SYNCHRONIZATION RULES FOR HMM-BASED AUDIO-VISUAL LAUGHTER SYNTHESIS

Hüseyin Çakmak , Jérôme Urbain , Thierry Dutoit

TCTS lab - University of Mons, Belgium

ABSTRACT

In this paper we propose synchronization rules between acoustic and visual laughter synthesis systems. This work follows up our previous studies on acoustics laughter synthesis and visual laughter synthesis. The need of synchronization rules comes from the constraint that in laughter, HMM-based synthesis of laughter cannot be performed using a unified system where common transcriptions may be used. Therefore acoustic and visual models are trained independently without any synchronization constraints. In this work, we propose simple rules derived from the analysis of audio and visual laughter transcriptions in order to generate visual laughter transcriptions starting from acoustic transcriptions. A perceptive Mean Opinion Score (MOS) test is conducted to evaluate the method.

Index Terms— Audio-visual, laughter, synthesis, HMM, synchronization

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 the synchronization between audio and synthesized visual laughter.

Acoustic synthesis of laughter using Hidden Markov Models (HMMs) has already been addressed in a previous work [1]. To characterize the acoustic laughter, phonetic transcriptions were used and the results outperformed the state of the art. Extensions of the latter work were done to perform automatic phonetic transcriptions [2] and to integrate the arousal in the system [3]. The goal of audio-visual laughter synthesis is to generate an audio waveform of laughter as well as its corresponding facial animation sequence. In statistical data-driven audiovisual speech synthesis, it is common that separate acoustic and visual models are trained [4, 5, 6, 7] sometimes with an additional explicit time difference model for synchronization purposes [8, 9].

In 2014, a visual laughter synthesis system has also been proposed and is the basis of the visual laughter synthesis in this work [10]. The latter work has shown that a separate segmentation of the laughter is needed to correctly model the visual trajectories meaning that phonetic transcriptions are not suited to describe the visual cues for laughter as it has been shown to be feasible for speech [8, 11, 12, 13]. Further developments have shown that the head motion should be modeled separately as well [14]. Modeling independently audio, facial data and head data means having specific transcriptions for each and thus, the need for synchronization arises. In [10], the synchronization between modalities was guaranteed by imposing synthesized durations to be the same as in the database, in which the transcriptions are synchronous in the first place. To bring this to the next level and to be able to synthesize audio-visual laughter with any wanted duration, we derived simple synchronization rules to model the relationships between transcriptions.

The basic principle lying under the proposed method in this work is the study of the relation between the audio and visual transcriptions. Rules are extracted from the study of temporal shift between the beginning of the perceptible visual laughter and the beginning of the audible laughter. Likewise, rules are extracted from the study of temporal shift between the end of visually perceptible laughter and the end of the post-laughter inhalation. This method makes it possible to generate visual transcriptions starting from audio transcriptions. Generated visual transcriptions may then be used to synthesize visual laughter trajectories that are applied to an avatar before final video rendering. An online MOS test is then conducted to rate the quality of the animation and the matching between audio and visual modalities.

The paper is organized as follows : Section 2 gives a brief overview on the database used in this work, Section 3 explains the audio and visual laughter synthesis methods, Section 4 explains the proposed synchronization rules between acoustic and visual synthesis, Section 5 describes the evaluation and Section 6 concludes and gives an overview of future work.

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 EU Seventh Framework Programme (FP7/2007-2013) under grant n°270780.

2. THE AVLASYN DATABASE

The AVLASYN Database [15] used in this work 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 have been used for audio and facial expression recordings respectively. Figure 1 gives an overview of the recording pipeline.

