Real-Time Speech Driven Gesture Animation

by

Kenan Kasarcı

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 Science and Engineering

Koç University

July, 2016

ABSTRACT

Gesticulation, which includes instinctive or planned hand, arm and head body gestures, is an essential component of face-to-face communication. Gesture and speech co-exist in time with a tight synchrony, and they are planned and shaped by the emotional state and produced together. In our early studies we have developed joint gesture-speech models and proposed algorithms for speech driven gesture animation. These algorithms are mainly based on Viterbi decoders and can not run in realtime. In this thesis we investigate real-time implementation of these algorithms via optimal adjustment of the parameters in the Viterbi algorithm and focus on synthesizing upper body gestures in real-time, directly from speech signals without need for additional input. Our framework generates upper body gesture animations by selecting gesture phrases, which are defined in terms of body motion extracted from motion capture data. The selection is driven by a pre-trained hidden semi-Markov model (HSMM) which uses prosody features extracted from speech. Experimental evaluations are performed to compare realtime and non-realtime speech driven gesture animations using both objective and subjective evaluations. Objective evaluations quantify the similarity between gesture phrases over the frames of the corresponding animations. Subjective evaluations are performed with A/B pair comparison test. The experimental results confirm that our system is able to produce realistic and compelling speech-driven body gestures in real-time.

ÖZETÇE

El, kol ve baş ile içgüdüsel veya planlı yapılan vücut jestleri yüz-yüze iletişimin önemli öğelerinden bir tanesidir. Vücut jestleri ve konuşma, zamanda birlikte yer alan, konuşmacı tarafından bilişsel olarak duygu ve etkileşim içeriğine bağlı planlanan ve birlikte üretilen iletişim mekanizmalarıdır. Geçmiş çalışmalarımızda jest–konuşma ortak modellerinden ürettiğimiz konuşma ile sürülen jest animasyonu yöntemleri geliştirdik. Bu yöntemler ağırlıklı olarak Viterbi çözücüler kullanmaktadır ve gerçek zamanlı çalışmamaktadır. Bu tez çalışmasında, Viterbi algoritmasındaki parametreleri en uygun şekilde ayarlayarak, bu yöntemlerin gerçek zamanda çalışır hale getirilmesi üzerine incelemeler yaptık. Aynı zamanda ek bir girdiye ihtiyaç olmadan doğrudan konuşma sinyalleri ile üst–vücut jestlerini sentezlemeye odaklandık. Jest ifadeleri, hareket yakalama verilerinden elde edilen beden devinim örüntülerine karşılık gelir ve üst–vücut animasyonlarının oluşturulmasında kullanılır. Kullandığımız yöntemde jest ifadelerinin seçilmesi işlemi, konuşmadan çıkarılan bürün özniteliklerini kullanarak önceden eğitilmiş yarı saklı-Markov modeli ile yürütülür. Gerçek zamanda oluşturulan animasyonları özgün yöntemle elde edilen animasyonlarla karşılaştırmak üzere hem nesnel hem de öznel ölçütler kullandık. Nesnel değerlendirme için, gerçek zamanda oluşturulan animasyonlar ile özgün yöntemle oluşturulan animasyonların çerçeve bazında ne kadar çakıştığını ölçtük. Öznel değerlendirmelerde ise A/B ikili karşılaştırma sınamasını kullandık. Hem nesnel hem de öznel değerlendirme sonuçlarının gösterdiği üzere sistemimiz gerçek zamanda gerçekçi ve doğal konuşma sürümlü jest animasyonları üretebilmektedir.

TABLE OF CONTENTS

List of Tables **vi**

List of Figures **vii**

Chapter 1: Introduction **1**

1.1 Gesture and Speech 2

1.2 Related Work 3

1.3 Contributions 6

1.4 Organization 7

Chapter 2: Real-time Speech-Driven Animation System **8**

2.1 System Overview 8

2.2 Gesture Generation Model 9

2.3 Gesture Synthesis 11

2.4 Gesture Animation 14

2.5 CreativeIT Database 16

2.6 Animation System 17

Chapter 3: Evaluations **18**

3.1 Objective Evaluations 18

3.2 Subjective Evaluations 21

3.3 Video Examples 22

Chapter 4: Conclusion **23**

Appendices **25**

Appendix A: Constructing the Animation System **26**

Bibliography **29**

LIST OF TABLES

3.1 Real-time factor α for varying values of step size m and number of best states n. 19
3.2 Similarity rate for varying values of step size m and number of best states n. 20
3.3 The Subjective A-B Comparison Results. 22

LIST OF FIGURES

1.1 General HSMM: i_0 and d_0 represents the initial state and duration. The first state i_1 and its duration d_1 are selected according to the transition probability. i_1 produces two observations o_1, o_2 according to the emission probability. It transits, according to the transition probability, to state i_2 6

2.1 The block diagram of general framework. 8

2.2 In a hidden semi-Markov process each state has a duration and emits a number of observations. 10

2.3 Viterbi Path Selection 15

2.4 Screenshot from synthesized animation in Unity3D 16

3.1 Gray scaled colormap of the real-time factor α 19

3.2 Gray scaled colormap of the similarity rate 20

3.3 Screenshot from the video which is used for subjective tests 22

A.1 Real-time speech driven gesture animation system 27

A.2 Animation Stream Diagram 28

Chapter 1

INTRODUCTION

Communication between human characters are generally the most interesting aspects of networked virtual environments. Despite the current graphics technology provides these characters a photo-realistic appearance, it is still incapable of generating the wide variety of motions that human beings exhibit. Gestures and speech are two distinct methods of expressing human emotions and thoughts. As a verbal and non-verbal human communication, they co-exist in time and are tightly intertwined [1], yet modern input devices are inadequate to allow body gestures to be conveyed as intuitively and smoothly as it would be in person. Current virtual environments use keyboard and mouse commands to allow participants to utilize a small set of pre-defined gestures, however this kind of interaction is unnatural for improvised gestures. Due to these restrictions, movements of the human characters must be synthesized automatically to produce natural animations.

The aim of this thesis is to present a data-driven method that automatically generates upper body gesture animation from prosody features extracted from the speech signal in real-time. The framework is trained on motion capture data and recorded audio of people in conversation [2]. The primary contribution of this thesis is a modified method for generating the gesture sequence that is appropriate for real-time synthesis, along with an application that uses this method to produce animations from live speech.

1.1 Gesture and Speech

Gesture and speech are two different forms for expressing emotions and thoughts. They co-exist in time and form a tightly integrated system during language production and comprehension. Besides gesticulation is an essential part of communication, not only the gesticulatory body language of everyday face-to-face communication, but also in the production of speech. It contributes notably to the perception of conversations.

In one of the pioneering studies, Kenton investigated the temporal relationship between speech and body movement and proposed a hierarchical model for gesture [3].In this model, gesture phases are described as the different movement phases observable in the execution of gestures. The meaningful part of the gesture –the part people rely on in their interpretation of a gesture– is the stroke. In order to perform a stroke, the hands need to prepared for their execution during the phase referred to as preparation. The stroke may be followed by a retraction, a phase in which the hands relax and move back into a rest position. These gesture phases build higher-level units, namely gesture phrase. Kendon's pioneer work on the description of gesture phases has been essential to the study of co-verbal gestures. It has shown that gestural movement sequences can be broken down into a succession of different phases, which correspond to units at speech level. Furthermore, it has provided a technique for detailed accounts of gestures and their relation to speech, and has proven the fact that "speech and movement appear together as manifestations of the same process of utterance."

