# AUDIO-VISUAL LAUGHTER SYNTHESIS SYSTEM

Hüseyin Çakmak, Kevin El Haddad, Thierry Dutoit

University of Mons (UMONS/Belgium)
{huseyin.cakmak},{kevin.elhaddad},{thierry.dutoit}@umons.ac.be

## ABSTRACT

In this paper we propose an overview of a project aiming at building an audio-visual laughter synthesis system. The same approach is followed for acoustic and visual synthesis. First a database has been built to have synchronous audio and 3D visual landmarks tracking data. Then this data has been used to build HMM models of acoustic laughter and visual laughter separately. Visual laughter modeling was further separated into a facial modeling and head motion modeling. An automatic laughter segmentation process has been used to annotate visual laughter. Finally, simple rules were defined to synchronize all the different modalities to be able to produce new durations.

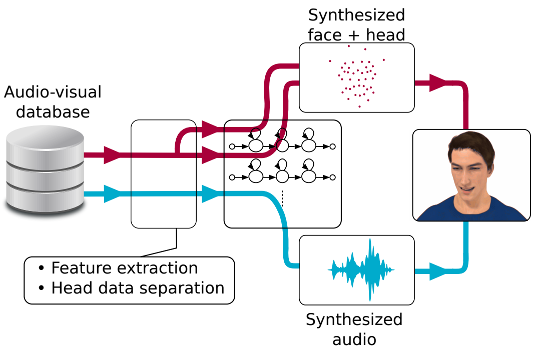
**Keywords:** audio, visual, laughter, synthesis, HMM-based

## 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.

As summarized in Figure 1 the main of this project is to build a complete audio-visual laughter synthesis system. From an audio-visual database, HMM-based modeling is done for audio data, face and head motion separately with their respective annotations. Phonetic transcriptions were made manually while face transcriptions are based on a GMM-based clustering of the visual data. Head transcriptions are derived from face transcriptions and from head motion data automatically as well.

**Figure 1:** Overview of the pipeline for HMM-based audio-visual laughter synthesis



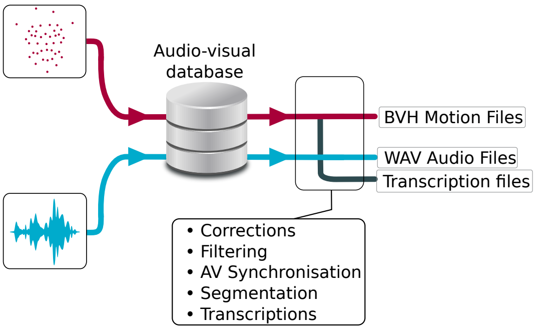
The paper is organized as follows: Section 2 gives a brief overview on the database used in the project, Section 3 explains the acoustic laughter synthesis method, Section 4 explains the visual synthesis, Section 5 describes synchronization rules between audio and visual modalities and Section 6 concludes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

## 2. THE AV-LASYN DATABASE

The AVLASYN Database [1] 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 (Shure SM58 micro and RME Fireface 400 Soundcard) and a marker-based motion capture system have been used for audio and facial expression recordings respectively. Figure 2 gives an overview of the 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. 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.

**Figure 2:** Data recording pipeline



This led to roughly 48 minutes of visual laughter and 13 minutes of audible laughter. Audible laughter is less than visual laughter because smile is also part of visual laughter and has not audible components. Also, facial expression begins before acoustic expression of laughter and continues after the last audible sound.

## 3. ACOUSTIC LAUGHTER SYNTHESIS

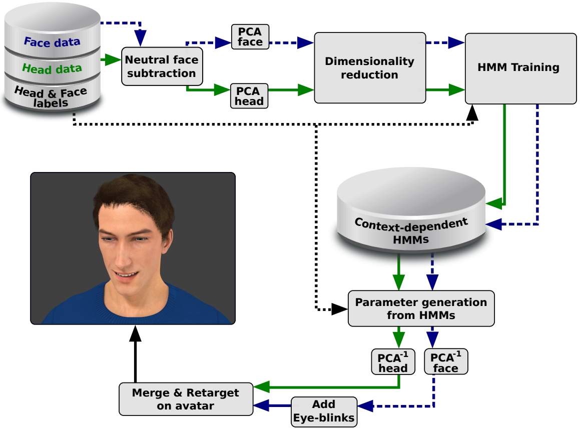
Acoustic synthesis of laughter using Hidden Markov Models (HMMs) has been addressed in 2013 [11]. To characterize the acoustic laughter, phonetic transcriptions [12] were used and the results outperformed the state of the art. Extensions of the latter work were done to perform automatic phonetic transcriptions [10] and to integrate the arousal in the system [9]. The goal of audio laughter synthesis is to generate an audio waveform of laughter.

