# Upper Body Animation Synthesis for a Laughing Character

Yu Ding[1], Jing Huang[1], Nesrine Fourati[1], Thierry Artières[2], and Catherine Pelachaud[1,3]

[1] Institut Mines-TELECOM, TELECOM ParisTech, Paris, France
[2]Université Pierre et Marie Curie (LIP6), Paris, France
[3] CNRS - LTCI UMR 5141, Paris, France

**Abstract.** Laughter is an important social signal in human communication. This paper proposes a statistical framework for generating laughter upper body animations. These animations are driven by two types of input signals, namely the acoustic segmentation of laughter as pseudo-phoneme sequence and acoustic features. During the training step, our statistical framework learns the relationship between the laughter human motion and the input signals. During the synthesis step, our trained framework synthesizes automatically natural head and torso animations from the input signals. Objective and subjective evaluations were conducted to validate this framework. The results show that our proposed framework is capable of generating laughing upper body movements.

**Keywords:** virtual character, head motion synthesis, torso motion synthesis, Hidden Markov model, laughter, character animation

## 1 Introduction

Embodied conversational agents, ECAs, are autonomous software characters with a human-like appearance and communicative capabilities. Several models of ECAs have been proposed [1],[2] but very few works focus on animation synthesis for laughing.

Laughter is frequently used in human communication. Laughter is strongly linked to positive emotions and even more to cheerful mood [3]. Humans laugh at humorous stimuli or to mark their pleasure when receiving praised statements[4]; they also laugh to mask embarrassment[5] or to be cynical. Laughter can act also as social indicator of in-group belonging; it can work as speech regulator during conversation; it can also be used to elicit laughter in interlocutors as it is very contagious [4].

Laughter morphology involves facial expressions, body movements and vocalizations [6]. For hilarious laughter [5], muscular activities include mainly the zygomatic major, mouth opening and jaw movement. Eyebrows may be raised or even frown in very intense laughter [6]. Saccadic movements affect the whole body. Torso may bend back and forth and shoulder may shake. Changes in respiration patterns are also prominent. Inhalation and exhalation phases are very

noticeable. All these movements are done very rhythmically and they are also highly correlated. Indeed they arise from the same physiological processes [6].

Darwin reported "During excessive laughter the whole body is often thrown backward and shakes, or is almost convulsed" [7]. Ruch and Ekman [6] described laughter movements as "rhythmic patterns", "rock violently sideways, or more often back and forth", "nervous tremor ... over the body", "twitch or tremble convulsively". Melo et al. [8] built a virtual character which "convulses the chest with each chuckle". It means that periodic motions of head and body are important and well-known features during laughter. The periodicity of body motion was used to distinguish between laughters in[9]. Ruch and Ekman [6] reported that rhythmical patterns during laughter were usually characterized by frequency around 5 $Hz$. Mancini et al. [9] observed 8 videos, which show people laughing while watching funny images. Laughing persons produce rhythmic body movements with frequencies in the range of [1.27$Hz$ 3.66$Hz$]. Using such findings of laughing behaviours, our main objective is to build an animation synthesis model of upper body movement during laughter.

The aim of this paper is to report head and torso animation synthesis for a hilarious laughing character. To achieve our aim, a data-driven animation model is proposed to first learn, from a collected laughter corpus, the relationship linking the input signals and human motions; then, this trained statistical model can be used as generator of laughter head and torso animations.

## 2 Dataset

We created a multimodal dataset of laughter. Three human subjects participated in the collection of laughter data. During recording session, the subjects watched funny movies for about 25-40 minutes. Since laughter motion occurs mainly during social interactions [10], [11], we propose an interactive setup where two subjects watch funny videos together. Only the movement of one person was gathered. Three-dimensional torso and head movements and audio signal are recorded by a motion capture system at 125 frames per second ($fps$) and a microphone at 44100 $Hz$, which were synchronized using the approach described in [12]. During data processing, all laughter episodes were manually extracted. In total, we obtain 259 laughter episodes; each one lasts from 1 to 37 seconds. Then phonetic transcription is extracted by Urbain et al [13], in which 12 laughter pseudo-phonemes are defined in reference to speech phoneme. Laughter involves very specific sounds that cannot be translated as speech phonemes. For simplicity, laughter pseudo-phoneme is called phoneme in this paper. Phonetic transcription contains phoneme (text signal) and its duration. An intensity value is also provided for each phoneme. Notice that, if one phoneme occurs successively several times, the sum of phoneme lasting time is viewed as one phoneme duration. Finally, PRAAT [13] is used to extract acoustic signals at 125 fps including pitch and energy.

