Towards a Listening Agent: A System Generating Audiovisual Laughs and Smiles to Show Interest

Kevin El Haddad*
University of Mons
31, Boulevard Dolez
Mons, Belgium
kevin.elhaddad@umons.ac.be

Hüseyin Çakmak
University of Mons
31, Boulevard Dolez
Mons, Belgium
huseyin.cakmak@umons.ac.be

Emer Gilmartin
Trinity College Dublin
College Green, Dublin 2
Dublin, Ireland
gilmare@tcd.ie

Stéphane Dupont
University of Mons
31, Boulevard Dolez
Mons, Belgium
stephane.dupont@umons.ac.be

Thierry Dutoit
University of Mons
31, Boulevard Dolez
Mons, Belgium
thierry.dutoit@umons.ac.be

ABSTRACT

In this work, we experiment with the use of smiling and laughter in order to help create more natural and efficient listening agents. We present preliminary results on a system which predicts smile and laughter sequences in one dialogue participant based on observations of the other participant's behavior. This system also predicts the level of intensity or arousal for these sequences. We also describe an audiovisual (AV) concatenative synthesis process used to generate laughter and smiling sequences, producing multi-level amusement expressions from a dataset of audiovisual laughs. We thus present two contributions: one in the generation of smiling and laughter responses, the other in the prediction of what laughter and smiles to use in response to an interlocutor's behaviour. Both the synthesis system and the prediction system have been evaluated via Mean Opinion Score tests and have proved to give satisfying and promising results which open the door to interesting perspectives.

CCS Concepts

•**Computing methodologies** → **Intelligent agents;** Discourse, dialogue and pragmatics; Natural language generation; •**Human-centered computing** → *HCI theory, concepts and models;*

*This work was partly supported by the Chist-Era project JOKER with contribution from the Belgian Fonds de la Recherche Scientifique (FNRS), contract no. R.50.01.14.F.

ICMI '16, November 12 - 16, 2016, Tokyo, Japan

ISBN 978-1-4503-4556-9/16/11…$15.00

DOI: http://dx.doi.org/10.1145/2993148.2993182

Keywords

HCI; Embodied Conversational Agent (ECA); Audiovisual Synthesis; Conditional Random Fields, Laugh, Smile

1. INTRODUCTION

To create more natural human-computer interaction (HCI) applications, particularly in domains where such a human-like interaction could be more effective such as entertainment, education, and healthcare, the incorporation of signals other than TTS-rendered speech and pre-existing graphical elements into the interface may be valuable. Such interfaces must be capable of understanding such signals in the user, and generating appropriate behaviour in the avatar or agent. There is a significant body of work addressing aspects of these challenges, particularly in the domain of social signal processing processing, with research into generating and understanding feedback signals, emotional speech, and gesture. Laughter and smiling are fundamental human behaviours, present in interaction, which can have a bonding function, and would therefore be very valuable additions to human machine interactions. To explore such interfaces, we are working on the creation of audiovisual (AV) virtual agents which incorporate natural face-to-face communication and especially humour and laughter.

In this paper we present two contributions on the inclusion of laughter and smiling in a conversational agent which will aid more natural HCI. The first is an AV synthesis system which can generate amused expressions at several levels of intensity or arousal. We also describe preliminary work on a probabilistic model for the prediction of appropriate agent smiles and laughs at different arousal levels based on a database of interactions. Our goal is to give an agent the ability to generate appropriate smiling and laughter behaviour in response to an interlocutor. This behaviour would be useful in simulating feedback or interest in the interlocutor without recourse to ASR, and could thus be used while the interlocutor is speaking or while a spoken response is being prepared.

