Speech Communication 98 (2018) 28–41

Contents lists available at ScienceDirect

Speech Communication

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

HMM-based generation of laughter facial expression

Hüseyin Çakmak*, Thierry Dutoit

University of Mons (UMONS), TCTS lab, Boulevard Dolez 31, Mons, 7000, Belgium

ARTICLE INFO

Keywords:
Visual
Laughter
Generation
Facial expression

ABSTRACT

This paper proposes a model for visual laughter generation by the means of speaker-dependent training of Hidden Markov Models (HMMs). It is composed of the following parts: 1) facial and 2) and head motions are modeled with separate HMMs while 3) eye-blink are added as a post-processing step on the generated eyelid trajectories.

The models are trained on a database of facial expressions recorded on one male subject watching humorous videos. A commercially available marker-based motion capture system was used to record the visual data. A preliminary study has shown that modeling head motion with the same transcriptions as for facial deformation is not the best choice due to the rigidness of the resulting head motion.

Finally, the generated facial laughter trajectories are used to animate a 3D face model and the corresponding animation is rendered in a video. An online perception MOS test is conducted to assess the improvement compared to the previous method and to compare with the perception of ground truth trajectories. Results show that the new approach significantly outperforms the previous one.

1. Introduction

Laughter is one of the most important signals of human interactions. It has various functions in the social context, among others conveying our emotions, back-channeling, displaying affiliation or mitigating an unpleasant comment (Glenn, 2003).

With advances in human-machine interaction and developments in speech processing, a growing interest in laughter processing has been seen mostly in the last decade. Detecting, analyzing and producing laughter have become tasks that a machine should be able to perform in order to enhance applications including human-machine interactions.

The basic aim of visual laughter generation is to produce an animation which looks natural and which corresponds to a given audio laughter sequence which may itself be generated. Studies in the field of visual laughter generation are rare (cf Section 2). Some focused on the chest motion or the shoulders' motion during laughter and very few on the facial deformations. Our approach presented here has the advantage of being based on motion capture data (high precision trajectories of facial deformations) and of being parametrized in such a way that the generated trajectories may be used to animate any correctly rigged 3D model.

This paper focuses on the visual part of a project aiming at building a system capable of producing audiovisual laughter with controls on intensity, type and duration. The approach followed here is to model facial deformations by means of landmark trajectories. The basic steps that we followed throughout the work are:

1. Recording of the 3D data using a motion capture system (Naturalpoint, 2013)
2. Post-processing to shape the data for training
3. Training Hidden Markov Models (HMMs) on this data. Different models are built for facial and head motion with their respective transcriptions
4. Generating trajectories from these models
5. Using the generated trajectories on a 3D face model
6. Rendering a video of the animation

A first method which was partially explained in a previous work (Cakmak et al., 2014) and a novel one are compared in this paper. In the old method, the same transcriptions are used to model the facial deformations and the head motion trajectories. In the new method, the facial deformation data and head motion data are modeled using separate transcriptions. In both cases, a PCA is applied on the data consisting of the trajectories along each axis for each marker (cf Section 3). However, in the new method, separate PCAs are applied on facial deformation data and head motion data (Fig. 2) to the contrary of the first method in which all the translations, including head motion trajectories, were transformed through a single PCA(Fig. 1). Although the resulting animations were acceptable, they did not match the scores obtained by the animations rendered using the real trajectories. This

* Corresponding author.
E-mail address: huseyin.cakmak@umons.ac.be (H. Çakmak).

https://doi.org/10.1016/j.specom.2017.12.006
Received 12 January 2017; Received in revised form 30 July 2017; Accepted 15 December 2017
Available online 16 December 2017
0167-6393/

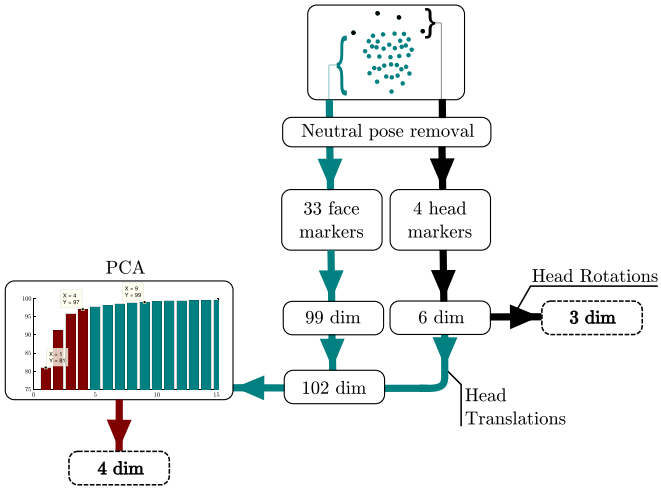


Fig. 1. Diagram of the data preparation pipeline before modeling used in the old method.

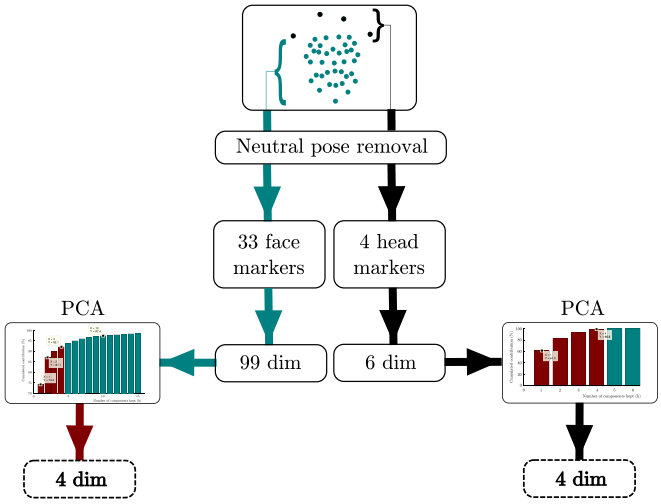


Fig. 2. Diagram of the data preparation pipeline before modeling used in the new method.

work tries to fill this gap in the perceived quality between the animations built using generated trajectories and real trajectories by a better modeling of the head motion which appeared to be the most penalizing part in the old method.

This paper is organized as follows: Section 2 gives information on related works, Section 3 gives information about the database used in this work as well as the post-processing that was done to shape the data for the next steps. Section 4 describes how and why an automatic segmentation was done. The HMM-based training and generation is presented in Section 5. In Section 6, the post-processing step to add eye-blinks is detailed. Sections 7 explains how the rendered videos were evaluated as well as the results of this evaluation. Finally, Section 8 concludes on the work and proposes directions for further improvements.

2. Related works

As visual laughter, which is the main matter of this work, and visual speech are produced by the same vocal apparatus, we provide a brief overview of the visual speech generation background.

Starting from rule-based systems (Cohen and Massaro, 1993), a variety of other systems following diverse approaches have been proposed. Video-based systems (Bregler et al., 1997; Ezzat et al., 2002) are based on the general idea to assemble video frames taken from a database of a speaking person. By doing this, a new video of that person speaking a new utterance is created.

Data-driven systems (Theobald et al., 2004; Deng and Neumann, 2006; Bailly et al., 2003) control parameters that deform a 3D object and build models of facial deformation. For HMM-based approaches, two main research directions may be found in the literature. The first category gathers image-based systems where the features are directly derived from video frames and the aim is to produce a realistic face (Sako et al., 2000; Wang et al., 2011). The second category gathers motion capture based systems where the features are derived from landmarks on the face that are tracked over time (Schabus et al., 2013; Masuko et al., 1998; Tamura et al., 1998; Govokhina et al., 2007; Bailly et al., 2009). The latter has the advantage to allow to drive any 3D face from the generated trajectories. The present work is based on a motion capture approach.

The goal of audiovisual laughter synthesis is to generate an audio waveform of laughter as well as its corresponding facial animation sequence. This paper focuses on the generation of the visual trajectories corresponding to the facial and head motion during laughter. An overview of studies related to visual laughter generation is given below.

In the last twenty years progresses have been mainly made in laughter detection (e.g. Petridis and Pantic, 2011; Petridis et al., 2013; Scherer et al., 2012; Kumar et al., 2009) and acoustic laughter generation (e.g. Urbain et al., 2013a; Oh and Wang, 2013; Cagampan et al., 2013; Thati et al., 2013; Sundaram and Narayanan, 2007; Beller, 2009; Lasarcyk and Trouvain, 2007). However, attempts to visual laughter generation, which is the main scope of this paper, remain scarce. There are only a few attempts to perform visual laughter generation in the literature.

In 2008, a parametric physical chest model which could be animated from laughter audio signals was proposed by DiLorenzo et al. (2008). The model is able to produce realistic upper body animation but facial animation is not addressed. DiLorenzo et al. propose a method that generates an animation from a sound track of an individual laughing. They also tried to apply their model to other non-verbal signals such as coughing and sneezing. For rendering, they developed a rigging tool within the DreamWorks Animations production pipeline.

In 2009, Cosker and Edge (2009) studied non-speech articulations including laughing, crying, sneezing and yawning. They explored the possible mapping between facial expressions and their related audio signals. HMM were used to model the audio-visual correlation. As for DiLorenzo et al., the produced animation is audio-driven. Cosker et al. used PCA in their parametrization process. The data used in that study consists of a few utterances (6 to 10) acted by 4 subjects.

In 2012, Niewiadomski and Pelachaud (2012) proposed models for different intensities of laughter and for the respiration behaviour during laughter. They used data-driven methods combined with a high-level animation control. Their approach for different laughter intensities generation was to apply a specific motion modulation for each available facial landmark.

In 2013, two studies (Urbain et al., 2013b; Niewiadomski et al., 2013) included the use of laughter capable agents for human-machine interactions. Two different agents animated from recorded data are proposed. One of them is the Greta Realizer (Niewiadomski et al., 2009) which takes as controls either high level commands using the FACS or low level commands using FAP. The other agent is the Living Actor [1] which plays a set of manually drawn animations.

The next year, in 2014, Niewiadomski et al. published a study on the rythmic body movements of laughter (Niewiadomski et al., 2014). They analysed body motion capture data and reconstructed it with appropriate harmonics. These harmonics were then reduced to a two-dimensional space which represents the inputs of the model that

[1] http://www.livingactor.com/Euro/fr/ .

generates continuous laughter body movements. Two types of movements were involved ; torso leaning and shoulder vibration.