## 3 Head and Torso Motion Synthesis

We propose a system to produce head and torso motions featured by 3D rotation angles (hence a 6 dimensional signal) from a number of input signals which are: the pseudo-phoneme sequence together with their duration and their intensity (low or high), and audio features (we use pitch and energy).

**Animation Generator** To do so we consider building one model of generating animation for every $(phoneme, intensity)$ pair, we name a model for each pair an Animation Generator (AG). Since *silence* phoneme is not labelled by intensity and the other 11 phonemes are labelled by low or high intensities, we build 23 AGs. Each of these 23 AGs is learned independently from the training corpus of corresponding (input, output) pairs where the input stands for all the above input features and the ouput stands for a sequence of animation motion for the 6 data streams we want to learn to synthesize (the 6 dimensions of the animation signal). Our modeling framework is based on three ideas that we detail now.

- Modeling one dimensional shaking-like movement with what we call *Loop HMM*.
- Introducing speech influence on motion through transition probability parameterization, yielding what we call Transition Parameterized Loop HMM (TPLHMM).
- Taking into account the dependencies between the 6 dimensions of the animation movements with coupled HMMs, yielding Coupled TPLHMM (CTPLHMM).

**Modelling Shaking Motion with a Loop HMM.** We propose a specific HMM that we call a Loop HMM (LHMM) to model (and synthesize) a one-dimensional shaking-like (and/or trembling) signal (Figure 1). It has an approximate left-to-right chain structure where transitions are allowed from one state to itself, to the previous and to the next state. Yet it is intended that the transition probability from one state to the previous state be very small so that a likely state sequence will depict the entire chain form the first state to the last state with some *hesitation* corresponding to few back transitions.

The HMM is designed so that an observation sequence produced along such a state sequence will correspond to one shake pattern (with some trembling effect coming from back transitions). There is one Gaussian distribution associated to each state of the chain, which are set by hand rather than learned, as follows. We first divide the range of the signal value in $N$ intervals and define $N$ Gaussian Probability Density Function (PDF), one for each interval. The mean of the Gaussian distribution for a given interval is the mean of this interval and its variance is defined according to the width of this interval. Then we assign one of the PDF to every state of the left-right HMM so that going from the first state to the last state corresponds to a trajectory of a shaking movement. For instance in Figure 1, the first state has PDF $p_2$ which outputs intermediate

values in the observation space, the second state has PDF $p_3$ which outputs higher values, it is followed by a state with PDF $p_2$, then by a state with PDF $p_1$ which outputs lower values. If a signal is produced by this HMM along a state sequence that goes from the first (left) to the last (right) state it will correspond to a shaking-like motion.

Finally, there is a loop from the last state to the first state to enable the repetition of such a shaking and trembling pattern. Figure 1 (top) shows one example of a synthesized motion stream by a LHMM. As can be seen, the animation inferred by a LHMM shows the repetition of a pattern.

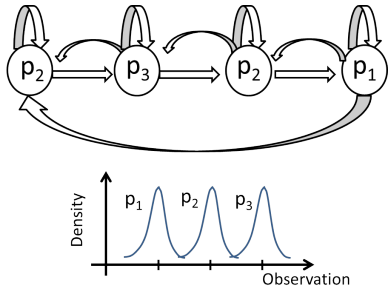


**Fig. 1.** A Loop HMM whose manual design allows us to model shaking and trembling one dimensional movements.

**Taking into account the dependency with speech** Some evidence about the motion pattern may be gained from taking into account the dependencies between audio signal and motion during laughter [14]. Audio signal (we use pitch and energy) may then be used to shape the synthesized animation stream. In addition to introducing some variability in the inferred animation such a strategy makes animation look more realistic because of an increased consistency with the audio signal.

