AFFECT-EXPRESSIVE HAND GESTURES SYNTHESIS AND ANIMATION

Elif Bozkurt, Engin Erzin and Yücel Yemez

Multimedia, Vision and Graphics Laboratory
College of Engineering, Koç University, Istanbul, Turkey
{ebozkurt, eerzin, yyemez}@ku.edu.tr

ABSTRACT

Speech and hand gestures form a composite communicative signal that boosts the naturalness and affectiveness of the communication. We present a multimodal framework for joint analysis of continuous affect, speech prosody and hand gestures towards automatic synthesis of realistic hand gestures from spontaneous speech using the hidden semi-Markov models (HSMMs). To the best of our knowledge, this is the first attempt for synthesizing hand gestures using continuous dimensional affect space, i.e., activation, valence, and dominance. We model relationships between acoustic features describing speech prosody and hand gestures with and without using the continuous affect information in speaker independent configurations and evaluate the multimodal analysis framework by generating hand gesture animations, also via objective evaluations. Our experimental studies are promising, conveying the role of affect for modeling the dynamics of speech-gesture relationship.

Index Terms— Prosody analysis, continuous affect, gesture animation, hidden semi-Markov models.

1. INTRODUCTION

Gestures expressing affect are responsible for the communication of aspects related to feelings, moods, and intensity of emotional experience. 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 hand gestures in synchrony with speech, which is expected to incorporate 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.

In this paper, we propose a speaker-independent framework for joint analysis of hand gestures with continuous affect attributes (i.e. activation, valence, and dominance) and speech prosody using the hidden semi-Markov models (HSMMs) [1]. We develop a multimodal system for data-driven, learning based synthesis and animation of affect-expressive gestures on the USC CreativeIT database [2]. Our work is on the basis of hand gesture phrases where a gesture phrase defines the basic gesture element [3]. Hence, we use gesture phrases to describe and model the relationship with prosody and affect information. To the best of our knowledge, this is the first attempt for synthesizing hand gestures using the continuous affect attributes. We particularly investigate the influence of affect information in a gesture synthesis and animation framework. Hence, in this study we use affect information as a driving factor of a gesture synthesis process and/or as an adjusting factor during the animation process. We report promising results demonstrating the importance of affect information for expressive hand gestures synthesis and animation.

2. RELATED WORK

Hand gesture is an important nonverbal behavior in human communication. Yang et al. show that individual's attitude as well as the interaction type as friendly or conflictive can be predicted using only the dynamics of the hand gesture phrases over an interaction [4]. Additionally, the expression of hand gesture often spontaneously accompanies speech production. Although some efforts have been devoted to exploring interactions between gestures and speech aiding communication [5], studies on the speech audio-driven gesture animation including arm movements are still limited so far. In [6], Levine et al. have introduced gesture controllers, availing a modular methodology to drive beat-like gestures with live speech via customized gesture repertoires. From a hierarchical perspective, the work of Levine et al. is mainly concentrated on the gesture phase level, whereas in a more recent study [7], Baena et al. present a framework that links speech prosody to beat gestures at phrase level based on manually annotated body motion and speech signals. They basically employ motion graphs to generate appropriate gestures with varying emphasis for a given speech input by modeling aggressive and neutral performances. Bozkurt et al. synthesize body gesture phrases from prosody observations using the hidden semi-Markov models (HSMMs) [8]. 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 [9].

