MULTIMODAL ANALYSIS OF SPEECH PROSODY AND UPPER BODY GESTURES USING HIDDEN SEMI-MARKOV MODELS

Elif Bozkurt, Shahriar Asta, Serkan Özkul, Yücel Yemez, and Engin Erzin

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

ABSTRACT

Gesticulation is an essential component of face-to-face communication, and it contributes significantly to the natural and affective perception of human-to-human communication. In this work we investigate a new multimodal analysis framework to model relationships between intonational and gesture phrases using the hidden semi-Markov models (HSMMs). The HSMM framework effectively associates longer duration gesture phrases to shorter duration prosody clusters, while maintaining realistic gesture phrase duration statistics. We evaluate the multimodal analysis framework by generating speech prosody driven gesture animation, and employing both subjective and objective metrics.

Index Terms— Prosody analysis, gesture segmentation, gesture animation

1. INTRODUCTION

Gesture and speech co-exist in time with a tight synchrony, and they are planned and shaped by the emotional state and produced together. In one of the pioneering studies on gesture and speech relationship, Kendon [1] proposed a widely accepted hierarchical model for gesture. In this model, the core gestural element is defined as gesture phase and combinations of gesture phrases form gesture units. In this hierarchical model, semantic expressiveness of hierarchy levels increases as we move further away from the core. On the other hand there are four widely referred types of gestures, which were proposed by McNeill [2]: iconics, metaphorics, deictics and beats. Iconic gestures illustrate images of an object or action, metaphoric gestures represent abstract ideas, deictic gestures relatively locate entities in physical space, and beat gestures are simple repetitive movements to emphasize speech.

Synchrony between gestural and phonological structures has been studied by various researchers. Kendon stated the synchrony between strokes and stressed syllables in [1], later McNeill [2] proposed the widely accepted phonological synchrony rule stating that the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech. Valbonesi et al. [3] investigates the nature of temporal relationship between speech and gestures. In a recent study, Loehr [4] presents a detailed investigation of temporal and structural synchrony between intonation and gesture. His findings verify the alignment of the pitch accents with the gestural strokes; furthermore he presents evidences of the synchrony between gesture phrases and intermediate intonational phrases.

This work was supported by Turk Telekom under Grant Number 11315-02.

Existing gesture synthesis methods can be classified into two groups: rule-based and data-driven approaches. The Embodied Conversational Agents (ECAs) of Cassell [5] is a pioneering rule-based full-body gesture synthesis system, which performs animations over a pre-defined gesture tree. The VirtualHuman research project [6] and the probabilistic approach of Neff et al. [7] are examples of audiovisual data-driven approaches for full-body gesture animation. The VirtualHuman project aims to develop interactive virtual characters with a personality profile, whereas Neff et al. [7] present a probabilistic approach to produce full-body gesture animation for a given input text in the style of a particular performer. Recently, Levine et al. [8] have introduced gesture controllers, which avails a modular methodology to drive beat-like gestures with live speech, using customized gesture repertoires. Gesture controllers infer hidden states from speech, and select the optimal gesture kinematics based on the inferred states. From a hierarchical perspective, the work of Levine et al. is mainly concentrated on the gesture phase level. Although motion capture systems are becoming widely available, there is limited work in the literature on processing of 3D motion data rather than using it for 3D reconstruction. Heloir et al. [9] provide technical setup, scenarios and challenges in building a motion capture database for virtual human animation. Similarly, Busso et al. [10] present the interactive emotional dyadic motion capture (IEMOCAP) database, which is a multimodal and multispeaker database of improvised dyadic interactions.

Early works on prosody driven gesture synthesis mostly concentrate on facial expression and head motion. Face animation with expressions using neural networks [11], and multimodal communication using affine transformations [12] are among the works on facial expression synthesis. An approach to synthesize emotional head motion sequences driven by prosodic features is presented in [13] by building hidden Markov models for emotion categories to model temporal dynamics of emotional head motion sequences. 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 has been studied in [14]. A recent paper [15] focuses on building a speech-driven facial animation framework to generate natural head and eyebrow motions using dynamic Bayesian networks (DBNs).