To exploit such a correlation between speech and movements we developed an extension of our LHMM, whose state transition probabilities depend on acoustic features. We call these models Transition Parameterized Loop HMM (TPLHMM). They may be used to model and synthesize one dimensional shaking movements that are linked in some way with speech. We implemented this idea in a similar way as proposed previously by [15] to take into account the dependency of observations sequences in the HMM framework to what was called contextual or external variables. The difference lies in that, while in [15] contextual variables were used to alter Gaussian PDF means, we use the speech features to alter the transition probabilities in our TPLHMMM. We consider that transition probablities from state $i$ to state $j$ at time $t$ are defined according to:

$$a_{i,j}(t) = \frac{e^{W_{i,j}\theta_t}}{\sum_{j'} e^{W_{i,j'}\theta_t}} \quad (1)$$

where $\theta_t$ and $W$' are $c$-dimensional vectors. $\theta_t$ stands for contextual features at time $t$ (e.g. pitch and energy) and $W$'s are parameter matrices (to be learned from data) associated to each possible transition. The parameters of a TPLHMM (the $W$'s) are learned via likelihood maximization with a Generalized EM algorithm. To ease learning it is initialized with a trained LHMM (HMM).

**Isolated and joint modeling of the 6 dimensional animation signal** A first possibility to model and synthesize the 6 dimensional animation signal is to assume the 6 signals are independent from each others and to learn independently one LHMM or one TPLHMM per dimension. Alternatively one could consider jointly modeling head and torso motions. For example, Ruch and Ekman [6] reported that the backward tilt of the head facilitates the forced exhalations, while exhalation directly influences torso motion as being done in DiLorenzo et al. [14]. Therefore, the relationship between head and torso motions should be modeled jointly for, to be tested, augmenting naturalness of synthesized animations. In our work, we used Coupled HMMs (CHMM) [16] which have been designed to model multiple interdependent streams of observations. In a CHMM with $K$ streams of observations, there is one HMM per stream and transition probabilities account for transiting from $K$-tuple of states (one state in each stream's HMM) to another $K$-tuple of states. In our experiments we use 6 trained TPLHMMs to initialize one CHMM, we then get a Coupled TPLHMM, whose transitions are parameterized with speech features. After initialization it is retrained through maximum likelihood estimation.

**Animation Synthesis** Given a phoneme sequence of length $T$, together with their intensity and duration, we independently synthesize $T$ segments of appropriate duration. Each of the segment is synthesized with the corresponding model of the (phoneme, intensity) pair, which is either a set of 6 LHMMs, or a set of 6 TPLHMMs, or a CTPLHMM with 6 streams. In case TPLHMMs or CTPLHMM are used the acoustic features are exploited to alter the transition probabilities.

Whatever the models used, the synthesis is performed simply by randomly generating a state sequence according to transition probability distribution, then by synthesizing the most likely observation sequence given the state sequence, which consists in the sequence of the means of the Gaussian distribution of the states in the sequence.

## 4 Experiments

Animation synthesis model is built from human data of 2 subjects. The data contains 205 laugh sequences and 25625 frames in total. Human data from another

subject is used for validation through subjective and objective evaluation studies. It contains 54 laugh sequences and 6750 frames. Objective and subjective evaluations are conducted to validate the proposed animation synthesis model.

### 4.1 Objective Evaluation

As described in Section 3, LHMM and TPLHMM treat separately each dimension motion of head and torso, while the coupled model can simulate the relationship between them. We first investigate whether such a coupling is relevant; then we compare the animations synthesized by LHMM, TPLHMM and CTPLHMM with respect to few quantitative criterion.

