Speech Driven Upper Body Gesture Analysis and Synthesis

by

Serkan Özkul

A Thesis Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computer Engineering

Koç University

September, 2012

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Serkan Özkul

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Yücel Yemez

_____

Assist. Prof. Metin Sezgin

_____

Assist. Prof. Sinem Ergen

Date: _____

ABSTRACT

In this thesis we present a new computational model for natural and believable upper-body gesture synthesis in synchrony with speech using statistical learning techniques over multimodal gesticulation data. The framework consists of four main tasks for: i) unimodal clustering of gesture and intonational phrases, ii) multimodal analysis of gesture and intonational phrases, iii) speech driven gesture synthesis, and iv) gesture animation. The first task consists of unimodal analysis of speech and upper body motion to learn temporal patterns of gesture and speech prosody. Body motion features, which are extracted from multi-channel synchronous video recordings, are used to define gesture phrases with a semi-supervised temporal clustering scheme. On the other hand prosody features, which are extracted from speech input, are used to define intonational phrases with an unsupervised temporal clustering scheme. The second task performs multimodal analysis to learn dependencies between gesture and intonational phrases by utilizing a hidden semi-Markov model (HSMM). Third, we perform gesture synthesis, that is extraction of gesture sequence and gesture durations, given the speech input. The final task is to perform gesture animation, where the synthesized gesture sequence is mapped into body motion sequences to maintain a natural looking animation. The performance of the proposed speech driven gesture synthesis system is tested over our MVGL-MUB Database. Experimental results demonstrate that our system is able to properly discover audiovisual correlations between speech and gesture thus it can synthesize realistic and natural body gestures along with 3D human model animation.

ÖZETÇE

Bu tez çalışmasında, çok kipli beden hareketi verisi üzerinde istatistiksel öğrenme teknikleri kullanarak, konuşma ile eşzamanlı, doğal ve inandırıcı üst beden hareketleri sentezi için yeni bir çatı yapısı ve sayısal model önerilmektedir. Önerilen çatı yapısı 4 ana kısımdan oluşmaktadır: i) üst beden hareketi ve prozodik bölütler üzerinde tek kipli kümeleme, ii) jest ve prozodik bölütler üzerinde çok kipli analiz, iii) konuşma güdümlü jest sentezi ve iv) beden jest animasyonu. İlk kısım, jestlerin ve konuşma prozodisinin zamansal örüntülerini öğrenmek için konuşma ve beden hareketlerinin tek kipli analizinden oluşmaktadır. Jest örüntülerinin belirlenmesi çok kanallı ve eş zamanlı video kayıtlarından çıkarılan beden hareketlerinin yarı denetlemeli zamansal kümelenmesi ile sağlanmıştır. Buna karşılık prozodi örüntüleri ise konuşma girdisinden çıkarılan prozodi özniteliklerinin denetimsiz zamansal kümelenmesiyle tanımlanmıştır. İkinci kısım, konuşma ve jestler arasındaki bağıntıları öğrenmek için gizli yarı Markov modellerine dayalı çok kipli bir analiz yöntemi kullanmaktadır. Üçüncü kısım beden hareketi sentezi problemini ele alır; bu da konuşma girdisi verildiğinde jest sekansının ve jest sürelerinin oluşturulmasına karşılık gelir. Son kısımda ise, sentezlenmiş hareket dizisinden doğal görünümlü bir üst beden hareketi animasyonunun oluşturulması hedeflenir. Önerdiğimiz konuşma güdümlü jest animasyon sisteminin başarımını oluşturmuş olduğumuz MVGL-MUB veritabanı üzerinde ölçüyoruz. Elde ettiğimiz deney sonuçları, önerdiğimiz sentez sisteminin, konuşma ile beden hareketleri arasındaki işitsel-görsel bağıntıyı uygun şekilde modellediğini ve böylece gerçekçi ve doğal üç boyutlu insan modeli animasyonları üretebildiğini göstermektedir.

ACKNOWLEDGMENTS

I would like to express my endless gratitude to Assoc. Prof. Yücel Yemez for his excellent advisory, reliable guidance and full support. This thesis have not been written without his profound knowledge. I thank to Assoc. Prof. Engin Erzin for his guidance, constructive comments, and help in every stage of this work. Moreover, I would like to express my special thanks to Prof. A. Murat Tekalp for being a great influence, inspiration and support. Without their reliable guidance and encouragement, this thesis would not have been a success.

I also thank to Assist. Prof. Metin Sezgin and Assist. Prof. Sinem Ergen for being in my thesis committee and for their valuable time. I specially thank to my friends Hilmi E. Eğilmez, Tolga Bağcı, Tuğtekin Turan, Shahriar Asta, Elif Bozkurt, Yalçın Şadi, Yusuf Salihlioğlu, Tahsin Dane and rest of my friends for their support, contribution in MVGL-MUB Database and providing me a two enjoyable years. I would like to show my thanks again to Shahriar for helping me in preparing Chapter 5.

Finally, I would like thank my family for their apprehension and support.

TABLE OF CONTENTS

List of Tables **viii**

List of Figures **ix**

Chapter 1: Introduction **1**

1.1 Related Work 4

1.2 Contributions 6

1.3 System Overview 6

Chapter 2: Motion Capture **9**

2.1 Motion Capture 9

2.2 Initialization of the system 12

2.3 3D Body Tracking 12

2.4 MVGL Motion Capture System 14

Chapter 3: Multimodal Database **20**

3.1 Existing multimodal databases 21

3.2 MVGL MUB Database 22

Chapter 4: Unimodal Clustering **25**

4.1 Unsupervised Clustering 25

4.2 Unsupervised Clustering of Prosody 27

4.3 Semi-Supervised Clustering of Gesture 29

Chapter 5: Multimodal Analysis and Synthesis of Prosody-Driven Gestures **33**

5.1 Gesture Generation Model 33

5.2 Gesture Synthesis 37

Chapter 6: Gesture Animation **39**

Chapter 7: Experimental Evaluations **45**

7.1 Objective Evaluations 45

7.2 Subjective Evaluations 50

Chapter 8: Conclusions **55**

Bibliography **58**

LIST OF TABLES

2.1 Optical Marker Placements on Actor 16

4.1 Gesture Labels and Descriptions 32

7.1 The Subjective A-B Comparison Results 51

LIST OF FIGURES

1.1 The block diagram of the general framework for the automatic hand gesture synthesis system. 7

2.1 MVGL Motion Capture System Setup 15
2.2 Demonstration of Marker Positions to Joint Angles and Marker Assignments in MotionBuilder 17
2.3 Simultaneous Camera Shot from the MoCap System 18

4.1 Parallel Branch HMM Structure 27
4.2 Upper body joint angles used in our system 30
4.3 *Gesture description figures* 32

5.1 *In a Hidden Semi-Markov Process each state has a duration and emits a number of observations.* 35

6.1 Gesture Animation Generation System 40
6.2 Viterbi Path Selection (Note that not all gestures have same duration. For each desired gesture class, there different number of gesture choices in the pool) 42
6.3 Smoothing Filter trials on Joint Angle Data 43
6.4 Screenshots from synthesized body animation 44

7.1 *The boundary distance between actual (g) and synthesized ($\hat{g}$) gestures.* 46
7.2 *A comparison of histogram similarity and boundary distance between various numbers of clusters (a) and states (b) for prosody patterns* . . 48

7.3 *Duration model comparison of synthesized and original gesture sequences (Actor #1)* 52
7.4 Duration model comparison of synthesized and original gesture sequences (Actor #2) 53
7.5 *Characteristic Gesture Frequency Comparison* 54

Chapter 1

INTRODUCTION

In casual face to face communication, mutimodality is the irrevocable part among the people. Human beings concurrently use verbal and non-verbal communication ways such as kinesics, occulesics, haptics, vocalics and similar methods to convey their thoughts and emotions in face to face. Ability of combining information from different senses enables humans to extract information from and understand complex environments. Similar examples can be observed in various HCI systems where speech, gestures, haptics and other human communication channels in establishing communicative interfaces. Multimodality provides us comprehension of environment in most cases meanwhile unimodal channel is insufficient to interpret. Accordingly, multimodal signal analysis and processing deals with the challenge of simultaneously handling with multiple sources of information.

Multimodal signal processing takes advantage of analyzing the underlying mutual relationship of signals of different sources which is a reinforcement that complements the weakness or insufficiency of one modality by using the strengths of other modality. For instance, in recognizing ambiguous syllables in speech recognition, lip reading algorithms to remove obscurity as an assistance. Moreover, this method can be used to generate speech driven face avatar animation. Also additional study for the underlying multimodal correlation of speech and facial expressions can lead to more realistic and natural speech driven face avatar animations. Yet another example is the integration of the visual information, robotics and haptics which can provide the necessary infrastructure for medical applications such as remote surgery.

The aim of this research is to investigate methods for combining different audio-

visual of information for applications in human computer interaction (HCI). Main motivation of this thesis is modeling the joint correlation between speech and upper body gestures, thereafter generating 3D human body animations driven by speech. We have formed a multimodal database to analyze, learn and synthesize audiovisual models and to generate of body gesture animations. The joint correlation model can be considered as a bidirectional mapping between speech and body gesture patterns. In gesture synthesis, this mapping can be used to estimate upper body gesture patterns from speech patterns.

Speech and gesture are two different ways of expressing human thoughts and emotions. As a verbal and non-verbal human communication, they co-exist in time with a tight synchrony and reveal information on verbal semantics and speaker emotions. Moreover, gesticulation is an essential component of face to face communication, and it contributes significantly to the natural and affective perception of human conversations. Human-centered HCI designs increasingly use interactions in virtual environments, where a natural, affective and believable gesticulation is often missing in virtual character animations. Automatic synthesis of gesticulation in synchrony with speech is expected to make non-verbal communication a natural part of virtual character animation, which can find wide range of applications in human-centered HCI, game industry and film industry. In this paper we present a new computational model for natural and believable upper-body gesture synthesis in synchrony with speech using statistical learning techniques over multimodal gesticulation data.

In one of the pioneering studies on gesture and speech relationship, Kendon proposed a widely accepted hierarchical model for gesture. In this model the core gestural element is defined as gesture phase. Gesture phases are further divided into active and passive phases. Active gesture phase includes strokes (the short and dynamic peak movement) and retractions. Passive gesture phase consists of gestures like hold, in which hand stays motionless. When a gesture phase is preceded by preparation, in which hand goes to the position of the gesture phase, it constitutes a gesture phrase. Furthermore, a combination of gesture phrases forms gesture units. In this hierarchi-

cal model, semantic expressiveness of hierarchy levels increases as we move further away from the core. In other words, gesture units are semantically more expressive than gesture phrases and gesture phrases include more semantic contents than gesture phases.

Synchrony between gestural and phonological structures has been studied. Kendon stated the synchrony between strokes and stressed syllables, later McNeill proposed the widely accepted phonological synchrony rule stating that the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech. In a recent study, Loehr presents a detailed investigation of temporal and structural synchrony between intonation and gesture. His findings verify the alignment of the pitch accents with the gestural strokes; furthermore he presents evidences of the synchrony between gesture phrases and intermediate intonational phrases.

There are four widely referred types of gestures, which were proposed by McNeill [3]: iconics, metaphorics, deictics and beats. Iconic gestures illustrate images of an object or action, metaphoric gestures represent abstract ideas, deictic gestures relatively locate entities in physical space, and beat gestures are simple repetitive movements to emphasize speech. In a later study, McNeill [4] points out Tuites proposal [5] that in every gesture there is a rhythmical beat-like pulse to carry significance beyond its immediate setting, and he suggests taking metaphoricity, iconicity, deixis, and emphasis as the dimensions of gesture rather than types of gesture. Hence nonverbal communication channel is expected to articulate a mixture of gesture dimensions for natural human-to-human communication. Similarly, automatic synthesis of gesture dimensions for animated characters (avatars) in virtual worlds would create natural and believable interactions over the nonverbal communication channel.

Human-computer interaction (HCI) is expected to benefit largely from human-centered designs rather than computer-centered designs. Human-centered designs should establish human-like information retrieval and information transfer to attain natural, affective and believable interactions. The extend and number of studies on affect recognition methods for human-like information retrieval have been growing

in the recent literature, an extensive survey on affect recognition technologies and methodologies can be found in (Zhihong). On the other hand, human-like information transfer, as well, poses many important research problems. Development of concepts and methodologies for human-like avatars appears as the challenging problems of human-like information transfer, which will eventually provide synthesis of full-body gestures in synchrony with emotion and speech.

1.1 Related Work

Previous literature on gesture and speech is governed by studies from psychology, computer vision, speech processing, linguistics, and machine learning. Full-body gesture synthesis can deliver valuable solutions for human-like information transfer. We can classify gesture synthesis into rule-based and data-driven approaches. The Embodied Conversational Agents (ECAs) of Cassell [1] is a pioneering rule-based full-body gesture synthesis system, which performs animations over a pre-defined gesture tree. Data-driven approaches either use gesture statistics from audiovisual data or 3D motion capture and animation technologies. Annotation of audiovisual data is commonly used to define gesture phase, phrase and unit statistics, as well as lexical and intonational characteristics of the accompanying text/speech. The VirtualHuman research project [2] and the probabilistic approach of Neff et al. [3] are examples of audiovisual data-driven approaches for full-body gesture animation. The VirtualHuman project aims to develop interactive virtual characters with a personality profile. Neff et al. [3] presents a probabilistic approach to produce full-body gesture animation for given input text in the style of a particular performer. They have a tool assisted annotation process over audiovisual data to define statistical style model of a particular performer. We shall further note that the focus of the ECAs and the VirtualHuman project is on iconic and deictic gestures where speech and gesture relationship is well defined. On the other hand the focus in [3] is on metaphoric gestures where speech and gesture relationship is performed with the statistical style model. Recently, Levine et al. [4] introduce gesture controllers, which avails a modular methodology to drive

beat-like gestures, with live speech using customized gesture repertoires. Gesture controllers infer hidden states from speech, and select the optimal gesture kinematics based on the inferred states. From a hierarchical perspective, the work of Levine et al. is mainly concentrated on the gesture phase level. Although motion capture is becoming widely available, there is limited number of studies in the literature on processing of 3D motion data rather than using it for 3D reconstruction. Heloir et al. [5] provides technical setup, scenarios and challenges in building a motion capture database for virtual human animation. Similarly, Busso et al. [6] presents the interactive emotional dyadic motion capture (IEMOCAP) database, which is a multimodal and multispeaker database of improvised dyadic interactions.

