Speech Communication 85 (2016) 29–42

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures

Elif Bozkurt*, Yücel Yemez, Engin Erzin

Multimedia, Vision and Graphics Laboratory, College of Engineering, Koç University, 34450 Sariyer, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 11 September 2015
Revised 3 October 2016
Accepted 9 October 2016
Available online 11 October 2016

Keywords:
Joint analysis of speech and gesture
Speech-driven gesture animation
Prosody-driven gesture synthesis
Speech rhythm
Unit selection
Hidden semi-Markov models

ABSTRACT

We propose a framework for joint analysis of speech prosody and arm motion towards automatic synthesis and realistic animation of beat gestures from speech prosody and rhythm. In the analysis stage, we first segment motion capture data and speech audio into gesture phrases and prosodic units via temporal clustering, and assign a class label to each resulting gesture phrase and prosodic unit. We then train a discrete hidden semi-Markov model (HSMM) over the segmented data, where gesture labels are hidden states with duration statistics and frame-level prosody labels are observations. The HSMM structure allows us to effectively map sequences of shorter duration prosodic units to longer duration gesture phrases. In the analysis stage, we also construct a gesture pool consisting of gesture phrases segmented from the available dataset, where each gesture phrase is associated with a class label and speech rhythm representation. In the synthesis stage, we use a modified Viterbi algorithm with a duration model, that decodes the optimal gesture label sequence with duration information over the HSMM, given a sequence of prosody labels. In the animation stage, the synthesized gesture label sequence with duration and speech rhythm information is mapped into a motion sequence by using a multiple objective unit selection algorithm. Our framework is tested using two multimodal datasets in speaker-dependent and independent settings. The resulting motion sequence when accompanied with the speech input yields natural-looking and plausible animations. We use objective evaluations to set parameters of the proposed prosody-driven gesture animation system, and subjective evaluations to assess quality of the resulting animations. The conducted subjective evaluations show that the difference between the proposed HSMM based synthesis and the motion capture synthesis is not statistically significant. Furthermore, the proposed HSMM based synthesis is evaluated significantly better than a baseline synthesis which animates random gestures based on only joint angle continuity.

1. Introduction

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

Gesture and speech co-exist in time with a tight synchrony; they are planned and shaped by the cognitive state and produced together. In one of the pioneering studies on gesture and speech relationship, Kendon (1980) proposed a widely accepted hierarchical model for gesture in terms of phases, phrases and units. In this model, the core gestural element is defined as gesture phase. Gesture phases can be active or passive. An active gesture phase can be a stroke (a short and dynamic peak movement) with a retraction or a preparation (in which arm goes to the start position of the stroke phase). Passive gesture phases are movements like hold and rest, in which arm stays motionless. Combinations of phases constitute gesture phrases, and then combinations of phrases form gesture units. In this hierarchical model, semantic expressiveness increases with the level of hierarchy. In other words, gesture units are semantically more expressive than gesture phrases,

* Corresponding author.
E-mail addresses: ebozkurt@ku.edu.tr (E. Bozkurt), yyemez@ku.edu.tr (Y. Yemez), eerzin@ku.edu.tr (E. Erzin).

http://dx.doi.org/10.1016/j.specom.2016.10.004
0167-6393/

and gesture phrases include more semantic content than gesture phases.

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

There are four widely referred types of gestures, which were proposed by McNeill (1992): iconics, metaphorics, deictics and beats. Iconic gestures illustrate images of objects or actions, metaphoric gestures represent abstract ideas, deictic gestures relatively locate entities in physical space, and beat gestures are simple repetitive movements to emphasize speech. In a later study, based on the Tuite's proposal (Tuite, 1993) that in every gesture there is a rhythmical beat-like pulse to carry significance beyond its immediate setting, McNeill (2006) suggested taking metaphoricity, iconicity, deixis, and emphasis as dimensions of gesture rather than types of gesture.

Although there seems to exist strong correlation between gestures and speech, there are several challenges and difficulties involved in modeling this relationship. The first challenge is due to the diversity of gestures related to speech semantic content. Iconic, metaphoric and deictic gestures belong to this category which are mainly related to semantic content. In this work, we exclude modeling these gestures and rather focus on modeling beat gestures in relation to speech prosody and rhythm. There are yet two main challenges in achieving this goal. The first is due to the difficulty in mapping local prosody information to relatively longer duration semantic gestures, e.g., gesture phrases. The second challenge is actually a temporal alignment problem: prosodic cues and corresponding gestures do not always co-occur at exactly the same time instant, and there may indeed be a lagged correlation in between. In this paper, to address the above challenges, we model the relationship between longer duration gesture phrases and shorter duration prosodic units using hidden semi-Markov models (HSMM), and attain an effective modeling and control of gesture durations for analysis and synthesis. In order to synthesize beat gesture sequences with optimal gesture durations, we employ Viterbi decoding over an HSMM structure with prosody observations. In the animation stage, synthesized gesture sequence with duration and speech rhythm information is mapped into body motion sequences by using a multiple objective unit selection algorithm.

1.1. Related work

Speech-driven gesture animation methods in the literature can be categorized by the type of input that they accept, such as textual or spoken. Text-driven methods target to generate and animate body gestures from a tagged text input (Cassell et al., 1994; 2001; Kopp and Wachsmuth, 2004; Marsella et al., 2013; Neff et al., 2008; Noma et al., 2000; Pelachaud, 2005; Reithinger et al., 2006; Stone et al., 2004). In this category, the Embodied Conversational Agents (ECAs) of Cassell et al. (1994) is a pioneering rule-based dialog system which can animate conversations between human-like agents with appropriate and synchronized speech, intonation, facial expressions, and hand gestures. In this system, the mapping from text to gestures is contained in a set of rules derived from nonverbal conversational behavior research, and gestures also collected in a pre-defined gesture dictionary. The Behavior Expression Animation Toolkit (BEAT) (Cassell et al., 2001) can be thought of as a more complex version of the gesture generation system proposed in Cassell et al. (1994), which uses linguistic and contextual information contained in the speech text to control movements of the hands, arms and face, and intonation of the voice. Another example to text-driven methods is the Virtual Human Presenter of Noma et al. (2000), which generates gestures using keyword triggered rules. More recent works such as VirtualHuman (Reithinger et al., 2006) and multimodal expressive ECAs (Pelachaud, 2005) aim to develop interactive virtual characters with personality profiles and full-body gesture animation by taking into account characters' affective states. Marsella et al. (2013) 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. In contrast to the rule-based methods mentioned above, two related methods Stone et al. (2004) and Neff et al. (2008) follow a data-driven approach to generate gesticulation of a particular speaker from speech text. Stone et al. (2004) use motion graphs to rearrange pre-recorded audio and motion segments, whereas Neff et al. (2008) develop a probabilistic framework to learn an abstract statistical model for the gesticulation of a speaker from annotated audiovisual data.

While text-driven gesture animation methods mostly focus on modeling metaphoric, deictic and iconic gestures, audio-driven methods are generally more suited to modeling beat gestures based on prosodic and intonational cues existing in speech signal. Yet there exist relatively very few works in the literature for beat gesture animation including arm movements driven by speech audio such as Fernandez-Baena et al. (2013); Levine et al. (2010). Levine et al. (2010) introduce gesture controllers, availing a modular methodology to drive beat-like gestures with live speech via customized gesture repertoires. Gesture controllers infer hidden states from speech using a conditional random field that analyzes acoustic features in the input and select the optimal gesture kinematics based on the inferred states. From a hierarchical perspective, the work of Levine et al. (2010) is mainly concentrated on the gesture phase level, whereas in a more recent study, Fernandez-Baena et al. (2013) 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.

Early works on audio-driven virtual character animation have mostly been concentrated on lip synchronization on which there exists currently a vast and quite mature literature (Bregler et al., 1997; Chen and Rao, 1998). Lip synchronization is basically formulated as a mapping from phonemes to visemes for which the state of the art methods commonly employ hidden Markov models (Li and Shum, 2006; Xue et al., 2006). Since facial motion is usually dominated by lip movements during speech, animation of facial expressions have so far received relatively less attention. Though there have been several attempts that address this challenging problem (Albrecht et al., 2002; Chuang and Bregler, 2005; Hong et al., 2002). In particular, Chuang and Bregler (2005) describe a method for creating expressive facial animation based on a statistical model learned from video for factoring the expression and lip speech. They also integrate head motion synthesis to their face animation scheme by first building a database of examples which relate audio pitch to motion and then matching new audio streams against segments in this database. The head motion synthesis problem, which is very crucial for generation of believable face animations, has been addressed in the recent literature. Busso and Narayanan (2007) present an approach to synthesize emotional head motion sequences driven by prosodic features, that

builds hidden Markov models for emotion categories to model the temporal dynamics of emotional head motion sequences. Sargin et al. (2008) propose a two-stage framework for joint analysis of head gesture and speech prosody patterns of a speaker towards automatic realistic synthesis of head gestures from speech prosody. In a more recent paper Mariooryad and Busso (2012) focuses on building a speech-driven facial animation framework to generate natural head and eyebrow motions using dynamic Bayesian networks (DBNs).

