Multimodal Analysis and Synthesis of Affective Human Body Gestures from Speech Prosody

by

Elif Bozkurt

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

Doctor of Philosophy

in

Electrical and Electronics Engineering

KOÇ
UNIVERSITY

September, 2016

Multimodal Analysis and Synthesis of
Affective Human Body Gestures from Speech Prosody

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Elif Bozkurt

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. Dr. Engin Erzin

_____
Prof. Dr. Çiğdem Eroğlu Erdem

_____
Assoc. Prof. Dr. Sinem Çöleri Ergen

_____
Assoc. Prof. Dr. Yücel Yemez

_____
Assoc. Prof. Dr. Hazım Kemal Ekenel

Date: _____

ABSTRACT

Conversational agents become more realistic as they utilize affective non-verbal communication, such as the use of gestures, in human computer interaction. Synthesis and animation of gestures to accompany affective verbal communication can help to create more naturalistic conversational agents. In human-to-human communication, speech signal carries rich emotional cues, which are further emphasized by affect-expressive gestures. Speech-driven gesture synthesis can map emotional cues of the speech signal into affect-expressive gestures by modeling complex variability and timing relationships of the joint articulation of speech and gesture. In this thesis, we first introduce a framework for joint analysis of speech prosody and human body gestures towards automatic synthesis and realistic animation of beat gestures from speech prosody and rhythm. Later, we investigate the use of continuous affect attributes, which are activation, valence and dominance, in speech-driven affective synthesis and animation of gestures. We present a statistical framework for multimodal analysis of gesture, speech and affect based on the hidden semi-Markov models, where gestures are representing the states, and speech-prosody and continuous affect attributes are representing the observations of the model. Different structures of the statistical model are evaluated for synthesis and animation of affect-expressive gestures given speech and affect attributes. Evaluations are performed over two multimodal datasets in speaker-dependent and independent settings. Among different statistical structures, the conditional structure, which models observation distributions as prosody given affect, achieved the best performance with objective evaluations for the gesture synthesis and with subjective evaluations for the gesture animation.

ÖZETÇE

Konuşabilen insansı sanal arayüzler, jestler gibi duygu yüklü sözsüz iletişimi kullandıkça insan bilgisayar etkileşiminde daha gercekçi hale gelirler. Duygu yüklü sözlü iletişime eşlik eden jestlerin sentez ve canlandırması, daha doğal görünen konuşabilen arayüzleri yaratmakta yardımcı olur. İnsan insana iletişimde zengin duygu bilgisi içeren konuşma sinyali, duygu ifade eden jestler ile daha da vurgulanır. Konuşma sürümlü jest sentezi, konuşma sinyalindeki duygu içeriğini konuşma ve jest karmaşık çeşitliliğini ve zamanlama ilişkisini modelleyerek duygu yüklü jestlere aktarır. Bu tezde, önce konuşma bürünü ve ritimi sürümlü vurgu jestleri otomatik sentezi ve gerçekçi canlandırması için konuşma bürünü ile jestlerin ortak modellemesini sunuyoruz. Daha sonra, aktivasyon, değerlik, ve baskınlık ile ifade edilen sürekli duygu özniteliklerini kullanarak konuşma sürümlü duygu yüklü jest sentezi ve canlandırmasını araştırıyoruz. Jestlerin durum ve konuşma bürünü ile duygu özniteliklerinin gözlem olarak ifade edildiği saklı yarı-Markov modelleri kullanarak jest, konuşma ve duygu sinyallerinin çok kipli analizini sunuyoruz. Konuşma ve duygu öznitelikleri sürümlü duygu ifadeli jest sentezi ve canlandırması için farklı istatiksel modeller değerlendirildi. Konuşmacıya bağlı ve konuşmacıdan bağımsız olarak düzenlenen sistemler çok kipli iki veri tabanı üzerinde denendi. Farklı istatiksel yapılar içinde, gözlem dağılımını duyguya bağlı bürün olarak modelleyen koşullu yapı, jest sentezi nesnel değerlendirmelerinde ve jest canlandırması öznel değerlendirmelerinde en iyi başarımı elde etti.

ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor, Prof. Engin Erzin, who has always been an inspiration for me. He has constantly offered his guidence and support, and spent time and energy in helping me during my research. I would like to thank Prof. Çiğdem Eroğlu Erdem and Prof. A. Tanju Erdem, who gave me the chance to participate in many exciting projects. I would like to thank Prof. Yücel Yemez, to whom I owe immense gratitude for his help and long, interesting discussions. I would like to thank Prof. Hazım Kemal Ekenel for his support during my undergraduate years, which encouraged me to pursue a career as a researcher. I would like to thank Prof. Sinem Çöleri Ergen, whose feedback and valuable suggestions helped me to improve this thesis. I would also like to thank present and former fellow researchers at the MVGL Lab, with whom it was a great pleasure to collaborate and share ideas.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Context and Motivation

Verbal and non-verbal behaviors are the two important aspects of human communication. While speech is the source of verbal communication, gestures as a channel for non-verbal communication, help to emphasize, complement, or clarify the intended thoughts, attitudes, and emotions. In this regard, computer generated gesture animations represent an elementary source of visual content for a wide variety of applications such as, movies and video games, and broader domains as human-computer interaction. The increasing role of computers in human life brings a demand for higher quality and more realistic animations and the main target for this demand is the representation of conversational agents (CAs). Although CA animations have reached a photorealistic level, representation of CAs is a challenging task when motion is involved.

Creating believable CAs engage multiple research directions ranging from motion control to cognitive aspects, where one important aspect is synthesis of gestures accompanying speech. Various types of systems have been proposed towards this goal and the main components considered include the following: source input signal (text, audio, video, motion capture), output signal (modality, visual quality, naturallness, expressiveness), visual representation (3D, image) and method used to achieve desired output. The most popular techniques for gesture synthesis use text or speech signal as input. Text-to-speech systems input text information to determine accompanying gesture types and their timings. Speech-driven animation systems use audio speech

signal as input and generate target behaviors based on features extracted from the input signal. Naturalistic speech carries emotional cues. Therefore, in order to create believable gestures animations for CAs, emotional content of speech should be also taken into account. Generation of expressive gestures using speech-driven systems is limited to the basic catagorical emotions. Overcoming this limitation represents one of our main motivations.

1.2 Goal and Contributions

This thesis presents an affective speech-driven gesture synthesis and animation framework for the expression of affective gestures in conversational agents. We are interested in fully automatic speaker-independent systems.

1.2.1 Goal

Our goal is to automatically generate expressive arm gestures accompanying affective speech.

For this purpose, we propose a system that inputs (1) speech audio data and (2) its affective attributes.

While expressive gesture generation remains as the main focus of thesis, comparison of methods with objective measures and subjective tests for perceptual experiments were also carried out.

1.2.2 Contributions

The main contributions presented within this work:

- We present a data-driven, learning-based computational model for affective speech-driven arm-gesture synthesis and animation.

- We propose the Hidden Semi-Markov Model based gesture model that we use to capture the relationship between gestures, affect and speech prosody.

- We present the unit selection based animation generation system.

- We investigate four models for integrating affect attributes into this general framework so as to generate novel gesture performances from affective speech: prosody-only, affect-only, joint affect and prosody, prosody conditioned on affect.

- We perform extensive experimental comparison of the four approaches using subjective and objective tests.

1.3 Thesis Organization

The remainder of this dissertation is organized as follows:

Chapter 2, Literature Review: presents notable work related to: (1) conversational agents, (2) visual text-to-speech, (3) speech-driven animations, (4) theories of emotion, (5) emotion and gestures. We overview the methods and systems related to these subjects and focus on affect expressiveness in each section.

Chapter 3, Hidden Semi-Markov Model Based Gesture Synthesis and Animation: presents the general framework with components multimodal analysis, gesture synthesis and animation generation. Multimodal datasets, which are used in the experiments, are also introduced.

Chapter 4, Speech Rhythm in Gesture Animation: presents speaker-dependent and independent framework using speech rhythm information during the animation generation process. We also present objective and subjective evaluation methods for the synthesis and animation results.

Chapter 5, Affective Synthesis and Animation of Arm Gestures from Speech Prosody: presents contribution of affect attributes in speech-driven synthesis and animation of gestures by using the following modeling strategies: (1) prosody-only, (2) affect-only, (3) joint affect and prosody, (4) prosody conditioned on affect. We also present objective and subjective evaluation methods for the synthesis and animation results.

Chapter 6, Conclusions: presents a summary and comments on the contributions.

1.4 Publications

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

1. **Multimodal Analysis of Speech Prosody and Upper Body Gestures Using Hidden Semi-Markov Models**, *poster presentation*, Elif Bozkurt, Shahriar Asta, Serkan Özkul, Yücel Yemez, and Engin Erzin, proceedings International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013 [Bozkurt et al., 2013].

2. **Affect-Expressive Hand Gestures Synthesis and Animation**, *poster presentation*, Elif Bozkurt, Engin Erzin, and Yücel Yemez, proceedings International Conference on Multimedia and Expo (ICME) 2015 [Bozkurt et al., 2015a].

3. **Multimodal Analysis of Speech and Arm Motion for Prosody-Driven Synthesis of Beat Gestures**, Elif Bozkurt, Yücel Yemez, and Engin Erzin, *submitted*, Speech Communication, 2015 [Bozkurt et al., 2015b].

4. **The JESTKOD Database: An Affective Multimodal Database of Dyadic Interactions**, Elif Bozkurt, Hossein Khaki, Sinan Keçeci, B. Berker Türker, Yücel Yemez, and Engin Erzin, *accepted for publication*, Language Resources and Evaluation (LREV), 2016 [Bozkurt et al., 2016a].

5. **Affective Synthesis and Animation of Arm Gestures from Speech Prosody**, Elif Bozkurt, Yücel Yemez, and Engin Erzin, *to be submitted* [Bozkurt et al., 2016b].

Chapter 2

LITERATURE REVIEW

This chapter presents a review of existing methods related to non-verbal behavior synthesis, more specifically, gesture synthesis for conversational agents. We will describe state of the art related to this topic, while maintaining a focus on the relation to affect expressiveness.

2.1 Conversational Agents

Conversational agents (CAs), which are used in various number of applications including e-learning, entertainment applications, and also as virtual psychotherapists, virtual presenters, characters in games, have been named under many terms such as virtual humans, avatars, virtual actors etc. We will use the term converstional agent to define the three-dimensional human representation in computer graphics intended to interact with humans.

Interacting with CAs is challenging since every human is an expert at observing details in human motion, facial expression and physical appearance. In addition, this familarity is also a reason for the dilemma in accepting synthesized human behavior and appearance, which known as the concept of *uncanny valley* [Mori, 1970]. According to Thalmann et al., achieving believability of CAs depends on three factors: appearance that is how realistic face and body shapes, skin textures, hair are; motion that is how gestures, facial expressions realistic, smooth, and flexible are; behavior that is how interactive, expressive the agent is [Magnenat-Thalmann and Thalmann, 2005]. These aspects are interdependent and equally important, and decades of research have brought great improvement. For example, photorealistic digital results are obtained by using scanning technologies such as the USC ICT's Light Stage [Debevec,

2012]. Real-time performance based facial animation has been carried to enable users control the facial expression of an agent in real-time [Weise et al., 2011].

Different techniques were introduced for representation of CAs, which made it difficult to collaborate and compare research methods. Therefore, standards have been defined such as FACS (Facial Action Coding System) and MPEG-4 [Ekman and Rosenberg, 1997, Ostermann, 1998]. FACS is a description of movements of facial muscles and jaw, tongue composed of 44 action units (AUs). Facial poses are represented as blendshapes and these blendshapes are interpolated during the animation. MPEG-4 is an object-based multimedia compression standard using facial animation parameters that combine AUs.

Creating believable behaviors for CAs remains as a difficult task. Human communicative skills include speech, coverbal gestures, eye gaze, and facial expressions. CAs can potentially display these non-verbal behaviors. Another fundamental aspect of believable behavior generation is perception and expression of emotion. Expressive behaviors are synthesized by visual text-to-speech or speech-driven methods. A review of visual text-to-speech and speech-driven animation systems are presented in sections 2.2 and 2.3, respectively.

2.2 Visual Text-to-Speech

Text-to-speech (TTS) synthesis generates speech from input text. Likewise, visual text-to-speech implies an additional text-driven animation module that targets to generate and animate body gestures from a tagged text input [Cassell et al., 1994, Cassell et al., 2001, Noma et al., 2000, Reithinger et al., 2006, Pelachaud, 2005, Stone et al., 2004, Neff et al., 2008, Kopp and Wachsmuth, 2004]. Text-driven systems use rule-based approaches to synthesize motion animation, where a set of rules are defined to simulate the desired behavior. Rules are mostly defined according to observations of visual cues and prosody. Generally, a few animation frames are predicted using the predefined rules, while the rest of the frames are obtained by interpolation. One of the rule-based systems was introduced by [Cassell et al., 1994], which automatically

generates and animates conversations of multiple agents accompanyed by speech, facial expressions and hand gestures. Arm, wrist and hand motions are coordinated to create meaningful gestures. Gestures and speech are synchronized at the phoneme, word, and sentence levels. In this system, the mapping from text to gestures is contained in a set of rules derived from nonverbal conversational behavior research, and gestures also collected in a pre-defined gesture dictionary. The Behavior Expression Animation Toolkit (BEAT) [Cassell et al., 2001] can be thought of as a more complex version of the gesture generation system proposed in [Cassell et al., 1994], which uses linguistic and contextual information contained in the speech text to control movements of the hands, arms and face, and intonation of the voice. Another example to text-driven methods is the Virtual Human Presenter of [Noma et al., 2000], which generates gestures using keyword triggered rules. More recent works such as VirtualHuman [Reithinger et al., 2006] and multimodal expressive CAs [Pelachaud, 2005] aim to develop interactive virtual characters with personality profiles and full-body gesture animation by taking into account characters' affective states. A common behavior specification framework was proposed for conversational agent behavior generation [Heylen et al., 2008], which is used by Greta [Niewiadomski et al., 2009]. Greta is capable of both verbal and non-verbal communication.

In contrast to the rule-based methods mentioned above, two related methods [Stone et al., 2004] and [Neff et al., 2008] follow a data-driven approach to generate gesticulation of a particular speaker from speech text. [Stone et al., 2004] use motion graphs to rearrange pre-recorded audio and motion segments, whereas [Neff et al., 2008] develop a probabilistic framework to learn an abstract statistical model for the gesticulation of a speaker from annotated audiovisual data.

Expressive gesture synthesis is a relatively new research area with a major focus on categorical representation of basic emotions. In the current literature, expressiveness of gestures are generally specified by using parameters [Hartmann et al., 2006], [Hartmann et al., 2005], [Pelachaud, 2009], [Niewiadomski et al., 2009], [Mancini et al., 2011], or by selecting appropriate gesture types [Noot and Ruttkay, 2005,Becker et al.,

2004, Becker et al., 2007, Gratch and Marsella, 2001] depending on the emotional state as defined in an input markup text file.

Text-based expressive animation systems generally specify the dynamics of movement based on a set of expressivity parameters for categorical emotions, defined in a markup language. For example, EMOTE is a 3D character animation system that takes an already existing neutral key pose animation and adds some impression of emotions into torso and arm movements [Chi et al., 2000]. However, EMOTE is for modifying but not for generating gestures, where the goal is designing affective expressions by high-level descriptors such as effort and shape transformations. Besides, modifying the expressivity of a gesture may alter the intended meaning in non-verbal communication. Greta on the other hand, is a conversational agent (CA) that communicates through her face and gestures while talking to the user via text-to-speech [Hartmann et al., 2006, Hartmann et al., 2005, Pelachaud, 2009, Niewiadomski et al., 2009, Mancini et al., 2011]. Based on the communicative functions contained in a markup input text, the gesture system chooses a matching prototype gesture to be executed. Baseline expressivity parameters, such as overall activation, spatial extent, temporal extent, fluidity, repetition and power, are defined for gestures, as well as for facial expressions, and the baseline is dynamically modified based on a set of hand-crafted rules depending on the communicative intentions (such as, wide vs. narrow gestures, and smooth vs. jerky movements). One of the comprehensive studies in creating expressive CAs is the SEMAINE system that employs Sensitive Artificial Listeners (SAL) scenario for studying emotional and non-verbal behavior [Schroder et al., 2012]. In this system a behavior lexicon associates communicative functions with the corresponding multimodal signals that a CA can produce.

Categorical emotion states do not only change movement quality, but also the type of movements in text-driven gesture animations. Several gesture synthesis studies have modeled the influence of emotional state on gesture selection. Gestyle is a markup language, which indicates parts of text where the CA must accompany with a gesture [Noot and Ruttkay, 2005]. Emotional states, which are defined as entries in

a style dictionary, are used for designing speech and non-verbal modalities. Several dictionaries can be used together such as, gender, culture, and profession. Max is a CA that uses a direct mapping between emotional states, mood, boredom level and behaviors to modulate gestures, facial expression and speech in time [Becker et al., 2004, Becker et al., 2007]. For example, yawning behavior is triggered for high levels of boredom. Additionally, high arousal emotional state leads to faster speech and associated gestures. Gratch and Marcella investigate impact of the emotional modes on the physical expressions through suitable choice of gestures by using the emotional state of a CA to drive the finite state machine that determines behaviors [Gratch and Marsella, 2001]. Action tendencies and type of non-verbal behaviors are linked to the emotional state.

2.3 Speech-driven Animations

In addition to text-driven approaches, speech-driven non-verbal behavior synthesis systems have been proposed, which utilize prosody features as input. Acoustic prosody, which refers to speech characteristics that do not portray vowels and consonants but longer units of speech like the syllables, reflects various features like the speaker's identity, emotional state, gender, or age. Prosody features can be used to determine how people are saying rather than what people are saying. Such set of features include: pitch, energy, rhythm, speech rate, pauses etc. These features are called as *suprasegmental* since these features are above the level of phoneme.

Prosody is crucial for spoken interaction and has been employed for non-verbal behavior synthesis purposes. An extensive study of prosody functions such as turn-taking, emotion and attitude are provided in [Xu, 2011]. Relation of prosody with eyebrow movements [Ding et al., 2013], head movements [Sargin et al., 2008, Busso et al., 2007], facial expresions [Busso and Narayanan, 2007, Mariooryad and Busso, 2012, Graf et al., 2002, Hong et al., 2002, Albrecht et al., 2002, Chuang and Bregler, 2005], and gestures [Levine et al., 2010, Baena et al., 2013, Chiu and Marsella, 2014, Bozkurt et al., 2013] have been investigated. Ding et al. investigated statistical

Markovian systems that take into account speech features as contextual information towards synthesis of eyebrow motion [Ding et al., 2013]. Granstrom and House investigated audio-visual prosody for creating a CA capable of displaying realistic head and eyebrow movements [Granström and House, 2006]. They tested the perceptual sensitivity of prominence and feedback to the timing of both eyebrow and head movement in relation to the syllable. The experimental results indicated that combined head and eyebrow movements are quite useful cues to prominence when synchronized with the stressed vowel. Busso and Narayanan investigate interrelation between facial expressions and prosody [Busso and Narayanan, 2007]. Their study reveals that the lower face region provides the highest activeness and correlation rates between the two modalities. The authors also show that the emotional content influences the relationship between facial expressions and speech, where a small set of prosodic features are used for emotion-dependent structures in the audio-visual mapping. They also mention that facial gestures and speech features are coupled at different time scales. For example, while lips are locally tightly coupled with the speech, the eyebrow and head motion are coupled at the sentence level. Sargin et al. introduced a two-stage framework for joint analysis of head gestures and speech prosody patterns towards synthesis of head gestures from speech prosody [Sargin et al., 2008]. [Busso et al., 2007] present an approach to synthesize emotional head motion sequences driven by prosodic features, that builds hidden Markov models for emotion categories to model the temporal dynamics of emotional head motion sequences. Mariooryad and Busso explored joint models of speech and facial expressions to preserve the coupling between the modalities. Various coupling models for head, eyebrow motions and speech are proposed and speech-driven facial animations are generated [Mariooryad and Busso, 2012]. The authors emphasize the interconnection of prosodic features and facial expressions based on the perceptual evaluations on the animation results, which are generated by using the joint models. Graf et al. show that head movements and facial expressions are well synchronized with main prosodic events such as pitch accents, phrase boundaries [Graf et al., 2002]. Albrect at al. use prosodic features for facial

expression generation from speech [Albrecht et al., 2002]. [Chuang and Bregler, 2005] describe a method for creating expressive facial animation based on a statistical model learned from video for factoring the expression and lip speech. They also integrate head motion synthesis to their face animation scheme by first building a database of examples which relate audio pitch to motion and then matching new audio streams against segments in this database.