Several versions of HMMs were developed, with varying contextual information and algorithms for estimating the parameters of the source-filter synthesis model. These methods were compared, in a perception test, to human laughs and copy-synthesis laughs. In this evaluation, participants were invited to rate the naturalness of the laughs they were listening to. The evaluation showed that 1) the addition of contextual information does not increase the naturalness, 2) the proposed method is significantly less natural than human and copy-synthesized laughs, but 3) significantly improves laughter synthesis naturalness compared to the state of the art. The evaluation also demonstrates that the duration of the laughter units can be efficiently learnt by the HMM-based parametric synthesis methods.

## 4. VISUAL LAUGHTER SYNTHESIS

In [4] we have proposed a visual laughter synthesis system. This 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 [5, 7, 6, 8]. Further developments have shown that the head motion should be modeled separately as well [2]. The general visual laughter synthesis pipeline is given in Figure 3. Principal Component Analysis (PCA) is performed independently on face and head motion data to uncorrelate the features and to reduce dimensions. Then the reduced data are trained separately for head and face with their specific transcriptions. As shown on Figure 4, the phonetic transcription are more granular than visual transcriptions. Facial transcription are mainly limited to two classes : neutral or laughter. Head motion transcription are based on the facial transcription in the sense that the neutral remains the same and the laughter is subdivided into a set of identical classes which represent each one period of the head oscillation occurring during laughter.

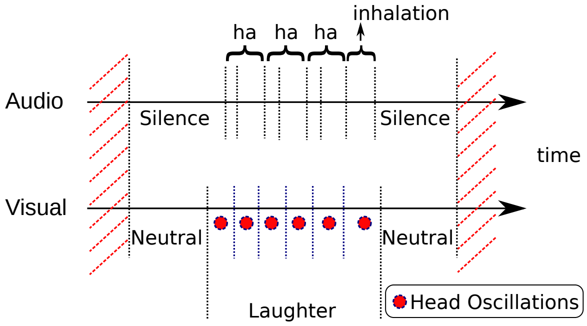
**Figure 3:** Visual laughter synthesis pipeline [2]



## 5. SYNCHRONIZATION RULES

In [4], 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 as explained below.

**Figure 4:** Audio, facial and head transcriptions



As explained in the previous sections, audio, facial data and head data are modeled separately with their own transcriptions and this leads to the need of synchronization techniques. The basic principle lying under the proposed method in [3] 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. An online MOS test is then conducted to rate the quality of the animation and the matching between audio and visual modalities. The results show that there is no significant difference between original animations and those using the synchronization rules of this section.

## 6. CONCLUSION AND FUTURE WORKS

This paper briefly presented recent advances towards building an audio-visual laughter synthesis system. The recording of a database, the development of the acoustic and visual parts of the system has been introduced. Synchronization rules used to be able to unify audio and visual parts has been presented as well.

Future works include the integration of the arousal as a control of the intensity of the laughter to be produced. The extension to more complex laughs is also planned, both in terms of synthesis and synchronization. Building the visual transcription directly from the audio track of laughter rather than being dependent on the availability of its phonetic transcription would be valuable as well.

## 7. ACKNOWLEDGEMENT

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

## 8. REFERENCES

[1] Çakmak, H., Urbain, J., Dutoit, T. 2014. The avlasyn 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).*

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

[3] Çakmak, H., Urbain, J., Dutoit, T. 2015. Synchronization rules for HMM-based audio-visual laughter synthesis. *IEEE International Conference on Acoustics Speech and Signal Processing ICASSP.*

[4] Çakmak, 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).*

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

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

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

[8] 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.*

[9] Urbain, J., Çakmak, H., Charlier, A., Denti, M., Dutoit, T., Dupont, S. 2014. Arousal-driven synthesis of laughter. *IEEE Journal of Selected Topics in Signal Processing* 8, 273–284.

[10] Urbain, J., Çakmak, H., Dutoit, T. 2013. Automatic phonetic transcription of laughter and its application to laughter synthesis. *Proc. Humaine Association Conference on Affective Computing and Intellignet Interaction (ACII).*

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

[12] Urbain, J., Dutoit, T. 2011. A phonetic analysis of natural laughter, for use in automatic laughter processing systems. *Proc. Humaine Association Conference on Affective Computing and Intellignet Interaction.*