**Investigating Relation between Head and Torso** To investigate the relevance of joint modeling of the 6 dimensions animation we tested the probabilistic independency between the 6 random variables corresponding to the states that are occupied at the same time in the 6 streams' LHMMs. For each pair of streams we built a contingency table for the two random variables of being in a state in the HMM for stream 1 while being in a state in the HMM for stream 2, then we computed a $\chi^2$ test to evaluate the independency between the two random variables. We found that whatever the two streams are and whatever the model is, i.e whatever the pair (phoneme, intensity) is, the two random variables were found statistically dependent at a p-value lower than 0.001. This means jointly modeling the multiple streams is actually relevant and should lead to improved animation.

Furthermore to quantify the degree of dependency between the multiple streams we computed relative mutual information. The mutual information between two random variables $X$ and $Y$, $I(X,Y)$, equals the difference between the entropy of $X$, $H(X)$ and the conditional entropy of $X$ given $Y$, $H(X|Y)$. If $X$ and $Y$ are independent, $Y$ does not bring any information about $X$ and $I(X,Y) = 0$. Alternatively, if $Y$ includes some information about $X$, the uncertainty on $X$ is reduced when knowing $Y$ so that the conditional entropy $H(X|Y)$ is lower than $H(X)$ and $I(X,Y) > 0$. Furthermore one can measure the amount of information $Y$ brings on $X$ by computing a normalized mutual information $\hat{I}(X,Y) = I(X,Y)/H(X)$ where $H(X)$ is the entropy of $X$. The normalized mutual information belongs to the range $[0,1]$. It equals 0 if $X$ and $Y$ are fully independent, while it equals 1 if $X$ may be deterministically predicted from $Y$.

In all the tests we performed we obtained normalized mutual information between 17% and 22% which shows that some uncertainty exists between the 6 dimensions of the animation but that it is not fully random either.

As a conclusion, the 6 dimensions of the animation are not independent. Hence, independent modeling of the 6 streams would be suboptimal, and these are not deterministically linked, meaning that a pure synchronous modeling of the 6 streams in a single LHMM or a single TPLHMM would not be a good option either. Finally these resuts justify our choice of modeling the 6 dimensional animation signal within a coupled HMM that enables modeling a weak dependency between the streams.

**Similarity between synthesized and real animations** We compared our models by computing 3 criteria which allow evaluating the similarity between a synthesized signal and a real signal. Basically we consider the quality of the synthesized signal with respect to three features: the main frequency of the signal, as extracted by the Periodicity Algorithm [17], the amplitude of this main frequency, and the energy of this frequency. These criteria allow investigating if the main features of a shaking-like movement are well modeled by the synthesis system.

For each of the three features we computed a normalized error (e.g. $\left|\frac{f^s - f^h}{f^h}\right|$ for the frequency feature, where $f^s$ and $f^h$ stand for the frequency of the synthesized and of the human animation signals averaged over all phonemes realizations. The lower such a measure is the closer the synthesized signal is from the original one. The frequency, amplitude and energy errors obtained for our various models are reported in Figure 1. According to these measures, TPLHMM and CTPLHMM do perform much better than LHMM while the difference of performance between TPLHMM and CTPLHMM is less clear.

**Table 1.** Performance of the models with respect to the synthesis quality (frequency, amplitude and energy errors). Performances are averaged results gained on 54 test sequences (standard deviations are given in brackets).

| Model | frequency | amplitude | energy |
|---|---|---|---|
| LHMM | 0.21 (0.074) | 0.24 (0.100) | 0.41 (0.071) |
| TPLHMM | 0.17 (0.063) | 0.19 (0.066) | 0.34 (0.057) |
| CTPLHMM | 0.17 (0.061) | 0.20 (0.059) | 0.31 (0.052) |

### 4.2 Subjective Evaluation

Two subjective evaluations were conducted through an online web application. First, we compare the animations synthesized by TPLHMM and CTPLHMM; then the best one is compared to human data. The participants were invited to watch 5 videos of laughing virtual character and to answer few questions for each video. They could control when to start the videos and could watch them as many times as they wish. Our aim is to evaluate the behaviors animation and not the appearance of the virtual agent. We used the same virtual agent to display motion data for both subjective evaluations. Motion data displayed with the virtual character consists of head and torso movements (motion capture or generated data) and facial expression. Facial expression of laughter was computed using our previous approach [18]. The 5 videos used in both subjective studies last respectively $9s$, $10s$, $18s$, $26s$ and $27s$.