Speech-driven gesture animation has been studied in the recent literature mostly without specific emphasis on affect-expressive models. In [Levine et al., 2010], Levine et al. have introduced *gesture controllers*, availing a modular methodology to drive beat-like gestures with live speech via customized gesture repertoires. They have a two-stage model, where in the first stage prosodic speech features are related to gestures, and then the gestures are related to motion features in the second stage. Gesture controllers infer hidden states from speech using a conditional random field that analyzes acoustic features in the input and select the optimal gesture kinematics based on the inferred states. Later, Baena et al. present a single stage model that links speech prosody to beat gestures based on manually annotated body motion and speech signals [Baena et al., 2013]. They employ motion graphs to generate appropriate gestures with varying emphasis for a given speech input to model aggressive and neutral performances. Chiu and Marsella employ a two step approach for speech-driven gesture animation [Chiu and Marsella, 2014]. They use conditional random fields (CRF) for mapping speech to gesture annotations. Then, Gaussian process latent variable models (GPLVMs), which represent gestures in a low-dimensional space, are used for motion synthesis. Finally, an interpolation algorithm models coarticulation of gestures.

Some recent studies utilize semantics in addition to prosody for non-verbal behavior synthesis. Marsella et al. consider agitation level and word stress of sentence audio to drive their rule-based character animation system that generates gestures including facial expressions, hand motion, head movements, eye saccades, blinks and gazes [Marsella et al., 2013]. Although their method produces promising results, it

requires significant setup of appropriate gesture motions for the gesture database. Sadoughi and Busso propose a hybrid system for head and hand gesture synthesis, combining the speech-driven and the rule-based system components [Sadoughi and Busso, 2015]. They detect similar gestures and study their relation with the semantic functions in the message. Then, a speech-driven system retrieves gesture samples for synthesizing behaviors constrained by the target gesture.

2.4 Theories of Emotion

In the psychology literature, three approaches have been introduced for modelling affect: categorical, dimensional and appraisal-based approaches [Grandjean et al., 2008]. The categorical approach mostly assumes a small number of universal emotion classes such as, anger, disgust, fear, happiness, sadness, and surprise [Ekman and Friesen, 1975]. However, this approach is limited since, humans have the ability to perceive and express more complicated affective states like depression [Russell, 1980]. On the other hand, dimensional approaches represent affect on a continous scale, which is useful for explaining non-categorical, complex affective states and also affective state transitions during human interactions. Dimensional approaches mostly concentrate on representation of affect in the two dimensional space as in activation and valence domains. Additionally, some researchers also include the third dimension as the dominance domain into their analysis. In these models, affect is described along the active-passive, positive-negative and dominant-submissive dimensions [Mehrabian, 1996]. A detailed overview on the continuous repsentation of affect can be found in [Gunes et al., 2011]. The third approach claims that emotions are elicited by continuous subjective evaluations (appraisals) of the outside world [Scherer et al., 2001]. However, it is still an open research area how to use the appraisal based affect approach for automatic measurement of affect, since it is challenging to evaluate such a complicated, multi-partite signal.

2.5 Emotion and Gestures

Neuroscientific and psychological studies reveal that body movement and its expressivity constitute an important modality of emotion communication [Tracy and Robins, 2007,De Gelder, 2006]. For example, fear brings to contract the body to be as small as possible, surprise brings to turn towards the object capturing our attention, joy may bring to openness and upward acceleration of the forearms [Boone and Cunningham, 1998]. Wallbott collected video recordings in which actors portrayed emotions within defined scenarios and observed discriminative features of emotions both in static descriptors of the body posture and in the body movement quality [Wallbott, 1998]. More specifically, Kipp and Martin investigated the relationship between basic gestural forms (handedness, hand shape and motion direction) and emotion, where the authors report that handedness is closely correlated with emotion categories [Kipp and Martin, 2009]. Similarly, Dael et al. also suggested that emotional attributes are associated with specific spatio-temporal characteristics perceived in arm gesture movements [Dael et al., 2013]. Within this view, processing of non-verbal signals for conversational agents can be translated into valuable information conveying affective messages. Karg et al. present an extensive survey on body movements for affective expression in [Karg et al., 2013].

Chapter 3

HIDDEN SEMI-MARKOV MODEL BASED GESTURE SYNTHESIS AND ANIMATION

3.1 Introduction

Gesticulation is an essential component of human communication. Speech and gestures form a composite communicative signal that boosts the naturalness and affectiveness of the communication. Although virtual environment designs in the human-computer interaction (HCI) field are increasingly adopting and emphasizing the human-centered aspect, a natural, affective and believable gesticulation is often missing in the virtual character animations. In this context, automatic synthesis of gesticulation in synchrony with speech, which incorporates nonverbal communication components into virtual character animation, can help improving the plausibility of animations and can find a wide range of applications in human-centered HCI, video gaming and film industries. We develop a multimodal system for speech-driven synthesis and animation of arm gestures using a statistical framework for joint analysis of speech and gesticulation.

Gesture and speech co-exist in time with a tight synchrony; they are planned and shaped by the cognitive state and produced together. In one of the pioneering studies on gesture and speech relationship, [Kendon, 1980] proposed a widely accepted hierarchical model for gesture in terms of phases, phrases and units. In this model, the core gestural element is defined as gesture phase. Gesture phases can be active or passive. An active gesture phase can be a stroke (a short and dynamic peak movement) with a retraction or a preparation (in which arm goes to the start position of the stroke phase). Passive gesture phases are movements like hold and rest, in

which arm stays motionless. Combinations of phases constitute gesture phrases, and then combinations of phrases form gesture units. In this hierarchical model, semantic expressiveness increases with the level of hierarchy. In other words, gesture units are semantically more expressive than gesture phrases, and gesture phrases include more semantic content than gesture phases.

Synchrony between gestural and phonological structures has previously been studied by various researchers [Wagner et al., 2014]. [Kendon, 1980] points out the synchrony between strokes and stressed syllables. Later [McNeill, 1992] proposes the widely accepted phonological synchrony rule: the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech. [Valbonesi et al., 2002] investigate the nature of temporal relationship between speech and gestures. In a recent study, [Loehr, 2012] presents a detailed investigation of temporal and structural synchrony between intonation and gesture. His findings verify the alignment of pitch accents with gestural strokes; furthermore, he presents evidences of the synchrony between gesture phrases and intermediate intonation phrases.

There are four widely referred types of gestures, which were proposed by [McNeill, 1992]: iconics, metaphorics, deictics and beats. Iconic gestures illustrate images of objects or actions, 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, based on the Tuite's proposal [Tuite, 1993] that in every gesture there is a rhythmical beat-like pulse to carry significance beyond its immediate setting, [McNeill, 2006] suggests taking metaphoricity, iconicity, deixis, and emphasis as dimensions of gesture rather than types of gesture.

Although there seems to exist strong correlation between gestures and speech, there are several challenges and difficulties involved in modeling this relationship. The first challenge is due to the diversity of gestures related to speech semantic content. Iconic, metaphoric and deictic gestures belong to this category which are mainly related to semantic content. We exclude modeling these gestures and rather focus on modeling beat gestures in relation to speech prosody . There are yet two

main challenges in achieving this goal. The first is due to the difficulty in mapping local prosody information to relatively longer duration semantic gestures, e.g., gesture phrases. The second challenge is actually a temporal alignment problem: prosodic cues and corresponding gestures do not always co-occur at exactly the same time instant, and there may indeed be a lagged correlation in between.

3.2 System Overview

The general block diagram of our speech-driven gesture synthesis and animation system is given in Figure 3.1. The system consists of three main tasks: analysis, synthesis, and animation. Within the analysis task we have two stages: (i) feature extraction and clustering of gesture phrases and prosodic units, and (ii) their multimodal analysis. Section 3.3 presents the first stage of the analysis task, where we perform feature extraction and unimodal clustering on speech and body motion data. The audio stream is processed to extract prosodic features of the speech, whereas the body motion data is expressed in terms of 3D joint angles. The gesture and audio feature streams are then segmented via temporal clustering into recurrent patterns, i.e, into gesture phrases and prosodic units, respectively. Section 3.4 describes the second stage of the analysis task, where we use hidden semi-Markov models (HSMM) to model the dependencies between speech prosody and gestures in a multimodal framework. Sections 3.5 and 3.6 mainly address the synthesis and animation tasks, respectively. In Section 3.5, we describe how an optimal sequence of gestures with durations is synthesized for a given speech input by using the Viterbi algorithm over HSMM. In Section 3.6, the synthesized gesture sequences with duration information are mapped into body motion sequences by using a multiple objective unit selection algorithm to generate animations.

3.3 Feature Extraction and Unimodal Clustering

We employ semi-supervised and unsupervised temporal clustering schemes to determine boundaries and categories of gesture phrases for speaker-dependent and inde-

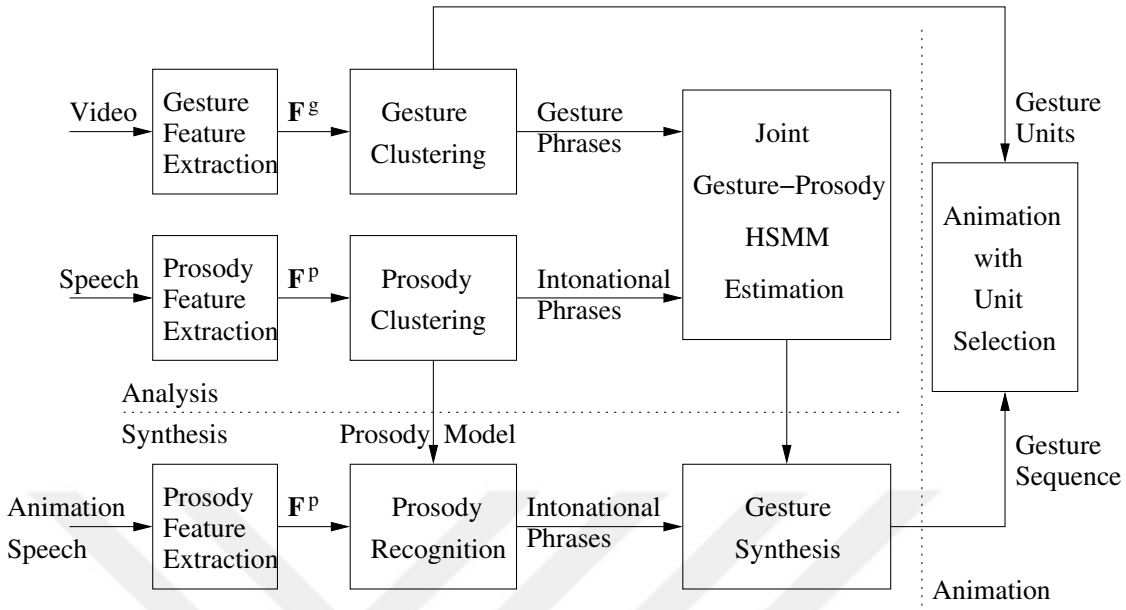


Figure 3.1: The block diagram of the general framework for the speech-driven gesture synthesis and animation system.

pendent scenarios, respectively. We use an unsupervised scheme to cluster speech into prosodic units.

3.3.1 Prosody Clustering

Characteristics of the prosody at the acoustic level, including intonation, rhythm, and intensity patterns, carry important temporal and structural synchrony with gesture phrases [Loehr, 2012]. Prosody has a marked effect on suprasegmental features such as pitch, energy, and timing in the vicinity of an prosodic event [Ananthakrishnan and Narayanan, 2008]. As we target to extract prosodic units through unsupervised clustering we choose to use pitch and energy related acoustic features to model the speech prosody. Note that here we define a prosodic unit as a speech segment with a recurrent prosodic pattern. We choose to include speech intensity, pitch, and confidence to pitch into the prosody feature vector. Prosody features are extracted over 50 ms analysis windows with 25 ms frame shifts. Speech intensity is defined as the

logarithm of the average signal energy in the analysis window,

$$I_k = \log(\frac{1}{W}\sum_{i=1}^{W} s_k[i]^2), \tag{3.1}$$

where s_k is the speech signal in the kth window, and W is the window size.

Pitch is extracted using the YIN fundamental frequency estimator, which is a robust pitch frequency estimator based on the well-known auto-correlation method [de Cheveigne and Kawahara, 2002]. Pitch feature, ν_k, is defined as logarithm of the fundamental frequency at the kth frame. The YIN estimator defines a difference function based on the auto-correlation function,

$$e_k(\tau) = \sum_{i=1}^{W}(s_k[i] - s_k[i+\tau])^2. \tag{3.2}$$

We define the confidence to pitch feature based on the normalized difference function as,

$$c_k = 1 - \frac{e_k(\tau^*)}{\frac{1}{\tau^*}\sum_{i=1}^{\tau^*} e_k(i)}, \tag{3.3}$$

where τ^* is the pitch lag corresponding to the fundamental frequency.

Since the prosody feature values are speaker and utterance dependent, we apply a mean and variance normalization to the prosody features. We compute the mean and variance of prosody features for each speech utterance, and perform mean and variance normalization to get the normalized prosody features $\bar{I}_k$, $\bar{\nu}_k$, and $\bar{c}_k$. Then the normalized intensity, pitch, confidence to pitch features and the first temporal derivative of these three parameters are used to define the prosody feature vector at frame k,

$$\boldsymbol{f}_k^p = [\bar{I}_k,\, \bar{\nu}_k,\, \bar{c}_k,\, \Delta\bar{I}_k,\, \Delta\bar{\nu}_k,\, \Delta\bar{c}_k], \tag{3.4}$$

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

We employ unsupervised temporal clustering using the parallel HMM architecture in [Sargin et al., 2008] to extract prosody clusters. 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 prosodic units through unsupervised learning. 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 HMM based unsupervised clustering process segments the prosody feature stream into prosodic units. We denote the lth prosodic unit in the stream by ε_l^p and the associated class label of the lth unit by ℓ_l^p, which is one of the M_p available prosody classes $\{p_1, p_2, ..., p_{M_p}\}$. A frame level label sequence is then defined for the prosodic unit sequence,

$$\xi_k = \ell_l^p \quad \text{for } k = k_l, k_l + 1, \ldots, k_{l+1} - 1, \tag{3.5}$$

where ξ_k is the prosody label of the kth speech frame and $[k_l, k_{l+1})$ spans the lth prosodic unit. These frame level prosody labels will eventually serve as the observations of the HSMM structure that we will describe in Section 3.4.

3.3.2 Gesture Clustering

We model gestures, specifically beat type arm gestures, at gesture phrase level to correlate with and emphasize speech prosody. For analysis of gestures, we employ joint angles from four body parts: left arm, left forearm, right arm, and right forearm as shown in Figure 3.2. We define the joint angle vector for the ith joint at frame k as $\theta_k^i = [\phi_x^{ik}, \phi_y^{ik}, \phi_z^{ik}]$, where $\phi_x^{ik}, \phi_y^{ik}, \phi_z^{ik}$ are the Euler angles respectively in the x, y, z directions, representing the orientation of the ith joint at frame k in degrees. Then, we define the gesture feature vector at frame k, $\boldsymbol{f}_k^{Ji}$, to include the joint angles from the ith body part and their first order derivatives,

$$\boldsymbol{f}_k^{Ji} = [\theta_k^i, \Delta\theta_k^i], \quad \text{for } i = 1, 2, 3, 4, \tag{3.6}$$

where $\Delta\theta_k^i$ denotes the first order derivative of the joint angle vector θ_k^i. The resulting gesture feature for the four joints at time frame k is defined as,

$$\boldsymbol{f}_k^g = [\boldsymbol{f}_k^{J1}, ..., \boldsymbol{f}_k^{J4}]. \tag{3.7}$$

Figure 3.2: The black circles correspond to the joints used for gesture representation as left forearm, left arm, right arm, and right forearm, respectively.

Semi-supervised Gesture Clustering

Temporal clustering of the gesture feature sequence is necessary for analysis of recurring beat gesture phrases. In the speaker-dependent case, we implement a semi-supervised clustering method using the parallel branch HMM structure, Λ^g, over the gesture feature stream $\boldsymbol{F}^g = \{\boldsymbol{f}_1^g, \boldsymbol{f}_2^g, ..., \boldsymbol{f}_T^g\}$, with duration of T frames. 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 $N_g = 10$ states corresponding to the minimum gesture phrase duration of 10 frames ($\frac{1}{3}$ seconds assuming 30 video frames/sec). The number of branches is iteratively increased to M_g in a semi-supervised manner using the following procedure:

(i) Initially set Λ^g to have two branches to model the *rest* position of arms and all the other remaining arm movements. Manually label examples of the rest event (as many as necessary to train an HMM structure) from gesture stream by inspecting the video.

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

(iii) Perform Viterbi decoding to get temporal clusters.

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

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

(vi) Manually 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).

The proposed semi-supervised clustering process segments the gesture feature stream into gesture phrases, ε_l^g, with label ℓ_l^g as one of the M_g available gesture classes $\{g_1, g_2, ..., g_{M_g}\}$. All gesture phrases ε^{g_i} from gesture class g_i are grouped together to build a gesture pool $G_i = \{\varepsilon^{g_i 1}, ..., \varepsilon^{g_i j}, ..., \varepsilon^{g_i N_i^G}\}$, where $\varepsilon^{g_i j}$ is the jth gesture phrase in the pool, and N_i^G is the number of gesture phrases in the gesture pool G_i.

Unsupervised Gesture Clustering

Unsupervised gesture clustering is applied for the speaker-independent setting, where a large scale multimodal dataset has been used for this purpose. Unlike the clustering results of the semi-supervised approach in Section 3.3.2, the resulting gesture patterns of unsupervised clustering are not explicitly compatible with the gesture phrase definition presented in [Kendon, 1980]. However, there is reported evidence that these patterns are meaningful for explaining the nature of gestures ([Yang et al., 2014], [Yang and Narayanan, 2016]). For simplicity, we will use the same notation with the gesture phrases in Section 3.3.2, as ε^{g_i} for the gesture patterns resulting from gesture class g_i.

Similar to the prosody clustering process in Section 3.3.1, we apply the unsupervised clustering method based on parallel-HMMs [Sargin et al., 2008] to the gesture feature sequence $\boldsymbol{F}^g = \{\boldsymbol{f}_1^g, \boldsymbol{f}_2^g, ..., \boldsymbol{f}_T^g\}$, with duration of T frames. We segment and cluster gesture sequences into gesture patterns with duration information. The HMM structure Λ^g is set to have M_g parallel branch HMMs, $\{\lambda_1^g, ..., \lambda_{M_g}^g\}$, where each λ_m^g is composed of $N_g = 10$ states corresponding to the minimum gesture pattern duration of 10 frames ($\frac{1}{3}$ seconds assuming 30 video frames/sec) considering the works by [Yang et al., 2014] and [Bozkurt et al., 2015a]. We also create a gesture pool G_i for each

gesture class g_i.

3.4 Multimodal Analysis of Gestures and Prosody

In this section we construct a multimodal analysis framework to model the relationship between beat gestures and speech prosody at the gesture phrase level. In general, prosodic units are much shorter than gesture phrases in duration, and the stroke of a gesture phrase precedes or ends at, but does not follow, the phonological peak syllable of speech as [McNeill, 1992] stated. A gesture phrase sequence, when accompanied by a sequence of prosodic units, forms a Markov random process. One useful mathematical model for multimodal analysis of gesture phrases and prosodic units can be constructed by taking gesture phrases as the states of a Markov chain and prosodic units as the observations of this Markov process. Hence, state transitions correspond to articulation of consecutive gesture phrases, and gesture phrases can be located in time according to the McNeill's phonological rule by observing prosodic units.

Synthesizing a gesture phrase sequence using the conventional Markov chain model given prosodic unit observations would however have a shortfall in modeling gesture phrase durations in time. A useful mathematical model to overcome this shortfall is to introduce a statistical state duration model so that one can better control gesture phrase durations in the synthesis process. Combination of these two useful mathematical models, i.e., Markov chain and state duration, yields the hidden semi-Markov model (HSMM) framework [Yu, 2010] that we use for multimodal analysis of gesture phrases and prosodic units. 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 [Barbu and Limnios, 2008]. Figure 3.3 shows how such an HSMM structure functions, where gesture phrase labels and frame-level prosody labels are depicted as states and observations, respectively. Gesture transitions define state transitions, whereas prosodic unit distributions per gesture class define the observation emission distributions. The state duration distributions are estimated over the gesture phrase duration informa-

tion. Note that the observations of the HSMM structure are frame-level prosody labels defined in (3.5). This is essential since in this way, hidden state durations, hence gesture durations, can be represented in terms of fixed length speech frames.