Early works on prosody driven gesture synthesis mostly concentrate on facial expression and head motion. Face animation with expressions using neural networks [7], and multimodal communication using affine transformations [8] are among the works on facial expression synthesis. An approach to synthesize emotional head motion sequences driven by prosodic features is presented in [9] by building hidden Markov models for emotion categories to model temporal dynamics of emotional head motion sequences. A two-stage framework for joint analysis of head gesture and speech prosody patterns of a speaker towards automatic realistic synthesis of head gestures from speech prosody has been studied in [10]. A recent paper [11] focuses on building a speech-driven facial animation framework to generate natural head and eyebrow motions using dynamic Bayesian networks (DBNs).

In this study, we employ hidden semi-Markov model (HSMM) for the multimodal analysis of gestures and prosody. The HSMM was first introduced by Ferguson [12] as the explicit duration hidden Markov models. The main intuition behind the HSMM idea is to extend hidden Markov models to processes where states have durations and thus emit a number of observations instead of a single one. This yields that the underlying process in the system is Markovian in certain jumps. Moreover, the state duration is allowed to follow a probabilistic distribution. Based on the work of Ferguson [12], several similar methods have been proposed [13–15]. Since the duration

of human body gestures are variable in nature, the problem of generating body gesture sequences from prosody observations fits into the concept of HSMMs. To our best knowledge this is the first time that HSMM is considered for the task of synthesizing body gestures from prosody observations.

1.2 Contributions

In this study we describe a framework for analyzing audio-visual data and synthesizing upper body gestures from speech prosody using HSMM. From a hierarchical perspective, our work is mainly concentrated on gesture phrases which are semantically more expressive than gesture phases studied in the work of Levine et al. [4]. Hence our framework provides a more personalized and natural speech-driven gesture synthesis, as also demonstrated by experiments.

The main contributions of this thesis work are:

- A framework for unimodal analysis of recurrent prosodic speech and upper-body gesture patterns
- Construction of the Multimodal Upper-Body (MVGL-MUB) Corpus
- A gesture animation method that maps synthesized gesture sequences to body motion sequences

We note that the HSMM-based gesture synthesis method that we use in this thesis work is adapted from [16].

1.3 System Overview

The general framework for our automatic hand gesture synthesis system is given in Fig. 1.1. The framework consists of four main functional blocks for: i) unimodal clustering of gesture and intonational phrases, ii) multimodal analysis of gesture and intonational phrases, iii) speech driven gesture synthesis, and iv) gesture animation.

The first functional block consists of the unimodal analysis of speech and body motion to learn temporal patterns of gesture and speech prosody. Body motion features, which are extracted from multi-channel synchronous video recordings, are used to define gesture phrases with a semi-supervised temporal clustering scheme. On the other hand prosody features, which are extracted from speech input, are used to define intonational phrases with an unsupervised temporal clustering scheme. The second functional block performs multimodal analysis to learn dependencies between gesture and intonational phrases by utilizing a hidden semi-Markov model (HSMM). In the third functional block, we perform gesture synthesis, that is extraction of gesture sequence and gesture durations, given the speech input. Finally the fourth functional block performs gesture animation, where the synthesized gesture sequence is mapped into body motion sequences to maintain a natural looking animation.

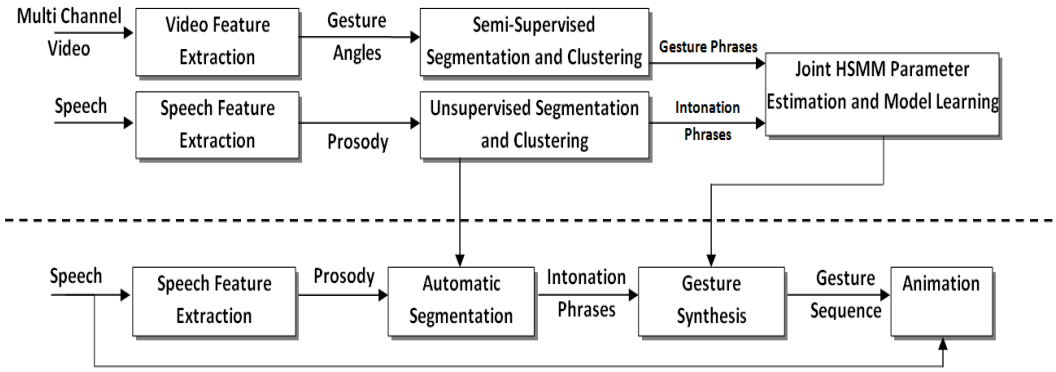


Figure 1.1: The block diagram of the general framework for the automatic hand gesture synthesis system.

The remainder of this thesis is organized as follows: Chapter 1 reviews the current literature on speech-driven gesture analysis and synthesis. Chapter 2 describes our Motion Capture system and provides details about the recording hardware and software specifications. Chapter 3 reviews current multimodal databases and presents our MVGL-MUB corpus in detail. Our framework for unimodal clustering of upper-body gestures and speech prosody is explained in Chapter 4. Multimodal analysis and synthesis of prosody-driven upper body gestures is described in Chapter 5. Chapter 6

describes our gesture animation method that maps synthesized gesture sequences to body motion sequences. Experimental evaluation of the proposed system is provided in Chapter 7. Finally, concluding remarks are given in Chapter 8.

Chapter 2

MOTION CAPTURE

2.1 Motion Capture

Motion capture systems have consistently been evolving and there are various techniques for capturing and modeling the motion. We can basically divide these motion capture techniques into 4 groups as stated below;

Types of Motion Capture Systems:

- Optical
- Inertial
- Mechanical
- Magnetic

Optical motion capture systems use visual data for tracking the motion of reference points that are clearly visible and previously attached on the points of interest. In this method, optical markers, which are inside of the field vision of mocap cameras, are used to mark objects for tracking. Marker based systems depend on the contrast of the marker color with the background color to maintain tracking operation. Optically active and passive markers are the types which are currently used in motion capture systems. Active optical capture systems such as LED markers that pulse flashes in synchronization with the cameras digital shutters. Passive optical systems, such as markers made of simple beam and IR light reflective matter which are visible in any lighting condition. These methods however cannot acquire the shape and texture properties of the object, which could also give supplementary information

about location of marker points. There exist a number of marker-based commercial systems as used for human motion capture but most of them rely on a high number of cameras to avoid occlusions, high frame rates or expensive hardware. [17] described method for low-cost accurate marker tracking system which previously analyzes skeleton structure parameters of the actor and maintain tracking of the body markers even if occlusion happens using skeleton this calibration parameters.

Markerless systems such as Microsoft Kinect do not require any indicators like markers to estimate motion of the body parts. They use computer vision algorithms to detect significant body parts such as face, torso, arms and legs and estimates the inter-limbs between them. Moreover kinect uses an IR light source and a receptor to measure depth of the player and combines depth data with visual data to distinguish body parts along with the motion of them. Wren [18] presented a markerless motion capture system Pfinder which tracks full body in real-time and interprets behaviors the person. Today in wide range of applications such as wireless interfaces, video databases and low-bandwidth coding, Pfinder features are being used.

For the other types of mocap systems such as inertial motion capture uses special sensors to capture motion of the object with less accuracy but do not use any visual data or cameras. In this system, inertial sensors such as gyroscopes are used to measure rotational rate of the attached objects. In human motion capture, rotational data received from the gyroscope ,that are attached to several limbs of the subject, can be transferred into a skeleton model in a limited accuracy. [19] proposed a new algorithm to accurately estimate linear acceleration of compulsive motions such as running. Since gyroscopes suffer from the rigorous movements, actor is still delimited to make swift motions.

Mechanical motion capture systems can directly track joint angles of the body which requires an exoskeleton (skeletal like articulated structure) to be worn for capturing relative motion of the performer. Potentiometers between articulated parts measures the Euler angles of the joints and transmits them to a base station for real time motion capture. Mechanical motion capture systems are relatively occlusion

free and have no limitation about volume of capture but restrict movements due to being heavy weight. Moreover, exoskeleton motion capture systems are widely used in robotic control applications such as robotic imitation of the gestures captured by the exoskeleton. [20] presents hardware implementation and control interface between exoskeleton and operational robot for teleoperation tasks with motion capture.

In magnetic systems of motion capture can calculate both range and the orientation between the attached magnetic sensors to body limbs. Measured relative magnetic flux between the sensors enables system to calculate both relative range and orientation of the sensors by mapping tracking volume. Hidden or occluded objects can be tracked by using the magnetic field of the markers. These sensor markers are not occluded by non-metallic objects but any likely magnetic or electrical inference will cause spikes and corruption in the sensor data. In addition, cross talk is the most common error between magnetic mocap sensors. [21] proposes a new mocap system with magnetic sensors which overcomes the cross-talk problem even in adjacent distances by usage and analysis of multiple axial sensors data.

In this work, we have used low-cost solution to visually track markers with synchronized multicamera system. For tracking the passive optical markers, we are using our own program TrackBuddy which is implemented on OpenCV library [22]. The tracking program tracks 2D pixel positions of the markers, attached to joints of the body parts, then calculates the 3D coordinates of the markers using calibration parameters and 2D projections on each camera's image plane. We make use of the multistereo correspondence information from multiple cameras to precisely estimate 3D positions of the markers. This provides us with a set of 3D point locations for each frame recorded. We employ Kalman filtering for smoothing out the observations and predicting future location of the points in that point cloud. This program also allows users to intervene into tracking process, therefore users can monitor whole tracking process and correct mis-tracking or lead tracking manually. However, the tracking process itself may become time-consuming and cumbersome.

2.2 Initialization of the system

For each frame in the recorded video sequence, a set of N images are obtained from the N cameras. All cameras are calibrated using a pinhole camera model based on perspective projection. Our camera calibration process is maintained by the MultiCamera Self-Calibration Tool [23] which requires additionally a laser pointer to calculate extrinsic and intrinsic parameters of the cameras.

Camera Calibration is a process of calculating true parameters of the camera by examining a captured video or photograph to deduce camera situation at the capture time. The extrinsic true parameters indicate the translation and rotation of the camera which determines the captured scene. The intrinsic parameters specifies focal length, principal axis, CCD dimensions (image resolution) and the skew parameter of the camera that varies among different cameras. In the calibration process, we have used a red laser pointer that roams in the stage while all cameras are simultaneously recording. After recording several hundreds of frames from each camera, we fed the calibration tool with these frames to solve multi-stereo correspondence problem [23].

2.3 3D Body Tracking

In order to track multiple markers from multiple cameras, all frames are converted to YCrCb color space which provides convenience and flexibility to track intensive colors and distinguish different colors in the frames captured by cameras from different views. Moreover, markers can be occluded by other objects therefore occluded markers may not be seen from the camera and the number of detected markers in every image may vary. However, we can overcome this problem by using the redundancy among multiple views. To calculate 3D coordinates of a marker, theoretically we need that marker to be visible to 2 cameras at least. So the usage of high number of cameras with different view angles, enables the marker likely to be seen by several cameras and a few camera occlusions would not cause problems. On the other hand a marker can get into occlusion by another marker and marker identification can be problematic in

the cease of occlusion but user can intervene the TrackBuddy and correct the marker identifications manually.

There are M_m markers attached to the actor and let W set of search windows to track markers where $W = \{w_1, w_2,..., w_{M_m}\}$. Each search window w_i is initially set on marker positions by the user to start tracking process. The search window set W is used to track all markers and store the 2D (x, y) positions of the markers for each frame.

Since both motion capture suit of the actor and dark color background plane have contrast colors with the markers, only the skin color of the actor and the markers should be meticulously distinguished. To differentiate markers with the skin color, we have used [22] OpenCV library skin color detection and removal functions which automatically detects skin color by using adaptive color thresholding. To track markers over black background, we apply predefined color thresholding to pixels inside each search window w_i. Among the pixels that overpass color threshold value, we calculate the mean coordinate point of elected pixels and set this point as the marker coordinate. We make use of the multistereo correspondence information to estimate 3D world coordinates of the markers and a set of 3D (x, y, z) points over time is obtained. 3D world coordinates of the markers are estimated via epipolar geometry based on marker projections on image plane of each camera. Moreover, we use Kalman filtering to smooth 3D marker motions and to predict occluded marker locations in similar method explained in [24].

In order to speed up the tracking process, we have embedded basic color comparison operations. Since displacement of the markers between successive frames are minor, spatial-temporal information is exploited to predict marker position and decreasing the marker searching area in latter frames.

3D marker locations are converted to joint angles to create more reliable and useful data by using a professional software [25]. In addition, most of the 3D character animating systems work with joint angles. Given set of 3D marker locations, [25] estimates joint angle motions by fitting a 3D human model into 3D marker motions.

In this process each marker should be attached to a body limb of the 3D human model (Fig. 2.2) thus [25] can calculate joint angle motions by using inverse kinematics. As an input/output relationship, [25] takes a set of 3D marker displacements over time and outputs a set of angle motions of the joints over time.

2.4 MVGL Motion Capture System

At Multimedia, Vision and Graphics Laboratory, an automated motion capture system with 8 cameras is available to collect and analyze multi-view video data, which is primarily used for human body motion modeling applications. The motion capture (MOCAP) system can track optical markers attached to the actors body and generate corresponding 3D world coordinates of the markers. In our experiments we have made recordings with 5 cameras as shown in Fig. 2.1, which are sufficient to capture upper and lower body movements of the actor. Each mocap camera can record 20 frames per second simultaneously at 1392x1040p resolution.

We use the 5 cameras of the MVGL 8-camera optical motion capture system, which are placed on a truss structure on the ceiling, positioned three meters away to capture the upper body movements of the participant in all directions for 3D tracking. These cameras are placed like a curve of a semicircle to prevent any marker getting out of the field of vision. This setting also increases the multicamera calibration accuracy.

