Modeling Multimodal Behaviors From Speech Prosody

Yu Ding[1], Catherine Pelachaud[1], and Thierry Artières[2]

[1] CNRS-LTCI, Institut Mines-TELECOM, TELECOM ParisTech, Paris, France
{yu.ding, catherine.pelachaud}@telecom-paristech.fr
[2]Université Pierre et Marie Curie (LIP6), Paris, France
thierry.artieres@lip6.fr

Abstract. Head and eyebrow movements are an important communication mean. They are highly synchronized with speech prosody. Endowing virtual agent with synchronized verbal and nonverbal behavior enhances their communicative performance. In this paper, we propose an animation model for the virtual agent based on a statistical model linking speech prosody and facial movement. A fully parameterized Hidden Markov Model is proposed first to capture the tight relationship between speech and facial movement of a human face extracted from a video corpus and then to drive automatically virtual agent's behaviors from speech signals. The correlation between head and eyebrow movements is also taken into account during the building of the model. Subjective and objective evaluations were conducted to validate this model.

Keywords: virtual agent, speech to motion synthesis, head motion synthesis, eyebrow motion synthesis, Hidden Markov model, speech driven

1 Introduction

Embodied conversational agents, ECAs, are autonomous software characters that often have a human-like appearance and endowed with communicative and expressive capabilities. They are capable of using speech and multimodal behaviors to convey intentions and to express emotions as humans do. The prevalence of embodied conversational agents in interactive systems such as online web applications has been motivated by the development of computational models to generate realistic, natural and believable virtual agents. In video games or cinematographic applications the animation of virtual characters are reproduced from large motion capture datasets of an actor's performance. This approach induces two non-negligible disadvantages: data acquisition expenses and restriction to movement reproduction of the recorded scenarios [1]. On the other hand, the control of the behaviors of autonomous virtual characters is either based on psychology literature [2] or on statistical approaches such as Markovian model [1].

Human-to-human communication is a multimodal process involving speech, facial expression, body gesture and gaze, etc. Humans are sensitive to subtle expressions during face-to-face conversation. For example, they are skilled in inferring their interlocutor's affective and mental states from their accompanied facial expression or body gestures. [3] reported that natural head motion significantly facilitates auditory speech perception. Speech and behaviors production are tightly coupled [4, 5]. For example, [6] found a strong correlation between the raise of pitch contour (F0) and of eyebrow movements.

Our work is focused on investigating a statistical model to infer eyebrow and head motions from speech signals. This model can be parameterized from training samples and can then synthesize natural animation motion from speech features. In our model, head and eyebrow motions are not separately synthesized from speech signals; rather it takes into account the relationships between these two-modal motions.

In the remaining of this paper we first describe related works, we introduce our approach and finally we report on experimental results from objective and subjective evaluations.

2 Related works

[7–9] proposed rule-based approaches to generate nonverbal communicative features, such as head motion and gesture. However, since human behaviors arise from and may be influenced by various factors, such as emotion, personality, gender, physiological state and social context [10], it is extremely difficult to define a large set of rules to fully capture the role of these factors onto human behaviors, even after decades of studies in psychology. Besides, multiple rules could result in synthesizing conflicting expressions [10].

Recently, data-driven models have been proposed as body or facial motion generators. Few predictive models have been investigated such as Conditional Restricted Boltzmann Machine (CRBM) [11] but the majority of systems are built using statistical models such as Hidden Markov models (HMMs) or more generally state space models. A state space model implements a probability density on observed sequences (e.g. a temporal series of head position). In a state space model the observed sequence is assumed to be produced in two steps: first a state sequence is chosen, second an observation sequence is produced given the state sequence. In both cases the choice is done by drawing a sample according to the corresponding model distribution. Such models are widely used to model speech, handwriting, etc. For instance to model speech phones one uses a state space model with one state for the beginning of the phone, one for the middle part and one for the end of the phone [12]. Within each state observation are produced according to a particular probability density (usually a Gaussian mixture).