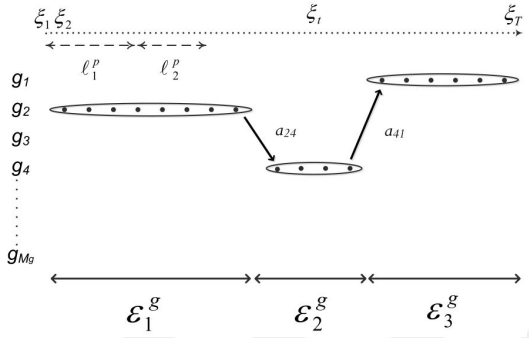


Figure 3.3: In the hidden semi-Markov model, prosodic units are labeled as ℓ_i^p, frame-level prosody labels (ξ_t) correspond to observations and gesture phrases (ε_k^g) correspond to states (g_m).

An HSMM representing frame-level prosody labels as observations with M_g fully connected states is modeled by $\Lambda^{gp} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{\Pi})$. The states of Λ^{gp} represent gesture classes, and the model parameters $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{D}$, $\boldsymbol{\Pi}$ respectively stand for 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}, each representing the state transition probability from gesture class g_i to g_j,

$$\boldsymbol{A} : \{a_{ij} = P(\ell_l^g = g_j | \ell_{l-1}^g = g_i)\} \quad i, j = 1, ..., M_g, \tag{3.8}$$

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

$$\boldsymbol{B} : \{b_i(p_j) = P(p_j | \ell_l^g = g_i)\} \quad i = 1, ..., M_g, \ j = 1, ..., M_p, \tag{3.9}$$

where $b_i(p_j)$ is the probability of observing prosodic unit p_j at gesture class g_i. The state duration distribution $\boldsymbol{D}$ is formed in terms of state dependent duration proba-

bility mass functions,

$$\boldsymbol{D} : \{d_i(n)\} \quad i = 1, ..., M_g,\ n = 1, ..., \frac{D_{max}}{\delta}, \tag{3.10}$$

where $d_i(n)$ is the probability of a gesture phrase from gesture class g_i lasting $n\delta$ sec, D_{max} is the maximum duration among all gesture phrases, and δ is the histogram bin size for the underlying probability mass function. In our experiments, we take the maximum duration as $D_{max} = 10$ s, and the histogram bin size as the speech frame duration $\delta = 25$ ms. The initial state probability vector $\mathbf{\Pi}$ is defined by entries π_i, each representing the probability of starting with gesture class g_i as the first gesture phrase,

$$\mathbf{\Pi} : \{\pi_i = P(\ell_1^g = g_i)\} \quad i = 1, ..., M_g. \tag{3.11}$$

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

$$\pi_i = P(\ell_1^g = g_i) \mathrel{\widehat{=}} \frac{C(1, i, j)}{\sum_{j'} C(1, i, j')}, \tag{3.12}$$

$$a_{ij} = P(\ell_l^g = g_j | \ell_{l-1}^g = g_i) \mathrel{\widehat{=}} \frac{\sum_l C(l, i, j)}{\sum_l \sum_{j'} C(l, i, j')}, \tag{3.13}$$

$$b_i(p_j) = P(p_j | \ell_l^g = g_i) \mathrel{\widehat{=}} \frac{O(i, j)}{\sum_{j'} O(i, j')}, \tag{3.14}$$

$$d_i(n) \mathrel{\widehat{=}} \frac{H(i, n\delta \leq t < (n+1)\delta)}{\sum_{n'} H(i, n'\delta \leq t < (n'+1)\delta)}, \tag{3.15}$$

where $C(l, i, j)$ is the number of times g_i appears as the label of the lth gesture phrase and g_j as the label of $(l+1)$st gesture phrase, $O(i, j)$ is the frame count of prosodic unit p_j at gesture class g_i, and $H(i, n\delta \leq t < (n+1)\delta)$ is the number of occurrences of gesture class g_i with duration t in $[n\delta,\ (n+1)\delta)$ interval.

3.5 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 prosodic unit labels, $\{\xi_1, \xi_2, \ldots, \xi_T\}$ (see (3.5)).

Note that the decoded optimal state sequence delivers a synthesized label sequence for gesture phrases, $\hat{\boldsymbol{\ell}}^g$, and a sequence of associated durations, $\boldsymbol{\kappa}$, where 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,

$$\psi_t(j) = \max_{\tau} \max_{i} \{\psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(\xi_k))\}, \tag{3.16}$$

where $\psi_t(j)$ is the accumulated logarithmic likelihood at time frame t in state g_j after observing prosody 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 to extract the optimal state sequence:

i. Initialize

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

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

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

   Repeat for $j = 1, 2, \ldots, M_g$

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

   for $i = 1, \ldots, M_g,\ \tau = 1, \ldots, T'$

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

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

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

iii. Backtrace the optimal gesture phrase sequence

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

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

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

While $t > 0$

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

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

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

The extracted optimal state sequence defines the optimal gesture label sequence $\hat{\boldsymbol{\ell}}^g = \{\hat{\ell}_1^g, \ldots, \hat{\ell}_L^g\}$ and the associated gesture phrase durations $\boldsymbol{\kappa} = \{\kappa_1, \ldots, \kappa_L\}$.

3.6 Gesture Animation

Animation of the synthesized gesture label sequence consists of three main tasks: Extraction of gesture phrase sequence with unit selection, smoothing gesture-to-gesture transitions, and finally graphical animation of the gesture phrase sequence.

The first task is to generate a synthesized sequence of gesture phrases, $\hat{\boldsymbol{\varepsilon}}^g$, given the synthesized gesture phrase label sequence $\hat{\boldsymbol{\ell}}^g$ along with the corresponding duration sequence $\boldsymbol{\kappa}$. This task is performed using unit selection over a pool of gesture phrases which are extracted during the gesture analysis in Section 3.3.2. The next task is to smooth joint angle discontinuities over a temporal window at gesture phrase boundaries, that is, at the boundary of each two consecutive synthesized gesture phrases $\hat{\varepsilon}_l^g$ and $\hat{\varepsilon}_{l+1}^g$, to extract a smoothed gesture sequence $\tilde{\boldsymbol{\varepsilon}}^g$. The smoothed gesture phrase sequence $\tilde{\boldsymbol{\varepsilon}}^g$ is finally used to animate beat gestures of a virtual character in synchrony with the input speech.

We employ a unit selection algorithm to generate the synthesized sequence of gesture phrases, $\hat{\boldsymbol{\varepsilon}}^g$, based on the gesture pool $G_i = \{\varepsilon^{g_i 1}, \varepsilon^{g_i 2}, \ldots, \varepsilon^{g_i N_i^G}\}$, which is constructed in Section 3.3.2 for each gesture phrase class g_i. The unit selection process is demonstrated in Figure 3.4 as the formation of an optimal gesture phrase sequence from the templates available in the gesture pools.

We minimize a mixture of penalty scores for duration mismatch and joint angle continuity mismatch during the unit selection process. The duration, joint angle

continuity mismatch penalties of a gesture phrase template $\varepsilon^{g_i j}$ for a synthesized gesture phrase with label $\hat{\ell}_l^g$ are respectively defined as,

$$D_\omega(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = ||\omega_e(\hat{\varepsilon}_{l-1}^g) - \omega_b(\varepsilon^{g_i j})||, \tag{3.17}$$

$$D_\kappa(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = ||\kappa_l - \kappa(\varepsilon^{g_i j})||, \text{and} \tag{3.18}$$

$$\tag{3.19}$$

where κ_l is the duration of the synthesized gesture phrase with label $\hat{\ell}_l^g$, $\kappa(\varepsilon^{g_i j})$ is the duration of the gesture phrase template $\varepsilon^{g_i j}$ in gesture pool G_i, $\omega_e(\hat{\varepsilon}_{l-1}^g)$ is the ending joint angle vector of the previously synthesized gesture phrase $\hat{\varepsilon}_{l-1}^g$, and $\omega_b(\varepsilon^{g_i j})$ is the beginning joint angle vector of the gesture phrase template $\varepsilon^{g_i j}$. The joint penalty score is defined as a mixture of three penalty scores for duration and joint angle:

$$\begin{aligned} D(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = \; & \alpha\overline{D_\omega}(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) + \\ & (1-\alpha)\overline{D_\kappa}(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i)] \end{aligned} \tag{3.20}$$

where $\overline{D_\omega}$, $\overline{D_\kappa}$ are min-max normalized penalty functions, and α and (1-α) are the mixture weights which we set experimentally over a validation set, as will be explained in Section 4.5.2.

The optimal path minimizing the above penalty score can be extracted by using the following Viterbi algorithm:

i. Initialization
   $V_1(j) = D(\varepsilon^{g_i j}|\hat{\ell}_1^g = g_i)$, for $j = 1, 2, ..., N_i^G$

ii. Recursion: Repeat for $l = 2, 3, ..., L$,
   Let $\hat{\ell}_{l-1}^g = g_{i'}$, for $j = 1, 2, ..., N_i^G$
   $V_l(j) = \min_{j'=1,...,N_{i'}^G}\{V_{l-1}(n) + D(\varepsilon^{g_i j'}|\hat{\ell}_l^g = g_i\}$,
   $Q_l(j) = \arg\min_{j'=1,...,N_{i'}^G}\{V_{l-1}(n) + D(\varepsilon^{g_i j'}|\hat{\ell}_l^g = g_i\}$,

iii. Backtrace the optimal path

$q_L = \arg\min_j \{V_L(j)\}$,
$q_l = Q_{l+1}(q_{l+1})$ for $l = L-1, L-2, ..., 1$,

iv. Construct the synthesized sequence of gesture phrases
$\hat{\varepsilon}_l^g = \varepsilon^{\hat{\ell}_l^g q_l}$ for $l = 1, 2, \ldots, L$.

The selected gesture phrases are resampled so as to fit the synthesized duration if necessary. Next, we smooth joint angle discontinuities at gesture boundaries over a temporal window. This is achieved by applying an exponential smoothing function on each pair of consecutive synthesized gesture phrases $\hat{\varepsilon}_l^g$ and $\hat{\varepsilon}_{l+1}^g$. Then, the smoothed gesture motion sequence $\tilde{\boldsymbol{\varepsilon}}^g$ is used to animate a virtual character based on the four skeleton joints mentioned in Section 3.3.2. The other joints of the body (e.g. spine, lower-body joints) are assumed to have no motion. For graphical animation, we use the MotionBuilder 3D Character Animation Software [Mot, 2012].

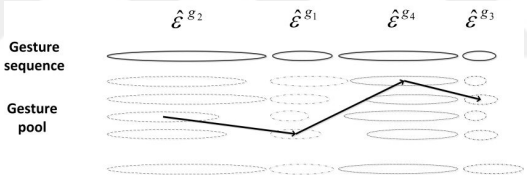


Figure 3.4: Unit selection based gesture animation generation: an optimal sequence of gesture phrases is formed from the gesture phrase templates available in the gesture pool for each gesture class.

3.7 Multimodal Datasets

sec:dataset) We use two datasets, one for speaker-dependent setting and the other for speaker-independent setting, respectively.

3.7.1 Dataset for Speaker-Dependent Setting

The multimodal MVGL-MUB dataset that we use to train our speaker-dependent, speech driven animation system consists of five recordings of a male native speaker

with a total duration of approximately 20 minutes, all in Turkish [Bozkurt et al., 2013]. We collect multiview video of the speaker using four synchronized cameras. The speaker wears a black suit with 15 color markers and a microphone placed close to mouth and synchronized with the cameras. We estimate the 3D positions of the body joints based on the markers' 2D positions tracked on each camera's image plane using color information [Ofli et al., 2008]. The resulting set of 3D points for joints' positions is then converted to a set of Euler angles extracted for each joint in its local frame using inverse kinematics. The motion capture data is recorded at 30 frames per second, and the audio signal is captured in PCM 44.1 kHz 16-bit stereo recording format. Five recording sessions are organized, where the speaker talks in standing pose under five different scenarios: i) telling a recollection of a past memory, ii) telling a fairy tale, iii) talking about a short documentary after watching it, iv) discussing on a spontaneous topic with a second participant, and v) commenting on series of photographs. During the recording sessions, the speaker does not receive any instructions on how to gesture or express himself.

Since gestures are in general person specific, the gesture phrases are determined by using the semi-supervised training scheme described in Section 3.3.2. We have identified the number of distinct gesture phrases as $M_g = 7$ for the given participant. A brief description of these gesture classes is provided in Table 3.1, whereas their distribution per recording is summarized in Table 3.2. In our experiments, we have spared the fourth recording as the validation set, and performed a leave-one-out training procedure such that one recording out of four is used for testing and the remaining three are used for training in four turns. The models resulting from training are used for synthesizing and animating gestures over the test recordings. Then, we perform subjective evaluations to assess naturalness and audio-visual synchrony of the animation results of the test recordings.

Table 3.1: Gesture phrases identified via semi-supervised clustering

Gesture	**Description**
g_1	Symmetric: Both arms move symmetrically
g_2	Rest: No motion
g_3	Left: Only left arm moves
g_4	Asymmetric: Both arms move asymmetrically
g_5	Contact: Hands touch to each other
g_6	Open: Stretching arms backwards
g_7	Right: Only right arm moves

Table 3.2: Gesture phrase distributions per recording

Rec.	**Gesture Phrase Counts**							**Total**	**Dur.**
ID	g_1	g_2	g_3	g_4	g_5	g_6	g_7	**Count**	**(s)**
(i)	52	64	9	22	1	0	19	167	239
(ii)	20	40	1	8	0	17	6	92	167
(iii)	22	49	1	23	10	21	40	166	265
(iv)	53	60	15	20	4	18	20	190	347
(v)	2	45	1	0	0	0	0	48	155
Total	149	258	27	73	15	56	85	663	1173

3.7.2 Dataset for Speaker-Independent Setting

In the speaker-independent animation system, we use the multimodal USC CreativeIT dataset that contains a variety of dyadic theatrical improvisations for studying expressive behaviors and natural human interaction [Metallinou et al., 2010], [Metallinou et al., 2015]. In this dataset the interactive performances are designed either as improvisations of scenes from theatrical plays or as theatrical exercises where actors repeat sentences in a manner that conveys specific intent such as, accepting or rejecting behavior towards the other.

The dataset contains vocal and body-language behavior information of the actors obtained through close-up microphones, Motion Capture (MoCap) and HD cameras. The MoCap data is provided as 3D coordinates of 45 marker positions in (x,y,z) directions at 60 fps and speech recordings at 48 kHz for each of the 16 (9 female) distinct actors. We use the MotionBuilder software [Mot, 2012] for converting 3D joint positions to Euler angle rotations of the arm and forearm in the (x,y,z) directions at 30 fps. We perform speaker-independent evaluations in a leave-one-pair-out manner using data from one actor-pair as the test set in turn, and the remaining data from the other pairs as the training data.

3.8 Summary

This chapter described a new multimodal analysis framework for modeling the relationship between intonational and gesture phrases using the hidden semi-Markov models (HSMMs). The proposed method effectively associates longer duration gesture phrases to shorter duration prosody clusters, while maintaining realistic gesture phrase duration statistics that leads to animations perceived as realistic.

Section 3.3 presented feature extraction and clustering steps for prosody and gesture motion features from audio-visual data. In Section 3.4, we presented joint modeling of intonational and gesture phrases using the hidden semi-Markov models. Section 3.5 presented gesture synthesis from a novel audio recording. The output of this step is synthesized gesture sequence with id and duration information. In

Section 3.6, we presented a unit selection based gesture animation algorithm, which inputs systhesis step output components.

Chapter 4

SPEECH RHYTHM IN GESTURE ANIMATION

4.1 Introduction

In a natural speaking style, beat gestures are articulated in synchrony with prosody and rhythm to emphasize the underlying speech [McNeill, 1992, Valbonesi et al., 2002, Loehr, 2012]. In this chapter, we construct a multimodal analysis framework to model the relationship between beat gestures, speech prosody and rhythm. Studies in diverse research areas suggest that human audio-visual communication is significantly rhythmic in nature, for example, in the way how spoken syllables and words are grouped together in time as in speech rhythm [Bolt, 1980, Ladd, 1996, Liberman, 1975] or how they are accompanied by body movements as in beat gestures [Tuite, 1993, Bos et al., 1994].

In practice it is difficult to interpret the notion of rhythm for speech. A large variety of measures have been proposed to characterize speech rhythm, which are mainly based on the durational characteristics of consonantal and vocalic intervals; for example, the percentage over which speech is vocalic [Ramus et al., 1999], the average durational difference between consecutive consonantal or vocalic intervals in an utterance, which defined as the Pairwise Variability Index [Grabe and Low, 2002]. Speech rhythm studies using these measures usually focus on the taxonomy of languages as stress-timed and syllable-timed languages [Ramus et al., 1999, Grabe and Low, 2002, Gibbon and Gut, 2001, Loukina et al., 2011].

In addition to the time-domain representations that we have discussed above, there exist also frequency domain representations of speech rhythm. Dynamic information extracted both at local and global level from frequency domain representation of speech rhythm has been used for assessment of emotion [Ringeval et al., 2012]. In

this chapter, we compile a dictionary of speech rhythm representations per gesture category to define a relational model between rhythmic similarities of speech and gesture modalities. We use the low-frequency Fourier analysis of speech rhythm as introduced by [Tilsen and Johnson, 2008] to investigate the relationship between speech rhythmicity and vowel, consonant deletions on the Buckeye corpus.

4.2 System Overview

The general block diagram of our speech-driven gesture synthesis and animation system using speech rhythm information is given in Figure 4.1. The system consists of three main tasks: analysis, synthesis, and animation. Within the analysis task we have two stages: (i) feature extraction and clustering of gesture phrases and prosodic units, and (ii) their multimodal analysis. First stage of the analysis task is executed as defined in Section 3.3, where we perform feature extraction and unimodal clustering on speech and body motion data. The audio stream is processed to extract prosodic features of the speech, whereas the body motion data is expressed in terms of 3D joint angles. The gesture and audio feature streams are then segmented via temporal clustering into recurrent patterns, i.e, into gesture phrases and prosodic units, respectively. In Section 4.3, in addition to prosodic features, we also describe extraction of speech rhythm features for the duration of each gesture phrase, which are to be used later in the animation generation stage. The second stage of the analysis task, where we use hidden semi-Markov models (HSMM) to model the dependencies between speech prosody and gestures in a multimodal framework is executed as defined in Section 3.4. We synthesize the optimal sequence of gestures with durations for a given speech input by using the Viterbi algorithm over HSMM as detailed in Section 3.5. In Section 4.4, the synthesized gesture sequences with duration information are mapped into body motion sequences by using a multiple objective unit selection algorithm to generate animations. In Section 4.5, we present the results of speaker-dependent and independent experiments conducted for objective and subjective evaluations of the proposed system.

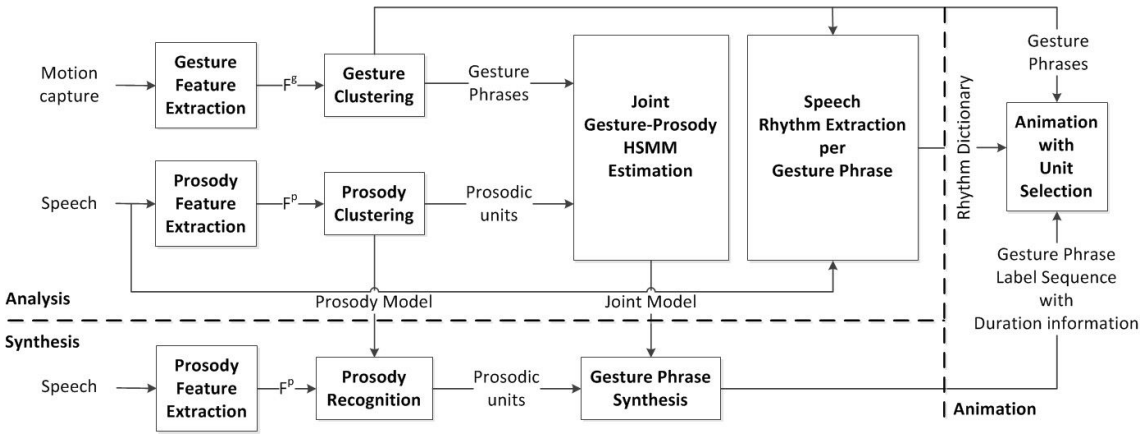


Figure 4.1: The block diagram of the general framework for the speech-driven gesture synthesis and animation system.

