EVALUATION OF HMM-BASED VISUAL LAUGHTER SYNTHESIS

Hüseyin Çakmak , Jérôme Urbain , Joëlle Tilmanne, Thierry Dutoit

TCTS lab - University of Mons, Belgium

ABSTRACT

In this paper we apply speaker-dependent training of Hidden Markov Models (HMMs) to audio and visual laughter synthesis separately. The two modalities are synthesized with a forced durations approach and are then combined together to render audio-visual laughter on a 3D avatar. This paper focuses on visual synthesis of laughter and its perceptive evaluation when combined with synthesized audio laughter. Previous work on audio and visual synthesis has been successfully applied to speech. The extrapolation to audio laughter synthesis has already been done. This paper shows that it is possible to extrapolate to visual laughter synthesis as well.

Index Terms**—** Audio, visual, laughter, synthesis, HMM

1. INTRODUCTION

Among features of human interactions, laughter is one of the most significant. It is a way to express our emotions and may even be an answer in some interactions. In the last decades, with the development of human-machine interactions and various progress in speech processing, laughter became a signal that machines should be able to detect, analyze and produce. This work focuses on laughter production and more specifically on visual laughter production. Acoustic synthesis of laughter using Hidden Markov Models (HMMs) has already been addressed in a previous work which is state-of-the-art and served as a basis for acoustic synthesis in this work [1].

The goal of audio-visual laughter synthesis is to generate an audio waveform of laughter as well as its corresponding facial animation sequence. This work follows a separated modeling approach.

Visual laughter synthesis systems are rare. DiLorenzo *et al* [2] proposed a parametric physical chest model which could be animated from laughter audio signals. Face animation was not part of the work. Cosker *et al* [3] studied the possible mapping between facial expressions and their related audio signals for non-speech articulations including laughter. The authors used HMMs to model the audio-visual correlation. As for DiLorenzo *et al*, the animation is audio-driven. More recent studies [4, 5] include the animation of laughter capable avatars in human-machine interaction. The proposed solutions include two different avatars animated from recorded data. One (Greta Realizer) is controlled either through high level commands using Facial Action Coding System (FACS) or low level commands using Facial Animation Parameters (FAPs) of the mpeg-4 standard for facial animation. The other avatar (Living Actor) plays a set of manually drawn animations.

In contrast with these works and following up our previous work on acoustic laughter synthesis, we investigated the extrapolation of HMM-based synthesis to visual laughter. The approach followed in the present work is to model facial expressions by means of facial landmark trajectories. First a 3D facial motion database has been recorded using the OptiTrack [1] motion capture system. Then this data has been modeled using an HMM-based approach. Synthesized trajectories were then retargeted to a 3D model into the MotionBuilder software where the animation was rendered. Results were evaluated through an online Mean Opinion Score (MOS) test where users were asked to rate the overall quality, the human-likeness and spontaneousness for each of the 27 videos presented in the evaluation.

The paper is organized as follows : Section 2 gives an overview on the database built for the purpose of this work, Section 3 explains the laughter synthesis method, Section 4 describes the evaluation and its results and Section 5 concludes and gives an overview of future work.

2. THE AV-LASYN DATABASE

The AV-LASYN Database is a synchronous audio-visual laughter database designed for laughter synthesis. The corpus contains data from one male subject and consists of 251 laughter utterances. Professional audio equipment and a marker-based motion capture system (OptiTrack) have been used for audio and facial expression recordings respectively. Figure 1 gives an overview of the recording pipeline.

The database contains laughter-segmented WAV audio

H. Çakmak receives a Ph.D. grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°270780.

[1] http://www.naturalpoint.com/optitrack

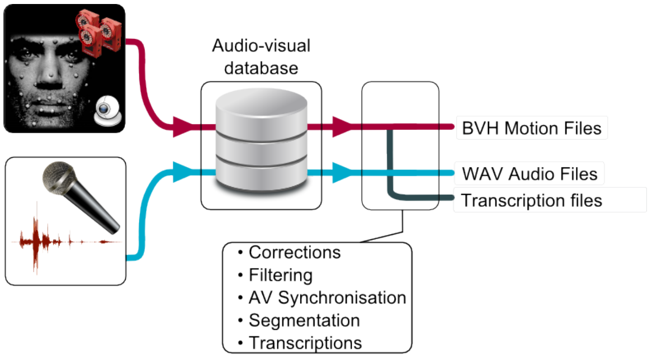

Fig. 1. Data recording pipeline

files and corresponding motion data in the Biovision Hierarchy (BVH) format. Transcription files are also available for the audio modality. Please refer to [6] for more information on transcriptions.