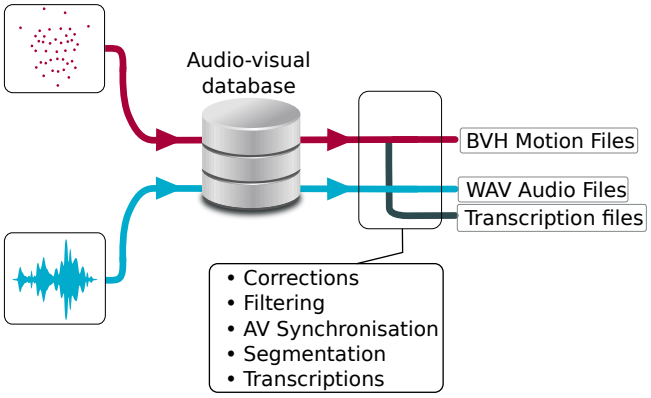

Fig. 1. Data recording pipeline

The database contains laughter-segmented audio files in WAV format and corresponding motion data in the Biovision Hierarchy (BVH) format. A visual segmentation was done on laughter files from which audible parts were phonetically annotated. Please refer to [16] for more information on transcriptions. The laughs were triggered by watching 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 led to roughly 48 minutes of visual laughter and 13 minutes of audible laughter.

3. HMM-BASED LAUGHTER SYNTHESIS

How the models used in this work are built is detailed in [1] for acoustic models and in [10] and [14] for visual models. A brief overview is given below. The HMM-based trajectory synthesis is based on the scripts of a tool called HTS [17].

Figure 2 gives the general pipeline followed to build the models. The main steps that must be introduced to understand the remainder of this paper are :

1. Features are extracted from the database (separate features for audio, face and head movements [10, 14]).
2. The audio features, facial deformation features and head movements features are modeled independently with their respective transcriptions.
3. Once the models are trained for each modality, trajectories are synthesized. For audio synthesis, the duration of each phone may either be estimated by the system or imposed. For visual synthesis, durations are imposed from rules based on the durations of acoustic phones to be synchronized with them.

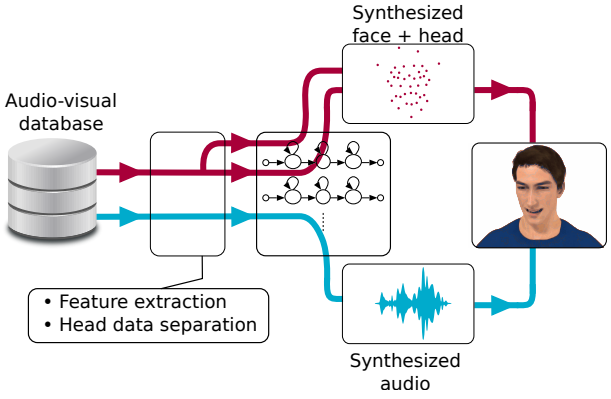

Fig. 2. Overview of the pipeline for HMM-based audio-visual laughter synthesis

The original visual transcriptions are build from an automated Gaussian Mixture Model-based segmentation detailed in [10, 14]. The aim of the present work is to be able to generate new such visual transcriptions that would fit a corresponding acoustic laughter transcription file. This would allow to synthesize visual laughter animation from existing HMM models that will be consistent with a given audio laughter. Since the synchronization method presented in this work does not rely on the acoustic synthesis itself but only on the phonetic transcriptions, we focus on the visual synthesis starting from given phonetic transcriptions in the remainder of this paper.

4. AV SYNCHRONIZATION RULES

In the HMM-based synthesis framework, the first stage is the training of the HMMs. HMM training needs to have relevant annotations so that the system knows what kind of information is contained in a given data file used for the training. In the case of acoustic data, the annotations are phonetic transcriptions. An ideal case would have been that these phonetic transcriptions could be used as annotations for the visual data as well. This was successfully applied to audio-visual speech synthesis in [11, 12, 13, 8]. Unfortunately, due to the fact that laughter is an inarticulate utterance [18], the correlation between the produced sound and the facial expression and particularly the mouth shape is much lower than in speech. This is why separate annotations were necessary for visual data training [10]. Instead of using phonetic classes, three specific classes related to the deformations on the face were used. A following study has shown that a third modality, the head motion, should be considered independently to better model the head motion by considering the shaking motion during laughter. This approach performed better in percep-

tive tests [14]. We finally end up with three different modalities (audio, face, head) all related to the same phenomenon (laughter). Each modality has its own transcriptions and therefore synchronization rules between these modalities are necessary since the three modalities are trained independently and nothing ensures the synchronization at the synthesis step.