A common approach for designing a speech to motion synthesis system relying on such statistical models consists in learning two state space models. There is one model for the speech stream and another model for the motion stream. Both models have the same number of states that are learnt jointly in such a way that these states are paired. Let us illustrate this approach with a simple case where the system consists in learning two HMMs where the first model is trained to model speech observation sequence whereas the second model is trained to model head movement observation sequences. First the speech HMM is learnt on speech observation sequences. Then this model is used to infer the most likely state sequence for every training speech observation sequence. Second the head move HMM model is trained by considering that head move observation sequence have been produced along the state sequences determined by the speech HMM. Doing so one learns which head movements are likely to occur for a given speech signal. The speech/motion mapping is somehow modeled by pairing the states. In presence of speech only, the speech model is first used to infer a state sequence then the head move model is used to syntesize a series of head positions along this state sequence via synthesis techniques as those proposed in [13]. This approach has been implemented in various ways. [14] used two Gaussian Mixture Models; while [15, 16, 1, 17–19] used two HMMs and [20] used a Conditional Random Field (CRF) for the speech and a HMM for motion.

Although this line of works has brought significant results it suffers from the weak interdependency between the input stream (e.g. speech) and the output stream (e.g. head movements) that is actually taken into account through the state space sharing strategy mentioned above.

3 Contextual Markovian Models

We have developed few variants of HMMs to simulate facial animation from speech described in a previous work [21]. In particular we proposed a new model that we name here fully parameterized Hidden Markov model (FPHMM) for learning a mapping function between speech and motion signals, where speech signals can influence directly the synthesized motion signals. We use such models here to simultaneously generate head and eyebrow motions from speech input. We first briefly introduce this model and we show how it can be used to synthesize a motion stream from a speech stream. More details on the models may be found in [21].

A FPHMM is an extension of a contextual HMM (named CHMM hereafter), which has been initially proposed in [22] for modeling and recognizing gestures. The idea behind CHMM is to exploit some contextual information that we know the observation we want to model depends on. Contextual variables may stand for the physiology of a person realizing the gesture, the amplitude of the gesture, the emotional state of a person speaking... [22, 23]. In CHMM contextual variables are used to alter probability density functions in states. More precisely, a CHMM is a HMM whose means and covariance matrices of Gaussian distribution depend on a set of contextual variables, noted by θ, that may vary with time [22, 23]. The contextual variables corresponding to a particular observation sequence are assumed to be known in the training stage as well as in the test stage.

A FPHMM is an extension of a CHMM where in addition to means and covariance matrices, transition probabilities and initial state distribution are also parameterized and depend on θ instead of being fixed at particular values.

In a FPHMM the transition probability a_{ij} from the i^{th} state to the j^{th} state at time t is defined as:

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

where θ_t is the c-dimensional vector of contextual features at time t and W's are matrices associated to transitions. Using such a modeling framework transition probabilities vary with time according to the values of contextual variables, meaning that a transition may be more likely to occur or not occur at some time according to the contextual information. As any statistical model a FPHMM is trained via likelihood maximization with a Generalized EM algorithm. To ease learning it is initialized with a trained CHMM.

In our work, a FPHMM is used to synthesize the motion stream from the speech stream as follows. We first learn a FPHMM that takes speech features as contextual variables and that produces motion features observation. During the training phase, motion and speech streams are both used to learn a FPHMM. During the motion synthesis phase (i.e. the animation generation phase), only the speech stream θ is known. It is used to compute the time dependent transition probabilities and the time dependent altered emission probability distributions in states (cf. e.g. equation (1)). Once all the parameters of the model are set, one can compute the most likely state sequence, or to get even a more accurate result one can infer the probability distribution over all state sequences. Finally, from this single state sequence or from the distribution over state sequences, one can synthesize a trajectory using techniques such as in [13].

4 Experiments

In this section, we present the corpus used in our experiments and how we extracted facial and prosodic features. Then we describe the conducted objective and subjective evaluations. At last, we report the results and discuss them.

4.1 Datasets

Experiments were performed on the Biwi 3D AudioVisual Corpus of Affective Communication database [24]. We used 240 sequences from 3 subjects extracted from this corpus. In this corpus, each subject tells 80 short English sentences. Each sentence lasts 4.67s long on average. 3D face geometries were captured at 25Hz, which comprise of a total of 23370 facial points including 3 head rotations.

