICASSP'93, pages 499–502. IEEE Computer Society, 1993. ISBN 0-7803-0946-4.

[108] JasonM. Saragih, Simon Lucey, and JeffreyF. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011. ISSN 0920-5691. doi: 10.1007/s11263-010-0380-4.

[109] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1330–1345, 2008.

[110] M. Schroeder and J. Trouvain. The MARY text-to-speech system, 2001.

[111] B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 2, pages II–1–4 vol.2, April 2003.

[112] W.A. Sethares and T.W. Staley. Periodicity transforms. *IEEE Transactions on Signal Processing*, 47(11):2953–2964, 1999.

[113] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *ICASSP*, pages 1315–1318, 2000.

[114] Jérome Urbain, Elisabetta Bevacqua, Thierry Dutoit, Alexis Moinet, Radoslaw Niewiadomski, Catherine Pelachaud, Benjamin Picart, Joëlle Tilmanne, and Johannes Wagner. The avlaughtercycle database. In *LREC*, 2010.

[115] Jérome Urbain, Huseyin Çakmak, and Thierry Dutoit. Automatic phonetic transcription of laughter and its application to laughter synthesis. In *Proceedings of ACII*, pages 153–158, 2013.

[116] Hannes Vilhjalmsson, Nathan Cantelmo, Justine Cassell, Nicolas E. Chafai, Michael Kipp, Stefan Kopp, Maurizio Mancini, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Zsofi Ruttkay, Kristinn R. Thï£¡risson, Herwin Van Welbergen, and Rick J. Van Der Werf. The behaviour markup language: recent developments and challenges. *LECTURE NOTES IN ARTIFICIAL INTELLIGENCE*, 4722:99–11, 2007.

[117] H.G. Wallbott and K. Scherer. Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, 24, 1986.

[118] K. Waters and T. Levergood. An automatic lip-synchronization algorithm for synthetic faces. In *Proceedings of the Second ACM International Conference on Multimedia*, MULTIMEDIA '94, pages 149–156, New York, NY, USA, 1994. ACM. ISBN 0-89791-686-7.

[119] T. Weise, B. Leibe, and L. Van Gool. Fast 3d scanning with automatic motion compensation. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007. doi: 10.1109/CVPR.2007.383291.

[120] J. West. Respiratory physiology: the essentials. 2004.

[121] A. D. Wilson and A. F. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:884–900, 1999.

[122] Chung-Hsien Wu and Wei-Bin Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21, 2011. ISSN 1949-3045.

[123] Jianxia Xue, J. Borgstrom, Jintao Jiang, L. Bernstein, and Abeer Alwan. Acoustically-driven talking face synthesis using dynamic bayesian networks. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1165–1168, 2006.

[124] Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43, 1998.

[125] F. E. Zajac. Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control. *Critical reviews in biomedical engineering*, 17(4):359–411, 1989.

[126] Victor Brian Zordan, Bhrigu Celly, Bill Chiu, and Paul C. DiLorenzo. Breathe easy: Model and control of simulated respiration for animation. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '04, pages 29–37, 2004. ISBN 3-905673-14-2.

Modèle Statistique de l’Animation Expressive de la Parole et du Rire pour un Agent Conversationnel Animé

RESUME : Texte

Notre objectif est de simuler des comportements multimodaux expressifs pour les agents conversationnels animés ACA. Ceux-ci sont des entités dotées de capacités affectives et communicationnelles; ils ont souvent une apparence humaine. Quand un ACA parle ou rit, il est capable de montrer de façon autonome des comportements multimodaux pour enrichir et compléter son discours prononcé et transmettre des informations qualitatives telles que ses émotions. Notre recherche utilise les modèles d’apprentissage à partir données. Un modèle de génération de comportements multimodaux pour un personnage virtuel parlant avec des émotions différentes a été proposé ainsi qu’un modèle de simulation du comportement de rire sur un ACA. Notre objectif est d'étudier et de développer des générateurs d'animation pour simuler la parole expressive et le rire d’un ACA. En partant de la relation liant prosodie de la parole et comportements multimodaux, notre générateur d'animation prend en entrée les signaux audio prononcés et fournit en sortie des comportements multimodaux.

Notre travail vise à utiliser un modèle statistique pour saisir la relation entre les signaux donnés en entrée et les signaux de sortie; puis cette relation est transformée en modèle d’animation 3D. Durant l'étape d’apprentissage, le modèle statistique est entrainé à partir de paramètres communs qui sont composés de paramètres d'entrée et de sortie. La relation entre les signaux d'entrée et de sortie peut être capturée et caractérisée par les paramètres du modèle statistique. Dans l'étape de synthèse, le modèle entrainé est utilisé pour produire des signaux de sortie (expressions faciale, mouvement de tête et du torse) à partir des signaux d'entrée (F0, énergie de la parole ou pseudo-phonème du rire). La relation apprise durant la phase d'apprentissage peut être rendue dans les signaux de sortie.

Notre module proposé est basé sur des variantes des modèles de Markov cachés (HMM), appelées HMM contextuels. Ce modèle est capable de capturer la relation entre les mouvements multimodaux et de la parole (ou rire); puis cette relation est rendue par l’animation de l’ACA.

Mots clés : Modèle de Markov caché, Agent Conversationnel Animé, Synthèse d’Animation, Animation de la Parole, Animation du Rire