On the other hand, one of the most accepted classification schemes in gesture literature was introduced by McNeill [1]. His taxonomy consists of four gesture types: beats, deictics, iconics and metaphorics. Beat gestures are simple, repetitive, oscillating hand movements. They are not only used to mark the tempo but also to introduce new themes in the utterance. Deictic gestures are classical pointing gestures that are typically performed with an extended index finger. Iconic gestures depict an action or object. Metaphorics resemble iconics but represent an abstract idea rather than an object.

Regarding beat gestures, McNeill observed that "beats tend to have the same

form regardless of content". He also observed that in most cases beats are used to emphasize main concepts or specific words and form nearly half of all gestures in narrative and non-narrative speech. Narrative speech contains many iconic gestures where about two-thirds of the gestures accompanying non-narrative speech contains beats. He suggested that synthesizing only beats gestures are sufficient enough to animate non-narrative speech.

McNeill also observed that metaphoric gestures vary across cultures. Narrative speech contains a great number of iconic gestures whereas non-narrative speech contains beats or metaphorics gestures over 80%. He suggested that synthesizing beats and metaphoric gestures are sufficient enough to animate non-narrative speech.

Prosody is mainly used to emphasize speech along with gestures, which are used to highlight words or word groups. Thus, prosody is a useful clue for selecting gestures to emphasize the underlying speech in natural looking animations.

1.2 Related Work

Literature on gesture and speech is based on diverse studies such as psychology, computer vision, speech processing, linguistics, and machine learning. Even though no system has been proposed which both synthesizes upper-body gestures by using hidden semi-Markov model (HSMM) in real time, several methods have been used to synthesize either full-body or facial animations from a variety of inputs. Goal of these methods is to animate Embodied Conversational Agents (ECAs), which operates on a pre-defined gesture tree [4]. Consequently this allows concurrent planning of synthesized speech and gesture. Frequently these methods depends on the content author to define gestures as part of the input by using a concise annotation scheme [5,6]. In addition, new annotation schemes are still proposed [7]. There are also methods to combine behavioral planning with gesture and speech synthesis [8,9]. But all of these methods depend on an annotation scheme that concisely and completely specifies the desired gestures.

To eliminate the need for annotation of the input text, Stone et al. [10] presented a

data-driven method that re-arranges recorded motion capture data to form the desired utterance. Yet, this method requires the hand annotation of all training data and is restricted to synthesize utterances from recorded phrases.

Since the proposed system must generate animations for arbitrary input, it also cannot require any form of annotation or specification from the user. Several methods have been proposed to animate characters from arbitrary text using natural language processing. Cassell et al. [11] propose an automatic rule-based gesture generation system for ECAs, while Neff et al. [12] present 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. On the other hand, coexisting formation of speech and body language of the text are what these two mechanisms are depending on. Neither non-verbal conversations nor the dialogues manufactured from the texts give as firm impression as a standart dialogue because of the fact that the element of feeling which is the significant part for gestures cannot be grasped by the texts [13].

A technique that relies on some presumption-based model is used for the investigation for fusing facial expressions and lip sync to movements straight from speech. Another design that works by interchanging frames in a video to synchronize it with letters that have been used in dialogue was introduced by Bregler et al. Brand [14] followed practicing Hidden Markov model and combined it with redirected animation onto advanced model to improve this model. Hidden Markov models are generally exercised to monitor the affiliation between speech and facial expressions [15, 16]. There are different automatic mechanisms that offer to manufacture facial expressions using more complicated modification of video sequences [17], mirroring muscles or by practicing a combination of rule-based and data driven mechanisms [17] [18].

Vocal prosody is not usually used by voice-based combination of facial expressions despite the fact that it is a regularly used method. Most voice-based mechanics practices techniques that chooses mouth shapes for relevant character as mouth actions form the most of facial expressions. Nonetheless, some mechanics that use

voice prosody to monitor vivid human movements other than lip movements have been introduced. Albrecht et al. [19] practice prosody aspects to drive a rule-based facial expression movement mechanism. On the other hand newer mechanisms exercise a data-driven path to develop head motion from pitch [20] and facial expressions from vocal intensity [21]. Sargin et at. [22], assimilates a more complex method using prosody features to directly drive head orientation with a hidden Markov model (HMM). Morency et al. [23] propose that prosody can be beneficial for guessing body languages even though these technics only animate head orientation from prosody. Recently, Levine et al. [24] used Hidden Markov Models to generate gestures automatically using speech. This work directly associates animation segments with prosody features and suffers from overfitting. Since the duration of human body gestures varies in nature, the problem of generating body gesture sequences from prosody observations fits into the concept of HSMMs.

The proposed system selects gestures needed for animation by employing a prosody-driven HSMM. Ferguson was the first to consider the HSMM, which is called HMM with variable duration. It allows the stochastic process to be a semi-Markov chain. Each state has variable duration and a number of observations being produced instead of a single one. This shows that the underlying process in the system is Markovian in certain jumps and makes it appropriate for use in a wider range of applications. Its forward–backward algorithms can be used to find the best state sequence of the underlying stochastic process.

While the methods described above are able to synthesize realistic body language for ECAs, they can't generate full-body animations from live speech signal by using HSMM . Animating virtual characters requires real-time speeds and a predictive model that does not rely on looking ahead in the audio stream. Such a model constitutes the main contribution of this work.

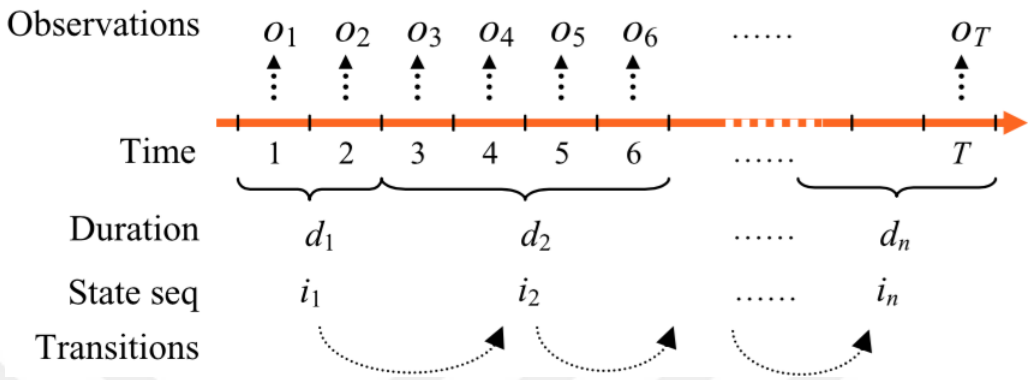

Figure 1.1: General HSMM: i_0 and d_0 represents the initial state and duration. The first state i_1 and its duration d_1 are selected according to the transition probability. i_1 produces two observations o_1, o_2 according to the emission probability. It transits, according to the transition probability, to state i_2

1.3 Contributions

In this thesis, we introduce a system that generates upper-body gesture animations for human characters from live speech signal in real-time. We note that the HSMM-based gesture synthesis method that we use in this thesis work is adapted from [25] [26]. In this early study we developed joint gesture-speech models and proposed algorithms for speech driven gesture animation. These algorithms which are based on Viterbi dynamic algorithm, produce gesture sequences with durations. In this work our goal is to generate believable and natural looking upper-body gesture animations in real-time. Our contributions are as follows:

- In the system, gesture phrases are defined as states of Markov chain and intonation phrases as observations of Markov process. The standard Viterbi dynamic algorithm is employed to decode the most likely gesture sequence and durations. Since Viterbi algorithm requires all past and future observations, it is not generally used to find the global optimum state sequence in real-time system. For this reason, to make it particularly useful for our real-time system, we have investigated necessary modifications to allow Viterbi algorithm to operate on

infinite time sequences and produce the optimum states with only a finite delay.