From such a large set of facial points, we extracted a subset of facial motion features that correspond to 3 head rotations and 8 eyebrow features (see Figure 1). These features coincide with the Facial Animation Parameters as defined by the norm MPEG-4 [25]. For sake of simplicity, we assume both eyebrows move identically and we take the mean of the right and the left eyebrows as the eyebrow

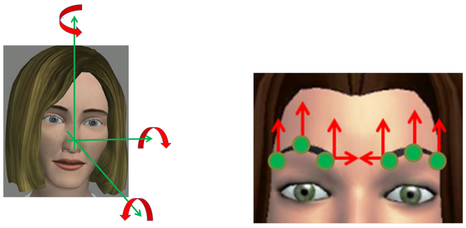

Fig. 1. Facial motion features - Left: 3 head rotations. Right: Eyebrow animation parameters (arrows illustrate displacements).

motion features. At the end a motion signal is transformed in a sequence of 7-dimensional (3-dimensional head and 4-dimensional eyebrow) feature vectors at a rate of 25 frames (i.e. feature vectors) per second (fps).

Concerning the speech features we consider 2 prosodic features (pitch and RMS energy), which were extracted with PRAAT software [26] at the same sample rate as for motion feature extraction (25 fps).

We call static features, the 7-dimentional motion features and the 2 prosodic ones. On top of these features, we include also the first and second order derivatives of static features (i.e. velocity and acceleration of the dynamic features).

4.2 Objective evaluation

Table 1. Performance of the models with respect to the synthesis quality (MSE). Performances are averaged results gained on 50 experiments (standard deviations are given in brackets).

Model	10 states	20 states	30 states	50 states
[18]	0.57 (0.054)	0.53 (0.049)	0.49 (0.061)	0.46 (0.053)
separate PFHMM	0.45 (0.069)	0.41 (0.066)	0.39 (0.059)	0.38 (0.051)
joint PFHMM	0.39 (0.071)	0.34 (0.059)	0.30 (0.045)	0.30 (0.036)

We first performed experiments for modeling head and eyebrow features separately with 2 PFHMMs, one for each motion stream. Then experiments were performed to jointly model head and eyebrow features with a single PFHMM, where the joint features of head and eyebrow were considered as FHMM's observation features. We compare our results with the baseline approach proposed in [18] that we have implemented and that we have tuned for the task at hand.

These two sets of experiments were evaluated by computing the reconstruction error defined as the mean square error between the synthesized motion signal (from the speech signal) and the real motion signal (MSE criterion). Table 1 reports the experimental performances with different numbers of states based on the averaged results and the standard deviation over 50 random splits of the dataset into 80% for training and 20% for testing. The experiments are configured using full covariance matrix and ergodic topology, which fully guarantee the model specified by samples of data training. As can be seen in Table 1, the joint model outperforms the baseline [18] as well as the other two using separate models.

In these experiments, a combination of covariance matrices were exploited for the observation function in FPHMM. In the joint model, the relationship between eyebrow and head features can be learned using the full covariance matrices. In synthesis phase (trajectory computation), generated eyebrow and head motions are defined not only from the speech features but also from their mutual influence. This influence is captured during the training phase through the combination of covariance matrices. On the other hand, when using two separate models, there is an underlying hypothesis implying that head and eyebrow movements are independent from each other that is carried out during the training phase and that produces poorer results.

As can be seen from Table 1, the joint model with 30 states achieves the best result. We use this model to compute the animation of the virtual agent. In the next section we present the subjective evaluation study we conducted where we looked at qualitative measures. It is a necessary complement of the objective measure we just presented. Indeed objective measure does not allow us to measure how motions are perceived by human eyes [27].

4.3 Subjective evaluation

In this section, we detail the subjective evaluation we conducted with human participants to evaluate the qualitative aspects of FPHMM as generator of head and eyebrow motion from speech signals. The evaluation was done through an online web application.