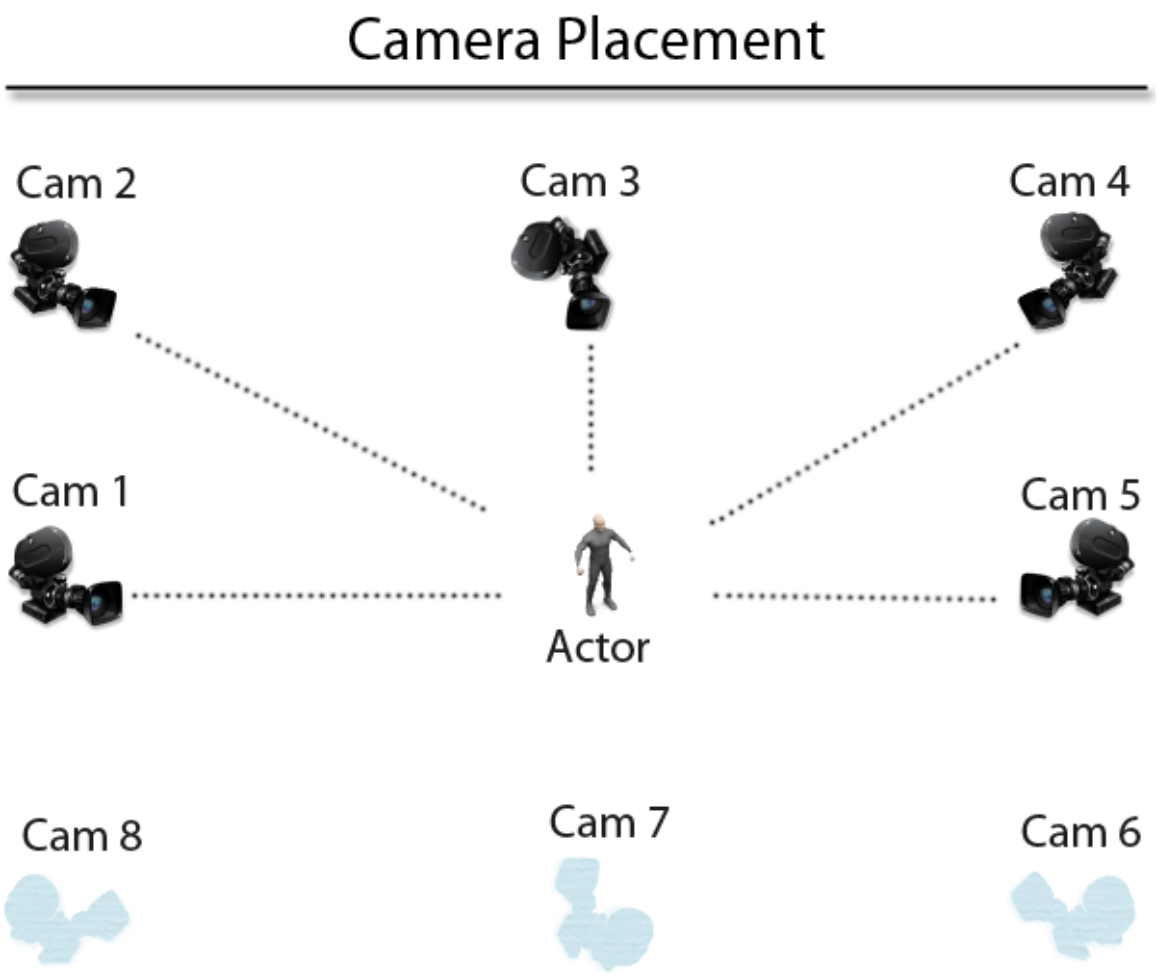


Figure 2.1: MVGL Motion Capture System Setup

Moreover, during the recordings participants wear a black motion capture suit that is covered with reflective optical markers, which are clearly visible in any lighting condition. Totally 15 markers are used in the upper body tracking process that are attached to $M_m = 15$ human upper body parts:

Body Parts	**Attached Marker Count**
Head	3
Chest	1
Hands	2 (Left Hand/Right Hand)
Elbows	2 (Left Elbow/Right Elbow)
Shoulders	2 (Left Shoulder/Right Shoulder)
Legs	2 (Left Leg/Right Leg)
Feet	2 (Left Feet/Right Feet)
Back	1

Table 2.1: Optical Marker Placements on Actor

Our motion capture process is a marker-based approach where a set of distinguishable color markers is attached to the joints of the participant. As a result of optical marker tracking process, 3D world coordinates for each frame and each marker are calculated by using the camera calibration parameters. After getting the 3D marker movement data, smoothing operations are applied to marker data to eliminate noises and latter a professional software, MotionBuilder [25], is used to convert 3D marker data to joint angles (Fig. 2.2).

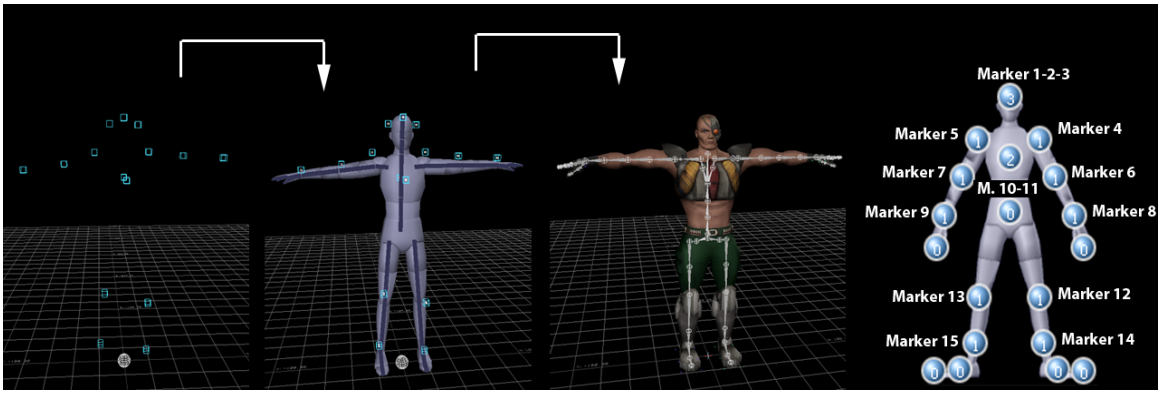


Figure 2.2: Demonstration of Marker Positions to Joint Angles and Marker Assignments in MotionBuilder

Optical markers with different colors are used in our recording to improve marker-tracking performance. Red optical markers are attached to the most active body parts such as hands. For more stable parts green and white optical markers are used. Searching and tracking of the markers are maintained inside of the corresponding search window w_i with 5x5 pixels wide.

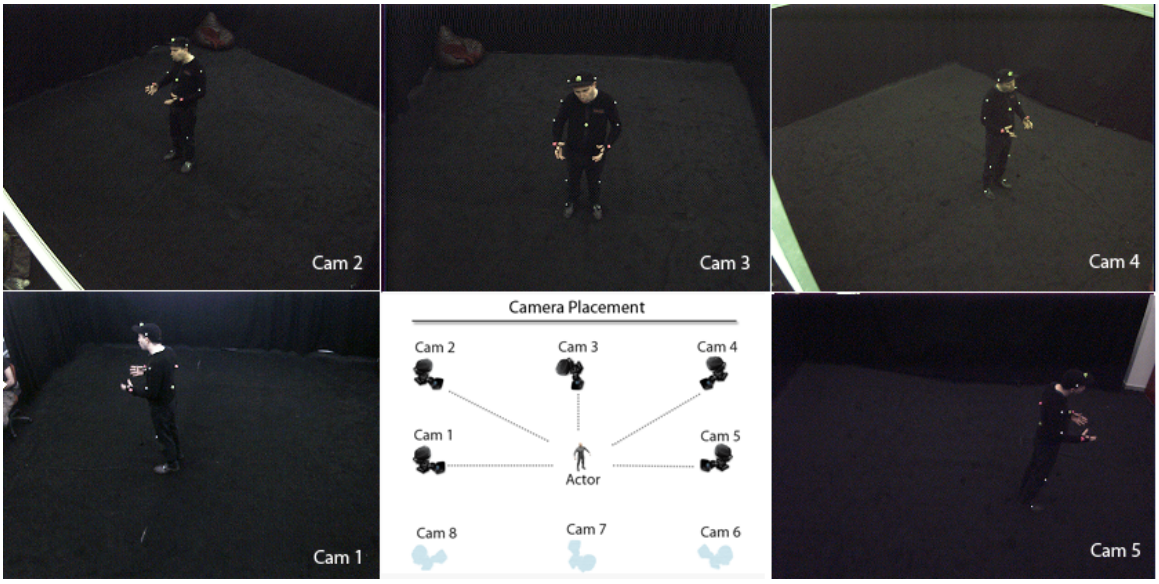


Figure 2.3: Simultaneous Camera Shot from the MoCap System

Speech is recorded with a high quality Sony ECM-166BMP lapel microphone worn by the participant close to mouth, so that movement of the participant would not cause volume alterations in the recordings. Moreover, it is small enough that the microphone does not occlude body markers. Lapel microphone and upper body recording system are both connected to a high capacity server that manages the synchronization between audio and video information. Speech of the actor is recorded at 16K sampling rate and stored in wav-formatted files.

The actor stands in the middle of the room during the recordings. We give the participant some degree of freedom in moving, but this freedom is limited to moving in place. The recordings are performed in the MVGL multi-camera studio with stable and good lighting conditions. A good lighting condition means that there is enough

diffuse light to leave no shadows on the actors body. All cameras are focused on the participant. Sample views from multi-camera are given in Fig. 2.3. The background and the floor of the room are covered with black color.

Chapter 3

MULTIMODAL DATABASE

An ideal system for automatic analysis and recognition of human affective information should be multimodal, as the human sensory system is. The integration of multiple sources of information would enhance the power for achieving a reliable emotional recognition that building multimodal databases is considered a very important issue for affective computing research. While this need is clearly acknowledged within research community, very few large multimodal databases are available. Most of the databases deal only with speech or facial expressions and even when considering few more complete multimodal databases available they mostly combine audio and visual (facial expression) information, very few uses 3D upper-body gesture information.

Moreover, naturalness of the emotional database is another issue. Acquiring realistic emotional data is a challenging task. In fact, many of the databases available ask subjects to act or pose emotions in order to extract speech and facial related features. In the last years, this lack of naturalism has been severely criticized. Recent research is oriented towards inducing emotions of speakers (elicited databases) or collecting real-life data. There have been a large number of studies on emotion and non-verbal communication of facial expressions and also on expressive body gestures. Yet, these studies were mostly based on acted basic emotions.

In addition, range of emotions in a database is also a concern. Early works considered only the basic seven emotion classes, namely, anger, happiness, sadness, surprise, fear, disgust and neutral assuming this classification is independent of cultural background. However, recent research include emotions exceeding traditional primary emotions such as shame, frustration and boredom. These are believed to play a key role both in learning processes and interpersonal relationships.

Data in a spontaneous database reflects a wide variety of emotions heavily dependent on the contextual information. In terms of annotations, affective databases can be annotated with categorical or dimensional labels. Categorical annotation uses discrete emotion classes to describe emotional content (e.g., happy, sad, etc.), while dimensional annotation uses the activation-valence coordinate space to describe emotional content (e.g., positive-negative valence vs. high-low arousal). Categorical labels are easy to use but they cannot be effectively used to cover the entire spectrum of emotions. On the other hand, dimensional labels are more flexible but they suffer from issues like inter-annotator agreement and emotional baseline detection.

3.1 Existing multimodal databases

HUMAINE database is one of the most comprehensive multimodal databases which was collected during the Third Summer School of the HUMAINE EU-IST project, held in Genova in September 2006. Acted emotional state recordings of anger, despair, interest, pleasure, sadness irritation, joy and pride incorporated facial expressions, body movements and gestures and speech. The database is also multi-lingual including recordings in languages English, French and Hebrew [26].

The CreativeIT database is a naturally induced affective database. The corpus is annotated with the continuous emotional descriptors (valence, activation) and collected using cameras, microphones and motion capture containing detailed audiovisual information of the actors' body language and speech cues. It serves two purposes. First, it provides insights into the creative and cognitive processes of actors during theatrical improvisation. Second, the database offers a well-designed and well-controlled opportunity to study expressive behaviors and natural human interaction [27].

IEMOCAP database is also multimodal including speech and facial expressions. Ten actors were asked to perform three selected scripts with clear emotional content. In addition to the scripts, the subjects were also asked to improvise dialogs in hypothetical scenarios, designed to elicit specific emotions (happiness, anger, sadness, frustration and neutral state). One participant of the pair was motion captured at a

time during each interaction. Fifty-three facial markers were attached to the subject being motion captured, who also wore wristbands and a headband with markers to capture hand and head motion, respectively [28].

Some of the most successful efforts to collect new emotional databases have been based on broadcasted television programs. Some of these examples are the Belfast natural database [29], the VAM database [30] and the EmoTV1 database [31]. Likewise, movie excerpts with expressive content have also been proposed for emotional corpora, especially for extreme emotions (e.g., SAFE corpus [32]).

Other attempts for collecting natural databases were based on recordings in place (Genova Airport Lost Luggage database [33]), recordings of spoken dialogs from real call centers (the CEMO [34], and CCD [35] corpora), asking the subjects to recall emotional experiences [36], inducing emotion with a Wizard of Oz approach in problem-solving settings using a human-machine interface (e.g., SmartKom database [37]), using games specially designed to emotionally engage the users (e.g., the EmoTaboo corpus [38]), and inducing emotion through carefully designed human-machine interaction (i.e., SAL [39]).

Our database is mainly focused on building a multimodal database by taking into account different sources of data such as upper-body movements, speech in a natural fashion. We systematically investigated the difference between emotions in the considered modalities and to verify the existence of systematic correlation between the different measures. In our database we considered categorical annotations.

3.2 MVGL MUB Database

The Multimodal Upper-Body (MVGL-MUB) Corpus consists of 42 recordings (about 2 hours long in total) from five pre-defined Turkish scenarios with 7 participants. Three scenarios (23 records) are monologue conversations. The rest of the conversations is dialogue. In this corpus different types of scenarios have been used to extract characteristic features from gestures. The recordings include storytelling of memories, documentaries, fairy tales as well as conversations and watch & tell scenarios. Each

scenario is designed for natural and transparent interaction of participants within the recordings. Detailed information can be accessed in [40]. A summary of the scenario descriptions are given as following:

Scenario 1: Storytelling a memory The first scenario consists of telling an exciting event that the subject/participant faced with. Each subject tells an incident about his/her experience avoiding pretended gestures in spontaneous manner.