The laughs were triggered by videos found on the web. The subject was free to watch whatever he wanted. A total amount of 125 minutes were watched by the subject to build this corpus. This lead to roughly 48 minutes of visual laughter and 13 minutes of audible laughter.

3. HMM-BASED LAUGHTER SYNTHESIS

HMM-based visual synthesis of laughter is an almost unexplored domain. In contrast, visual speech synthesis is quite well established and different approaches have been developed. Among the existing techniques, we can find rule-based systems [7], video-based systems [8, 9] and data-driven approaches [10, 11, 12]. HMM-based visual speech systems may be split into two groups. On the one hand we have image based systems where features are collected from videos and which aim is to synthesize new video realistic sequences [13]. On the other hand we have motion capture based approaches where the features are coordinates of tracked facial feature points [14, 15]. This work is based on the motion capture data approach.

The work presented here follows a separate modeling approach (see section 3.2.2 for details). This means that audio and motion data are trained separately and then merged together for the final rendering. An overview of the pipeline can be seen in Figure 2.

3.1. Audio features modeling

The audio modeling was done following the same pipeline as in [1]. HMMs were trained with the standard speaker-dependent, left-to-right, 5-state configuration using HTS tools [16, 17]. Window length was set to 25 ms and frameshift to 5 ms. Thirty-five Mel cepstral coefficients (MFCCs) as well as log F0 were used as features. Both MFCCs and F0 were extracted using STRAIGHT tools [18]. At the synthesis stage, the excitation source of the source-filter model was changed using DSM [19] to reduce buzziness in the produced sound. The laughs were synthesized using transcriptions from the training data and the durations were imposed to be the same as in these transcriptions.

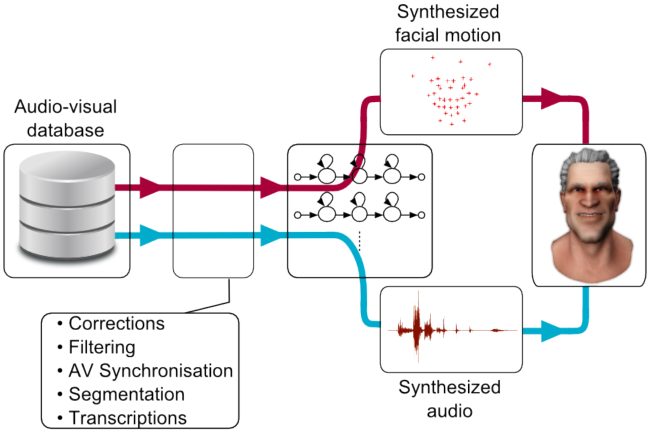

Fig. 2. Overview of the pipeline for HMM-based audio-visual laughter synthesis

3.2. Visual features modeling

3.2.1. Data preparation

For visual features modeling, a similar approach has been used. The data consists in coordinates of each tracked marker on the face i.e. 33 markers times 3 coordinates at 100 Hz. This results in a 99-dimensional space for facial features. To these 99 dimensions, we add 6 mores dimensions to model the head movements (3 translations values xyz and 3 rotations around the same axes).

A step of post-processing has been applied to the visual data. First the head motion was extracted so as to be available independently from the facial deformation data. Then the neutral face was subtracted to keep only the deformation of each facial point relatively to the neutral face. Finally, a PCA [20] has been applied on the translations and the dimensionality was reduced by keeping the first 4 principal components. As shown on Figure 3, these 4 PCs represent more than 97% of the variability in the data. Figure 4 gives the Root Mean Squared Error values in centimeters for the reconstruction of the data from the reduced PCA space as a function of the number of components kept. The RMSE values reported on Figure 4 are computed as a function of the number of components kept k using

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

where n is the number of frames, m is the data dimension, M is the original data and $M_{REC,k}$ is the reconstructed data using k components.

