Audio Affect Burst Synthesis: A Multilevel Synthesis System for Emotional Expressions

Kevin El Haddad[1], Hüseyin Çakmak[1], Martin Sulír[2], Stéphane Dupont[1], Thierry Dutoit[1]
[1]TCTS lab - University Of Mons, Mons, Belgium
[2]Technical University of Košice, Košice, Slovakia
kevin.elhaddad@umons.ac.be

***Abstract*—Affect bursts are short, isolated and non-verbal expressions of affect expressed vocally or facially. In this paper we present an attempt at synthesizing audio affect bursts on several levels of arousal. This work concerns 3 different types of affect bursts: disgust, startle and surprised expressions. Data are first gathered for each of these affect bursts at two different levels of arousal each. Then, each level of each emotion is modeled using Hidden Markov Models. A weighted linear interpolation technique is then used to obtain intermediate levels from these models. The obtained synthesized affect bursts are then evaluated in a perception test.**

I. Introduction

Affect bursts is a term describing very short, discrete and non-verbal expressions of affects expressed vocally or facially. They were introduced by Scherer in [1] and also studied by Schröder in [2]. They take a big part in social interactions since they are used to communicate feelings caused by precise events (e.g. a suddenly opened mouth with wide open eyes accompanied by a short ”oh!” sound could express amazement). They, therefore, help one to express (voluntarily or not) their reaction towards certain events, or, to understand one’s feelings and sentiments towards that event.

This makes affect bursts important in Human-Computer Interactions (HCI). Indeed, being able to detect such expressions would give us more information on someone’s emotional state and sentiments. On the other hand, in the framework of conversational agents, being able to synthesize them could potentially create more natural interactions since the virtual agent would be using common, emotional human-like expression.

In this paper, we present our work on audio affect bursts synthesis. Previous work on synthesizing mono- or multimodal emotional speech and even isolated affect expressions, such as laughter, can be found [3], [4], [5]. But to the best of our knowledge, very few work focused on affect bursts synthesis. In fact, some work can be found for instance, regarding synthesis of filled pauses [6]. We will present, here, Hidden Markov Model [7] (HMM)-based multilevel audio affect burst synthesis systems. In fact, this is an attempt at developing a synthesis system capable of generating 3 different affect bursts sounds on several levels of arousal. First, we attempt to accurately model 3 types of affect bursts (representing disgust, startle and surprise) using HMMs, each of them in two different levels of arousal. Then, for each affect burst, the HMMs representing each level are used with a linear interpolation technique in order to synthesize intermediate levels. In what follows we will first present the data that will be used for this work in Section II. We will then present our synthesis systems in Section III. In Section IV, we will expose the perceptual tests that were carried on in order to evaluate our system. We will also present and discuss, in that same section, the results obtained. We will finally conclude and give our perspectives for future work in Section V.

II. Audio Affect Burst Dataset

The data used here are taken from the AudioVisual Affect Burst (AVAB) database presented in [8]. This database contains audio, video and motion capture data of the three previously mentioned affect bursts each recorded on three different arousal levels. To record this database, an actor was asked to produce acted disgust, startle and surprise expressions both vocally and visually. Motion capture and grey scale video data were recorded using the NaturalPoint’s Optitrack system while the audio data was recorded using a Rode podcaster microphone. The motion capture and audio data were later synchronized together as explained in [8]. Optitrack is a marker based system. Reflective markers were placed on the actors face and on a headband. It recorded the data at 100 fps using 12 infrared emitting cameras. Thus, the motion capture data is, in fact, the recorded 3D trajectories of each marker. The audio data was recorded at 44.1 kHz and stored in 16 bit PCM WAV files. From this database and for the work presented here, only audio data of two levels from each affect burst were used. Table I gives the amount of instances used for each affect burst and at each level (Level 1 being the lowest arousal level and Level 2 being the highest). Each instance being the utterance of a single affect burst.

Affect Burst	Arousal Level	Instances
Disgust	Level 1	40
	Level 2	19
Surprise	Level 1	37
	Level 2	34
Fear	Level 1	34
	Level 2	39

TABLE I
Data instances

Fig. 1 shows the density distribution of the affect burst durations. The duration are plotted per affect burst and per

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

arousal level. We first notice that, as described in the literature, these affect burst durations are short and depend on the level. The duration is apparently higher for the higher levels for all three emotions. We also note that the disgust sounds are typically longer than the startle, which themselves are longer than the surprise ones. The first observation we can make from this graph is that disgust sounds have a longer duration than surprise sounds. This latter having longer durations than startle sounds ones. This is true whether we consider L1 or L2. The second observation is that, for all emotions, L2 tends to have a longer duration than L1.