In this work, we employ hidden semi-Markov models (HSMM) for multimodal analysis of gestures and prosody. The HSMM was first introduced by Ferguson (1980) as explicit duration hidden Markov models. The main intuition behind the HSMM idea is to extend hidden Markov models to processes where states have durations and thus emit a number of observations instead of a single one. This assumes that the underlying process is Markovian in certain jumps. Moreover, the state duration is allowed to follow a probabilistic distribution.

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

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

In addition to the time-domain representations that we have discussed above, there exist also frequency domain representations of speech rhythm. Dynamic information extracted both at local and global level from frequency domain representation of speech rhythm has been used for assessment of emotion (Ringeval et al., 2012). In this study, we compile a dictionary of speech rhythm representations per gesture category to define a relational model between rhythmic similarities of speech and gesture modalities. We use the low-frequency Fourier analysis of speech rhythm as introduced by Tilsen and Johnson (2008) to investigate the relationship between speech rhythmicity and vowel, consonant deletions on the Buckeye corpus.

1.2. Contributions

Our primary contribution in this paper is the HSMM based gesture model that we use to capture the relationship between gestures and speech prosody. In this model, gestures are hidden states with duration distributions, hence each gesture instance spans a random number of observations (prosodic units). The benefit of this model is two-fold. First, it allows us to effectively map shorter duration prosodic units to longer duration gestures, hence we can synthesize semantically high-level gestures, i.e., gesture phrases. Second, since we observe a number of prosodic units to decide on gesture type and timing, we can handle to some extent the temporal misalignment problem for correlating prosodic cues to gestures.

Another contribution of the paper is the unit selection based gesture animation system. Given speech prosody, our HSMM based gesture model generates a sequence of gesture labels (each indicating the type of a gesture phrase) as synthesis output. Our animation system then maps this gesture sequence with duration information to a body motion (joint angle) sequence by minimizing a multiple objective cost function. This cost function penalizes mismatches in speech rhythm as well as discontinuities in gesture transitions and deviations from optimal gesture durations.

We can compare our contributions to the two closely related works of Levine et al. (2010) and Fernandez-Baena et al. (2013). Gesture controllers of Levine et al. (2010) model the relationship between speech prosody and kinematic parameters of the motion capture data stream using a conditional random field. Based on the kinematic parameters inferred from this conditional random field given speech, the best gesture segments (gesture phases) are selected from a gesture repertoire via dynamic programming. In this sense gesture controllers model the correlation between speech and gesture at a lower level of semantics compared to our system, that is, primarily at the level of kinematic parameters, and then at the level of gesture phases. In our case, the use of HSMM allows us to analyze and model gestures at a higher level of semantics, i.e., directly at the level of gesture phrases. We also note that gesture controllers cannot model gesture transition probabilities due to their low-level inference modeling. As a result, our system yields more personalized and hence more consistent and expressive synthesis results. Fernandez-Baena et al. (2013) also use gesture phrases in their synthesis system. However, unlike our fully automatic synthesis approach, their system requires manually annotated prosody input. They use gesture motion graphs (GMG) permitting connections between consecutive and similar gestures, where gestures with smooth transitions are considered to be similar. The gestures are synthesized on this graph by selecting each time the gesture clip which is the most suited for a given pitch accent. Hence, in contrast to our approach, they do not use any statistical inference model and disregard the actual gesture transition probabilities. Moreover, the synchrony window, which models temporal alignment of gestures and speech, does not take into account the durational statistics of gestures. Fernandez-Baena et al. (2013) report that their animation system may produce poor results in a limited dataset due to excessive warping.

Furthermore we should note that we have presented preliminary results of our prosody-driven gesture synthesis system based on HSMM in an earlier paper (Bozkurt et al., 2013). In this current paper, we include an extensive description of the unit-selection based animation generation system within the HSMM framework, extend our original framework by introducing the speech rhythm information as a third modality to overcome any rhythmic mismatches in animations, evaluate our framework on two datasets for speaker-dependent and independent settings, and present objective evaluation methods to fine-tune gesture synthesis and animation parameters prior to subjective evaluations.

1.3. Overview

The general block diagram of our speech-driven gesture synthesis and animation system is given in Fig. 1. The system consists of three main tasks: analysis, synthesis, and animation. Within the analysis task we have two stages: (i) feature extraction and clustering of gesture phrases and prosodic units, and (ii) their multimodal analysis. Section 2 presents the first stage of the analysis task, where we perform feature extraction and unimodal clustering on speech and body motion data. The audio stream is processed to extract prosodic features of the speech, whereas the body motion

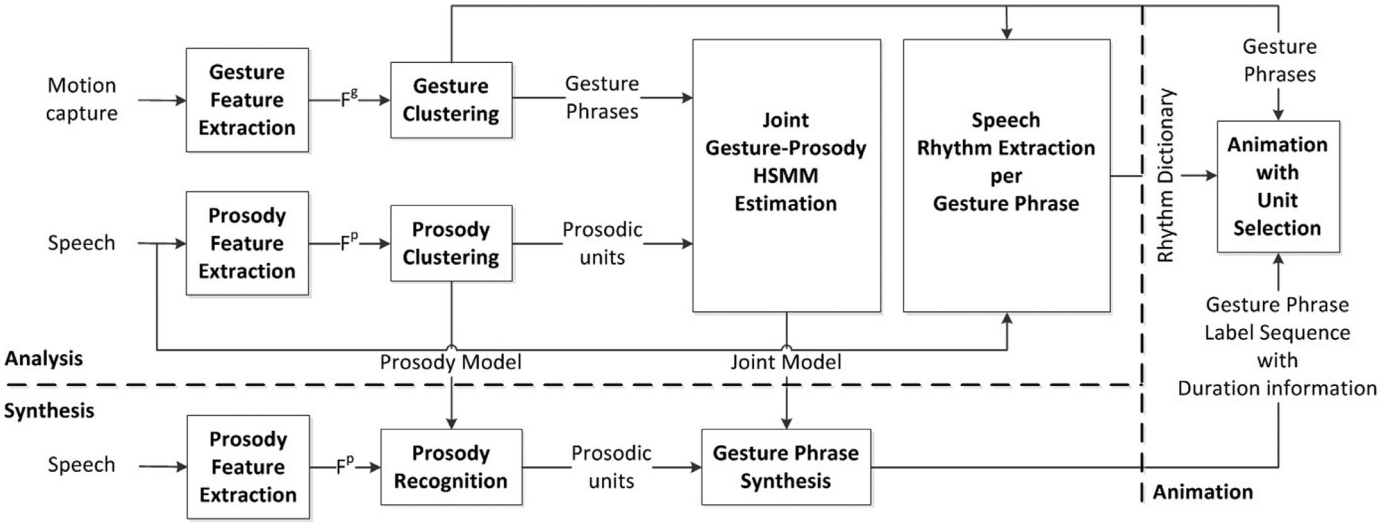


Fig. 1. The block diagram of the general framework for the speech-driven gesture synthesis and animation system.

data is expressed in terms of 3D joint angles. The gesture and audio feature streams are then segmented via temporal clustering into recurrent patterns, i.e, into gesture phrases and prosodic units, respectively. In Section 2, in addition to prosodic features, we also describe extraction of speech rhythm features for the duration of each gesture phrase, which are to be used later in the animation generation stage. Section 3 describes the second stage of the analysis task, where we use hidden semi-Markov models (HSMM) to model the dependencies between speech prosody and gestures in a multimodal framework. Sections 4 and 5 mainly address the synthesis and animation tasks, respectively. In Section 4, we describe how an optimal sequence of gestures with durations is synthesized for a given speech input by using the Viterbi algorithm over HSMM. In Section 5, the synthesized gesture sequences with duration information are mapped into body motion sequences by using a multiple objective unit selection algorithm to generate animations. In Section 6, we present the results of the speaker-dependent and independent experiments conducted for objective and subjective evaluations of the proposed system, and finally in Section 7, we provide our concluding remarks.

2. Feature extraction and unimodal clustering

We employ semi-supervised and unsupervised temporal clustering schemes to determine boundaries and categories of gesture phrases for speaker-dependent and independent scenarios, respectively. We use an unsupervised scheme to cluster speech into prosodic units. We also extract speech rhythm feature along with each gesture phrase.