**TPLHMM and CTPLHMM Comparison:** To compare TPLHMM and CTPLHMM, both trained models were applied to the 5 test samples. For each

test sample, a pair of videos was recorded in which the virtual agent's head and torso motions were driven respectively by these models. Each pair of video clips was displayed on the same web page and randomly arranged on the right or on the left. After watching each pair of video clips, participants were invited to select the best animation along four dimensions: naturalness of the animation, synchronization of head and torso movements with laugh sound, correlation of laughter intensity and torso movements, inter-correlation of head and torso movement.

This evaluation study involved 120 participants, 67 males and 53 females with age ranging from 18 to 65 years old (Mean=33.5 years, SD=9.6 years). We computed 95% confidence intervals that show that CTPLHMM is significantly better than TPLHMMs with respect to the 4 questions: we obtained a confidence interval equal to [66% 77%] for CTPLHMM being better than TPLHMMS with respect to Naturalness, [60% 72%] for Synchronisation, [58% 70%] for Intensity correlation and [63% 74%] for Head and Torso inter-correlation.

**Synthesized and Human Data Comparison:** With respect to the results above, CTPLHMM is perceived as the best animation synthesis framework; so we use the animations obtained with CTPLHMM in the comparison test with human data. This subjective evaluation was conducted to investigate how similar is the perception of the virtual agent displaying head and torso motions synthesized by CTPLHMM to the perception of the virtual agent displaying head and torso motions synthesized by CTPLHMM is similar to the perception of the virtual agent animated directly by human data. As the previous study, a comparison test was conducted.

In total, there were 80 participants consisting of 46 males and 34 females with age ranging from 12 to 78 (M=40.65 years, SD=17.91 years). To verify the hypothesis, 2 versions (conditions) of the virtual agent animations were created for each selected test sample. They are human and synthesized motions. There are a total of 10 video clips (5 input samples $\times$ 2 conditions). Each participant watched 5 video clips, each of which is randomly selected from the 2 conditions. Each video clip has been evaluated 40 times (i.e., by 40 participants). After watching each video clip, each participant was invited to answer the same four questions as in the first evaluation study, but this time the participant answered using a 5 point Likert scale.

The results are shown in Figure 2. As can be seen, synthesized motion obtains score less than human motions along the four dimensions: naturalness, synchronization, correlation of laughter intensity and torso movements, inter-correlation of head and torso movement. T-test shows that there are significant differences in all terms between human and synthesized data.

### 4.3 Discussion

The objective evaluation for comparing LHMM, TPLHMM and CTPLHMM shows that TPLHMM and CTPLHMM perform better than LHMM. It highlights that acoustic features and motions are linked. Thus acoustic features can

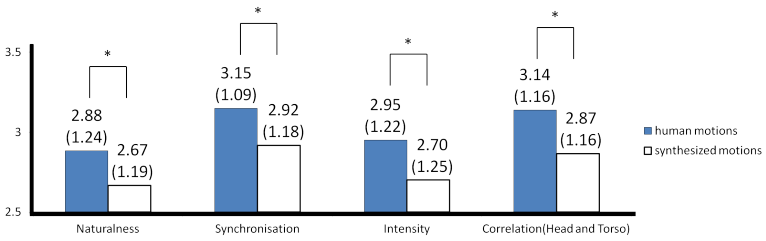


**Fig. 2.** Averaged values of virtual agent animated by animations from human and synthesized. Significant differences are identified by $\star$ ($P < .05$). The averaged values are shown with an histogram and the standard deviation is specified in parenthesis.

be used to capture motion trajectories. In LHMM, inputs of text signals, such as phoneme, intensity and duration, are global-level features. They do not contain enough information to characterize dynamic motion variance at each time frame. While, in TPLHMM and CTPLHMM, for each time frame, additional acoustic features are used to characterize dynamic variance of human motion. In LHMM and TPLHMM models, head and torso motions are modelled separately. In other words, they are considered as being independent. However, through the objective evaluation investigating the relation between head and torso, we found that head and torso motions are dependent with each other; relationship which is ignored in the other two models. In our work, coupled model is used to learn this dependent relation between head and torso movements.