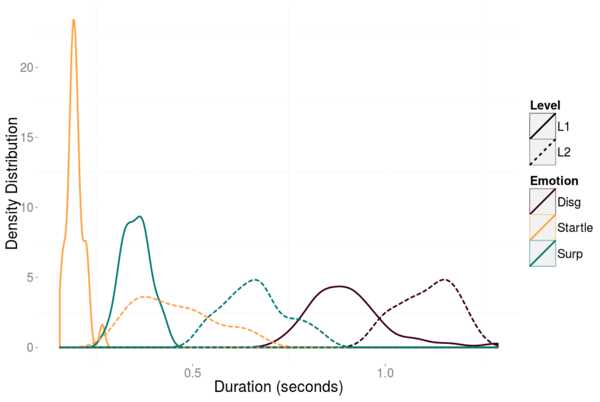

Fig. 1. Duration of the affect burst sounds per emotion and per arousal level.

The probability of voicing was computed using the Snack library [9]. A window of length 10 ms shifted by 10 ms was used for that. The percentage of voicing was then computed per instance. This means that for each affect burst instance, of each emotion, the sum of the voicing probabilities obtained was divided by the length of the total vector. The voicing percentage per instance is given in Fig. 2 per level and per emotion. A plot of the data distribution was preferred here, over a density distribution plot for the sake of a better readability. From this graph, we can see that the disgust sounds in this database tend to have a higher voicing probability than startle and surprise sounds. The startle and surprise sounds tend to contain mostly unvoiced components since a big part of their voicing percentage distribution is equal to zero. Another observation we can make from this graph is that the previous analysis is true for L1 and L2 since they have overlapping voicing percentage distributions.

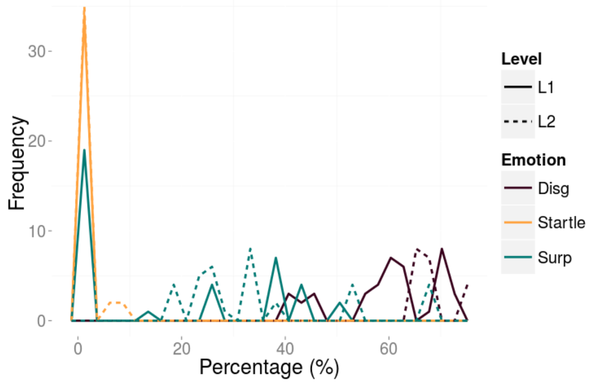

Fig. 2. Probability of voicing per emotion and arousal level

III. Affect Burst Synthesis System

HMM models, with a 5 state left-to-right topology, were trained for each emotion and each level separately. This topology was previously successfully used to represent other types of affect bursts like laughter. The features extracted and used were 25 order MFCCs and the log value of the pitch with their first and second order derivatives. They were modeled in each state with a multivariate single Gaussian distribution. The implementation of this system was made using the freely available HMM-based Speech Synthesis System (HTS) [10]. In order to obtain intermediate levels to those we already have, a weighted linear interpolation technique is used. After the models for each affect burst and each level was trained, the HTS-engine software was used to interpolate between two levels of the same affect burst. The interpolation technique is explained in [11]. The output signal is finally an linearly interpolated intermediate level affect burst.

Using these trained HMMs, 5 levels of each affect bursts were synthesized. First, the two extreme levels were synthesized without interpolation using their corresponding models. They are referred to as int1 (lowest level) and int5 (highest level). Also, three other intermediate levels were synthesized using the already mentioned weighted interpolation feature of hts-engine. The interpolation weights with respect to each synthesized level is shown in Table II.

Synthesized levels	L1	L2
int1	100%	0%
int2	75%	25%
int3	50%	50%
int4	25%	75%
int5	0%	100%

TABLE II
Weights used to obtain intermediate arousal levels from two extreme levels

Column 1 in Table II contains the names of the synthesized levels and the two other columns contain the interpolation weights used with each initial model to generate them.

IV. Evaluations

In order to evaluate our system, three evaluations were set up, one for each of the three emotions. The goals are to evaluate two things:

1) Whether our system is truly capable of synthesizing the affect bursts for the three emotions on different arousal levels.
2) The quality of the synthesized affect bursts, or in other words, whether they can be related to the emotions they are meant to express.

