Affective Synthesis and Animation of Arm Gestures from Speech Prosody

Elif Bozkurt, Yücel Yemez, and Engin Erzin

Multimedia, Vision and Graphics Laboratory, College of Engineering, Koç University, Sarıyer, İstanbul, 34450, Turkey

Abstract

In human-to-human communication, speech signals carry rich emotional cues that are further emphasized by affect-expressive gestures. In this regard, automatic synthesis and animation of gestures accompanying affective verbal communication can help to create more naturalistic virtual agents in human-computer interaction systems. Speech-driven gesture synthesis can map emotional cues of the speech signal to affect-expressive gestures by modeling complex variability and timing relationships of speech and gesture. In this paper, we investigate the use of continuous affect attributes, which are activation, valence and dominance, for speech-driven affective synthesis and animation of arm gestures. To this effect, we present a statistical framework based on hidden semi-Markov models (HSMM), where states are gestures and observations are speech-prosody and continuous affect attributes. The proposed framework is evaluated considering four distinct HSMM structures which differ by their emission distributions. Evaluations are performed over the USC CreativeIT database in a speaker-independent setup. Among the four statistical structures, the conditional structure, which models observation distributions as prosody given affect, achieves the best performance under both objective and subjective evaluations.

Keywords: Prosody analysis, gesture segmentation, arm gesture animation,

*Engin Erzin

Email address: ebozkurt,yyemez,eerzin@ku.edu.tr (Elif Bozkurt, Yücel Yemez, and Engin Erzin)

[1]This work was supported by TUBITAK under Grant Number 113E102.

unit selection, hidden semi-Markov models, speech-driven affective gesture synthesis

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 [1, 2]. Likewise, virtual agents (VA) 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 [3, 4]. For example, an empathic VA can encourage and persuade students while using e-learning systems [5]. Such VA 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 arm 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 [6, 7]. McNeil defines four types of gestures: iconics, metaphorics, deictics and beats [7]. 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. One early study discusses the synchrony between strokes and stressed syllables [8]. Later, the phonological synchrony rule, which says the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech, has been widely accepted in the literature [7]. The nature of temporal relationship between speech and gestures has been investigated based on this rule in [9].

Existing methods for synthesizing gestures in VAs can mainly be grouped as text-driven [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21] and audio speech-

driven [22, 23, 24, 25, 26, 27, 28, 29]. Some of these studies also consider affect expressiveness of gestures in emotional categories [10, 12, 13, 14, 15, 17, 18, 19, 20, 21, 24, 23]. However, categorical emotion models would fail to accurately describe non-basic states like thinking, embarrassment and depression. On the other hand, dimensional models of affect, such as activation, valence and dominance attributes, deliver continuous scale representations, which are useful for explaining non-categorical complex affective states [30]. In the dimensional model, activation, valence and dominance are respectively described along the active-passive, positive-negative and dominant-submissive dimensions [31].

Currently, linguistic features (e.g. lexical or syntactic features) are predominant for automatic synthesis of affect expressive gestures in VAs [30]. For example, in case of specifying prosody and non-verbal signals of synthesized speech for an answer to a yes/no question, VA's speech would be accompanied with a head nod for a yes answer [32, 33]. The affective expression of the VA 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 [13, 14, 17, 18, 19, 21]. However, targeting VAs' gestures based only on the lexical content input 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 affective VAs. In this regard, speech-driven gesture synthesis approaches are useful in creating realistic gestures given the rich information conveyed in human speech such as emotional cues and prosodic patterns, which are essential for modeling the variability and timing of the VAs' gestures [6]. Although there are studies analyzing gestures in relation to affect attributes [34], speech-driven gesture synthesis methods mostly consider neutral speech [22, 25, 26, 29], or agitation [24] and intensity [23] levels of speech. In the current literature, use of speech and gesture to enhance affective expressiveness of VAs remains as an open problem.

The main objective of our current work is to investigate ways of integrating

affect attributes (i.e., activation (A), valence (V) and dominance (D)) into a general framework so as to generate more affect-expressive gesture synthesis from affective speech. Our computational model is based on the framework that we previously developed in our earlier works [25, 27, 28] for speech-driven gesture synthesis and animation. We jointly analyze gestures with continuous affect attributes and speech prosody using hidden semi-Markov models (HSMMs). 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 system (along with [27] which is a preliminary version of this current work). In this work, we also introduce new objective evaluation methods for quantifying the quality of synthesis results and perform 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 other strategies that we consider for incorporating affect into our framework.

2. Related Work

Neuroscientific and psychological studies reveal that arm gesture and its expressivity constitute an important modality of emotion communication [35]. For example, joy may bring to openness and upward acceleration of the forearms [36]. More specifically, Kipp and Martin investigate the relationship between basic gestural forms and emotion, where the authors report that handedness is closely correlated with emotion categories [37]. Similarly, Dael et al. also suggest that emotional attributes are associated with specific spatio-temporal characteristics perceived in arm gesture movements [38].

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 is generally specified by using parameters [13, 14, 17, 18, 19], or by selecting appropriate gesture types [21, 12, 15, 10] depending

on the emotional state as defined in an input markup text file. In addition to text-driven approaches, speech-driven gesture synthesis systems have been proposed, which utilize prosody features as input yet exclude explicit emphasis on affect [22, 23, 26, 25]. There are also studies that combine semantics with prosody for synthesizing typical gestures [24, 29, 39]. In the remaining part of this section, we first discuss the literature on expressive gesture synthesis and animation, which is mostly text-based, and then give a summary of the existing speech-based methods, including a discussion of our previous work on gesture synthesis and animation.

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 [40]. However, EMOTE is for modifying but not for generating gestures. Besides, modifying the expressivity of a gesture may alter the intended meaning in non-verbal communication. Greta on the other hand, is an embodied conversational agent (ECA) that communicates through her face and gestures while talking to the user via text-to-speech [13, 14, 17, 18, 19]. 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 are 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 VAs is the SEMAINE system that employs Sensitive Artificial Listeners (SAL) scenario for studying emotional and non-verbal behavior [20]. In this system a behavior lexicon associates communicative functions with the corresponding multimodal signals that a VA 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 specifies when and how the VA shall accompany certain gestures [21]. Emotional states, which are defined as entries in a style dictionary, are used for designing speech and non-verbal modalities. Max is a VA that uses a direct mapping between emotional states, mood, boredom level and behaviors to modulate gestures, facial expression and speech in time [12, 15]. 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 VA to drive the finite state machine that determines behaviors [10].

Speech-driven gesture animation has been studied in the recent literature mostly without specific emphasis on affect-expressive models. In [22], 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. Later, Baena et al. have presented a single stage model that links speech prosody to beat gestures based on manually annotated body motion and speech signals [23]. 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 [26]. They use conditional random fields (CRF) for mapping speech to gesture annotations. Then, Gaussian process latent variable models (GPLVMs) are used for motion synthesis.

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 [24]. 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, where 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 [29]. Chiu et al. present a deep and temporal model that realizes the mapping relation between speech and gestures as well as temporal relations among gestures by using linguistic and prosodic features [39].