The same year, Ding et al. (2014b) proposed a laughter animation synthesis system that generates face and body motions. Their model takes as input a sequence of phones of laughter along with their duration. Lip and jaw movements were further driven by laughter prosodic features. The proposed generator had 3 separate modules which handled the synthesis of specific body parts : 1) lip and jaw , 2) head and eyebrow and 3) torso and shoulder.

Still in 2014, Ding et al. (2014a) published another work in which they studied a different way to synthesize the head and torso motions of a laughing character. The movements are represented by 3 rotation angles for head and torso (6 angles in total). They investigated the use of HMM with 3 different structures ; 1) loop HMM (LHMM) for simple one dimensional shaking-like movement modeling, 2) Transition Parametrized Loop HMM (TPLHMM) in which they introduce the influence of speech on motion and 3) Coupled TPLHMM in which they take into account the dependencies between the parameters defining the head and torso movements. They conducted objective and subjective tests and among the 3 tested HMM structures, the Coupled modeling received globally the best scores. However, compared to human motion data, results of the synthesis are significantly lower.

In 2015, Niewiadomski and Pelachaud (2015) studied the effect of adding wrinkles on the perception of facial actions and full-face expressions of laughter. In their study, they selected 6 episodes of laughter facial expressions from a female subject of the AVLC database. They applied these facial expressions to a 3D agent and conducted a perception test to evaluate possible influences. They used animation with and without wrinkles and with different levels of intensity. The results showed that the addition of wrinkles increased the perceived intensity and amusement of laughter but not the perceived naturalness.

Still in 2015, Ding and Pelachaud (2015) published a study on the lip animation synthesis in which they also included laughter along with speech. They focused on building a real-time system consisting of a unified statistical model to infer speech and laughter lip shape from speech/laughter text information. They evaluated the results by calculating the difference between the generated lip trajectories and the original ones. They compared the results with a linear regression approach for synthesis. The results of their evaluation led to conclude that the proposed statistical model outperforms the linear regression one. They also stated that the laughter lip animation using their method is not as accurate as for speech lip animation.

3. Data

The database used in this work is the AVLASYN DB (Çakmak et al., 2014). It contains 250 utterances of laughter, for a total amount of 48 minutes. The corpus includes motion capture recordings. Thirty-three markers on the face (see Fig. 3 for marker positions) of one male subject (French speaker) were tracked and their 3D coordinates are available at a framerate of 100 fps. This represents 99 dimensions. In addition, 6 values represent the head motion by the means of 3 spatial coordinates as well as 3 Euler rotation angles. Synchronized audio is also available.

A few transformations were performed on that data before training. For each of the recording sessions, the neutral face (one frame where there are no facial deformation) was subtracted from all other frames so that the remaining data consists in deformation deltas related to the neutral position. Proceeding this way will help avoiding morphological differences when retargeting on a 3D face. The neutral face subtraction step is performed on the 99 dimensions of facial data as well as the 3 translation values on the head data but not on the Euler rotation angles of the head motion since the rotation values may be applied directly on a 3D face model without having to deal with morphology differences.

Since the human face has strong deformation constraints, one can expect that the motion of the different parts of the face are highly correlated. This assumption was already verified

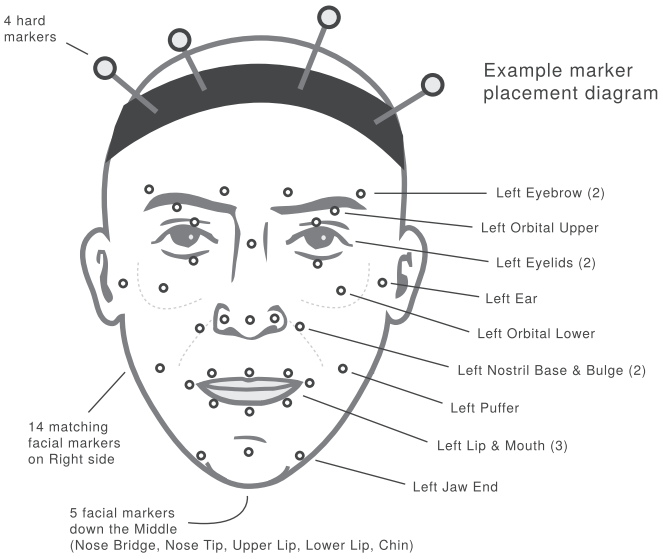


Fig. 3. Reference positions for the markers on the face of the subject. Diagram from Naturalpoint (2013).

in Cakmak et al. (2014) for the same laughter data and similar conclusions were shown for speech as well in Schabus et al. (2013). In this work, we apply the Principal Components Analysis (PCA) (Jollife, 2002) independently on the 99 dimensions of the face data and on the 6 dimensions of the head data.

Fig. 4 shows the cumulated percentage of variability of the facial motion as a function of the number on Principal Components kept (only the first 15 components are shown). We can see that with the first component only we already have almost 75% of the variability. With the first four components we reach 92% and with ten components we are above 97%. Fig. 5 displays the evolution of the Root Mean Square Error (RMSE) of reconstruction computed with Eq. (1) as a function of the number of Principal Components kept.

$$RMSE(k) = \sqrt{\frac{1}{99 \cdot n} \sum_{i=1}^{n} \sum_{j=1}^{99} (M_{ij} - M_{REC,k,ij})^2} \quad (1)$$

Where

- M is the matrix containing the original data (each row represents a frame and each column one dimension)
- $M_{REC,\,k}$ is the matrix containing the data after applying the PCA and reconstructing the original data with k components
- n is the number of frames in the data, namely rows in M and M_{REC}
- i and j are indices of the row and column considered in the matrices

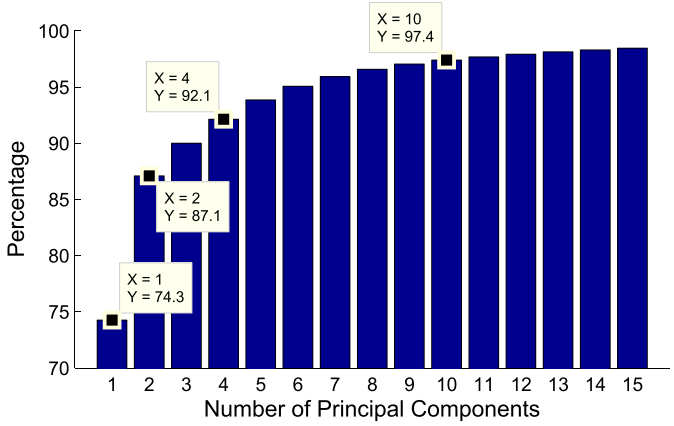


Fig. 4. Cumulated percentage of variability of the facial deformation data as a function of the number of principal components considered.

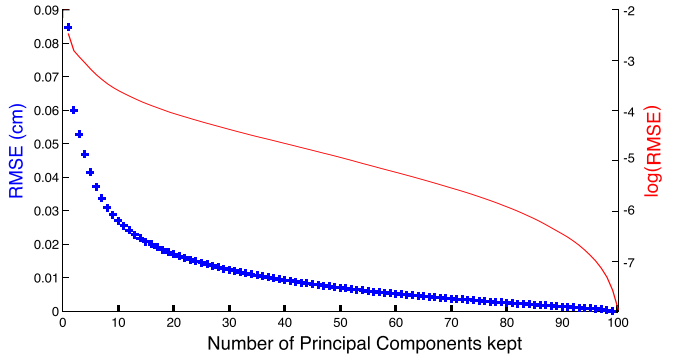


Fig. 5. RMS error of data reconstruction as a function of the number of principal components considered.

Since our final goal is retargeting this data on a 3D face model, we have conducted an informal perceptive test where videos were rendered with incremental numbers of components kept and people were asked to tell at which video they do not see changes compared to the previous one. Out of the ten people doing the test, one judged the number of components to keep as 2, seven voted for 4 components and two for 5 components. We therefore decided to keep 4 components for facial deformation data.

A similar analysis is made on the 6 dimensions of the head motion and as it may be seen on Fig. 6, keeping the first 4 components allows to cover almost 99% of the variability in the data. Prior to the PCA, the range of the head rotations is normalized to match the range of the head translations: the maximum range (maximum value minus minimum value) of translations is compared to the maximum range of rotations and the rotation values are scaled once and for all. This is to avoid performing a PCA on features that have different ranges due to their difference of nature. Another approach might be to perform the PCA separately on translation and rotations but in our application, it is relevant to keep them together since they are correlated and performing the PCA on the six dimensions will ensure the resulting features to be uncorrelated which might not be the case if separate PCA are performed on translations and rotations (cf Section 5).

To analyze the influence of each principal component independently, we have applied the inverse PCA transformation by keeping only the analyzed feature.

Fig. 7 gives the contribution for each of the first four principal components. For each , four images corresponding to the two extreme values that the principal component can take as well as two intermediate values are displayed. Also, Fig. 8 gives the variability of each marker when the data is reconstructed with one PC only. The variability of each marker is computed as follows :

1. Firstly, the norm of each marker in the XYZ reconstructed space is computed at each frame, using only one PC.

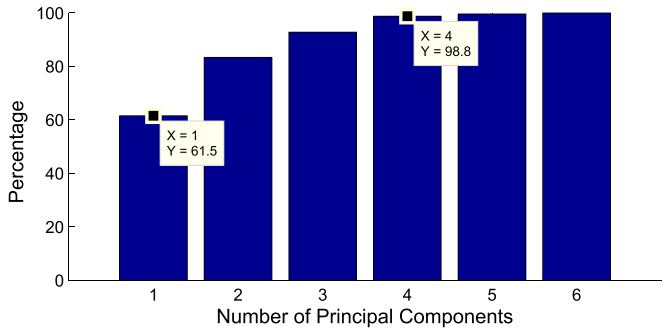


Fig. 6. Cumulated percentage of variability of the head motion data as a function of the number of principal components considered.

2. Secondly, for each marker the std of the norm is calculated.
3. Finally, *round*(100**std*) is plotted on a figure to see which markers' motion are related to the considered PC. The markers have 3 different colors on the figure depending on how much they move relatively to the motion range ; if the markers' motion std is located in the interval [*min*; *min* + 0.2**range*], the marker is colored in green. If it is located in the interval [*min* + 0.2**range*; *min* + 0.8**range*], the marker is colored in blue. Finally, if it is located in the interval [*min* + 0.8**range*; *max*], the marker is colored in red. See Fig. 8.