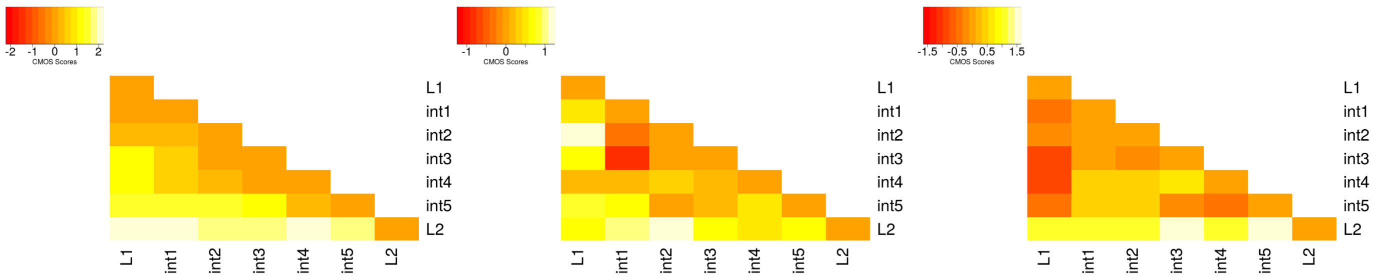

Fig. 3. CMOS results: Heatmaps representing the comparison scores obtained from a perceptual evaluation. L1 and L2 correspond to stimuli using natural audio and motion; int1 to int5 to stimuli with synthesized audio and natural motion.

A. CMOS and MOS tests

To do so, we decided to add facial expressions to the synthesized audio to be evaluated. This is because the synthesized affect bursts, just as the recorded ones, are very short sounds and thus, presenting them to subjects without contextual information such as the facial expression that goes with it would be somewhat confusing for a subject. This statement is based on preliminary informal perception evaluations. So, the facial expressions will be added in the form of a 3D animation. Previously recorded facial expression trajectories from the AVAB database will be directly applied onto the avatar. These facial expressions are thus not synthesized. Please note that the work presented here concerns only audio synthesis and so the synthesis of the facial expressions trajectories is out of the scope of this paper. Each of the synthesized audio affect burst, of each emotion was coupled with a single facial expression of the emotion to which it corresponds, i.e. all the synthesized disgust affect bursts were coupled with the same disgust facial expression (same goes for startle and surprise). There are thus three facial expressions to be applied onto the avatar. For each emotion, the facial expression was chosen from the L1 level in the AVAB database. Facial expressions from the L2 level might have been too expressive and thus might have been too dominant during the perception. To answer our evaluation goals, each of the three evaluation setups was composed of a Comparative Mean Opinion Score (CMOS) test and a Mean Opinion Score (MOS) test. Each setup will proceed as follows. Each participant was shown two videos and two questions at a time. The first question instructed the participants to say which of the two presented videos contains a more intense emotion expression (intense in a general sense) and by how much. A scale below the 2 videos was given with 7 possible choices: 3 on the right in favor of the video on the right, 3 on the left in favor of the one on the left and one in the middle which represents neutrality (both videos are the same). The further on the scale to the right (alt. left), the more the video on the right (alt. left) is perceived as intense. Each of the choices corresponded to a score of integers going from -3 to 3. The negative values corresponding to the video on the left, the positive ones to the one on the right. This question serves as the CMOS test. The second instruction was: Taking into account the animation on the right only, rate how well does the animation suite the emotional expression corresponding to the setup (disgust, startle or surprised). The participants had the choice between one of the following: ”Not at all”, ”a little”, ”average”, ”a lot”, ”completely”. Each of these corresponded respectively to a score or integers going from 1 to 4. This instruction will serve as the MOS test.

The CMOS will help answer our first evaluation goal while the MOS will help answer the second one.

Two supplementary videos are also created for each emotion (therefore for each setup). Similarly to the previously mentioned videos, these videos are created using the same facial expression as the one used in the setup to which they are dedicated, coupled with two randomly selected audio affect bursts taken from the AVAB database. These latter were selected from each of the two considered levels of each emotion. The lower level will be referred to as L1 and the higher as L2. These 2 videos created for each setup will serve as ground truth in the following.

We thus end up with seven videos per setup. As mentioned previously, two videos at a time are presented to a given subject for evaluation and comparison. The combinations of the videos are created randomly in each session and in such a way to avoid that a certain combination is repeated during the same session. Therefore, each participant will be given 21 combinations during his/her session.

B. Results and Discussion

The evaluation platform was online. Each new participant was appointed a different setup. The participants could stop the evaluation process whenever they wanted. At the end, we counted a total of 39 participants making a total of 783 evaluations (294 disgust, 252 startle and 237 surprise). The results of the CMOS and MOS tests are given in Fig. 3 and Fig 4 respectively.