Hypothesis The subjective evaluation was conducted to investigate two hypotheses: 1) the perception of the virtual agent displaying head and eyebrow motions synthesized by FPHMM is similar to the perception of the virtual agent animated directly by human data; 2) the two-modal motions (head and eyebrow) outperform single model motion (either head or eyebrow) at a perceptual level. Through the first hypothesis we aim to measure if FPHMM is capable of capturing and of rendering the sophisticated relationship between motion and speech streams and that the animation of a virtual character offers similar results when driven by FPHMM and by real human data. The second one is to verify that multimodal motions facilitate human perception over monomodal motions. To verify the first hypothesis, we compare the perceptions resulted from real and generated motions. To answer the second hypothesis, we compare how animations of the virtual agent driven either by one of the modal motions (either head or eyebrow motions) or by two modal motions (head and eyebrow motions) are perceived.

Our work focuses on nonverbal communication and not on the appearance of the virtual agent. Therefore, motions from both, FPHMM types and human data, are displayed through the identical virtual agent.

Protocol The participants went on a web page where after answering few questions about themselves, they have to view videos of the virtual agents. Their task consisted in answering few questions. We provide elements of the protocol we follow for the perceptive evaluation study.

(1) Participants: in total, there were 280 participants consisting of 136 males and 144 females with age ranging from 18 to 65 (M=32.89 years, SD=7.99 years).

(2) Stimuli: 7 spoken utterances were randomly selected from the testing database. They were given as input to the trained FPHMM. Then the synthesized motions (sequences of MPEG-4 FAPs frames) and the corresponding WAV file of the spoken utterances are used to drive a virtual agent. In all the animations the lip shapes and body movements of the virtual agents are reproduced from real data. The final animations including eyebrow and head motion as well as lip shape and body movements are stored as video clips.

To study both hypotheses, 7 versions (conditions) of the virtual agent animations were created for each selected sentence. In all conditions the lip shapes and body movements remain constant and are duplicated from real human data:

1^{st} condition (cond1): No eyebrow and head motion;

2^{nd} condition (cond2): Only human eyebrow motion (no head motion);

3^{rd} condition (cond3): Only synthesized eyebrow motion (no head motion);

4^{th} condition (cond4): Only human head motion (no eyebrow motion);

5^{th} condition (cond5): Only synthesized head motion (no eyebrow motion);

6^{th} condition (cond6): Human eyebrow and head motion;

7^{th} condition (cond7): Synthesized eyebrow and head motion;

Therefore, there are a total of 49 video clips (7 sentences × 7 conditions), each of which lasts about 4.5s.

(3) Design and Procedure: Subjective evaluations were conducted online. At first, each participant fills out a demographic questionnaire concerning their age, gender, education level, occupation and country in which participant spent the majority of his/her life. Then, the participant watches 7 randomly selected video clips out of 49. The 7 video clips watched by any participant are comprised of the 7 sentences and of the 7 conditions. After watching each video clip, each participant is invited to answer the following questions using a 5 point Likert scale:

1. Do you think the animation of the virtual character is intelligible?
2. Do you think the animation of the virtual character is natural?
3. Do you think the correlation between the speech and the facial expression of the virtual character is coherent?
4. Do you think the correlation between the speech and the facial expression of the virtual character is synchronized?
5. Which emotion(s) does the virtual character display? You should grade each emotion separately.

The same 12 emotional states as used in the BIWI experiments are considered [24]: anger, sadness, fear contempt, nervousness, disgust, frustration, stress, excitement, confidence, surprise and happiness. Each video clip has been evaluated 40 times (i.e., by 40 participants).

Table 2. Results of F-statistic from repeated measures ANOVA

	intelligible	natural	coherent	synchronized
only eyebrow	F=20.6,p<.001	F=1.79, p>.05	F=2.02, p>.05	F=2.57, p>.05
only head	F=1.7, p>.05	F=0.39, p>.05	F=0.02, p>.05	F=2.57, p>.05
eyebrow&head	F=3.51, p>.05	F=0.93, p>.05	F=1.9, p>.05	F=0.01, p>.05

Result To investigate the differences of participants perception from human and from synthesized motions, we conducted three pairwise comparisons in term of intelligibility, naturalness, coherency and synchronization: only eyebrow motion (2nd and 3rd conditions), only head motion (4th and 5th conditions) and both (6th and 7th conditions). The comparison results based on repeated measures ANOVA are shown in Table 2. The results show no significant differences between human and generated motions in almost all pairwise conditions except for the only eyebrow condition in term of intelligibility. In this latter case, the mean scores of this pairwise condition are 2.67 and 2.27 for only human and synthesized motions, respectively. While the difference is significative, they are still not too highly different.