In line with the work of Levine et al. [22] and in an attempt to better model the relationship of prosodic patterns with longer duration gesture phrases, we previously introduced a hidden semi-Markov model (HSMM) based speech-driven upper-body gesture synthesis system [25]. We later extended this work to include rhythm information from speech as well as prosody with extensive objective and subjective evaluations [28]. More recently in [27], we identified affect and prosody feature fusion as the most correlated feature set with the original gesture trajectories based on the canonical correlation analysis (CCA). This observation was encouraging to investigate deeper for the affective gesture synthesis and animation. Except this preliminary study, the problem of affective speech-driven gesture synthesis has not yet explicitly been addressed in the literature, and remains as an open research challenge that we intend to tackle in this current paper. In summary, our contributions are as follows:

1. We investigate ways of exploiting affective content of speech to enhance expressiveness of synthesized gestures. We use continuous description of affect in combination with speech to drive a gesture synthesis and animation system. To this effect, we employ ground-truth annotated affect attributes as well as affect attributes estimated from speech prosody.
2. We propose four distinct emission model structures, which are i) unimodal prosody, ii) unimodal affect, iii) joint affect and prosody, and iv) prosody given affect conditional structures, to investigate contribution of affect information in our HSMM-based gesture synthesis and animation framework. Among them the prosody given affect conditional structure yields the best performance.
3. We propose three new objective metrics to assess the quality of the synthe-

sized affect expressive gestures. The first metric is a weighted correlation score between gesture and affect to set the number of gesture classes. The other two metrics are mean symmetric Kullback-Leibler divergence (KLD) distances to evaluate modeling performance of the HSMM structure. The second objective metric is the weighted mean duration KLD distance between the original and the synthesized gesture duration statistics. We also provide a weighted standard deviation for the duration KLD distance. Finally, the third objective metric is the gesture-prosody KLD distance, which tests goodness of the joint gesture and prosody distribution as an output of the gesture synthesis framework.

4. We perform subjective tests to evaluate the quality of the speech and affect driven animations.

The remainder of this paper is organized as follows. Section 3 presents an overview of the HSMM based gesture synthesis and animation framework. Section 4 describes the proposed affect expressive gesture synthesis and animation system. Section 5 presents the experimental evaluations and discussions. Finally, Section 6 concludes the paper.

3. HSMM based Gesture Synthesis and Animation Framework

In this section, we present a brief summary of our speech-driven gesture synthesis and animation framework that was previously introduced in [25, 28]. Our framework has three components in general: (i) gesture-speech multimodal analysis, (ii) speech-driven gesture synthesis and (iii) gesture animation. The affect-expressive gesture synthesis and animation system that we will later describe in Section 4 is based on this general HSMM-based framework.

3.1. Gesture-Speech Multimodal Analysis

Our gesture input is a feature stream of joint angles of arms and fore-arms (See Fig. 1), which are extracted from motion capture data. We perform an unsupervised temporal gesture clustering based on the parallel-branch HMM

on the joint angle feature stream, so as to map arm motion data to a sequence of gesture segments, where each gesture segment belongs to one of the M gesture classes [41, 27, 28]. Gesture clustering groups similar gestures based on their joint angles and velocities, hence characterizes them with objective properties like handedness, arm orientation and motion direction.

The parallel-branch HMM is composed of M branches, where each branch models a gesture class and has left-to-right states. In this architecture, transitions from any branch (gesture class) to another, as well as self-transitions, are allowed. The number of branches (classes) is predetermined prior to the clustering process. The input feature stream is segmented using the Viterbi decoding algorithm, where each temporal gesture segment is best modeled by one of the M branches.

There exist close temporal and structural relationships between gestures and speech prosody, including intonation, rhythm, and intensity patterns, as reported in recent works such as [6, 42]. Due to this, we parameterize speech signals as prosody features composed of speech intensity, pitch, and confidence-to-pitch as in [27, 28]. We also apply mean and variance normalization to prosody features to ensure speaker and utterance independence.

We analyze the relationship between gesture and speech modalities by constructing a hidden semi-Markov model (HSMM), which replaces self-transition probabilities of the regular hidden Markov model by state duration distributions [43]. We take gestures as states of a Markov chain with prosody as observations, where state transitions correspond to articulation of consecutive gestures. We represent the state sequence in the HSMM structure by $\boldsymbol{\ell}$, which is a sequence of labels where each ℓ_l is one of the M available gesture classes $\{g_1, g_2, ..., g_M\}$.

An HSMM depicting continuous observations with M fully connected states is defined 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 matrix, observation emission distribution, state duration distribution matrix, and initial state distribution matrix. The $M \times M$ 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. The observation emission distribution $\boldsymbol{B}$ is modeled by continuous probability distribution functions $b_i(\boldsymbol{f})$ for each gesture class g_i, where $\boldsymbol{f}$ is the observation feature vector.

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

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

where $d_i(\tau)$ is the probability of a gesture segment from gesture class g_i lasting $\tau\delta$ 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, based on our gesture clustering results, and the histogram bin size as the speech frame duration $\delta = 25$ msec.

The model parameters $\boldsymbol{A}$, $\boldsymbol{D}$ and $\boldsymbol{\Pi}$ are all estimated in terms of probability distributions over the training corpus. As for the emission distribution model $\boldsymbol{B}$, GMMs (Gaussian Mixture Models) are among the mostly preferred models for continuous feature representations in generative models such as HMMs and their variants. Another option would be using DNNs (Deep Neural Networks). However, GMMs are faster to compute and easier to learn compared to DNNs. We further discuss estimation of $\boldsymbol{B}$ in Section 4.2 by considering different structures to model the emission distributions of observations.

3.2. Gesture Synthesis

Gesture synthesis is characterized as decoding an optimal state sequence $\hat{\boldsymbol{\ell}}$ over the HSMM Λ, given a sequence of observations $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_T\}$ of length T time frames. In our framework, the decoded optimal state sequence yields a sequence of synthesized gesture class labels along with their durations, and the HSMM framework secures to have realistic gesture segment durations [25, 27, 28]. Unlike the HMM, in the HSMM framework, the state duration distributions are also considered during the Viterbi decoding process, based on a forward

likelihood function:

$$\psi_t(i)=\max_{\tau,j}\left\{\psi_{t-\tau}(j)+\log\left(a_{ji}d_i(\tau)\prod_{k=t-\tau+1}^{t}b_i(\boldsymbol{f}_k)\right)\right\}, \quad (2)$$

where $\psi_t(i)$ is the accumulated logarithmic likelihood at time frame t in state g_i after observing $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_t\}$. Note that the maximization in (2) finds the optimal duration τ^* and the optimal previous state g_{j^*}, and ties the accumulated likelihood $\psi_{t-\tau^*}(j^*)$ to $\psi_t(i)$. Based on the likelihood function $\psi_t(i)$, we use the modified Viterbi decoding algorithm to extract the optimal state sequence, which is equivalent to the optimal gesture segment sequence $\hat{\boldsymbol{\ell}} = \{\hat{\ell}_1, \ldots, \hat{\ell}_L\}$, and the associated gesture segment durations $\hat{\boldsymbol{\tau}} = \{\hat{\tau}_1, \ldots, \hat{\tau}_L\}$, where L is the number of gesture segments in the synthesized sequence.

3.3. Gesture Animation

Animation of the synthesized gesture sequence is composed of three main tasks: generation of gesture motion sequences, smoothing gesture transitions, and graphical animation.