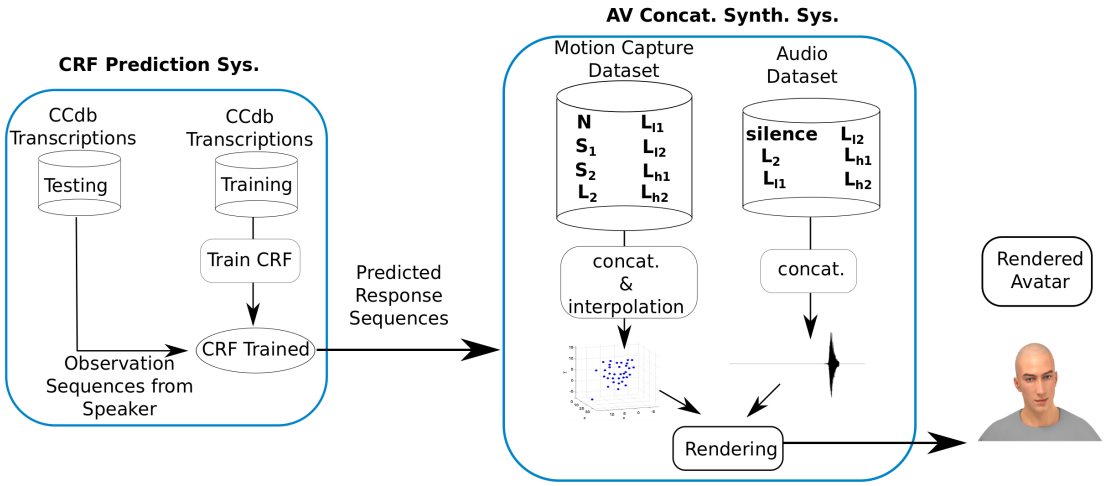

Figure 1: Overview of our prediction/synthesis systems mechanism

The smiling and laughter synthesis is made by an audio and visual concatenation system inspired by the familiar concatenative synthesis paradigm in TTS. The system can generate amused behaviour at several levels of intensity or arousal from a database of such behaviour. The transition from one level to another is managed by a linear interpolation technique to obtained smoothed transitions.

We are thus able to generate successive smiles and laughter episodes at different arousal levels. With the concatenation/interpolation technique, we also have a precise control over the sequence duration. The system concatenates motion capture data segments or parts of them on one side and audio data on the other. The concatenation is made in such a way that both audio and visual cues are synchronous.

A system capable of generating smiling and laughter must also be able to decide when to output such behaviour and how intense the speech or laughter should be. To this end, a Conditional Random Field (CRF) was trained using some of the labels of the Cardiff Conversational Database (CCDb) [2]. Our goal is to obtain a model able to provide appropriate smile and laughs responses by an artificial agent at several levels of arousal to smile and laughter stimuli. Since this part is a preliminary work, the stimuli are the CCDb participants' smiles and laughs labels with their levels. Fig. 1 gives an overview of both system's working pipeline.

Below we describe existing work in the field. We then describe the design, implementation, and evaluation of our audiovisual synthesis and of our predictive model for laughter and smile responses. We conclude with a discussion and outline future work.

1.1 Feedback in Dialogue Systems and Conversational Agents

Spoken interaction is far more than a verbal rendition of linguistic text, and includes elements in several modalities contributing to the smooth progression of dialogue. Prosodic, syntactic, and gestural/postural cues can contribute to turntaking decisions and discrimination between backchannel feedback and utterances on the main floor. Several of these cues have been the subject of dialogue system research, particularly in the domain of turntaking. Emotive and affective elements in terms of tone of voice, lexical choice, and facial expression have been studied in social signal processing projects such as SEMAINE [26], which resulted in the creation of an artificial listener demonstrating different emotive or personality types. Probabilistic models of feedback and turn taking have also been implemented in dialogue systems. Raux and Eskenazi's turntaking module based on prosodic and syntactic cues has been successfully implemented in the Let's Go system [25]. Meena et al's probabilistic model provides suitably positioned feedback responses and natural turntaking in a Map Task system [18]. This feedback is in the form of tokens such as 'yeah', 'mmhmm', and 'okay', and is thus tied to the interlocutor's linguistic input. Non-verbal feedback, vocal and gestural feedback based on probabilistic models has been incorporated into systems providing training in public speaking [9], roleplay support for training in effective negotiation, and psychological screening using the SimSensei platform [11]. Our work focuses on implementing laughter and smiling in such agents, using synthesis. We are building on prior work on synthesis of laughter and smiled behaviour.

1.2 Laughter in Spoken Interaction

Laughter and smiling are fundamental to human interaction, particularly in the social or affective dimension, but also help to manage information transfer [13]. Laughter is predominantly a social rather than a solo activity, is universally present in humans, part of the 'universal human vocabulary', innate, instinctual, and inherited from primate ancestors [6, 23]. It has been described as a social cohesion or bonding mechanism used since our primate days [15]. It has been suggested that laughter can provide clues to dialogue structure [16], and this has been demonstrated particularly around topic changes where laughter seems to provide an interlude of social bonding and ward off uncomfortable silence [14]. Laughter episodes take a range of forms – from loud bouts to short, often quiet chuckles. It is multimodal, comprising a stereotyped exhalation of air from the mouth in conjunction with rhythmic head and body movement [3, 4]. However, while this stereotypical laugh is generally produced upon asking an informant to laugh, it has been shown to be only one of several manifestations of laughter present in social interaction, and often not the most prevalent [28].

In conversation, laughter generally punctuates rather than interrupts speech, although it can occur within speech as speech-laughs [27].

2. AV AMUSEMENT SYNTHESIS SYSTEM BY CONCATENATION

This section gives an overview of our system. Inspired by the synthesis by concatenation method in speech synthesis, the idea here is to synthesize an AV sequence of amused expressions of different levels by concatenating recorded expressions of amusement.