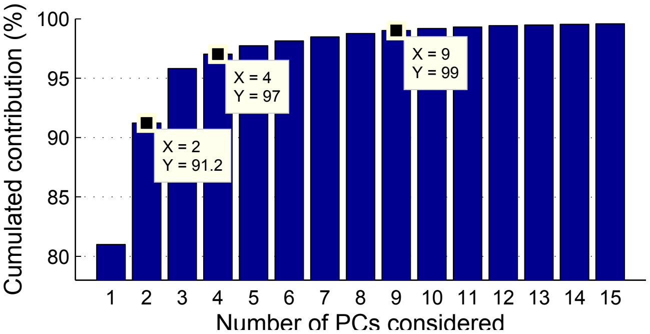

Fig. 3. Cumulated contributions for the first 15 principal components

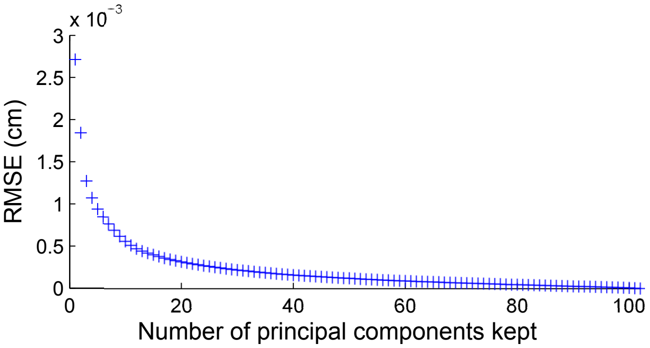

Fig. 4. RMSE of reconstruction as a function of the number of principal components kept

3.2.2. *Visual annotations*

Our first attempt of visual laughter synthesis relied on phonetic annotations as a basis for visual data modeling. Unfortunately, for laughter, temporal segmentation based on phones does not suit facial expression. Indeed, laughter begins to appear on the face before it becomes audible and disappears after the last sounds. Moreover, facial deformation is reduced during laughter compared to speech. In speech, for example, the mouth motion is more directly linked to the emitted phoneme than in laughter. Based on these findings, we annotated visual data separately using a three-class basis. The classes were *Neutral*, *Smile* and *Audible_Laugh*. We first manually annotated a subset of the database and tried to model this subset with HMMs. The outcome was much better than with phonetic annotations. Since manual annotations are time-consuming, we decided to build automatic clusters based on the raw data. This should be possible since there is a distinguishable difference between the facial expression during laughter and during the neutral pose. Indeed, for a common visual laughter sequence like [*Neutral*, *Smile*, *Audible_Laugh*, *Smile*, *Neutral*], the data is quite still during the neutral pose (except for the eyes), then we have a rising transition where facial expression changes to reach a laughing pose and finally the facial expression goes back to a neutral pose after a falling transition. The automatic clustering was performed using a GMM-based approach [21]. All the frames of the corpus were clustered into 3 classes by fitting GMMs on the space reduced by PCA [22].

After the fitting process, each frame is assigned a cluster. Based on these assignments, we generated transcriptions in a HTS compatible format. To see if the clusters correspond to relevant classes in terms of facial deformation, we played the motions with a distinguishable color for each cluster. This analysis showed that the three clusters correspond to a neutral state (N), a laughter with mouth open state (L) and a last state between the two others (B). Concretely, the laughter sequences are most of the time [N, L, N] and [B, L, B]. The third state (B) corresponds roughly to smile, meaning that the mouth is not yet open but that there is some movement compared to the neutral state.

These transcriptions are used in the HMM-based modeling framework implemented in HTS demo scripts. Five-state left-to-right HMMs are trained to represent visual data including facial deformations and head movements. In this work audio and visual modalities are trained and synthesized independently. Each modality has its own transcriptions that are used both for training and for synthesis. At the synthesis stage, we synthesize separately the audio and visual trajectories of the same input laugh (using corresponding audio and visual transcriptions). We ensure that the synthesized trajectories for both modalities are synchronous by forcing durations to the values in the transcriptions. We can thus put the synthesized trajectories back together to form the synthesized audio-visual version of the input laugh. Facial motion trajectories were transformed back to the initial high dimensional space. Finally coordinates of the neutral face are added to the synthesized facial deformations and the result is saved in BVH format.