The subjective evaluation compared TPLHMM and CTPLHMM. CTPLHMM obtains higher score than TPLHMM. In the subjective evaluation on comparing synthesized and human motions, human data is perceived significantly better than synthesized data in terms of naturalness, synchronisation, intensity and correlation of head and torso movements. However the difference in perception is not so severe (less than 1 on a 5 likert scale). This suggests that the proposed CTPLHMM is somehow capable of synthesizing human-like head and body motions.

## 5 Conclusion

In this paper we have presented an approach to model laughter head and torso movements, which are very rhythmic and show saccadic patterns. To capture laughter motion characteristics, we have developed a statistical approach to reproduce frequency movements, such as shaking and trembling. Our statistical model takes as input such phoneme sequences and acoustic features of laughter sound. Then it outputs the head and torso animations of the virtual agent. In the training model, not only the relation between input and output features is modelled, but also the relation between head and torso movements is captured. Experiments show that our model is able to capture the dynamism of laughter movement, but do not overcome animation from human data.

## References


1. Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., Shapiro, A.: Virtual character performance from speech. In: Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation. (2013) 25–35
2. Cassell, J., Vilhjlmsson, H., Bickmore, T.: Beat: The behavior expression animation toolkit, Proceedings of SIGGRAPH (2001) 477–486
3. Ruch, W., Kohler, G., Van Thriel, C.: Assessing the 'humorous temperament': Construction of the facet and standard trait forms of the state-trait-cheerfulness-inventory - stci. Humor: International Journal of Humor Research **9** (1996) 303–339
4. Provine, R.R.: Laughter: A scientific investigation. Penguin books edn. (2001)
5. Huber, T., Ruch, W.: Laughter as a uniform category? A historic analysis of different types of laughter. In: Congress of the Swiss Society of Psychology. (2007)
6. Ruch, W., Ekman, P.: The Expressive Pattern of Laughter. Emotion qualia, and consciousness (2001) 426–443
7. Darwin, C.: The expression of the emotions in man and animals. London: John Murray (1872)
8. de Melo, C.M., Kenny, P.G., Gratch, J.: Real-time expression of affect through respiration. Computer Animation and Virtual Worlds **21**(3-4) (2010) 225–234
9. Mancini, M., Varni, G., Glowinski, D., Volpe, G.: Computing and evaluating the body laughter index. In: Proceedings of Human Behavior Understanding. (2012) 90–98
10. McKeown, G., Curran, W., McLoughlin, C., Griffin, H.J., Bianchi-Berthouze, N.: Laughter induction techniques suitable for generating motion capture data of laughter associated body movements. FG (2013) 1–5
11. Niewiadomski, R., Mancini, M., Baur, T.: MMLI: Multimodal multiperson corpus of laughter in interaction. In proceeding of: 4th international workshop on Human Behavior Understanding **8212** (2013) pp 184–195
12. Fourati, N., Pelachaud, C.: Emilya: Emotional body expression in daily actions database. In: LREC 2014, the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland (2014) 3486–3493
13. Urbain, J., Çakmak, H., Dutoit, T.: Automatic phonetic transcription of laughter and its application to laughter synthesis. In: Proceedings of Affective Computing and Intelligent Interaction. (2013) 153–158
14. DiLorenzo, P.C., Zordan, V.B., Sanders, B.L.: Laughing out loud: control for modeling anatomically inspired laughter using audio. ACM Trans. Graph. **27**(5) (2008) 125
15. Wilson, A., Bobick, A.: Parametric hidden markov models for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(9) (1999) 884–900
16. Brand, M.: Coupled hidden markov models for modeling interacting processes. Technical report (1997)
17. Sethares, W., Staley, T.: Periodicity transforms. IEEE Transactions on Signal Processing **47**(11) (1999) 2953–2964
18. Ding, Y., Prepin, K., Huang, J., Pelachaud, C., Artières, T.: Laughter animation synthesis. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems. (2014) 773–780