In this study, we employ hidden semi-Markov model (HSMM) for multimodal analysis of gestures and prosody. The HSMM was first introduced by Ferguson [16] as the 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 state duration is allowed to follow a probabilistic distribution. We employ the HSMM framework to realistically model the gesture phrase durations for the problem of generating body gesture sequences from prosody observations. To our best knowledge, this is the first time that HSMM is considered for the task of synthesizing

body gesture phrases from prosody observations. Moreover, from a hierarchical perspective, our work is mainly concentrated on gesture phrases which is semantically more expressive than gesture phases studied in the work of Levine et al. [8]. Hence our framework provides a more personalized synthesis. Our experiments show that the animated gestures generated by our method are plausible and look natural.

2. MULTIMODAL ANALYSIS AND SYNTHESIS OF GESTURE PHRASES

The general framework for our automatic hand gesture synthesis system is given in Fig. 1. The framework consists of three main functional blocks for analysis, synthesis and animation. The analysis functional block consists of unimodal analysis of speech and body motion to extract intonational and gesture phrases, as well as multimodal analysis to learn dependencies between intonational and gesture phrases by utilizing an HSMM. In the synthesis functional block, we generate a gesture sequence along with gesture durations given a speech input. Finally, in the animation functional block, the synthesized gesture sequence is mapped into a body motion sequence so as to obtain a natural looking animation.

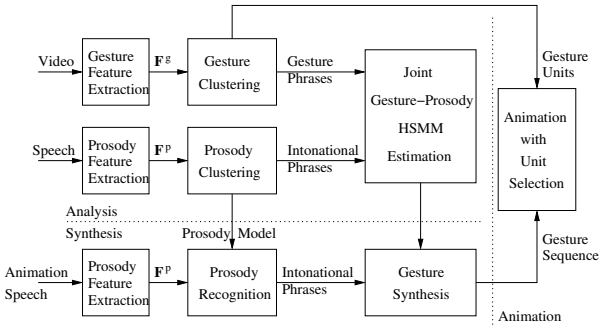

Fig. 1. The block diagram of the general framework for the automatic upper body gesture synthesis system.

2.1. Prosody Clustering

Prosodic voice characteristics at the acoustic level, including intonation, rhythm and intensity patterns, carry important temporal and structural synchrony with gesture phrases [4]. Acoustic features such as pitch and speech intensity can be used to model underlying intonational phrases of speech. We choose to include normalized speech intensity, normalized pitch and confidence to pitch, which is represented with pitch gain, in the prosody feature vector. We estimate the prosody feature vector for each speech frame of 25 msec duration centered on a 50 msec analysis window. Speech intensity, I_k, is extracted as the logarithm of the signal energy in the kth analysis window. The normalized speech intensity, $\bar{I}_k$, is then extracted with mean and variance normalization. Pitch, τ_k, is computed using the auto-correlation method [17]. Confidence to pitch, r_k, is set to the normalized auto-correlation value at pitch lag τ_k. Since pitch values differ for each speaker and the system is desired to be speaker-independent, speaker normalization is applied. For each speech segment, we compute the mean pitch value over the pitch values with pitch confidence higher than 0.4. Then the mean pitch value is removed from the pitch values, which are computed for each segment, to obtain the normalized pitch $\bar{\tau}_k$. Then normalized intensity, normalized pitch, pitch confidence and the first derivative of these three parameters are used to define the prosody feature vector,

$$\boldsymbol{f}_k^p = [\bar{I}_k,\, \bar{\tau}_k,\, r_k,\, \Delta\bar{I}_k,\, \Delta\bar{\tau}_k,\, \Delta r_k], \tag{1}$$