2.1. Prosody clustering

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

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

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

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

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

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

$$c_k = 1 - \frac{e_k(\tau^*)}{\frac{1}{\tau^*}\sum_{i=1}^{\tau^*} e_k(i)}, \tag{3}$$

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

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

$$\boldsymbol{f}_k^p = [\bar{I}_k,\ \bar{\nu}_k,\ \bar{c}_k,\ \Delta\bar{I}_k,\ \Delta\bar{\nu}_k,\ \Delta\bar{c}_k], \tag{4}$$

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

We employ unsupervised temporal clustering using the parallel HMM architecture in Sargin et al. (2008) to extract prosody clusters. The prosody feature stream $\boldsymbol{F}^p = \{\boldsymbol{f}_1^p, \boldsymbol{f}_2^p, \ldots, \boldsymbol{f}_T^p\}$ is used to train a parallel branch HMM structure, Λ^p, which clusters the prosody feature stream and captures recurrent prosodic units through unsupervised learning. The HMM structure Λ^p is composed of M_p parallel left-to-right HMMs, $\{\lambda_1^p, \lambda_2^p, \ldots, \lambda_{M_p}^p\}$, where

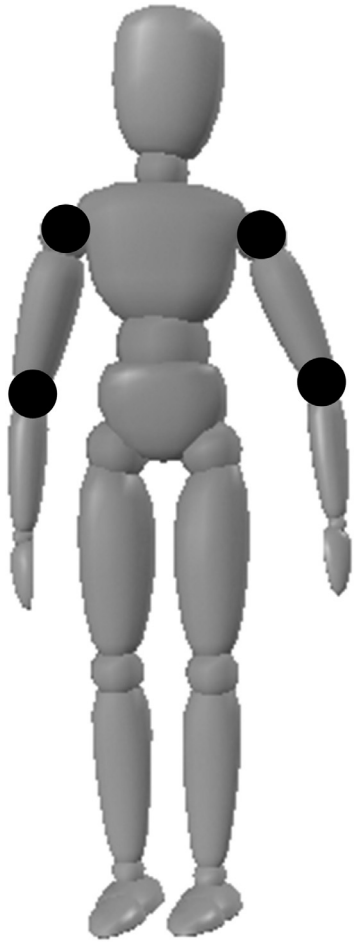


Fig. 2. The black circles correspond to the joints used for gesture representation as left forearm, left arm, right arm, and right forearm, respectively.

each λ_m^p is composed of N_p states. The HMM-based unsupervised clustering process segments the prosody feature stream into prosodic units. We denote the lth prosodic unit in the stream by ε_l^p and the associated class label of the lth unit by ℓ_l^p, which is one of the M_p available prosody classes $\{p_1, p_2, \ldots, p_{M_p}\}$. A frame level label sequence is then defined for the prosodic unit sequence,

$$\xi_k = \ell_l^p \quad \text{for} \quad k = k_l, \quad k_l + 1, \ldots, k_{l+1} - 1, \tag{5}$$

where ξ_k is the prosody label of the kth speech frame and $[k_l, k_{l+1})$ spans the lth prosodic unit. These frame level prosody labels will eventually serve as the observations of the HSMM structure that we will describe in Section 3.

2.2. Gesture clustering

We model gestures, specifically beat type arm gestures, at gesture phrase level to correlate with and emphasize speech prosody. For analysis of gestures, we employ joint angles from four body parts: left arm, left forearm, right arm, and right forearm as shown in Fig. 2. We define the joint angle vector for the ith joint at frame k as $\theta_k^i = [\phi_x^{ik}, \phi_y^{ik}, \phi_z^{ik}]$, where $\phi_x^{ik}, \phi_y^{ik}, \phi_z^{ik}$ are the Euler angles respectively in the x, y, z directions, representing the orientation of the ith joint at frame k. 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,

$$\boldsymbol{f}_k^{Ji} = [\theta_k^i, \Delta\theta_k^i], \quad \text{for } i = 1, 2, 3, 4, \tag{6}$$

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

$$\boldsymbol{f}_k^g = [\boldsymbol{f}_k^{J1}, \ldots, \boldsymbol{f}_k^{J4}]. \tag{7}$$

2.2.1. Semi-supervised gesture clustering

Temporal clustering of the gesture feature sequence is necessary for analysis of recurring beat gesture phrases. In the speaker-dependent case, we implement a semi-supervised clustering method using the parallel branch HMM structure, Λ^g, over the gesture feature stream $\boldsymbol{F}^g = \{\boldsymbol{f}_1^g, \boldsymbol{f}_2^g, \ldots, \boldsymbol{f}_T^g\}$, with duration of T frames. The HMM structure Λ^g initially is set to have two parallel branch HMMs, $\{\lambda_1^g, \lambda_2^g\}$, where each λ_m^g is composed of $N_g = 10$ states corresponding to the minimum gesture phrase duration of 10 frames ($\frac{1}{3}$ s assuming 30 video frames/sec). The number of branches is iteratively increased to M_g in a semi-supervised manner using the following procedure:

(i) Initially set Λ^g to have two branches to model the *rest* position of arms and all the other remaining arm movements. Manually label examples of the rest event (as many as necessary to train an HMM structure) from gesture stream by inspecting the video.
(ii) Perform the Baum–Welch training of the Λ^g.
(iii) Perform Viterbi decoding to get temporal clusters.
(iv) Visually inspect and correct clusters as needed. Repeat steps (ii) and (iii) until convergence.
(v) If a new gesture phrase, which is recurrent in the data and not covered by the Λ^g, exists, go to step (vi), otherwise stop.
(vi) Manually label several examples of the new gesture phrase. Add a branch to the Λ^g for the new gesture phrase with initial training. Go to step (ii).

The proposed semi-supervised clustering process segments the gesture feature stream into gesture phrases, ε_l^g, with label ℓ_l^g as one of the M_g available gesture classes $\{g_1, g_2, \ldots, g_{M_g}\}$. All gesture phrases ε^{g_i} from gesture class g_i are grouped together to build a gesture pool $G_i = \{\varepsilon^{g_i 1}, \ldots, \varepsilon^{g_i j}, \ldots, \varepsilon^{g_i N_i^G}\}$, where $\varepsilon^{g_i j}$ is the jth gesture phrase in the pool, and N_i^G is the number of gesture phrases in the gesture pool G_i. For each gesture phrase in the gesture pool, we also extract a speech rhythm feature as defined in the Section 2.3.

2.2.2. Unsupervised gesture clustering

Unsupervised gesture clustering is applied for the speaker-independent setting, where a large scale multimodal dataset has been used for this purpose. Unlike the clustering results of the semi-supervised approach in Section 2.2.1, the resulting gesture patterns of unsupervised clustering are not explicitly compatible with the gesture phrase definition presented in Kendon (1980). However, there is reported evidence that these patterns are meaningful for explaining the nature of gestures (Yang et al., 2014; Yang and Narayanan, 2016). For simplicity, we will use the same notation with the gesture phrases in Section 2.2.1, as ε^{g_i} for the gesture patterns resulting from gesture class g_i.

Similarly to the prosody clustering process in Section 2.1, we apply the unsupervised clustering method based on parallel-HMMs (Sargin et al., 2008) to the gesture feature sequence $\boldsymbol{F}^g = \{\boldsymbol{f}_1^g, \boldsymbol{f}_2^g, \ldots, \boldsymbol{f}_T^g\}$, with duration of T frames. We segment and cluster gesture sequences into gesture patterns with duration information. The HMM structure Λ^g is set to have $M_g = 40$ parallel branch HMMs, $\{\lambda_1^g, \ldots, \lambda_{M_g}^g\}$, where each λ_m^g is composed of $N_g = 10$ states corresponding to the minimum gesture pattern duration of 10 frames ($\frac{1}{3}$ s assuming 30 video frames/sec) considering the works by Yang et al. (2014) and Bozkurt et al. (2015). We also create a gesture pool G_i for each gesture class g_i and extract a speech rhythm feature for each gesture pattern in the pool.

2.3. Speech rhythm feature extraction

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

Speech is a rhythmic and temporally structured source where the acoustic signal is transmitted as syllables in which most of the energy fluctuations occur in the range between 3 to 20 Hz (Greenberg, 1999; Greenberg and Arai, 2004). When we refer to the term rhythm, we do not mean that these energy terms are perfectly periodic, but rather that there are regulations on syllable duration and energy patterns within and across prosodic phrases, which are important for intelligibility and naturalness of the spoken speech (Ladd, 1996; Liberman, 1975). For example, from a phonetic point of view, we cannot fully define speech as sequences of phonemes, syllables or words. When we listen to speech, we hear that segments or syllables are shortened or lengthened in accordance with an underlying pattern. However, characterization of speech rhythm is not an easy task itself and most of the methods rely on measurements of segmental durations to describe the temporal patterns of speech (Gibbon and Gut, 2001; Grabe and Low, 2002; Loukina et al., 2011; Ramus et al., 1999).

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

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

$$\boldsymbol{r}(i,j) = \frac{1}{\sum_n e_n}[e_1, e_2, \ldots, e_d], \quad \text{for} \quad i = 1,2,\ldots,M_g;\ j = 1,2,\ldots,N_i^G, \tag{8}$$

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

The collection of speech rhythm feature vectors, $\boldsymbol{r}(i, j)$, are compiled as a dictionary of speech rhythm representations per gesture class and expressed as $R_i = \{\boldsymbol{r}(i,1), \ldots, \boldsymbol{r}(i, N_i^G)\}$. In other words, the speech rhythm dictionary R_i is linked with the gesture pool G_i defined in Section 2.2, where each gesture phrase has a coinciding speech rhythm representation. In the animation generation step, R_i is used to reduce rhythmic mismatches between the input speech and the gesture phrases selected from the gesture pool G_i.