Conclusions from these figures are summarized in Table 1.

4. Segmentation

HMM-based training is a process which needs labels (annotations) in addition to the data itself. In our application, labels are text files containing the sequences of classes present in a given laughter occurrencez as well as the temporal boundaries of the classes.

4.1. Facial motion

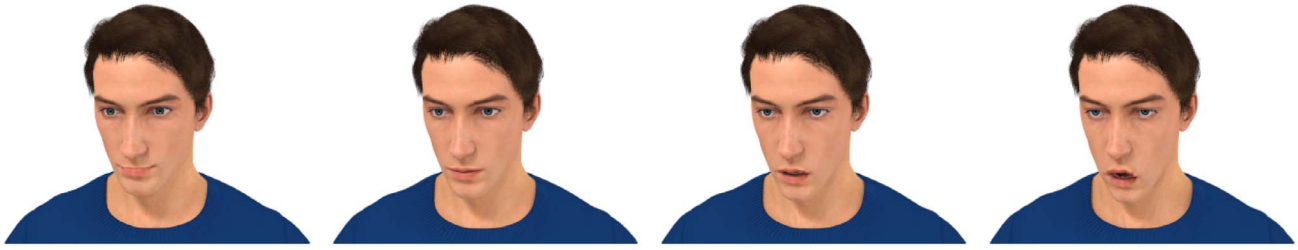
A recent work (Urbain et al., 2014a) focused on the acoustic synthesis of laughter using an HMM-based approach such as in the present work. The used labels were phonetic classes mostly based on speech phonetics (e.g., fricatives, vowels, ...). As part of a preliminary study, we modeled the facial expressions using these phonetic classes. A similar approach has recently been applied for visual speech generation with decent results for the mouth motion only (Schabus et al., 2012,Schabus et al., 2013). Unfortunately, the correlation between facial deformation and the utterance's phonetics is not as important in laughter as it is in speech due to the inarticulate nature of laughter (Ruch and Ekman, 2001). Indeed, the generated trajectories were erratic and noisy and the generated trajectories meant nothing in terms of natural motion. This confirmed that new visual labels have to be defined. To define new visual classes, we manually annotated a small subset of the data. Three classes were considered: Neutral, Smile and Laugh. As their names suggest, the three classes are related to facial pose. The criterion used to delimit the transition from one class to the other was based on the shape of the mouth. If the mouth is closed and neutral then the chosen class was *Neutral*. At the frame where a smile appears on the mouth, which is still closed, we switch to the *Smile* class. Finally, if the mouth opens it is considered as the transition to the *Laugh* class. The new labels built using this protocol on a subset of the available data were used to train HMMs. The generated trajectories were smooth and when the resulting 3D point cloud animation representing each marker was played, it looked like a laughing face.

Since the manual annotation of each file is a time-consuming task, an automatic clustering approach was implemented. A GMM-based clustering approach (McLachlan and Peel, 2000) was used. Since the idea was to replicate the manual process done before and automate it, we trained GMMs on the data corresponding to the visual cues that we considered while annotating manually (shape and position of the mouth). We used the first two PCs for GMM-based clustering as they are closer to the visual cues used for manual annotation (cf Table 1).

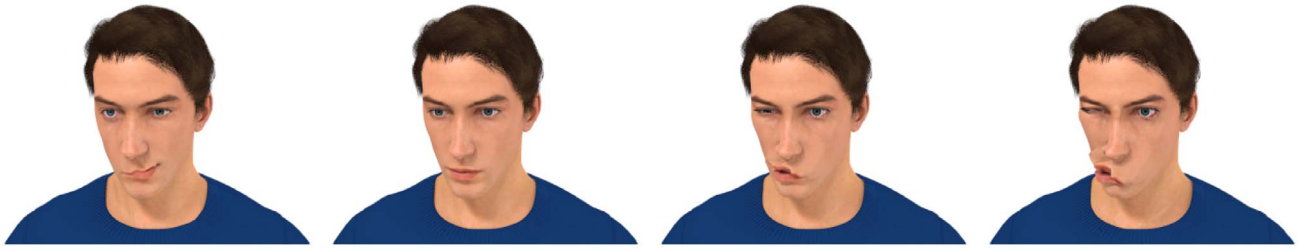
Automatic segmentation is performed as follows:

1. All the facial laughter segments data are concatenated (n_i by 99 matrices concatenated to build an overall N by 99 matrix M_{99} where $N = \sum_i n_i = 288,900$ frames and where n_i denotes the number of frames in the laughter utterance number i).
2. A PCA analysis is done on M_{99} and the first two components are kept so that we end up with a N by 2 matrix $M_{PCA,\ 2}$.
3. GMM fitting is applied on $M_{PCA,\ 2}$ to cluster frames into three classes.
4. Label files are built from this classification by grouping all the successive frames that were classified as being in the same cluster.

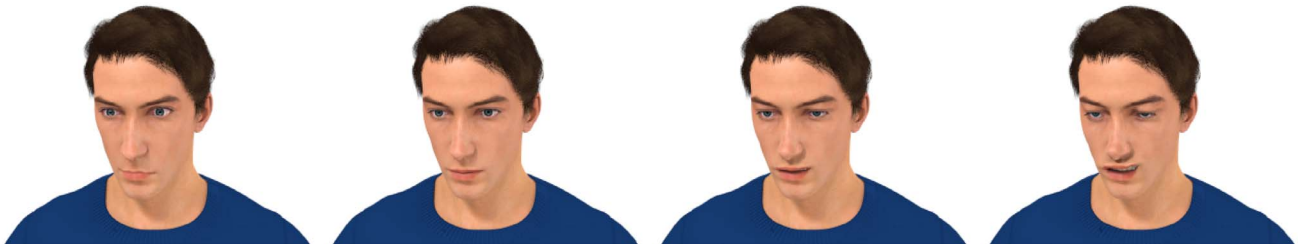
(a) First PC only. This is the global deformation of the face during laughter.



(b) Second PC only. The motion is mainly on the jaw and thus on the mouth opening.



(c) Third PC only. The deformation is asymmetric and mainly on the right half of the face.



(d) Fourth PC only. Also an asymmetric contribution more on the left half of the face and slighter than PC_3.

Fig. 7. Representation of the contribution to the facial deformation of each of the first four principal components

This method has shown to segment appropriately the laughs in the sense that the classes are consistent between segmented files and the fact that we notice a posteriori that this segmentation is well fitted for HMM modeling. The most common sequences being A-B-A, A, C-B-C and C. The B class corresponds to the laughter onset while the A class is the neutral pose. The C class is close to the A class (neutral) but has slight variations compared to A. C might be considered as a “shaky neutral pose” or as an A class with more variance in its trajectories. Fig. 9 shows an example of the A-B-A and C-B-C segmentation.

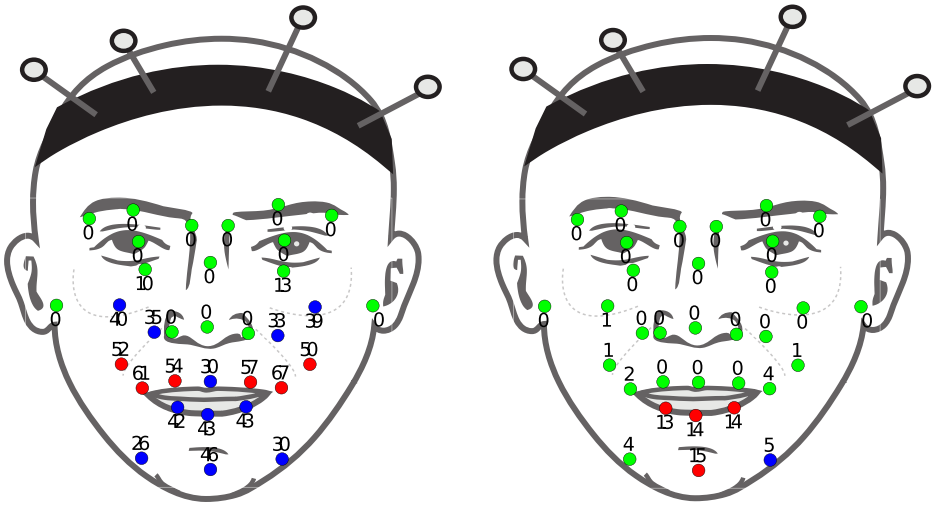
Among the 250 files, 230 are correctly annotated by this method. This represents 92% of our data. The remaining 8% (20 files) correspond mostly to longer utterances.

The whole segmentation of these files are not erroneous. Indeed, many parts of these long utterances are well segmented and it seems possible to improve the segmentation by applying simple post-processing rules. For example, we have noticed that a recurrent error in the segmentation is characterized by too short classes. They represent 7 files among the 20 files that are not perfectly annotated. With a simple constraint on the minimum length of classes, we increase the segmentation accuracy from 92% to 94.8%. Therefore, about 5% of the files are not completely well annotated and none of them are part of the most common sequences in the corpus.

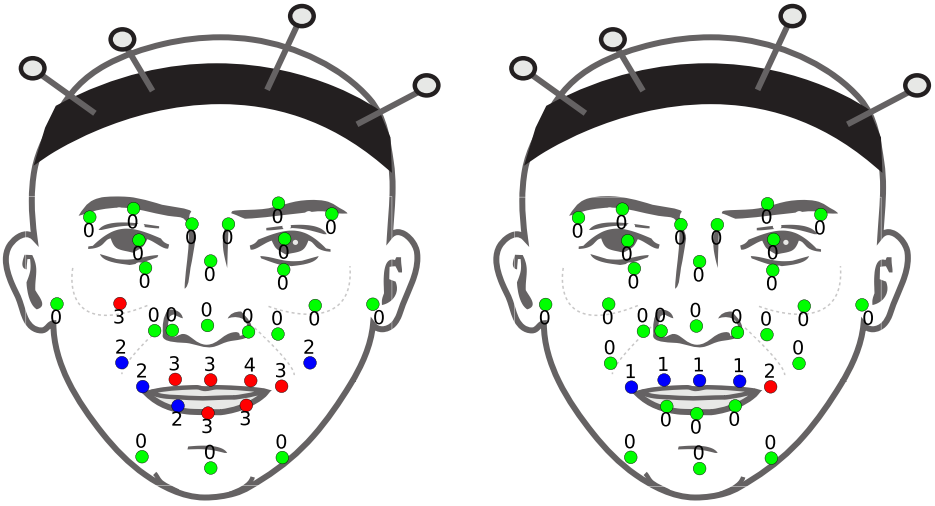
Table 2 gives the number of occurrences of each sequence in the database. Less frequent sequences were gathered in the “Other” category. Since HMM modeling needs a significant number of occurrences of each class to train relevant models, sequences with less than 10 occurrences (that are in the “Other” class) were not considered for modeling.