Scenario 2: Storytelling a documentary The second scenario is storytelling a documentary film. Each subject watches the same documentary film and talks about it to another person in front of the cameras.

Scenario 3: Storytelling a fairy tale In the third scenario, subject is asked to read/watch a fairy tale from a text/video, and tells the story as he/she remembers.

Scenario 4: Conversation with an agent In the fourth scenario, subject is given a text from phone call between an angry client and customer support service employee who is asked to act part of the client and forced the subject to make overpowered gestures.

Scenario 5: Watch & Tell As for the fifth case, subject is expected to watch and tell thoughts about various previously prepared videos and images.

Scenario	Length	Conversation (Monolog / Dialog)	Record Count
Scenario 1	1-3 mins	Monolog	9
Scenario 2	2-4 mins	Monolog	7
Scenario 3	2-5 mins	Monolog	7
Scenario 4	2-3 mins	Dialog	11
Scenario 5	3-4 mins	Dialog	8

All the recordings in our database is kept meticulously for easiness of future processing. For each recording in our database, there exist;

- 5 videos captured from 5 different cameras
- An audio recording from the lapel microphone
- 3D marker locations data over time

- Joint angle data over time
- 3D model animation of the actor
- Comparison video of model animation with original video

We generate 3D marker location sequences by tracking the videos captured from 5 different viewpoint cameras and using multistereo correspondence information to estimate 3D marker world coordinates for each frame. These 3D marker coordinates are then converted to joint angles in MotionBuilder [25] to create more reliable and useful data. 3D human model animation is generated in MotionBuilder based on joint angle movement data.

Chapter 4

UNIMODAL CLUSTERING

4.1 Unsupervised Clustering

We extract intonational phrases through unsupervised and gesture phrases through semi-supervised temporal clustering. For the purpose of temporal clustering we employ the parallel branch HMM structure that we defined in [10]. The parallel branch HMM structure Λ has M parallel branches and N states as shown in Fig. 4.1. In the HMM structure Λ, observation probability densities are modeled by a single Gaussian with diagonal covariance. The states labeled as s_s and s_e are non emitting start and end states of the parallel HMM structure. As it can be observed from Fig. 4.1, the parallel HMM Λ is composed of M parallel left-to-right HMMs, $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$, where each λ_m is composed of N states, $\{s_{m,1}, s_{m,2}, \ldots, s_{m,N}\}$. The state transition matrix A_{λ_m} of each λ_m is associated with a sub-diagonal matrix of A_Λ. The feature stream is a sequence of feature vectors, $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_T\}$, where $\boldsymbol{f}_t$ denotes the feature vector at frame t. Unsupervised temporal segmentation using HMM structure Λ yields L number of segments $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L\}$. The l-th temporal segment is associated with the following sequence of feature vectors,

$$\varepsilon_l = \{\boldsymbol{f}_{t_l}, \boldsymbol{f}_{t_l+1}, \ldots, \boldsymbol{f}_{t_{l+1}-1}\} \quad l = 1, 2, \ldots, L \tag{4.1}$$

where $\boldsymbol{f}_{t_1}$ is the first feature vector $\boldsymbol{f}_1$ and $\boldsymbol{f}_{t_{L+1}-1}$ is the last feature vector $\boldsymbol{f}_T$.

The segmentation of the feature stream is performed using Viterbi decoding to maximize the probability of model match, which is the probability of feature sequence

$\boldsymbol{F}$ given the parallel HMM Λ,

$$\mathrm{P}(\boldsymbol{F}|\Lambda) = \max_{\varepsilon_l, m_l} \prod_{l=1}^{L} \mathrm{P}(\varepsilon_l|\lambda_{m_l}) \tag{4.2}$$

where ε_l is the l-th temporal segment, which is modeled by the m_l-th branch of the parallel HMM Λ. Since, the temporal segment ε_l from frame t_l to $(t_{l+1} - 1)$ is associated with segment label ℓ_l, where we have a label sequence $\boldsymbol{\ell} = \{\ell_1, \ell_2, \ldots, \ell_L\}$ corresponding to the temporal segments $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L\}$.

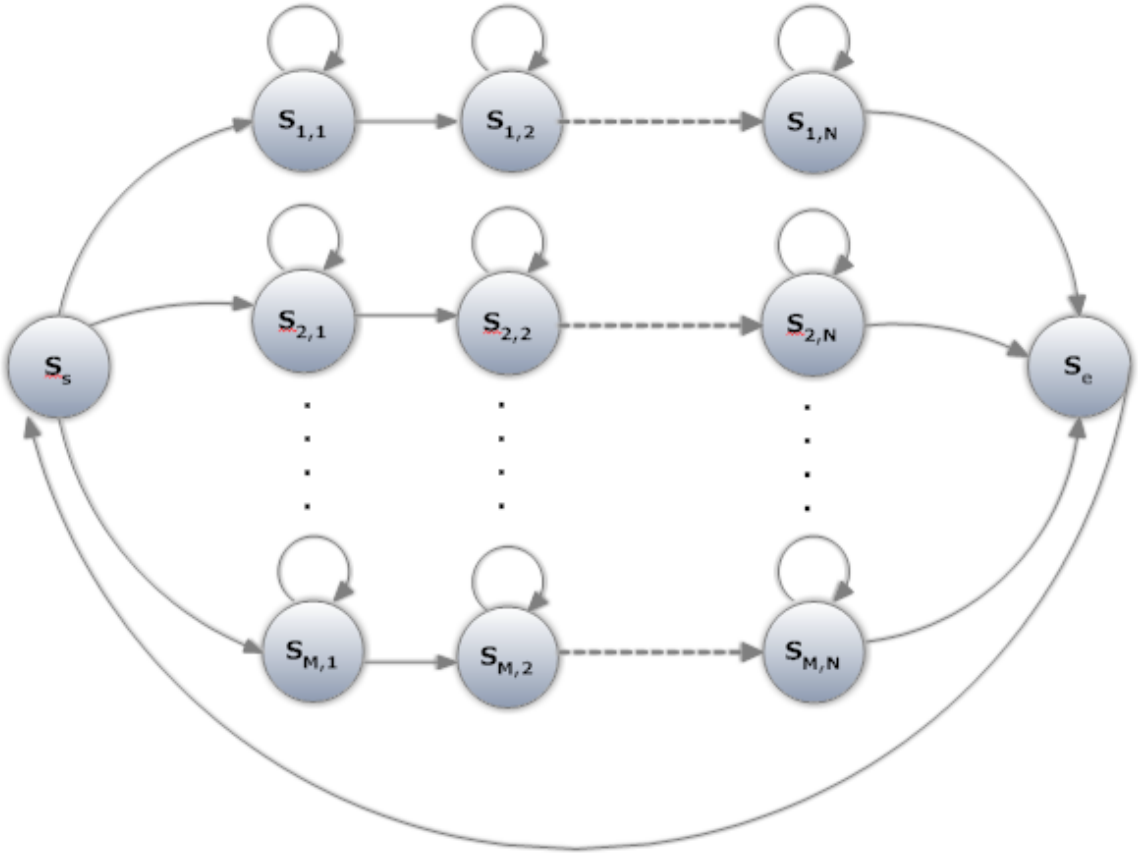


Figure 4.1: Parallel Branch HMM Structure

4.2 Unsupervised Clustering of Prosody

Prosodic voice characteristics at the acoustic level, including intonation, rhythm and intensity patterns, carry important temporal and structural synchrony with gesture phrases [41]. Acoustic features such as pitch and speech intensity can be used to model underlying intonational phrases of speech. We choose to include normalized speech intensity, normalized pitch and confidence to pitch, which is represented with pitch gain, into the prosody feature vector. We estimate the prosody feature vector for each speech frame of 25 msec duration centered on a 50 msec analysis window. Speech intensity is extracted as the logarithm of the signal energy in the analysis window,

$$I_k = \log(\frac{1}{W}\sum_{i=0}^{W} s_k[i]^2), \tag{4.3}$$

where s_k is the speech signal in the kth window, and W is the window size. The normalized speech intensity, $\bar{I}_k$, is then extracted with mean and variance normalization. Pitch is extracted using the auto-correlation method [42]. The normalized auto-correlation function for the kth speech frame can be defined as,

$$r_k(\tau) = \frac{\sum_i (s_k[i] * s_k[i-\tau])}{\sqrt{\sum_i s_k[i]}\sqrt{\sum_i s_k[i-\tau]}}. \tag{4.4}$$

Then the lag value which maximize the auto-correlation function is set as the pitch feature, $\tau_k^* = \arg\max_\tau r_k(\tau)$, and the corresponding auto-correlation value is set as the pitch gain, $r_k^* = \max_\tau r_k(\tau)$. Since pitch values differ for each speaker and the system is desired to be speaker-independent, speaker normalization is applied. For each speech segment, we compute the mean pitch value over the pitch values with pitch gain higher than 0.4. Then the mean pitch value is removed from the pitch values, which are computed for each segment, to obtain the normalized pitch $\bar{\tau}_k^*$. Then normalized intensity, normalized pitch, pitch gain and the first derivative of

these three parameters are used to define the prosody feature vector,

$$\boldsymbol{f}_k^p = [\bar{I}_k,\ \bar{\tau}_k^*,\ r_k^*,\ \Delta\bar{I}_k,\ \Delta\bar{\tau}_k^*,\ \Delta r_k^*], \tag{4.5}$$

where Δ defines the first order derivative for the corresponding features.

The prosody feature stream $\boldsymbol{F}^p = \{\boldsymbol{f}_1^p, \boldsymbol{f}_2^p, ..., \boldsymbol{f}_T^p\}$ is used to train a parallel branch HMM structure, Λ^p, which clusters the prosody feature stream and captures recurrent intonational phrases. The HMM structure Λ^p is composed of M_p parallel left-to-right HMMs, $\{\lambda_1^p, \lambda_2^p, ..., \lambda_{M_p}^p\}$, where each λ_m^p is composed of N_p states. The unsupervised clustering process defines temporal intonational segments, ε_l^p, where each segment label ℓ_l^p is assigned to one of the M_p available segment classes $\{p_1, p_2, ..., p_{M_p}\}$. Furthermore, frame level label sequence is defined for intonational phrase sequence,

$$\xi_t = \ell_l^p \quad \text{for } t = t_l, t_l + 1, \ldots, t_{l+1} - 1, \tag{4.6}$$

where ξ_t is the intonation label of the tth speech frame.

Note that optimal choice of HMM branch count M and state count N will later be investigated in Chapter 5.

4.3 Semi-Supervised Clustering of Gesture

We model upper-body gestures, specifically hand gestures, at gesture phrase level to emphasize speech intonation. In the analysis of gestures we employ joint angles as the gesture features from four body parts: left arm, left forearm, right arm, and right forearm. Note that angle values of each joint are defined in the local frame of reference on the ancestor joint (see Fig. 4.2).

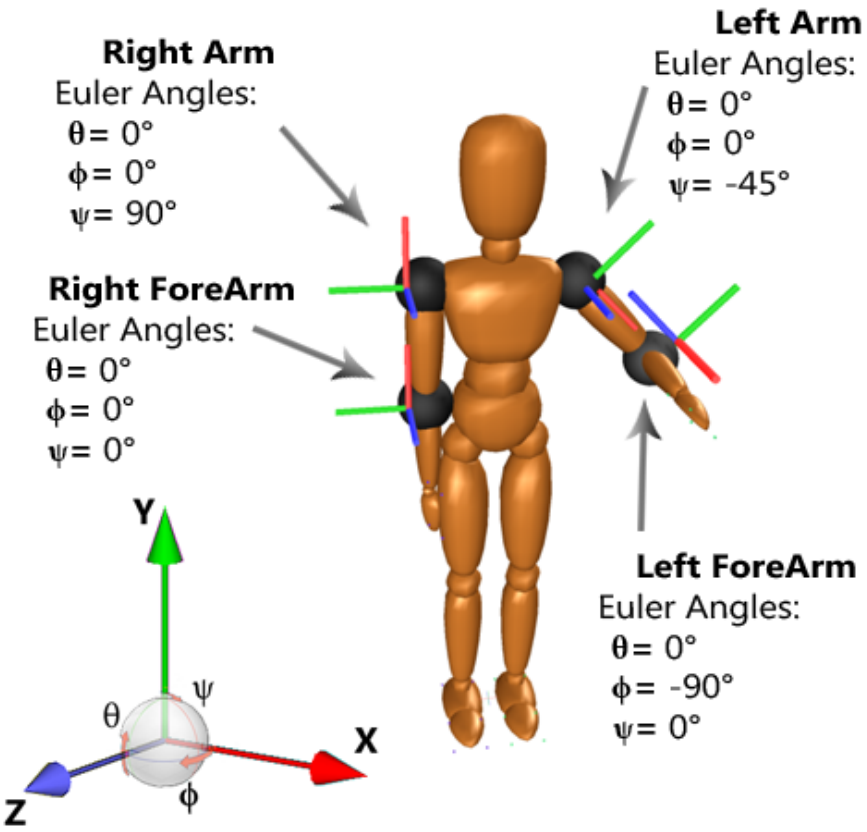


Figure 4.2: Upper body joint angles used in our system

We define the gesture feature vector for the ith joint at frame k, $\boldsymbol{f}_k^{Ji}$, to include the joint angles from the ith body part and their first order derivatives,

$$\boldsymbol{f}_k^{Ji} = [\theta_k^i, \phi_k^i, \psi_k^i, \Delta\theta_k^i, \Delta\phi_k^i, \Delta\psi_k^i], \quad \text{for } i = 1, 2, 3, 4, \tag{4.7}$$

where θ_k^i, ϕ_k^i and ψ_k^i are the Euler angles respectively in x, y and z directions, representing the posture of the ith joint at frame k, and $\Delta\theta_k^i$, $\Delta\phi_k^i$, $\Delta\psi_k^i$ denote their respective first order derivatives. The resulting gesture feature (Fig. 4.2) for the four joints at time frame k is defined as,

$$\boldsymbol{f}_k^g = [\boldsymbol{f}_k^{J1}, ..., \boldsymbol{f}_k^{J4}]. \tag{4.8}$$

We implement a semi-supervised clustering using the parallel branch HMM structure, Λ^g, over the gesture feature stream $\boldsymbol{F}^g = \{\boldsymbol{f}_1^g, \boldsymbol{f}_2^g, ..., \boldsymbol{f}_T^g\}$ to extract recurrent gesture phrases. The HMM structure Λ^g initially is set to have two parallel branch HMMs, $\{\lambda_1^g, \lambda_2^g\}$, where each λ_m^g is composed of N_g states. The number of branches increased iteratively to M_g in a semi-supervised manner using the following procedure:

(i) Initially set Λ^g to have two branches to model rest position of hands and all the other hand movements. Hand label several examples of rest event.