For the visual part, the chunks of data needed for synthesis are extracted from the available dataset. They are then concatenated using linear interpolation to smooth the transitions between different expressions and to control the length of a single expression while keeping it as natural as possible.

For the audio cues, laughter sounds are concatenated with "silence" in such a way that the audio sequence is temporally synchronized with the visual one. Since smiles emit no sound and are just facial expressions, the audio cue corresponding to them is silence.

Previous work on AV laughter synthesis can be found. In [7], a parametric approach is presented. In [12], a combination of parametric and concatenative approach is presented.

2.1 Smiles and Laughs AV Dataset

The data used comes from the AVLASYN database [8]. It is a multimodal database containing 203 laughter events recorded from a single male subject. This database has motion capture data of the subject's facial expressions as well as the audio.

During these recordings the subject watched funny videos while being recorded. Laughter eventually occurred as well as smiles. The Optitrack system was used with facial landmarks for the motion capture recordings as shown in figure 2. Although a parametric synthesizer would add more flexibility (and will be considered in future work), we preferred a concatenative approach in this work for more naturalness. This will indeed facilitate the evaluation of the prediction system presented in Section 3.

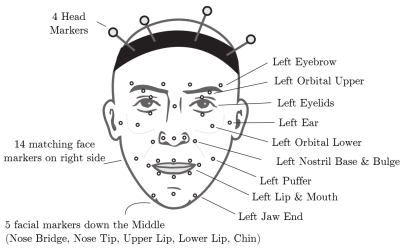

Figure 2: Optitrack facial makers position

2.2 Level Assignment

For the purpose of this work, not all the laughter events were used. For our concatenation system, we need amusement expressions at different intensity levels. In this work, we consider smiles and laughter to be different amusement types. We also consider the intensity levels of these expressions.

Indeed, smile arousal intensity levels were assessed based on the width of spreading of the lips and the degree of opening of the mouth. The more spread and/or opened the mouth is the higher the level.

All the laughter samples in AVLASYN were presented to participants in an online experiment which required them to quantify the amusement level of each laughter sample (this experiment is out of the scope of this article). A total of 226 participants were asked to grade the intensity as they perceived it of the laughter presented to them. They graded the laughs on a scale of 0 to 4 based on both the audio and the visual. At the end of the evaluation, the mean value of the scores given to each laugh was calculated, thus giving us an amusement intensity score for each of them.

2.3 Data Selection

As mentioned in Section 2.2, for this work we need amusement expressions on several levels of arousal. Thus, two smiles and five laughs were selected from the database and the selection criteria were their estimated intensity as well as their duration.

Concerning the smiles, AVLASYN contains an instance of a neutral face (no specific emotion expressed) to which succeeds two consequent smiles with an increasing arousal level followed by a laugh. This sequence was segmented (the segmentation process will be explained in the following) and each of the four expressions kept to be used in our concatenation system. This choice was motivated by the fact that this sequence contains a succession of different levels of amusement expressed naturally, which is precisely what we aim to obtain using our concatenative approach. We thus have two levels of smiling (referred to as S_1 for the lower level and S_2 for the higher level), one level of laughter (referred to as L_2) and a neutral expression (referred to as N). Laughs of two other levels (one lower and the other higher than the one we had already selected) were chosen based on their amusement score mentioned in Section 2.2 and their duration. We then extracted two laughs from each of these levels from different recordings of a different duration each (L_{l1} and L_{l2} for the lower levels and L_{h1} and L_{h2} for the higher level). This was done to obtain the most varied laughs possible with the minimum number of samples. Such a dataset would allow us to have control over the laughs intensity and also its duration which is also an important parameter as will be seen later on in Section 2.6.

We thus gathered a subset of the previously mentioned dataset with an AV content of one neutral expression, two smile expressions on two different levels and five laughs classified in three different arousal levels.

Table 1: AV dataset content summary

Data	Type	Duration (sec.)	Intensity
N	neutral	0.6	-
S_1	smile	1.36	low
S_2	smile	0.77	high
L_2	laugh	2.51	medium
L_{l1}	laugh	0.52	low
L_{l2}	laugh	1	low
L_{h1}	laugh	1.64	high
L_{h2}	laugh	5.34	high

Table 1 summarizes our dataset content with more details about the data selected.

2.4 Segmentation

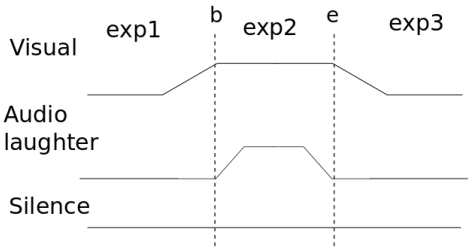

Figure 3: Data segmentation process. The dotted lines show the segmentation beginning (b) and ending (e) points which are based on the visual data. The audio segments are extracted synchronously with the visual.