4.2. Head motion

Natural head motion is an important part of realistic facial animation. Among existing approaches are rule-based methods (Cassell et al., 1994; Pelachaud et al., 1996) that produce head motion (nodding) from labeled text by pre-defined rules. A different approach was proposed by Graf et al. (2002). They estimate the conditional probability distribution of the most significant head movements based on their collected head motion data. They use pitch accents occurrences in the estimation. Concatenative approaches were proposed by Chuang and Bregler (2002) and Deng et al. (2004). More recent studies from



(a) PC_1 per marker variability (b) PC_2 per marker variability



(c) PC_3 per marker variability (d) PC_4 per marker variability

Fig. 8. Variability of each marker when the motion is reconstructed only with one PC. Markers have 3 different colors depending on how much they move relatively to the range of motion for the considered PC. The least moving markers are in *green*, markers with an intermediate variability are in *blue* and the most moving markers are in *red*.

Table 1
Comments on the single PC reconstruction images from Figs. 7 and 8.

Figures	PC	Comments
7(a) and 8(a)	PC_1	The first PC represents the global deformation of the face during laughter. In Fig. 7(a), at the extreme value on the right, the eyes are closed and the shape of the mouth is clearly in a laughing shape while on the other extreme on the left the eyes are excessively opened and the mouth is excessively closed. In Fig. 8, we can see that all the markers related to the lower part of the face are moving as well as the marker around the cheeks and also lower eyelids. This confirms the hypothesis that the first PC is related to the overall laughter facial expression.
7(b) and 8(b)	PC_2	The contribution of the second PC is more related to the jaw motion and hence to the mouth opening. This is also verified in Fig. 8 where we see the motion is mainly located on the lower lip and jaw, hence the mouth opening.
7(c) and 8(c)	PC_3	The third PC has an asymmetric contribution to the facial deformation with some motion on the right cheek. It also appears that it has an impact on the details of the mouth motion.
7(d) and 8(d)	PC_4	The fourth PC has an effect on the upper lip with a slightly more important contribution on the left mouth corner.

Busso et al. (2005, 2007) proposed an HMM-based framework to generate head motion directly from acoustic features.

In this paper, we followed an HMM-based approach but the head motion is not derived from acoustic features in contrast with previously cited works. We are extracting PCA components, which are used as features, from our head motion data as it is explained in Section 3. These features, along with their delta and delta-delta, are then modeled using HMMs following the same framework as for the facial data but with different annotations.

In addition to the head going backwards during the laugh, the head also exhibits a shaking motion (Ruch and Ekman, 2001). A new training scheme is thus necessary to model the head motion more accurately. This means that a specific labeling for the head motion is needed as well. To do so, we have studied the six dimensions of the head motion

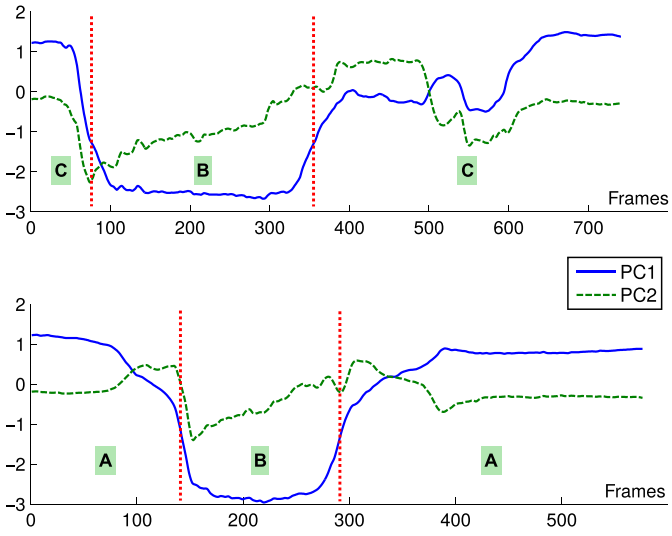


Fig. 9. Example of a segmentation of the A-B-A and C-B-C sequences for face motion.

Table 2
Number of occurrences of each sequence.

A	36
A-B-A	124
C	21
C-B-C	23
Other	46
Total	250

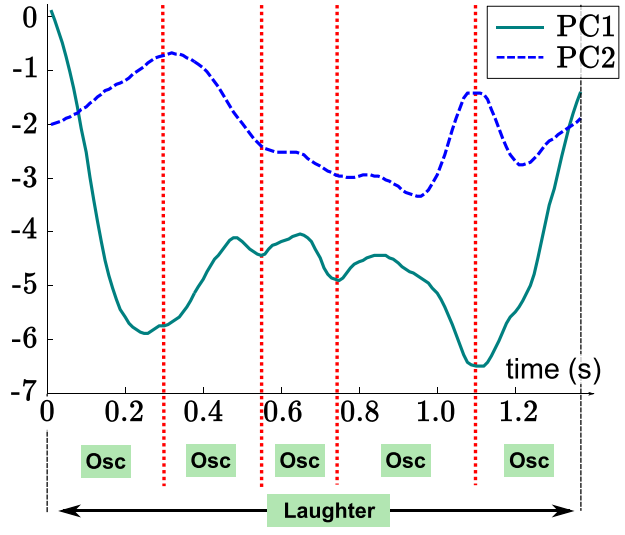


Fig. 10. Example of oscillations on the first two principal components of the head motion. Vertical (red) lines show the boundaries of each oscillation that occur within a laughter event. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

data. We also studied the PCA components of the head motion data and it appeared that the relevant oscillations are quite clearly distinguishable on the first two principal components as it may be seen on Fig. 10. We have chosen to model the head motion as the repetition of a shaking pattern by considering that one occurrence of the pattern is one period of the oscillation. All the files as those on Fig. 10 were annotated by putting boundaries at the local minima of the PC1 (corresponding to local maxima on PC2). This annotation process was done only on data

Table 3
Number of occurrences (N_{occ}) of laughs with N_{osc} oscillations.

N_{osc}	N_{occ}
1	13
2	24
3	23
4	39
≥ 5	164

segments corresponding to laughter (Cluster B). For the case of head motion in this particular work, it is assumed that parts of data not corresponding to laughter do not contain head oscillations (which is true in our data) and are therefore annotated as "Neutral". In other words, the head oscillation class is defined as the portion between two successive maxima on PC1 and PC2 withing a laughter event and everything else is tagged as the Neutral class. Those annotations were then transformed into HTK-compatible label files to use in the training stage.

Therefore, only two classes are used for the head motion transcriptions. The neutral class and the oscillation class. The number of HMMs that are trained is however higher as contextual HMMs are trained. Contextual information contains the two adjacent gestural units (classes) before and after the considered class. That way, the model for an oscillation is different according to its position in the laughter.

Table 3 gives the number of occurrences (N_{occ}) of laughs that contain a certain amount of oscillations (N_{osc}). Values of N_{osc} greater or equal to 5 are grouped as they will be considered as the same context in the training stage.

At the generation stage, the input of the head motion generation system is the number of oscillations wanted in the laugh.

Duration of each class may be given as input or the system can estimate them from the duration models built during the training phase.

5. HMM-based visual laughter generation

5.1. Pipeline

To train the HMM models, we have used HTS version 2.2 (Zen et al., 2011) which is a set of tools released as a patch to HTK (HMM toolkit) (Young et al., 2006). Context-dependent, five-state, single stream, Hidden Markov Models (HMMs) were trained. Separate models were trained for facial deformation and for head motion. Fig. 11 gives an overview of the overall pipeline. The HMM-based motion trajectories generation has already been tested in previous studies (Schabus et al., 2013,Tilmanne et al., 2012). Out of the 147 files of the type C-B-C and A-B-A (cf Table 2), 19 files were randomly chosen as the test base and were not included in the files used for training the models.

5.2. HMM definition

The HMM definition is done with a finite number of parameters. These parameters are a set of transition probabilities (HMM topology), the form of the distribution of the observation in each state and the number of states in the HMM. The next paragraphs explain our decisions on these three parameters, respectively.

HMM topology: We have opted for a topology called left-to-right HMMs with no state skips nor back transitions which means that from a given state the only possible transitions are going to the succeeding one or staying in the same state. The chosen topology should be able to appropriately model our three classes A, B and C (cf Section 4). For classes A and C, the choice of the topology is not crucial since these classes refer to small deformations on the face or even no deformation at all. In contrast, for the class B which is the most relevant event in our