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

We extract intonational phrases through unsupervised temporal clustering. For the purpose of temporal clustering we employ the parallel branch HMM structure described in [14]. The prosody feature stream $\boldsymbol{F}^p = \{\boldsymbol{f}_1^p, \boldsymbol{f}_2^p, ..., \boldsymbol{f}_T^p\}$ is used to train a parallel branch HMM structure, Λ^p, which clusters the prosody feature stream and captures recurrent intonational phrases. The HMM structure Λ^p is composed of M_p parallel left-to-right HMMs, $\{\lambda_1^p, \lambda_2^p, ..., \lambda_{M_p}^p\}$, where each λ_m^p is composed of N_p states. The unsupervised clustering process defines temporal intonational segments, where each segment label ℓ_l^p for the lth segment is assigned to one of the M_p available segment classes $\{p_1, p_2, ..., p_{M_p}\}$. The frame level label sequence for an intonational phrase sequence is then defined by

$$\xi_t = \ell_l^p \quad \text{for } t = t_l, t_l + 1, \ldots, t_{l+1} - 1, \tag{2}$$

where ξ_t is the intonation label of the tth speech frame, and t_l is the first frame of the lth intonational segment.

2.2. Gesture Clustering

In this study we model upper-body gestures, specifically hand gestures, at gesture phrase level to emphasize speech intonation. For analysis of gestures, we employ joint angles as the gesture features from four body parts: left arm, left forearm, right arm, and right forearm. We define the gesture feature vector for the ith joint at frame k, $\boldsymbol{f}_k^{Ji}$, to include the joint angles from the ith body part and their first order derivatives,

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

where θ_k^i, ϕ_k^i and ψ_k^i are the Euler angles respectively in x, y and z directions, representing the posture of the ith joint at frame k, and $\Delta\theta_k^i$, $\Delta\phi_k^i$, $\Delta\psi_k^i$ denote their respective first order derivatives. The resulting gesture feature for the four joints at time frame k is defined as

$$\boldsymbol{f}_k^g = [\boldsymbol{f}_k^{J1}, ..., \boldsymbol{f}_k^{J4}]. \tag{4}$$

We perform semi-supervised clustering by using the parallel branch HMM structure, Λ^g, over the gesture feature stream $\boldsymbol{F}^g = \{\boldsymbol{f}_1^g, \boldsymbol{f}_2^g, ..., \boldsymbol{f}_T^g\}$ to extract recurrent gesture phrases. The HMM structure Λ^g is initially set to have two parallel branch HMMs, $\{\lambda_1^g, \lambda_2^g\}$, where each λ_m^g is composed of N_g states. The number of branches increased iteratively to M_g by using both human supervision and the parallel branch HMM temporal clustering in a semi-supervised manner. Eventually the semi-supervised clustering process defines gesture phrase segments, where each segment label ℓ_l^g for the lth segment is assigned to one of the M_g available gesture phrase classes $\{g_1, g_2, ..., g_{M_g}\}$.

2.3. HSMM for Intonational and Gesture Phrases

In a natural speaking style, beat gestures are articulated in synchrony with speech prosody to emphasize the underlying speech [2–4]. In this section we construct a multimodal analysis framework to model the relationship between beat gestures and speech prosody. In general, intonational phrases last much shorter than gesture phrases in duration, and a gesture stroke precedes or ends at, but does not follow, phonological peak syllable of speech as McNeill stated in [2].

A gesture phrase sequence, when accompanied by a sequence of intonation phrases, forms a random process which can be seen as

a Markov process. One useful mathematical model for the multimodal analysis of gesture and intonation phrases can be constructed by taking gesture phrases as the states of a Markov chain and intonation phrases as the observations of this Markov process. Hence state transitions correspond articulation of consecutive gesture phrases, and gesture phrases are expected to localize in time with respect to the McNeill's phonological rule by observing intonation phrases. Since the relationship between gesture and intonation phrases is not strong however, if a gesture phrase sequence were to be synthesized using this model, given intonation phrase observations, gesture phrases would have a shortfall in modeling duration in time. Another useful mathematical model to overcome this shortfall is to introduce a state duration model so that one can better control gesture phrase durations in the synthesis process. Combination of these two useful mathematical models yields the hidden semi-Markov model (HSMM) framework [18], that we use for multimodal analysis of gesture and intonation phrases. Fig. 2 shows how such an HSMM structure works.