- We have investigated required duration for synthesizing gestures by changing step size in duration model and the number of best states in Viterbi algorithm.

- We have generated animations in real-time using Unity3D[1]. The animation process consists of three main tasks: *i*) extraction of prosody features from speech signal in real-time *ii*) selection of appropriate gestures corresponding to live speech signal in real-time *iii*) lining the appropriate gestures up in time to form a continuous animation stream.

1.4 Organization

Organization of this thesis is as follows: In chapter 2 we introduce the proposed system. We first summarize the previously developed joint gesture-speech model for speech driven gesture animation and then present our modifications in order to make it run in real time. In Chapter 3 presents the experimental evaluations using both objective and subjective measures. Finally, in Chapter 4 we discuss conclusions and future directions.

[1]Unity, http://unity3d.com

Chapter 2

REAL-TIME SPEECH-DRIVEN ANIMATION SYSTEM

2.1 System Overview

The general framework for the automatic upper body gesture synthesis system, from which this thesis work is adapted, is given in Figure 2.3. The framework consists of three main functional blocks: *i*) analysis of gesture and intonational phrases, *ii*) speech driven gesture synthesis, and *iii*) gesture animation.

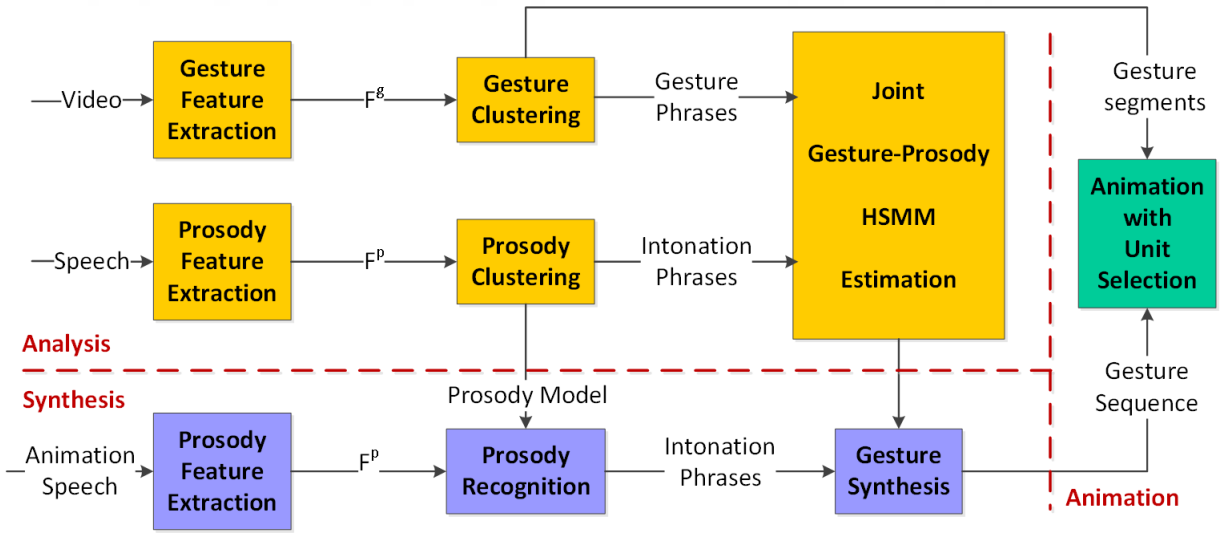

Figure 2.1: The block diagram of general framework.

Unimodal analysis of speech and body motion, which is performed in the analysis functional block, is used to learn temporal patterns of body motion features along with speech prosody features. Body motion features, which are extracted from motion capture data, are used to define gesture phrases with a semi-supervised temporal

clustering scheme. Moreover prosody features, which are extracted from speech input, are used to define intonational phrases with an unsupervised temporal clustering scheme. In addition to this, analysis functional block performs multimodal analysis to learn dependencies between gesture and intonational phrases by utilizing an HSMM. Synthesis functional block performs gesture synthesis which is an extraction of gesture sequence and gesture durations, given the speech signal. Finally, animation functional block performs gesture animation, where the synthesized gesture sequence is mapped into body motion sequences so as maintain a natural looking animation.

2.2 Gesture Generation Model

In our system, speech signal is described with its prosody features whereas gesture features are described with Euler rotation angles of left-right forearm and left-right arm. Gesture patterns, i.e., gesture phrases, that we use are obtained via semi-supervised clustering over joint-angle stream, and are each expected to correspond to a meaningful movement (i.e. spreading arms). Gestures are modeled as HSMM states and corresponding prosody features are observations. The trained HSMM model and Viterbi algorithm are used for synthesizing real-time gesture types with duration information given a prosody sequence as input. Gesture type and duration information are picked from gesture pool with unit selection method for producing optimum gesture sequence and an animation is created for this sequence.

There are two parts for this system to work in real-time: Speech driven gesture synthesis and body gesture animation. Body movement synthesis problem is taken in consideration in the first part and the second part focuses on creating natural looking body motion animation from synthesized gesture sequence.

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

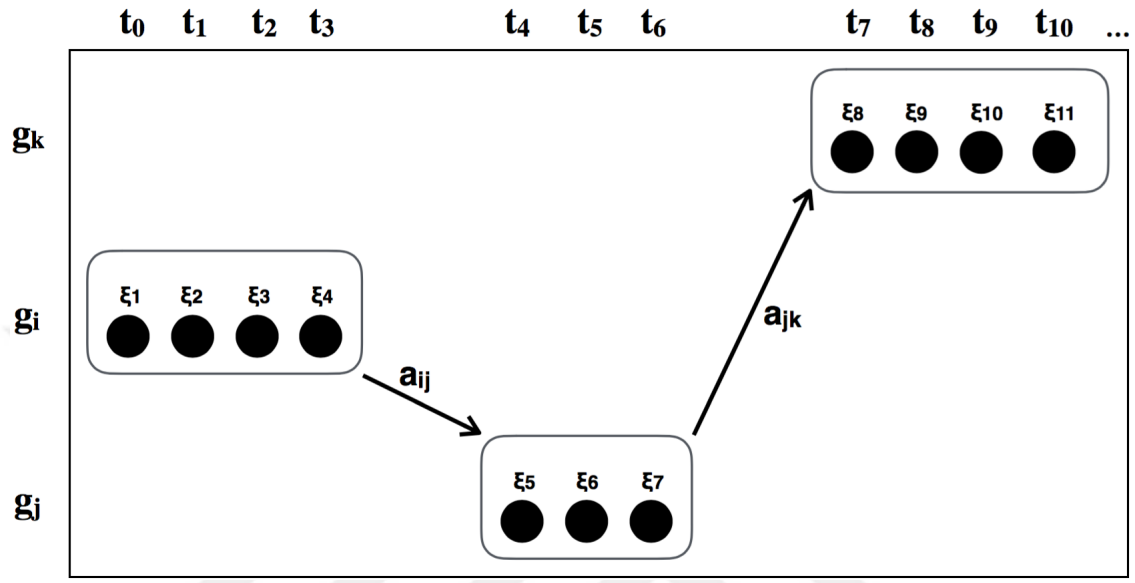

Figure 2.2: In a hidden semi-Markov process each state has a duration and emits a number of observations.