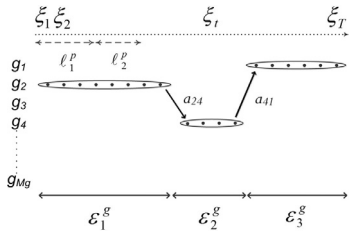


Fig. 3. In the hidden semi-Markov model, prosodic units are labeled as ℓ_i^p, frame-level prosody labels (ξ_t) correspond to observations and gesture phrases (ε_k^g) correspond to states (g_m).

Although rhythm could be seen as a timing aspect of speech prosody along with intonation and stress, it represents phrase-level timing characteristics. Hence rather than including it in the short-term prosody feature representation, we rather use it as a similarity factor while creating gesture animations as will be described in Section 5.

3. Multimodal analysis of gestures and prosody

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

Synthesizing a gesture phrase sequence using the conventional Markov chain model given prosodic unit observations would however have a shortfall in modeling gesture phrase durations in time. A useful mathematical model to overcome this shortfall is to introduce a statistical state duration model so that one can better control gesture phrase durations in the synthesis process. Combination of these two useful mathematical models, i.e., Markov chain and state duration, yields the hidden semi-Markov model (HSMM) framework (Yu, 2010) that we use for multimodal analysis of gesture phrases and prosodic units. HSMM is an extension of HMM, which allows the underlying process to be a semi-Markov chain with states having variable durations. This is to say that the underlying process is Markovian at certain jump instants (Barbu and Limnios, 2008). Fig. 3 shows how such an HSMM structure functions, where gesture phrase labels and frame-level prosody labels are depicted as states and observations, respectively. Gesture transitions define state transitions, whereas prosodic unit distributions per gesture class define the observation emission distributions. The state duration distributions are estimated over the gesture phrase duration information. Note that the observations of the HSMM structure are frame-level prosody labels defined in (5). This is essential since in this way, hidden state durations, hence gesture durations, can be represented in terms of fixed length speech frames.

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

probability, observation emission distribution, state duration distribution, and initial state distribution matrices.

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

$$\boldsymbol{A} : \{a_{ij} = P(\ell_l^g = g_j | \ell_{l-1}^g = g_i)\} \quad i, j = 1, \ldots, M_g, \tag{9}$$

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

$$\boldsymbol{B} : \{b_i(p_j) = P(p_j | \ell_l^g = g_i)\} \quad i = 1, \ldots, M_g, \ j = 1, \ldots, M_p, \tag{10}$$

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

$$\boldsymbol{D} : \{d_i(n)\} \quad i = 1, \ldots, M_g, \ n = 1, \ldots, \frac{D_{max}}{\delta}, \tag{11}$$

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

$$\boldsymbol{\Pi} : \{\pi_i = P(\ell_1^g = g_i)\} \quad i = 1, \ldots, M_g. \tag{12}$$

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

$$\pi_i = P(\ell_1^g = g_i) \ \widehat{=} \ \frac{C(1, i, j)}{\sum_{j'} C(1, i, j')}, \tag{13}$$

$$a_{ij} = P(\ell_l^g = g_j | \ell_{l-1}^g = g_i) \ \widehat{=} \ \frac{\sum_l C(l, i, j)}{\sum_l \sum_{j'} C(l, i, j')}, \tag{14}$$

$$b_i(p_j) = P(p_j | \ell_l^g = g_i) \ \widehat{=} \ \frac{O(i, j)}{\sum_{j'} O(i, j')}, \tag{15}$$

$$d_i(n) \ \widehat{=} \ \frac{H(i, n\delta \le t < (n+1)\delta)}{\sum_{n'} H(i, n'\delta \le t < (n'+1)\delta)}, \tag{16}$$

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

4. Gesture synthesis

Gesture synthesis is defined as decoding an optimal state sequence, $\hat{\boldsymbol{\ell}}^g$, over the HSMM Λ^{gp} given a sequence of frame level prosodic unit labels, $\{\xi_1, \xi_2, \ldots, \xi_T\}$ (see (5)). Note that the decoded optimal state sequence delivers a synthesized label sequence for gesture phrases, $\hat{\boldsymbol{\ell}}^g$, and a sequence of associated durations, $\boldsymbol{\kappa}$, where the HSMM framework secures to have realistic gesture phrase durations. In HMM framework, where the underlying process is Markov, given an observation sequence, the Viterbi algorithm is employed to decode the most likely state sequence. In HSMM framework however, states have variable durations and a sequence of observations are emitted at a single state. This requires us to define a forward likelihood function, which incorporates state duration model,

$$\psi_t(j) = \max_{\tau} \max_{i} \left\{ \psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(\xi_k)) \right\}, \tag{17}$$

where $\psi_t(j)$ is the accumulated logarithmic likelihood at time frame t in state g_j after observing prosody labels $\{\xi_1, \xi_2, \ldots, \xi_t\}$. Based on the forward likelihood function $\psi_t(j)$, we define the following modified Viterbi decoding algorithm to extract the optimal state sequence:

i. Initialize
   $\psi_1(i) = \log(\pi_i b_i(\xi_1)) \quad i = 1, 2, \ldots, M_g$
ii. Recursion: Repeat for $t = 2, 3, \ldots, T$
   $T' = \min(D_{max}, t)/\delta$
   Repeat for $j = 1, 2, \ldots, M_g$
   $\Psi_{t\tau}^{ij} = \psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(\xi_k))$
   for $i = 1, \ldots, M_g, \ \tau = 1, \ldots, T'$
   $\psi_t(j) = \max_{\tau \in [1,T']} \max_{i \in [1,M_g]} \{\Psi_{t\tau}^{ij}\}$
   $\varphi_t(j) = \arg\max_{i \in [1,M_g]} \max_{\tau \in [1,T']} \{\Psi_{t\tau}^{ij}\}$
   $\nu_t(j) = \arg\max_{\tau \in [1,T']} \max_{i \in [1,M_g]} \{\Psi_{t\tau}^{ij}\}$
iii. Backtrace the optimal gesture phrase sequence
   $\hat{\ell}_L^g = \arg\max_j \psi_T(j)$
   $\kappa_L = \nu_T(\hat{\ell}_L^g)$
   $l = L - 1; \quad t = T$
   While $t > 0$
   $\hat{\ell}_l^g = \varphi_t(\hat{\ell}_{l+1}^g)$
   $\kappa_l = \nu_{t-\kappa_{l+1}}(\hat{\ell}_l^g)$
   $t = t - \kappa_{l+1}; \quad l = l - 1$

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

5. Gesture animation

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

The first task is to generate a synthesized sequence of gesture phrases, $\hat{\boldsymbol{\varepsilon}}^g$, given the synthesized gesture phrase label sequence $\hat{\boldsymbol{\ell}}^g$ along with the corresponding duration sequence $\boldsymbol{\kappa}$ and the input speech rhythm information. This task is performed using unit selection over a pool of gesture phrases which are extracted during the gesture analysis in Section 2.2. The next task is to smooth joint angle discontinuities over a temporal window at gesture phrase boundaries, that is, at the boundary of each two consecutive synthesized gesture phrases $\hat{\varepsilon}_l^g$ and $\hat{\varepsilon}_{l+1}^g$, to extract a smoothed gesture sequence $\tilde{\boldsymbol{\varepsilon}}^g$. The smoothed gesture phrase sequence $\tilde{\boldsymbol{\varepsilon}}^g$ is finally used to animate beat gestures of a virtual character in synchrony with the input speech.

We employ a unit selection algorithm to generate the synthesized sequence of gesture phrases, $\hat{\boldsymbol{\varepsilon}}^g$, based on the gesture pool $G_i = \{\varepsilon^{g_i 1}, \varepsilon^{g_i 2}, \ldots, \varepsilon^{g_i N_i^G}\}$, which is constructed in Section 2.2 for each gesture phrase class g_i. The unit selection process is demonstrated in Fig. 4 as the formation of an optimal gesture phrase sequence from the templates available in the gesture pools.

We minimize a mixture of penalty scores for duration mismatch, joint angle continuity mismatch and speech rhythm mismatch during the unit selection process. Speech rhythm similarity of gestures is used to avoid rhythmic mismatches between the input speech and the synthesized gesture motions during animations. We use the rhythm dictionary R_i for gesture class g_i, constructed as described in Section 2.3. The duration, joint angle continuity and speech rhythm mismatch penalties of a gesture

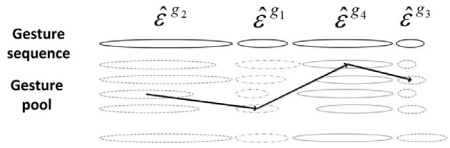


Fig. 4. Unit selection based gesture animation generation: an optimal sequence of gesture phrases is formed from the gesture phrase templates available in the gesture pool for each gesture class.