As one of the major elements that control and influence the multimodal communicative channels, emotion has been widely studied in terms of its relation to speech and gesture. Various existing emotion descriptive models can be summarized as: categorical and dimensional. The theory of universal emotions and interpretation of affective expressions in terms of basic categories has been the most commonly accepted approach among the researchers. Early works on audio-driven emotional virtual character animation synthesis have mostly been concentrated on the head motion synthesis problem. Busso et al. present an approach in [10] 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. A more recent paper [11] focuses on building a speech-driven facial animation framework to generate natural head and eyebrow motions using dynamic Bayesian networks (DBNs). Niewiadomski et al. [12] present *multimodal sequential expressions* that are composed of different non-verbal behaviors. Their system is capable of synthesizing combinations of signals defined within a *behavior set* for eight emotional states. Garcia et al. classify gestures by associating them to

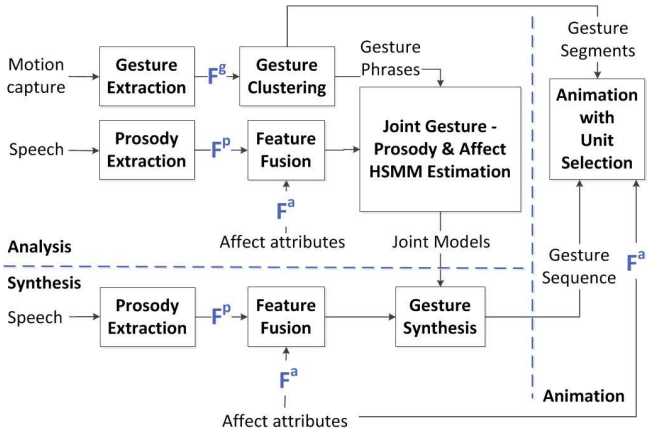

Fig. 1. The block diagram of the general framework for the affect dependent, speech-driven gesture synthesis system.

emotions (e.g. hand clapping and joy, arms crossing and anger) [13] EMOTE is a kinematic system for expressive variation of arm and torso movements that is based on the analysis of Laban [14].

An alternative approach to the categorical representation of affect is the dimensional description. The use of dimensional emotion description has been explored in multimodal affect recognition and synthesis [15]. The expression of basic emotions or discrete emotion categories are not flexible enough for modeling the complex facial expressions in human life. Jia et al. propose a text-to-visual-speech system, where expressive head and facial gestures are synthesized based on *pleasure*, *arousal*, and *dominance* descriptors of semantic expressivity [16]. Hartmann et al. define gesture expressivity in terms of *overall activation, spatial extent, temporal extent, fluidity, power, and repetition* [17] (see [18] for an extensive survey on body movements for affective expression).

3. SYSTEM OVERVIEW

The general framework for our automatic hand gesture synthesis system, given in Figure 1, consists of three main functional blocks for analysis, synthesis, and animation. The analysis step consists of extraction and clustering of joint angles representing hand gestures, speech prosody feature extraction and early fusion with continuous affect attributes, and joint analysis of gestures and prosody-affect fusion by using HSMMs. The synthesis step consist of extraction of prosody features and fusion with continuous affect attributes followed by HSMM-Viterbi algorithm based gesture synthesis. Finally, the animation part consists of animation generation via unit selection applied on a gesture pool with regard to a multi-objective cost function.

3.1. Gesture Clustering

In one of the pioneering studies on gesture and speech relationship, Kendon [3] proposed a widely accepted hierarchical model for gesture. In this model, the core gestural element is defined as gesture phase. We model gestures at gesture phrase level to emphasize speech intonation and affect. Gestures in motion capture data are represented by joint angles of arms and fore-arms. 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. 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,

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

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 forearms at time frame k is defined as,

$$\boldsymbol{f}_k^g = [\boldsymbol{f}_k^{J1}, ..., \boldsymbol{f}_k^{J4}]. \tag{2}$$

There is evidence that hand and arm movements are significant for distinguishing between affective states [4]. To this end, temporal clustering of the gesture feature sequence is necessary for analysis of recurring gesture phrases. For this purpose, we implement the unsupervised clustering method based on parallel-branch HMM, Λ^g, over the gesture feature stream $\boldsymbol{F}^g = \{\boldsymbol{f}_1^g, \boldsymbol{f}_2^g, ..., \boldsymbol{f}_T^g\}$, as in [19]. We segment and cluster gesture sequences into gesture phrases with duration information. The HMM structure Λ^g initially 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).