The data gathered were segmented based mainly on the visual cues and more precisely on the transition between facial expressions. For the visual data, the beginning of the segment was at the end of the transition from the previous facial expression to the one being extracted. The end of the segment was at the beginning of the transition from the expression being extracted to the following one. The audio cue corresponding to each facial expression was extracted based on the visual segments so that both have the same length and are synchronized. Thus obtaining laughs sounds and silence for the smiles and the neutral expressions. Figure 3 illustrates the segmentation procedure.

2.5 AV Concatenation System Overview

This system is based on a concatenation/interpolation method. This method allows us to control two important aspects. First, the duration of each expression can be controlled thus also controlling the duration of the total sequence. Second, the arousal level can be chosen based on the data available. The primary purpose of the interpolation is to obtain a smooth transition in the visual cue from one facial expression to another. The whole system will be explained in more details in this section.

2.6 Concatenation Duration Control

As seen in Section 2.3, the data used here have fixed lengths, but the duration of the neutral and smiling expressions can still be controlled by concatenation. In order to do so, we either concatenate the expression with itself or take a slice of it.

We express the required duration as a function of the actual duration of the desired expression to be synthesized:

$$Td = kT + r$$

Td being the required duration, T the actual one, k the factor and r the remainder.

Using this and to obtain an expression of duration Td, we first concatenate k entire expressions. We then concatenate the result with a slice of the same expression. This slice would have a temporal length r extracted from the beginning of this expression.

Concerning the motion capture data, and for the neutral and smiling data, a linear interpolation, which will be explained in more details in the next section, is then applied between the last and first frames of two frame chunks being concatenated. Indeed, without this interpolation and even for the same expression, the concatenated expression is very likely to present discontinuities at the transitions. This is true even when concatenating the same expression since that the beginning and ending frames of a same expression in our dataset are not exactly the same.

Concerning laughter motion capture data, this concatenation/interpolation is not applied here. This is because laugh lengths and intensities affect the pattern of the laugh itself, i.e., the sound and facial expression related to a longer laugh can be different from a shorter one when being expressed by the same person. In this work, in order to have some control over the laugh lengths while choosing the laughs with respect to their intensities, their durations in the dataset are also considered when picking a laugh instance to be concatenated. That is why several laughter instances were picked and not only the their intensity but also their durations were considered. In our case (in our dataset) the shorter the laugh, the lower its intensity.

We can therefore have some control of the laugh duration even though, with the current database, this control is restrained due to the limited number of samples. The bigger the database, the more control over the intensity and the length we will have. Even though a concatenation technique to control the laughs length has not been presented here, it is being investigated and will be the subject of future work.

2.7 Interpolation of Visual Data for Smooth Concatenation

As explained in Section 2.6, a linear interpolation is needed to have a smooth transition when concatenating AV expressions (an AV expression with itself of with another one). When expression B is to be concatenated to expression A while B temporally follows A, this interpolation is applied between the last frame a of A and the first frame b of B. This is done as follows:

$$f = w1 * a + w2 * b$$

$w1$, and $w2$ being the interpolation weight vectors. A $w1$-$w2$ pair should always add up to 1. f being the interpolated frame. To obtain a smooth enough transition between two expressions, several interpolation frames should be created. The average transition duration from one expression to another in our dataset was 0.53 seconds (corresponding to 53 frames if the frame rate is 100 fps). To calculate this value, all the transitions between neutral, smiles and laughs from the above-mentioned selected data were considered. Therefore, using the interpolation, 53 new frames are inserted to create a smooth transition every time the expression (or level of expression) changes and 10 frames are inserted when concatenating a neutral of smile expression with itself. These numbers of frames inserted can be increased (alt. decreased) to obtain a longer (alt. shorter) transition duration.

2.8 System Evaluation

In this section we intend to evaluate our concatenative synthesis system with a Mean Opinion Score (MOS) test.The motion capture trajectories of the expressions are therefore synthesized and then stored in videos of MP4 format and at 30 fps rates using the Blender software [5]. These videos will serve as stimuli in the MOS test. Since we have 3 levels of laughter and 2 levels of smiling in our dataset, we synthesized several combinations of expressions in order to

evaluate the efficiency of the interpolation technique. All smile samples were used for this test but only one laughter sample from each level was chosen randomly to simplify the test. Thus, 15 videos were created to serve as stimuli for the MOS test and Table 2 shows the content of these videos. The second column shows the succession of expressions contained in each file (e.g. N-L-N indicates that the corresponding file begins with an N expression followed by an L and ends with an N, the "-" indicates a transition in the visual cue that is created by interpolation). In this table, R1 represents the complete file containing N, S_1, S_2 and L_2. R2 and R3 correspond to the complete files from which L_l1 (referred to as L_1 in what follows) and L_h2 (referred to as L_3 in what follows) were taken. By complete we mean, containing the original transitions with the expressions preceding and following the laughs. These will help us compare the natural transitions with the synthetic ones.