4.3 Speech Rhythm Feature Extraction

A multimodal system that combines speech and gesture modalities requires an explicit understanding of how these modalities co-occur and how they are jointly perceived. Generally, there is an underlying periodic pattern of pulses in speech (sometimes reinforced by coupled periodic motions of hands), and prominent events in the speech are approximately aligned in time with these pulses [Port, 2003]. In other words, the rhythmic production of speech, marked by pitch accents and stressed syllables, influences the temporal pattern of coinciding gestures [Iverson and Thelen, 1999]. Therefore, the rhythmic harmony of speech and accompanying gestures is important when synthesizing natural-looking virtual character animations.

Speech is a rhythmic and temporally structured source where the acoustic signal is transmitted as syllables in which most of the energy fluctuations occur in the range between 3 to 20 Hz [Greenberg, 1999, Greenberg and Arai, 2004]. When we refer to the term rhythm, we do not mean that these energy terms are perfectly periodic, but rather that there are regulations on syllable duration and energy patterns within and across prosodic phrases, which are important for intelligibility and naturalness of the spoken speech [Ladd, 1996, Liberman, 1975]. For example, from a phonetic point of view, we cannot fully define speech as sequences of phonemes, syllables or

words. When we listen to speech, we hear that segments or syllables are shortened or lengthened in accordance with an underlying pattern. However, characterization of speech rhythm is not an easy task itself and most of the methods rely on measurements of segmental durations to describe the temporal patterns of speech [Ramus et al., 1999, Grabe and Low, 2002, Gibbon and Gut, 2001, Loukina et al., 2011].

On the other hand, frequency domain representation of the speech rhythm, which characterizes how the energy of speech is distributed in the frequency domain, can be as useful in our framework. In this study we analyze speech rhythm using Fourier analysis of the amplitude envelope of bandpass filtered speech rather than computing rhythm with time domain measurements of interval durations. The frequency domain approach pays much less attention to where intervals begin and end, and more attention to the acoustic contents of those intervals by analyzing the power spectrum of the amplitude envelope of speech [Tilsen and Johnson, 2008].

We used the speech rhythm features as defined in [Tilsen and Johnson, 2008] in our multimodal framework. The input speech signal s^{ij}, which corresponds to $\varepsilon^{g_i j}$ (the jth gesture phrase in the gesture pool G_i), is filtered with a passband of 700 - 1300 Hz to capture mostly vocalic energy and filter out glottal energy and obstruent noise. Then, envelope of the band-pass filtered signal is low-pass filtered and down-sampled. Then, the normalized spectral energy distribution of this down-sampled signal over $d = 8$ bands is defined as the speech rhythm feature vector as,

$$\begin{aligned} \boldsymbol{r}(i,j) &= \frac{1}{\sum_n e_n}[e_1, e_2, ..., e_d], \\ &\text{for} \quad i = 1, 2, ..., M_g; j = 1, 2, ..., N_i^G, \end{aligned} \tag{4.1}$$

where e_n is the spectral energy for the nth band and $\boldsymbol{r}(i,j)$ is the speech rhythm feature vector corresponding to jth gesture phrase of ith gesture class, $\varepsilon^{g_i j}$, and M_g is the number of gesture classes. Further details of the speech rhythm feature extraction can be found in [Tilsen and Johnson, 2008].

The collection of speech rhythm feature vectors, $\boldsymbol{r}(i,j)$, are compiled as a dictionary of speech rhythm representations per gesture class and expressed as $R_i = \{\boldsymbol{r}(i,1), ..., \boldsymbol{r}(i,N_i^G)\}$. In other words, the speech rhythm dictionary R_i is linked with

the gesture pool G_i defined in Section 3.3.2, where each gesture phrase has a coinciding speech rhythm representation. In the animation generation step, R_i is used to reduce rhythmic mismatches between the input speech and the gesture phrases selected from the gesture pool G_i.

Although rhythm could be seen as a timing aspect of speech prosody along with intonation and stress, it represents phrase-level timing characteristics. Hence rather than including it in the short-term prosody feature representation, we rather use it as a similarity factor while creating gesture animations as will be described in Section 4.4.

4.4 Gesture Animation

Animation of the synthesized gesture label sequence consists of three main tasks: Extraction of gesture phrase sequence with unit selection, smoothing gesture-to-gesture transitions, and finally graphical animation of the gesture phrase sequence.

The first task is to generate a synthesized sequence of gesture phrases, $\hat{\boldsymbol{\varepsilon}}^g$, given the synthesized gesture phrase label sequence $\hat{\boldsymbol{\ell}}^g$ along with the corresponding duration sequence $\boldsymbol{\kappa}$ and the input speech rhythm information. This task is performed using unit selection over a pool of gesture phrases which are extracted during the gesture analysis in Section 3.3.2. The next task is to smooth joint angle discontinuities over a temporal window at gesture phrase boundaries, that is, at the boundary of each two consecutive synthesized gesture phrases $\hat{\varepsilon}^g_l$ and $\hat{\varepsilon}^g_{l+1}$, to extract a smoothed gesture sequence $\tilde{\boldsymbol{\varepsilon}}^g$. The smoothed gesture phrase sequence $\tilde{\boldsymbol{\varepsilon}}^g$ is finally used to animate beat gestures of a virtual character in synchrony with the input speech.

We employ a unit selection algorithm to generate the synthesized sequence of gesture phrases, $\hat{\boldsymbol{\varepsilon}}^g$, based on the gesture pool $G_i = \{\varepsilon^{g_i 1}, \varepsilon^{g_i 2}, \ldots, \varepsilon^{g_i N_i^G}\}$, which is constructed in Section 3.3.2 for each gesture phrase class g_i. The unit selection process is demonstrated in Figure 3.4 as the formation of an optimal gesture phrase sequence from the templates available in the gesture pools.

We minimize a mixture of penalty scores for duration mismatch, joint angle continuity mismatch and speech rhythm mismatch during the unit selection process.

Speech rhythm similarity of gestures is used to avoid rhythmic mismatches between the input speech and the synthesized gesture motions during animations. We use the rhythm dictionary R_i for gesture class g_i, constructed as described in Section 4.3. The duration, joint angle continuity and speech rhythm mismatch penalties of a gesture phrase template $\varepsilon^{g_i j}$ for a synthesized gesture phrase with label $\hat{\ell}_l^g$ are respectively defined as,

$$D_\omega(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = ||\omega_e(\hat{\varepsilon}_{l-1}^g) - \omega_b(\varepsilon^{g_i j})||, \quad (4.2)$$

$$D_\kappa(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = ||\kappa_l - \kappa(\varepsilon^{g_i j})||, \text{and} \quad (4.3)$$

$$D_r(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = ||r_l - r(i,j)||, \quad (4.4)$$

where κ_l is the duration of the synthesized gesture phrase with label $\hat{\ell}_l^g$, $\kappa(\varepsilon^{g_i j})$ is the duration of the gesture phrase template $\varepsilon^{g_i j}$ in gesture pool G_i, $\omega_e(\hat{\varepsilon}_{l-1}^g)$ is the ending joint angle vector of the previously synthesized gesture phrase $\hat{\varepsilon}_{l-1}^g$, and $\omega_b(\varepsilon^{g_i j})$ is the beginning joint angle vector of the gesture phrase template $\varepsilon^{g_i j}$, r_l is the input speech rhythm feature for the lth phrase $\hat{\varepsilon}_l^g$ and $r(i,j)$ is the speech rhythm feature of the gesture phrase template $\varepsilon^{g_i j}$. The joint penalty score is defined as a mixture of three penalty scores for duration, joint angle and rhythm:

$$\begin{aligned} D(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = {} & \beta\alpha\overline{D_\omega}(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) + \\ & \beta(1-\alpha)\overline{D_\kappa}(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i)] + \\ & (1-\beta)\overline{D_r}(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i), \end{aligned} \quad (4.5)$$

where $\overline{D_\omega}$, $\overline{D_\kappa}$ and $\overline{D_r}$ are min-max normalized penalty functions, and α and β are the mixture weights which we set experimentally over a validation set, as will be explained in Section 4.5.2.

The selected gesture phrases are resampled so as to fit the synthesized duration if necessary. Next, we smooth joint angle discontinuities at gesture boundaries over a temporal window. This is achieved by applying an exponential smoothing function on each pair of consecutive synthesized gesture phrases $\hat{\varepsilon}_l^g$ and $\hat{\varepsilon}_{l+1}^g$. Then, the smoothed gesture motion sequence $\tilde{\boldsymbol{\varepsilon}}^g$ is used to animate a virtual character based on the four

skeleton joints mentioned in Section 3.3.2. The other joints of the body (e.g. spine, lower-body joints) are assumed to have no motion.

4.5 Experimental Results

We use two datasets for synthesizing and animating prosody-driven arm gestures in speaker-dependent and independent settings. Prior to subjective and objective evaluations, we fine-tune the parameters of our speech-driven gesture synthesis scheme based on objective metrics. These parameters are the number of prosody clusters and the penalty score weight parameters used in the unit selection algorithm for animation generation. Finally we present subjective and objective evaluations of the proposed framework.

4.5.1 The Number of Prosody Clusters

One of our primary goals in this work is to synthesize gesture sequences with realistic gesture durations. Hence, one possible objective evaluation of our HSMM based gesture synthesis is to consider the similarity between the original and the synthesized gesture duration statistics. To this effect, we perform the prosody clustering process under different parameter settings, and for each setting we synthesize a different gesture sequence. We then estimate the duration distribution of each synthesized sequence as in (5.3). Next, we compute the symmetric Kullbeck-Leibler (KL) divergence between the original duration distribution $d_i(k)$ and the synthesized distribution $\hat{d}_i(k)$ over the validation set to measure the duration similarity of the synthesized gesture sequence with the original one, where smaller KL divergence values indicate more consistent duration distributions.

We perform unsupervised prosody clustering using parallel-branch HMM structures, as described in Section 3.3.1, with branch numbers ranging from 10 to 18 and state numbers per branch ranging from 3 to 5 for the speaker-dependent setting. An over-segmented prosody stream with larger number of branches would produce redundant and similar clusters while under-segmentation with less branches would

have to merge distinct clusters. The range of values for setting number of states on the other hand is selected considering the minimum duration of temporal prosody clusters. Table 4.1 presents the symmetric KL divergence scores on the validation set for various parameter settings. We observe that with small number of branches, $M_p < 14$, the KL divergence has minimum values at $N_p = 4$ number of states per branch. As the number of branches gets larger, $M_p \geq 14$, the optimal KL divergence values are attained at $N_p = 3$ number of states per branch. We use the $M_p = 16$ and $N_p = 3$ as the optimal setting with a KL divergence value of 1.0498 for subjective evaluation of our gesture synthesis system, which is presented in Section 4.5.3.

Moreover, we perform unsupervised prosody clustering on the CreativeIT dataset for the speaker-independent evaluations. We select number of states per branch as 3 and vary the number of branches ranging from 10 to 32. Table 4.2 presents the symmetric KL divergence scores of the original and synthesized gesture duration distributions for the speaker-independent setting in a leave-one-actor pair-out manner. As in Table 4.1, using $M_p = 16$ prosody clusters gives the optimal KL-divergence value as 7.5270. This value is higher than the value obtained in the speaker-dependent setting, which is expected since the speaker-independent system has higher variability.

Table 4.1: The symmetric KL divergence of the original and synthesized gesture duration distributions for various prosody clustering settings in the speaker-dependent system

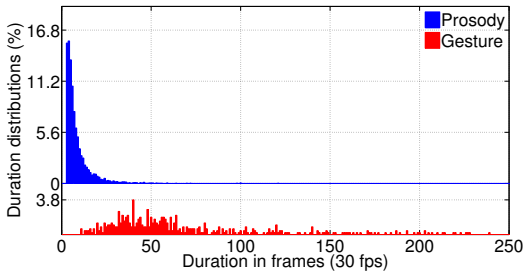
N_p	M_p				
	10	**12**	**14**	**16**	**18**
3	2.1251	1.1501	1.3333	**1.0498**	1.3544
4	1.1052	1.0698	1.6390	1.5562	1.8460
5	1.5551	1.9665	1.5705	1.7472	1.4214

In addition, for the optimal symmetric KL-divergence settings, we compare histograms of gesture phrase and pattern durations with prosodic unit durations (measured in number of frames with 30 fps) in Figure 4.2, for speaker-dependent (top) and independent (bottom) settings, respectively. In the figures, the discrepancy in

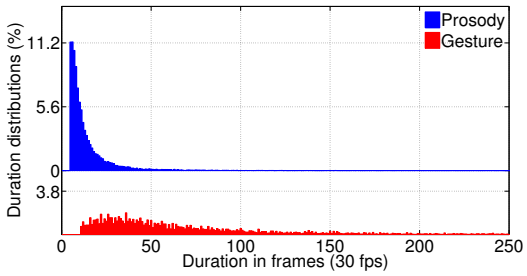
Table 4.2: The symmetric KL divergence of the original and synthesized gesture duration distributions for various prosody clustering settings in the speaker-independent system

N_p	M_p			
	10	**16**	**24**	**32**
3	7.8959	**7.5270**	7.7871	7.9978

duration distributions for the two modalities is clear. The observation that gesture phrases and patterns have longer durations compared to prosodic units is inline with our choice of using HSMMs for joint modeling of the two modalities.



(a) Speaker-dependent setting



(b) Speaker-independent setting

Figure 4.2: Duration histograms of prosodic units (blue) and gesture phrases or patterns (red).

On the other hand, the KL-divergence value for the baseline synthesis method is calculated as 1.8376 and 7.8365 for speaker-dependent and independent settings, respectively. A higher KL-divergence in this case is expected since, unlike the HSMM-

based synthesis, statistical duration information is simply ignored with the baseline synthesis as well as any gesture transition statistics and prosody-gesture correlations.

4.5.2 The Penalty Score Weight Parameters

The proposed gesture animation system employs unit-selection to minimize a mixture of three different penalty scores while mapping synthesized gesture sequences into motion sequences. These scores are the duration difference penalty, the joint angle continuity and the speech rhythm similarity as defined in Section 4.4. Hence prior to the gesture motion smoothing step in Section 4.4, the penalty score weights α and β in (4.5) are to be set experimentally. We consider two scoring functions to set α and β values. The first function is based on a windowed cross-lagged correlation (WCLC) score [Boker et al., 2002]. The second function evaluates smoothness of the animation through a jerkiness score as defined by [Hogan and Sternad, 2009].

Correlation is a commonly used tool to evaluate synchrony in interacting time-series coordination [Delaherche et al., 2012]. Canonical correlation analysis (CCA) is a statistical analysis technique for measuring the linear relationship between two multi-dimensional variables. CCA seeks a pair of basis vectors, u_x and u_y, one for each multi-dimensional variable, $\boldsymbol{x}$ and $\boldsymbol{y}$, such that the correlations between the projections of these variables onto basis vectors are mutually maximized. We define a CCA based correlation coefficient,

$$\rho^{cca}(\boldsymbol{x}, \boldsymbol{y}) = corr(u_x^T \boldsymbol{x}, u_y^T \boldsymbol{y}) \tag{4.6}$$

where u_x and u_y are the canonical basis vectors maximizing correlation of the first pair of canonical variables and $()^T$ is matrix transpose.

We use CCA to correlate kinetic energy of the synthesized gesture sequences ($\boldsymbol{E}^s$) to kinetic energy of the original gesture sequences ($\boldsymbol{E}^o$) and to the speech prosody ($\boldsymbol{F}^p$). We define the kinetic energy per joint as the square of joint angles' angular velocity,

$$e_k^i = \left[\frac{\sum_{j=1}^{2} [\boldsymbol{\theta}_{k+j}^i - \boldsymbol{\theta}_{k-j}^i] j}{2 \sum_{j=1}^{2} j^2} \right]^2 \tag{4.7}$$

where $\boldsymbol{\theta}_k^i$ is the joint angle vector in radians for the i^{th} joint at frame k. The resulting kinetic energy sequence is defined as $\boldsymbol{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_T\}$, where $\boldsymbol{e}_k = [e_k^1, \ldots e_k^4]$ is the four dimensional kinetic energy vector at frame k.

The correlation coefficient between kinetic energy of the synthesized ($\boldsymbol{E}^s$) and original ($\boldsymbol{E}^o$) gesture sequences is defined at frame k with time lag τ as,

$$\rho_{k,\tau}(\boldsymbol{E}^s, \boldsymbol{E}^o) = \begin{cases} \rho^{cca}(\boldsymbol{E}_k^s, \boldsymbol{E}_{k+\tau}^o), & \text{if } -\tau_{max} < \tau \leq 0 \\ \rho^{cca}(\boldsymbol{E}_{k-\tau}^s, \boldsymbol{E}_k^o), & \text{if } \tau_{max} > \tau > 0, \end{cases} \tag{4.8}$$

where $\boldsymbol{E}_k = \{\boldsymbol{e}_k, \boldsymbol{e}_{k+1}, \ldots, \boldsymbol{e}_{k-1+W}\}$ is the kinetic energy vectors over a time window of size W starting at time frame k and τ_{max} is the maximum time lag. By selecting the windows with lag value τ ranging from -τ_{max} to +τ_{max}, we guarantee a mirror symmetry such that the resulting set of correlations will contain the same values when the two series are swapped as for $\rho_{k,\tau}(\boldsymbol{E}^o, \boldsymbol{E}^s)$. Then the maximum correlation coefficient between $\boldsymbol{E}^s$ and $\boldsymbol{E}^o$ is calculated as,

$$\rho^*(\boldsymbol{E}^s, \boldsymbol{E}^o) = \mathcal{E}\{\max_{\tau} \rho_{k,\tau}(\boldsymbol{E}^s, \boldsymbol{E}^o)\}, \tag{4.9}$$

where $\mathcal{E}$ is the expectation over time windows of the highest correlation coefficients over the time lags. Note that a similar correlation coefficient can be extracted between kinetic energy of the synthesized gesture sequences ($\boldsymbol{E}^s$) and the speech prosody ($\boldsymbol{F}^p$) as $\rho^*(\boldsymbol{E}^s, \boldsymbol{F}^p)$.

Then the WCLC-based score function given the penalty score weights α and β of unit-selection is defined as,

$$S_{WCLC}(\alpha, \beta) = \frac{\rho^*(\boldsymbol{E}^s, \boldsymbol{E}^o|\alpha, \beta) + \rho^*(\boldsymbol{E}^s, \boldsymbol{F}^p|\alpha, \beta)}{2}, \tag{4.10}$$

where the maximum correlation coefficients can also be computed for the given penalty score weights α and β in interval [0, 1]. We target to set α and β values to maximize the WCLC-based score, which will help us to measure and preserve correlations of the synthesized arm motion behaviors to the original motion and speech prosody behaviors in our animations. In our experiments, we use an analysis window size W as 6 sec and a maximum lag value τ_{max} as 1 sec. The WCLC-based score function is

extracted on the validation set for each training set in turn and then the average is used for the speaker-dependent setting, whereas for the speaker-independent setting leave-one-actor pair-out method is employed.

Our experience from subjective tests shows that humans' sensitivity to errors in gesture animation is highly correlated with its smoothness. Hence, noticeable artifacts introduced by motion editing, such as sudden jumps of the joints, should be avoided for a natural looking animation. We adopt the jerkiness measure, which is defined as the derivative of joints' acceleration in [Hogan and Sternad, 2009], to measure the smoothness of a gesture motion sequence. For the i^{th} joint at frame k, given the joint angle vector $\boldsymbol{\theta}_k^i$, the jerkiness is calculated as

$$\boldsymbol{J}_k^i = \frac{\boldsymbol{\theta}_{k+1}^i - 3\boldsymbol{\theta}_k^i + 3\boldsymbol{\theta}_{k-1}^i - \boldsymbol{\theta}_{k-2}^i}{\Delta^3} \quad \text{for} \quad k = 1, \ldots, K; \;\; i = 1, \ldots, 4, \tag{4.11}$$

where K is the total number of frames, and $\Delta = t_k - t_{k-1}$ is the frame duration of the animation. The overall jerkiness of the synthesized animation is then computed as:

$$J = \sqrt{\sum_{k=1}^{K} \frac{\sum_{i=1}^{4} (\boldsymbol{J}_k^i)^2}{4} \frac{K^3}{v_a^2}} \tag{4.12}$$

where v_a is the average angular velocity and calculated over the whole sequence and all arm joint angles. We measure the overall jerkiness value J on the validation set for each training set in turn and then take the average ($\bar{J}$) for the speaker-dependent setting whereas, for the speaker-independent case we employ leave-one-actor pair-out method.