3.2. Prosody Extraction

Prosody characteristics at the acoustic level, including intonation, rhythm, and intensity patterns, carry important temporal and structural synchrony with gesture phrases [20]. 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. We extract prosody feature vector for each speech frame of 25 ms duration centered on a 50 ms analysis window. Speech intensity is extracted as the logarithm of the average signal energy in the analysis window,

$$I_k = \log(\frac{1}{W}\sum_{i=1}^{W} s_k[i]^2), \tag{3}$$

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 [21]. 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,

$$e_k(\tau) = \sum_{i=1}^{W}(s_k[i] - s_k[i+\tau])^2. \tag{4}$$

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

$$c_k = 1 - \frac{e_k(\tau^*)}{\frac{1}{\tau^*}\sum_{i=1}^{\tau^*} e_k(i)}, \tag{5}$$

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 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,

$$\boldsymbol{f}_k^p = [\bar{I}_k,\, \bar{\nu}_k,\, \bar{c}_k,\, \Delta\bar{I}_k,\, \Delta\bar{\nu}_k,\, \Delta\bar{c}_k], \tag{6}$$

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

3.3. Affect Features

Prosody conveys intonation which is important for modeling the variability and complexity of timings of the gestures [5]. However, using only prosody to drive the synthesis may not capture the affective information, ignoring the affective expressions of the gestures. We use the ground truth continous affect attributes of the dataset that are available in *activation* (A), *valence* (V), and *dominance* (D) domains. The annotations are normalized to the range $[-1, 1]$ and used as the 3D affect features, denoted as f^a in this study. Moreover, an affect estimation system if provided, may replace the groundtruth affect attributes, as well. In our system we assume each domain has equal contribution to synthesis and animation steps.

3.4. Feature Fusion

Feature level data fusion is one main information combining scheme for closely coupled and synchronized modalities. In this work, feature level fusion of prosody and continuous affect attributes is defined as concatenation of the modalities. Prior to fusion, both feature sets are normalized to ensure that the contribution of each component to the final representation is comparable. In addition, feature dimension reduction using Principal Component Analysis (PCA) is employed for a more compact representation. We retain the first two principal components as they preserve 98% of the total variance. The fused 9D feature set is denoted as f^{ap} and after PCA as f^{AP}.

3.5. HSMM Modeling

Hidden Markov models have proven to be useful models for approaching learning problems in sequential data. However, one disadvantage is that state duration distributions are restricted to geometric form. In many real world data the strict Markovian constraints are undesirable, particularly if we wish to learn or encode non-geometric state durations. We use the hidden semi-Markov model (HSMMs) [1] that allow construction of highly interpretable models by admitting use of prior knowledge on state durations for modeling the relationship between hand gestures, speech prosody and affect. In our framework, we take gestures as the states of a Markov chain and fusion of prosody and affect (or prosody only, or affect only) signals as the observations of this Markov process. Hence state transitions correspond articulation of consecutive gestures. Introducing a state duration model allows us to better control gesture phrase durations in the synthesis process. Figure 2 shows how such an HSMM structure works.

An HSMM representing continuous observations with M_g fully connected states is modeled as $\Lambda^{gp} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{\Pi})$. The states of Λ^{gp} represent gesture phrase classes, and the model parameters $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{D}$, $\boldsymbol{\Pi}$ 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,

$$\boldsymbol{A} : \{a_{ij} = P(\ell_l^g = g_j | \ell_{l-1}^g = g_i)\} \quad i, j = 1, ..., M_g, \tag{7}$$

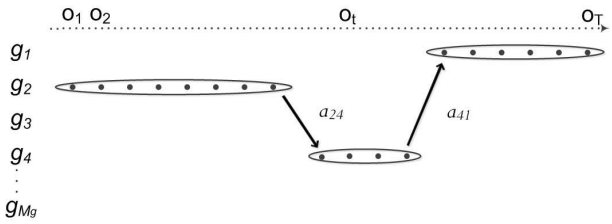

Fig. 2. In a Hidden Semi-Markov Process, each state has a duration and emits a number of observations.