The first task is generating a sequence of gesture motion, based on the synthesized gesture segment label $\hat{\boldsymbol{\ell}}$ and duration $\hat{\boldsymbol{\tau}}$ sequences, where we use unit selection over the gesture segments extracted for gesture analysis as mentioned in Section 3.1. First, we gather all gesture segments from gesture class g_i into a gesture subset $G_i = \{\varepsilon_n^{g_i}\}$, where $\varepsilon_n^{g_i}$ represents the n-th gesture segment in the subset. The union of all these disjoint sets gives us the overall gesture pool that we denote by G. Then, we apply a dynamic programming algorithm to minimize a joint distortion function R over the gesture pool G. R function penalizes duration differences of candidate gestures from gesture pool with the synthesized duration and joint angle differences during gesture transitions. The gesture segments selected from the pool G are further resampled to fit the synthesized duration $\hat{\boldsymbol{\tau}}$. Next, we smooth joint angle discontinuities over a temporal window at gesture segment transition boundaries by applying an exponential smoothing function on the synthesized gesture motion sequence of joint angles as shown in

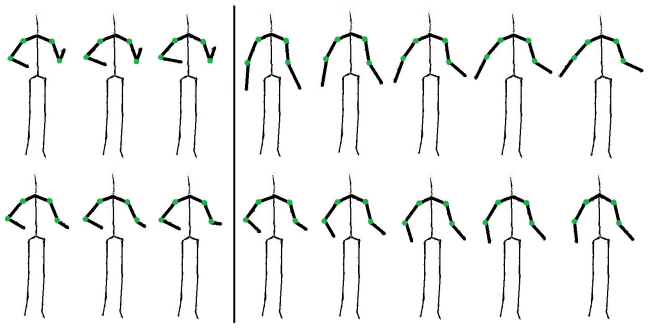


Figure 1: Gesture transition smoothing during animation generation step (vertical line denotes a gesture boundary): (Top) Animation sequence without smoothing. (Bottom) Windowed smoothing effect for the same animation sequence. Green circles denote arm and fore-arm joints.

Figure 1. Finally, we use the MotionBuilder 3D Character Animation Software for animating the smoothed gesture motion sequence[2].

4. Affect Expressive Gesture Synthesis and Animation

Arms are among the most expressive human body parts as informative indicators of human affective state, as well as intensity of emotion [35]. We use affect attribute annotations in addition to prosody features for synthesizing expressive arm gestures. We present the general framework for our affect-expressive speech-driven gesture synthesis and animation system in Figure 2, which consists of three main functional phases: analysis, synthesis and animation. General HSMM framework of the three main functional phases have been presented in Section 3. In Section 4, we present affective gesture modeling, objective evaluation metrics to be used and affective gesture animation.

[2]Autodesk MotionBuilder: 3D Character Animation for Virtual Production, http://www.autodesk.com

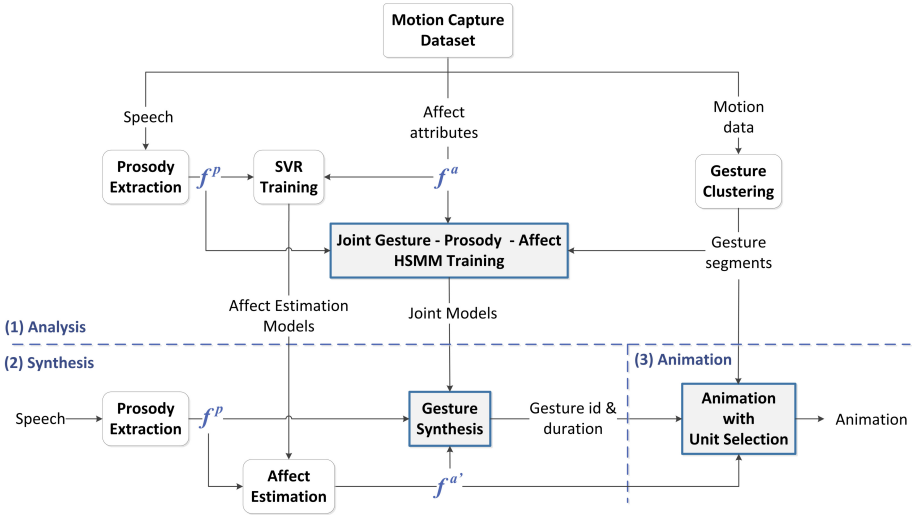


Figure 2: Block diagram of the general framework for the proposed speech-driven affective gesture animation system. The framework has three phases: (1) analysis, (2) synthesis and (3) animation. Inputs of analysis step are speech, motion-capture data and their affective attributes and the outputs are trained models to be used in the synthesis phase. The only input for synthesis step is speech and output is gesture segment sequence with gesture id and duration information. Animation phase inputs this sequence and generates arm gesture animations, final output of the framework, by selecting gestures from a pool of gesture segments created in the analysis phase.

4.1. Features Driving the Synthesis

Psychological and physiological changes due to emotional experience elicit modulations in the speech production system [44]. Based on these changes, affect can be explained with a three dimensional representation: activation (active vs. passive), valence (positive vs. negative), dominance (dominant vs. submissive) [1, 44]. Additionally, prosody provides a rich source of information that complements linguistic message reflecting the emotional state of the speaker. We use prosody and affect attributes (i.e. activation, valence, dominance) for driving the gesture synthesis process.

The normalized intensity, pitch and confidence-to-pitch features along with their first-order temporal derivatives are used to define the 6D prosody feature vector, which we denote by $\boldsymbol{f}^p$. Note that we use an indexed subscript as $\boldsymbol{f}^p_t$ to

represent the prosody vector at time frame t.

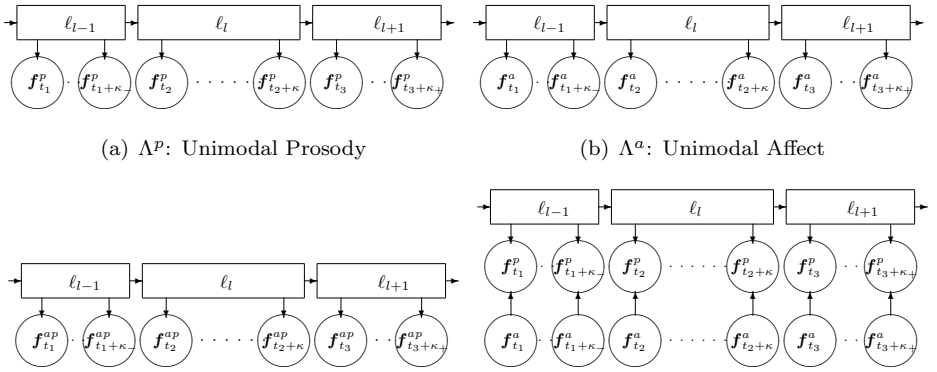
In this study we use a multimodal database that has been annotated continuously in time in *activation* (A), *valence* (V) and *dominance* (D) attributes. Annotations were performed by several annotators watching the video recordings (including audio) to comprehensively capture each actor's affective state [45]. The multimodal database delivers a processed average of individual annotations as the ground-truth annotation of the AVD attributes, which define affective transcriptions of verbal and non-verbal characteristics of the actors. The ground-truth affect attribute annotations together with extracted speech prosody are expected to model more realistic gesture patterns when driving an affect-expressive gesture synthesis and animation system. For this purpose, we explore the role of affect in speech-driven gesture animation by primarily using the ground-truth affect attribute annotations. We represent the 3D affect feature vector by $\boldsymbol{f}^{a}$.

We also investigate the performance of our speech-driven gesture animation system with the affect attributes estimated from speech. For this purpose, Support Vector Regression (SVR) models are employed with a Gaussian kernel to estimate the activation, valence and dominance attributes of affect from prosody features [46]. The ground-truth AVD annotations are used to train the SVR models. Then, each affect attribute is estimated separately, since SVR performs non-linear regression from a multi-dimensional input to a uni-dimensional output. We represent the estimated affect attributes by $\boldsymbol{f}^{\hat{a}}$.