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

$$D_\omega(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = ||\omega_e(\hat{\varepsilon}_{l-1}^g) - \omega_b(\varepsilon^{g_i j})||, \quad (18)$$

$$D_\kappa(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = ||\kappa_l - \kappa(\varepsilon^{g_i j})||, \text{and} \quad (19)$$

$$D_r(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = ||r_l - r(i,j)||, \quad (20)$$

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

$$\begin{aligned} D(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) = \beta\alpha\overline{D_\omega}(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i) + \beta(1-\alpha)\overline{D_\kappa}(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i)] \\ + (1-\beta)\overline{D_r}(\varepsilon^{g_i j}|\hat{\ell}_l^g = g_i), \end{aligned} \quad (21)$$

where $\overline{D_\omega}$, $\overline{D_\kappa}$ and $\overline{D_r}$ are min-max normalized penalty functions, and α and β are the mixture weights which we set experimentally over a validation set, as will be explained in Section 6.4.

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

i. Initialization
   $V_1(j) = D(\varepsilon^{g_i j}|\hat{\ell}_1^g = g_i)$, for $j = 1, 2, \ldots, N_i^G$
ii. Recursion: Repeat for $l = 2, 3, \ldots, L$,
   Let $g_{i'} = \hat{\ell}_{l-1}^g$, for $j = 1, 2, \ldots, N_i^G$
   $V_l(j) = \min_{j'=1,\ldots,N_{i'}^G}\{V_{l-1}(n) + D(\varepsilon^{g_i j'}|\hat{\ell}_l^g = g_i)\}$,
   $Q_l(j) = \arg\min_{j'=1,\ldots,N_{i'}^G}\{V_{l-1}(n) + D(\varepsilon^{g_i j'}|\hat{\ell}_l^g = g_i)\}$,
iii. Backtrace the optimal path
   $q_L = \arg\min_j\{V_L(j)\}$,
   $q_l = Q_{l+1}(q_{l+1})$ for $l = L-1, L-2, \ldots, 1$,
iv. Construct the synthesized sequence of gesture phrases
   $\hat{\varepsilon}_l^g = \varepsilon^{\hat{\ell}_l^g q_l}$ for $l = 1, 2, \ldots, L$.

The selected gesture phrases are resampled so as to fit the synthesized duration if necessary. Next, we smooth joint angle discontinuities at gesture boundaries over a temporal window. This is achieved by applying an exponential smoothing function on each pair of consecutive synthesized gesture phrases $\hat{\varepsilon}_l^g$ and $\hat{\varepsilon}_{l+1}^g$. Then, the smoothed gesture motion sequence $\tilde{\boldsymbol{\varepsilon}}^g$ is used to animate a virtual character based on the four skeleton joints mentioned in Section 2.2. The other joints of the body (e.g. spine, lower-body joints) are assumed to have no motion. For graphical animation, we use the MotionBuilder 3D Character Animation Software[1]

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

Table 1
Gesture phrases identified via semi-supervised clustering.

Gesture	Description
g_1	Symmetric: both arms move symmetrically
g_2	Rest: no motion
g_3	Left: only left arm moves
g_4	Asymmetric: both arms move asymmetrically
g_5	Contact: hands touch to each other
g_6	Open: stretching arms backwards
g_7	Right: only right arm moves

Table 2
Gesture phrase distributions per recording.

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

6. Experimental results

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

6.1. Dataset for speaker-dependent setting

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

Since gestures are in general person specific, the gesture phrases are determined by using the semi-supervised training scheme described in Section 2.2. We have identified the number of distinct gesture phrases as $M_g = 7$ for the given participant. A brief description of these gesture classes is provided in Table 1, whereas their distribution per recording is summarized in Table 2. In our

experiments, we have spared the fourth recording as the validation set, and performed a leave-one-out training procedure such that one recording out of four is used for testing and the remaining three are used for training in four turns. The models resulting from training are used for synthesizing and animating gestures over the test recordings. Then, we perform subjective evaluations to assess naturalness and audio-visual synchrony of the animation results of the test recordings.

The proposed HSMM based synthesis is subjectively evaluated in pairwise comparisons with baseline and motion-capture synthesis results. Our baseline gesture synthesis method creates an animation from a sequence of random gestures via gesture phrase selection based solely on joint angle continuity, hence discarding any gesture duration and transition statistics as well as prosody-gesture correlations. The motion-capture synthesis on the other hand uses the captured true motion in the animations; that is the speaker's gestures are directly copied to the animation.

6.2. Dataset for speaker-independent setting

In the speaker-independent animation system, we use the multimodal USC CreativeIT dataset that contains a variety of dyadic theatrical improvisations for studying expressive behaviors and natural human interaction (Metallinou et al., 2010, 2016). In this dataset the interactive performances are designed either as improvisations of scenes from theatrical plays or as theatrical exercises where actors repeat sentences in a manner that conveys specific intent such as, accepting or rejecting behavior towards the other.

The dataset contains vocal and body-language behavior information of the actors obtained through close-up microphones, Motion Capture (MoCap) and HD cameras. The MoCap data is provided as 3D coordinates of 45 marker positions in (x, y, z) directions at 60 fps and speech recordings at 48 kHz for each of the 16 distinct actors (9 of whom are female). We use the MotionBuilder software[2] for converting 3D joint positions to Euler angle rotations of the arm and forearm in the (x,y,z) directions at 30 fps. We perform speaker-independent evaluations in a leave-one-pair-out manner using data from one actor-pair as the test set in turn, and the remaining data from the other pairs as the training data.

6.3. Number of prosody clusters

One of our primary goals in this work is to synthesize gesture sequences with realistic gesture durations. Hence, one possible objective evaluation of our HSMM based gesture synthesis is to consider the similarity between the original and the synthesized gesture duration statistics. To this effect, we perform the prosody clustering process under different parameter settings, and for each setting we synthesize a different gesture sequence. We then estimate the duration distribution of each synthesized sequence as in (11). Next, we compute the symmetric Kullbeck–Leibler (KL) divergence between the original duration distribution $d_i(k)$ and the synthesized distribution $\hat{d}_i(k)$ over the validation set to measure the duration similarity of the synthesized gesture sequence with the original one, where smaller KL divergence values indicate more consistent duration distributions.

We perform unsupervised prosody clustering using parallel-branch HMM structures, as described in Section 2.1, with branch numbers ranging from 10 to 18 and state numbers per branch ranging from 2 to 5 for the speaker-dependent setting. An over-segmented prosody stream with larger number of branches would produce redundant and similar clusters while under-segmentation with less branches would have to merge distinct clusters. The range of values for setting number of states on the other hand is selected considering the minimum duration of temporal prosody clusters. Table 3 presents the symmetric KL divergence scores on the validation set for various parameter settings. We observe that with small number of branches, $M_p < 14$, the KL divergence has its minimum at $N_p = 4$ number of states per branch. As the number of branches gets larger, $M_p \geq 14$, the optimal KL divergence value is attained for smaller number of states per branch. We use the $M_p = 16$ and $N_p = 3$ as the optimal setting with a KL divergence value of 1.0498 for subjective evaluation of our gesture synthesis system, which is presented in Section 6.5.

Table 3
The symmetric KL divergence of the original and synthesized gesture duration distributions for various prosody clustering settings in the speaker-dependent system.

N_p	M_p				
	10	12	14	16	18
2	2.1013	1.2358	1.3212	1.1848	1.2793
3	2.1251	1.1501	1.3333	**1.0498**	1.3544
4	1.1052	1.0698	1.6390	1.5562	1.8460
5	1.5551	1.9665	1.5705	1.7472	1.4214

Table 4
The symmetric KL divergence of the original and synthesized gesture duration distributions for various prosody clustering settings in the speaker-independent system.

N_p	M_p			
	10	16	24	32
2	7.8353	7.8267	7.9192	7.7408
3	7.8959	**7.5270**	7.7871	7.9978
4	7.7411	7.8443	7.7154	7.9411
5	7.7343	7.8207	7.8125	7.9100

Moreover, we perform unsupervised prosody clustering on the CreativeIT dataset for the speaker-independent evaluations. We vary the number of states per branch ranging from 2 to 5 and the number of branches ranging from 10 to 32. Table 4 presents the symmetric KL divergence scores of the original and synthesized gesture duration distributions for the speaker-independent setting in a leave-one-actor pair-out manner. As in Table 3, using $M_p = 16$ prosody clusters with 3 states per branch gives the optimal KL-divergence value as 7.5270. This value is higher than the value obtained in the speaker-dependent setting, which is expected since the speaker-independent system has higher variability.