4. EVALUATION

4.1. Preparing videos

The synchrony between all input files (audio, mocap, transcriptions) enables us to create audiovisual examples with only one synthesized modality (the other modality is taken back from the original data): for example we only synthesize the facial trajectories (with forced durations) and we add the corresponding original (as opposed to synthesized) audio track. In total, 190 laughs were rendered within MotionBuilder. Among the 251 laughter utterances available in the database, we did not include those with no sound at all which correspond to 46 files with only smile as well as 15 files with less than 3 classes in their transcriptions to avoid errors during HMMs training. Motion data and audio were available both as they were originally recorded and as synthesized data. By combining original and synthesized as summarized in Table 1, we end up with 4 categories of videos and a total of 760 videos.

		audio	
		original	synthesized
Visual	original	C1	C2
	synthesized	C3	C4

Table 1. Four different combinations evaluated in the test

4.2. Online MOS test

An online Mean Opinion Score (MOS) test was conducted to evaluate the results. Forty-six participants (33 male, 13 female, aged 15 to 49 with mean age 26.5) evaluated each 27 videos of laughter. The videos presented to users were picked up randomly from the 760 available videos with the constraint that a given video cannot be shown more than once to a participant. Prior to the evaluation, participants were asked to fill a questionnaire to provide us with information such as age, gender, whether they are audio processing experts, image processing experts or working on laughter. They were also asked to use headphones but since it might not be possible for everyone, the ones that did use loudspeakers were asked to specify they did. During the test, for each video, the participants were asked to rate it on a 5-point Likert scale (0-very poor to 4-excellent). Three characteristics were evaluated : overall quality (Q1), human-likeness (Q2) and spontaneousness (Q3). Finally, after the videos, a last questionnaire was filled by participants to give their impressions on some defects that we suspected to cause bad evaluations. A free comment area was also available for participants wishing to express themselves about the videos.

4.3. Results

A total of 1245 evaluations were collected. Figure 5 gives the mean scores for each question and for each combination of audio and visual data (cf Table 1) as well as the corresponding standard errors. It is noticeable in this figure that for a given combination, mean scores are quite similar for each question. To verify this, we calculated the Pearson's correlation coefficient of each pair of questions and obtained high correlation values ($C_{Q1Q2} = 0.89$, $C_{Q1Q3} = 0.82$, $C_{Q2Q3} = 0.80$). Further statistical analysis is needed but this shows that keeping those three questions in future evaluations might not be relevant. To assess statistical significance of pairwise comparisons between the different combinations (C1 to C4), we have used the Tukey HSD test with a confidence level of 99%. The result is that all combinations are significantly different from each other (and this remains true whatever the considered question is). As shown on Figure 5, audio has an important impact on the results. Scores are much higher when audio is original (C1,C3). In a sense, this might mean that the visual modality is not taken into account by participants and that only the audio has an impact on scores. However, we also have a significant difference between C1 et C3, which would not be the case if the visual modality had no impact on scores. This indicates that either the proposed visual synthesis surpasses state-of-the-art acoustic laughter synthesis, or that shortcomings on the visual track are less penalizing than audio defects.