Then, to test our second hypothesis (the two-modal motions outperforms single model motion), we compare the 4 different conditions involving synthesized motions with each other, namely: no head and eyebrow motions (1st condition), only synthesized eyebrow motion (3rd condition), only synthesized head motion (5th condition) and both synthesized (head and eyebrow) motions (7th condition). The comparison results presented in Figure 2 show that when eyebrow and head motions are modeled together, the human perception improves in the term of intelligibility, naturalness, coherency and synchronization. The results based on repeated measures ANOVA show significant differences between any two different types of motions among the 4 cases (synthesized eyebrow or head, both synthesized eyebrow and head, no motion of eyebrow and head). The same conclusions are supported by the similar pairwise for the animations from human data (see details in Figure 2).

At last, we investigated how synthesized motion conveyed emotional information. To do this, we extracted the scores in term of 12 emotions from both synthesized head and eyebrow motions (7th condition) and from both human motions (6th condition), respectively. Moreover, the BIWI corpus provides the emotion recognition rate for the videos of real humans for these 12 emotions using a 5 point Likert scale. We consider these results as reference when evaluating the virtual agent's performance. Figure 3 reports the recognition rate of

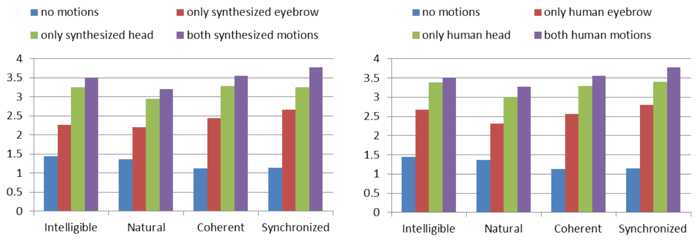

Fig. 2. Comparison results between animations from synthesized data (left) and human data (right).

participants for the virtual agent when driven from human data and from synthesized model, as well as for the videos of real humans. The results are average over the recognition rate for 5 of the 7 sentences spoken by the virtual agent. Two of the sentences have to be disregarded for this test as no result is provided for them in the case of the evaluation of the real human videos. The number of participants differ in the case of the study made with videos of real humans as with videos made with virtual agents. In case of the BIWI study, the first set of videos was evaluated when the BIWI corpus was built. There is a very low number of participants that evaluated each video (often around 4 participants per sentence) while in our study we have 40 participants evaluating all the sentences in average. As can be seen in Figure 3, in all three cases, the faces, be of a real human or of a virtual agent, are able to convey some emotional communication. However, in term of the perception of emotional expressiveness, the videos of the real humans outperform the videos of the virtual character driven from our statistical model; in turn, the videos of the virtual character driven from our statistical model outperform the videos of the virtual character driven from human data.

Due to the too big difference between participants number, we only compare results between the conditions of the virtual agent driven from our model and from human data. The animations with synthesized motions are perceived as showing more emotions in a statistically significant way for the emotion: fear, contempt, nervousness, stress, excitement, surprise, happiness. There is no significant difference for the emotions: anger, sadness, frustration. For the emotions disgust and confidence, the animations from human data are ranked higher than the animations from synthesized data.

Discussion The post-hoc pairwise comparisons between identical modal motions show no significant difference between the human and synthesized motions. In the training phase, the mapping between human audio-visual signals is captured and recorded in FPHMM, and thus rendered in the output synthesized animations. Therefore the perception of the animation of the virtual agent with synthesized motions is similar to the perception of animation with human motions.