Fig. 3 shows heatmaps representing the comparison results between the affect bursts at different arousal levels and per emotion. In these graphs, and as previously explained the arousal levels are compared per emotion. The colors in the heatmaps represent the total score obtained by each affect burst at a certain arousal level when compared to another. Since, as explained previously, the affect bursts did not receive the exact same amount of votes, the scores were normalized by the total amount of votes each final score received. The more the colors

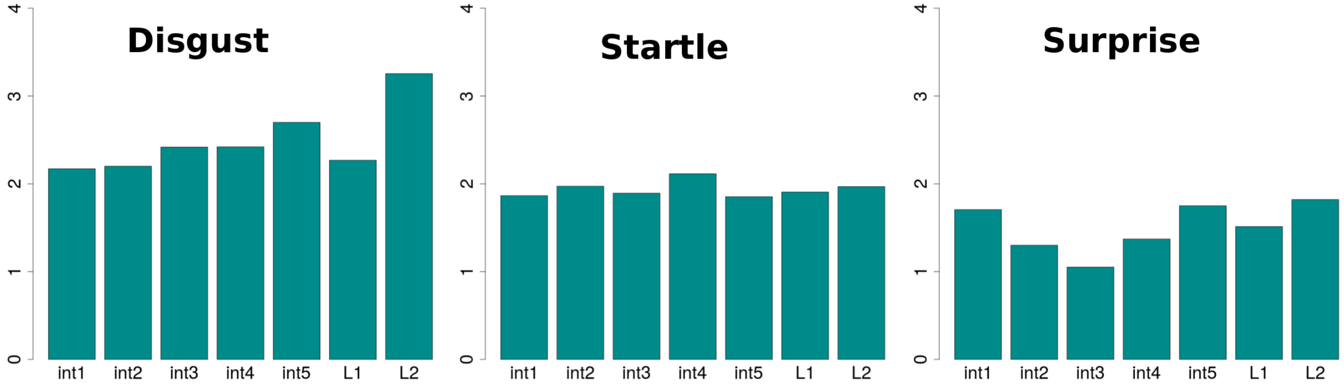

Fig. 4. MOS results: Mean scores obtained per emotion and arousal level from a perceptual test assessing the ability of the sounds to express ”disgust”, ”startle” and ”Surprise”. L1 and L2 correspond to stimuli using natural audio and motion; int1 to int5 to stimuli with synthesized audio and natural motion.

in the heatmaps tend towards the white (and therefore the score is a higher positive number), the more intense the level on the y axis is, compared to the one on the x axis.

Concerning the synthesized affect bursts (int1 to int5), in general, we can see that the higher the arousal level is, the better the difference of level is perceived. We also note that L2 was, in all three cases, perceived as more intense than all other arousal levels. That, even when compared to int5 which is supposed to have the same arousal level. This suggests that the signal quality and naturalness perceived, which we suppose, is higher for L2 (which contains natural audio), plays a role into the perception of the arousal level. On the other side, L1 was correctly perceived for the ”disgust” and ”startle” affect bursts as least intense, even with respect to int1 which is supposed to be at the same arousal level as L1. But L1 was wrongly perceived as more intense than the others in the ”surprise” affect burst case, except when compared to L2. When looking in more details at the ”surprise” heatmap, we can notice that even the synthesized affect bursts scores, when compared to each other, are not that high. So, the difference in levels is more hardly perceived in the case of ”surprise” than in the case of ”disgust” and ”startle”.

It is also important to note that the scores obtained for the ”disgust” affect burst, show clearly that the interpolation did indeed successfully produced intermediate arousal levels for this affect bursts. Indeed, when looking at the scores obtained by the synthesized affect bursts (int1 to int5), we can see that the bigger the difference in level of arousal, the bigger the score obtained by the higher level is. This is not true in all cases concerning the ”startle” affect burst. We can even see some errors in the latter (when comparing int1 to int2 and int3).

Since the facial expressions used for the evaluations were the same for each emotion, and only the audio to which the visual was coupled changed, we can conclude that the weighted linear interpolation performed well in general. The best results being most probably for the ”disgust” affect burst followed by the ”startle” and then the ”surprise” affect burst.