4.2. Affective Gesture Modeling

In this section, we introduce our proposed emission distribution models for multimodal analysis of gesture, affect and speech using the HSMM framework explained in Section 3.1. Affect attributes and prosody features, as defined in Section 4.1, constitute the observations of the HSMM structure. We introduce four distinct structures to model emission distributions of the observations. Figure 3 visualizes these four structures, where the observations are assumed to be generated from a gesture state sequence $\boldsymbol{\ell}$. Figures 3(a) and 3(b) present the

baseline models which use prosody-only and affect-only features as observations, respectively. The other two structures, which consider joint and conditional distributions of affect and prosody, are shown in Figure 3(c) and Figure 3(d), respectively. In particular, the structure in Figure 3(d) utilizes conditional observation distributions of prosody given affect, which defines a non-homogeneous emission model. Note that for all the four structures in Figure 3, the state transition, initial distribution and duration probabilities are calculated based on the gesture state sequences in the training data, and the four structures 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 3: 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 ($\Lambda^{p|a}$). All the four structures are assumed to be generated from the same state sequence, $\boldsymbol{\ell}$, and have the same state transition, initial distribution and duration probabilities, which are computed on the training data. The only difference between the four HSMMs is their emission probabilities.

4.2.1. Unimodal Structures

We consider two baseline HSMM structures, Λ^p and Λ^a, with unimodal emission distributions as given in Figure 3(a) and 3(b), where the first one uses only prosody features $\boldsymbol{f}^p$, and the second one uses 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

$$b_i(\boldsymbol{f})=\sum_{k=1}^{K}\omega_{ik}\mathcal{N}(\boldsymbol{f};\boldsymbol{\mu}_{ik},\boldsymbol{\Sigma}_{ik}),\ \text{ for } i=1,...,M, \tag{3}$$

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

4.2.2. 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 3(c) with joint emission models. The concatenated 9D feature set is denoted by $\boldsymbol{f}^{ap}=[\boldsymbol{f}^a\boldsymbol{f}^p]$

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

$$b_i(\boldsymbol{f}^{ap})=\sum_{k=1}^{K}\alpha_{ik}\mathcal{N}(\boldsymbol{f}^{ap};\boldsymbol{\mu}_{ik}^{(ap)},\boldsymbol{C}_{ik}^{(ap)}), \tag{4}$$

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

$$\boldsymbol{\mu}_{ik}^{(ap)}=\begin{bmatrix}\boldsymbol{\mu}_{ik}^{(a)}\\ \boldsymbol{\mu}_{ik}^{(p)}\end{bmatrix},\quad \boldsymbol{C}_{ik}^{(ap)}=\begin{bmatrix}\boldsymbol{\Sigma}_{ik}^{(aa)}\boldsymbol{\Sigma}_{ik}^{(ap)}\\ \boldsymbol{\Sigma}_{ik}^{(pa)}\boldsymbol{\Sigma}_{ik}^{(pp)}\end{bmatrix}, \tag{5}$$

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

4.2.3. Conditional Structure

We model the conditional structure of prosody given affect in Figure 3(d), $\Lambda^{p|a}$, through the conditional emission probability function by using a GMM based model, 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 [47] and has been applied to a large range of problems, including emotional state tracking [48]. We model the conditional probability distribution of prosody $\boldsymbol{f}^p$ given affect $\boldsymbol{f}^a$ as,

$$b_i(\boldsymbol{f}_t^p|\boldsymbol{f}_t^a) = \sum_{k=1}^{K} P(k|\boldsymbol{f}_t^a, i) b_i(\boldsymbol{f}_t^p|\boldsymbol{f}_t^a, k), \tag{6}$$

where the indices t, i and k are respectively running over the frames, states and mixture components. In (6) $P(k|\boldsymbol{f}_t^a, i)$ is the occupancy probability of the k-th mixture at state g_i and defined as

$$P(k|\boldsymbol{f}_t^a, i) = \frac{\alpha_{ik}\mathcal{N}(\boldsymbol{f}_t^a; \boldsymbol{\mu}_{ik}^{(a)}, \boldsymbol{\Sigma}_{ik}^{(a)})}{\sum_{n=1}^{K} \alpha_{in}\mathcal{N}(\boldsymbol{f}_t^a; \boldsymbol{\mu}_{in}^{(a)}, \boldsymbol{\Sigma}_{in}^{(a)})}, \tag{7}$$

where α_{ik} is the k-th mixture weight. The conditional distribution is also defined as a Gaussian,

$$b_i(\boldsymbol{f}_t^p|\boldsymbol{f}_t^a, k) = \mathcal{N}(\boldsymbol{f}_t^p; \mathbf{X}_{ikt}^{(p)}, \mathbf{Y}_{ik}^{(p)}), \tag{8}$$

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

$$\mathbf{X}_{ikt}^{(p)} = \boldsymbol{\mu}_{ik}^{(p)} + \boldsymbol{\Sigma}_{ik}^{(pa)}{\boldsymbol{\Sigma}_{ik}^{(aa)}}^{-1}(\boldsymbol{f}_t^a - \boldsymbol{\mu}_{ik}^{(a)}), \tag{9}$$

$$\mathbf{Y}_{ik}^{(p)} = \boldsymbol{\Sigma}_{ik}^{(pp)} - \boldsymbol{\Sigma}_{ik}^{(pa)}{\boldsymbol{\Sigma}_{ik}^{(aa)}}^{-1}\boldsymbol{\Sigma}_{ik}^{(ap)}. \tag{10}$$

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

4.3. Objective Evaluation Metrics for Gesture Clustering and Synthesis

In this section, we define three objective metrics for evaluation of the multimodal HSMM structure of gesture, affect and speech. The first metric is a

weighted correlation score between gesture and affect to set the number of gesture classes, M. The other two metrics are mean symmetric Kullback-Leibler divergence (KLD) scores to evaluate the modeling performance of the HSMM structure.

4.3.1. Weighted Correlation of Gesture and Affect

The unsupervised clustering of gestures, described in Section 3.1, is performed over joint angles of arms and fore-arms, and should ideally result in M gesture classes each of which represents a distinct temporal motion pattern while preserving variety of affective gesture forms. In order to preserve variety of affective gesture forms while optimizing the value of M, we consider the correlation of motion features in each gesture class with the corresponding affect features. While we aim for an M value as low as possible considering the computational cost of HSMM-based synthesis, we would also like to maximize the correlation between motion and affect features within gesture classes. To this effect, all the gesture segments belonging to gesture class g_i are concatenated so as to form a joint angle feature vector denoted by $\boldsymbol{F}^{g_i}$. Similarly, the corresponding affect feature sequence for a given affect attribute a, which can be A, V or D, is denoted by $\boldsymbol{F}^{a_i}$ for each gesture class g_i. Then, the canonical correlation analysis (CCA) score is computed between $\boldsymbol{F}^{g_i}$ and $\boldsymbol{F}^{a_i}$, denoted as $\rho(\boldsymbol{F}^{g_i}, \boldsymbol{F}^{a_i})$. We then define a weighted gesture-affect correlation score, $\bar{\rho}^a_M$, over all the gesture classes as

$$\bar{\rho}^a_M = \sum_{i=1}^{M} \omega_i \rho(\boldsymbol{F}^{g_i}, \boldsymbol{F}^{a_i}) \tag{11}$$

where M is the total number of gesture classes and ω_i is the weight of each gesture class g_i. Each ω_i is computed as the ratio of the duration, T_i, of the gesture segments belonging to class g_i to the total duration of gestures in the database as $\omega_i = T_i/(\sum_j T_j)$.