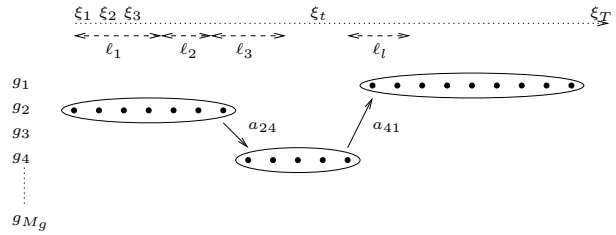

Fig. 2. *In a Hidden Semi-Markov Process, each state has a duration and emits a number of observations.*

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

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

$$\boldsymbol{A} : \{a_{ij} = P(\ell_l^g = g_j | \ell_{l-1}^g = g_i)\} \quad i, j = 1, ..., M_g, \quad (5)$$

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

$$\boldsymbol{B} : \{b_i(p_k) = P(p_k | \ell_l^g = g_i)\} \quad k = 1, ..., M_p,\ i = 1, ..., M_g, \quad (6)$$

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

$$\boldsymbol{D} : \{d_i(k)\} \quad i = 1, ..., M_g,\ k = 1, ..., \frac{D_{max}}{\delta}, \quad (7)$$

where $d_i(k)$ is the probability of gesture g_i lasting $k\delta$ sec, D_{max} is the maximum duration among all gestures, and δ is the histogram bin size for the underlying probability mass function. We take the maximum duration as $D_{max} = 10$ sec, and the histogram bin size as the speech frame duration, $\delta = 25$ msec. The initial state probability vector $\boldsymbol{\Pi}$ is defined by entries π_i representing the probability of starting with gesture g_i as the first gesture phrase,

$$\boldsymbol{\Pi} : \{\pi_i = P(\ell_1^g = g_i)\} \quad i = 1, ..., M_g. \quad (8)$$

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

$$\pi_i = P(\ell_1^g = g_i) \widehat{=} \frac{C(1, i, j)}{\sum_{j'} C(1, i, j')}, \quad (9)$$

$$a_{ij} = P(\ell_l^g = g_j | \ell_{l-1}^g = g_i) \widehat{=} \frac{\sum_l C(l, i, j)}{\sum_l \sum_{j'} C(l, i, j')}, \quad (10)$$

$$b_i(p_k) = P(p_k | \ell_l^g = g_i) \widehat{=} \frac{O(i, k)}{\sum_{k'} O(i, k')}, \quad (11)$$

$$d_i(k) \widehat{=} \frac{H(i, k\delta \le \tau < (k+1)\delta)}{\sum_{k'} H(i, k'\delta \le \tau < (k'+1)\delta)}, \quad (12)$$

where $C(l, i, j)$ is the number of times g_i is the lth gesture and g_j is the $(l+1)$st gesture, $O(i, k)$ is the number of frame count of intonation phrase p_k at gesture g_i, and $H(i, k\delta \le \tau < (k+1)\delta)$ is the number of occurrences of gesture g_i with duration τ in $[k\delta,\ (k+1)\delta)$ interval.

2.4. 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 intonational phrase labels, $\{\xi_1, \xi_2, \ldots, \xi_T\}$. Note that the decoded optimal state sequence delivers synthesized sequence of gesture phrases and their durations, where the HSMM framework secures to have realistic gesture phrase durations. 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 is emitted at a single state. This requires to define a forward likelihood function, which incorporates the state duration model,

$$\psi_t(j) = \max_{\tau} \max_{i} \{\psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \prod_{k=t-\tau+1}^{t} b_j(\xi_k))\}, \quad (13)$$

where $\psi_t(j)$ is the accumulated logarithmic likelihood at time frame t in state g_j after observing intonational phrase labels $\{\xi_1, \xi_2, \ldots, \xi_t\}$. Based on the forward likelihood function $\psi_t(j)$, we define the following modified Viterbi decoding algorithm to extract the optimal state sequence, that is the optimal gesture phrase sequence $\hat{\boldsymbol{\ell}}^g = \{\hat{\ell}_1^g, \ldots, \hat{\ell}_L^g\}$, and the associated gesture phrase durations $\boldsymbol{\kappa} = \{\kappa_1, \ldots, \kappa_L\}$.