(ii) Perform the Baum-Welch training of the Λ^g.

(iii) Perform the Viterbi decoding to get temporal clusters.

(iv) Visually inspect clusters, correct clusters as needed. Repeat steps (ii) and (iii) until convergence.

(v) If a new gesture phrase, which is recurrent in the training data and not covered by the Λ^g, exists, go to step (vi), otherwise stop.

(vi) Hand label several examples of the new gesture phrase. Add a branch to the Λ^g for the new gesture phrase with initial training. Go to step (ii).

Eventually the semi-supervised clustering process defines gesture phrase segments, ε_l^g, where each segment label ℓ_l^g is assigned to one of the M_g available gesture phrase classes $\{g_1, g_2, ..., g_{M_g}\}$. We currently use $M_G = 7$ parallel branches (hence number of gesture phrase clusters) and $N_G = 4$ states for the parallel branch HMM structure Λ^g. Table 4.1 and figure 4.3 describes the gesture classes obtained as the result of semi-supervised clustering when performed on our MVGL-MUB database.

Label:	Description:	Details:
g_1	Simultaneous hand raise	Actor raises both hands at the same time
g_2	Rest	Stand by position of the actor
g_3	Left hand gestures	Actor raises left hand only
g_4	Perpendicular gestures	Asynchronously moves one hand in vertical and other in horizontal direction
g_5	Hands contact	Actor touches his/her hands one another)
g_6	Arms open	Actor stretches his/her arms to backward)
g_7	Right hand gestures	Actor raises right hand only

Table 4.1: Gesture Labels and Descriptions

Figure 4.3: *Gesture description figures*

Chapter 5

MULTIMODAL ANALYSIS AND SYNTHESIS OF PROSODY-DRIVEN GESTURES

5.1 Gesture Generation Model

In a natural speaking style beat gestures are articulated in synchrony with speech prosody to emphasize the underlying speech [41, 43, 44]. In this chapter we construct a multimodal analysis framework to form a relationship between beat gestures and speech prosody. Unimodal clustering and labeling of gesture and intonation phrases have been discussed in Chapter 4. In general intonation phrases last much shorter than gesture phrases in duration, and the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech as McNeill stated in [43].

A gesture phrase sequence, when accompanied by a sequence of intonation phrases, forms a random process which can be seen as a Markov process. One useful mathematical model for the multimodal analysis of gesture and intonation phrases can be constructed by taking gesture phrases as the states of a Markov chain and intonation phrases as the observations of this Markov process. Hence state transitions correspond articulation of consecutive gesture phrases, and gesture phrases are expected to localize in time with respect to the McNeill's phonological rule by observing intonation phrases. Since the relationship between gesture and intonation phrases is not strong, once you decode gesture phrase sequence given the intonation phrase observations, gesture phrases have a shortfall in modeling duration in time. Another useful mathematical model to overcome this shortfall is to introduce a state duration model so that we can better control gesture phrase durations in the decoding process. Combination of these two useful mathematical models yields hidden semi-Markov

model (HSMM) framework [15] for the multimodal analysis of gesture and intonation phrases. HSMM is an extension of HMM which allows the underlying process to be a semi-Markov chain with states having variable durations. This is to say that, the underlying process is Markovian at certain jump instants [45]. Fig. 5.1 shows how such an HSMM structure works. The design implies that speech and body gesture features should be clustered prior to the multimodal analysis. The temporal clustering scheme, described in Chapter 4 is employed for this purpose.

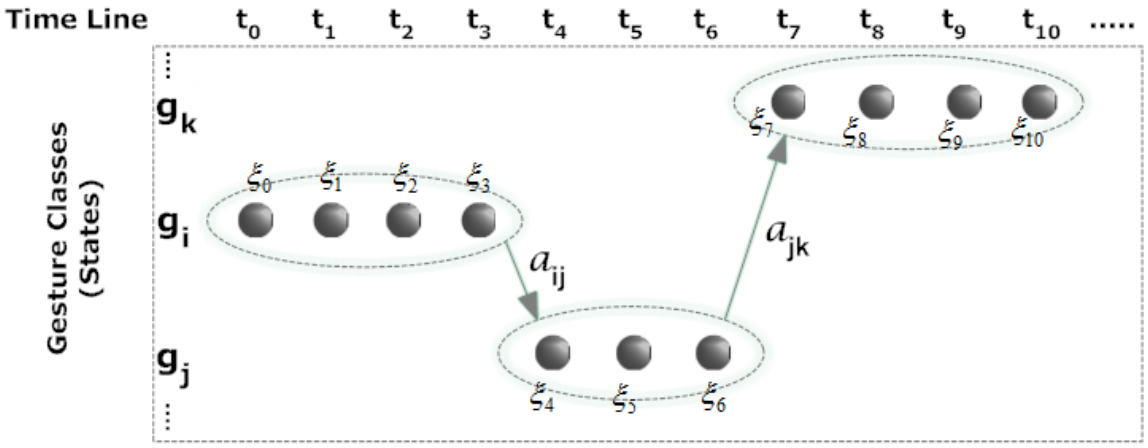


Figure 5.1: *In a Hidden Semi-Markov Process each state has a duration and emits a number of observations.*

An HSMM representing intonation phrases as observations with M_g fully connected states is modeled as $\Lambda^{gp} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{\Pi})$. The states of Λ^{gp} represent gesture phrase classes, and the model parameters $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{D}$, $\boldsymbol{\Pi}$ respectively are state transition probability, observation emission distribution, state duration distribution, and initial state distribution matrices.

The $M_g \times M_g$ state transition matrix $\boldsymbol{A}$ is defined by entries a_{ij} representing the state transition probability from gesture g_i to g_j,

$$\boldsymbol{A} : \{a_{ij} = P(\ell_l^g = g_j | \ell_{l-1}^g = g_i)\} \quad i, j = 1, ..., M_g, \tag{5.1}$$

where ℓ_l^g represents the lth gesture in the sequence of gesture phrases. The observation

emission distribution $\boldsymbol{B}$ is modeled by discrete probability mass functions for each gesture g_i,

$$\boldsymbol{B} : \{b_i(p_k) = P(p_k|\ell_l^g = g_i)\} \quad k = 1, ..., M_p,\ i = 1, ..., M_g, \tag{5.2}$$

where $b_i(p_k)$ is the probability of observing intonation phrase p_k at gesture g_i. The state duration distribution $\boldsymbol{D}$ is formed as state dependent duration probability mass functions,

$$\boldsymbol{D} : \{d_i(k)\} \quad i = 1, ..., M_g,\ k = 1, ..., \frac{D_{max}}{\delta}, \tag{5.3}$$

where $d_i(k)$ is the probability of gesture g_i lasting $k\delta$ sec, D_{max} is the maximum duration among all gestures, and δ is the histogram bin size for the underlying probability mass function. We take the maximum duration as $D_{max} = 10$ sec, and the histogram bin size as the speech frame duration, $\delta = 25$ msec. The initial state probability vector $\boldsymbol{\Pi}$ is defined by entries π_i representing the probability of starting with gesture g_i as the first gesture phrase,

$$\boldsymbol{\Pi} : \{\pi_i = P(\ell_1^g = g_i)\} \quad i = 1, ..., M_g. \tag{5.4}$$

The Λ^{gp} model is extracted by estimating the statistical parameters of the model over a training corpus. Statistical parameter estimations are given as:

$$\pi_i = P(\ell_1^g = g_i) \mathrel{\widehat{=}} \frac{C(1, i, j)}{\sum_{j'} C(1, i, j')}, \tag{5.5}$$

$$a_{ij} = P(\ell_l^g = g_j|\ell_{l-1}^g = g_i) \mathrel{\widehat{=}} \frac{\sum_l C(l, i, j)}{\sum_l \sum_{j'} C(l, i, j')}, \tag{5.6}$$

$$b_i(p_k) = P(p_k|\ell_l^g = g_i) \mathrel{\widehat{=}} \frac{O(i, k)}{\sum_{k'} O(i, k')}, \tag{5.7}$$

$$d_i(k) \mathrel{\widehat{=}} \frac{G(i, k\delta \leq \tau < (k+1)\delta)}{\sum_{k'} G(i, k'\delta \leq \tau < (k'+1)\delta)}, \tag{5.8}$$

where $C(l, i, j)$ is the number of times g_i is the lth gesture and g_j is the $(l+1)$st

gesture, $O(i,k)$ is the number of frame count of intonation phrase p_k at gesture g_i, and $G(i, k\delta \leq \tau < (k+1)\delta)$ is the number of occurrences of gesture g_i with duration τ in $[k\delta,\ (k+1)\delta)$ interval.

5.2 Gesture Synthesis

Gesture synthesis is defined as decoding an optimal state sequence, $\hat{\boldsymbol{\ell}}^g$, over the HSMM Λ^{gp} given a sequence of frame level intonational phrase labels, $\{\xi_1, \xi_2, \ldots, \xi_T\}$. Note that the decoded optimal state sequence delivers synthesized sequence of gesture phrases and their durations, and the HSMM framework secures to have realistic gesture phrase durations. In HMM framework, where the underlying process is Markov, given an observation sequence, the Viterbi algorithm is employed to decode the most likely state sequence. In HSMM framework however, states have variable durations and a sequence of observations are emitted at a single state. This requires us to define a forward likelihood function, which incorporates state duration model,

$$\psi_t(j) = \max_{\tau} \max_{i} \{\psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(\xi_k))\}, \tag{5.9}$$

where $\psi_t(j)$ is the accumulated logarithmic likelihood at time frame t in state g_j after observing intonational phrase labels $\{\xi_1, \xi_2, \ldots, \xi_t\}$. Based on the forward likelihood function $\psi_t(j)$, we define the following modified Viterbi decoding algorithm (Alg. 1) to extract the optimal state sequence, that is the optimal gesture phrase sequence $\{\hat{\ell}_1^g, \ldots, \hat{\ell}_L^g\}$, and the associated gesture phrase durations $\{\kappa_1, \ldots, \kappa_L\}$.

Algorithm 1 The modified Viterbi decoding algorithm for gesture synthesis

Require: : Λ^{gp} and $\{\xi_1, \xi_2, \ldots, \xi_T\}$

$\psi_1(i) = \log(\pi_i b_i(\xi_1)) \quad i = 1, 2, \ldots, M_g$

for $t = 2$ to T **do**

for $j = 1$ to M_g **do**

$T' = \min(D_{max}, t)/\delta$

$\psi_t(j) = \max_{\tau\in[1,T']} \max_{i\in[1,M_g]}\{\psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(\xi_k))\}$

$\varphi_t(j) = \arg\max_{i\in[1,M_g]} \max_{\tau\in[1,T']}\{\psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(\xi_k))\}$

$\nu_t(j) = \arg\max_{\tau\in[1,T']} \max_{i\in[1,M_g]}\{\psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(\xi_k))\}$

end for

end for

$\hat{\ell}_L^g = \arg\max_j \psi_T(j)$

$\kappa_L = \nu_T(\hat{\ell}_L^g)$

$l = L - 1; \quad t = T$

while $t > 0$ **do**

$\hat{\ell}_l^g = \varphi_t(\hat{\ell}_{l+1}^g)$

$\kappa_l = \nu_{t-\kappa_{l+1}}(\hat{\ell}_l^g)$

$t = t - \kappa_{l+1}; \quad l = l - 1$

end while

Chapter 6

GESTURE ANIMATION

Animation of the synthesized gesture sequence has been performed over three main tasks: extraction of gesture motion sequence with unit selection, smoothing gesture-to-gesture transitions, and finally animation of the gesture motion sequence. The first task is to generate a synthesized sequence of gesture motion segments, $\hat{\boldsymbol{\varepsilon}}^g$, given the synthesized gesture phrase $\hat{\boldsymbol{\ell}}^g$ and duration $\boldsymbol{\kappa}$ sequences. Then the next task is to smooth joint angle discontinuities over a temporal window at gesture segment boundaries, that is at the boundary of two consecutive synthesized gesture segments $\hat{\varepsilon}_l^g$ and $\hat{\varepsilon}_{l+1}^g$, to extract a smoothed gesture motion segment sequence $\tilde{\boldsymbol{\varepsilon}}^g$. Then the smoothed gesture motion sequence $\tilde{\boldsymbol{\varepsilon}}^g$ is used to animate upper-body gestures of a virtual character.