In order to set optimal values for the α and β parameters, we evaluate the WCLC-based correlation score function S_{WCLC} and the average jerkiness value $\bar{J}$ as a function of the α and β parameters. Figure 4.3 plots these two evaluation metrics in both speaker-dependent and independent settings. In the WCLC-based score function higher correlations values, which are plotted in whiter colors, are preferred. However, in the average jerkiness metric, lower jerkiness values, which are plotted in darker

colors, are preferred. Note that, speaker-dependent and independent settings have similar tendencies. Lower values of the β parameter, which weights the rhythm difference penalty more, are not preferred by the WCLC-based score function in both settings. Similarly, lower values of α and higher values of β are preferred by the jerkiness metric in both settings. For the speaker-dependent setting, we set two possible parameter settings at (0.2, 0.5) and (0.2, 0.9) values of the (α, β). Note that these two points have lower jerkiness and higher correlation score values. Similarly, a single parameter point (α, β) is set for the speaker-independent setting at (0.3, 0.9).

4.5.3 Subjective Evaluations

Gestures can accompany speech in various different ways. Objective evaluations are incapable of qualifying such variabilities, whereas subjective tests can evaluate realism and naturalness of the animation by reflecting human perception. We conduct subjective tests using the animations of speaker-dependent setting, where each participant is shown side-by-side animation pairs and asked to perform A/B comparisons so as to evaluate the naturalness of gesture animations on a scale of $(-2, -1, 0, 1, 2)$. The values in this scale represent A much better, A better, no preference, B better and B much better, respectively. Animation clips in the test are designed to be long enough to allow participants to be able to evaluate transitions between gestures. Each comparison consists of a pair of animation clips of 40 to 80 second duration generated for a given utterance by using two of the following three methods: the HSMM-based synthesis, the baseline synthesis, and the motion-capture synthesis. The proposed HSMM based synthesis is subjectively evaluated in pairwise comparisons with baseline and motion-capture synthesis results. Our baseline gesture synthesis method creates an animation from a sequence of random gestures via gesture phrase selection based solely on joint angle continuity, hence discarding any gesture duration and transition statistics as well as prosody-gesture correlations. The motion-capture synthesis on the other hand uses the captured true motion in the animations; that is the speaker's gestures are directly copied to the animation. In the subjective evaluations,

the HSMM-based synthesis method employs the unit selection penalty score weight parameters (α, β) with values (0.2, 0.5) and (0.2, 0.9), which are fine-tuned via objective evaluations over the validation set as described in sections 4.5.1 and 4.5.2. The first setup, in which $\beta = 0.5$, emphasizes rhythm penalty score more compared to the latter one. So, the influence of using rhythm is also evaluated in the subjective tests.

Table 4.3: Results of the subjective A/B pair comparison test

A/B Pair	**Average Preference**	**p-value <**
Motion-capture / Baseline	-0.613	0.0001
Rhythm emphasized in HSMM-based ($\alpha = 0.2, \beta = 0.5$)		
Motion-capture / HSMM-based	-0.017	0.9186
Baseline / HSMM-based	0.343	0.0152
Rhythm less influencial in HSMM-based ($\alpha = 0.2, \beta = 0.9$)		
Motion-capture / HSMM-based	-0.574	0.0016
Baseline / HSMM-based	0.242	0.1210

In the subjective tests, each of the 26 participants is shown 22 pairs of animation clips in random order from a pool of 66 animation clips. The clip pool consists of 12 samples for each of the five pairs presented in Table 4.3. Each test includes 4 samples for each pairwise comparison, plus a pair of identical clips, which is used to ensure the participants' engagement in the test. The left-right display order of the animation pairs in clips and sequence order of clips are set randomly. All participants are native Turkish speakers of ages in the range 22-40 (16 female, 10 male). Table 4.3 presents the average preference scores and their statistical significances in the subjective evaluations. The preference score for each pair of the five cases is calculated as the average of participants' evaluation scores. A negative average preference score implies the method on the left side is preferred over the one on the right side and vice versa. A paired two-tailed t-test is used to evaluate the significance of test takers' preferences. We observe in Table 4.3 that while motion-capture synthesis is significantly favored over the baseline, it is not significantly discriminated from

the proposed HSMM-based synthesis when rhythm is emphasized ($\beta = 0.5$) in the animation generation step. Additionally, HSMM-based synthesis results, emphasizing rhythm, are assessed to be significantly more realistic and natural than baseline synthesis results with a p-value less than 0.0152. On the other hand, HSMM-based synthesis results are assessed to be less natural when rhythm is less influential ($\beta = 0.9$) in the animation generation step.

4.5.4 Objective Evaluations

In the proposed framework, we target to jointly model beat gestures, which are used to emphasize speech, and speech prosody. The HSMM-based synthesis and animation framework targets to match prosodic units and gestures. Gesture phrase labels and durations are extracted from the HSMM model, then the unit selection based animation system sets the sequence of the synthesized gestures from a large dictionary of gestures by applying the three penalty scores for duration, joint angle continuity and rhythm. Rather than matching the original joint angle sequence in the synthesis, the model tries to match prosody, which is emphasis on speech, and beat gestures, which emphasize motion of arms. Considering that the same source, i.e., the speech prosody, is driving the motion of arms for both the original gestures and the synthesized gestures, the kinetic energy difference between the original and synthesized gestures can be defined as a pose-invariant objective metric for animation of beat gestures from speech-prosody.

This pose-invariant objective metric is defined as the root mean square error (RMSE) between the kinetic energies of the original and synthesized joints:

$$RMSE = \sqrt{\frac{\frac{1}{4}\sum_{k=1}^{K} ||\boldsymbol{e}_k^o - \boldsymbol{e}_k^s||^2}{K}}, \tag{4.13}$$

where K is the total extent of the data and the $\boldsymbol{e}_k^o$ and $\boldsymbol{e}_k^s$ are respectively 4D kinetic energy vectors of the original and synthesized gesture sequences as defined in (4.7). The dataset used in the speaker-independent setting is dyadic, where speakers frequently take turns and mostly hold floor for a short time duration. In the RMSE

evaluations, we set a test dataset (TSTIND) by segmenting audio recordings into semantically meaningful utterances when the speaker holds the floor. For the speaker-dependent setting, the audio segments in the subjective test, which we refer them as TSTDEP, are used for the objective RMSE evaluations. The total duration of the TSTDEP and TSTIND evaluation sets are 458 and 444 seconds for speaker-dependent and independent settings, respectively. The RMSE scores of the HSMM-based rhythm emphasized synthesis and the baseline synthesis over the speaker-dependent and independent settings are presented in Table 4.4. Note that, RMSE scores of the HSMM-based synthesis is lower in both settings. While the RMSE scores are lower for the speaker-dependent setting, the scores and the score difference between the baseline and HSMM-based synthesis are larger for the speaker-independent setting. This is mainly due to the fact that the gesture animation pool for the speaker-dependent set is smaller, on the other hand the gesture pool in the CreativeIT dataset is larger with more speaker and gesture variability.

Table 4.4: Objective evaluations based on the RMSE between the kinetic energies of the original and synthesized joint angles over the speaker-dependent (TSTDEP) and speaker-independent (TSTIND) datasets

	HSMM-based	**Baseline**
TSTDEP	0.25	0.27
TSTIND	0.67	0.86

Another objective measure that we use to assess the quality of the resulting animations quantifies how good the proposed model to transfer speech prosody into beat gesture is. This goodness measure can be computed by correlating the speech prosody to the kinetic energy of the joints. Recall that, we defined the CCA based correlation coefficient $\rho_{k,\tau}(\cdot,\cdot)$ at time frame k with time lag τ in (4.8) that correlates two streams of information. Using this CCA-based correlation coefficient, we define

the following three correlation metrics as

$$\gamma^{po} = \rho_{k,0}(\boldsymbol{F}^p, \boldsymbol{E}^o), \tag{4.14}$$

$$\gamma^{ps} = \rho_{k,0}(\boldsymbol{F}^p, \boldsymbol{E}^s), \tag{4.15}$$

$$\gamma^{pb} = \rho_{k,0}(\boldsymbol{F}^p, \boldsymbol{E}^b), \tag{4.16}$$

where they define the correlation between the speech prosody and the kinetic energy, respectively for the original mocap (γ^{po}), for the HSMM-based synthesized (γ^{ps}), and for the baseline synthesized (γ^{pb}) joint angles.

Table 4.5 presents mean, standard deviation and percent of outliers, i.e., the percentage of correlation values that are greater than 0.2, for the three correlation metrics. Note that, statistics in this table are extracted over all the contents of the CreativeIT and MVGL-MUB datasets. The mean correlation values for the original mocap joint angles are not high for both datasets, where they are 0.24 and 0.07 for the CreativeIT and MVGL-MUB datasets, respectively. Based on these reference correlation values for the original mocap joint angles, the proposed HSMM-based synthesis yields higher mean correlation values than the baseline synthesis system. The main reason of the low mean correlation values is the sparsity of temporal windows, in which speech prosody and kinetic energy of the joint angles are strongly correlated. Although strongly correlated instances are sparse, subjective evaluations suggest that these highly correlated instances have an important role on the perception of naturalness in animations. Hence, we also compute the percentage of windows with relatively higher correlation values as outliers. The correlation threshold value is set as 0.2, which is a value close to the mean plus standard deviation for the synthesized joint angles. Note that, the percent of the outliers for the γ^{po} in the CreativeIT dataset is 60.6%, which is a relatively high percentage and this is probably due to the affective theatrical improvisations of the dataset. The proposed HSMM-based synthesis has 19.5% and 12.8% outliers respectively in the CreativeIT and MVGL-MUB datasets. Note that, the outliers of the HSMM-based synthesis are twice as high as the baseline synthesis. This is again a valuable objective evidence for the quality

improvement of the proposed HSMM-based animation system.

Table 4.5: Objective evaluations based on the correlations between the speech prosody and the kinetic energy of the original mocap (γ^{po}), the HSMM-based synthesized (γ^{ps}), and the baseline synthesized (γ^{pb}) joint angles. The last row presents percentage of correlation values that are greater than 0.2.

	CreativeIT			**MVGL-MUB**		
	γ^{po}	γ^{ps}	γ^{pb}	γ^{po}	γ^{ps}	γ^{pb}
Mean	0.24	0.10	0.04	0.07	0.08	0.06
Std	0.21	0.13	0.13	0.14	0.14	0.11
> 0.2 **(%)**	60.6	19.5	8.3	13.6	12.8	7.5

4.6 Summary

We presented a framework for speech prosody and rhythm driven synthesis and animation of beat gesture sequences. The main challenge in a gesture synthesis and animation system is modeling the relationship between speech and gesture modalities in a meaningful way and using this model to create new speech-synchronous animations. Our system employs hidden semi-Markov models (HSMMs) to explore the multimodal relationship and to synthesize speech-driven gesture sequences as detailed in Section 3.5. Then, gesture animations are generated from the synthesized gesture sequences using the unit selection algorithm as described in Section 4.4. We evaluated our framework in speaker-dependent and independent settings in Section 4.5. Building blocks of the speaker-dependent animations are the gesture phrases, which are extracted from motion capture data using semi-supervised segmentation as presented in Section 3.3.2. Gesture phrases are expressive enough to generate plausible motion sequences from an available motion capture dataset. They also remain intact during animation generation and significantly contribute to consistency and naturalness of the resulting animations. The speaker-independent animation system employs unsupervised clustering to segment the motion capture data, where the building blocks of the speaker-independent animations are defined as gesture patterns as detailed in

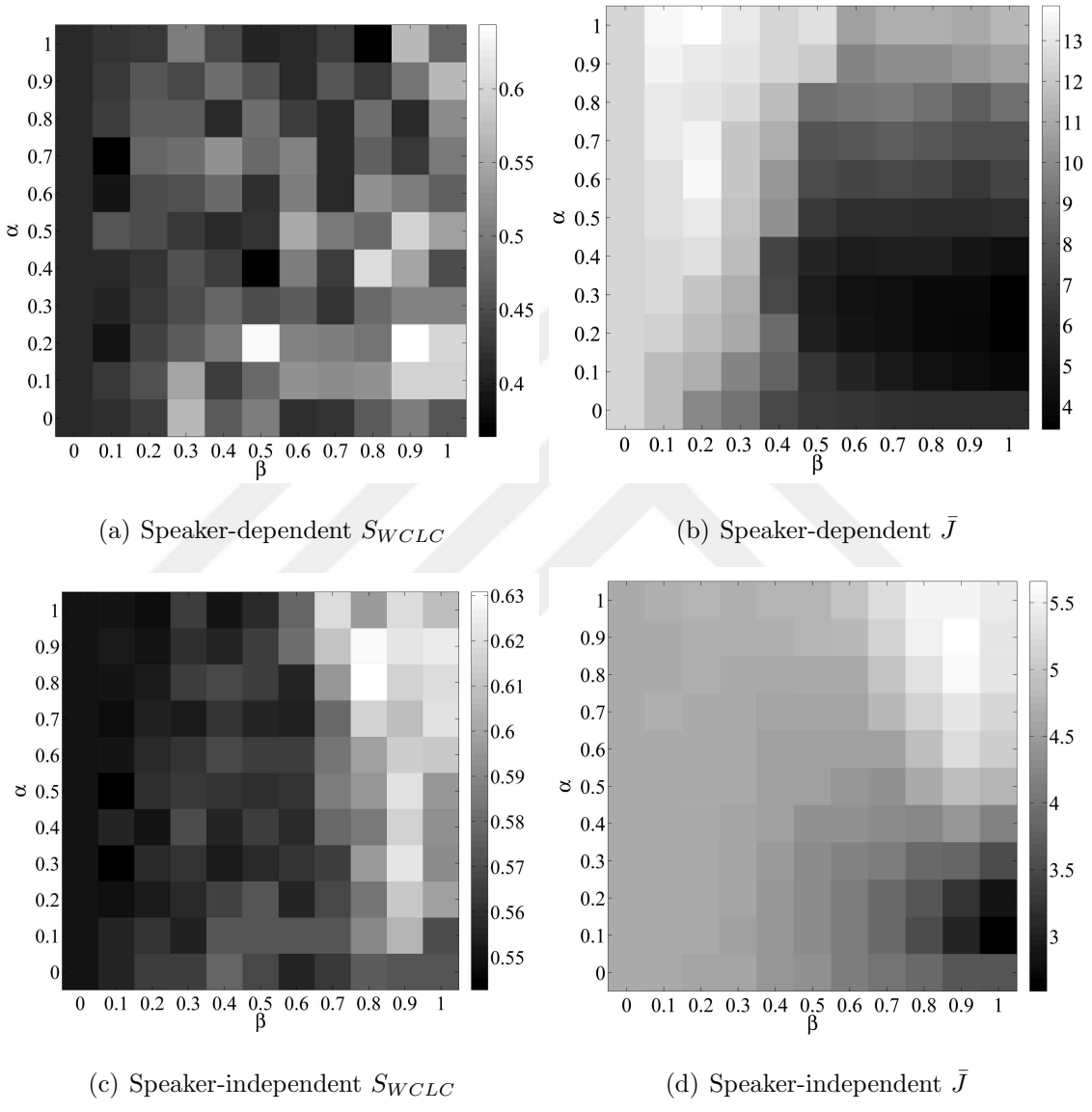
Section 3.3.2.

The proposed system first segments speech prosody and motion capture data by clustering them into prosodic units and gestures (phrases or patterns), respectively. In the multimodal analysis of gestures and prosodic units, gestures are defined as the states of a Markov chain and prosodic units are taken as the observations of this Markov process as presented in Section 3.4. Hence, state transitions model articulation of consecutive gestures. Alignment of gestures and prosodic units is captured by the HSMM. The proposed HSMM-based synthesis method effectively associates longer duration gestures to shorter duration prosody clusters while maintaining the realistic gesture duration and transition statistics as detailed in Section 3.5. Hence, our multimodal HSMM-based framework provides an effective solution for modeling the relationship between gestures and prosodic units, both at the temporal level and on a multimodal basis.

We use a unit selection method to map synthesized gesture sequences with duration information into motion sequences as presented in Section 4.4. The gestures from all gesture clusters are gathered in a gesture pool and the unit selection algorithm picks a sequence of optimal gesture realizations while minimizing a multiple objective cost. One limitation of our current framework is hence the representativeness and quality of the available gesture pool. We assume that gesture phrase samples in the gesture pool are statistically comparable to the ones used in the multimodal analysis step in terms of gesture category, duration distribution, and rhythm similarity.

We use objective and subjective evaluation methods to set the system parameters and to assess animation quality in two datasets. Our objective parameter setting results on the datasets are coherent for speaker-dependent and independent settings as presented in Section 4.5.2. Subjective evaluations indicate that the proposed system, when rhythm is emphasized in the animation, is significantly better than the baseline synthesis, and statistically similar to the animations created by using the original motion capture data as presented in Section 4.5.3. Furthermore, the RMSE between the kinetic energies of the original and the synthesized joints are reported, and we show

that the proposed HSMM-based synthesis yields lower RMSE scores than the baseline synthesis in Section 4.5.4. We also present the correlation of the speech prosody and the kinetic energy of joint angles as a valuable objective score. In subjective evaluations, we observe that strongly correlated instances of prosody and kinetic energy have an important role on the perception of naturalness in animations. We show that the proposed HSMM-based synthesis has twice as high strongly correlated instances than the baseline synthesis.



(a) Speaker-dependent S_{WCLC}

(b) Speaker-dependent $\bar{J}$

(c) Speaker-independent S_{WCLC}

(d) Speaker-independent $\bar{J}$

Figure 4.3: The WCLC-based score function $S_{WCLC}(\alpha, \beta)$ and the average jerkiness $\bar{J}$ plotted as a function of α and β values.

Chapter 5

AFFECTIVE SYNTHESIS AND ANIMATION OF ARM GESTURES FROM SPEECH PROSODY

5.1 Introduction

Non-verbal behavior provides information on the emotional state and personality of interlocutors, and is an integral part of human-to-human communication [Mehrabian, 1972, Liu et al., 2016]. Humans express non-verbal behaviors through body movements, gestures, head movements, eyebrow movements, facial expressions and gaze [Kopp et al., 2006]. Likewise, conversational agents (CA) can potentially display non-verbal behavior as in human-to-human communication and make human-computer interaction (HCI) easier and more fulfilling by enhancing believability and realism and by increasing the sense of empathy and attachment to synthetic characters [Vinayagamoorthy et al., 2006, Vinciarelli et al., 2012]. For example, an empathic CA can encourage and persuade students while using e-learning systems [Gwo-Dong et al., 2012]. Moreover, CAs with naturalistic behaviors can be used in mobile interfaces [Kang et al., 2015], intelligent tutoring systems [Rickel and Johnson, 2000], entertainment [Rehm and Wissner, 2005], and also as virtual interpreters [Baledassarri and Cerezo, 2012]. Such CA based interactions are desired to have automated generation of non-verbal behaviors rather than hand-crafted animations tuned to specific scenarios. In this paper we focus on automatic synthesis of gestures from affective speech as means of non-verbal communication.

Gestures are performed mainly by hands, arms and head to convey non-verbal cues in affective human communication, and arm-gesticulation is one of the most frequently used non-verbal behavior in human communication [Wagner et al., 2014,

McNeill, 1992]. Existing methods for synthesizing gestures in CAs can mainly be grouped as text-driven [Gratch and Marsella, 2001, Stone et al., 2004, Becker et al., 2004, Hartmann et al., 2006, Hartmann et al., 2005, Becker et al., 2007, Neff et al., 2008, Pelachaud, 2009, Niewiadomski et al., 2009, Mancini et al., 2011, Schroder et al., 2012, Noot and Ruttkay, 2005] and audio speech-driven [Levine et al., 2010, Baena et al., 2013, Marsella et al., 2013, Bozkurt et al., 2013, Chiu and Marsella, 2014, Bozkurt et al., 2015a, Bozkurt et al., 2015b, Sadoughi and Busso, 2015]. Some of these studies also consider affect expressiveness of gestures in emotional categories [Gratch and Marsella, 2001, Becker et al., 2004, Hartmann et al., 2006, Hartmann et al., 2005, Becker et al., 2007, Pelachaud, 2009, Niewiadomski et al., 2009, Mancini et al., 2011, Schroder et al., 2012, Noot and Ruttkay, 2005, Marsella et al., 2013, Baena et al., 2013]. However, categorical emotion models would fail to accurately describe non-basic states like thinking, embarrassment and depression. On the other hand, dimensional approaches represent affect on a continuous scale, which is useful for explaining non-categorical, complex affective states and also affective state transitions [Gunes et al., 2011].