4.3.2. Mean Duration KLD

The multimodal HSMM structure models the gesture duration statistics in terms of state duration distributions. The gestures synthesized from the gener-

ative HSMM structure should have realistic gesture duration statistics. Hence, we set a second objective metric to test the fit between the original and the synthesized gesture duration statistics. Recall that the state dependent duration probability mass function $d_i(\tau)$, as defined in (1) for each gesture class g_i, carries the statistics of the original gesture durations. Similarly, we represent the duration probability mass functions for the synthesized gestures by $\hat{d}_i(\hat{\tau})$. The symmetric Kullback-Leibler divergence (KLD) can be employed to measure the similarity between these two probability mass functions. We define a mean duration KLD distance measure, $\mathcal{D}^d_{\mathrm{KL}}$, over all gesture classes as

$$\begin{aligned}\mathcal{D}^d_{\mathrm{KL}} &= \sum_{i=1}^{M} \frac{\omega_i}{2}[\mathcal{D}^d_{\mathrm{KL}}(O||S) + \mathcal{D}^d_{\mathrm{KL}}(S||O)] \\ &= \sum_{i=1}^{M} \frac{\omega_i}{2}[\mathcal{D}_{\mathrm{KL}}(d_i(\tau), \hat{d}_i(\hat{\tau})) + \mathcal{D}_{\mathrm{KL}}(\hat{d}_i(\hat{\tau}), d_i(\tau))],\end{aligned} \tag{12}$$

where $\mathcal{D}^d_{\mathrm{KL}}(O||S) = \mathcal{D}_{\mathrm{KL}}(d_i(\tau), \hat{d}_i(\hat{\tau}))$ is the KLD distance from synthesized to the original sequence (and $\mathcal{D}^d_{\mathrm{KL}}(O||S)$ is the distance from the original to the synthesized sequence) for duration statistics of gesture class g_i, and ω_i is the weight of each gesture class g_i as used in (11). Similarly, we also define a weighted standard deviation for the duration KLD distance as

$$\mathcal{S}^d_{\mathrm{KL}} = \sqrt{\sum_{i=1}^{M} \omega_i [\frac{1}{2}(\mathcal{D}^d_{\mathrm{KL}}(O||S) + \mathcal{D}^d_{\mathrm{KL}}(S||O)) - \mathcal{D}^d_{\mathrm{KL}}]^2} \tag{13}$$

to observe the deviation of the KLD distance across gesture classes.

4.3.3. Mean Gesture-Prosody KLD

We set the third objective measure to evaluate the goodness of the joint gesture and prosody distribution as an output of the gesture synthesis framework. Hence, we construct the joint probability distributions of gesture classes and speech prosody over the original and synthesized recordings, and compute the symmetric KLD distance between these two distributions.

In our framework, the gesture classes are discrete, whereas the prosody representation is continuous. In order to compute joint probability mass function,

we first quantize prosody features into Q clusters using the k-means clustering algorithm. The joint probability of gesture and prosody can then be computed over original and synthesized gesture labels (ℓ and $\hat{\ell}$) by

$$P(g_i, p_j) = \frac{C(i,j)}{\sum_{i'} \sum_{j'} C(i',j')} \tag{14}$$

where $C(i,j)$ is the total count of frames with gesture class g_i and prosody cluster p_j. The KLD distance for gesture-prosody statistics, $\mathcal{D}_{\mathrm{KL}}^{gp}$, is defined by

$$\mathcal{D}_{\mathrm{KL}}^{gp}(Q) = \mathcal{D}_{\mathrm{KL}}(\mathcal{P}_{\ell}(g,p), \mathcal{P}_{\hat{\ell}}(g,p)), \tag{15}$$

where Q refers the number of clusters for prosody quantization, and $\mathcal{P}_{\ell}(g,p)$ and $\mathcal{P}_{\hat{\ell}}(g,p))$ represent the joint probability mass functions, respectively for the original and synthesized gesture (label) sequences.

4.4. Affective Gesture Animation

We use the unit selection algorithm to generate affective gesture animations with a joint distortion function as defined in Section 3.3. In the proposed affective gesture animation system, we choose the distortion function R to penalize i) gesture transition distortion, ii) duration difference and iii) mismatch of affect attributes. The joint distortion function R is defined for each gesture label $\hat{\ell}_l$ of the synthesized state sequence as

$$\begin{aligned} R(\varepsilon_n^{g_i} | \hat{\ell}_l = g_i) \quad = \quad & R_{\omega}(\varepsilon_n^{g_i} | \hat{\ell}_l = g_i) + \\ & R_{\tau}(\varepsilon_n^{g_i} | \hat{\ell}_l = g_i) + \\ & R_a(\varepsilon_n^{g_i} | \hat{\ell}_l = g_i), \end{aligned} \tag{16}$$

where R_{ω}, R_{τ} and R_a are the min-max normalized distortion functions [27]. The distortion function R_{ω} is the normalized mean squared difference between the joint angles of the gesture segments at the transition from state $\hat{\ell}_{l-1}$ to state $\hat{\ell}_l$. R_{τ} is the normalized absolute-valued difference between the durations of the candidate gesture segment $\varepsilon_n^{g_i}$ and the synthesized duration $\hat{\tau}_l$. Similarly, R_a is the normalized mean squared difference between the (annotated) affect

attributes of the gesture segment $\varepsilon_n^{g_i}$ and the (annotated or estimated) affect attributes of the input speech where 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 distortion function as in [27]:

$$\varphi_l(\varepsilon_n^{\hat{\ell}_l}) = \min_j \left\{ \varphi_{l-1}(\varepsilon_j^{\hat{\ell}_{l-1}}) + R(\varepsilon_n^{g_i} | \hat{\ell}_l = g_i) \right\}, \tag{17}$$

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 $\{\hat{\ell}_1, \hat{\ell}_2, \ldots, \hat{\ell}_l\}$.

5. Experimental Evaluations

We use the multimodal USC CreativeIT database that contains vocal and body-language behavior information of the actors obtained through close-up microphones, motion capture and HD cameras [45]. The interactions performed by pairs of actors in this database are either improvisations of scenes from theatrical plays or theatrical exercises where each actor repeats a short sentence emphasizing specific emotions. Each recording is annotated by multiple annotators with dimensional affect attributes (activation, valence, dominance) in the range [-1,1]. Then, the average of the attributes computed over the annotators are fixed as the ground-truth annotations.

There are eight pairs of speakers in the database and we perform speaker-independent evaluations in a leave-one-pair-out manner using each time the data from one actor pair as the test set, and the remaining data from the other pairs as the training. At each turn, we use all the data available for training (both theatrical play and theatrical exercise recordings), but for testing we synthesize gestures only for theatrical play recordings since each actor expresses more consecutive gestures in these recordings. In each of the theatrical exercise interactions, two actors utter short sentences in turns, repeating the same sentences

throughout the recording (e.g. '*Marry me*', '*I do not know*'). So, each actor holds the floor only for a short time compared to the theatrical play recordings where actors improvise scenes from a theatrical play. We consider the theatrical play recordings for objective and subjective evaluations.

5.1. Affect Estimation from Speech Prosody

We estimate affect attributes from speech prosody. Prosody features are extracted within windows of 50 msec with 25 msec shift (40 fps) whereas, affect attributes are annotated per frame by watching videos at 60 fps frame rate. We down-sample the affect attribute features to 40 fps and use the Support Vector Regression (SVR) technique [46] with the radial basis kernel to estimate activation, valence and dominance attributes of affect from speech prosody. We calculate the estimation results for the parameter C with varying values (1, 10, 20) and fix its value as 10, based on mean square error criterion. We set the parameter ϵ to 0.1, which specifies the margin within which no penalty is associated for the training loss function.

Affect estimation is evaluated using the leave-one-actor-pair-out approach. The MSE values are obtained as 0.045, 0.125, 0.112 for the activation, valence and dominance attributes, respectively. We also measure the Pearson correlation values as 0.40, 0.29, and 0.29, for the activation, valence, and dominance, respectively.