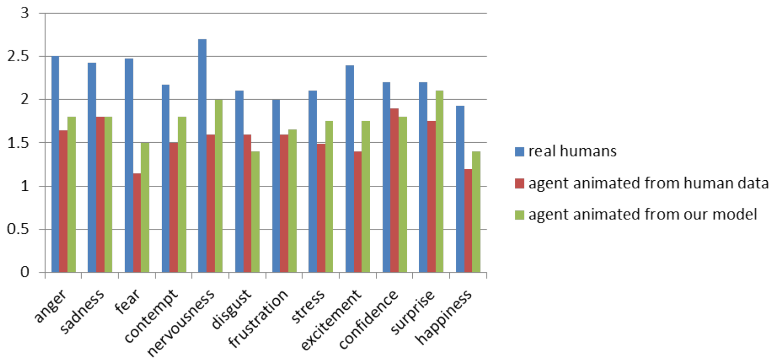

Fig. 3. Perceived emotions: average values over 5 sentences.

The post-hoc pairwise comparisons of conditions show that there are significant differences among the monomodal and multimodal conditions. The animations created in the multimodal conditions are perceived as more intelligible, natural, coherent and synchronized than the animations created with one modality only. Humans are very skilled in reading nonverbal signals. So the lack of either eyebrow or head motions are negatively perceived along these 4 qualitative dimensions. The comparisons of the perceptions between only eyebrow and only head motions reveals that head motion plays a more important role than eyebrow motion at the perception level. Similar results have also been reported in [18].

Rather than examining the recognition rate of emotions from the videos in different cases, we look at the level of emotional expressiveness. Indeed even in the reference case, namely the videos of real humans speaking the various utterances in emotionally-colored fashion, we remark that the recognition rate level of emotion is rather low and that there are a lot of confusion. That is human participants did not show a strong agreement in their perception of which emotion the human actor aimed to convey. Such a result is reproduced with the virtual agent. In both cases, animation of the virtual character from real human data and from our model, a lot of confusion can be noticed. However we can remark that the virtual agent driven by our model is perceived as speaking with emotions with a higher level than when the virtual agent is driven by human data. This can be interpreted as the virtual agent driven by our model is able to exhibit more expressive behaviors; that is, it can communicate in a more emotionally colored manner. Our statistical model relying on FPHMM is capable of capturing the speech/motion relationship. It is also able to render the quality of emotional behaviors: not only its types of movements (i.e. which head movements and eyebrow shapes) but also the dynamism of the movements. Our model computes the types of the visual cues but also their trajectory that carries out dynamics characteristics.

The results of both evaluation studies, the objective and subjective studies, show that our model is able to capture pertinent information that are conveyed through the nonverbal behavior animation of the virtual agent. Considering the link between prosody and both head and eyebrow motions together allows seizing their tight coupling. This is in link with results found in studies from psychology domain [28, 5].

We can conclude that the machine learning approach (FPHMM) can capture the link between speech prosody and facial movements and can reproduce the movement dynamism in the output synthesized animations. Thus our animation model based on FPHMM is able to gather these both aspects that are important to compute when animating a virtual agent.

5 Conclusion

In this paper, we have presented a data-driven approach to generate head and eyebrow motions for a virtual agent from speech prosody. The full parameterized HMM is used to capture the direct mapping between audio and visual information. The trained PFHMM allows defining visual animation as a function of the speech signal. The objective evaluation study shows that considering that simultaneously eyebrow and head motions increases the precision of the resulting animation. It also confirms that eyebrow and head motions are not independent from each other but rather are connected; the multimodal signals reinforce the communicative meaning. On the other hand, the subjective evaluation shows that our proposed model enhances the perception of the virtual agent animation at the level of emotional expressiveness.