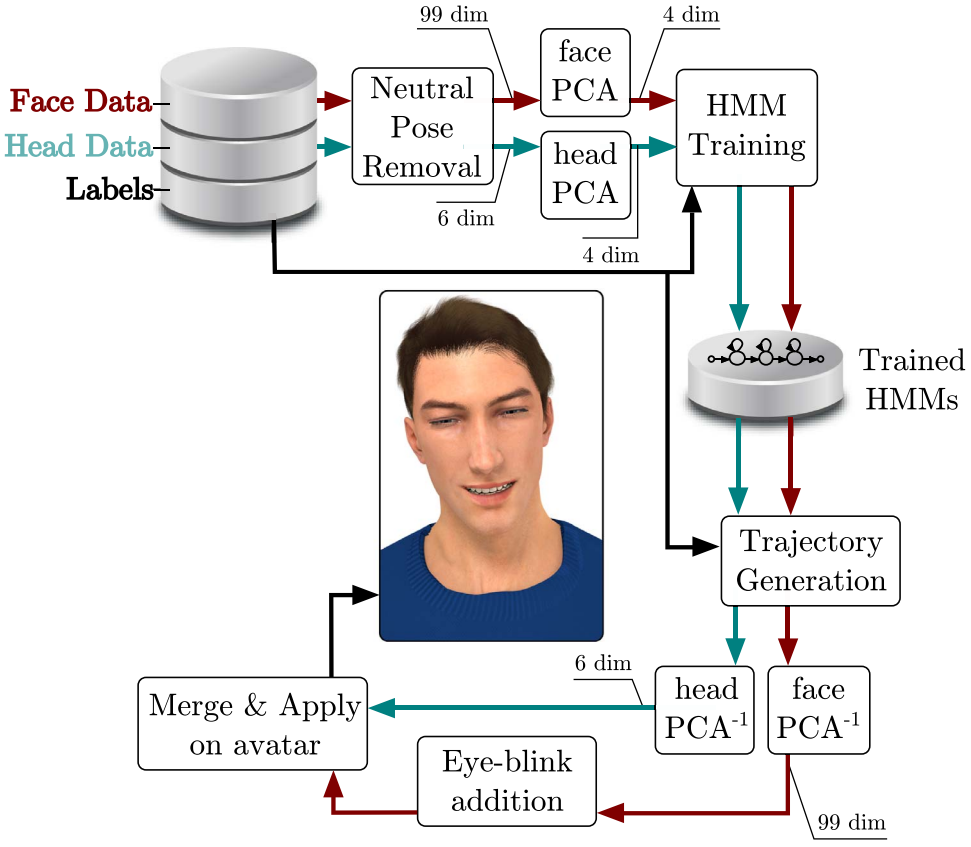


Fig. 11. Overview of the speaker-dependent visual laughter generation system.

application, the appropriateness of the topology is of importance. The left-to-right HMM is relevant for motion modeling since a motion sequence is the succession of basic steps that are recurrent from one occurrence to the other. In our application, the facial deformation during laughter always begins with a rather neutral pose and when laughter occurs the face changes gradually to reach a "laughing" pose then remains in this pose for some time and finally comes back to a pose close to the neutral one. Durations of each step (neutral-to-laughter, laughter and laughter-to-neutral) may differ but these basic steps are always present and in the same order even for a smile which may be considered as a visual laugh of small intensity. Therefore a left-to-right topology for the HMM that will model these motion appears to be appropriate.

Observation probability specification: The features that we will model are continuous parameters as they are derived from continuous trajectories. They can therefore be modeled with Gaussian Mixture densities as done in this work.

Number of states for facial deformation data modeling: In our application, we want to build a model of facial deformation trajectories. Increasing the number of states of the HMM might increase the accuracy of the models but a too large number of states will lead to overfitting to the training data and decrease the quality. An optimal number of states should thus be determined. Since our final product is a rendered movie of a laughing 3D face model, we have trained models for all facial classes with different number of states and rendered the resulting trajectories in videos to compare them visually. It was asked to ten people to choose their preferred videos. No additional information was given to evaluators. They could watch videos in the wanted order and as many times as desired. Five different models with increasing numbers of states (from three to seven) have been trained. People were given the videos and simply asked to tell which one they prefer with no constraint on the order of visioning. Out of the ten people who analyzed the rendered videos, seven have concluded that there is no difference while three expressed a preference for the video related to 5-state models. Fig. 12 shows the generated trajectories of the first two PCs from models built with three, five and seven states. As it may be noticed, the differences are small. The most significant difference is on the first PC (Fig. 12(a)) between the case with three states and the case with five states. The difference between five states and seven states is almost inexistent for PC_1 and quite small on PC_2 (Fig. 12(b)). Based on the previous considerations, a 5-state topology has been chosen to model facial deformation.

Number of states for head motion data modeling: A similar protocol has been followed for head motion data modeling. Models with increasing numbers of states from three to seven have been trained and used for head motion generation. The resulting trajectories have then been applied on the 3D face model and presented to ten people who were asked to tell their preference, if any. Out of the ten people evaluating the videos, six did not perceive differences in the videos, one preferred the 3-state case and three persons preferred the 7-state case. Based on these evaluations. The 7-state topology was chosen for the head motion data modeling.

5.3. Features

The features trained here are, in the case of facial deformation data, the first 4 Principal Components of the PCA applied to the 99-dimensional space. In the case of head motion data, we also used the first 4 components of the PCA applied to the 6-dimensional space (cf Section 3). Facial deformation models and head motion models are trained following the same pipeline but separately. In each case, all features are augmented by their first and second order dynamic features (Δ and Δ^2). The models are clustered using decision trees to build contextual models based on questions on the neighborhood of each class. For facial data, the directly neighboring classes are taken into account in the clustering process (cf Table 2). For head data, the preceding two and following two classes are considered.

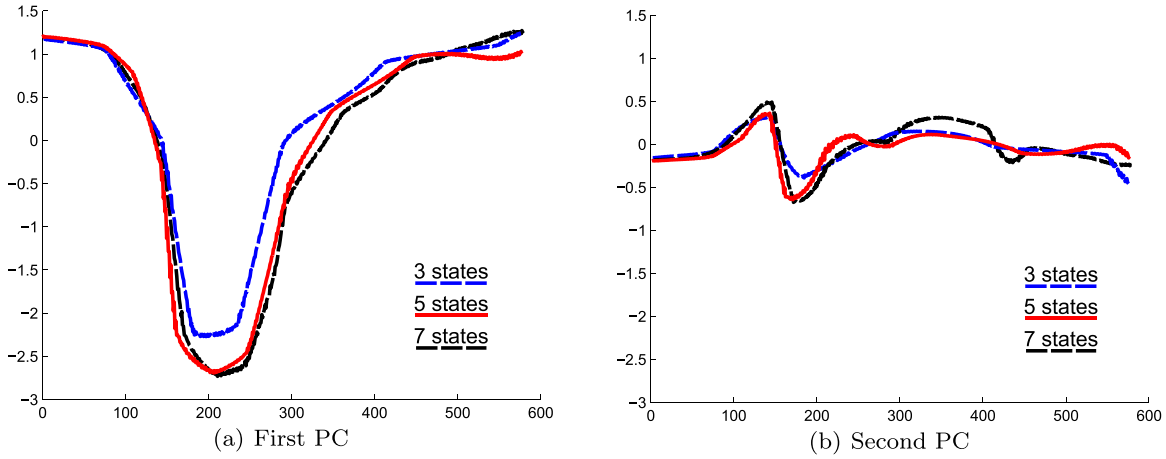


Fig. 12. Comparison of generated trajectories for the cases with 3 (blue dashed), 5 (red continuous) and 7 (black dashed) states. Value on X axis are visual frames of 10ms and on Y axis the amplitude of the generated trajectories for facial deformation data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As mentioned in Section 3, having uncorrelated features has an advantage in the training process. Indeed, if we assume the features to be uncorrelated, which is the case for principal components, then working with diagonal covariance matrices for Gaussian Mixtures is a good approximation. This alleviates computation since for each observation variable we have separate Gaussians i.e. only two parameters have to be estimated, namely the mean and variance of the Gaussian for each observation variable. If we did not make the PCA to work with uncorrelated features and to reduce the dimensionality, we would have had 105 features to model instead of 8 and we could not ignore the non-diagonal elements of the covariance matrix since some features are strongly correlated due to strong constraints in the human face deformations. This would have lead to 105 by 105 covariance matrices which represent 11,025 elements to estimate instead of 8 elements as done in this work. Another important advantage of reducing the total amount of parameters to estimate is that we have quite limited training data and the more parameters to estimate the more data we need to build accurate models. Therefore reducing the number of parameters to estimate helps building more accurate models given the fixed amount of training data available.

6. Adding eye blinks

Eye-blinks have an important contribution to the naturalness of the rendered 3D face model, in particular during idle phases where there is almost no facial deformation (Trutoiu et al., 2011). In Trutoiu et al. (2011), Trutoiu et al. show that animation generated from actual human data are better rated than conventional animation techniques which may suffer from incorrect assumptions such as the symmetry of blinks. Other studies include speech-driven synthesis where eyelid motion is derived from speech signals such as in Dziemianko et al. (2009) and Le et al. (2012). Methods that include the use of the gaze information also exist as in Le et al. (2012) and Ma and Deng (2009).

Although the eyeblink generation method proposed here is based on important conclusions from related works, it is important to mention that this work does not aim at making a comparison with other methods. The main scope of this work is the facial and head motion generation during laughter. However, participants to previous tests suggested that the addition of eye-blinks would increase the perceived quality of the resulting animations. This is why a simple method to add blinks has been developed to assess the possible effect of eye-blink presence in the rendered animations.

In this work, eye-blinks are added as a post-processing step on the generated eyelid trajectories. Three main parameters are used to model eye-blinking:

1. The shape of the blinking motion
2. The time delay between blinks
3. The duration of the blinking motion

6.1. Determining the shape of the blink

In this work, eye-blinks are represented by the evolution of the distance between upper and lower eyelids during the eye-blinking events. By studying these distance trajectories, the eye-blinks are annotated manually. A generic eye-blink shape (Fig. 13) is built from these distance trajectories extracted from the database as it is shown in Trutoiu et al. (2011) that real blinking trajectories are better perceived than conventional animation techniques. They are first resampled to fit the average eye-blink length. Then the mean eye-blink trajectory is calculated. Finally, the amplitude of the mean trajectory is normalized to be between 0 and 1 so that it represents a factor that will modulate the distance between upper and lower eyelids.

When adding an eye-blink, the distance between upper and lower eyelid trajectories is multiplied frame by frame by the generic eye-blink shape of Fig. 13 so that when its value is 1, the upper eyelid position remains unchanged and the eye is opened and when it is 0, the upper eyelid position is the same as the position of the lower eyelid and the eye is closed.

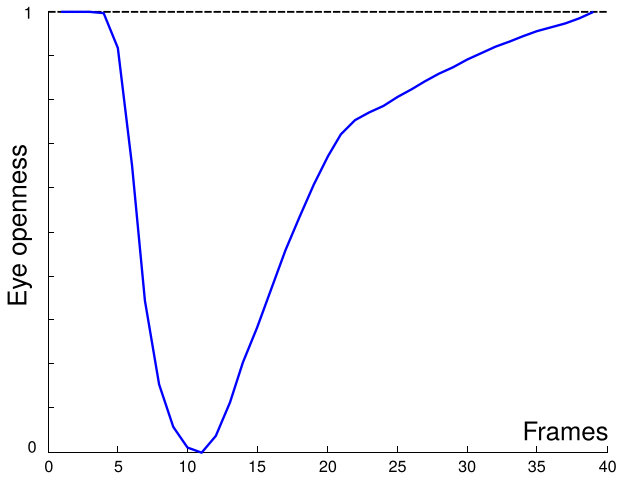


Fig. 13. Generic eye-blink shape used in the model (1 frame = 10 ms).