Currently, linguistic features are predominant for automatic synthesis of affect expressive gestures in CAs. Existing methods aim to retain the observed correlations and alignments between speech and gestures while maintaining the predominant role of linguistic choices [Gunes et al., 2011]. For example, when a text-to-speech system drives the CA, based on the lexical content, affective expression of the CA is usually guided by selecting a desired affective state and a movement type described in a markup language, and then by adding expressiveness to the movement via modulation [Hartmann et al., 2006], [Hartmann et al., 2005], [Pelachaud, 2009], [Niewiadomski et al., 2009], [Mancini et al., 2011], [Noot and Ruttkay, 2005]. However, targeting gestures based only on the lexical content may fail to capture the large variability of gestures and its dependency on speech. Modeling cross-modal aspects (e.g., joint modeling, causal modeling, modality alignment) of speech, gestures and affect is required for creating CAs which are capable of expressing non-verbal communication.

On the other hand, speech-driven gesture synthesis approaches are useful in creat-

ing realistic gestures given the rich information conveyed in speech such as emotional cues and prosodic patterns, which are essential for modeling the variability and timing of the gestures [Wagner et al., 2014]. Although there are studies analyzing gestures in relation to affect attributes [Kleinsmith and Bianchi-Berthouze, 2013], speech-driven gesture synthesis methods mostly consider neutral speech [Levine et al., 2010,Bozkurt et al., 2013,Chiu and Marsella, 2014,Sadoughi and Busso, 2015], or agitation [Marsella et al., 2013] and intensity [Baena et al., 2013] levels of speech. In the current literature, use of speech and gesture to enhance affective expressiveness of CAs remains as an open problem.

In this chapter, we present a data-driven, learning-based computational model for affective speech-driven arm-gesture synthesis and animation. We jointly analyze gestures with continuous affect attributes (i.e., activation (A), valence (V) and dominance (D)) and speech prosody using hidden semi-Markov models. Our computational model is based on the framework that we previously developed in our earlier works [Bozkurt et al., 2013, Bozkurt et al., 2015b] for speech prosody-driven gesture synthesis and animation. The main objective of our current work is to investigate ways of integrating affect attributes into this general framework so as to generate novel gesture performances from affective speech. To the best of our knowledge, this is the first study that uses continuous description of affect in combination with speech to drive a gesture synthesis and animation framework (along with [Bozkurt et al., 2015a] that is a preliminary version of this current work). We present a fully automatic speaker-independent system which takes speech as input and generates synchronous animations of expressive arm gestures. We conduct experiments on the USC CreativeIT dataset [Metallinou et al., 2010, Metallinou et al., 2015], which contains synchronous speech and motion capture data recorded from improvised dyadic interactions, and perform objective and subjective evaluations to assess the contribution of incorporating affect into gesture synthesis. Our findings suggest that modeling gesture observations with a conditional distribution of prosody given affect sustains the best performance in objective and subjective evaluations compared to several

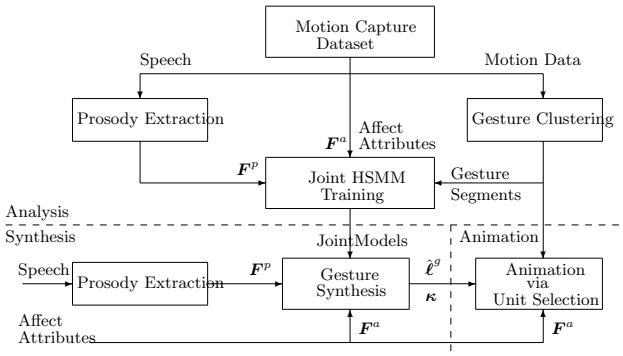


Figure 5.1: Block diagram of the general framework for the proposed speech-driven affective gesture animation system.

other strategies that we consider to incorporate affect into our framework.

5.2 System Overview

The general framework for our speech-driven gesture synthesis system is given in Figure 5.1, which consists of three main functional phases for analysis, synthesis and animation. In the analysis phase, we perform unsupervised clustering over 3D arm motion data in order to construct a dictionary of gestures. Then, we train a hidden semi-Markov model using speech prosody and affect attributes as observations over the gesture classes. In the synthesis phase, the HSMM is used as a generative model to synthesize a gesture sequence given prosody and affect observations, where each gesture is expressed with a gesture class id and duration information. Finally in the animation phase, we find an optimal sequence of arm motion from the synthesized gesture sequence by using a unit selection algorithm. The unit selection algorithm performs a multi-objective cost minimization task on a pool of recorded gestures so as to construct an optimal motion sequence over the synthesized gestures with duration information.

5.2.1 Gesture Clustering

We represent gestures in motion capture data by the joint angles of arms and fore-arms, which are significant for distinguishing affective states [Yang et al., 2014]. Accordingly, we define the joint angle vector for the i^{th} joint at frame k as $\theta_k^i = [\phi_{ik}^x, \phi_{ik}^y, \phi_{ik}^z]$, where $\phi_{ik}^x, \phi_{ik}^y, \phi_{ik}^z$ are the Euler angles respectively in the x, y, z directions, representing the orientation of the i^{th} joint at frame k. Then, we define the gesture feature vector at frame k, $\boldsymbol{f}_k^{Ji}$, to include the joint angles from the i^{th} body part and their first order derivatives,

$$\boldsymbol{f}_k^{Ji} = [\theta_k^i, \Delta\theta_k^i], \quad \text{for } i = 1, 2, 3, 4, \tag{5.1}$$

where $\Delta\theta_k^i$ denotes the first order derivative of the joint angle vector θ_k^i. The resulting gesture feature for the four joints of arms and fore-arms at time frame k is defined as,

$$\boldsymbol{f}_k^g = [\boldsymbol{f}_k^{J1}, ..., \boldsymbol{f}_k^{J4}]', \tag{5.2}$$

where $[\ ]'$ representing the transpose operator.

We choose to represent arm movements in a finite discrete gesture space, where the movements are mapped to a sequence of gesture segments and each gesture segment belongs to a gesture class. We perform temporal clustering of gesture feature sequences to define recurring gesture segments, which are referred as gesture classes in this study. Gesture clustering aims to group similar gestures based on their joint angles and velocities, which can be interpreted as describing gestures with objective features like handedness, arm orientation or motion direction. For this purpose, we employ unsupervised clustering of gestures based on the parallel-branch HMM model as defined in [Sargin et al., 2008]. In this model, a parallel-branch HMM structure is defined over the gesture feature stream $\boldsymbol{F}^g = \{\boldsymbol{f}_1^g, \boldsymbol{f}_2^g, ..., \boldsymbol{f}_T^g\}$, where T is the length of feature sequence. The HMM structure initially is set to have M_g parallel branches, where each branch represents a gesture class. Each branch is a left-to-right HMM, which is composed of $N_g = 10$ states corresponding to the minimum gesture segment duration of 10 frames ($\frac{1}{3}$ seconds assuming 30 video frames/sec). At the end of the

clustering process, the gesture feature stream is divided into gesture segments, where the l^{th} segment in the sequence is denoted by ε_l^g, and labeled with ℓ_l^g as one of the M_g available gesture classes $\{g_1, g_2, ..., g_{M_g}\}$.

5.2.2 Features Driving the Synthesis

We use prosody and affect attributes to drive the gesture synthesis. Prosody characteristics at the acoustic level, including intonation, rhythm, and intensity patterns, carry important temporal and structural synchrony with gestures [Loehr, 2012]. Acoustic features such as pitch and speech intensity can be used to model the underlying intonation of speech. We choose to include speech intensity, pitch, and confidence-to-pitch into the prosody feature vector as in [Bozkurt et al., 2015a]. Since the prosody feature values are speaker and utterance dependent, we apply mean and variance normalization to the prosody features to get the normalized prosody features. Then the normalized intensity, pitch, confidence-to-pitch features and their first-order temporal derivatives are used to define the 6D prosody feature vector as $\boldsymbol{f}^p$.

Prosody conveys intonation which is important for modeling the variability and complexity of timings of gestures [Wagner et al., 2014]. However, using only prosody to drive the synthesis may not capture whole extent of the affective information. For this purpose, we explore the use of continuous affect attributes of speech data, which are *activation* (A), *valence* (V), and *dominance* (D) for affect-expressive gesture synthesis and animation. The continuous affect attributes, AVD, are extracted through manual ratings of audio-visual recordings by expert raters. It is common to extract a ground truth annotation of AVD by averaging the ratings of the expert raters. Since this study is a proof of concept to identify the role of affect in speech-driven gesture animation, we primarily choose to use the ground truth annotations of the affect attributes. We represent the 3D affect feature by $\boldsymbol{f}^a$.

On the other hand, ground truth annotation of affect attributes is hard to extract for an automated speech-driven gesture animation system. Hence, we also investigate the performance of our speech-driven gesture animation system with the affect at-

tributes estimated from speech as well. We estimate affect attributes from the speech prosody and use these estimated attributes in the gesture synthesis step. We use support vector regression (SVR) models with a Gaussian kernel to estimate the activation, valence, dominance attributes of affect from prosody features. We use the ground truth AVD annotations to train these SVR models [Chang and Lin, 2011]. We estimate each affect attribute separately, since SVR performs non-linear regression from a multi-dimensional input to a uni-dimensional output. We represent the estimated affect attributes by $\boldsymbol{f}^{\hat{a}}$.

5.2.3 HSMM Modeling

The hidden semi-Markov model (HSMM) replaces self-transition probabilities in the regular hidden Markov model by state duration distributions [Yu, 2010]. We construct an HSMM model to relate gesture, prosody and affect modalities. In our framework, we take gestures as the states of a Markov chain, where prosody and affect features are the observations. Hence, state transitions correspond to articulation of consecutive gestures. Introducing a state duration model allows us to better control gesture segment durations as well as timing of the gestures in the synthesis process.

An HSMM representing continuous observations with M_g fully connected states is modeled as $\Lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{\Pi})$. The states of Λ represent gesture classes, and the model parameters $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{D}$, $\boldsymbol{\Pi}$ are respectively 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}, each representing the state transition probability from gesture g_i to g_j.

The state duration distribution $\boldsymbol{D}$ is formed in terms of state dependent duration probability mass functions,

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

where $d_i(n)$ is the probability of a gesture segment from gesture class g_i lasting $n\delta$ sec. Here, D_{max} is the maximum duration among all gesture segments, and δ is the histogram bin size for the underlying probability mass function. In our experiments,

we take the maximum duration as $D_{max} = 5$ sec, and the histogram bin size as the speech frame duration $\delta = 25$ msec.

5.2.4 Gesture Synthesis

Gesture synthesis is defined as decoding an optimal state sequence, $\hat{\boldsymbol{\ell}}^g$, over the HSMM Λ given a sequence of feature vectors, $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_T\}$. Note that the decoded optimal state sequence delivers synthesized sequence of gesture segments and their durations, where the HSMM framework secures to have realistic gesture segment durations. Unlike the HMM, in the HSMM framework, state duration distributions are employed in the decoding process. This requires to define a forward likelihood function incorporating the state duration model for the Viterbi decoding algorithm,

$$\psi_t(j) = \max_{\kappa,i} \left\{ \psi_{t-\kappa}(i) + \log\left(a_{ij} d_j(\kappa) \prod_{k=t-\kappa+1}^{t} b_j(f_k) \right) \right\}, \tag{5.4}$$

where $\psi_t(j)$ is the accumulated logarithmic likelihood at time frame t in state g_j after observing $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_t\}$. Note that the maximization in (5.4) finds an optimal duration κ^* and the previous state i^*, and ties the accumulated likelihood $\psi_{t-\kappa^*}(i^*)$ to $\psi_t(j)$. Based on the likelihood function $\psi_t(j)$, we use the modified Viterbi decoding algorithm to extract the optimal state sequence, that is the optimal gesture segment sequence $\hat{\boldsymbol{\ell}}^g = \{\hat{\ell}^g_1, \ldots, \hat{\ell}^g_L\}$, and the associated gesture segment durations $\boldsymbol{\kappa} = \{\kappa_1, \ldots, \kappa_L\}$, where L is the number of gesture segments in the synthesized sequence.

5.2.5 Gesture Animation

Animation of the synthesized gesture sequence consists of three main tasks: generation of gesture motion sequences, smoothing gesture transitions and graphical animation of gestures. The first task is to generate a synthesized sequence of gesture segments, $\hat{\boldsymbol{\varepsilon}}^g$, given the synthesized gesture segment label $\hat{\boldsymbol{\ell}}^g$ and duration $\boldsymbol{\kappa}$ sequences. This task is performed using unit selection over the gesture segments, which are extracted during the gesture analysis in Section 5.2.1. We apply a dynamic programming algorithm to

minimize a joint distortion function, R, which penalizes gesture transition distortion, duration difference and mismatch of affect attributes. The joint distortion function R is defined for each gesture label $\hat{\ell}_l$ of the synthesized state sequence as

$$\begin{aligned} R(\varepsilon_n^{g_i}|\hat{\ell}_l = g_i) &= R_\omega(\varepsilon_n^{g_i}|\hat{\ell}_l = g_i) + \\ & \quad R_\tau(\varepsilon_n^{g_i}|\hat{\ell}_l = g_i)] + \\ & \quad R_a(\varepsilon_n^{g_i}|\hat{\ell}_l = g_i), \end{aligned} \tag{5.5}$$

where R_ω, R_τ and R_a are the min-max normalized distortion functions. The distortion function R_ω is the normalized mean squared difference between the joint angles of the animated gesture segments at the transition from state $\hat{\ell}_{l-1}$ to state $\hat{\ell}_l$. The distortion function R_τ is the normalized absolute-valued difference between the durations of the candidate gesture segment $\varepsilon_n^{g_i}$ and the synthesized duration $\hat{\kappa_l}$. Similarly, the distortion function R_a is the normalized mean squared difference between the (annotated) affect features of the gesture segment $\varepsilon_n^{g_i}$ and the (annotated or estimated) affect features of the input speech (each of the attributes A, V and D is averaged over the temporal window of the synthesized gesture segment). The unit selection algorithm performs minimization of the following cost function:

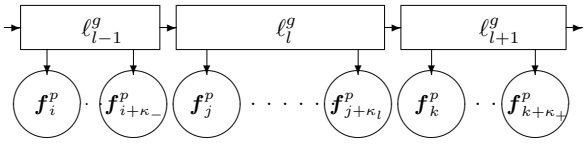
$$\varphi_l(\varepsilon_n^{\hat{\ell}_l}) = \min_j \left\{ \varphi_{l-1}(\varepsilon_j^{\hat{\ell}_{l-1}}) + R(\varepsilon_n^{g_i}|\hat{\ell}_l = g_i) \right\}, \tag{5.6}$$

where $\varphi_l(\varepsilon_n^{\hat{\ell}_l})$ gives the minimum accumulated distortion at the l-th label $\hat{\ell}_l$ of the gesture state sequence when the n-th segment is selected from the corresponding gesture subset, after observing gestures labeled as $\{\hat{\ell}_1, \hat{\ell}_2, \ldots, \hat{\ell}_l\}$. The gesture segments selected from the pool are further resampled to fit the synthesized duration $\hat{\boldsymbol{\kappa}}$.

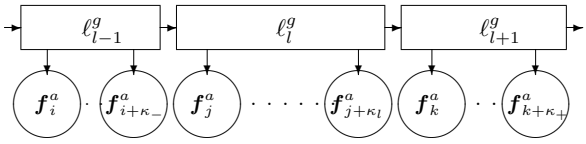
The second task is to smooth joint angle discontinuities over a temporal window at gesture unit boundaries. This is achieved by applying an exponential smoothing function on the synthesized gesture motion sequence. Finally, the smoothed gesture motion sequence is animated using the MotionBuilder 3D Character Animation Software[1].

5.3 Affect-Expressive Gesture Modeling

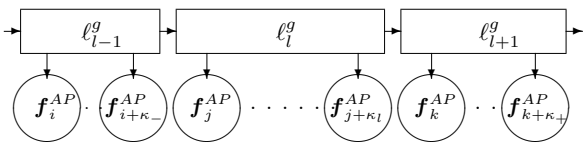
In this paper, we investigate the use of continuous affect attributes in speech-driven affective synthesis and animation of arm-gestures. As we defined in Section 5.2.2, the observations of the HSMM structure are affect attributes and prosody features. In this section, we define four distinct structures to model emission distributions of the observations. Figure 5.2 presents these four structures, where the features are assumed to be generated from the gesture state sequence $\boldsymbol{\ell}^g$ after unsupervised segmentation as defined in Section 5.2.1. Figures 5.2(a) and 5.2(b) present baseline models, which use prosody-only and affect-only features as observations, respectively. We consider joint distribution of affect and prosody as shown in Figure 5.2(c). The fourth structure in Figure 5.2(d) uses conditional observation distributions of prosody given affect, which defines a non-homogeneous emission model. Note that for all the four structures in Figure 5.2, state transition and duration probabilities are calculated based on the gesture state sequences in the training data, and they only differ by their emission distributions.



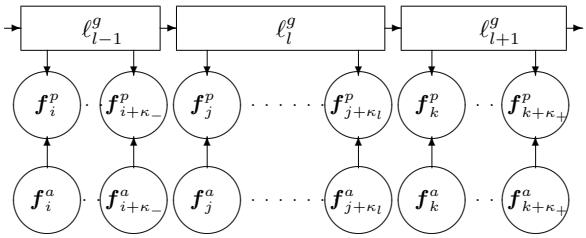
(a) Λ^p: Unimodal Prosody



(b) Λ^a: Unimodal Affect



(c) Λ^{ap}: Joint Affect and Prosody



(d) $\Lambda^{p|a}$: Prosody given Affect

Figure 5.2: Different emission models for the HSMM: (a) unimodal prosody (Λ^p), (b) unimodal affect (Λ^a), (c) joint affect and prosody (Λ^{ap}), and (d) prosody given affect ($\Lambda^{p|a}$).

5.3.1 Unimodal Structures

We consider two baseline structures, Λ^p and Λ^a, with unimodal emission distributions as given in Figure 5.2(a) and 5.2(b), where the first one using only prosody features, $\boldsymbol{f}^p$, and the second one using only affect features, $\boldsymbol{f}^a$, as observations. We call them as unimodal structures and define their observation distributions using the Gaussian mixture model (GMM) density functions, as

$$b_j(\boldsymbol{f})=\sum_{k=1}^{K}\omega_{jk}\mathcal{N}(\boldsymbol{f};\boldsymbol{\mu}_{jk},\boldsymbol{\Sigma}_{jk}),\ \text{ for } j=1,...,M_g, \tag{5.7}$$

where the observation $\boldsymbol{f}$ can be defined as $\boldsymbol{f}^a$ or $\boldsymbol{f}^p$, $b_j(\boldsymbol{f})$ is the observation probability distribution for state j, $\mathcal{N}(:\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the Gaussian density function with mean $\boldsymbol{\mu}$ and diagonal covariance matrix $\boldsymbol{\Sigma}$, K is the number of mixtures, M_g is the number of gesture classes, and ω_{jk} is the weight of the k-th Gaussian component, such that they sum up to 1 over all the components ($\sum_{k=1}^{K}\omega_{jk}=1$).

5.3.2 Joint Structure

Feature level data fusion is one main information combining scheme for closely coupled and synchronized modalities. In this study, feature level fusion of the prosody and the continuous affect attributes is defined in the joint structure Λ^{ap} as given in Figure 5.2(c) with the joint emission models. The concatenated 9D feature set is denoted as

$$\boldsymbol{f}^{AP}=\begin{bmatrix}\boldsymbol{f}^a\\ \boldsymbol{f}^p\end{bmatrix}. \tag{5.8}$$

The joint emission model is defined using the GMM density function as

$$b_j(\boldsymbol{f}^{AP})=\sum_{k=1}^{K}\alpha_{jk}\mathcal{N}(\boldsymbol{f}^{AP};\boldsymbol{\mu}_{jk}^{(AP)},\boldsymbol{\Sigma}_{jk}^{(AP)}), \tag{5.9}$$

where j runs over the states of the HSMM and α_{jk} values are the mixture weights over K components. The mean vector and the covariance matrix are respectively defined

as

$$\boldsymbol{\mu}_{jk}^{(AP)} = \begin{bmatrix} \boldsymbol{\mu}_{jk}^{(a)} \\ \boldsymbol{\mu}_{jk}^{(p)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{jk}^{(AP)} = \begin{bmatrix} \boldsymbol{\Sigma}_{jk}^{(aa)} \boldsymbol{\Sigma}_{jk}^{(ap)} \\ \boldsymbol{\Sigma}_{jk}^{(pa)} \boldsymbol{\Sigma}_{jk}^{(pp)} \end{bmatrix}, \tag{5.10}$$

where $\boldsymbol{\mu}_{jk}^{(a)}$ and $\boldsymbol{\mu}_{jk}^{(p)}$ are the unimodal mean vectors, $\boldsymbol{\Sigma}_{jk}^{(aa)}$ and $\boldsymbol{\Sigma}_{jk}^{(pp)}$ are the unimodal full-covariance matrices, and the $\boldsymbol{\Sigma}_{jk}^{(ap)}$ is the cross-covariance matrix.