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

$$\mathbf{A} : \left\{a_{ij} = P(\ell_l^g = g_j \mid \ell_{l-1}^g = g_i)\right\} \quad i, j = 1, ..., M_g, \tag{2.1}$$

where ℓ_l^g represents the ℓth gesture in the sequence of gesture phrases. The observation emission distribution $\mathbf{B}$ is modeled by discrete probability mass functions for each gesture g_i,

$$\mathbf{B} : \{b_i(p_k) = P(p_k \mid \ell_l^g = g_i)\} \quad k = 1, ..., M_p, \;\; i = 1, ..., M_g, \tag{2.2}$$

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

$$\mathbf{D} : \{d_i(k)\} \quad i = 1, ..., M_g, \;\; k = 1, ..., \tfrac{D_{max}}{\delta}, \tag{2.3}$$

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 $\mathbf{\Pi}$ is defined by entries π_i representing the probability of starting with gesture g_i as the first gesture phrase,

$$\mathbf{\Pi} : \{\pi_i = P(\ell_l^g = g_i)\} \quad i = 1, ..., M_g. \tag{2.4}$$

2.3 Gesture Synthesis

Synthesizing gesture animations in real time is an important issue with the current system that can animate believable and natural looking body gestures. Viterbi algorithm must include all past and future states to find the optimal gesture sequence. For this reason most of the times it can not be used on systems working in real time.

In order to make Viterbi algorithm to be able to operate in real time, instead of evaluating every states at each frame, n best states which have the highest scores are analyzed and future frames are evaluated based on these best states. Moreover duration possibilities in the model are evaluated using a step size denoted by m.

Gesture synthesis is defined as decoding the most likely state sequence, $\hat{\ell}^g$, over the HSMM Λ^{gp} given a sequence of frame level intonational phrase labels, $\{\xi_1, \xi_1, ..., \xi_T\}$. The optimal state sequence contains synthesized sequence of gesture phrases and their durations, where the HSMM framework produces realistic gesture durations. In HMM framework, where the underlying process is Markov, given an observation sequence, the Viterbi algorithm is employed to decode an optimal state sequence. In HSMM framework however, states have variable durations and a sequence of observations is emitted from a single state. Therefore the forward likelihood function, which incorporates state duration model, is defined as,

$$\psi_t(j) = \max_{\tau} \max_{i} \left\{ \psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(\xi_k)) \right\} \tag{2.5}$$

where $\psi_t(j)$ is the accumulated logarithmic likelihood at time frame t in state g_j after observing intonational phrase labels $\{\xi_1, \xi_1, ..., \xi_T\}$. We defined maximum time as $T = 200$.

Step size, m, is defined as the step size between $[1, 200]$ in duration model. $L_{BestStates}$ represents the list of n number of states having the best cumulative scores from $[1, M_g]$ at frame t.

Viterbi algorithm that produces the optimum gesture sequence $\hat{\ell}^g = \left\{ \hat{\ell}_1^g, ..., \hat{\ell}_L^g \right\}$ and gesture duration sequence $\kappa = \{\kappa_1, ..., \kappa_L\}$ are given below.

Algorithm 1 The modified Viterbi decoding algorithm for real-time gesture synthesis

i. Initialize

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

$\mathbf{L_{bestStates}} = \{\mathbf{1}, \mathbf{2}, ..., \mathbf{M_g}\}$

ii. Recursion: Repeat for: $t = 2, 3, ..., T$

$\mathbf{T'} = \mathbf{1}, \mathbf{m+1}, \mathbf{2m+1}, \mathbf{3m+1}, ..., \min(\mathbf{D_{max}}, \mathbf{t})/\delta$

Repeat for: $j = 1, 2, ..., M_g$

$$\Psi_{t\tau}^{ij} = \psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \Pi_{k=t-\tau+1}^{t} b_j(\xi_k))$$

$$\psi_t(j) = \max_{\tau \in T'} \max_{i \in [1, M_g]} \left\{ \Psi_{t\tau}^{ij} \right\}$$

$$\varphi_t(j) = \arg\max_{i \in \mathbf{L_{bestStates}}} \max_{\tau \in [1, T']} \left\{ \Psi_{t\tau}^{ij} \right\}$$

$$\nu_t(j) = \max_{\tau \in T'} \max_{i \in [1, M_g]} \left\{ \Psi_{t\tau}^{ij} \right\}$$

Update $L_{bestStates}$ to have the n best indices of $\arg\max_{j \in [1, M_g]} \{\nu_t(j)\}$

iii. Backtrace the optimal gesture phrase sequence

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

$\kappa_L = \nu_T(\hat{\ell}_L^g); \quad l = L - 1; \quad t = T$

While $t > 0$

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

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

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

2.4 Gesture Animation

Gesture animation contains two main tasks

- Extraction of gesture sequence with unit selection
- Animating gesture motion sequence in Unity3D.

In the extraction phase, the synthesized gesture phrase $\hat{\ell}^g$ and duration κ sequences are used to generate a synthesized sequence of motion segments $\hat{\epsilon}^g$. To achieve this goal, unit selection is performed. For each gesture g_i, a set of representative temporal segment templates is formed as $G_i = \left\{\varepsilon^{g_i 1}, \varepsilon^{g_i 2}, ..., \varepsilon^{g_i K_i}\right\}$ where K_i represented as the number of templates in the collection of gesture g_i. In this process, the penalty score of the duration and joint angle continuity is aimed to be minimum.

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

$$\upsilon_k\left(\varepsilon^{g_i k} \mid \hat{\ell}_l^g = g_i\right) = \left\|\kappa_l - \kappa\left(\varepsilon^{g_i k}\right)\right\|, \tag{2.6}$$

$$\upsilon_\omega\left(\varepsilon^{g_i k} \mid \hat{\ell}_l^g = g_i\right) = \left\|\omega_e\left(\hat{\varepsilon}_{l-1}^g\right) - \omega_b\left(\varepsilon^{g_i k}\right)\right\| \tag{2.7}$$

where κ_l is the duration of the synthesized gesture phrase $\hat{\ell}_l^g, \kappa\left(\varepsilon^{g_i k}\right)$ is the duration of the gesture segment template $\varepsilon^{g_i k}, \omega_e\left(\hat{\varepsilon}_{l-1}^g\right)$ is the ending joint angle vector of the synthesized gesture segment $\hat{\varepsilon}_{l-1}^g$ and $\omega_b\left(\varepsilon^{g_i k}\right)$ 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,

$$\upsilon\left(\varepsilon^{g_i k} \mid \hat{\ell}_l^g = g_i\right) = \alpha\upsilon_\omega\left(\varepsilon^{g_i k} \mid \hat{\ell}_l^g = g_i\right) + (1-\alpha)\,\upsilon_\kappa\left(\varepsilon^{g_i k} \mid \hat{\ell}_l^g = g_i\right) \tag{2.8}$$