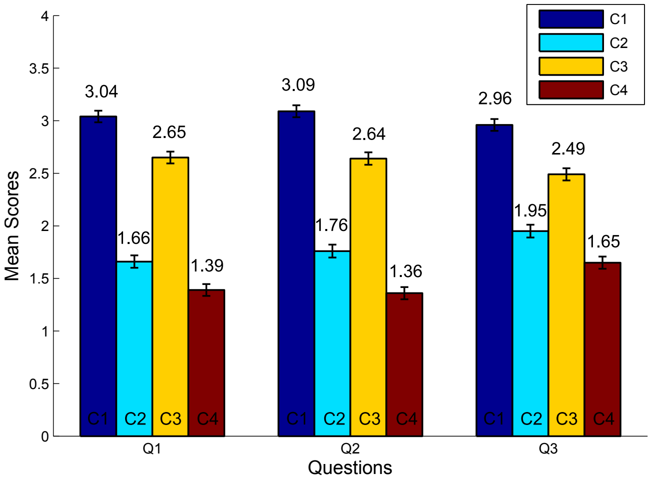

Fig. 5. Mean scores for each combination and each question

5. CONCLUSION AND FUTURE WORKS

In this paper we have presented a way of synthesizing visual laughter based on motion capture data recorded for this purpose. To the best of our knowledge, this paper presents the first attempt to synthesize visual laughter based on motion capture data and following a HMM-based framework for trajectory synthesis. An online MOS test has been conducted to evaluate the results and showed that the presented visual synthesis appears plausible to participants. The evaluation also showed that audio defects have an important impact on the perception of quality. Future works on visual synthesis include improvements by modeling head motion in a more suitable way, adding eye blinking models and going deeper into automatic GMM-based clustering. Improvements on audio laughter synthesis are also planned by investigating new features to better model acoustic laughter and the development of automatic phonetic transcriptions based on HMMs. Finally, joint audiovisual modeling will be investigated to be able to synthesize audiovisual laughter in a unified framework and not separately. As part of joint audiovisual modeling, ways of using duration models to ensure synchronization between audio and visual modalities will be studied.

6. REFERENCES

[1] J. Urbain, H. Cakmak, and T. Dutoit, "Evaluation of HMM-based laughter synthesis," in *Acoustics Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013.

[2] P.C. DiLorenzo, V.B. Zordan, and B.L. Sanders, "Laughing out loud: control for modeling anatomically inspired laughter using audio," *ACM Trans. Graph*, 2008.

[3] D. Cosker and J. Edge, "Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations ," in *Computer Animation and Social Agents (CASA)*, 2009.

[4] J. Urbain, R. Niewiadomski, M. Mancini, H. Griffin, H. Cakmak, L. Ach, and G. Volpe, "Multimodal analysis of laughter for an interactive system," in *Proceedings of the INTETAIN 2013*, 2013.

[5] R. Niewiadomski, J. Hofmann, J. Urbain, T. Platt, J. Wagner, B. Piot, H. Cakmak, S. Pammi, T. Baur, S. Dupont, M. Geist, F. Lingenfelser, G. McKeown, O. Pietquin, and W. Ruch, "Laugh-aware virtual agent and its impact on user amusement," in *Proc. int. conf. on Autonomous agents and multi-agent systems*, 2013, AAMAS.

[6] J. Urbain and T. Dutoit, "A phonetic analysis of natural laughter, for use in automatic laughter processing systems," in *ACII (1)*, Sidney K. D'Mello, Arthur C. Graesser, Björn Schuller, and Jean-Claude Martin, Eds. 2011, vol. 6974 of *Lecture Notes in Computer Science*, pp. 397–406, Springer.

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

[8] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *Proc. of the 24th annual conf. on Computer graphics and interactive techniques*, 1997, SIGGRAPH '97.

[9] T. Ezzat, G. Geiger, and T. Poggio, "Trainable video-realistic speech animation," in *Proc. of the 29th annual conf. on Computer graphics and interactive techniques*, 2002, SIGGRAPH '02.

[10] B-J. Theobald, J.A. Bangham, I.A. Matthews, and G.C. Cawley, "Near-videorealistic synthetic talking faces: Implementation and evaluation," *Speech Communication*, vol. 44, no. 1, 2004.

[11] Z. Deng and U. Neumann, "efase: expressive facial animation synthesis and editing with phoneme-isomap controls," in *Symposium on Computer Animation*, 2006.

[12] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *International Journal of Speech Technology*, vol. 6, no. 4, 2003.

[13] L. Wang, Y-J. Wu, X. Zhuang, and F.K. Soong, "Synthesizing visual speech trajectory with minimum generation error," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011.

[14] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda, "Text-to-visual speech synthesis based on parameter generation from hmm," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, 1998, vol. 6.

[15] D. Schabus, M. Pucher, and G. Hofer, "Joint audio-visual hidden semi-markov model-based speech synthesis," *Selected Topics in Signal Processing, IEEE Journal of*, 2013.

[16] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," 1999.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0," in *Proc. of Sixth ISCA Workshop on Speech Synthesis*, 2007.

[18] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, 2006.

[19] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech*, 2009.

[20] I. T. Jollife, *Principal Component Analysis*, October 2002.

[21] G. McLachlan and D. Peel, *Finite Mixture Models*, 2000.

[22] H. Cakmak, J. Urbain, and T. Dutoit, "The av-lasyn database : A synchronous corpus of audio and 3d facial marker data for audio-visual laughter synthesis," in *Proc. of the 9th Int. Conf. on Language Resources and Evaluation (LREC'14)*, 2014.