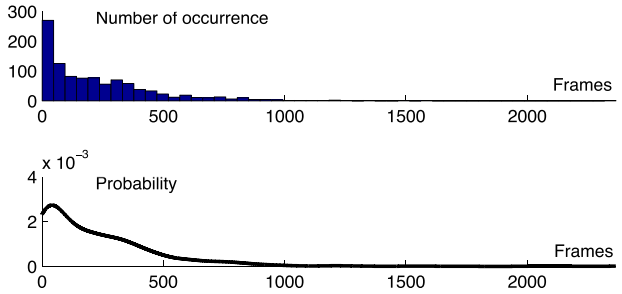


Fig. 14. Histogram of the original data (up) and fitted probability distribution (down) of time between two occurrences of eye-blinking (1 frame = 10 ms).

6.2. Determining the timing of eye-blinks

The positions where eye-blinks are added are determined from the probability distribution of the time delay between two successive eye-blinks. The corresponding probability density is built by fitting on the actual eye-blink delays in the database. For the fitting process, kernel density estimation which is a non-parametric way to estimate the probability density function of a random variable was used (Bowman and Azzalini, 1997; Johnson et al., 1995). Random time delays that follow the fitted distribution are then generated to determine the timing of the eye-blinks in the laugh. Fig. 14 displays the fitted distribution as well as the histogram of the actual data on which the fitting processes are performed.

6.3. Determining the duration of each eye-blink

A similar approach is used to define the duration of the eye-blink itself. A probability distribution is fitted on the eye-blink durations data (cf Fig. 15) and a random duration based on the fitted distribution is generated for each eye-blink.

This duration is used to modify the generic shape in length by re-sampling. This simple approach allows us to add eye-blinks on generated trajectories with consistent delays and durations while ensuring some randomness to avoid monotony both in the time between eye-blinks and in the duration of the eye-blinks themselves.

7. Evaluation

Even if the best way of demonstrating the methods would be to generate random laugh sequences with estimated durations, we have chosen to generate visual laughs corresponding to a test subset of the available database. Two reasons motivated this choice:

1. To have a more relevant comparison with the original facial expression from the database (included in the test as Method 1) and with the previous method (included in the test as Method 2).
2. To be able to add an audio track to the video. Indeed, if we do not take the original sequences and durations as in the database, we need to generate the corresponding audio as well, which is out of the scope of this paper. Even if the facial models presented in this paper are capable of generating different sequences and to estimate the durations, we have generated the facial expression corresponding to the available audio track for the commodity of the test and for the sake of comparison between methods.

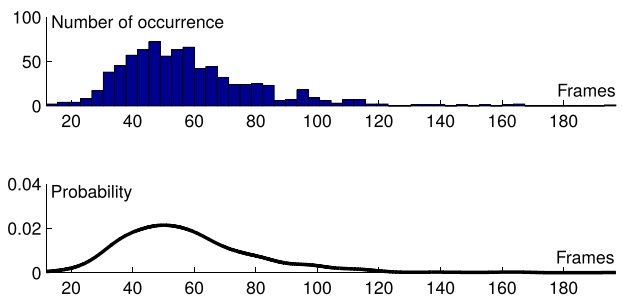


Fig. 15. Histogram of the original data (up) and fitted probability distribution (down) of the duration of eye-blinks (1 frame = 10 ms).

To evaluate the system proposed in this work, an online test was built. As explained in Section 5.1, the evaluation was performed on a test set of 19 files not used in the training process. Four different methods were rendered and included in the test. The four different cases are summarized in Table 4. Method 1 is simply the recorded motion trajectories applied on the 3D face model. Including the original data will serve as the ground truth for comparison with generated trajectories. Method 2 consists in generated trajectories using the system implemented in Cakmak et al. (2014). The third and fourth methods include the developments added in the present work. Method 3 is rendered with the system where the head motion is modeled separately from the facial deformations. Method 4 relies on Method 3 with the addition of eye-blinks. For the three generated methods (Methods 2 to 4), the same 19 transcription files were used as input to the respective systems.

In all the methods, the corresponding original audio files were added to the final videos [2]. Original class durations available in transcriptions were imposed to ensure synchronization with the audio tracks.

7.1. Online MOS test

An online Mean Opinion Score (MOS) test was used. Prior to the evaluation, explanations were given to the visitor regarding the test itself. Then the evaluation began and at each step, one randomly chosen video was shown. The user could play the video as many times as wished until (s)he chooses to go to the next video. The random picking of video had the constraint to avoid previously shown videos for a given user. For each video, two questions were asked to the user. The first question is related to the quality of the animation and the second to the appropriateness of the visual animation given the audio signal. For each question, the user was asked to give a rating on a 5-point scale (0-very poor to 4-excellent). A free comment area was also available at the end of the evaluation for the participants wishing to leave a comment.

Seventy-two participants (48 males and 24 females, aged 19 to 68 with mean age 34) did the evaluation. Seventy of them completed the test fully by evaluating 20 videos and two of them stopped the test when they had evaluated 11 and 13 videos respectively. A total amount of 1424 evaluations were given.

7.2. Results

Table 5 gives the mean scores as well as the corresponding standard errors for each question and each type of video (method). The score distribution for question 1 and question 2 may be found in Figs. 16 and17 respectively.

To assess the statistical significance of pairwise comparisons between the results by method, the Tukey's Honest Significant Difference test with a confidence level of 95% is used. As we can see on Tables 6 and 7, there are only significant differences between Method 2 and all the other for both questions. Non-parametric statistics (Mu et al., 2012) are also used to analyze the results. Ordered logistic regression using the logit and probit cumulative link models are used to study the influence of the methods on the given scores. In both cases, the only significant method is Method 2 with a confidence level of at least 99% for both questions.

[2] Sample videos at http://tcts.fpms.ac.be/~cakmak/personal/?page_id=202.

Table 4
Methods tested in the evaluation.

Headline	Method 1	Method 2	Method 3	Method 4
Motion data	Original data	Generated from Cakmak et al. (2014)	Generated with separate head models	Method 3 + eye-blinks
Features	99 facial translations + 3 head translations + 3 head rotations	**4 PCs** from 102 dim (99 facial translations + 3 head translations) **3** head rotations	**4 PCs** from 99 facial translations **4 PCs** from 6 head dimensions (3 head translations + 3 head rotations)	Cf Method 3
Transcriptions	/	Facial deformation-based	Facial deformation-based for facial data Head oscillations-based for head data	Cf Method 3
Eye-blinks	Original data	No	No	Yes (proposed method)

Table 5
Mean scores and standard errors for each method.

Method	Av. Score Q1 (Quality)	AV. Score Q2 (AV appropriateness)
Method 1	3.01 (std err. = 0.052)	3.00 (0.048)
Method 2	2.84 (0.044)	2.59 (0.050)
Method 3	3.07 (0.042)	2.96 (0.044)
Method 4	3.12 (0.052)	3.00 (0.062)

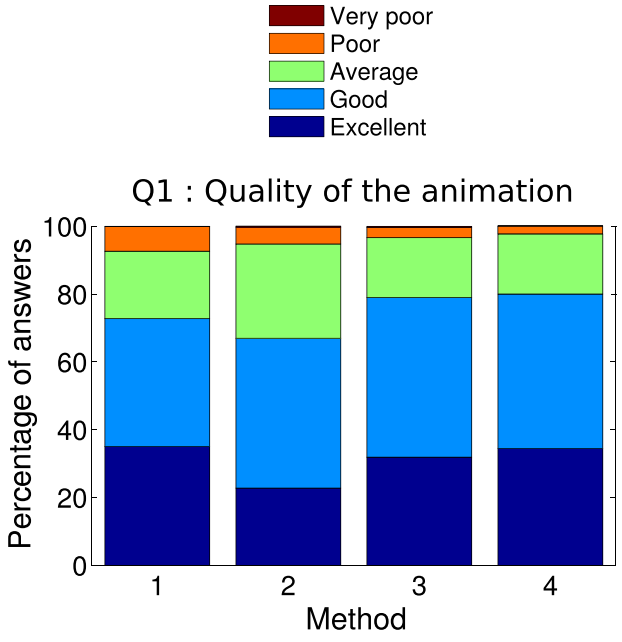


Fig. 16. Score distribution for each method obtained for question 1.

It is to notice that the original trajectories (Method 1) did not receive significantly better scores than the best generation methods. This is a good result that shows that the perceived quality between recorded data and generated data is not significantly different in our specific setup at least.

Among the tested generation methods, Method 2 has the worst results. The gap with other methods is even larger for question 2 which was related to the appropriateness of the animation considering the audio signal. While the mean of other methods are similar to their respective mean scores for question 1, the mean score of Method 2 for question 2 (2.59) is lower than for question 1 (2.84). This shows that the proposed methods in this work are perceived as better matching the audio signal than for Method 2.

These results show that the methods presented in this work are better than the previous one and also that the perceived quality of the new methods is comparable to the original trajectories. However, based on these results only, the addition of eye-blinks does not seem to significantly improve the animations even though the obtained mean scores for Method 4 (including eye-blinks) are slightly better than those of Method 3 (without eye-blinks).

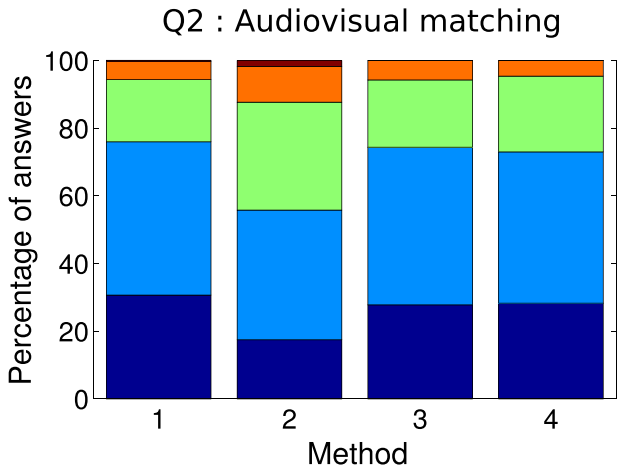


Fig. 17. Score distribution for each method obtained for question 2 (Cf legend of Fig. 16).