The transcriptions for the audio modality consist in several successive phones such as fricatives, different vowels, nasal sounds, nareal fricatives, silence and inhalation (cf [15] for more details). The most common transcriptions for audio laughter are similar to : silence-h-a-h-a-h-a-inhalation-silence. In the case of facial data, three classes are used in the visual transcriptions : Neutral, Laughter and Semi-Neutral. The latter is a facial expression between no expression at all and a slight smile (cf [10] for more details). The majority of the laughs in the database are a succession of the first two classes in the following order : Neutral-Laughter-Neutral. Finally, the head motion transcriptions are the result of a sub-segmentation of the facial "Laughter" class defined above. Each occurrence of one class during a laughter sequence represents one period of the head oscillation that occurs in laughter (cf [14] for more details). Figure 3 gives a schematic overview of the different transcriptions. As we can see, the beginning of the audio laughter (end of the silence before) is not exactly aligned with the beginning of visual laughter (end of the neutral face). Similarly, the visual laughter class ends some time after the last audible contribution. This shows that visual laughter is temporally wider than acoustic voiced laughter. Figure 3 also shows head oscillations with red circles. As we can see, they are defined such that the Neutral class before and after the laughter on the face remain the same in head motion transcriptions and the laughter class becomes a succession of "oscillation periods".

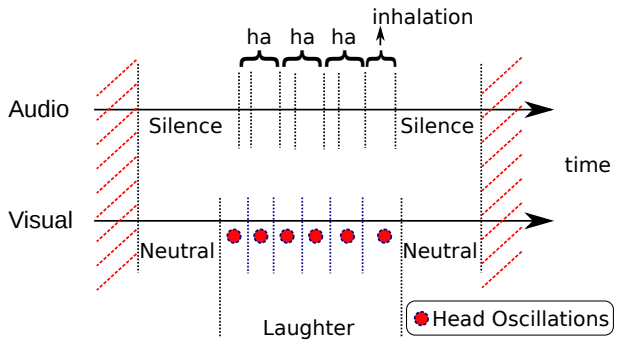

Fig. 3. Schematic representation of the different transcriptions (audio, face, head)

The synchronization method proposed in this work is to study the relation between the audio and facial data transcriptions and derive rules from it to later use these rules to produce facial transcriptions corresponding to phonetic transcriptions. In this work, the facial transcriptions are assumed to be a sequence of type "Neutral-Laughter-Neutral" which is the most common sequence ($\pm$80% of the database). Once facial transcriptions are generated, head motion transcriptions are generated from these facial transcriptions. Finally, the generated visual transcriptions are used as input to their respective HMM models for synthesis and trajectories are produced. The synthesized visual data for face and head are then merged and transformed appropriately [14] before application on a 3D face. Finally, video animations are produced with the corresponding audio data. The next sub-sections explain how exactly the facial transcriptions and head motion transcriptions are generated.

4.1. Visual transcriptions generation

One particularity of laughter is that its visual expression usually begins before the audible sound and finishes after the last voiced sound [18]. Starting from this observation, we have studied the time shift between the end of the initial silence in the phonetic transcriptions and the end of the neutral expression in the visual transcriptions. Figure 4 shows the histogram of this time shift in the most common case in our database (an audio laughter beginning with a nasal sound). A positive value on the X axis means that the visual laughter begins before the audio laughter. The most common diphones at the beginning of a laughter in the used database are "silence-nasal" and "silence-nareal_fricative" where nareal_fricative is a phone representing the sound produced when one laughs by expelling air from the nose. One probability function for each of both cases were built as well as a global probability function for all other cases. These functions are used to generate random time shifts to determine the starting time of the visual laughter with respect to the start of the audible laugh. For example, if the first silence ends at time t_1 in a given audio laughter file and if the generated random time shift from the appropriate probability function is $t_{\Delta,Left}$ then the left boundary of the visual laughter will be set to be $t_{visual,Left} = t_1 - t_{\Delta,Left}$.