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

i. Initialization

$$V_1(k) = \upsilon_\omega \left(\varepsilon^{g_i k} \mid \hat{\ell}_l^g = g_i \right), \text{ where } k = 1, 2, ..., K_i$$

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

$$V_1(k) = min_{j=1,...,K_i} \left\{ V_{l-1}(j) + \upsilon \left(\varepsilon^{g_i k} \mid \hat{\ell}_l^g = g_i \right) \right\}$$

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

iii. Backtrace the optimal path

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

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

iv. Construct the synthesized sequence of gesture motion segments

$$\hat{\varepsilon}^{\hat{l}_l^g} = \hat{\varepsilon}^{\hat{l}_l^g g_l} \text{ for } l = 1, 2, ..., L.$$

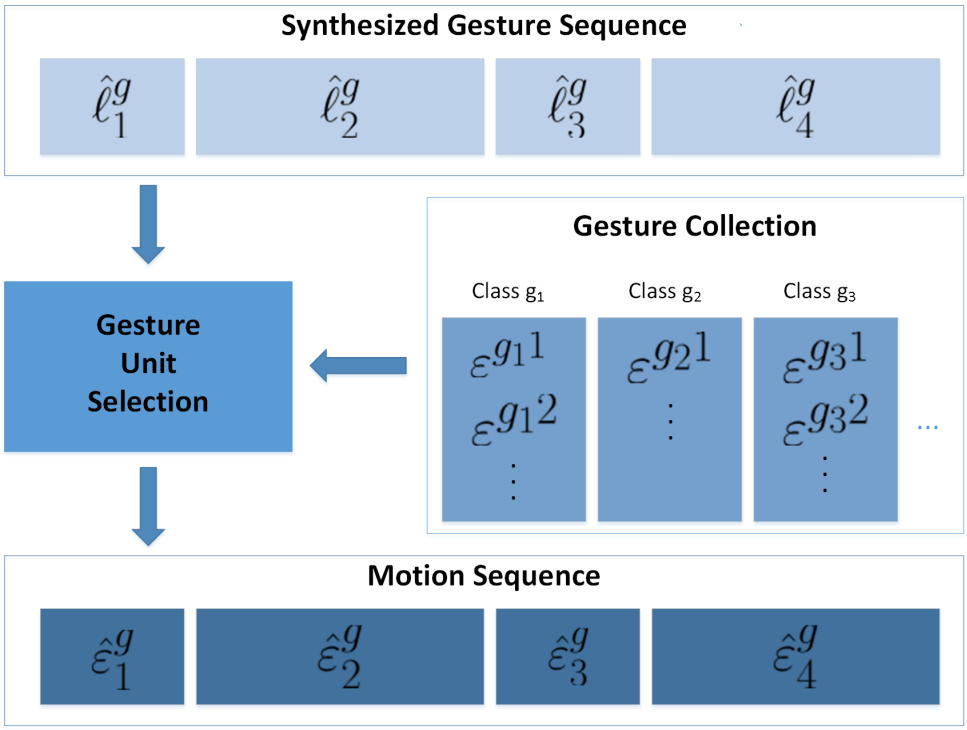

Figure 2.3: Viterbi Path Selection

In the second phase, gesture motion sequence is converted to character animation file format and animated using Unity3D.

Figure 2.4: Screenshot from synthesized animation in Unity3D

2.5 CreativeIT Database

In this work, we use the multimodal CreativeIT database that consists of goaldriven improvised interactions [27]. The database is collected using cameras and microphones and contains detailed full body motion capture data in addition to audio data. The actors' gestures were kept as natural and intuitive as possible and the setup itself did not influence the actors' expressive behavior. The actors were not instructed to produce specific emotions, instead a variety of improvised emotional expressions and interaction dynamics occur as part of the performance. The database contains the recording of nine full sessions, each of which contains approximately one hour of audiovisual data.

2.6 Animation System

The animation system is designed for animating virtual character as we proposed in this thesis. The system consists of four main tasks: i) live speech record, ii) extraction of the prosody features, iii) speech driven gesture synthesis, and iv) gesture animation. The system is written almost entirely in C# and has been ported to run on both Windows, OS X. *Appendix*1 shows the design and implementation of the system.

Chapter 3

EVALUATIONS

3.1 Objective Evaluations

In the experiments, the synthesized gestures and their durations have been obtained by changing the step size in duration model and the best states number in Viterbi algorithm. Due to the fact that states have averagely 12 transitions between each other, best states number is taken between $[3, 12]$. At the end of each iteration, n number of states with highest cumulative scores are added to $L_{bestStates}$ list and used in next iteration. Step size is selected in between $[1, 10]$, where the default value is 1.

Real-time factor, α, is the metric that we define to measure the real time performance of the system and it is calculated as the ratio of gesture synthesis duration to speech recording duration. The duration of unit selection is not included as it takes much less time compared to Viterbi decoder. Real-time animations can be generated when real-time factor is below 1. Gray scaled colormap related to best states number and step size is given in Figure 3.1. The region on the right of the white line represents the cases when real-time factor is below 1. When you get closer to black, real time factor drops below 1. Table 3.1 shows the real-time factor α for varying values of step size m and number of best states n.

We have also defined an objective similarity metric to measure the quality of animations we obtained in real-time. This similarity metric gives the percentage of the amount of overlap of frame–based gesture types between two gesture sequences.

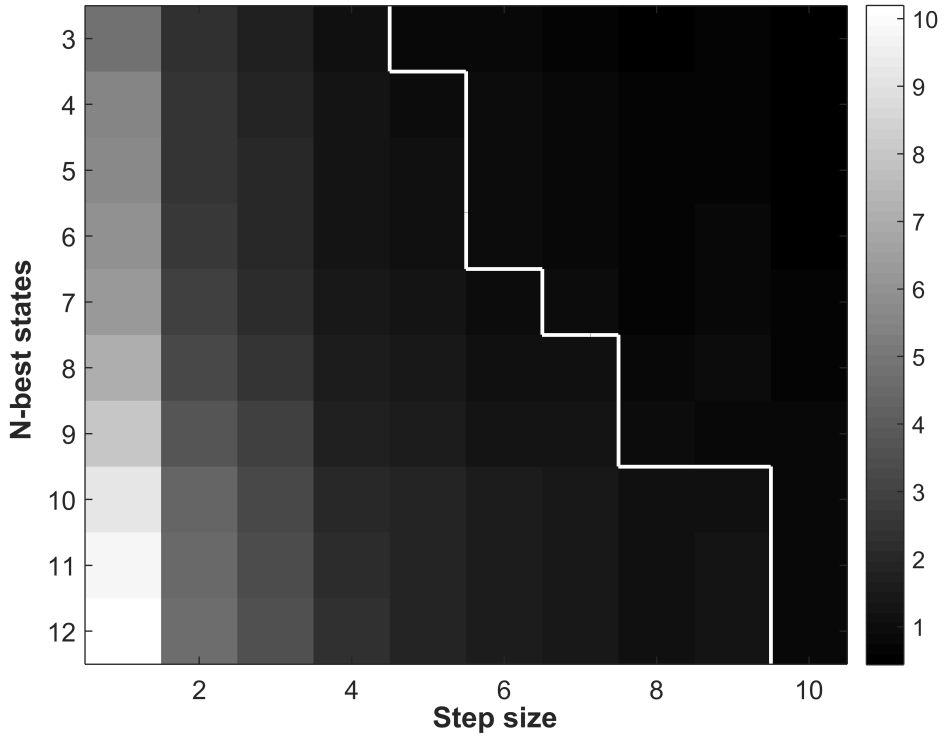

Figure 3.1: Gray scaled colormap of the real-time factor α

Table 3.1: Real-time factor α for varying values of step size m and number of best states n.