Fig. 4 shows barplots for each of the 3 emotions. These barplots represent the mean score obtained per arousal level. A 95% confidence interval Student t-test was conducted in order to check whether or not the mean values of the obtained scores is statistically significantly different from 0. The p-values obtained were all smaller than 0.01. The mean values of the scores obtained here are therefore all significantly different from 0. After observing the raw data presented in these graphs we can draw some conclusions. First, we can see that the ”disgust” affect bursts, is generally associated to the disgust emotion better than the ”startle” and ”surprise” affect bursts are associated to startle and surprise. This is true even for the L1 and L2 in all cases. All the levels of the ”disgust” affect burst also obtained a mean score above the average (2) while the mean scores are mostly below the average for ”surprise” and below or equal to the average for ”startle”. Although the mean scores are low for ”startle”, it is important to note that the scores obtained by the synthesized affect bursts are roughly equal to the ones obtained by L1 and L2. In order to take into account the statistical significance of these results, a 95% confidence interval Student's t-test was also applied here comparing the mean scores of L1 and L2 with each other and each with int1 to int5 for each of the 3 emotions. Only the results obtained when comparing L2 of ”disgust” with L1 and int1 to int5 was significant ($p - value < 0.01$). All the other test gave a p-value higher than 0.05. This means, on one side, that the synthesized sounds for ”startle” and ”surprise” are not significantly worse than the natural audio and on the other that having more participants might ameliorate the results obtain from this MOS test.

In conclusion, our synthesis system is able to express the selected emotions as well as the natural audio, but with the advantage of controlling the arousal level, as shown in the previous paragraph. This is less clear concerning ”surprise”.

Another important conclusion can be drawn from the difference in efficiency of our HMM-based synthesis systems to synthesize the three types of affect bursts. Indeed, this difference might be explained by the nature of the sounds we attempted to model in this work. Indeed, the surprise and startle are, as presented in Section II, mostly unvoiced in this database on the contrary of the disgust sounds. They are also shorter than the latter. Therefore, the HMMs must model the surprise and startle fundamental frequencies and MFCCs less efficiently than for the disgust sounds.

V. Conclusion and Perspectives

In this paper we presented a first attempt at audio affect burst synthesis using an HMM-based system. We also present a first attempt at synthesizing intermediate arousal levels for different affect bursts using a weighted linear interpolation technique. Three types of affect bursts representing the emotions of disgust, startle and surprise were considered. Our system was tested using a CMOS and a MOS test proving the efficiency of our system mostly for the ”disgust” and ”startle” affect bursts. Our HMM-based synthesis system, proved to be more efficient to model and synthesized voiced sounds rather than unvoiced sounds. In future work, our goals are to obtain a multimodal multilevel affect bursts synthesis system by synthesizing the facial expressions and controlling their level of arousal using the same method described here. We also intend to improve the current audio synthesis system by using more complex interpolation techniques rather than a simple weighted linear interpolation.

Even though we are able to control the duration (not studied in this work) and arousal level of the synthesized sounds using our system, one of its limitations, is that each emotion was expressed using a single type of sound coming (since the data used were similar for each emotion and expressed by the same speaker). Another perspective we have is, therefore, to be able to synthesize a certain emotion, not only on different arousal levels, but also using different types of sounds.

Acknowledgment

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

References

[1] K. Scherer, *Emotions: Essays on emotion theory*, . J. S. E. S. van Goozen, N.E. van de Poll, Ed., Hillsdale, USA, 1994.

[2] M. Schröder, “Experimental study of affect bursts,” *Speech Communication*, vol. 40, no. 12, pp. 99 – 116, 2003.

[3] S. Sundaram and S. Narayanan, “Automatic acoustic synthesis of human-like laughtera),” *The Journal of the Acoustical Society of America*, vol. 121, no. 1, 2007.

[4] 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*, May 2013, pp. 7835–7839.

[5] R. Niewiadomski and C. Pelachaud, *Intelligent Virtual Agents: 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. Towards Multimodal Expression of Laughter, pp. 231–244.

[6] J. Adell, A. Bonafonte, and D. Escudero, *Text, Speech and Dialogue: 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ch. Filled Pauses in Speech Synthesis: Towards Conversational Speech, pp. 358–365.

[7] L. R. Rabiner and B.-H. Juang, “An introduction to hidden markov models,” *ASSP Magazine, IEEE*, vol. 3, no. 1, pp. 4–16, 1986.

[8] K. E. Haddad, H. Cakmak, S. Dupont, and T. Dutoit, “AVAB-DBS: an Audio-Visual Affect Bursts Database for Synthesis,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. C. Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), may 2016.

[9] K. Sjölander, “The Snack Sound Toolkit [computer program webpage],” consulted on September, 2014.

[10] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, “The HMM-based speech synthesis system (hts),” *h ttp://hts. ics. nitech. ac. jp*, 2008.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker interpolation for hmm-based speech synthesis system,” *Acoustical Science and Technology*, vol. 21, no. 4, pp. 199–206, 2000.