Table 6
Pairwise p-values between the generation methods for question Q1 (Quality of the animation).

Method	1	2	3	4
1	–	**0.05**	0.72	0.39
2	**0.05**	–	**0.00**	**0.00**
3	0.72	**0.00**	–	0.90
4	0.39	**0.00**	0.90	–

Table 7
Pairwise p-values between the generation methods for question Q2 (Matching between audio and animation).

Method	1	2	3	4
1	–	**0.00**	0.93	1.00
2	**0.00**	–	**0.00**	**0.00**
3	0.93	**0.00**	–	0.98
4	1.00	**0.00**	0.98	–

To verify the influence of eye-blinks, an additional test was conducted to compare Method 3 and Method 4 by building a Comparative Mean Opinion Score (CMOS) test in which participants were presented 10 pairs of videos. Each pair consists in the same laugh generated with Method 3 and Method 4. The pairs were randomly picked. Method 3 and 4 were presented side by side for the picked video and the participant was asked to choose which one (s)he prefers. Participants were not told that the only difference between videos was the presence of eye-blinks nor that there were only two kinds of videos. The position of

Table 8
Frequency of each answer given in the CMOS test. Positive values are in favor of Method 4.

Score	Frequency	Percentage	Cumulated percentage
−3	0	0.00	0.00
−2	3	2.31	2.31
−1	7	5.38	7.69
0	36	27.69	35.38
1	46	35.38	70.77
2	36	27.69	98.46
3	2	1.54	100.00
Total	130	100.00	100.00

methods (on the left or on the right) was also randomly chosen for each new pair of videos to compare. The participants were asked to give their preference on a 7-point scale. The 7-point scale goes from -3 to 3 with step 1 and the value 0 means that no difference is seen. A positive value means that the video with blinks was preferred and a negative value indicates that the video without blinks was preferred. Thirteen participants took this comparison test and the results show that there is a slight preference for the videos with eye-blinks. The overall mean score is located in the interval [0.77;1.11] with a confidence level of 95%. After the test, participants were asked if they think that eye-blinks has a positive effect on the quality and all participants answered that yes. As it may be seen in Table 8, out of the 130 evaluations given in total, 84 (64.61%) were in favor of the Method 4, 36 (27.69%) were 0 and only 10 (7.69%) were in favor of Method 3.

7.3. Discussion

The fact that original trajectories (Method 1) did not receive significantly better results than the generated ones (Methods 3 and 4) may possibly be due to two factors. Firstly the original trajectories may contain some local artifacts like markers oscillating locally or bumping because of a loss of tracking for a few frames. Even if post-processing has been done on the recorded data to correct these kind of artifacts, a few may still exist since it is not realistically feasible to verify manually the whole set of thousands of frames recorded for each of the 105 trajectories. These artifacts are not present in the generated trajectories. Indeed the tracking errors are local and therefore are smoothed during the HMM training stage. Secondly, the re-hydrating motion which is due to the presence of markers on the lips of the subject is not in the generated trajectories neither since they were not annotated as relevant information in the first place. Indeed, we have seen that this motion does not appear when the same subject is laughing without markers on his lips which means that this is an artifact related to the recording protocol and not to laughter itself. These artifacts may have influenced the raters negatively for some of the videos of the Method 1.

Another possible factor that influenced the overall results is the fact that we included audio in the evaluation videos. It is possible that the presence of audio tends to flatten the differences between the methods. It would be interesting to investigate if we obtain bigger differences between methods if we compare them without audio. However, the obtained results in this paper are still consistent as the audio was the same for all methods for a given animation. Moreover, applications in which laughter would be presented without audio should be very rare in real life scenarios.

8. Conclusion

In this paper, a visual laughter generation system was proposed based on HMM modeling of 3D motion capture data. A principal component analysis was conducted on the facial deformation data and on the head motion data separately. The facial deformation and the head motion were then modeled separately with their specific annotations. The separate modeling between audio and visual data was necessary because the phonetic transcriptions were not appropriate for visual laughter modeling. Indeed our attempts to model visual laughter using phonetic transcriptions resulted in chaotic generated trajectories. We therefore modeled visual data with new specific annotations. Since annotating manually is a highly time-consuming process, we have developed an automatic GMM-based annotation process. This work shows that modeling head motion separately improves the perceived quality of the animations. Head motion has its own annotations as well but they are dependent on facial annotations and are based on positions of local minima on the first principal component. The head annotations may therefore easily be created automatically (although the local minima were annotated manually in the present work). It is important to note that the method used for head motion presented in this paper has proved to be relevant in the case of our data but it would need more investigation to be able to confirm that the method is relevant for other subjects.

From the models, new trajectories were generated and post-processed to add eye-blinks before application on a 3D face model. Finally, videos were rendered and an online perceptive evaluation was conducted. The results showed an improvement in terms of mean scores given by raters compared to a joint modeling of head motion and facial deformation. An additional comparative test showed that the addition of eye-blinks is slightly preferred compared to the case without blinking.

This work showed that it is possible to build models that can generate perceptually acceptable face and head motion animations for laughter using HMMs. The current system already allows us to impose the durations of the gestural units during laughter. Following the same pipeline, different laughter types may also be modeled if the corresponding corpus is available.

Future works include the development of synchronization rules between the audio and the visual laughter so that this work can be used to generate a laughter animation that matches a given acoustic laughter signal. A preliminary work towards this objective have been presented in Çakmak et al. (2015b) and Çakmak et al. (2015a). In the present work, to ensure time synchronization between audio and visual animation in the videos, we have used the original recorded audio files with corresponding visual transcriptions to generate facial and head trajectories. However the visual laughter generation system presented in this work is not limited to these corpus-specific durations. Any duration may be generated for facial laughter animation but the system is currently lacking a synchronization module to match any given audio laughter file.

The integration of laughter intensity (i.e. arousal) would also be an important improvement of the current work. It could be used as a driving signal for the synthesizer to build intensity-driven visual laughter synthesis. This was introduced in Urbain et al. (2014b) for acoustic laughter synthesis and El Haddad et al. (2015) for amused speech synthesis. Intensity could be used as a feature to build intensity-dependent models as well. This would give an additional control parameter to generate more specific laughs. If the intensity data is available (i.e the information about how intense a given laugh is), we can easily duplicate the visual classes in as many intensity levels as available. For example, if we could classify every laugh in 3 intensity categories, then we can build new annotations (on top of the existing ones without changing how the current annotations are done) by duplicating the “Laughter” class into 3 classes : “Laughter_level1”, “Laughter_level2”, “Laughter_level3” where the “*_levelx” specifies the intensity of the laugh. Doing so, different models would be built depending on the intensity which would directly be taken into account by the labels/classes in the first place. Once these intensity-dependent models are available, we could easily apply interpolation between HMM models to generate intermediate intensity levels. The exact same approach can be followed to produce different types of laughter. The only needed resource to do it with the current system is the availability of such data.

Reactive visual laughter generation is another field to investigate. Reactive generation is characterized by the ability to change the input of the generation system on the fly and to see the effect on the output immediately. Reactive generation was made possible for HMM-based speech synthesis with the mage project (Astrinaki et al., 2013). Its extrapolation to acoustic and visual laughter were already tested in d'Alessandro et al. (2014) and gave promising results. Combining this latter work with the present work would be a significant contribution to the visual laughter generation field.

An other important perspective is the integration of acoustic laughter models (acoustic laughter generation) in the system rather than using original audio tracks. There is still room for improvements on visual laughter models as well. In their current state, visual models do not take into account subtleties such as head motion during inhalation periods, which are characterized by the rise and strengthening of the trunk and therefore have an impact on head motion.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.specom.2017.12.006.

References

Astrinaki, M., Moinet, A., Dutoit, T., Reboursière, L., 2013. Mage 2.0: new features and its application in the development of a talking guitar.

Bailly, G., Bérar, M., Elisei, F., Odisio, M., 2003. Audiovisual speech synthesis. Int. J Speech Technol 6 (4).

Bailly, G., Govokhina, O., Elisei, F., Breton, G., 2009. Lip-synching using speaker-specific articulation, shape and appearance models. EURASIP J. Aud. Speech Music Process. 2009.

Beller, G., 2009. Analyse et Modèle Génératif de l'expressivité: Application à la Parole et à l'Interprétation Musicale. Université Pierre et Marie Curie-Paris VI Ph.D. thesis.

Bowman, A., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis : The Kernel Approach with S-Plus Illustrations: The Kernel Approach with S-Plus Illustrations. OUP Oxford.

Bregler, C., Covell, M., Slaney, M., 1997. Video rewrite: driving visual speech with audio. Proc. of the 24th Annual Conf. on Computer Graphics and Interactive Techniques.

Busso, C., Deng, Z., Grimm, M., Neumann, U., Narayanan, S., 2007. Rigid head motion in expressive speech animation: analysis and synthesis. IEEE Trans. Audio Speech Lang. Process. 15 (3), 1075–1086.

Busso, C., Deng, Z., Neumann, U., Narayanan, S., 2005. Natural head motion synthesis driven by acoustic prosodic features. J. Vis. Comput. Animat. 16 (3–4), 283–290.

Cagampan, B., Ng, H., Panuelos, K., Uy, K., Cu, J., Suarez, M., 2013. An exploratory study on naturalistic laughter synthesis. Proceedings of the 4th International Workshop on Empathic Computing (IWEC'13). Beijing, China.

Çakmak, H., El Haddad, K., Dutoit, T., 2015. Gmm-based synchronization rules for hmm-based audio-visual laughter synthesis. Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on. IEEE, pp. 428–434.

Çakmak, H., Urbain, J., Dutoit, T., 2015. Synchronization rules for hmm-based audio-visual laughter synthesis. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2304–2308.

Cakmak, H., Urbain, J., Tilmanne, J., Dutoit, T., 2014. Evaluation of hmm-based visual laughter synthesis. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4578–4582.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., Stone, M., 1994. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques. ACM, New York, NY, USA, pp. 413–420. http://dx.doi.org/10.1145/192161.192272.