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

$$\boldsymbol{B} : \{b_i(\mathbf{O}) = P(\mathbf{O} | \ell_l^g = g_i)\} \quad i = 1, ..., M_g, \tag{8}$$

where $b_i(\mathbf{O})$ is the probability of observing vector $\mathbf{O}$ at gesture g_i. We use diagonal-covariance Gaussian Mixture Models (GMMs) for modeling the feature sets per gesture cluster with a fixed number of mixtures. The state duration distribution $\boldsymbol{D}$ is formed as state dependent duration probability mass functions,

$$\boldsymbol{D} : \{d_i(k)\} \quad i = 1, ..., M_g,\ k = 1, ..., \frac{D_{max}}{\delta}, \tag{9}$$

where $d_i(k)$ is the probability of gesture g_i lasting $k\delta$ sec, D_{max} is the maximum duration among all gestures, and δ is the histogram bin size for the underlying probability mass function. We take the maximum duration as $D_{max} = 5$ sec, and the histogram bin size as the speech frame duration, $\delta = 25$ msec. The initial state probability vector $\boldsymbol{\Pi}$ is defined by entries π_i representing the probability of starting with gesture g_i as the first gesture phrase,

$$\boldsymbol{\Pi} : \{\pi_i = P(\ell_1^g = g_i)\} \quad i = 1, ..., M_g. \tag{10}$$

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

$$\pi_i = P(\ell_1^g = g_i) \mathrel{\widehat{=}} \frac{C(1, i, j)}{\sum_{j'} C(1, i, j')}, \tag{11}$$

$$a_{ij} = P(\ell_l^g = g_j | \ell_{l-1}^g = g_i) \mathrel{\widehat{=}} \frac{\sum_l C(l, i, j)}{\sum_l \sum_{j'} C(l, i, j')}, \tag{12}$$

$$b_i(\mathbf{O}) = P(\mathbf{O} | \ell_l^g = g_i) \mathrel{\widehat{=}} \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{O}, \mu_{ik}, \Sigma_{ik}), \tag{13}$$

$$d_i(k) \mathrel{\widehat{=}} \frac{H(i, k\delta \le \tau < (k+1)\delta)}{\sum_{k'} H(i, k'\delta \le \tau < (k'+1)\delta)}, \tag{14}$$

where $C(l, i, j)$ is the number of times g_i is the lth gesture and g_j is the $(l+1)$st gesture, $\mathbf{O}$ is the feature vector at gesture g_i, $K = 10$ is the number of mixture components, and $H(i, k\delta \le \tau < (k+1)\delta)$ is the number of occurrences of gesture g_i with duration τ in $[k\delta,\ (k+1)\delta)$ interval.

3.6. 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 continuous features, $\{o_1, o_2, \ldots, o_T\}$. Note that the decoded optimal state sequence delivers synthesized sequence of gesture phrases and their durations,

where the HSMM framework secures to have realistic gesture phrase durations. Unlike the HMM framework, in HSMM framework states have variable durations and a sequence of observations is emitted at a single state. This requires to define a forward likelihood function for the Viterbi decoding algorithm, which incorporates the state duration model,

$$\psi_t(j) = \max_{\tau} \max_{i} \{\psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(o_k))\}, \quad (15)$$

where $\psi_t(j)$ is the accumulated logarithmic likelihood at time frame t in state g_j after observing observations $\{o_1, o_2, \ldots, o_t\}$. Based on the forward likelihood function $\psi_t(j)$, we use the modified Viterbi decoding algorithm to extract the optimal state sequence, that is the optimal gesture phrase sequence $\hat{\boldsymbol{\ell}}^g = \{\hat{\ell}_1^g, \ldots, \hat{\ell}_L^g\}$, and the associated gesture phrase durations $\boldsymbol{\kappa} = \{\kappa_1, \ldots, \kappa_L\}$.