5.2. Setting Number of Gesture Clusters

In order to set the number of gesture clusters, M, we use the weighted correlation score, $\bar{\rho}_M^a$ defined in (11) in Section 4.3, for each affect attribute a, as a function of M varying from 10 to 90, as given in Figure 4. The weighted correlation score $\bar{\rho}_M^a$ rapidly increases with the first four values of M from 10 to 40. Then, the increase slows down for larger M values. Note that for large values of M the computational cost of the gesture synthesis task also increases by M^2. Based on these observations, we set the number of gesture classes as

$M = 40$, which maintains acceptable correlation values of 0.48, 0.48 and 0.52 for activation, valence and dominance attributes, respectively.

We also explore gesture cluster joint angle means and AVD distributions for M=40 in Figure 5. The red dots in the figure correspond to gesture cluster AVD mean values and axes denote activation, valence, and dominance. For clarity, we only show gesture cluster mean joint angle samples that have the maximum or the minimum AVD values, such as clusters g_7, g_{18}, g_{24}, g_{28}, and g_{32}, as shown in the figure. Based on this figure, we observe that, as the activation value increases, gestures become more open and raised (see for instance g_7, g_{28}, g_{32} on the left pane) in the dataset. Similarly in the case of dominance, open arms raised from elbows are perceived as more dominant (see g_{24}, g_7, g_{28}, g_{32} on the middle pane). On the other hand, it is hard to perceive the impact of valence by just observing the figure. The reason for this outcome could be that human perception of valence may rely more on direction than on location as pointed in [49].

5.3. Objective Evaluations of the HSMM Gesture Synthesis

We consider similarity of gesture duration statistics and joint statistics of gesture-prosody between original and synthesized gestures as the objective evaluation criteria for synthesis as defined in Section 4.3. In all objective evaluations, we select Λ^p and Λ^a as the baseline methods and compare joint and conditional models against these baselines for statistical significance of the KLD distances by randomly sampling from distributions 100 times and calculating two-tailed t-test scores. We use mean of all iterations in the tables. Table 1 presents mean and standard deviation (std) of the duration KLD distances, $\mathcal{D}^d_{\mathrm{KL}}$ and $\mathcal{S}^d_{\mathrm{KL}}$, resulting from the unimodal, joint and conditional emission structures of the HSMM and all methods significantly differ from the baselines with p-value smaller than the significance threshold 0.01. The conditional structure with the estimated AVD attributes, $\Lambda^{p|\hat{a}}$, attains the smallest mean and std for the duration KLD distances. The conditional structure with the ground-truth affect annotations, $\Lambda^{p|a}$, performs very close to the $\Lambda^{p|\hat{a}}$ structure and attains the

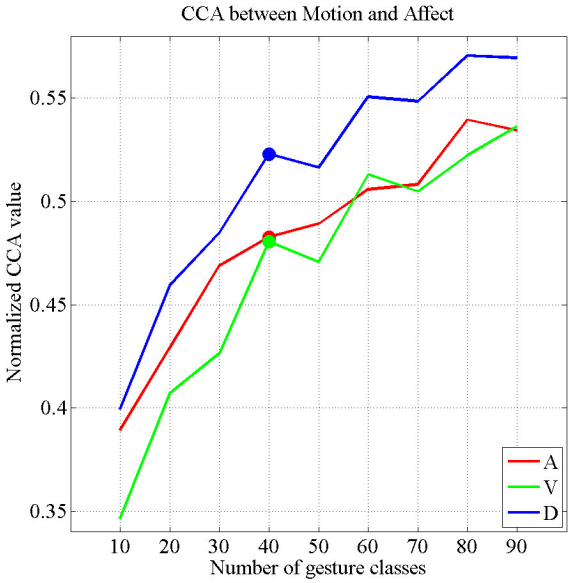


Figure 4: The weighted global CCA-based correlation scores between motion features and affect attributes as a function of number of gesture classes: (A) Activation, (V) Valence, and (D) Dominance.

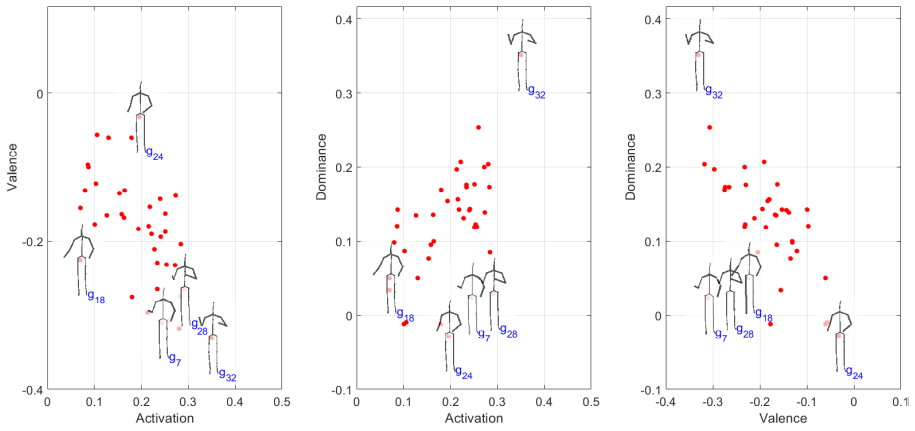


Figure 5: Red dots represent mean activation (A), valence (V), and dominance (D) values for individual gesture clusters. Mean joint angles for gesture clusters g_7, g_{18}, g_{24}, g_{28}, and g_{32} are also plotted with stick figures since they have the maximum and the minimum mean AVD values in the dataset for number of clusters $M = 40$.

second smallest mean and std. On the other hand, Λ^a has the largest mean duration KLD distance and Λ^{ap} has the largest std among all structures. These observations suggest that the conditional emission probability model performs better than the unimodal and joint structure models.

Table 1: The mean and std of the duration KLD distances for the proposed HSMM structures

	Λ^p	Λ^a	Λ^{ap}	$\Lambda^{p\|a}$	$\Lambda^{p\|\hat{a}}$
$\mathcal{D}^d_{\mathrm{KL}}(O\|\|S)$	5.71	6.50	5.95	5.67	5.66
$\mathcal{D}^d_{\mathrm{KL}}(S\|\|O)$	5.62	7.01	5.70	5.42	5.41
$\mathcal{D}^d_{\mathrm{KL}}$	5.67	6.76	5.82	5.56	5.54
$\mathcal{S}^d_{\mathrm{KL}}(O\|\|S)$	1.01	0.79	1.15	0.76	0.78
$\mathcal{S}^d_{\mathrm{KL}}(S\|\|O)$	0.97	1.12	1.90	0.95	0.74
$\mathcal{S}^d_{\mathrm{KL}}$	0.92	0.86	1.46	0.74	0.68

As the second objective measure, we employ the mean gesture-prosody KLD distance, $\mathcal{D}^{gp}_{\mathrm{KL}}$, to evaluate the joint statistics of gesture and prosody as defined in (15) in Section 4.3. Table 2 presents the KLD distances of the gesture-

prosody statistics with varying Q (of values 16, 32, 64) for different HSMM structures. We observe that the multimodal HSMM structures, which include affect and prosody, have consistently smaller $\mathcal{D}_{\mathrm{KL}}^{gp}$ distances compared to the unimodal structures. Furthermore, the conditional structures, $\Lambda^{p|a}$ and $\Lambda^{p|\hat{a}}$, attain the statistically significant smallest gesture-prosody KLD distances over all prosody quantization levels. Although the affect-only structure attains the largest gesture-prosody KLD distance, adding affect information together with prosody enables us to better model the relationship between the gesture and prosody modalities.