The modified Viterbi decoding algorithm for gesture synthesis:

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

$$\Psi_{t\tau}^{ij} = \psi_{t-\tau}(i) + \log(a_{ij} d_j(\tau) \textstyle\prod_{k=t-\tau+1}^{t} b_j(\xi_k))$$
$$\psi_t(j) = \max_{\tau \in [1, T']} \max_{i \in [1, M_g]} \{\Psi_{t\tau}^{ij}\}$$
$$\varphi_t(j) = \arg\max_{i \in [1, M_g]} \max_{\tau \in [1, T']} \{\Psi_{t\tau}^{ij}\}$$
$$\nu_t(j) = \arg\max_{\tau \in [1, T']} \max_{i \in [1, M_g]} \{\Psi_{t\tau}^{ij}\}$$

iii. Backtrace the optimal gesture phrase sequence
$\hat{\ell}_L^g = \arg\max_j \psi_T(j)$
$\kappa_L = \nu_T(\hat{\ell}_L^g); \quad l = L - 1; \quad t = T$
While $t > 0$

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

2.5. Gesture Animation

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

The first task is to generate a synthesized sequence of joint angles, $\hat{\boldsymbol{F}}^g$, given the synthesized gesture phrase $\hat{\boldsymbol{\ell}}^g$ and duration $\boldsymbol{\kappa}$ sequences. This task is performed using unit selection over the gesture phrases which are extracted during the gesture analysis. To select gesture units with low concatenation distortion and low duration difference, we apply a dynamic programing algorithm that minimizes a joint distortion function of the joint angle differences at transitions and the gesture duration differences. The selected gesture units are interpolated to fit the synthesized duration. The next task is to smooth joint angle discontinuities over a temporal window at gesture unit boundaries. This is achieved by applying an exponential smoothing function on the synthesized gesture motion sequence. Finally, the smoothed gesture motion sequence is animated using the *MotionBuilder 3D Character Animation Software* [19].

3. EXPERIMENTAL RESULTS

We use the multimodal upper-body (MVGL-MUB) corpus for modeling the relationship between intonational and gesture phrases with HSMMs [20]. In our experimental evaluations, we only consider a single subject from the MVGL-MUB database, who has 5 recordings with a total duration of 20 minutes. Since the variety of upper-body expressions depends on the particular speaker, the upper body gesture clusters for a given speaker are determined experimentally by using the semi-supervised training set up described in Section 2.2. The number of distinct upper-body expressions is identified as $M_g = 8$ for the selected subject. The multimodal analysis and synthesis steps are carried out by leave-one-recording-out. We evaluate the proposed methods with objective and subjective tests.

3.1. Objective Evaluation

We consider similarity of the original and the synthesized gesture duration statistics as a possible objective metric to evaluate our HSMM based gesture synthesis framework. We use the symmetric Kullbeck-Leibler (KL) divergence to measure the similarity of the synthesized gesture sequence with the original one, where a smaller divergence value indicates that the duration distributions of the two sequences are consistent. Unsupervised clustering of the prosody is performed using the parallel-branch HMM structures as described in Section 2.1, with branch numbers ranging from 8 to 16 and state numbers per branch ranging from 3 to 5. The resulting HSMM for intonational and gesture phrases is evaluated with the KL divergence metric for different parameter settings of the unsupervised prosody clustering process. Table 1 presents the symmetric KL divergence scores for various parameter settings. The minimum KL divergence is observed with the $M_p = 8$ branches and $N_p = 4$ states. We use this setting in the subjective evaluations of our gesture synthesis system.

Table 1. The symmetric KL divergence of the original and synthesized gesture duration distributions for various prosody clustering settings.

N_p \ M_p	**8**	**10**	**12**	**16**
3	1.171	1.121	0.956	0.761
4	**0.649**	1.293	1.093	1.147
5	1.272	1.221	1.021	1.574

3.2. Subjective Evaluations

Gesture sequences accompanying speech may be expressed in various different ways. Objective evaluations are incapable of qualifying such variabilities whereas, subjective tests would evaluate realism and naturalness of the animation better by reflecting human perception. We conducted subjective tests where each participant was shown side-by-side pairwise A/B comparisons and was asked to evaluate the naturalness of the upper body gesture animations on a scale of (-2, -1, 0, 1, 2), where the scale corresponds to (A much better, A better, no preference, B better, B much better). Each comparison consisted of a pair of animation clips in 40-120 second duration for the same utterance generated by two of the three methods: the HSMM-based synthesis, random synthesis, and motion-capture synthesis. The random synthesis has been created from a random gesture sequence, where in the animation phase unit selection is performed based on the minimum concatenation distortion. The motion-capture synthesis on the other hand uses the true motion capture data in the animations.