These files contain all possible transitions from a smiling or neutral expression to a smiling or laughter expression as well as the opposite. We made sure the videos had all the same length (6 seconds) by controlling the lengths of the neutral and smiling expressions. This contributes to test the efficiency of the duration control of our system.

Table 2: Characteristics of the stimuli created and scores obtained on the 2 questions of the MOS test

Stimuli	Config.	Scores Q1	Scores Q2
1	N-S_1-N	2.71	3.06
2	N-S_2-N	2.69	2.86
3	S_1-S_2-S_1	2.74	3.03
4	N-L_1-N	2.83	2.89
5	N-L_2-N	3.17	3.11
6	N-L_3-N	2.29	3
7	S_1-L_1-S_1	2.86	2.66
8	S_1-L_2-S_1	3.11	3.2
9	S_1-L_3-S_1	2.31	2.91
10	S_2-L_1-S_2	2.91	2.89
11	S_2-L_2-S_2	3.23	3.29
12	S_2-L_3-S_2	2.14	2.97
13	R_1	3.2	3.14
14	R_2	3.14	3.23
15	R_3	2.16	2.97

The videos were separately shown to 34 participants. These later were asked to view reply to two questions for each video:

1. Q1: What do you think about the quality of the animation (please do not take the sound into account)?
2. Q2: How coherent are the audio (or the silence) and the video in the following animation?

We hoped that the first question would incite the participants to focus more on the quality of the animation and thus on the transitions too with the least bias from the audio as possible. The second question would complement the first one by assessing the audio's coherence with the video and therefore the synchronization quality between the audio and the video.

The participants could reply to each question by choosing one of the following responses: "Very high/Very coherent", "High/Coherent", "Average", "Low/Not so coherent", "Very low/Not coherent at all". The answers were each mapped to the scores 4, 3, 2, 1, 0 respectively.

The mean scores obtained on the 2 questions for each file are shown in Table 2.

From these results we can conclude that our concatenative synthesis system is efficient and synthesizes reactions that are well perceived on the naturalness scale. Indeed, all the scores (for both Q1 and Q2) are on average above 2 (on a 0 to 4 scale).

Another conclusion could be drawn concerning the context surrounding the laughs in the files by concatenation. Indeed, in this study, the laughs were surrounded either by neutral of smiles expression (2 smiles at different levels). It appears that the surroundings of the laughs do not directly affect the naturalness perceived. In fact, surprisingly, the files containing laughs surrounded by N, S_1 and S_2 obtained each equal mean score values (2.76) even if 95% CI Student's t-tests showed that these equalities are not significant.

When comparing the files obtained by concatenation and the ones created directly from the real data, we can see that some file obtained on average higher scores while others lower scores. A 95% CI Student's t-test applied showed that some of the difference between the scores obtained were significant and some weren't. It is thus safe to conclude that, in general, our concatenative synthesis system performs on a naturalness scale as good as the natural data.

3. LAUGHS AND SMILES PREDICTION: A PROBABILISTIC FEEDBACK

The goal of this prediction system is to generate the smiles or laughs sequences as well as their intensity levels. These sequences will then be used as input to the concatenative synthesis system described previously to generate feedback to the speaker. Previous work was done generating laughter as a feedback using mimicking [29, 21]. Instead of mimicking laughs, we predict smiles and laughs based on non-verbal expression observations from the speaker. We intend to compare these two techniques later on, but we expect a prediction model to give better results mainly because of its ability to generate laughs/smiles in response to a variety of expressions and not just to the mimicked ones. Probabilistic sequential models have the ability to model sequential data, i.e. to generate a sequence of states from observations. Hidden Markov Models (HMM) [24] and Conditional Random Fields (CRF) [17] are some of the techniques mostly used to do so. In this study our predictive system will be based on CRF. Indeed, Morency et. al. [20], compared the two techniques to generate backchannel sequences based on multimodal features extracted from real life conversation data. In that work, CRFs gave better results than HMM. Nevertheless, the choice of the CRF was also motivated by the work of de Kok and Heylen in [10], in which CRF are used to predict smile types of a listener based on the listener's smiles context and the speaker's smiles. CRF showed to give poor results for this task. But they were evaluated by comparing the generated sequence with the real sequence in the test set. The generated sequence was never synthesized and evaluated subjectively. Indeed, as previously mentioned, the goal of our work here, is not to copy the CCDb's participants' reactions but to be able to generate a sequence seeming as natural as possible and convenient with the context in which they will be synthesized. Also, the observed features and generated data used in our work are different from the one presented in [10] since we do not take into

account the types of smiles but rather consider smiles and laughs along with their intensity levels. We therefore expect the CRF to give good enough results.