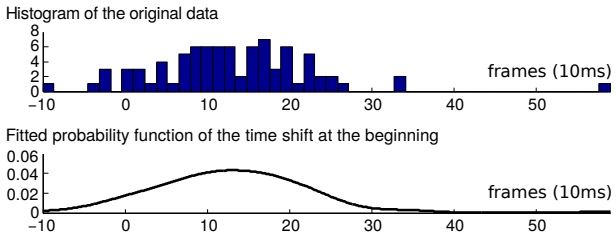

Fig. 4. Histogram and fitted probability function of the time shift between the end of the first facial neutral pose and the end of the first audio silence

A similar approach is followed to model the time shift at the end of the laughter. The histogram of the time shift

between the end of the visual laughter and the end of the audible inhalation sound is modeled by probability functions. Three probability functions are also built. Two corresponding to the two most common inhalation phones (either fricative or nareal_fricative sounds respectively corresponding to an unvoiced inhalation from the mouth or from the nose) and one additional global probability function for all other cases. For example, if the inhalation phone ends at time t_2 and the generated time shift from the appropriate probability function is $t_{\Delta,Right}$ then the right boundary of the visual laughter will be set to be $t_{visual,Right} = t_2 + t_{\Delta,Right}$.

4.2. Head transcriptions generation

The head motion transcriptions are such that they are identical to facial transcriptions but with the laughter class divided into a certain amount of oscillations. Figure 5 gives the number of oscillations as a function of the length of the laughter for the files in the database. A linear model is fitted on this data and is represented as a red line on the Figure and its equation is $y = 0.022x + 2.9$. The Pearson's correlation coefficient between the two axis on Figure 5 is 0.93. We therefore assume that a linear model is relevant to determine the number of oscillations that should occur for a laughter of a given length.

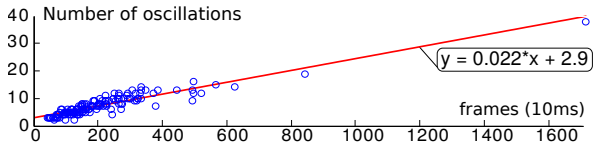

Fig. 5. Number of oscillation as a function of the length of the laughter and a linear model fitted on the scatter plot

Once the number of oscillations is determined, the total length of the laughter is divided by this number of oscillations and the result is used as the period for sub-segments.

5. EVALUATION

An online perceptive test was conducted to assess the appropriateness of the synchronization method presented in the previous section. The original audio tracks are used and three different types of videos are included in the test :

1. ORI = Videos rendered by applying original visual trajectories on the 3D avatar
2. SYN = Videos rendered by applying synthesized visual trajectories (from original visual transcriptions) on the 3D avatar
3. SYN-GEN = Videos rendered by applying synthesized visual trajectories (from generated visual transcriptions) on the 3D avatar

Thirty-nine participants took the online test. Fourteen women aged from 25 to 67 (average = 35.9) and twenty-five men aged from 22 to 49 (average = 32.6). Each participant was asked to rate 20 videos randomly picked from a set of 42 videos. The set of 42 videos consists in 14 different laughs (rendered with the three methods ORI, SYN and SYN-GEN) picked randomly from the available database and they were not used in the training process of the visual models. For each video, participants were asked to rate on a 5-point scale from very poor (0) to Excellent (4) the overall quality (Q1) and how well the animation matches the audio in the video (Q2). Figure 6 gives the mean scores obtained for each type of video and each of both questions. Standard errors are also given. The TUKEY Honestly Significant Difference (HSD) test with a confidence level of 95% shows that there are no significant differences between the mean scores for Q1. The same test is performed for Q2 and shows that there is a significant difference only between the original trajectories and the synthesized ones (marked with an asterisk on the Figure). Although the mean score for Q2 for SYN-GEN is slightly lower (2.64) than for SYN (2.80), the TUKEY HSD test does not show a significant difference between them.