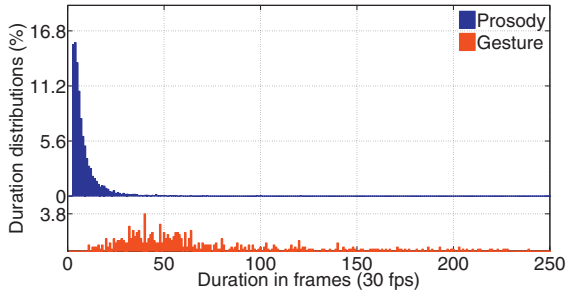
In addition, for the optimal symmetric KL-divergence settings, we compare histograms of gesture phrase and pattern durations with prosodic unit durations (measured in number of frames with 30 fps) in Fig. 5, for speaker-dependent (top) and independent (bottom) settings, respectively. In the figures, the discrepancy in duration distributions for the two modalities is clear. The observation that gesture phrases and patterns have longer durations compared to prosodic units is inline with our choice of using HSMMs for joint modeling of the two modalities.

On the other hand, the KL-divergence value for the baseline synthesis method is calculated as 1.8376 and 7.8365 for speaker-dependent and independent settings, respectively. A higher KL-divergence in this case is expected since, unlike the HSMM-based synthesis, statistical duration information is simply ignored with the baseline synthesis as well as any gesture transition statistics and prosody-gesture correlations.

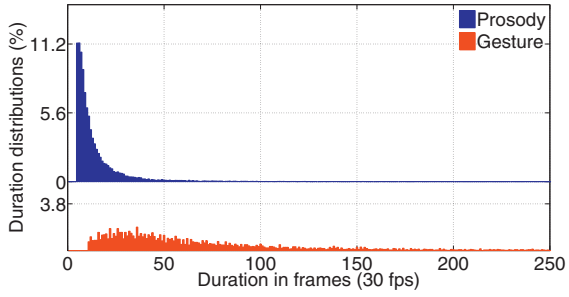
6.4. Penalty score weight parameters

The proposed gesture animation system employs unit-selection to minimize a mixture of three different penalty scores while

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



(a) Speaker-dependent setting



(b) Speaker-independent setting

Fig. 5. Duration histograms of prosodic units (blue) and gesture phrases or patterns (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mapping synthesized gesture sequences into motion sequences. These scores are the duration difference penalty, the joint angle continuity and the speech rhythm similarity as defined in Section 5. Hence prior to the gesture motion smoothing step in Section 5, the penalty score weights α and β in (21) are to be set experimentally. We consider two scoring functions to set α and β values. The first function is based on a windowed cross-lagged correlation (WCLC) score (Boker et al., 2002). The second function evaluates smoothness of the animation through a jerkiness score as defined by Hogan and Sternad (2009).

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

$$\rho^{cca}(\boldsymbol{x},\boldsymbol{y}) = corr(u_x^T\boldsymbol{x}, u_y^T\boldsymbol{y}) \tag{22}$$

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

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

$$e_k^i = \left[\frac{\sum_{j=1}^{2}[\boldsymbol{\theta}_{k+j}^i - \boldsymbol{\theta}_{k-j}^i]j}{2\sum_{j=1}^{2} j^2}\right]^2 \tag{23}$$

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

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

$$\rho_{k,\tau}(\boldsymbol{E}^s,\boldsymbol{E}^o) = \begin{cases} \rho^{cca}(\boldsymbol{E}_k^s, \boldsymbol{E}_{k+\tau}^o), & \text{if } -\tau_{max} < \tau \le 0 \\ \rho^{cca}(\boldsymbol{E}_{k-\tau}^s, \boldsymbol{E}_k^o), & \text{if } \tau_{max} > \tau > 0, \end{cases} \tag{24}$$

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

$$\rho^*(\boldsymbol{E}^s,\boldsymbol{E}^o) = \mathcal{E}\left\{\max_{\tau} \rho_{k,\tau}(\boldsymbol{E}^s,\boldsymbol{E}^o)\right\}, \tag{25}$$

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

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

$$S_{WCLC}(\alpha,\beta) = \frac{\rho^*(\boldsymbol{E}^s,\boldsymbol{E}^o|\alpha,\beta) + \rho^*(\boldsymbol{E}^s,\boldsymbol{F}^p|\alpha,\beta)}{2}, \tag{26}$$

where the maximum correlation coefficients can also be computed for the given penalty score weights α and β in interval $[0,1]$. We target to set α and β values to maximize the WCLC-based score, which will help us to measure and preserve correlations of the synthesized arm motion behaviors to the original motion and speech prosody behaviors in our animations. In our experiments, we use an analysis window size W of 6 s and a maximum lag value τ_{max} of 1 s. The WCLC-based score function is extracted on the validation set for each training set in turn and then the average is used for the speaker-dependent setting, whereas for the speaker-independent setting leave-one-actor pair-out method is employed.

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

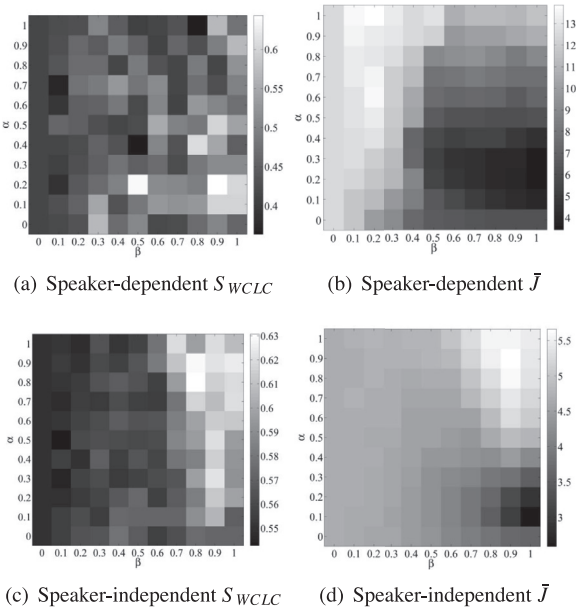
$$\boldsymbol{J}_k^i = \frac{\boldsymbol{\theta}_{k+1}^i - 3\boldsymbol{\theta}_k^i + 3\boldsymbol{\theta}_{k-1}^i - \boldsymbol{\theta}_{k-2}^i}{\Delta^3} \quad \text{for} \quad k = 1,\ldots,K; \quad i = 1,\ldots,4, \tag{27}$$

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

$$J = \sqrt{\sum_{k=1}^{K} \frac{\sum_{i=1}^{4} (\boldsymbol{J}_k^i)^2}{4} \frac{K^3}{v_a^2}} \tag{28}$$

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

In order to set optimal values for the α and β parameters, we evaluate the WCLC-based correlation score function S_{WCLC} and the



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

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

Fig. 6. The WCLC-based score function $S_{WCLC}(\alpha, \beta)$ and the average jerkiness $\bar{J}$ plotted as a function of α and β values.

average jerkiness value $\bar{J}$ as a function of the α and β parameters. Fig. 6 plots these two evaluation metrics in both speaker-dependent and independent settings. The WCLC-based evaluations target higher correlations values, which are plotted in lighter colors. On the other hand, lower average jerkiness values, which are plotted in darker colors, are better for the animation quality. In both speaker-dependent and independent settings, the right-bottom quadrant optimizes both evaluation metrics. That is, higher quality animations based on the correlation score S_{WCLC} and the average jerkiness $\bar{J}$ are expected to be created with the lower values of α and higher values of β in both settings. For the speaker-dependent setting, we set two possible parameter settings at (0.2, 0.5) and (0.2, 0.9) values of the (α, β). Note that these two points have lower jerkiness and higher correlation score values. Furthermore, the first setting with a lower β value is expected to create rhythm emphasized animations, whereas rhythm is less influential with the second setting. Similarly, a single parameter point (α, β) is set for the speaker-independent setting at (0.3, 0.9).

6.5. Subjective evaluations

Gestures can accompany speech in various different ways. Objective evaluations are incapable of qualifying such variabilities, whereas subjective tests can evaluate realism and naturalness of the animation by reflecting human perception. We conduct subjective tests using the animations of speaker-dependent setting, where each participant is shown side-by-side animation pairs and asked to perform A/B comparisons so as to evaluate the naturalness of gesture animations on a scale of $(-2, -1, 0, 1, 2)$. The values in this scale represent A much better, A better, no preference, B better and B much better, respectively. Animation clips in the test are designed to be long enough to allow participants to be able to evaluate transitions between gestures. Each comparison consists of a pair of animation clips of 40 to 80 second duration generated for a given utterance by using two of the following three methods: the HSMM-based synthesis, the baseline synthesis, and the motion-capture synthesis. In the subjective evaluations, the HSMM-based synthesis method employs the unit selection penalty score weight parameters (α, β) with values (0.2, 0.5) and (0.2, 0.9), which are fine-tuned via objective evaluations over the validation set as described in Sections 6.3 and 6.4. The first setup, in which $\beta = 0.5$, emphasizes rhythm penalty score more when compared to the latter one. In this way, the influence of using rhythm is better evaluated in the subjective tests.

Table 5
Results of the subjective A/B pair comparison test for speaker-dependent setting.