3.1 System Overview

As previously mentioned the focus of this paper is predicting the sequences and then synthesizing them. A recognition/detection system is not used to avoid the errors it might have induced. So we bypass it by directly using sequences of labels from the transcriptions of the CCDb. We thus focus our study on the the prediction and synthesis tasks.

The CCDb contains dyadic conversations and transcriptions of the speech and several non-verbal and paralinguistic audio-visual events such as smiles and laughs. The goal here being to create a listening agent, in the CCDb and for a pair of interlocutors, the speaker and listener roles are alternated with respect to who is talking. The person uttering a sentence is considered as the "speaker" at this moment of the conversation while his interlocutor is considered as the "listener". Thus, the "speaker"'s labels are fed to the CRF as its inputs and the "listener"'s labels as its output.

The goal of this prediction system is not to predict a sequence similar to the one of the "listener" in the test set when the observations from the corresponding "speaker", but rather to predict a sequence appropriate to the context in which they were predicted (i.e. a sequence that would seem natural to a speaker that would be interacting with the agent). So, this system cannot be evaluated by comparing the sequence predicted from the observations of a given "speaker"'s sequence in the test set, with the sequence of the corresponding "listener". It will rather be tested subjectively by synthesizing the output sequences and evaluate the obtained results under a Mean Opinion Score test.

3.2 Dataset Used

The data used here came entirely from the CCDb. It contains 30 dyadic conversations only 8 of which were fully annotated. We therefore had access to only 8 of them for this study. The conversations were recorded audiovisually (2D video and audio). The annotations were made manually and temporarily segmented the speech as well as several paralinguistic and non-verbal expressions (facial, audio and body gestures) such as agreement, surprise, etc... Annotations were also made for smiles and laughs which, along with the speech are the most important for this study. Indeed, with these transcriptions, we were able to label, at a specific time in a conversation, the two interlocutors as "speaker" or "listener" as previously explained using the speech labels. We can also determine the smiling/laughter (or neutral) state of each of them using the corresponding labels. The transcriptions for the smiles also contained information about the intensity levels: 3 intensity levels were annotated. But the laugh labels didn't contain any information about the intensity level. Before using the transcriptions to train our prediction model, we therefore completed the annotations by adding 3 intensity levels to the laughs labels and the missing intensity levels to the smiles labels. The smiles intensities were assessed based on the lips spreading width and on the mouth opening. The laughs intensity estimation was made subjectively: 2 subjects gave a score between 1 and 3 for each laugh, at the end the average score was computed and rounded. The obtained score determined the intensity level (1 or lower for low, 2 for medium and 3 for high).

3.3 Implementation

To train the CRF, sequences of states were created from the smiling and laughter annotations. The states could take a value corresponding to whether it is a smile or laugh and its intensity level. The sequences coming from the "speaker" are the observations and the ones coming from the "listener" the outputs. Thus, each conversation were divided in segments. Each segment is delimited in the beginning by the start of the participant taking the role of the "speaker" and at the end by the change of turns, i.e. the "speaker" becomes "listener". Each segment is divided into 100 ms wide subsegments, each of which is assigned one of the previously mentioned states to which it corresponds (smiles or laugh and the corresponding intensity level) The features or observations used to train the the CRF are formed by the current and two previous subsegment states of the speaker and the two previous listener subsegment states.

The python-crfsuite [1] which is a Python binding of the CRFSuite implementation [22] was used here. The CRF was trained with a Stochastic Gradient Descent algorithm.

3.4 System Results and Evaluation

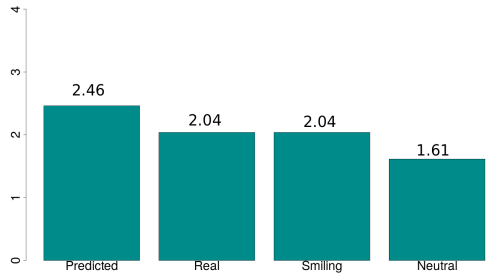

Figure 4: Mean scores obtained per behaviour synthesized for Q1

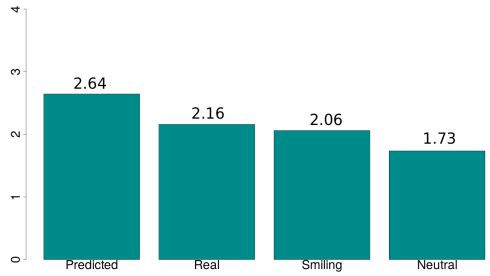

Figure 5: Mean scores obtained per behaviour synthesized for Q2

Taking into account the limited amount of data, seven conversations were use for training and one for testing. Training and testing processes were repeated three times. A different conversation was randomly picked for the testing set. This was done to test our system with different speakers and listeners. At the end we thus obtained predicted sequences for three conversations (thus 6 listeners since the role of listener is alternated between the two interlocutors per conversation) and for three different systems (since they were trained with different datasets each time).