5.3.3 Conditional Structure

We can consider affect as the cause of modifications on the prosody of speech and gesticulation of body motion. Such a causal relationship can be modeled through conditional emission probability function of the prosody given affect observations. The conditional structure $\Lambda^{p|a}$ in Figure 5.2(d) defines this causal relationship. We model the conditional dependency using a GMM based mapping, which estimates an optimal statistical mapping from a set of observed continuous random variables to a target continuous variable. This method was originally introduced for the articulatory-to-acoustic mapping [Toda et al., 2008] and has been applied to a large range of problems, including emotional state tracking [Metallinou et al., 2013]. In this study, our interest is modeling conditional distribution rather than the mapping problem. We model the conditional probability distribution of prosody, $\boldsymbol{f}^p$, given affect, $\boldsymbol{f}^a$, as

$$b_j(\boldsymbol{f}_i^p|\boldsymbol{f}_i^a) = \sum_{k=1}^{K} P(k|\boldsymbol{f}_i^a, j) b_j(\boldsymbol{f}_i^p|\boldsymbol{f}_i^a, k), \tag{5.11}$$

where indexes i, j and k are respectively running over the frames, states and mixture components. In (5.11) $P(k|\boldsymbol{f}_i^a, j)$ is the occupancy probability of the k^{th} mixture at state j and it is defined as

$$P(k|\boldsymbol{f}_i^a, j) = \frac{\alpha_{jk} \mathcal{N}(\boldsymbol{f}_i^a; \boldsymbol{\mu}_{jk}^{(a)}, \boldsymbol{\Sigma}_{jk}^{(a)})}{\sum_{n=1}^{K} \alpha_{jn} \mathcal{N}(\boldsymbol{f}_i^a; \boldsymbol{\mu}_{jn}^{(a)}, \boldsymbol{\Sigma}_{jn}^{(a)})}. \tag{5.12}$$

The conditional distribution is also defined as a Gaussian,

$$b_j(\boldsymbol{f}_i^p|\boldsymbol{f}_i^a, k) = \mathcal{N}(\boldsymbol{f}_i^p; \mathbf{X}_{jki}^{(p)}, \mathbf{Y}_{jk}^{(p)}), \tag{5.13}$$

where the mean vector $\mathbf{X}_{jki}^{(p)}$ and covariance matrix $\mathbf{Y}_{jk}^{(p)}$ are defined as

$$\mathbf{X}_{jki}^{(p)} = \boldsymbol{\mu}_{jk}^{(p)} + \boldsymbol{\Sigma}_{jk}^{(pa)}\boldsymbol{\Sigma}_{jk}^{(aa)^{-1}}(\boldsymbol{f}_i^a - \boldsymbol{\mu}_{jk}^{(a)}), \tag{5.14}$$

$$\mathbf{Y}_{jk}^{(p)} = \boldsymbol{\Sigma}_{jk}^{(pp)} - \boldsymbol{\Sigma}_{jk}^{(pa)}\boldsymbol{\Sigma}_{jk}^{(aa)^{-1}}\boldsymbol{\Sigma}_{jk}^{(ap)}. \tag{5.15}$$

As we point out earlier, this study is a proof of concept to identify the role of affect in speech-driven gesture animation. Hence, the ground truth annotations of the affect attributes are set as the primary affect features. However, for an automated speech-driven gesture animation system, we also investigate performance of the estimated affect attributes, as well. The estimated AVD attributes are used to drive the affect-expressive gesture animation system with the conditional structure, which is denoted as $\Lambda^{p|\hat{a}}$ in our experimental evaluations.

5.4 Experimental Results

We use the multimodal USC CreativeIT dataset that contains a variety of dyadic theatrical improvisations for studying expressive behaviors and natural human interaction [Metallinou et al., 2010], [Metallinou et al., 2015]. The interactions performed by the pairs of actors are either improvisations of scenes from theatrical plays or theatrical exercises where actors each repeat a short sentence to express the interaction goals that emphasize specific emotions.

The dataset contains vocal and body-language behavior information of the actors obtained through close-up microphones, motion capture and HD cameras. Each recording is annotated with dimensional emotional attributes (activation, valence, dominance), where multiple annotaters annotate the videos and average of the attributes over the annotators are defined as the ground truth annotations. There are eight pairs of speakers in the dataset and we perform speaker-independent evaluations in a leave-one-pair-out manner using data from one actor pair as the test set in turn, and the remaining data from the other pairs as the training data. At each turn, we use all data from training set actors (theatrical play and theatrical exercise recordings), and for test purposes we synthesize gestures only for theatrical play recordings

since each actor expresses more consecutive gestures in these recordings due to longer duration interactions.

5.4.1 Gesture Clustering Evaluations

Gesture segments, which are the building block of our animation system and output of the gesture clustering step, are useful for representing meaningful arm motion behaviors [Yang et al., 2014, Yang and Narayanan, 2016]. Gesture clustering, which segments motion data and constructs gesture classes, is crucial for the quality of our animations. Each gesture class should ideally represent a distinct temporal motion pattern and variety of affective gesture forms should be preserved. We apply unsupervised clustering over the gesture sequences using parallel-branch HMMs as described in Section 5.2.1, where we investigate an optimal value for the number of gesture classes, M_g. We consider correlation of motion features in each gesture class with the corresponding affect features for setting the value of M_g. While we aim for a M_g value as low as possible considering computational costs, we like to maximize the correlation between motion and affect features within gesture classes as much as possible that is measured using the normalized Canonical Correlation Analysis (CCA). All the gesture segments, which belong to gesture class g_i, are concatenated and the corresponding rotation angle information is represented as a feature vector sequence $\boldsymbol{F}^{g_i}$. Similarly, the corresponding affect feature sequence for gesture g_i and for affect attribute a, which can be A, V, or D, is constructed as $\boldsymbol{F}^{a_i}$. Then, the CCA score between motion and affect features can be computed and represented as $\rho_i^{cca}(\boldsymbol{F}^{g_i}, \boldsymbol{F}^{a_i})$ for gesture class g_i and for affect attribute a. Then, we can define a weighted global correlation score over all the gesture classes as

$$\bar{\rho}_m^a = \sum_{i=1}^{m} \omega_i \rho_i^{cca}(\boldsymbol{F}^{g_i}, \boldsymbol{F}^{a_i}) \tag{5.16}$$

where m is the total number of gesture classes varying in the range [10,90], and ω_i is the weight of each gesture class g_i. The weight values are computed as the ratio of gesture class g_i's duration to the total duration of gestures, $\omega_i = T_i/(\sum_j T_j)$, where

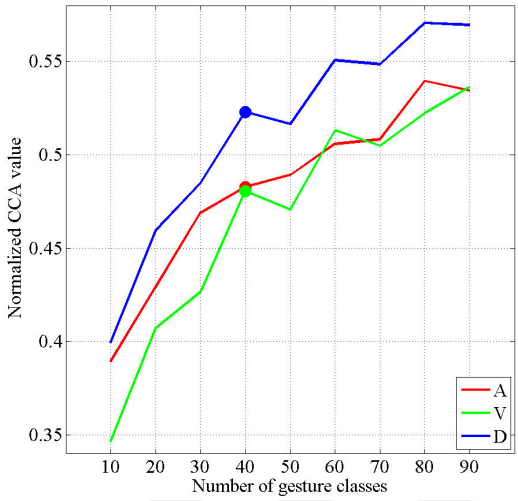


Figure 5.3: The weighted global correlation scores between motion features and affect attributes as a function of number of gesture classes: (A) Activation, (V) Valence, and (D) Dominance.

T_i is the total duration of gesture class g_i.

The weighted global correlation scores, $\bar{\rho}_m^a$, for each affect attribute, a, as a function of number of gesture classes, m, from 10 to 90 are given in Figure 5.3. Note that the weighted global correlation score rapidly increases with the first four number of gesture classes from 10 to 40. Then, the increase slows down for higher number of classes. Additionally, number of classes higher than 50 greatly increases the computational cost for our gesture synthesis approach. We observe that lower number of classes fail to preserve motion and affect correlation, whereas higher number of classes introduce a considerable amount of computational cost to the gesture synthesis. Hence, we set the number of gesture classes as $M_g = 40$, which maintains acceptable correlation values for activation, valence and dominance attributes, which are computed respectively as 0.48, 0.48, and 0.52.

5.4.2 Correlation-based Evaluations

In this study we investigate several approaches for incorporating affective information in a gesture synthesis-animation system as shown in Table 5.1 and objectively evaluate

each approach. We design six scenerios (S) where affect fused with prosody (f^{ap}), prosody only (f^{p}) and affect only (f^{a}) features drive the gesture synthesis and affect may/not ($\alpha = 1/\alpha = 0$) contribute to unit selection based animation generation step. In other words, we consider two options: affect attributes can be directly included as an input feature to the synthesis system (S1, S2, S5, S6) and/or they can be influential during the selection of the gesture samples in the animation generation part after the synthesis step (S1, S3, S5). We set coefficient values managing the contribution of cost scores in Section 5.2.5 as $\beta_1 = 0.30$, $\beta_2 = \beta_3 = 0.35$ and perform speaker-independent evaluations in a leave-one-pair-out manner using data from one actor-pair as the test set in turn, and the remaining data from the other pairs as the training data.

Table 5.1: Experimental set-up for evaluating the contribution of affect in gesture synthesis and animation.

		Affect Contribution in Animation Generation	
		$\alpha = 1$	$\alpha = 0$
Features for Gesture Synthesis	f^{ap}	S1	S2
	f^{p}	S3	S4
	f^{a}	S5	S6

We employ CCA and kinetic energy differences for evaluating the animation results. First, we compare synthesized and the original gesture sequence trajectories based on first order-canonical correlation means and standard deviations for different scenarios as shown in Table 5.2. We observe that feature sets including affect attributes as the driving factor in the synthesis step (S1,S2,S5, and S6) yield higher first-order canonical correlation mean values compared to the ones that do not include (S3, S4). Moreover, including affect information in the animation step as an adjusting factor also yields higher correlations (S1,S3) compared to ones that do not (S2, S4) except the scenerios where affect is the only factor driving the synthesis process (S5

vs. S6). This small decay may be due to forcing the candidate gesture selection in a smaller pool in S5 than in S6.

Table 5.2: Average first-order canonical correlation mean and (std) values for synthesized and original gesture sequences

	S1	**S2**	**S3**	**S4**	**S5**	**S6**
mean	0.566	0.516	0.495	0.461	0.614	0.627
std	0.111	0.119	0.104	0.095	0.040	0.080

Secondly, we calculate the CCA values for the synthesized gesture trajectories and affect attributes A, V, and D in Table 5.3. Similar to results in the previous table, system driven by affect features perform better compared to other systems. Moreover, using affect information in both gesture synthesis and animation generation steps gives better results. Valence (V) has the lowest correlation value compared to activation (A) and dominance (D) domains for systems including prosody as the driving factor for synthesis and highest for the system driven by only affect. This result can be interpreted as gestures in our database convey valence better than prosody.

Table 5.3: CCA for synthesized gesture sequences and affect attributes over the whole sequence

	S1	**S2**	**S3**	**S4**	**S5**	**S6**
A	0.520	0.449	0.354	0.292	0.585	0.545
V	0.482	0.456	0.329	0.287	0.593	0.558
D	0.516	0.483	0.308	0.299	0.588	0.541

Lastly, we compare kinetic energy (KE) differences of the synthesized sequences with the original ones. We compute kinetic energy as the sum of angular velocity values' squares per joint. Then, we calculate frame-level energy differences between the original and the synthesized sequences and normalize the total value by dividing with the length of the sequence. In parallel to results in Tables 5.2 and 5.3 affect only

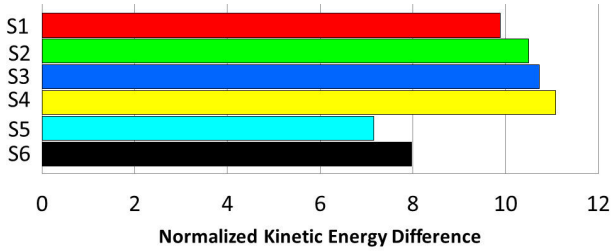


Figure 5.4: Normalized kinetic energy difference of synthesized and the original gesture sequences for the six scenarios.

driven synthesis systems have less KE difference in Figure 5.4. Moreover, employing affect information either during the synthesis or animation process decreases the KE difference.

5.4.3 Objective Evaluations

In the proposed HSMM-based affect expressive gesture synthesis system, we both target to model gesture durations and relationship between gesture and prosody in a realistic manner. To this extent, we first evaluate the fit between the original and the synthesized gesture duration statistics. Recall that the state dependent duration probability mass functions, $d_i(n)$, are defined in (5.3), which are statistics of the original gesture durations. Similarly, duration probability mass functions for the synthesized gestures can be computed and represented as $\hat{d}_i(n)$ for each gesture g_i. Then, the symmetric Kullbeck-Leibler divergence (KLD) between these two duration probability mass functions can measure the fit between the original and the synthesized gesture duration statistics. A mean KLD distance is defined over all gesture classes for the duration statistics as

$$KLD^d = \sum_{i=1}^{M_g} \omega_i KLD(d_i(n), \hat{d}_i(n)), \tag{5.17}$$

where $KLD(d_i(n), \hat{d}_i(n))$ is the symmetric KLD distance for duration statistics of gesture class g_i and ω_i is the weight of each gesture class g_i as used in (5.16).

Table 5.4 presents the mean KLD distance values of duration statistics for the

proposed HSMM structures in Section 5.3. The conditional structure with the ground-truth affect annotations, $\Lambda^{p|a}$, attains the smallest mean KLD distance. Also note that the conditional structure with the estimated AVD attributes, $\Lambda^{p|\hat{a}}$, performs close to the $\Lambda^{p|a}$ structure and attains the second smallest mean KLD distance. These two observations suggest that conditional emission probability model performs better that the unimodal and joint structure models.

Table 5.4: Mean KLD distances of duration statistics for the proposed HSMM structures

KLD^d				
Λ^p	Λ^a	Λ^{ap}	$\Lambda^{p\|a}$	$\Lambda^{p\|\hat{a}}$
0.147	0.188	0.162	0.139	0.140

The second objective evaluation investigates the relationship between gesture classes and speech prosody. Note that the main functionality of the HSMM structure is to model joint articulation of speech prosody and arm gestures. Hence, we construct joint probability distributions of gesture classes and speech prosody over the original and synthesized recordings and compute the mean symmetric KLD distances between these two distributions.

In our framework gesture classes are discrete, whereas, the prosody representation is continuous. In order to compute joint probability mass function, we first extract a discrete representation for the prosody features. This is performed by quantizing prosody features into Q clusters using the k-means clustering algorithm. The joint probability mass function of gesture and prosody is extracted as

$$P(g_i, p_j) = \frac{C(i,j)}{\sum_{i'} \sum_{j'} C(i',j')} \tag{5.18}$$

where $C(i,j)$ is the total count of frames with gesture class g_i and prosody cluster p_j. Then, the KLD distance for gesture-prosody relationship is defined as

$$KLD^{gp} = KLD(P(g,p), P(\hat{g},p)), \tag{5.19}$$

where $P(g,p)$ and $P(\hat{g},p)$ are the joint probability mass functions respectively for the original and synthesized gesture sequences.

Table 5.5 presents the KLD distances of gesture-prosody relationship for the proposed HSMM structures in Section 5.3. Again, the conditional structures, $\Lambda^{p|a}$ and $\Lambda^{p|\hat{a}}$, attain consistently the smallest KLD distances over all prosody quantization levels.

Table 5.5: KLD distances of gesture-prosody relationship for the proposed HSMM structures over different prosody quantization levels

	KLD^{gp}				
Q	Λ^{p}	Λ^{a}	Λ^{ap}	$\Lambda^{p\|a}$	$\Lambda^{p\|\hat{a}}$
16	3.42	3.86	3.29	2.95	2.83
32	3.13	3.53	3.01	2.72	2.59
64	2.83	3.18	2.73	2.48	2.36

5.4.4 Subjective Evaluation

Gestures can accompany speech in various different ways. Objective evaluations are incapable of qualifying such variabilities, whereas subjective tests can evaluate realism and naturalness of the animation by reflecting human perception. We use mean opinion score (MOS) tests to subjectively evaluate the four synthesis methods mentioned in Section 5.3 in addition to the original motion sequence. We segment audio recordings based on speaker turns in the dialogs and ensure only a single speaker speaks at a time. The total number of test audio segments is 29 with an average duration of 15s. The test contains 27 clips per test session, where the test takers assess how expressive and coherent arm gestures are with speech. We include two training clips at the beginning of the test, one of which is a high quality sample and the other is a low quality in terms of consistency with speech. Each scenario has five samples in every test session and all clips, except the training clips are shown

in random order. A five-point assessment is adopted (1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Very Good). Then, 25 test takers were asked to judge whether the animation was expressive and natural. We did not perform simultaneous side-by-side presentations (A/B test) of the synthesized gesture animations, since the test takers would have shifted their gaze from one to the other while the utterance was played. Our previous experience with the A/B tests show that the test takers may focus on segments of animations, while missing the impression animations would have given throughout the video.

Table 5.6 shows distribution and mean of test scores per scenario, where test takers give relatively high scores to the synthesized gesture animations generated by models using prosody features conditioned on the affect features, $\Lambda^{p|a}$, compared to those of generated with unimodal or jointly modeled prosody, affect features. On average, 33 % of the original motion sequence animations and 21 % of $\Lambda^{p|a}$ models synthesized animations are given the highest score. Moreover, among the three scenarios, which synthesize gesture animations from speech, the proposed conditional model, $\Lambda^{p|a}$, performs the best with a mean opinion score (MOS) of 3.752. The MOS test score for the animations generated using original motion sequence is 3.928. The lowest test scores belong to animations based on the Λ^{a} models. We deduce this low performance may be related to failure of affect attributes in modeling timing of the gestures, since gesture sequences evolve more rapidly than affect attributes in time.

We use Analysis of Variance (ANOVA) tests to analyze subjective test results given in Table 5.6. The test reveals that subjective test scores are significantly different for the five methods ($F[4,620]=29.48$, $p = 1.1102e\text{-}16$). Finally, post-hoc tests with Tukey-Kramer method was used for multiple comparisons between scenarios to assess which methods exhibit statistically significant differences between one another as shown in Table 5.7. The test scores based on the proposed $\Lambda^{p|a}$ models are not significantly different from the original motion sequence scenario scores, and significantly different from other models' scores. Additionally, scores based on models, Λ^{p} and Λ^{ap} are not significantly different from each other.

Table 5.6: Mean opinion score (MOS) subjective test results (Scores, 5=Very good, 1=Bad).

	Distribution of scores per method					
Method	**5**	**4**	**3**	**2**	**1**	**MOS**
Original	0.33	0.39	0.18	0.1	0.0	3.928
$\Lambda^{p\|a}$	0.21	0.45	0.24	0.1	0.0	3.752
Λ^{p}	0.13	0.33	0.30	0.17	0.07	3.304
Λ^{ap}	0.08	0.26	0.34	0.20	0.12	2.968
Λ^{a}	0.01	0.26	0.30	0.30	0.13	2.736

Table 5.7: p-values for the ANOVA (* The mean difference is significant at p=0.01 level)

	Λ^{ap}	Λ^{p}	Λ^{a}	Original
$\Lambda^{p\|a}$	0.001(*)	0.006(*)	0.001(*)	0.646
Λ^{ap}		0.080	0.397	0.001(*)
Λ^{p}			0.001(*)	0.001(*)
Λ^{a}				0.001(*)

5.5 Summary

We proposed a framework for affective speech driven arm gesture synthesis and animation. The proposed system is speaker-independent and fully-automatic, which inputs speech with its affective attributes and outputs speech-synchronous animations of expressive arm gestures. Our framework is composed of analysis, synthesis and animation stages. In the analysis stage, we first segment and cluster gesture sequences into meaningfull patterns using the parallel branch HMM structures (Section 5.2.1) and model unimodal affect, unimodal prosody, joint prosody and affect, and prosody conditioned on affect features for each gesture class (Section 5.3). We

analyze the relationship between gestures, speech and affect attributes (i.e. activation, valence, dominance) using the hidden semi-Markov models. The synthesis stage in Section 5.2.4, uses these models and inputs speech prosody and its affective attributes to synthesize gestures with id and duration information. Then in the animation stage, which is detailed in Section 5.2.5, gesture animations are generated from the synthesized gesture sequences using the unit selection method.