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

In the subjective tests, each of the 26 participants (16 of whom are female) is shown 22 pairs of animation clips in random order from a pool of 66 animation clips. The clip pool consists of 12 samples for each of the five pairs presented in Table 5. Each test includes 4 samples for each pairwise comparison, plus a pair of identical clips, which is used to ensure the participants' engagement in the test. The left-right display order of the animation pairs in clips and sequence order of clips are set randomly. All participants are native Turkish speakers of ages in the range 22–40. Table 5 presents the average preference scores and their statistical significance in the subjective evaluations. The preference score for each pair of the five cases is calculated as the average of participants' evaluation scores. A negative average preference score implies that the method on the left side is preferred over the one on the right side and vice versa. A paired two-tailed t-test is used to evaluate the significance of test takers' preferences. We observe in Table 5 that while motion-capture synthesis is significantly favored over the baseline, it is not significantly discriminated from the proposed HSMM-based synthesis when rhythm is emphasized ($\beta = 0.5$) in the animation generation step. Additionally, the HSMM-based synthesis results, emphasizing rhythm, are assessed to be significantly more realistic and natural than the baseline synthesis results with a p-value less than 0.0152. On the other hand, HSMM-based synthesis results are assessed to be less natural when rhythm is less influential ($\beta = 0.9$) in the animation generation step. Samples of animation clips from the subjective A/B comparison tests are available for online demonstration[3].

6.6. Objective evaluations

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

[3] Sample clips for the prosody-driven synthesis of beat gestures are at http://mvgl.ku.edu.tr/prosodybeatrhythm

Table 6
Objective evaluations based on the RMSE between the kinetic energies of the original and synthesized joint angles over the speaker-dependent (TSTDEP) and speaker-independent (TSTIND) datasets.

	HSMM-based	Baseline
TSTDEP	0.25	0.27
TSTIND	0.67	0.86

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

$$RMSE = \sqrt{\frac{\frac{1}{4}\sum_{k=1}^{K} ||\boldsymbol{e}_k^o - \boldsymbol{e}_k^s||^2}{K}}, \tag{29}$$

where K is the total extent of the data, and the $\boldsymbol{e}_k^o$ and $\boldsymbol{e}_k^s$ are respectively 4D kinetic energy vectors of the original and synthesized gesture sequences as defined in (23). The dataset used in the speaker-independent setting is dyadic, where speakers frequently take turns and mostly hold floor for a short time duration. In the RMSE evaluations, we set a test dataset (TSTIND) by segmenting audio recordings into semantically meaningful utterances when the speaker holds the floor. For the speaker-dependent setting, the audio segments in the subjective test, which we refer them as TSTDEP, are used for the objective RMSE evaluations. The total duration of the TSTDEP and TSTIND evaluation sets are 458 and 444 s for speaker-dependent and independent settings, respectively. The RMSE scores of the HSMM-based rhythm-emphasized synthesis and the baseline synthesis over speaker-dependent and independent settings are presented in Table 6. Note that RMSE scores of the HSMM-based synthesis is lower in both settings. While the RMSE scores are lower in the speaker-dependent case, the scores as well as the score difference between the baseline and HSMM-based synthesis are larger for the speaker-independent setting. This is mainly due to the fact that the gesture animation pool for the speaker-dependent set is smaller, whereas the gesture pool in the CreativeIT dataset contains more speaker and gesture variability.

Another objective measure that we use to assess the quality of the resulting animations quantifies how good the proposed model is in transferring speech prosody into beat gestures. This goodness measure can be computed by correlating the speech prosody to the kinetic energy of the joints. Recall that we have defined the CCA based correlation coefficient $\rho_{k,\tau}(\cdot,\cdot)$ at time frame k with time lag τ in (24), that correlates two streams of information. Using this CCA-based correlation coefficient, we define the following three correlation metrics:

$$\gamma^{po} = \rho_{k,0}(\boldsymbol{F}^p, \boldsymbol{E}^o), \tag{30}$$

$$\gamma^{ps} = \rho_{k,0}(\boldsymbol{F}^p, \boldsymbol{E}^s), \tag{31}$$

$$\gamma^{pb} = \rho_{k,0}(\boldsymbol{F}^p, \boldsymbol{E}^b), \tag{32}$$

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

Table 7 presents mean, standard deviation and percent of outliers, i.e., the percentage of correlation values that are greater than values 0.15, 0.2, and 0.25 for the three correlation metrics. Note that the statistics in this table are extracted over all the contents of the CreativeIT and MVGL-MUB datasets. The mean correlation values for the original mocap joint angles, γ^{po}, are not high for any of the datasets, where they are calculated as 0.24 and 0.07 for the CreativeIT and the MVGL-MUB datasets, respectively. Taking into account these reference correlation values for the original mocap joint angles, we can state that the proposed HSMM-based synthesis yields higher mean correlation values than the baseline synthesis system. The main reason of the low mean correlation values is the sparsity of temporal windows, in which speech prosody and kinetic energy of the joint angles are strongly correlated. Although strongly correlated instances are sparse, subjective evaluations suggest that these highly correlated instances have an important role on the perception of naturalness in animations. Hence, we also compute the percentage of the windows with relatively higher correlation values as outliers. The correlation threshold values are set as 0.25, 0.20, and 0.15, which are values close to the mean plus standard deviation for the synthesized joint angles. Note that the percentage of the outliers for the γ^{po} in the CreativeIT dataset are 51.9%, 60.6%, and 67.8% respectively for threshold values 0.25, 0.20, and 0.15. These values are relatively high ratios compared to the percentage of outliers in the MVGL-MUB dataset. We think that this is probably due to the affective theatrical improvisations of the CreativeIT dataset. The proposed HSMM-based synthesis has 12.4%, 19.5%, 34.4% and 4.8%, 12.8%, 18.5% outliers respectively in the CreativeIT and MVGL-MUB datasets. Note also that, in most cases, the outliers of the HSMM-based synthesis are almost twice more than those of the baseline synthesis. This is again a valuable objective evidence for the quality improvement obtained with the proposed HSMM-based animation system.

Table 7
Objective evaluations based on the correlations between speech prosody and kinetic energy for original mocap (γ^{po}), HSMM-based synthesized (γ^{ps}), and baseline synthesized (γ^{pb}) joint angles. The last three rows present percentage of correlation values that are greater than 0.25, 0.20, and 0.15.

	CreativeIT			MVGL-MUB		
	γ^{po}	γ^{ps}	γ^{pb}	γ^{po}	γ^{ps}	γ^{pb}
Mean	0.24	0.10	0.04	0.07	0.08	0.06
std	0.21	0.13	0.13	0.14	0.14	0.11
>0.25 (%)	51.9	12.4	5.7	2.9	4.8	2.5
>0.20 (%)	60.6	19.5	8.3	13.6	12.8	7.5
>0.15 (%)	67.8	34.4	19.34	27.7	18.5	15.2

7. Conclusion

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

The proposed system first segments speech prosody and motion capture data by clustering them into prosodic units and gestures (phrases or patterns), respectively. In the multimodal analysis of gestures and prosodic units, gestures are defined as the states of a Markov chain and prosodic units as the observations of this Markov process. Hence, state transitions model the articulation

of consecutive gestures. Alignment of gestures and prosodic units is captured by the HSMM. The proposed HSMM-based synthesis method effectively associates longer duration gestures to shorter duration prosodic units while maintaining the realistic gesture duration and transition statistics. Hence, our multimodal HSMM-based framework provides an effective solution for modeling the complex relationship between gestures and prosodic units, both at the temporal level and on a multimodal basis.

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

We use objective and subjective evaluation methods to set the system parameters and to assess animation quality over two datasets. Our objective evaluation results on these datasets are coherent for speaker-dependent and independent settings. Subjective evaluations indicate that the proposed system, when rhythm is emphasized in the animation, is significantly better than the baseline synthesis, and statistically similar to the animations created by using the original motion capture data. Furthermore, the reported RMSE between the kinetic energies of the original and the synthesized joints shows that the proposed HSMM-based synthesis framework yields better results than the baseline synthesis. We also present the correlation between speech prosody and the kinetic energy of joint angles as a valuable objective evaluation metric. In subjective evaluations, we observe that strongly correlated instances of prosody and kinetic energy have an important role on the perception of naturalness in animations. We show by experiments that the strongly correlated instances in the proposed HSMM-based synthesis are almost twice more than those in the baseline synthesis.

As future work, we plan to develop additional components to extend our current framework, such as semantic analysis of speech, synthesis of head motion and lip-sync, that would help achieving more realistic animation results. We will also expand our work by considering affective modeling on a multimodal dataset for speech-driven expressive gesture synthesis. We plan to investigate affective relationships between speech and gesture modalities in the domain of conversational interactions.

Acknowledgments

This work was supported by Turk Telekom under Grant Number 11315-02 and by TUBITAK under Grant Number 113E102.