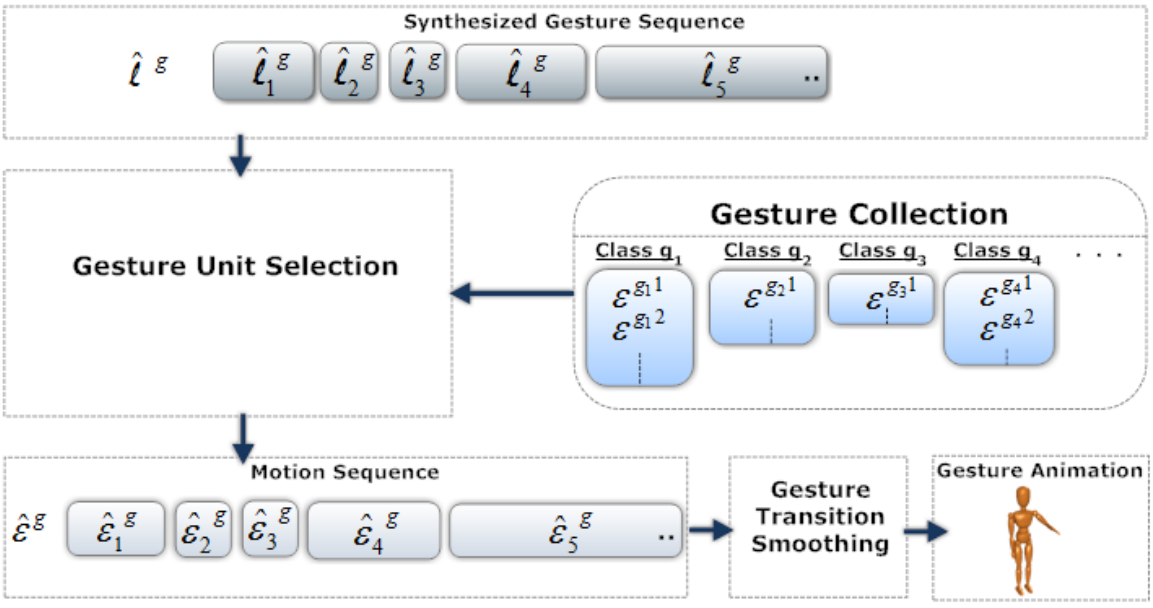


Figure 6.1: Gesture Animation Generation System

We employ a unit selection algorithm to generate the synthesized sequence of gesture motion segments $\hat{\boldsymbol{\varepsilon}}^g$. In order to employ unit selection, a collection of representative temporal segment templates for each gesture g_i is constructed as $G_i = \{\varepsilon^{g_i 1}, \varepsilon^{g_i 2}, \ldots, \varepsilon^{g_i K_i}\}$, where K_i is the number of templates in the collection of gesture g_i. We target to minimize a mixture of duration and joint angle continuity penalty scores during the unit selection. The duration penalty and joint angle continuity scores of a gesture segment template $\varepsilon^{g_i k}$ for a synthesized gesture phrase $\hat{\ell}_l^g$ are respectively defined as,

$$\upsilon_\kappa(\varepsilon^{g_i k}|\hat{\ell}_l^g = g_i) = ||\kappa_l - \kappa(\varepsilon^{g_i k})||, \quad \text{and} \tag{6.1}$$

$$\upsilon_\omega(\varepsilon^{g_i k}|\hat{\ell}_l^g = g_i) = ||\omega_e(\hat{\varepsilon}_{l-1}^g) - \omega_b(\varepsilon^{g_i k})||, \tag{6.2}$$

where κ_l is the duration of the synthesized gesture phrase $\hat{\ell}_l^g$, $\kappa(\varepsilon^{g_i k})$ is the duration of the gesture segment template $\varepsilon^{g_i k}$, $\omega_e(\hat{\varepsilon}_{l-1}^g)$ is the ending joint angle vector of the synthesized gesture segment $\hat{\varepsilon}_{l-1}^g$, and $\omega_b(\varepsilon^{g_i k})$ is the beginning joint angle vector of the gesture segment template $\varepsilon^{g_i k}$. Then the overall penalty score to be minimized in the unit selection is set as,

$$\upsilon(\varepsilon^{g_i k}|\hat{\ell}_l^g = g_i) = \alpha\upsilon_\omega(\varepsilon^{g_i k}|\hat{\ell}_l^g = g_i) + (1-\alpha)\upsilon_\kappa(\varepsilon^{g_i k}|\hat{\ell}_l^g = g_i), \tag{6.3}$$

where α is the mixture weight of the joint angle penalty score.

Unit selection based gesture motion sequence extraction is configured as finding an optimal path on a lattice of temporal gesture templates by minimizing the accumulated penalty score. The optimal path can be extracted by the following Viterbi algorithm:

i. Initialization

   $V_1(k) = \upsilon_\kappa(\varepsilon^{g_i k}|\hat{\ell}_1^g = g_i)$, where $k = 1, 2, ..., K_i$

ii. Recursion: Repeat for $l = 2, 3, ..., L$,

   $V_l(k) = \min_{j=1,\ldots,K_i}\{V_{l-1}(j) + \upsilon(\varepsilon^{g_i k}|\hat{\ell}_l^g = g_i\}$,

$$Q_l(k) = \arg\min_j\{V_{l-1}(j) + \upsilon(\varepsilon^{g_i k}|\hat{\ell}_l^g = g_i\}$$

iii. Backtrace the optimal path

$q_L = \arg\min_k\{V_L(k)\}$,

$q_l = Q_{l+1}(q_{l+1})$ for $l = L-1, L-2, ..., 1$,

iv. Construct the synthesized sequence of gesture motion segments

$\hat{\varepsilon}_l^g = \varepsilon^{\hat{\ell}_l^g q_l}$ for $l = 1, 2, \ldots, L$.

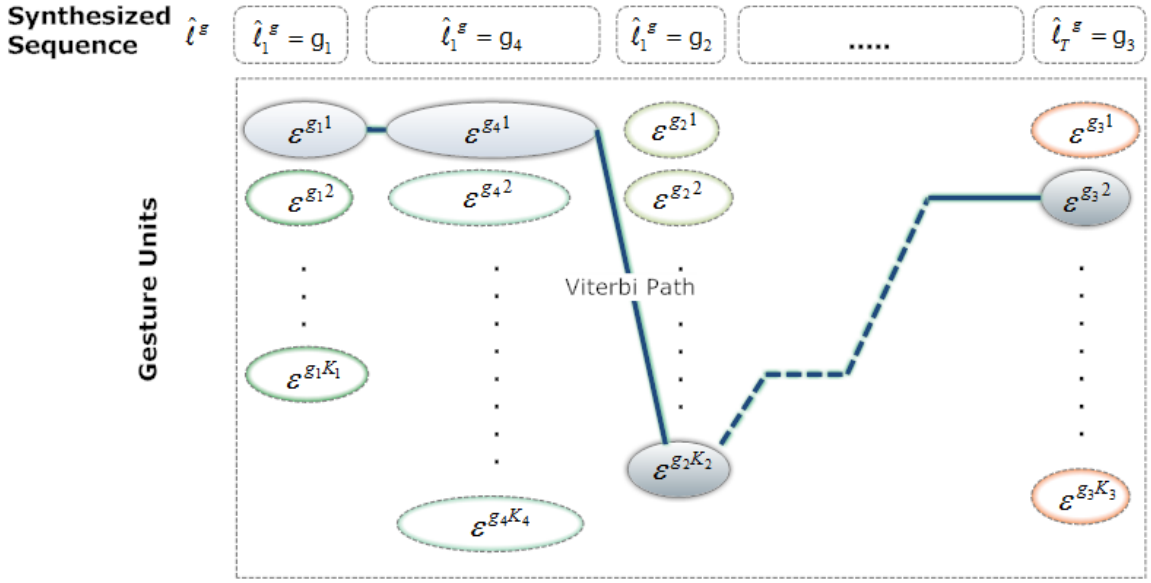


Figure 6.2: Viterbi Path Selection (Note that not all gestures have same duration. For each desired gesture class, there different number of gesture choices in the pool)

As the second task of the gesture animation, an exponential smoothing function is applied on the synthesized gesture motion sequence $\hat{\boldsymbol{\varepsilon}}^g$, and the smoothed gesture motion sequence $\tilde{\boldsymbol{\varepsilon}}^g$ is extracted. The smoothing window size is adjusted to generate passivated gestures or energetic gestures by respectively increasing or decreasing smoothing window size. Smoothing window size ($\zeta_s = 5$) is decided based on length of gestures in the pool. Besides, different smoothing schemes are applied and consequences on motion sequence are observed.

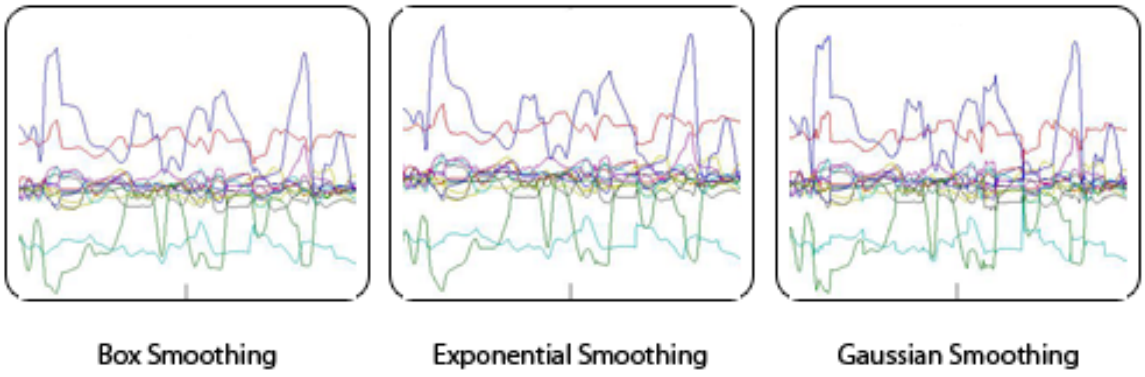


Figure 6.3: Smoothing Filter trials on Joint Angle Data

For more realistic skeleton animation, motion of spine and lower body joints are also transferred from the original skeleton animation. Then resulting motion sequence is animated using the *MotionBuilder 3D Character Animation Software* [25].

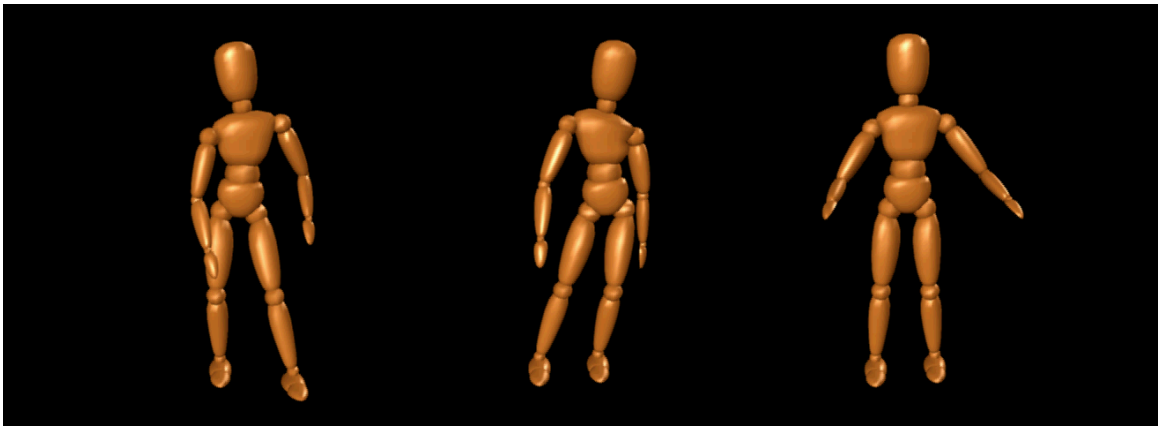


Figure 6.4: Screenshots from synthesized body animation

Chapter 7

EXPERIMENTAL EVALUATIONS

A database of various records belonging to two different actors from the MVGL-MUB corpus is considered for evaluations. In order to synthesize a record in the dataset, the HSMM machine is trained on the dataset from which the record considered for synthesis is excluded. Following training, the intonational phrases of the test record is produced and fed to the HSMM machine where the modified Viterbi algorithm is used to generate a synthesized gesture sequence. This is repeated for all the records in the database in a leave-one-out fashion. Such an experiment where all the records of the data collection for a given actor are synthesized in the manner described above will be referred to as a synthesis session.

7.1 Objective Evaluations

It is expected that the generated synthesis sequence follows the duration model of the actual gesture sequence. Also, the gesture boundaries in synthesized gesture sequence are expected to be close to those of the actual gesture sequence. Hence, to determine the quality of the generated synthesis sequence two performance criteria are considered. The first criteria measures the similarities between the probabilistic duration model histograms of the synthesized and actual gesture sequences according to the following equation:

$$S_i = \frac{\sum_{k=1}^{D_{max}} P(i,k)\hat{P}(i,k)}{\sqrt{\sum_{k=1}^{D_{max}} P(i,k)^2}\sqrt{\sum_{k=1}^{D_{max}} \hat{P}(i,k)^2}} \tag{7.1}$$

Here, $P(i,k)$ and $\hat{P}(i,k)$ are the probabilities that gesture type i has a duration equal to k for actual and synthesized gesture sequences respectively. Also, $0 \leq S_i \leq 1$ and

a perfect match between synthesized and actual gesture sequences in terms of the duration model yields $S_i = 1$ for each gesture type. The value of S_i decreases as the duration model resemblance of gesture type i in the synthesized gesture sequence to actual gesture sequence diminishes.

As for the second performance criteria, it computes the boundary distance of the gestures in the synthesized sequence to those in the actual gesture sequence (Figure 7.1). That is, for each gesture in the synthesized sequence, the time distance to its closest gesture boundary in the actual sequence is measured. Note that, the distance between two boundary positions are measured regardless of the type of the gesture in any of the two sequences. The sum of these distances is then normalized by the count of the gestures along the synthesized sequence. This criterion demonstrates how well gesture boundaries are preserved while generating the synthesized gesture sequence. The boundary distance between the two sequences is computed according to the following equation:

$$D = \frac{1}{L} \sum_{i} |b_i - \hat{b}_i| \tag{7.2}$$

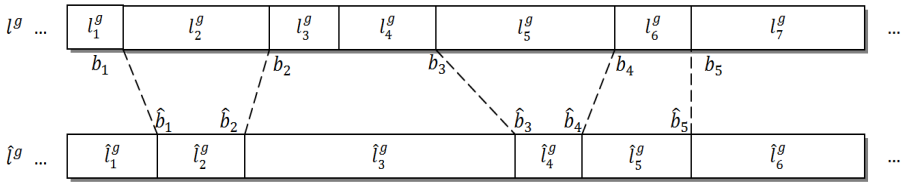


Figure 7.1: *The boundary distance between actual (g) and synthesized ($\hat{g}$) gestures.*

Here, L is the number of gestures in the synthesized sequence while b_i and $\hat{b}_i$ are the boundary times of the actual and synthesized gestures respectively. To evaluate the synthesis performance of a session, duration model similarity and boundary distance measures are considered. A synthesis session on a database of R records containing M distinguishable gesture types results in $R \times M$ duration model similarity values. Thus, the duration model similarity of a synthesis session is given by the following

equation:

$$\mathbf{S} = \frac{1}{M \times R} \sum_{i=1}^{M} \sum_{r=1}^{R} S_{ir} \tag{7.3}$$

Here, S_{ir} is the histogram similarity measure (Eq. 7.1) corresponding to gesture type i in record r. Since the chosen data collection has 4 records for each actor and the number of available gestures acquired is equal to 7 (excluding the start pose) then $R = 4$ and $M = 7$ in our experiments.