3.7. Gesture Animation

Animation of the synthesized gesture sequence consists of three main tasks. The first task is to generate a synthesized sequence of gesture segments, $\hat{\boldsymbol{\varepsilon}}^g$, given the synthesized gesture phrase $\hat{\boldsymbol{\ell}}^g$ and duration $\boldsymbol{\kappa}$ sequences. This task is performed using unit selection over the gesture phrases which are extracted during the gesture analysis in Section 3.1. To select gesture units with low concatenation distortion and low duration (and low affect attributes) difference, we apply a dynamic programing algorithm that minimizes a joint distortion function, D,

$$\begin{aligned} D(\varepsilon^{g_i j} | g_i = \hat{\ell}_l^g) \quad = \quad & \beta_1 D_\omega(\varepsilon^{g_i j} | g_i = \hat{\ell}_l^g) + \\ & (\beta_2 + \beta_3(1-\alpha)) D_\kappa(\varepsilon^{g_i j} | g_i = \hat{\ell}_l^g)] + \\ & \beta_3 \alpha D_a(\varepsilon^{g_i j} | g_i = \hat{\ell}_l^g), \end{aligned} \quad (16)$$

where $\varepsilon^{g_i j}$ is the j^{th} gesture segment candidate in the pool for cluster g_i, D_ω is the mean square error (MSE) of joint angle differences at transitions, D_κ is gesture duration difference, D_a is MSE of affect attribute differences, and α, β_1, β_2, β_3 are the coefficients for managing contribution of each element.

The selected gesture units are interpolated to fit the synthesized duration. The next 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 [22].

4. EXPERIMENTAL RESULTS

We use the multimodal USC CreativeIT database that contains a variety of dyadic theatrical improvisations for studying expressive behaviors and natural human interaction [2]. 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 other.

The database contains vocal and body-language behavior information of the actors obtained through close-up microphones, Motion Capture (MoCap) and HD cameras. Each recording is annotated with dimensional emotional descriptors (activation, valence, dominance), where multiple annotaters annotate the videos and average of these attributes per recording is used in this study. 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 MotionBuilder software [22] for converting 3D joint positions to Euler angle rotations of the arm and forearm in the (x,y,z) directions.

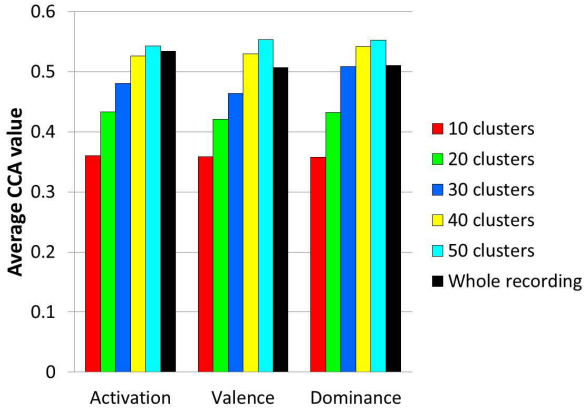

Fig. 3. CCA on gesture trajectories and affect attributes as a function of gesture clusters.

4.1. Gesture Clustering Evaluation

Gesture clustering is crucial where the building block of the animation system is gesture phrase. It should neither degrade affective information existent in the sequence nor complicate the overall system. We cluster gesture sequences in an unsupervised manner as summarized in Section 3.1 and expect it to capture gesture dynamics based on affective expression of the hand gestures. In order to decide on the optimal number of gesture clusters, we examine Canonical Correlation Analysis (CCA) scores of the four joints' Euler angles per cluster group with the corresponding affect attributes in domains activation, valence, and dominance, respectively. Average CCA scores per domain as a function of number of clusters from 10 to 50 are shown in Figure 3 where clustering gestures into 40 clusters maintains correlation requirements as high as the correlations calculated over the whole sequences as shown in the last bar columns per affect domain.