References

Albrecht, I., Haber, J., peter Seidel, H., 2002. Automatic generation of non-verbal facial expressions from speech. In: Proc. Computer Graphics International 2002, pp. 283–293.
Ananthakrishnan, S., Narayanan, S., 2008. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. Audio Speech Lang. Process. IEEE Trans. 16, 216–228.
Barbu, V., Limnios, N., 2008. Semi-Markov Chains and Hidden Semi-Markov Models toward Applications: Their Use in Reliability and DNA Analysis, first ed. Springer Publishing Company, Incorporated.
Boker, S.M., Rotondo, J.L., Xu, M., King, K., 2002. Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series.. Psychol. Methods 7, 338.
Bolt, R.A., 1980. Put-that-there: voice and gesture at the graphics interface. In: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques. ACM, New York, NY, USA, pp. 262–270.
Bos, E., Huls, C., Claassen, W., 1994. EDWARD: full integration of language and action in a multimodal user interface. Int. J. Hum. Comput. Stud. 40, 473–495.
Bozkurt, E., Asta, S., Ozkul, S., Yemez, Y., Erzin, E., 2013. Multimodal analysis of speech prosody and upper body gestures using hidden semi-Markov models. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3652–3656. Vancouver
Bozkurt, E., Erzin, E., Yemez, Y., 2015. Affect-expressive hand gestures synthesis and animation. In: 2015 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6.
Bregler, C., Covell, M., Slaney, M., 1997. Video rewrite: driving visual speech with audio. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, pp. 353–360.
Busso, C., Narayanan, S., 2007. Interrelation between speech and facial gestures in emotional utterances: a single subject study. IEEE Trans. Audio Speech Lang. Process. 15, 2331–2347.
Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Douville, B., Prevost, S., Stone, M., 1994. Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques. ACM, New York, NY, USA, pp. 413–420.
Cassell, J., Vilhjálmsson, H.H., Bickmore, T., 2001. BEAT. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '01. ACM Press, New York, New York, USA, pp. 477–486.
Chen, T., Rao, R.R., 1998. Audio-visual integration in multimodal communication. Proc. IEEE 86, 837–852.
de Cheveigne, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. 111, 1917.
Chuang, E., Bregler, C., 2005. Mood swings: expressive speech animation. ACM Trans. Graph. 24, 331–347.
Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D., 2012. Interpersonal synchrony: a survey of evaluation methods across disciplines. Affect. Comput. IEEE Trans. 3, 349–365.
Ferguson, J., 1980. Variable duration models for speech. In: Symp. Application of Hidden Markov Models to Text and Speech. Princeton, NJ, pp. 143–179.
Fernandez-Baena, A., Montano, R., Antonijoan, M., Roversi, A., Miralles, D., Alias, F., 2013. Gesture synthesis adapted to speech emphasis. Speech Commun. 57, 331–350.
Gibbon, D., Gut, U., 2001. Measuring speech rhythm. In: Eurospeech, pp. 95–98.
Grabe, E., Low, E.L., 2002. Duration variability in speech and the rhythm class hypothesis. Lab. Phonol. 7.
Greenberg, S., 1999. Speaking in shorthand: a syllable-centric perspective for understanding pronunciation variation.. Speech Commun. 29, 159–176.
Greenberg, S., Arai, T., 2004. What are the essential cues for understanding spoken language? IEICE Trans. Inf. Syst. E87, 1059–1070.
Hogan, N., Sternad, D., 2009. Sensitivity of smoothness measures to movement duration, amplitude, and arrests. J. Mot. Behav. 41, 529–534.
Hong, P., Wen, Z., Huang, T.S., 2002. Real-time speech-driven face animation with expressions using neural networks. IEEE Trans. Neural Netw. 13, 916–927.
Iverson, J.M., Thelen, E., 1999. Hand, mouth and brain: the dynamic emergence of speech and gesture. J. Conscious. Stud. 6, 19–40.
Kendon, A., 1980. Gesticulation and speech: two aspects of the process of utterance. In: The Relationship of Verbal and Nonverbal Communication. Mounton Publishers, The Hague, The Netherlands, pp. 207–227.
Kopp, S., Wachsmuth, I., 2004. Synthesizing multimodal utterances for conversational agents. Comput. Animat. Virtual Worlds 15, 39–52.
Ladd, R., 1996. Intonational Phonology. Cambridge University Press.
Levine, S., Krähenbühl, P., Thrun, S., Koltun, V., 2010. Gesture controllers. ACM Trans. Graph. 29, 124:1–124:11.
Li, Y., Shum, H.Y., 2006. Learning dynamic audio-visual mapping with input-output hidden markov models. IEEE Trans. Multimed. 8, 542–549.
Liberman, M., 1975. The Intonational System of English PhD Thesis.
Loehr, D., 2012. Temporal, structural, and pragmatic synchrony between intonation and gesture. Lab. Phonol. 3, 71–89.
Loukina, A., Kochanski, G., Rosner, B., Keane, E., 2011. Rhythm measures and dimensions of durational variation in speech.. J. Acoust. Soc. Am. 129, 3258–3270.
Marioryad, S., Busso, C., 2012. Generating human-like behaviors using joint , speech-driven models for conversational agents. IEEE Trans. Audio Speech Lang. Process. 20, 2329–2340.
Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., Shapiro, A., 2013. Virtual character performance from speech. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. ACM, New York, NY, USA, pp. 25–35.
McNeill, D., 1992. Hand and Mind: What Gestures Reveal about Thought. University Of Chicago Press.
McNeill, D., 2006. Gesture and Thought.
Metallinou, A., Lee, C.C., Busso, C., Carnicke, S., Narayanan, S.S., 2010. The USC creativeIT Database : A Multimodal Database of Theatrical Improvisation. Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC).
Metallinou, A., Yang, Z., Lee, C.C., Busso, C., Carnicke, S., Narayanan, S., 2016. The USC creativeit database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. Lang. Resour. Eval. 50, 497–521.
Neff, M., Kipp, M., Albrecht, I., Seidel, H.P., 2008. Gesture modeling and animation based on a probabilistic recreation of speaker style. ACM Trans. Graph. 27, 5:5–5:24.

Noma, T., Zhao, L., Badler, N., 2000. Design of a virtual human presenter. IEEE Comput. Graph. Appl. 20, 79–85.
Ofli, F., Demir, Y., Yemez, Y., Erzin, E., Tekalp, A.M., Balci, K., Kizoglu, I., Akarun, L., Canton-Ferrer, C., Tilmanne, J., Bozkurt, E., Erdem, A.T., 2008. An audio-driven dancing avatar. J. Multimodal User Interfaces 2, 93–103.
Pelachaud, C., 2005. Multimodal expressive embodied conversational agents. In: Proceedings of the 13th Annual ACM International Conference on Multimedia. ACM, New York, NY, USA, pp. 683–689.
Port, R.F., 2003. Meter and speech. J. Phon. 31, 599–611.
Ramus, F., Nespor, M., Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal.. Cognition 265–292.
Reithinger, N., Gebhard, P., Markus, L., Ndiaye, A., Pfleger, N., Klesen, M., 2006. Virtual human dialogic and affective interaction with virtual characters. In: Proceedings of the International Conference on Multimodal Interfaces (ICMI06), pp. 51–58.
Ringeval, F., Chetouani, M., Schuller, B., 2012. Novel metrics of speech rhythm for the assessment of emotion. In: INTERSPEECH: Annual Conference of the International Speech Communication Association, pp. 2763–2766.
Sargin, M.E., Yemez, Y., Erzin, E., Tekalp, A.M., 2008. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. IEEE Trans. Pattern Anal. Mach. Intell. 30, 1330–1345.
Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., Bregler, C., 2004. Speaking with hands: creating animated conversational characters from recordings of human performance. In: ACM SIGGRAPH 2004 Papers. ACM, New York, NY, USA, pp. 506–513.
Tilsen, S., Johnson, K., 2008. Low-frequency fourier analysis of speech rhythm.. J. Acoust. Soc. Am. 124, 34–39.
Tuite, K., 1993. The production of gesture. Semiotica 93, 83–106.
Valbonesi, L., Ansari, R., Mcneill, D., Quek, F., Duncan, S., McCullough, K.E., Bryll, R., 2002. Multimodal signal analysis of prosody and hand motion: temporal correlation of speech and gestures,. In: Proc. Eur. Signal Process. Conf. (EUSIPCO 02), pp. 75–78.
Wagner, P., Malisz, Z., Kopp, S., 2014. Gesture and speech in interaction: an overview. Speech Commun. 57, 209–232.
Xue, J., Borgstrom, J., Jiang, J., Bernstein, L.E., Alwan, A., 2006. Acoustically-driven talking face synthesis using dynamic bayesian networks. In: in Proc. of the Int. Conf. on Multimedia and Expo 2006 (ICME), pp. 1165–1168.
Yang, Z., Metallinou, A., Erzin, E., Narayanan, S., 2014. Analysis of interaction attitudes using data-driven hand gesture phrases. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 699–703.
Yang, Z., Narayanan, S.S., 2016. Modeling dynamics of expressive body gestures in dyadic interactions. IEEE Trans. Affect. Comput..
Yu, S.Z., 2010. Hidden semi-markov models. Artif. Intell. 174, 215–243.