m	n	α	m	n	α	m	n	α
10	3	0.4383	9	4	0.7050	8	8	0.8436
10	4	0.4779	9	5	0.7303	8	9	0.9610
10	5	0.5043	9	6	0.7660	7	3	0.7331
10	6	0.5472	9	7	0.8501	7	4	0.8065
10	7	0.6165	9	8	0.9724	7	5	0.8393
10	8	0.7067	9	9	0.7723	7	6	0.8805
10	9	0.7689	8	3	0.5723	7	7	0.9754
10	10	0.8103	8	4	0.6360	6	3	0.8513
10	11	0.8548	8	5	0.6525	6	4	0.9066
10	12	0.8615	8	6	0.6872	6	5	0.9351
9	3	0.6520	8	7	0.7417	6	6	0.9663
5	3	0.8930						

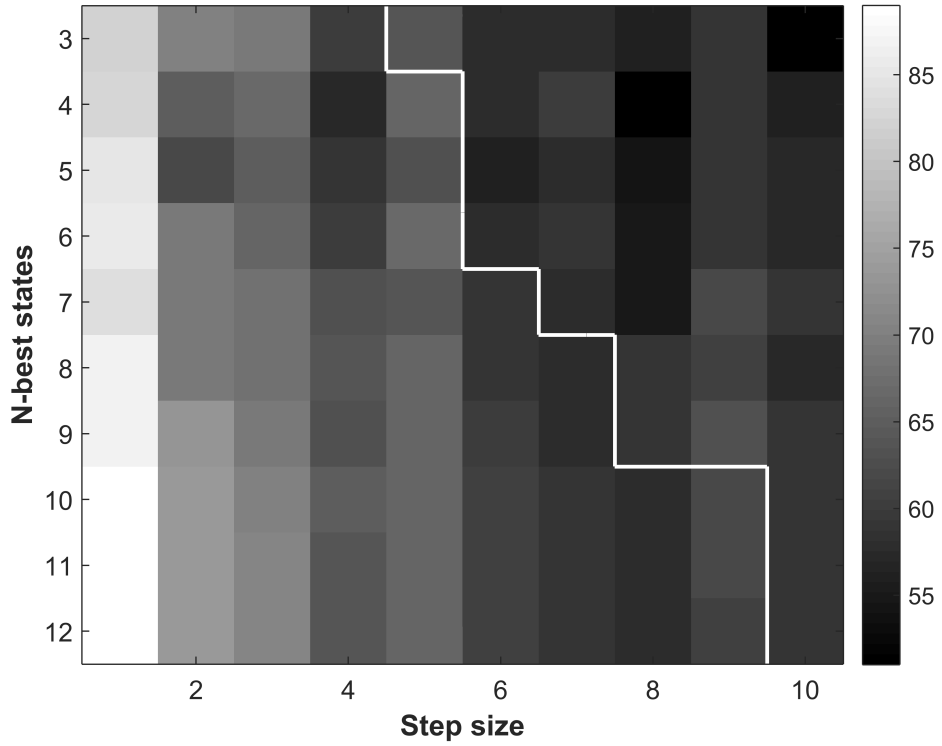

Figure 3.2: Gray scaled colormap of the similarity rate

Table 3.2: Similarity rate for varying values of step size m and number of best states n.

m	n	**Similarity** (%)	m	n	**Similarity** (%)	m	n	**Similarity** (%)
10	3	51	9	4	59	8	8	59
10	4	56	9	5	59	8	9	59
10	5	57	9	6	59	7	3	58
10	6	57	9	7	62	7	4	60
10	7	59	9	8	61	7	5	58
10	8	57	9	9	63	7	6	59
10	9	59	8	3	56	7	7	58
10	10	59	8	4	51	6	3	58
10	11	59	8	5	54	6	4	58
10	12	59	8	6	55	6	5	56
9	3	59	8	7	55	6	6	58
5	3	64						

Figure 3.2 shows gray scaled colormap of similarity metric according to number of best states and step size. The regions on the right of the white line represents the cases when the system can run in real-time. Closer you get to white color, higher the similarity metric value. Table 3.2 shows the similarity rate for varying values of step size m and number of best states n.

The operating points $(3, 5)$ and $(9, 9)$ are determined as the points where real-time factor is below 1 and the similarity metric has the highest score. At these points, 64% similarity score is obtained. Gesture synthesis and animations are generated in Unity3D by using these points.

3.2 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 in real-time. 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 7 segments, where each segment is approximately 1 minute. For each segment, there exists real-time, non-real time and random body gesture animations. Random body gesture animation is generated by selecting random gestures from the gesture pool and appending them in sequence.

The overall evaluation comprises 7 steps of video-pair comparison. Steps are formed as two (Real-time vs. Non-real time), two (Real-time vs. Random), two (Non-real time vs. Random) and one pair of identical animations comparisons in a random sequence.

The subjective tests are performed over 9 subjects. The average preference scores for the three comparison types are presented in Table 3.3. We observe that the real-time animations are favored over the random ones while the non-real time animations are favored over the real-time animations as expected.

Table 3.3: The Subjective A-B Comparison Results.

A–B Pair	**Preference Score**
Non-Real Time versus Random	-1.22
Real-Time versus Random	-0.66
Non-Real Time versus Real-Time	-0.77