4.2. Experimental Set-up

In this study we investigate several approaches for incorporating affective information in a gesture synthesis-animation system as shown in Table 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 3.7 as $\beta_1 = 0.30$, $\beta_2 = \beta_3 = 0.35$ and perform speaker-independent evaluations in a leave-one-session-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 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

4.3. Evaluation of Synthesis Results

Inputs to our gesture synthesis system can be prosody, affect attributes or fusion of these two with/out PCA as shown in Table 1. We first evaluate compatibility of the individual feature sets for driving the gesture synthesis process. Since we aim for synthesis results that would reflect affective expressions, we analyze feature and the original gesture joint angle sequences by using CCA. In Table 2, we present mean and standard deviation (std) values of the first-order canonical correlations. We observe that feature fusion of prosody and affect attributes, f^{ap}, is the most correlated feature set with the original gesture trajectories. However, dimension reduction with PCA does not provide any improvement over the fused set.

Table 2. Average first-order canonical correlation mean and (std) values for each feature set and original gesture joint angle sequences.

f^{ap}	f^{p}	f^{a}	f^{AP}
0.66 (0.095)	0.41 (0.097)	0.64 (0.104)	0.47 (0.125)

HSMM framework allows to incorporate duration distributions when modeling gestures. So, one possible method to evaluate synthesis results could be measuring the similarity between the original and the synthesized gestures and gesture duration statistics. To this effect, for each feature set in Table 1 we synthesize a different gesture sequence. Then, we estimate the gesture distributions as the frame-level representation ratio in the sequence and duration distribution of each synthesized sequence via (9). In Table 3, we present the symmetric Kullbeck-Leibler (KL) divergence values between the original gesture sequence $\boldsymbol{\ell}^g$ and the synthesized one $\hat{\boldsymbol{\ell}}^g$, $KL_{\boldsymbol{\ell}}$, as well as the symmetric KL divergence values between the original duration distribution $d_i(k)$ and the synthesized one $\hat{d}_i(k)$, KL_d, where smaller KL divergence values indicate more consistent distributions. Smallest $KL_{\boldsymbol{\ell}}$ value for f^{ap} feature set suggests that affect attiributes and prosody together provide better gesture id predictions. Similarly, the same feature set provides better duration predictions with the smallest KL_d value. In particular, feature sets containing the affect information have better gesture id prediction results and feature sets containing prosody information have better gesture duration prediction results, respectively.

Table 3. Symmetric KL divergence scores for synthesis results

	f^{ap}	f^{p}	f^{a}	f^{AP}
$KL_{\boldsymbol{\ell}}$	**3.69**	5.28	5.19	4.25
KL_d	**7.26**	7.30	8.29	7.71

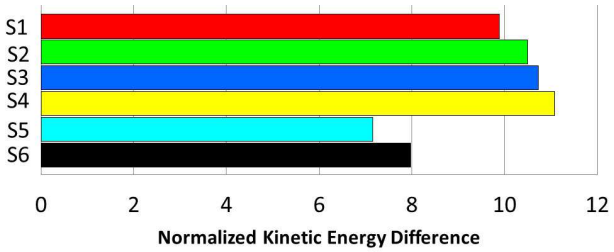

Fig. 4. Normalized kinetic energy difference of synthesized gesture sequences for the six scenarios.

4.4. Evaluation of Animation results

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 4. 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 4. 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. 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. 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 4 and 5 affect only driven synthesis systems have less KE difference

in Figure 4. Moreover, employing affect information either during the synthesis or animation process decreases the KE difference.