In order to evaluate our prediction system, some of the predicted sequences will be compared to sequences of the same length containing:

- the real behavior of the "listener" in the test set corresponding to the "speaker" from which the prediction was made. The "listener"s transcriptions will be used directly to generate these sequences using our concatenative synthesis system.
- two unchanging behaviors: one constantly neutral and the other constantly smiling.

This perception test will thus be a comparison between four classes of behaviours: "Predicted", "Real", "Smiling" and "Neutral".

This would help us evaluating the quality of our predictive system. Also, we are taking advantage of this evaluation to study the impact of the facial smiling/laughter variability in on the perception of artificial agent's behaviour. Indeed our hypothesis is that a higher variability of facial expressions will be perceived as more human-like and therefore more natural.

To simplify the evaluation process, from the predicted sequences, only the ones containing smiling or laughter are considered. From these, four sequences are selected randomly. For each of these selected sequences, a sequence for each of the three other classes mentioned above are also generated. We end up with a total of 16 sequences: 4 Predicted, 4 Real, 4 Neutral and 4 Smiling.

The motion capture trajectories of the expressions corresponding to these sequences are then synthesized using the above described concatenative synthesis system. These are used to create separate video files in which the rendered avatar is juxtaposed with the corresponding "speaker" video sequence as shown in Fig 6. At the end, we have a total of 16 videos and thus 4 sets, each of them representing one of the previously mentioned classes.

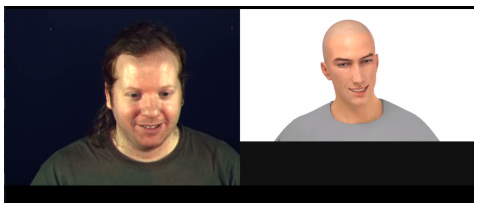

Figure 6: Stimuli video example

The obtained videos are used in a Mean Opinion Score (MOS) test. The idea is to let third party observers judge of the quality of the avatar's reactions in his fictive interaction with the "speaker" in each case.

In this test, 34 participants were asked to view the 16 videos one after another and reply to two questions every time:

1. Q1 : How well does the avatar seem to understand the speaker?
2. Q2 : How natural is the avatar's reactions?

The first question would show how much the variability of the AV expressions during a sequence in general, in our predictive system in particular, affects the way the avatar's attention and interest towards the speaker is perceived. The second question would show how natural sequences of smiling and laughter expression generated randomly without understanding what the speaker is saying, are perceived.

The participants could reply to each question by choosing one of the following responses: "Very well/Very natural", "Well/Natural", "Average", "Bad/Not so natural", "Very bad/ Not natural at all". The answers were each mapped to the scores 4, 3, 2, 1, 0 respectively. Fig. 4 and Fig. 5 show the mean score values obtained per class for Q1 and Q2 respectively.

As can be seen in both figures the scores obtained for both questions are similar. Indeed the "Predicted" class obtained the highest average score and the "Neutral" class obtained the lowest one. Concerning Q1, the "Real" and "Smiling" class surprisingly obtained the same results. For Q2, "Real" obtained a higher average score than the "Smiling" one.

Here again, a 95% CI Student's t-test was applied to check the significance of the results when comparing pairwise all the sets of scores (each set corresponding to a class). The tests showed significant results (p-value$<$0.05) for all comparisons except for "Real" vs "Smiling", for both questions.

From these results we can draw the following conclusions. First, the "Predicted" class made the avatar seem like it was understanding the speaker more than the other classes did. It was also perceived as more natural than the others even when compared to the "Real" class. Second, it appears that our hypothesis is verified here. Indeed, data containing more variability in the facial expressions ("Predicted" and "Real") were perceived as more natural than the unchanging ones ("Smiling" and "Neutral"). It is also safe to consider that it made also the avatar look like he was understanding more with the more variant expressions than with the constant ones. In fact, the variability of the expressions generated were also compared between the real and predicted sequences. The predicted sequences have more variability than the real ones. This might also be a cause for the predicted being perceived as more natural than the real ones (1.5 laughs and 0.5 smiles on average in the predicted stimuli compared to 1.2 laughs and 0.25 smiles in the real ones). Another interesting point we can notice, is that adding emotional expressions do indeed improve the naturalness perceived and make the avatar seem to understand better. Indeed, the classes with emotions ("Predicted", "Real" and "Smiling") obtained significantly better scores than the neutral ones.