Table 2: The mean gesture-prosody KLD distances ($\mathcal{D}_{\mathrm{KL}}^{gp}(Q)$) for the proposed HSMM structures over different quantization levels of prosody (Q)

Q	Λ^{p}	Λ^{a}	Λ^{ap}	$\Lambda^{p\|a}$	$\Lambda^{p\|\hat{a}}$
16	2.69	3.03	2.60	2.28	2.22
32	2.47	2.78	2.38	2.11	2.04
64	2.23	2.50	2.15	1.92	1.86

In the proposed HSMM structures, the affect observation is taken as a 3D feature vector of the activation, valence and dominance attributes. Table 3 and Table 4 present the KLD distances of duration, $\mathcal{D}_{\mathrm{KL}}^{d}$, and gesture-prosody statistics, $\mathcal{D}_{\mathrm{KL}}^{gp}(64)$, for the proposed HSMM structures with 1D and 2D affect attribute observations. Table 3 presents results for prosody conditioned on the ground-truth affect attribute features, whereas Table 4 presents results for prosody conditioned on the estimated affect features. We observe that the dominance attribute has the smallest distances for both $\mathcal{D}_{\mathrm{KL}}^{d}$ and $\mathcal{D}_{\mathrm{KL}}^{gp}$ among 1D affect attribute observations for both ground-truth and estimated affect attributes. On the other hand, the HSMM structure with the 2D activation and dominance attribute observations attains the smallest duration KLD distance, $\mathcal{D}_{\mathrm{KL}}^{d}$, for both ground-truth and estimated affect attributes. Interestingly, the smallest gesture-prosody KLD distance, $\mathcal{D}_{\mathrm{KL}}^{gp}$, is attained by the $\Lambda^{p|D}$ and $\Lambda^{p|\hat{D}}$ structures with 1D dominance attributes.

The least contribution of an affect attribute is obtained when prosody features are conditioned only on the activation attribute for both the ground-truth and the estimated activation attributes. An explanation for this result could be that activation is more explicitly represented in speech prosody, such as with speech intensity, compared to valence and dominance attributes. Hence, including activation into the gesture synthesis model may not bring any additional information and can even introduce redundancy, as in Tables 1 and 2 we observe smaller $\mathcal{D}^{d}_{\mathrm{KL}}$ and $\mathcal{D}^{gp}_{\mathrm{KL}}$ distances for the unimodal Λ^{p} structure when respectively compared to $\Lambda^{p|A}$ and $\Lambda^{p|\hat{A}}$ structures in Table 3 and Table 4.

In most of the models, conditioning on the estimated attributes does better than conditioning on the ground-truth attributes. A possible reason of this behaviour can be the fact that estimated attributes have more well-behaved characterization, such as being low-passed and smooth. On the other hand, the ground truth annotations often have high frequency components with abrupt changes.

Table 3: KLD distances of duration and gesture-prosody statistics for the proposed HSMM structures with the ground-truth affect attributes

	$\Lambda^{p\|A}$	$\Lambda^{p\|V}$	$\Lambda^{p\|D}$	$\Lambda^{p\|AV}$	$\Lambda^{p\|AD}$	$\Lambda^{p\|VD}$
$\mathcal{D}^{d}_{\mathrm{KL}}(O\|\|S)$	5.87	5.75	5.62	5.86	5.64	5.67
$\mathcal{D}^{d}_{\mathrm{KL}}(S\|\|O)$	5.98	5.69	5.58	5.83	5.44	5.43
$\mathcal{D}^{d}_{\mathrm{KL}}$	5.92	5.72	5.60	5.85	5.54	5.55
$\mathcal{S}^{d}_{\mathrm{KL}}(O\|\|S)$	0.99	1.04	0.97	0.89	0.90	0.89
$\mathcal{S}^{d}_{\mathrm{KL}}(S\|\|O)$	1.27	1.23	1.28	1.14	1.23	0.83
$\mathcal{S}^{d}_{\mathrm{KL}}$	1.05	1.07	1.02	0.96	0.99	0.77
$\mathcal{D}^{gp}_{\mathrm{KL}}(64)$	2.96	2.60	2.19	3.03	2.50	2.55

5.4. Subjective Evaluation of Affective Gesture Animation

Subjective tests can evaluate realism and naturalness of the animation by reflecting human perception, unlike the objective evaluations. We use a mean

Table 4: KLD distances of duration and gesture-prosody statistics for the proposed HSMM structures with the estimated affect attributes

	$\Lambda^{p\|\hat{A}}$	$\Lambda^{p\|\hat{V}}$	$\Lambda^{p\|\hat{D}}$	$\Lambda^{p\|\hat{A}\hat{V}}$	$\Lambda^{p\|\hat{A}\hat{D}}$	$\Lambda^{p\|\hat{V}\hat{D}}$
$\mathcal{D}^d_{\mathrm{KL}}(O\|\|S)$	5.89	5.82	5.73	5.83	5.62	5.68
$\mathcal{D}^d_{\mathrm{KL}}(S\|\|O)$	5.77	5.50	5.55	5.58	5.45	5.57
$\mathcal{D}^d_{\mathrm{KL}}$	5.83	5.66	5.64	5.71	5.54	5.62
$\mathcal{S}^d_{\mathrm{KL}}(O\|\|S)$	1.08	1.00	0.98	0.87	0.84	0.94
$\mathcal{S}^d_{\mathrm{KL}}(S\|\|O)$	1.29	1.28	1.17	0.84	0.93	0.77
$\mathcal{S}^d_{\mathrm{KL}}$	1.08	1.07	1.02	0.80	0.82	0.79
$\mathcal{D}^{gp}_{\mathrm{KL}}(64)$	2.95	2.38	2.36	2.98	2.60	2.57

opinion score (MOS) test to subjectively evaluate animations of the four synthesis methods defined in Section 4.2 in comparison to the original mocap motion. We segment audio recordings based on speaker turns in the dyadic interactions and ensure only a single speaker speaks at a time. Total number of the extracted audio test segments is 29 with an average duration of 15s. The test is run with 25 participants using 5 methods, which are the animations of the four synthesis methods and the original mocap. Each test session for a participant contains 27 clips, where each method is tested with 5 clips and two other clips are used to train the participant at the beginning of the test. One of the two training clips is of high quality and the other is of low quality in terms of speech-gesture synchrony. During the test, all clips, except the training clips, are shown in random order. In the test, each participant is asked to evaluate how expressive and coherent arm gesture animations are with the speech using a five-point assessment scale (1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Very Good).

Table 5 shows the distribution of test scores and mean opinion scores with standard deviation (std) values for each method, where methods are sorted top-to-bottom in decreasing MOS order. The original mocap attains the highest MOS as 3.928. Among the proposed four synthesis methods, the conditional structure $\Lambda^{p|a}$ attains the highest MOS as 3.752. These observations are in

line with the objective evaluations, and suggest that the conditional emission probability model performs better than the unimodal and joint structure models. Note also that the MOS gap between the conditional structure and the unimodal prosody structure is relatively larger compared to the gap with the original mocap. The distribution of the test scores shows that 33% of the original mocap animations and 21% of the $\Lambda^{p|a}$ synthesized animations are evaluated with the highest score. Moreover, 45% of the $\Lambda^{p|a}$ synthesized animations are assessed as "good" by the participants of the subjective tests. The lowest test scores belong to the animations based on the unimodal affect structure Λ^{a}, which may be due to the failure of affect attributes in modeling timing of the gestures.

Table 5: Mean opinion score (MOS) subjective test results with the score scale 1: Bad, 2: Poor, 3: Fair, 4: Good, and 5: Very Good.