References

1. Busso, C., Deng, Z., Neumann, U., Narayanan, S.: Natural head motion synthesis driven by acoustic prosodic features. Journal of Visualization and Computer Animation **16**(3-4) (2005) 283–290
2. Bevacqua, E., Prepin, K., Niewiadomski, R., de Sevin, E., Pelachaud, C.: GRETA: Towards an Interactive Conversational Virtual Companion. In: Artificial Companions in Society: perspectives on the Present and Future. (2010) 1–17
3. Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Bateson, E.V.: Visual prosody and speech intelligibility: Head movement improves auditory speech perception. Psychological Science **15(2)** (2004) 133–137
4. Kendon, A.: Gesture : Visible Action as Utterance. Cambridge University Press (2004)
5. Ekman, P.: About brows: Emotional and conversational signals. In von Cranach, M., Foppa, K., Lepenies, W., Ploog, D., eds.: Human ethology: Claims and limits of a new discipline: contributions to the Colloquium. Cambridge University Press, Cambridge, England; New-York (1979) 169–248
6. Bolinger, D.: Intonation and Its Uses: Melody in Grammar and Discourse. University Press (1989)
7. Pelachaud, C., Badler, N.I., Steedman, M.: Generating facial expressions for speech. Cognitive Science **20** (1996) 1–46

8. Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Bechet, T., Douville, B., Prevost, S., Stone, M.: Animated conversation: Ruled-based generation of facial expression gesture and spoken intonation for multiple conversational agents. In: Computer Graphics. (1994) 413–420
9. Beskow, J.: Rule-based visual speech synthesis. In: ESCA - EUROSPEECH '95. 4th European Conference on Speech Communication and Technology, Madrid (September 1995)
10. Lee, J., Marsella, S.: Modeling speaker behavior: A comparison of two approaches. In: IVA. (2012) 161–174
11. Chiu, C.C., Marsella, S.: How to train your avatar: A data driven approach to gesture generation. In: IVA. (2011) 127–140
12. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. In: Proceedings of the IEEE. (1989) 257–286
13. Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T.: Speech parameter generation algorithms for hmm-based speech synthesis. In: ICASSP. (2000) 1315–1318
14. Costa, M., Chen, T., Lavagetto, F.: Visual prosody analysis for realistic motion synthesis of 3d head models. In: Proc. of ICAV3D. (2001) 343–346
15. Dziemianko, M., Hofer, G., Shimodaira, H.: Hmm-based automatic eye-blink synthesis from speech. In: INTERSPEECH. (2009) 1799–1802
16. Hofer, G., Shimodaira, H., Yamagishi, J.: Speech driven head motion synthesis based on a trajectory model. In: ACM SIGGRAPH 2007 posters. (2007)
17. Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S.: Rigid head motion in expressive speech animation: Analysis and synthesis. IEEE Trans. on Audio, Speech & Language Processing **15**(3) (2007) 1075–1086
18. Mariooryad, S., Busso, C.: Generating human-like behaviors using joint, speech-driven models for conversational agents. IEEE Trans. on Audio, Speech & Language Processing **20**(8) (2012) 2329–2340
19. Xue, J., Borgstrom, J., Jiang, J., Bernstein, L., Alwan, A.: Acoustically-driven talking face synthesis using dynamic bayesian networks. In: Multimedia and Expo, 2006 IEEE International Conference on. (2006) 1165–1168
20. Levine, S., Krähenbühl, P., Thrun, S., Koltun, V.: Gesture controllers. ACM Trans. Graph. **29**(4) (2010)
21. Ding, Y., Radenen, M., Artières, T., Pelachaud, C.: Speech-driven eyebrow motion synthesis with contextual markovian models. In: ICASSP. (2013) 3756–3760
22. Wilson, A.D., Bobick, A.F.: Parametric hidden markov models for gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **21** (1999) 884–900
23. Radenen, M., Artières, T.: Contextual hidden markov models. In: ICASSP. (2012) 2113–2116
24. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L.: A 3-D Audio-Visual Corpus of Affective Communication. Multimedia, IEEE Transactions on **12**(6) (October 2010) 591–598
25. Pandzic, I., Forcheimer, R.: MPEG4 Facial Animation - The standard, implementations and applications. John Wiley & Sons (2002)
26. Boersma, P., Weeninck, D.: Praat, a system for doing phonetics by computer. Glot International **5**(9/10) (2001) 341–345
27. Lee, J., Marsella, S.: Predicting speaker head nods and the effects of affective information. Multimedia, IEEE Transactions on **12**(6) (2010) 552–562
28. McNeill, D.: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago (1992)