We evaluate our system in a leave-one-pair speaker-out manner on the USC CreativeIT dataset, which contains speech and motion-capture data recorded during goal-driven improvised dyadic interactions. These recordings are annotated in the activation, valence, and dominance domains. However, it is costly to manually annotate new input speech data for synthesizing novel gestures. Therefore, we also estimate values of affect attributes in our framework using the Support Vector Regressors. One of the challenges for synthesizing expressive gestures from affective speech include possible differences in the distribution of affect information for training and test data. Since we employ a speaker-independent setting, the distributions of affect for training and test sets may differ. Unless vast amount of affective multimodal data is available, joint models of prosody and affect attributes may fail to cover whole emotion space. Therefore, we use conditional models of prosody given affect in our framework. Then, in Section 5.4.3, we objectively evaluate gesture synthesis results using models for prosody-only, affect-only, joint prosody and affect, prosody conditioned on affect and prosody conditioned on estimated affect, on the basis of their duration distribution similarity to the original duration distributions, and on the basis of joint gesture-prosody class distributions. We use normalized Kullbeck - Leibler divergence scores to measure the similarity of synthesized and the original distributions. Models using prosody conditioned on affect and prosody conditioned on estimated affect yield the best synthesis results in the objective evaluations. On the other hand, we observe synthesis results based on affect only-driven models achieve the highest KL-divergence scores in both of the objective assessments. We infer this shortcoming may be related to failure of affect attributes in modeling the timing of

gestures with respect to speech; gesture sequences evolve more rapidly than affect attributes in time and synchrony between prosody and gestures is lost for this set up. This result also indicates the importance of accurate timing and alignment of gestures and speech for the believability gesture animations in CAs.

Subjective evaluations, in Section 5.4.4, indicate that conditional modeling of prosody given the affect attributes in an HSMM based gesture synthesis and animation framework can generate good quality animations. 21 % of the animations were assessed to have very high quality in terms of their expressiveness and their coherence with the accompanying speech and 45 % of them as good quality in the mean opinion score tests. Moreover, the same setting was not significantly different from the original motion sequence according to the ANOVA test results. Besides, affect-driven animations were mostly rated at average quality. We believe animation quality in our framework mostly depends on the synthesis results. Reliable synthesis results yield to high quality animations, whereas animations generated from imprecise synthesis results fail to achieve required expressivity in arm gestures.

Chapter 6

CONCLUSIONS

6.1 Summary

This section presents a summary of results and contributions presented in this thesis. As mentioned in the introduction chapter, we present an affective speech-driven gesture synthesis and animation framework.

Chapter 3 presents speech-driven gesture synthesis and animation framework components including: feature extraction and clustering, multimodal analysis of gestures and prosody, gesture synthesis and animation. We use two multimodal datasets for speaker-dependent and independent applications, respectively. We use gesture phrases and gesture segments as the building block of speaker-dependent and independent animation systems, respectively. In our framework, all analysis, synthesis and animation generation steps revolve around the gesture phrase (segment) patterns.

Chapter 4 presents a gesture animation process using speech rhythm information. Speech rhythm features extraction steps and contribution of using rhythm information to animation results are explained. We also propose objective evaluation methods and perform subjective tests to assess the quality of the animations. Both objective and subjective tests show that animations using rhythm information are more realistic.

Chapter 5 presents affect attributes as a driving factor for the gesture synthesis process in addition to prosody features. We investigate the contribution of using affect information in four different models: prosody-only, affect-only, joint affect and prosody and prosody conditioned on affect. We objectively evaluate the synthesis results of the four methods, and models using prosody-conditioned on affect have the most contribution to the synthesis results. Even if, we use estimated values of affect attributes, we get better performance with the models using prosody conditioned

on affect compared to other three methods. Subjective test results also reveal that conditional modeling yields more realistic animations, which are coherent with the accompanying speech.

As mentioned in the introductory chapter, this manuscript contains

- an affective speech-driven gesture synthesis and animation framework
- investigation of ways of including affect information in a general gesture synthesis framework
- objective and subjective evaluations to assess contributions of including affect into the system

BIBLIOGRAPHY

[Mot, 2012] (2012). MotionBuilder: 3D Character Animation for Virtual Production, http://www.autodesk.com.

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

[Ananthakrishnan and Narayanan, 2008] Ananthakrishnan, S. and Narayanan, S. (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):216–228.

[Baena et al., 2013] Baena, A. F., Montano, R., Antonijoan, M., Roversi, A., Miralles, D., and Alias, F. (2013). Gesture synthesis adapted to speech emphasis. *Speech Communication*, 57:331–350.

[Baledassarri and Cerezo, 2012] Baledassarri, S. and Cerezo, E. (2012). Maxine: Embodied conversational agents for multimodal emotional communication. *Computer Graphics*.

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

[Becker et al., 2004] Becker, C., Kopp, S., and Wachsmuth, I. (2004). Simulating the emotion dynamics of a multimodal conversational agent. In *Affective Dialogue Systems*, pages 154–165. Springer.

[Becker et al., 2007] Becker, C., Kopp, S., and Wachsmuth, I. (2007). Why emotions should be integrated into conversational agents. *Conversational informatics: an engineering approach*, pages 49–68.

[Boker et al., 2002] Boker, S. M., Rotondo, J. L., Xu, M., and King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, 7(3):338.

[Bolt, 1980] Bolt, R. A. (1980). Put-that-there: Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '80, pages 262–270, New York, NY, USA. ACM.

[Boone and Cunningham, 1998] Boone, R. T. and Cunningham, J. G. (1998). Children's decoding of emotion in expressive body movement: the development of cue attunement. *Developmental psychology*, 34(5):1007.

[Bos et al., 1994] Bos, E., Huls, C., and Claassen, W. (1994). EDWARD: full integration of language and action in a multimodal user interface. *International Journal of Human-Computer Studies*, 40(3):473–495.

[Bozkurt et al., 2013] Bozkurt, E., Asta, S., Ozkul, S., Yemez, Y., and Erzin, E. (2013). Multimodal Analysis of Speech Prosody and Upper Body Gestures using Hidden Semi-Markov Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3652–3656, Vancouver.

[Bozkurt et al., 2015a] Bozkurt, E., Erzin, E., and Yemez, Y. (2015a). Affect-expressive hand gestures synthesis and animation. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE.

[Bozkurt et al., 2016a] Bozkurt, E., Khaki, H., Kececi, S., Turker, B. B., Yemez, Y., and Erzin, E. (2016a). The jestkod database: An affective multimodal database of dyadic interaction. *Language Resources and Evaluation*.

[Bozkurt et al., 2015b] Bozkurt, E., Yemez, Y., and Erzin, E. (2015b). Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *submitted*.

[Bozkurt et al., 2016b] Bozkurt, E., Yemez, Y., and Erzin, E. (2016b). Affective synthesis and animation of arm gestures from speech prosody. *to be submitted*.

[Busso et al., 2007] Busso, C., Deng, Z., Grimm, M., Neumann, U., and Narayanan, S. S. (2007). Rigid Head Motion in Expressive Speech Animation : Analysis and Synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1075–1086.

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

[Cassell et al., 1994] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '94, pages 413–420, New York, NY, USA. ACM.

[Cassell et al., 2001] Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). BEAT. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, SIGGRAPH'01, pages 477–486, New York, New York, USA. ACM Press.

[Chang and Lin, 2011] Chang, C. and Lin, C. (2011). Libsvm: A library for support vector machines. *ACM Tran. on Intelligent Systems and Technology (TIST)*, 2(3):27.

[Chi et al., 2000] Chi, D., Costa, M., Zhao, L., and Badler, N. (2000). The emote model for effort and shape. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 173–182, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

[Chiu and Marsella, 2014] Chiu, C. C. and Marsella, S. (2014). Gesture generation with low-dimensional embeddings. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 781–788. International Foundation for Autonomous Agents and Multiagent Systems.

[Chuang and Bregler, 2005] Chuang, E. and Bregler, C. (2005). Mood swings: Expressive speech animation. *ACM Trans. Graph.*, 24(2):331–347.

[Dael et al., 2013] Dael, N., Goudbeek, M., and Scherer, K. (2013). Perceived gesture dynamics in nonverbal expression of emotion. *Perception*, 42(6):642–657.

[de Cheveigne and Kawahara, 2002] de Cheveigne, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917.

[De Gelder, 2006] De Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience*, 7(3):242–249.

[Debevec, 2012] Debevec, P. (2012). The light stages and their applications to photoreal digital actors. *SIGGRAPH Asia Technical Briefs*, 2.

[Delaherche et al., 2012] Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., and Cohen, D. (2012). Interpersonal synchrony: A survey of evaluation methods across disciplines. *Affective Computing, IEEE Transactions on*, 3(3):349–365.

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

[Ekman and Friesen, 1975] Ekman, P. and Friesen, W. (1975). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues.* Spectrum books. Prentice-Hall.

[Ekman and Rosenberg, 1997] Ekman, P. and Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press, USA.

[Gibbon and Gut, 2001] Gibbon, D. and Gut, U. (2001). Measuring speech rhythm. In *Eurospeech*, pages 95–98.

[Grabe and Low, 2002] Grabe, E. and Low, E. L. (2002). Duration variability in speech and the rhythm class hypothesis. *Laboratory Phonology*, 7.

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

[Grandjean et al., 2008] Grandjean, D., Sander, D., and Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and cognition*, 17(2):484–495.

[Granström and House, 2006] Granström, B. and House, D. (2006). Measuring and modeling audiovisual prosody for animated agents. In *Proc. of Speech Prosody*. Citeseer.

[Gratch and Marsella, 2001] Gratch, J. and Marsella, S. (2001). Tears and fears: Modeling emotions and emotional behaviors in synthetic agents. In *Proceedings of the fifth international conference on Autonomous agents*, pages 278–285. ACM.

[Greenberg, 1999] Greenberg, S. (1999). Speaking in shorthand: a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29(2-4):159–176.

[Greenberg and Arai, 2004] Greenberg, S. and Arai, T. (2004). What are the essential cues for understanding spoken language? *IEICE Trans. Inf. Syst.*, E87:1059–1070.

[Gunes et al., 2011] Gunes, H., Schuller, B., Pantic, M., and Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 827–834. IEEE.

[Gwo-Dong et al., 2012] Gwo-Dong, C., Lee, J.-H., Chin-Yeh, W., Po-Yao, C., Liang-Yi, L., and Tzung-Yi, L. (2012). An empathic avatar in a computer-aided learning program to encourage and persuade learners. *Journal of Educational Technology & Society*, 15(2):62.

[Hartmann et al., 2005] Hartmann, B., Mancini, M., and Pelachaud, C. (2005). Towards affective agent action: Modelling expressive eca gestures. In *International conference on Intelligent User Interfaces-Workshop on Affective Interaction, San Diego, CA*.

[Hartmann et al., 2006] Hartmann, B., Mancini, M., and Pelachaud, C. (2006). Implementing expressive gesture synthesis for embodied conversational agents. In *Proceedings of the 6th International Conference on Gesture in Human-Computer Interaction and Simulation*, GW'05, pages 188–199, Berlin, Heidelberg. Springer-Verlag.

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

[Hogan and Sternad, 2009] Hogan, N. and Sternad, D. (2009). Sensitivity of smoothness measures to movement duration, amplitude, and arrests. *Journal of Motor Behavior*, 41(6):529–534.

[Hong et al., 2002] Hong, P., Wen, Z., and Huang, T. S. (2002). Real-time speech-driven face animation with expressions using neural networks. *IEEE Transactions on Neural Networks*, 13(4):916–927.

[Iverson and Thelen, 1999] Iverson, J. M. and Thelen, E. (1999). Hand, mouth and brain: The dynamic emergence of speech and gesture. *Journal of Consciousness Studies.*, 6:19–40.

[Kang et al., 2015] Kang, S.-H., Feng, A. W., Leuski, A., Casas, D., and Shapiro, A. (2015). Smart mobile virtual humans:chat with me!. In *Intelligent Virtual Agents*, pages 475–478. Springer.

[Karg et al., 2013] Karg, M., Samadani, A.-A., Gorbet, R., Kuhnlenz, K., Hoey, J., and Kulic, D. (2013). Body movements for affective expression: A survey of automatic recognition and generation. *Affective Computing, IEEE Transactions on*, 4(4):341–359.

[Kendon, 1980] Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In Key, M. R., editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. Mounton Publishers, The Hague, The Netherlands.

[Kipp and Martin, 2009] Kipp, M. and Martin, J.-C. (2009). Gesture and emotion: Can basic gestural form features discriminate emotions? In *Affective Computing*

and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, pages 1–8. IEEE.

[Kleinsmith and Bianchi-Berthouze, 2013] Kleinsmith, A. and Bianchi-Berthouze, N. (2013). Affective body expression perception and recognition: A survey. *Affective Computing, IEEE Transactions on*, 4(1):15–33.

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

[Kopp and Wachsmuth, 2004] Kopp, S. and Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation And Virtual Worlds*, 15:39–52.

[Ladd, 1996] Ladd, R. (1996). *Intonational Phonology*. Cambridge University Press.

[Levine et al., 2010] Levine, S., Krähenbühl, P., Thrun, S., and Koltun, V. (2010). Gesture controllers. *ACM Transactions on Graphics*, 29(4).

[Liberman, 1975] Liberman, M. (1975). *The Intonational system of English*. PhD Thesis.

[Liu et al., 2016] Liu, K., Tolins, J., Fox Tree, J., Neff, M., and Walker, M. (2016). Two techniques for assessing virtual agent personality. *IEEE Transactions on Affective Computing*, 7(1):94–105.

[Loehr, 2012] Loehr, D. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, 3(1):71–89.

[Loukina et al., 2011] Loukina, A., Kochanski, G., Rosner, B., and Keane, E. (2011). Rhythm measures and dimensions of durational variation in speech. *The Journal of the Acoustical Society of America*, 129(5):3258–3270.

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

[Mancini et al., 2011] Mancini, M., Castellano, G., Peters, C., and McOwan, P. W. (2011). Evaluating the communication of emotion via expressive gesture copying behaviour in an embodied humanoid agent. In *International Conference on Affective Computing and Intelligent Interaction*, pages 215–224. Springer.

[Mariooryad and Busso, 2012] Mariooryad, S. and Busso, C. (2012). Generating Human-Like Behaviors Using Joint , Speech-Driven Models for Conversational Agents. *IEEE Transactions on Audio, Speech and Language Processing*, 20(8):2329–2340.

[Marsella et al., 2013] Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., and Shapiro, A. (2013). Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '13, pages 25–35, New York, NY, USA. ACM.

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

[McNeill, 2006] McNeill, D. (2006). *Gesture and Thought*.

[Mehrabian, 1972] Mehrabian, A. (1972). *Nonverbal communication*. Transaction Publishers.

[Mehrabian, 1996] Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

[Metallinou et al., 2013] Metallinou, A., Katsamanis, A., and Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137 – 152. Affect Analysis In Continuous Input.

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

[Metallinou et al., 2015] Metallinou, A., Yang, Z., Lee, C.-c., Busso, C., Carnicke, S., and Narayanan, S. (2015). The usc creativeit database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. *Language Resources and Evaluation*, pages 1–25.

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

[Neff et al., 2008] Neff, M., Kipp, M., Albrecht, I., and Seidel, H.-P. (2008). Gesture Modeling and Animation Based on a Probabilistic Recreation of Speaker Style. *ACM Transactions on Graphics*, 27(1):5:5–5:24.

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

[Noma et al., 2000] Noma, T., Zhao, L., and Badler, N. (2000). Design of a virtual human presenter. *IEEE Computer Graphics and Applications*, 20(4):79–85.

[Noot and Ruttkay, 2005] Noot, H. and Ruttkay, Z. (2005). Variations in gesturing and speech by gestyle. *International Journal of Human-Computer Studies*, 62(2):211–229.

[Ofli et al., 2008] Ofli, F., Demir, Y., Yemez, Y., Erzin, E., Tekalp, A. M., Balci, K., Kizoglu, I., Akarun, L., Canton-Ferrer, C., Tilmanne, J., Bozkurt, E., and Erdem, A. T. (2008). An audio-driven dancing avatar. *Journal on Multimodal User Interfaces*, 2(2):93–103.

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

[Pelachaud, 2005] Pelachaud, C. (2005). Multimodal expressive embodied conversational agents. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 683–689, New York, NY, USA. ACM.

[Pelachaud, 2009] Pelachaud, C. (2009). Studies on gesture expressivity for a virtual agent. *Speech Communication*, 51(7):630–639.

[Port, 2003] Port, R. F. (2003). Meter and speech. *Journal of Phonetics*, 31:599–611.

[Ramus et al., 1999] Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, pages 265–292.

[Rehm and Wissner, 2005] Rehm, M. and Wissner, M. (2005). Gamble a multiuser game with an embodied conversational agent. In *Entertainment Computing-ICEC 2005*, pages 180–191. Springer.

[Reithinger et al., 2006] Reithinger, N., Gebhard, P., Markus, L., Ndiaye, A., Pfleger, N., and Klesen, M. (2006). VirtualHuman Dialogic and Affective Interaction with

Virtual Characters. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI06)*, pages 51–58.

[Rickel and Johnson, 2000] Rickel, J. and Johnson, W. L. (2000). Task-oriented collaboration with embodied agents in virtual worlds. *Embodied conversational agents*, pages 95–122.

[Ringeval et al., 2012] Ringeval, F., Chetouani, M., and Schuller, B. (2012). Novel Metrics of Speech Rhythm for the Assessment of Emotion. In *INTERSPEECH: Annual Conference of the International Speech Communication Association*, pages 2763–2766.

[Russell, 1980] Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

[Sadoughi and Busso, 2015] Sadoughi, N. and Busso, C. (2015). Retrieving target gestures toward speech driven animation with meaningful behaviors. In *International conference on Multimodal interaction (ICMI 2015)*, pages 115–122, Seattle, WA, USA.

[Sargin et al., 2008] Sargin, M. E., Yemez, Y., Erzin, E., and Tekalp, A. M. (2008). Analysis of Head Gesture and Prosody Patterns for Prosody-Driven Head-Gesture Animation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1330–1345.

[Scherer et al., 2001] Scherer, K. R., Schorr, A., and Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

[Schroder et al., 2012] Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Ter Maat, M., McKeown, G., Pammi, S., Pantic, M., et al. (2012). Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165–183.

[Stone et al., 2004] Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., and Bregler, C. (2004). Speaking with hands: Creating animated conversational characters from recordings of human performance. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, pages 506–513, New York, NY, USA. ACM.

[Tilsen and Johnson, 2008] Tilsen, S. and Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *The Journal of the Acoustical Society of America*, 124(2):34–39.

[Toda et al., 2008] Toda, T., Black, A. W., and Tokuda, K. (2008). Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. *Speech Communication*, 50(3):215 – 227.

[Tracy and Robins, 2007] Tracy, J. L. and Robins, R. W. (2007). The prototypical pride expression: development of a nonverbal behavior coding system. *Emotion*, 7(4):789.

[Tuite, 1993] Tuite, K. (1993). The production of gesture. *Semiotica*, 93(1-2):83–106.

[Valbonesi et al., 2002] Valbonesi, L., Ansari, R., Mcneill, D., Quek, F., Duncan, S., McCullough, K. E., and Bryll, R. (2002). Multimodal signal analysis of prosody and hand motion: temporal correlation of speech and gestures,. In *Proc. Eur. Signal Process. Conf. (EUSIPCO 02)*, pages 75–78.

[Vinayagamoorthy et al., 2006] Vinayagamoorthy, V., Gillies, M., Steed, A., Tanguy, E., Pan, X., Loscos, C., Slater, M., et al. (2006). Building expression into virtual characters. In *Eurographics Conference State of the Art Report*, Vienna, Austria.

[Vinciarelli et al., 2012] Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., and Schroeder, M. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87.

[Wagner et al., 2014] Wagner, P., Malisz, Z., and Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, 57:209 – 232.

[Wallbott, 1998] Wallbott, H. G. (1998). Bodily expression of emotion. *European journal of social psychology*, 28(6):879–896.

[Weise et al., 2011] Weise, T., Bouaziz, S., Li, H., and Pauly, M. (2011). Realtime performance-based facial animation. In *ACM Transactions on Graphics (TOG)*, volume 30, page 77. ACM.

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

[Yang et al., 2014] Yang, Z., Metallinou, A., Erzin, E., and Narayanan, S. (2014). Analysis of interaction attitudes using data-driven hand gesture phrases. In *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*.

[Yang and Narayanan, 2016] Yang, Z. and Narayanan, S. S. (2016). Modeling dynamics of expressive body gestures in dyadic interactions. *IEEE Transactions on Affective Computing*.

[Yu, 2010] Yu, S. Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, 174(2):215–243.