In the subjective tests, each of the 9 participants was shown 20 pairs of animation clips in random order from a pool of 72 animation clips. Table 2 presents the average preference scores and statistical significance of the preferences in the subjective evaluations. While the motion-capture synthesis is favored significantly over the random synthesis, it is not discriminated significantly from the HSMM-based synthesis. The proposed HSMM-based synthesis is assessed to be significantly more realistic and natural than the random synthesis with a p-value less than 0.004. Animation clip samples from the subjective A/B comparison test are available for online demonstration [21].

Table 2. Results of the subjective A/B pair comparison test.

A/B Pair	Average Preference	p-value $<$
Motion-capture / Random	-1.130	0.0006
Motion-capture / HSMM-based	0.019	0.9396
HSMM-based / Random	-0.796	0.004

4. CONCLUSIONS

We have presented a new multimodal analysis framework for modeling the relationship between intonational and gesture phrases using the hidden semi-Markov models (HSMMs). The proposed method effectively associates longer duration gesture phrases to shorter duration prosody clusters, while maintaining realistic gesture phrase duration statistics that leads to animations perceived as realistic by subjective evaluations. Future work would consider further investigation of objective metrics for the evaluation of the multimodal analysis framework of prosody and gesture streams, as well as investigation of affective interactions between prosody and gesture.

5. REFERENCES

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

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

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

[4] Dan Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology*, vol. 3, no. 1, 2012.

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

[6] Norbert Reithinger, Patrick Gebhard, L Markus, Alassane Ndiaye, Norbert Pfleger, Martin Klesen, and H Multimedia Information Systems, "VirtualHuman Dialogic and Affective Interaction with Virtual Characters," in *Proceedings of the International Conference on Multimodal Interfaces (ICMI06)*, 2006, pp. 51–58.

[7] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel, "Gesture Modeling and Animation Based on a Probabilistic Recreation of Speaker Style," *ACM Transactions on Graphics*, vol. 27, no. 1, pp. 5:5—-5:24, Mar. 2008.

[8] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun, "Gesture controllers," *ACM Transactions on Graphics*, vol. 29, no. 4, July 2010.

[9] Alexis Heloir, Michael Neff, and Michael Kipp, "Exploiting Motion Capture for Virtual Human Animation: Data Collection and Annotation Visualization," in *Proc. of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Mul- timodality*, 2010.

[10] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Nov. 2008.

[11] Pengyu Hong, Zhen Wen, and T S Huang, "Real-time speech-driven face animation with expressions using neural networks.," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 916–927, Jan. 2002.

[12] Tsuhan Chen and Ram R. Rao, "Audio-Visual Integration in Multimodal Communication," *Proc. of IEEE*, vol. 86, no. 5, pp. 837—–852, 1998.

[13] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan, "Rigid Head Motion in Expressive Speech Animation : Analysis and Synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1075–1086, 2007.

[14] Mehmet Emre Sargin, Yucel Yemez, Engin Erzin, and Ahmet Murat Tekalp, "Analysis of Head Gesture and Prosody Patterns for Prosody-Driven Head-Gesture Animation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1330–1345, Aug. 2008.

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

[16] J.D. Ferguson, "Variable duration models for speech," in *Symp. Application of Hidden Markov Models to Text and Speech*, Princeton, NJ, 1980, pp. 143—-179.

[17] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, New York, NY, USA, 1993.

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

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

[20] S. Ozkul, E. Bozkurt, S. Asta, Y. Yemez, and E. Erzin, "Multimodal analysis of upper-body gestures, facial expressions and speech," in *Proc. of 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3)*, 2012.

[21] E. Bozkurt, S. Asta, S. Ozkul, Y. Yemez, and E. Erzin, "Sample clips for the speech-driven upper body animation. http://mvgl.ku.edu.tr/demos," Dec. 2012.