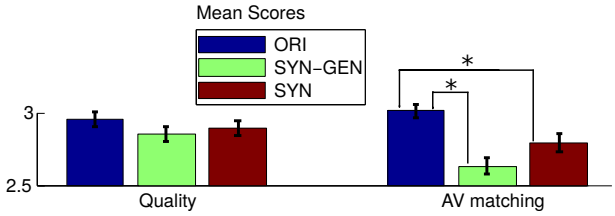

Fig. 6. Mean scores and standard errors for each method

6. CONCLUSION AND FUTURE WORKS

In this work, we have proposed simple rules to synchronize the audio and visual synthesis of laughter based on the study of the relations between audio and facial transcriptions and between facial and head transcriptions. Trajectories were synthesized using the proposed method and animation videos were rendered to conduct an online perceptive test. The test showed that there are no significant difference between the videos rendered from original transcriptions and generated transcriptions using the method presented in this paper.

Future work include improvements on the rules to reach a more precise and specific synchronization. Extending the rules to be able to deal with more complex laughs than the typical ones considered in this paper is also in the scope. In this work, it was assumed that audio transcriptions contain an inhalation which might not always be the case. Also, only the most typical laughs were considered. Extending this first work to more complex laughs is necessary to build a more flexible audio-visual laughter synthesis system.

7. REFERENCES

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

[2] J. Urbain, H. Çakmak, and T. Dutoit, "Automatic phonetic transcription of laughter and its application to laughter synthesis," in *Proceedings of the 5^{th} biannual Humaine Association Conference on Affective Computing and Intellignet Interaction (ACII)*, Geneva, Switzerland, 2-5 September 2013, pp. 153–158.

[3] J. Urbain, H. Çakmak, Aurélie Charlier, Maxime Denti, T. Dutoit, and Stéphane Dupont, "Arousal-driven synthesis of laughter," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, pp. 273–284, 2014.

[4] G. Hofer, J. Yamagishi, and H. Shimodaira, "Speech-driven lip motion generation with a trajectory hmm," 2008.

[5] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hmm-based text-to-audio-visual speech synthesis," in *ICSLP*, 2000.

[6] L. Wang, Y. Wu, X. Zhuang, and F. Soong, "Synthesizing visual speech trajectory with minimum generation error," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4580–4583.

[7] G. Hofer and K. Richmond, "Comparison of hmm and tmd methods for lip synchronisation," 2010.

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

[9] G. Bailly, O. Govokhina, F. Elisei, and G. Breton, "Lip-synching using speaker-specific articulation, shape and appearance models," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, pp. 5, 2009.

[10] H. Çakmak, J. Urbain, J. Tilmanne, and T. Dutoit, "Evaluation of hmm-based visual laughter synthesis," in *Acoustics Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014.

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

[12] 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.

[13] M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuda, "Visual speech synthesis based on parameter generation from hmm: Speech-driven and text-and-speech-driven approaches," in *AVSP'98 Int. Conf. on Auditory-Visual Speech Processing*, 1998.

[14] H. Çakmak, J. Urbain, and T. Dutoit, "Hmm-based synthesis of laughter facial expression," *Transactions on Affective Computing (TAC)*, 2015, [Submitted].

[15] H. Çakmak, 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.

[16] J. Urbain and T. Dutoit, "A phonetic analysis of natural laughter, for use in automatic laughter processing systems," in *ACII 2011*, 2011, pp. 397–406.

[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] W. Ruch and P. Ekman, "The expressive pattern of laughter," in *Emotion, qualia and consciousness*, A. Kaszniak, Ed., pp. 426–443. World Scientific Publishers, 2001.