To measure the boundary distance for an entire synthesis session, the boundary distance between the synthesized and actual gesture sequences of each record in the data collection is computed. This results in R boundary distance values. Consequently, the boundary distance for the whole session can be computed by Eq. 7.4 which is a weighted sum of the distances acquired for each record, where the duration of each record is used as the weight:

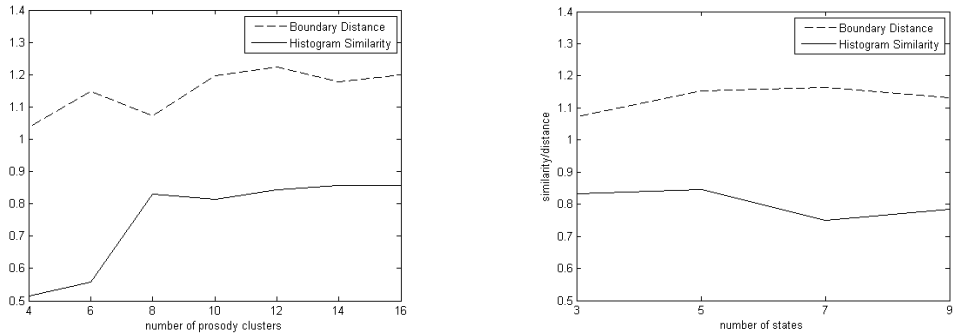
$$\mathbf{D} = \frac{\sum_{r=1}^{R} \tau_r D_r}{\sum_{r=1}^{R} \tau_r} \tag{7.4}$$

Here, D_r is the boundary distance (Eq.7.2) for record r and τ_r is the duration of the record r in seconds. In our experiments, the bin length of all histograms from which the duration model is computed (Eq. 5.8) is set to be 0.025 sec, and the maximum gesture duration (D_{max}) is set to 10 sec.

Since prosody patterns are extracted by unsupervised clustering, it is crucial to determine the number of states used in the HMM machine which clusters prosody stream. Utilizing small number of states favors long prosodic patterns while a larger number of states results in smaller prosody patterns. Also, the number of resulting prosody clusters should be determined prior to synthesis. An over-segmented prosody stream produces redundant and similar prosody clusters while under-segmentation combines two or more clusters into a single segment. In such cases lack of generalization or loss of data affects the synthesis process and results in a poor performance.

As for the number of states, one can argue that using 3 states is the most suitable

choice since it simplifies the whole process. This is our assumption in the synthesis phase and further experimental analysis proofs its validity. To overcome the ambiguities related to the optimum number of prosody clusters, an experiment has been conducted in which the unsupervised prosody clustering unit of the framework is adjusted to segment the prosody stream of the records in the data collection into $4, 6, 8, 10, 12, 14$ and 16 clusters. For each number of prosody clusters a synthesis session is performed and the session performance is evaluated resulting in **S** and **D** values (equations 7.3 and 7.4 respectively) for each synthesis session. The optimum number of clusters is expected to maximize the duration model similarity and minimize the boundary distance. Figure 7.2(a) gives the resulting plot of the experiment conducted on the video records belonging to a single actor.



(a) (b)

Figure 7.2: *A comparison of histogram similarity and boundary distance between various numbers of clusters (a) and states (b) for prosody patterns*

It is clear from the plot that when speech is segmented into 8 clusters, histogram similarity between the synthesized and real gesture is maximized (0.8313) while the boundary distance is almost minimized (1.0714). The boundary distance for the case when prosody patterns are segmented to 4 clusters are slightly less than that of 8 cluster prosody patterns. However, the histogram similarity for the 4 cluster prosody patterns is the minimum similarity value among all other test cases. This yields that clustering speech stream into 8 segments results in the best synthesis performance.

A comparative analysis, similar to the analysis made for various number of prosody patterns, is performed for various number of states of the HMM machine which segments the prosody patterns. That is, the number of prosody clusters is fixed and set to the optimum value (8) found in the previous experiment and intonational phrases are produced using a HMM machine with $3, 5, 7$ and 9 number of states. For each number of HMM states a synthesis session is performed. Subsequently the synthesis performance for each session is evaluated resulting in **S** and **D** values for each synthesis session. The optimum number of states which results in minimum boundary distance (**D**) as well as maximum duration model similarity (**S**) is then chosen as the optimum number of states to segment prosody patterns. The result is shown in Figure 7.2(b).

The histogram similarity measure is more or less the same when 3 or 5 states are considered. However, the boundary distance is clearly less when 3 states are considered while clustering prosodies. This outcome verifies our assumption that employing a 3 state HMM machine for clustering speech into prosody patterns is an optimum choice, since it not only simplifies the resulting model but also it provides the framework with a synthesized gesture sequence which is more similar to the actual one in terms of duration model and the alignment of gesture boundaries.

Choosing the resulting optimum number of states and clusters, a gesture sequence is generated for each record. The resulting histograms for one of the records of scenario 2 (telling fairy tales) and their corresponding actual gesture sequences can be seen in Figure 7.3, and respectively for a different actor in Figure 7.4. It is obvious that the generated gesture sequence follows the duration model of the training set. Note that, in each sub-plot of Figures 7.3 and 7.4, the upper histogram represents the duration model of the synthesized gesture sequence and the lower one represents the actual gesture sequence. Finally in Figure 7.5, we plot the occurrence frequencies of specific gesture types for the two actors.

7.2 Subjective Evaluations

Subjective A-B comparisons are performed using the speech driven body gesture animations to analyze opinions on how natural and realistic body gestures are synthesized. The participants are asked to evaluate the naturalness of the body gesture animations for an A-B test pair on a scale of (-2, -1, 0, 1, 2), where the scale corresponds to (A much better, A better, no preference, B better, B much better).

The whole test database consists of meaningful 20 segments, where each segment is approximately 15 seconds. For each segment, there exists original, synthesized and random body gesture animations. Random body gesture animation is generated by selecting random gestures from the gesture pool and appending them in sequence (see Chapter 6). The overall evaluation comprises 12 steps of video-pair comparison. The first two steps constitute the out-of-evaluation training phase which is designed for getting familiar with the test system. The remaining steps are formed as three (Original vs. Synthesized), three (Original vs. Random), three (Synthesized vs. Random) and one pair of identical animations comparisons in a random sequence. The test segments for the 10 steps are selected randomly out of 20 segment choices without any repetition.

The subjective tests are performed over 35 subjects via our online A-B test system. The average preference scores for the three comparison types are presented in Table 7.1. Samples of the audio-visual sequences for the speech-driven body gesture animations are available online http://mvgl.ku.edu.tr. We observe that the synthesized animations are favored over the random ones while the originals are favored over the synthesized as expected.

A-B Pair	**Preference Score**
Original versus Synthesis	-0,43
Original versus Random	-1,06
Synthesis versus Random	-0,54

Table 7.1: The Subjective A-B Comparison Results

According to the subjective test results, participants favored *Original* over *Synthesized* animation 65.7% of the time which implies a p-value less than 0.05 where p-value quantifies the strength of the evidence in favor of hypothesis. Moreover, *Original* animation is preferred over *Random* one 88.5% of the time which is significantly above the chance, p-value, less than 0.001. Finally, *Synthesized* animation was favored over *Random* animation 71.4% of the time which is also majorly above the p-value less than 0.01.

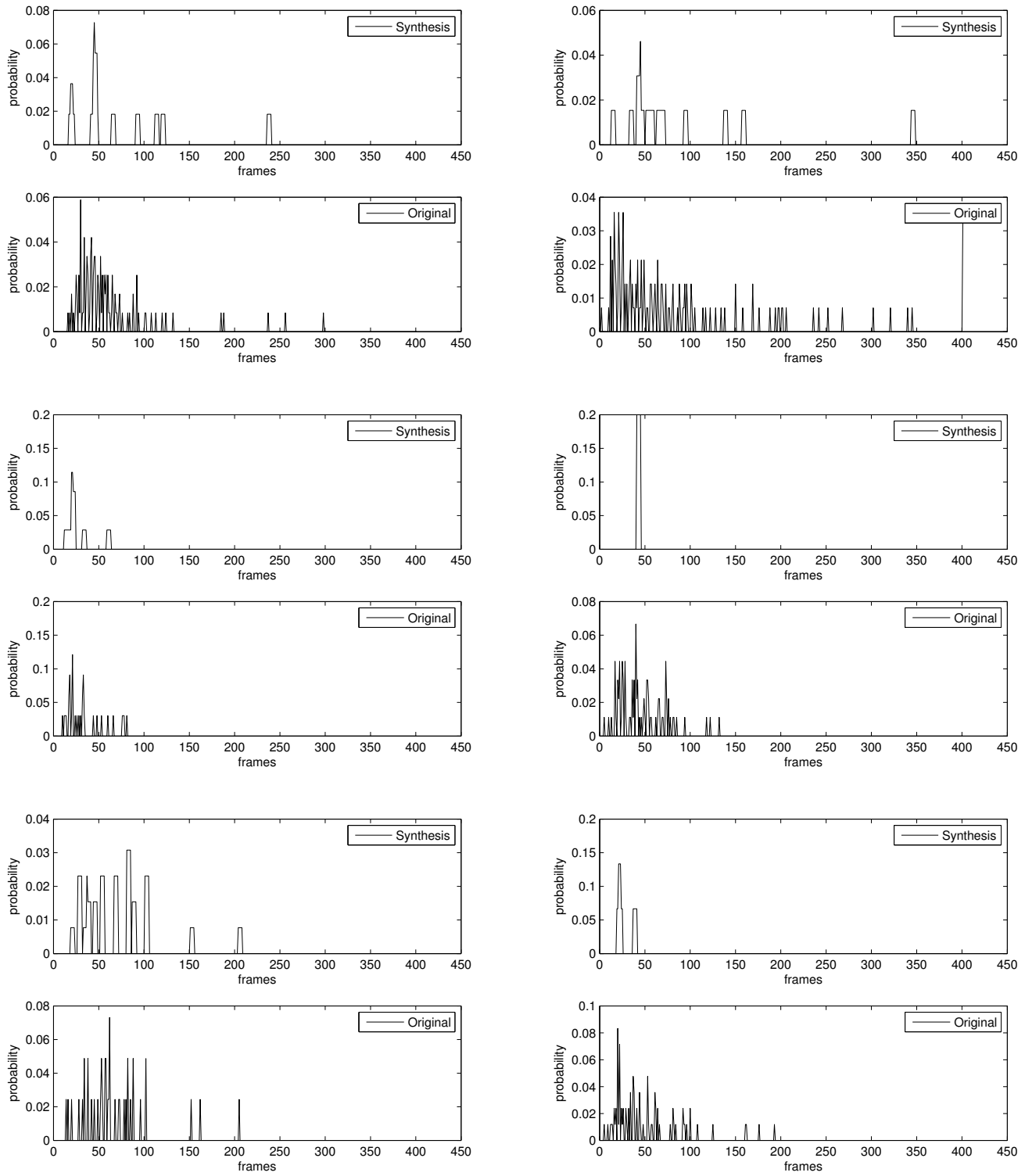


Figure 7.3: *Duration model comparison of synthesized and original gesture sequences (Actor #1)*