Çakmak, H., Urbain, J., Dutoit, T., 2014. The AV-LASYN database : a synchronous corpus of audio and 3d facial marker data for audio-visual laughter synthesis. Proc. of the 9th Int. Conf. on Language Resources and Evaluation (LREC'14).

Chuang, E., Bregler, C., 2002. Performance Driven Facial Animation using Blendshape Interpolation. 2 (2), 3.

Cohen, M., Massaro, D., 1993. Modeling coarticulation in synthetic visual speech. Models and Techniques in Computer Animation. Springer-Verlag, pp. 139–156.

Cosker, D., Edge, J., 2009. Laughing, crying, sneezing and yawning: automatic voice driven animation of non-speech articulations. Computer Animation and Social Agents (CASA).

d'Alessandro, N., Tilmanne, J., Astrinaki, M., Hueber, T., Dall, R., Ravet, T., Moinet, A., Cakmak, H., Babacan, O., Barbulescu, A., Parfait, V., Huguenin, V., Kalayc, E.S., Hu, Q., 2014. Reactive statistical mapping: towards the sketching of performative control with data. In: Rybarczyk, Y., Cardoso, T., Rosas, J., Camarinha-Matos, L. (Eds.), Innovative and Creative Developments in Multimodal Interaction Systems. IFIP Advances in Information and Communication Technology 425. Springer Berlin Heidelberg, pp. 20–49. http://dx.doi.org/10.1007/978-3-642-55143-7_2.

Deng, Z., Narayanan, S., Busso, C., Neumann, U., 2004. Audio-based head motion synthesis for avatar-based telepresence systems. Proceedings of the 2004 ACM SIGMM Workshop on Effective Telepresence. ACM, pp. 24–30.

Deng, Z., Neumann, U., 2006. eFASE: expressive facial animation synthesis and editing with phoneme-isomap controls. Symposium on Computer Animation.

DiLorenzo, P., Zordan, V., Sanders, B., 2008. Laughing out loud: control for modeling anatomically inspired laughter using audio. ACM Trans. Graph.

Ding, Y., Huang, J., Fourati, N., Artieres, T., Pelachaud, C., 2014. Upper body animation synthesis for a laughing character. Intelligent Virtual Agents. Springer, pp. 164–173.

Ding, Y., Pelachaud, C., 2015. Lip animation synthesis: a unified framework for speaking and laughing virtual agent.

Ding, Y., Prepin, K., Huang, J., Pelachaud, C., Artières, T., 2014. Laughter animation synthesis. Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems. International Foundation for Autonomous Agents and Multiagent Systems, pp. 773–780.

Dziemianko, M., Hofer, G., Shimodaira, H., 2009. HMM-based automatic eye-blink synthesis from speech. INTERSPEECH. ISCA, pp. 1799–1802.

El Haddad, K., Moinet, A., Dutoit, T., et al., 2015. An hmm approach for synthesizing amused speech with a controllable intensity of smile. 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). IEEE, pp. 7–11.

Ezzat, T., Geiger, G., Poggio, T., 2002. Trainable videorealistic speech animation. Proc. of the 29th Annual Conf. on Computer Graphics and Interactive Techniques.

Glenn, P.J., 2003. Laughter in Interaction. Cambridge University Press.

Govokhina, O., Bailly, G., Breton, G., et al., 2007. Learning optimal audiovisual phasing for a HMM-based control model for facial animation. 6th ISCA Workshop on Speech Synthesis (SSW6).

Graf, H.P., Cosatto, E., Strom, V., Huang, F.J., 2002. Visual prosody: facial movements accompanying speech. Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on. IEEE, pp. 396–401.

Johnson, N., Kotz, S., Balakrishnan, N., 1995. Continuous Univariate Distributions. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics Wiley & Sons. Number vol. 2

Jollife, I.T., 2002. Principal Component Analysis.

Kumar, K.S., Mallidi, S.H.R., Murty, K.S.R., Yegnanarayana, B., 2009. Analysis of laugh signals for detecting in continuous speech. Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech). Brighton, UK, pp. 1591–1594.

Lasarcyk, E., Trouvain, J., 2007. Imitating conversational laughter with an articulatory speech synthesis. Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter. Saarbrücken, Germany, pp. 43–48.

Le, B.H., Ma, X., Deng, Z., 2012. Live speech driven head-and-eye motion generators. IEEE Trans. Vis. Comput. Graph. 18, 1902–1914. http://dx.doi.org/10.1109/TVCG.2012.74.

Ma, X., Deng, Z., 2009. Natural eye motion synthesis by modeling gaze-head coupling. IEEE, pp. 143–150.

Masuko, T., Kobayashi, T., Tamura, M., Masubuchi, J., Tokuda, K., 1998. Text-to-visual speech synthesis based on parameter generation from HMM. Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on. 6.

McLachlan, G., Peel, D., 2000. Finite Mixture Models.

Mu, M., Mauthe, A., Tyson, G., Cerqueira, E., 2012. Statistical analysis of ordinal user opinion scores. Consumer Communications and Networking Conference (CCNC), 2012 IEEE. IEEE, pp. 331–336.

Naturalpoint, 2013. Optitrack.

Niewiadomski, R., Bevacqua, E., Mancini, M., Pelachaud, C., 2009. Greta: an interactive expressive eca system. Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2. International Foundation for Autonomous Agents and Multiagent Systems, pp. 1399–1400.

Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., Dupont, S., Geist, M., Lingenfelser, F., McKeown, G., Pietquin, O., Ruch, W., 2013. Laugh-aware virtual agent and its impact on user amusement. Proc. Int. Conf. on Autonomous Agents and Multi-Agent Systems.

Niewiadomski, R., Mancini, M., Ding, Y., Pelachaud, C., Volpe, G., 2014. Rhythmic body movements of laughter. Proceedings of the 16th International Conference on Multimodal Interaction. ACM, pp. 299–306.

Niewiadomski, R., Pelachaud, C., 2012. Towards multimodal expression of laughter. Intelligent Virtual Agents. Springer, pp. 231–244.

Niewiadomski, R., Pelachaud, C., 2015. The effect of wrinkles, presentation mode, and intensity on the perception of facial actions and full-face expressions of laughter. ACM Trans. Appl. Percept. 12 (1), 2.

Oh, J., Wang, G., 2013. Lolol: laugh out loud on laptop. Proceedings of the International Conference on New Interfaces for Musical Expression. Graduate School of Culture Technology, KAIST, Daejeon, Republic of Korea, pp. 190–195.

Pelachaud, C., Badler, N.I., Steedman, M., 1996. Generating facial expressions for speech. Cogn. Sci. 20 (1), 1–46.

Petridis, S., Leveque, M., Pantic, M., 2013. Audiovisual detection of laughter in human-machine interaction. Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, Geneva, Switzerland, pp. 129–134.

Petridis, S., Pantic, M., 2011. Audiovisual discrimination between speech and laughter: why and when visual information might help. IEEE Trans. Multimedia 13 (2), 216–234.

Ruch, W., Ekman, P., 2001. The expressive pattern of laughter. In: Kaszniak, A. (Ed.), Emotion, qualia and consciousness. World Scientific Publishers, pp. 426–443.

Sako, S., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2000. HMM-based text-to-

audio-visual speech synthesis. INTERSPEECH'00. pp. 25–28.
Schabus, D., Pucher, M., Hofer, G., 2012. Speaker-adaptive visual speech synthesis in the HMM-framework. Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH 2012). Portland, OR, USA, pp. 979–982.
Schabus, D., Pucher, M., Hofer, G., 2013. Joint audiovisual hidden semi-markov model-based speech synthesis. Select. Top. Sig. Process. IEEE J.
Scherer, S., Glodek, M., Schwenker, F., Campbell, N., Palm, G., 2012. Spotting laughter in natural multiparty conversations: a comparison of automatic online and offline approaches using audiovisual data. ACM Trans. Interact. Intell. Syst. 2 (1), 4.
Sundaram, S., Narayanan, S., 2007. Automatic acoustic synthesis of human-like laughter. J. Acoust. Soc. Am. 121 (1), 527–535.
Tamura, M., Masuko, T., Kobayashi, T., Tokuda, K., 1998. Visual speech synthesis based on parameter generation from HMM: Speech-driven and text-and-speech-driven approaches. AVSP'98 Int. Conf. on Auditory-Visual Speech Processing.
Thati, S.A., Kumar, S., Yegnanarayana, B., 2013. Synthesis of laughter by modifying excitation characteristics. J. Acoust. Soc. Am. 133 (5), 3072–3082.
Theobald, B.-J., Bangham, J., Matthews, I., Cawley, G., 2004. Near-videorealistic synthetic talking faces: implementation and evaluation. Speech Commun. 44 (1).
Tilmanne, J., Moinet, A., Dutoit, T., 2012. Stylistic gait synthesis based on hidden markov models. EURASIP J. Adv. Sig. Proc. 72.
Trutoiu, L.C., Carter, E.J., Matthews, I., Hodgins, J.K., 2011. Modeling and animating eye blinks. ACM Trans. Appl. Percept. 8 (3).
Urbain, J., Çakmak, H., Charlier, A., Denti, M., Dutoit, T., Dupont, S., 2014. Arousal-driven synthesis of laughter. IEEE J. Sel. Top. Sig. Process 8 (2), 273–284.
Urbain, J., Çakmak, H., Charlier, A., Denti, M., Dutoit, T., Dupont, S., 2014. Arousal-driven synthesis of laughter. IEEE J. Sel. Top. Sig. Process 8, 273–284. http://dx.doi.org/10.1109/JSTSP.2014.2309435.
Urbain, J., Çakmak, H., Dutoit, T., 2013. Evaluation of HMM-based laughter synthesis. Acoustics Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on.
Urbain, J., Niewiadomski, R., Mancini, M., Griffin, H., Çakmak, H., Ach, L., Volpe, G., 2013. Multimodal analysis of laughter for an interactive system. Proceedings of the INTETAIN 2013.
Wang, L., Wu, Y.-J., Zhuang, X., Soong, F., 2011. Synthesizing visual speech trajectory with minimum generation error. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on.
Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2006. The HTK Book Version 3.4. Cambridge University Press.
Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., Tokuda, K., 2011. The HMM-based speech synthesis system (HTS) version 2.2.