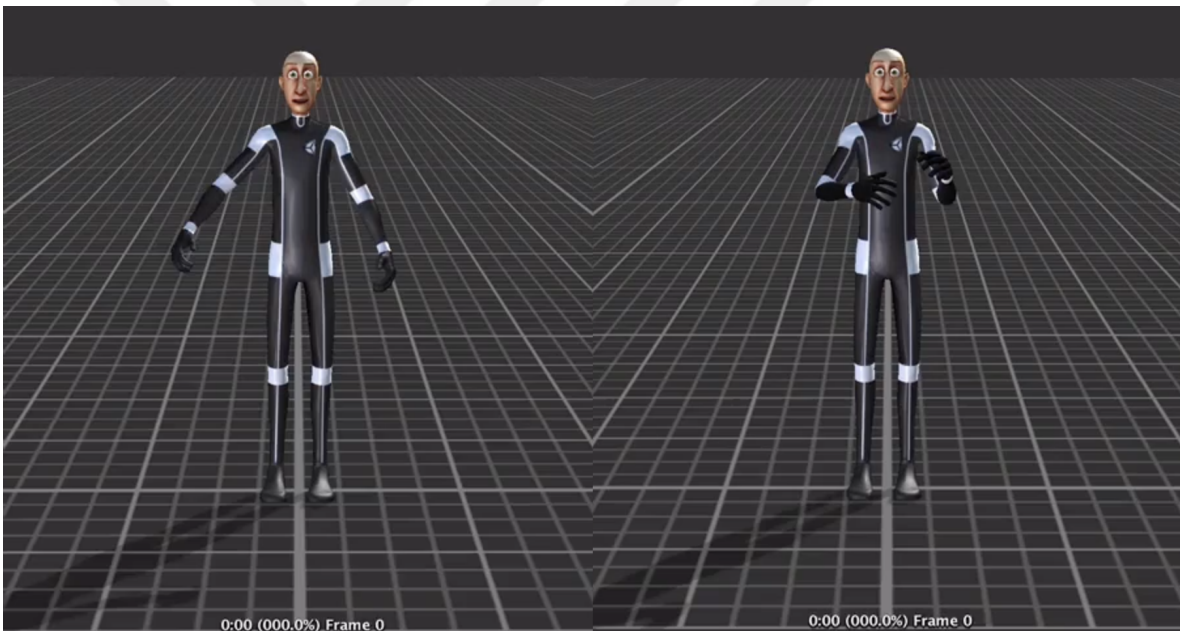

Figure 3.3: Screenshot from the video which is used for subjective tests

3.3 Video Examples

A number of videos, besides the images provided in this thesis, were synthesized to demonstrate the quality of animations generated with the proposed system. The videos present examples of animations and may be viewed on the website [1].

[1]Sample videos for real-time animation. http://mvgl.ku.edu.tr/demos

Chapter 4

CONCLUSION

In this thesis we have presented a new system that generates upper-body gesture animations for human character from live speech signal in real-time. Especially, we focused on developing algorithms to generate gesture animations in real-time. We note that HSMM-based gesture synthesis method that we used in this thesis work is adapted from our early study [25]. Main contributions of this thesis can be summarized as follows:

- We did necessary modifications to allow Viterbi dynamic algorithm to operate on infinite time sequences and produce the optimal gesture sequence only a finite delay.

- We optimized the parameters in Viterbi algorithm so that the system can run in real-time. By changing the step size on the duration model and using n numbered best states at the frame t, we were able to reduce real-time factor below 1.

- We constructed continuous animation stream which extracts the prosody features from the speech signal and selects the appropriate gestures corresponding to live speech. The system is developed by using Unity3D which has a rich and sophisticated animation system.

Since the beginning, all of our studies about Viterbi algorithm pointed out that creating a real-time speech driven gesture animation is crucial, that determines the performance and quality of the overall system. Therefore we put significant effort to optimize the parameters. In our results, we observed that evaluating the best states is

an effective way of reducing the real-time factor below 1, however the fact remains that changing the step size is much more effective on both the real-time factor and similarity values.

Although we tested our system on monologs we strongly believe that proposed system can be expanded to dyadic interactions. We also believe that the proposed system can be easily adapted to other multimodal applications such as speech driven facial expression synthesis in real-time.

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

- Video games and movies which require realistic body animations that are 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.

- Communication applications, specifically in visual teleconferencing where the receiver 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.

- Anchors or video jockeys can be replaced with 3D human models with realistic gestures.

As future research for real-time speech driven gesture synthesis, a new model, facial expressions, can be added to our system. Therefore from live speech signal, upper-body and face animations could be generated.

Appendices

Appendix 1

CONSTRUCTING THE ANIMATION SYSTEM

This appendix presents the design of the animation system. Since the system uses existing motions, the gestures it produces must be included in the motion database. Therefore, the angle information of every gesture within the gesture phrase pool is extracted and converted into Biovision Hierarchy (BVH) character animation file format. BVH files contain skeleton hierarchy information in addition to the motion data. To be able to use the character animations in Unity3D each BVH file is converted to FilmBox (FBX) binary file format with Blender[1] program by using Python and Bash scripting.

FBX files are loaded as resource into Unity3D and the skeleton hierarchy information is mapped into humanoid skeletons. At the last phase, animation clips are extracted from the humanoid animations to use in the system.

Basically, system's working principles are as follows:

- In Main Thread, speech signals are recorded every minute and converted to Waveform Audio (WAV) file format which contains uncompressed audio.
- A new thread is created for each WAV file and prosody features are extracted in this thread.
- Gesture id and duration information are synthesized via Viterbi algorithm from prosody features given to the system as an input.
- Animation clips and gesture durations that paired with gesture ids are sent to animation queue in Main Thread.

[1]Blender, https://www.blender.org/

- Each animation clip in animation queue is animated in order while its duration is also taken in consideration.