5. CONCLUSIONS

We investigate the role of affect in a gesture synthesis and animation framework. We jointly model phrase-level gestures with continuous affect attributes (activation, valence, dominance) and prosody using hidden semi-Markov models in speaker independent fashion on the USC CreativeIT dataset. Gesture sequences for affect and prosody feature fusion, prosody only and affect only configurations are synthesized. Our gesture synthesis evaluations based on CCA analysis and symmetric KL-divergence, respecively suggest that affect prosody fusion is the most correlated feature set with the original gesture trajectories and provide the best gesture and gesture duration modeling. On the other hand, we observe animations based affect only-driven synthesis results achieve the highest correlation scores with the original gesture trajectories, and affect attributes, especially for the valence domain. This feature set also has the least kinetic energy difference with the original sequence. On the contrary animations of the synthesis results including prosody as the driving factor demonstrate low valence correlations with the original sequence. The deficiency of prosody in revealing valence related cues degrades animation results in the database we use. Our future work will include subjective tests on the generated animations for analyzing influence of affect on the animations reflected by human perception.

6. ACKNOWLEDGMENTS

This work is supported by TÜBİTAK under Grant Number 113E102.

7. REFERENCES

[1] S. Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, Feb. 2010.

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

[3] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, Mary Ritchie Key, Ed., pp. 207–227. Mounton Publishers, The Hague, The Netherlands, 1980.

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

[5] P. Wagner, Z. Malisz, and S Kopp, "Gesture and speech in interaction: An overview," *Speech Communication*, vol. 57, pp. 209 – 232, 2014.

[6] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," *ACM Transactions on Graphics*, vol. 29, no. 4, July 2010.

[7] A. F. Baena, R. Montano, M. Antonijoan, A. Roversi, D. Miralles, and F. Alias, "Gesture synthesis adapted to speech emphasis," *Speech Communication*, vol. 57, pp. 331–350, 2013.

[8] E. Bozkurt, S. Asta, S. Ozkul, Y. Yemez, and 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.

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

[10] C. Busso and S. S. Narayanan, "Interrelation Between Speech and Facial Gestures in Emotional Utterances : A Single Subject Study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, Nov. 2007.

[11] S. Mariooryad and C. Busso, "Generating Human-Like Behaviors Using Joint , Speech-Driven Models for Conversational Agents," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 8, pp. 2329–2340, Oct. 2012.

[12] R. Niewiadomski, S.J. Hyniewska, and C. Pelachaud, "Constraint-based model for synthesis of multimodal sequential expressions of emotions," *Affective Computing, IEEE Transactions on*, vol. 2, no. 3, pp. 134–146, July 2011.

[13] A. Garcia-Rojas, F. Vexo, D. Thalmann, A. Raouzaiou, K. Karpouzis, and S. Kollias, "Emotional Body Expression Parameters In Virtual Human Ontology," in *Proceedings of 1st Int. Workshop on Shapes and Semantics*, 2006, pp. 63–70.

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

[15] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *Affective Computing, IEEE Transactions on*, vol. 4, no. 1, pp. 15–33, Jan 2013.

[16] J. Jia, Z. Wu, S. Zhang, H.. Meng, and L. Cai, "Head and facial gestures synthesis using pad model for an expressive talking avatar," *Multimedia Tools and Applications*, vol. 73, no. 1, pp. 439–461, 2014.

[17] B. Hartmann, M. Mancini, and 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*, Berlin, Heidelberg, 2006, GW'05, pp. 188–199, Springer-Verlag.

[18] M. Karg, A.-A. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic, "Body movements for affective expression: A survey of automatic recognition and generation," *Affective Computing, IEEE Transactions on*, vol. 4, no. 4, pp. 341–359, Oct 2013.

[19] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Analysis of Head Gesture and Prosody Patterns for Prosody-Driven Head-Gesture Animation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1330–1345, Aug. 2008.

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

[21] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917, 2002.

[22] "MotionBuilder: 3D Character Animation for Virtual Production, http://www.autodesk.com," 2012.