4. CONCLUSION AND PERSPECTIVE

Two main contributions were presented here. First, a multilevel AV synthesis system by concatenation for amusement expressions, smiles and laughs. Then, a system predicting the behaviour of an artificial agent that should give AV feedback to a speaker with different levels of smiles and laughs. The agent's reactions are predicted based on observations of the speaker's behavior. Both systems were evaluated subjectively using MOS tests and gave satisfying and encouraging results. In our future works we plan on making this system work in real time in an HCI experiment. A real-time reactive virtual agent has already been built that is compatible with our concatenative synthesis system and will be used for this purpose. It is also possible to develop an AV real-time and efficient laughter and smiling detection system ([19, 21]) which could generate the speaker observation sequence.

5. REFERENCES

[1] python-crfsuite. http://python-crfsuite.readthedocs.io/en/latest/. Accessed: 2016-05-14.

[2] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham, and C. Wallraven. Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 277–282, June 2013.

[3] J. A. Bachorowski and M. J. Owren. Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, pages 219–224, 1995.

[4] D. W. Black. Laughter. *JAMA: the journal of the American Medical Association*, 252(21):2995–2998, 1984.

[5] Blender Online Community. Blender - a 3d modelling and rendering package,

[6] R. Burling. *The talking ape: How language evolved*, volume 5. Oxford University Press, USA, 2007.

[7] H. Çakmak. *Audiovisual Laughter Synthesis - A Statistical Parametric Approach*. PhD thesis, University of Mons, February 2016.

[8] 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.

[9] M. Chollet, T. Wortwein, L.-P. Morency, A. Shapiro, and S. Scherer. Exploring Feedback Strategies to Improve Public Speaking: An Interactive Virtual Audience Framework. In *Proceedings of UbiComp 2015*. ACM, 2015.

[10] I. de Kok and D. Heylen. *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I*, chapter When Do We Smile? Analysis and Modeling of the Nonverbal Context of Listener Smiles in Conversation, pages 477–486. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[11] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, and others. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[12] Y. Ding, K. Prepin, J. Huang, C. Pelachaud, and T. Artières. Laughter animation synthesis. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pages 773–780, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems.

[13] K. El Haddad, H. Cakmak, S. Dupont, and T. Dutoit. Laughter and Smile Processing for Human-Computer Interactions. In *Just talking - casual talk among humans and machines*, Portoroz, Slovenia, 23-28 May 2016.

[14] E. Gilmartin, F. Bonin, C. Vogel, and N. Campbell. Laughter and Topic Transition in Multiparty Conversation. In *Proc. SigDial*, pages 304–308, Metz, France, 2013.

[15] P. J. Glenn. *Laughter in interaction*. Cambridge University Press Cambridge, 2003.

[16] E. Holt. Conversation Analysis and Laughter. *The Encyclopedia of Applied Linguistics*, 2013.

[17] J. Lafferty. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. pages 282–289. Morgan Kaufmann, 2001.

[18] R. Meena, G. Skantze, and J. Gustafson. Data-driven models for timing feedback responses in a Map Task dialogue system. *Computer Speech & Language*, 28(4):903–922, 2014.

[19] W. A. Melder, K. P. Truong, M. D. Uyl, D. A. Van Leeuwen, M. A. Neerincx, L. R. Loos, and B. S. Plum. Affective multimodal mirror: Sensing and eliciting laughter. In *Proceedings of the International Workshop on Human-centered Multimedia*, HCM '07, pages 31–40, New York, NY, USA, 2007. ACM.

[20] L.-P. Morency, I. de Kok, and J. Gratch. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84, 2010.

[21] 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 *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '13, pages 619–626, Richland, SC, 2013. International Foundation for Autonomous Agents and Multiagent Systems.

[22] N. Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.

[23] R. R. Provine. Laughing, tickling, and the evolution of speech and self. *Current Directions in Psychological Science*, 13(6):215–218, 2004.

[24] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[25] A. Raux and M. Eskenazi. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. *Proceedings of SIGdial 2008*, 2008.

[26] M. Schroder. The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. *Advances in human-computer interaction*, 2010:2, 2010.

[27] J. Trouvain. Phonetic aspects of "speech laughs". In *Oralité et Gestualité: Actes du colloque ORAGE, Aix-en-Provence. Paris: L'Harmattan*, pages 634–639, 2001.

[28] J. Trouvain. Segmenting phonetic units in laughter. In *Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona: Universitat Autònoma de Barcelona*, pages 2793–2796, 2003.

[29] J. Urbain, R. Niewiadomski, E. Bevacqua, T. Dutoit, A. Moinet, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner. Avlaughtercycle. *Journal on Multimodal User Interfaces*, 4(1):47–58, 2010.