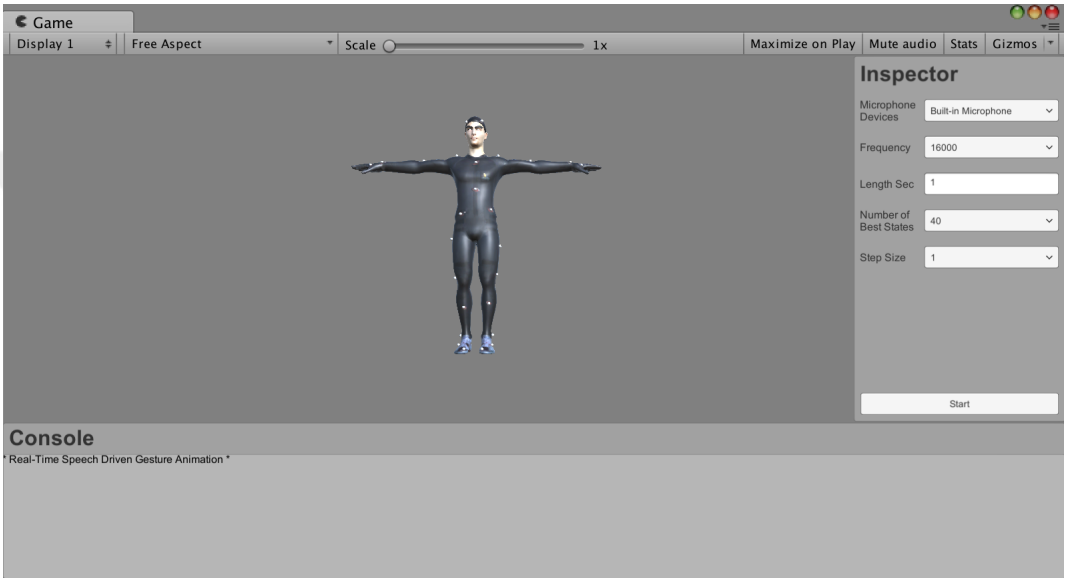

Figure A.1: Real-time speech driven gesture animation system

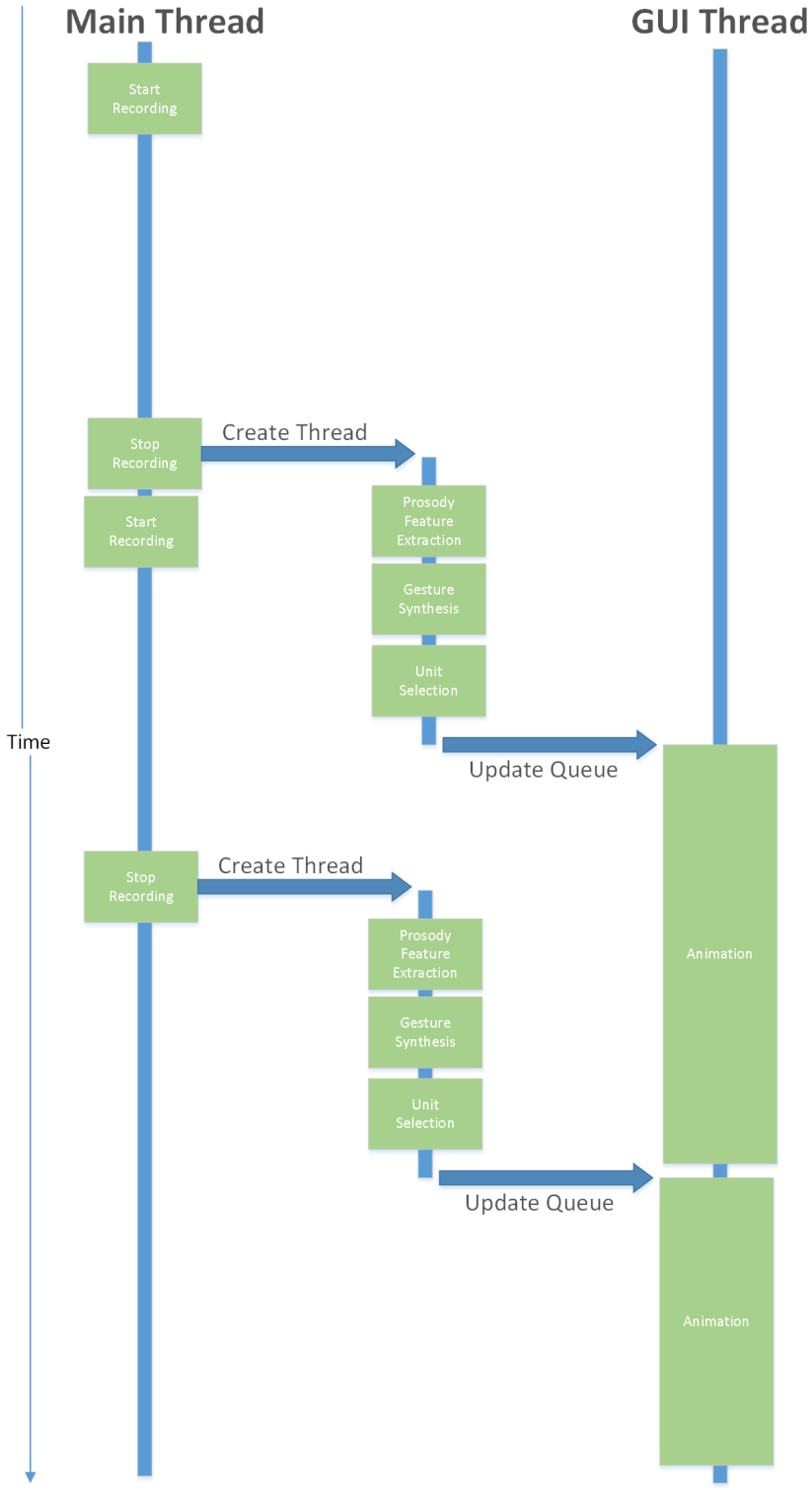

Figure A.2: Animation Stream Diagram

BIBLIOGRAPHY

[1] D. McNeill, *Hand and mind: What gestures reveal about thought.* University of Chicago Press, 1992.

[2] A. Metallinou, C. C. Lee, C. Busso, S. Carnicke, and S. S. Narayanan, "The USC CreativeIT Database : A Multimodal Database of Theatrical Improvisation," in *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC)*, May 2010.

[3] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," *The relationship of verbal and nonverbal communication*, vol. 25, no. 1980, pp. 207–227, 1980.

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

[5] B. Hartmann, M. Mancini, and C. Pelachaud, "Formational parameters and adaptive prototype instantiation for mpeg-4 compliant gesture synthesis," in *Computer Animation, 2002. Proceedings of.* IEEE, 2002, pp. 111–119.

[6] S. Kopp and I. Wachsmuth, "Synthesizing multimodal utterances for conversational agents," *Computer animation and virtual worlds*, vol. 15, no. 1, pp. 39–52, 2004.

[7] M. Kipp, M. Neff, and I. Albrecht, "An annotation scheme for conversational gestures: How to economically capture timing and form," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 325–339, 2007.

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

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

[10] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, and C. Bregler, "Speaking with hands: Creating animated conversational characters from recordings of human performance," in *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3. ACM, 2004, pp. 506–513.

[11] J. Cassell, H. H. Vilhjálmsson, and T. Bickmore, "Beat: the behavior expression animation toolkit," in *Life-Like Characters*. Springer, 2004, pp. 163–185.

[12] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel, "Gesture modeling and animation based on a probabilistic re-creation of speaker style," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 1, p. 5, 2008.

[13] C. Jensen, S. D. Farnham, S. M. Drucker, and P. Kollock, "The effect of communication modality on cooperation in online environments," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 2000, pp. 470–477.

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

[15] Y. Li and H.-Y. Shum, "Learning dynamic audio-visual mapping with input-output hidden markov models." *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 542–549, 2006.

[16] J. Xue, J. Borgstrom, J. Jiang, L. E. Bernstein, and A. Alwan, "Acoustically-driven talking face synthesis using dynamic bayesian networks," in *2006 IEEE International Conference on Multimedia and Expo*. IEEE, 2006, pp. 1165–1168.

[17] Y.-J. Chang and T. Ezzat, "Transferable videorealistic speech animation," in *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*. ACM, 2005, pp. 143–151.

[18] J. Beskow, "Talking heads-models and applications for multimodal speech synthesis," 2003.

[19] O. Arikan and D. A. Forsyth, "Interactive motion generation from examples," in *ACM Transactions on Graphics (TOG)*, vol. 21, no. 3. ACM, 2002, pp. 483–490.

[20] E. Chuang and C. Bregler, "Mood swings: expressive speech animation," *ACM Transactions on Graphics (TOG)*, vol. 24, no. 2, pp. 331–347, 2005.

[21] E. Ju and J. Lee, "Expressive facial gestures from motion capture data," in *Computer Graphics Forum*, vol. 27, no. 2. Wiley Online Library, 2008, pp. 381–388.

[22] M. E. Sargin, E. Erzin, Y. Yemez, A. M. Tekalp, A. T. Erdem, C. Erdem, and M. Ozkan, "Prosody-driven head-gesture animation," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 2. IEEE, 2007, pp. II–677.

[23] L.-P. Morency, C. Sidner, C. Lee, and T. Darrell, "Head gestures for perceptual interfaces: The role of context in improving recognition," *Artificial Intelligence*, vol. 171, no. 8, pp. 568–585, 2007.

[24] S. Levine, C. Theobalt, and V. Koltun, "Real-time prosody-driven synthesis of body language," in *ACM Transactions on Graphics (TOG)*, vol. 28, no. 5. ACM, 2009, p. 172.

[25] E. Bozkurt, S. Asta, S. Özkul, Y. Yemez, and E. Erzin, "Multimodal analysis of speech prosody and upper body gestures using hidden semi-markov models," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3652–3656.

[26] K. Kasarcı, E. Bozkurt, Y. Yemez, and E. Erzin, "Real-time speech driven gesture animation," in *2016 24th Signal Processing and Communication Application Conference (SIU)*. IEEE, 2016, pp. 1917–1920.

[27] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The usc creativeit database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.