	Distribution of scores (%)						
Method	**5**	**4**	**3**	**2**	**1**	**MOS**	**std**
Original	33	39	18	10	0	3.928	0.985
$\Lambda^{p\|a}$	21	45	24	10	0	3.752	0.922
Λ^{p}	13	33	30	17	7	3.304	1.101
Λ^{ap}	8	26	34	20	12	2.968	1.128
Λ^{a}	1	26	30	30	13	2.736	1.048

We use analysis of variance (ANOVA) tests to analyze the subjective evaluation results given in Table 5. Post-hoc tests are then performed using Tukey-Kramer method [50] to assess the statistical significance of the differences between the scores as shown in Table 6. We observe that the subjective test scores obtained with the proposed $\Lambda^{p|a}$ structure are not significantly different from the original mocap scores, but significantly different from the scores of the other three synthesis methods. Furthermore, the scores using the Λ^{p} and Λ^{ap} structures are not significantly different from each other. Samples of animation clips from the subjective tests are available for online demonstration[3].

[3]Sample clips for the affective synthesis and animation of arm gestures from speech prosody

Table 6: Significance test of the subjective evaluations: 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(*)

6. Conclusions and Future Work

We proposed a framework for the incorporation of continuous affect attributes into an affective speech driven arm gesture synthesis and animation system. We introduced new objective evaluation methods for the comparison of synthesized and the original gestures' duration and co-occurance statistics with prosody clusters. Finally, we performed subjective tests to evaluate human perception of the proposed system. The main finding of these evaluations is that the conditional structure $\Lambda^{p|a}$, which models the observation distribution as prosody given affect, achieves the best performances under both objective and subjective evaluations. Our experiments also show that, among the affect attributes, dominance is the one that contributes the most to the performance of our conditional HSMM structure.

We note that we have also considered an alternative HSMM structure where affect influences gesture state transitions, which does not however yield any performance improvement over the proposed conditional structure (hence not reported). This could be due to the fact that gestures may already be clustered into affective classes during the unsupervised clustering process.

One major factor that constrains the quality of the generated animations is the size of the available gesture pool. Recall that the gesture segments selected

are available at http://mvgl.ku.edu.tr/affective_gesture_synthesis

from the pool are often needed to be resampled to fit the synthesized durations. In case the difference between the durations of the selected and the synthesized segments is very large, this resampling process may create observable motion artifacts in the animations. This however becomes less of a problem when the gesture pool is sufficiently large.

In the proposed framework, the synthesized gestures do not necessarily match the particular semantic content of the input speech, as the gesture synthesis process is mainly based on prosody and affect attributes. Although this is a major limitation for the proposed framework, it can be addressed as an interesting future research to incorporate semantic analysis of speech to include the use of iconic, metaphoric and deictic gestures within the framework.

References

[1] A. Mehrabian, Nonverbal communication, Transaction Publishers, 1972.

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

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

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

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

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

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

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

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

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

[11] M. Stone, D. DeCarlo, I. Oh, C. Rodriguez, A. Stere, A. Lees, C. Bregler, Speaking with hands: Creating animated conversational characters from recordings of human performance, in: ACM SIGGRAPH 2004 Papers, ACM, New York, NY, USA, 2004, pp. 506–513. doi:10.1145/1186562.1015753.

[12] C. Becker, S. Kopp, I. Wachsmuth, Simulating the emotion dynamics of a multimodal conversational agent, in: Affective Dialogue Systems, Springer, 2004, pp. 154–165.

[13] B. Hartmann, M. Mancini, C. Pelachaud, 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, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 188–199.

[14] B. Hartmann, M. Mancini, C. Pelachaud, Towards affective agent action: Modelling expressive eca gestures, in: International conference on Intelligent User Interfaces-Workshop on Affective Interaction, San Diego, CA, 2005.

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

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

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

[18] R. Niewiadomski, S. Hyniewska, C. Pelachaud, Modeling emotional expressions as sequences of behaviors, in: International Workshop on Intelligent Virtual Agents, Springer, 2009, pp. 316–322.

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

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

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

[22] S. Levine, P. Krähenbühl, S. Thrun, V. Koltun, Gesture controllers, ACM Transactions on Graphics 29 (4). doi:10.1145/1833351.1778861.

[23] A. F. Baena, R. Montano, M. Antonijoan, A. Roversi, D. Miralles, F. Alias, Gesture synthesis adapted to speech emphasis, Speech Communication 57 (2014) 331–350. doi:10.1016/j.specom.2013.06.005.

[24] S. Marsella, Y. Xu, M. Lhommet, A. Feng, S. Scherer, A. Shapiro, Virtual character performance from speech, in: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '13, ACM, New York, NY, USA, 2013, pp. 25–35. doi:10.1145/2485895.2485900.

[25] E. Bozkurt, S. Asta, S. Ozkul, Y. Yemez, E. Erzin, Multimodal Analysis of Speech Prosody and Upper Body Gestures using Hidden Semi-Markov Models, in: IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, 2013, pp. 3652–3656.

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

[27] E. Bozkurt, E. Erzin, Y. Yemez, Affect-expressive hand gestures synthesis and animation, in: Multimedia and Expo (ICME), 2015 IEEE International Conference on, IEEE, 2015, pp. 1–6.

[28] E. Bozkurt, Y. Yemez, E. Erzin, Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures, Speech Communication 85 (2016) 29 – 42.

[29] N. Sadoughi, C. Busso, Retrieving target gestures toward speech driven animation with meaningful behaviors, in: International conference on Multimodal interaction (ICMI 2015), Seattle, WA, USA, 2015, pp. 115–122. doi:10.1145/2818346.2820750.

[30] H. Gunes, B. Schuller, M. Pantic, R. Cowie, Emotion representation, analysis and synthesis in continuous space: A survey, in: Automatic Face &

Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE, 2011, pp. 827–834.

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

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

[33] N. Sadoughi, Y. Liu, C. Busso, Msp-avatar corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents, in: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, Vol. 7, IEEE, 2015, pp. 1–6.

[34] A. Kleinsmith, N. Bianchi-Berthouze, Affective body expression perception and recognition: A survey, Affective Computing, IEEE Transactions on 4 (1) (2013) 15–33. doi:10.1109/T-AFFC.2012.16.

[35] M. Karg, A.-A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, D. Kulic, Body movements for affective expression: A survey of automatic recognition and generation, Affective Computing, IEEE Transactions on 4 (4) (2013) 341–359. doi:10.1109/T-AFFC.2013.29.

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

[37] M. Kipp, J.-C. Martin, 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, IEEE, 2009, pp. 1–8.

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

[39] C.-C. Chiu, L.-P. Morency, S. Marsella, Predicting co-verbal gestures: a deep and temporal modeling approach, in: International Conference on Intelligent Virtual Agents, Springer, 2015, pp. 152–166.

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

[41] M. E. Sargin, Y. Yemez, E. Erzin, A. M. Tekalp, Analysis of Head Gesture and Prosody Patterns for Prosody-Driven Head-Gesture Animation, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (8) (2008) 1330–1345. doi:10.1109/TPAMI.2007.70797.

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

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

[44] R. Cowie, R. R. Cornelius, Describing the emotional states that are expressed in speech, Speech communication 40 (1) (2003) 5–32.

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

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

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

[48] A. Metallinou, A. Katsamanis, S. Narayanan, Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information, Image and Vision Computing 31 (2) (2013) 137 – 152.

[49] M. de Meijer, The contribution of general features of body movement to the attribution of emotions, Journal of Nonverbal Behavior 13 (4) (1989) 247–268.

[50] J. W. Tukey, The Collected Works of John W. Tukey, Belmont, Calif.: Wadsworth Advanced Books and Software, 1984.