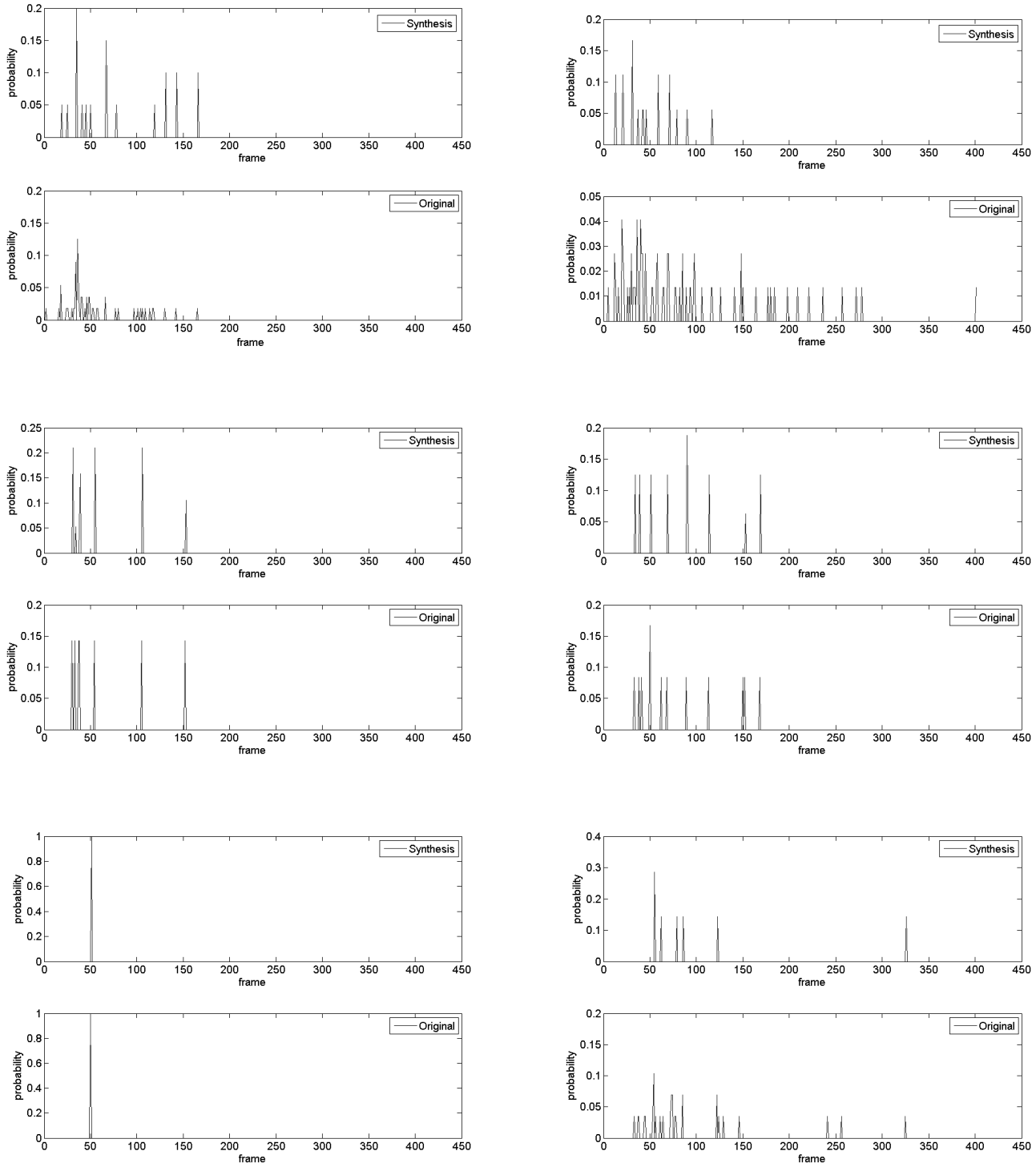
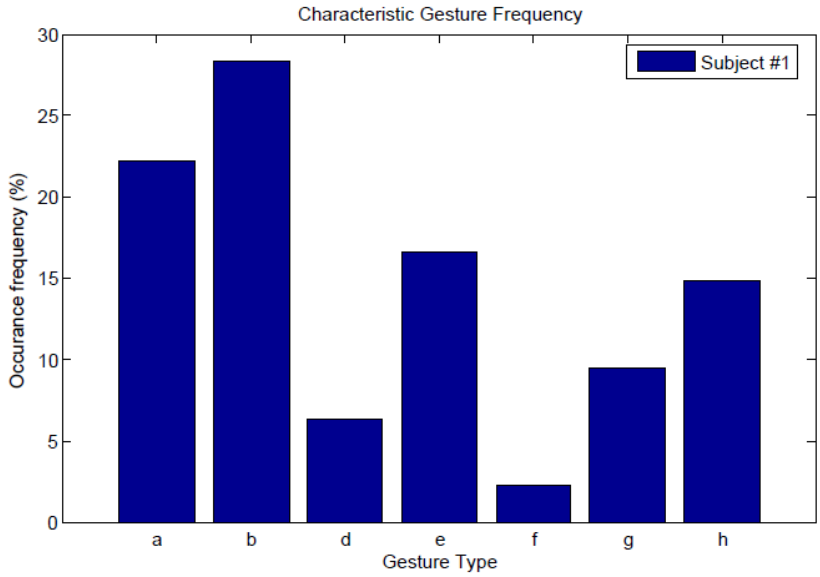
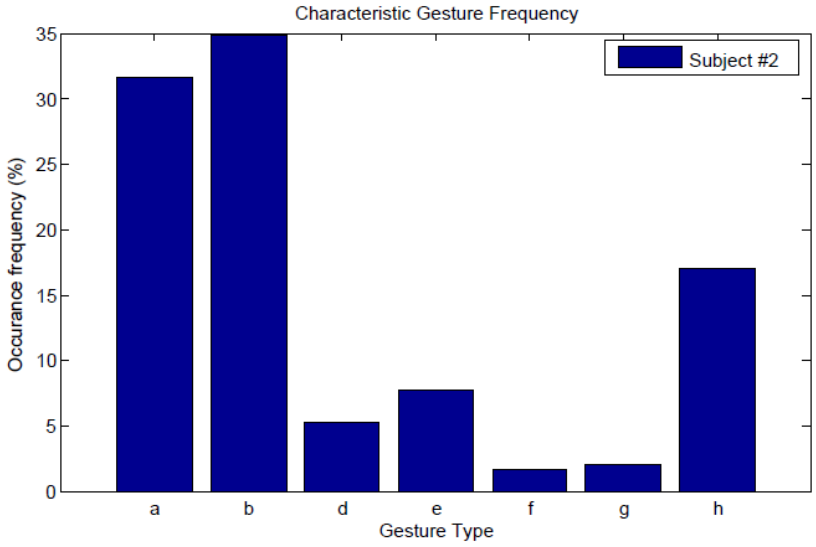


Figure 7.4: Duration model comparison of synthesized and original gesture sequences (Actor #2)



(a)



(b)

Figure 7.5: *Characteristic Gesture Frequency Comparison*

Chapter 8

CONCLUSIONS

In this thesis we have presented a new computational model for natural and plausible upper-body gesture synthesis in synchrony with speech using statistical learning techniques over multimodal gesticulation. Especially, we focused on finding mappings between speech and gesture patterns, which are modeled by taking into account lag between modalities and the duration of gestures. The main contributions of this thesis can be summarized as follows:

- An automated overall system is developed that observes and learns characteristic gestures then synthesizes natural and realistic body gestures which are driven by speech.

- We have modeled the mapping between speech and upper body gestures along with gesture durations. Both the correlation of audio-visual modalities and the correlation of successive gestures are evaluated in the gesture synthesis system. Furthermore, characteristic gesture duration models are generated for synthesizing realistic gestures.

- We developed a gesture animation system which generates smooth and realistic gesture animations by making use of duration and joint angle based gesture selection metric.

- An objective evaluation scheme is constituted for the audio-visual modality synthesis systems

- A low-cost, high performance motion capture software has been developed for

marker tracking and human modeling. In addition, a video supervision tool has been implemented for monitoring, segmenting and grouping gesture patterns.

Since the beginning, all of our studies about multimodal analysis and synthesis systems pointed out that creating a multimodal corpus is crucial, that determines the performance and quality of the overall system. Therefore we put significant effort to generate a useful audio-visual database and meticulously formed our recording scenarios to capture natural behaviors. In addition, gesture segmentation and labeling should be done meticulously because temporal gesture clusters significantly affect gesture animation quality (see Section 4.3).

Although we tested our system on recordings of particular scenarios belonging to two different actors, we strongly believe that proposed system can be applied to any person with challenging scenarios such as dancing with music. We also believe that the proposed framework can be easily adapted to other multimodal applications such as speech driven facial expression synthesis. Moreover, our system is currently able to generate multiple character profiles and synthesize body gestures of a person when driven by another person's speech.

The experimental results show that the proposed framework is successful at synthesizing natural and realistic body gestures which can be used in several application areas such as:

- Communication applications, specifically in visual teleconferencing where the receiver side can generate visual body animation based on incoming speech data. Thus transmitting only the speech data will significantly decrease bandwidth usage of common visual teleconferencing applications.

- Movies and video games which requires realistic body animations synchronized with speech. Motion capture systems are used to generate 3D realistic body animations in both movies and video games. In movies, using mocap recordings for dozens of scenes is a cumbersome process which can be facilitated by learning gestures of the actor and synthesizing it in future scenes. Personalized 3D body

gesture animations can be used especially in online role-playing games such as Second Life or World of Warcraft thus they will have more reality.

- Visual media where anchors or video jockeys can be replaced with 3D human models with realistic gestures.

As future research for speech driven gesture synthesis, a third modality, facial expressions, can be added to examine the correlation triangle between the three modalities so as to generate speech driven body and face animations.

BIBLIOGRAPHY

[1] J. Cassell, "Embodied conversational agents." Cambridge, MA, USA: MIT Press, 2000, ch. Nudge nudge wink wink: elements of face-to-face conversation for embodied conversational agents, pp. 1–27.

[2] N. Reithinger, P. Gebhard, M. Löckelt, A. Ndiaye, N. Pfleger, and M. Klesen, "Virtualhuman:dialogic and affective interaction with virtual characters," in *Proceedings of the International Conference on Multimodal Interfaces (ICMI'06*, 2006, pp. 2–4.

[3] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel, "Gesture modeling and animation based on a probabilistic re-creation of speaker style," *ACM Trans. Graph.*, vol. 27, no. 1, pp. 5:1–5:24, Mar. 2008.

[4] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," in *ACM SIGGRAPH 2010 papers*, ser. SIGGRAPH '10. New York, NY, USA: ACM, 2010, pp. 124:1–124:11.

[5] A. Heloir and M. Neff, "Exploiting motion capture for virtual human animation: Data collection and annotation visualization," in *In Proc. of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, 2010.

[6] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[7] P. Hong, Z. Wen, and T. S. Huang, "Real-time speech-driven face animation with expressions using neural networks," *IEEE Trans. Neural Networks*, vol. 13, pp. 916–927, 2002.

[8] T. Chen and R. Rao, "Audio-visual integration in multimodal communication," *Proc. of the IEEE*, vol. 86, no. 5, pp. 837 –852, 1998.

[9] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1075 –1086, March 2007.

[10] M. Sargin, Y. Yemez, E. Erzin, and A. Tekalp, "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation," *IEEE Trans. on Pattern Analysis and Machine Intelligence,*, vol. 30, no. 8, pp. 1330 –1345, 2008.

[11] S. Mariooryad and C. Busso, "Generating human-like behaviors using joint, speech-driven models for conversational agents," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2329 –2340, Oct. 2012.

[12] J. Ferguson, "Variable duration models for speech," in *Symp. Application of Hidden Markov Models to Text and Speech*, 1980, pp. 143–179.

[13] S. Levinson, "Continuously variable duration hidden markov models for automatic speech recognition," *Computer Speech & Language*, vol. 1, no. 1, pp. 29 – 45, 1986.

[14] M. Ostendorf, V. Digalakis, and O. A. Kimball, "From hmms to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 360–378, 1995.

[15] S.-Z. Yu, "Hidden semi-Markov models," *Artif. Intell.*, vol. 174, no. 2, pp. 215–243, 2010.

[16] S. Özkul, S. Asta, Y. Yemez, and E. Erzin, "HSMM based Multimodal Gesture Synthesis driven by Speech," *IEEE Trans. on Multimedia (To be submitted)*, 2012.

[17] H. K. Christian Schnauer, Thomas Pintaric, "Full body motion capture a flexible marker-based solution," in *Joint Virtual Reality Conference (JVRC 2011), Nottingham, UK*, 2011.

[18] T. D. A. P. P. Christopher Richard Wren, Ali Azarbayejani, "Pfinder: Real-time tracking of the human body," in *IEEE Trans. on Pattern Anal. Mach. Intell.*, 2000, pp. 797–808.

[19] A. Young, "Use of body model constraints to improve accuracy of inertial motion capture," in *Body Sensor Networks (BSN), 2010 International Conference on*, june 2010, pp. 180 –186.

[20] W. K. Lee and S. Jung, "Fpga design for controlling humanoid robot arms by exoskeleton motion capture system," in *Robotics and Biomimetics, 2006. ROBIO '06. IEEE International Conference on*, december 2006, pp. 1378 –1383.

[21] S. Yabukami, H. Kikuchi, M. Yamaguchi, K. Arai, K. Takahashi, A. Itagaki, and N. Wako, "Motion capture system of magnetic markers using three-axial magnetic field sensor," *Magnetics, IEEE Transactions on*, vol. 36, no. 5, pp. 3646 –3648, sep 2000.

[22] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[23] T. P. Tomas Svoboda, Daniel Martinec, "A convenient multi-camera self-calibration for virtual environments," in *PRESENCE: Teleoperators and Virtual Environments, MIT Press*, 2005, pp. 407–422.

[24] D. Comaniciu and V. Ramesh, "Mean shift and optimal prediction for efficient object tracking," in *Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 3, 2000, pp. 70 –73 vol.3.

[25] "Autodesk motionbuilder 2012." [Online]. Available: http://www.autodesk.com/motionbuilder

[26] "Humaine project portal." [Online]. Available: http://emotion-research.net

[27] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," in *Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010)*, Valletta, Malta, May 2010.

[28] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[29] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," 2003.

[30] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10-11, pp. 787–800, Oct. 2007.

[31] S. Abrilian, L. Devillers, S. Buisine, and J. C. Martin, *EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces*, 2005.

[32] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette, and C. Sedogbo, "The safe corpus: illustrating extreme emotions in dynamic situations," in *LREC Workshop Corpora and Emotion*, Gênes, Italie, May 2006.

[33] K. R. Scherer and G. Ceschi, "Lost luggage: A field study of emotion-antecedent appraisal," *Motivation and Emotion*, vol. 21, pp. 211–235, 1997.

[34] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs," Pittsburgh, September 2006, pp. 801–804.

[35] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293 – 303, march 2005.

[36] N. Amir, S. Ron, and N. Laor, "Analysis of an emotional speech corpus in Hebrew based on objective criteria," in *ITRW on Speech and Emotion*, Sep. 2000, pp. 29–33.

[37] F. Schiel, S. Steininger, U. Trk, and U. T. Urk, "The smartkom multimodal corpus at bas."

[38] A. Zara, V. Maffiolo, J. C. Martin, and L. Devillers, "Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics," in *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*. Springer-Verlag, 2007, pp. 464–475.

[39] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition."

[40] S. Özkul, E. Bozkurt, S. Asta, Y. Yemez, and E. Erzin, "Multimodal analysis of upper-body gestures, facial expressions and speech," in *4th International Workshop on Corpora for Research on EMOTION SENTIMENT SOCIAL SIGNALS*, 2012.

[41] D. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," in *Laboratory Phonology*, vol. 3, no. 1, 2012, pp. 71–89.

[42] H. J. P. J. Deller, J., "Discrete-time processing of speech signals." Macmillan Publishing Company, New York.

[43] D. Mcneill, "Hand and mind: What gestures reveal about thought." University Of Chicago Press, 1992.

[44] L. Valbonesi, R. Ansari, D. McNeill, F. Queck, S. Duncan, and K. E. McCullough, "Multimodal signal analysis of prosody and hand motion: temporal correlation of speech and gestures," in *Proc. of the European Signal Processing Conf. (EUSIPCO02)*, vol. 1, 2002, pp. 75–78.

[45] V. Barbu and N. Limnios, *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications: Their Use in Reliability and DNA Analysis*, 1st ed. Springer Publishing Company, Incorporated, 2008.