AVAB-DBS: Audio-Visual Affect Bursts Database for Synthesis

Kevin El Haddad, Hüseyin Çakmak, Stéphane Dupont, Thierry Dutoit
TCTS lab - University of Mons
31 Boulevard Dolez, 7000, Mons Belgium
{kevin.elhaddad}{huseyin.cakmak}{stephane.dupont}{thierry.dutoit}@umons.ac.be

Abstract

It has been shown that adding expressivity and emotional expressions to an agent's communication systems would improve the interaction quality between this agent and a human user. In this paper we present a multimodal database of affect bursts, which are very short non-verbal expressions with facial, vocal, and gestural components that are highly synchronized and triggered by an identifiable event. This database contains motion capture and audio data of affect bursts representing disgust, startle and surprise recorded at three different levels of arousal each. This database is to be used for synthesis purposes in order to generate affect bursts of these emotions on a continuous arousal level scale.

Keywords: Affect burst, Multimodal Database, Synthesis

1. Introduction

It has been mentioned in several recent studies that adding expressivity and emotional expressions to an agent's communication systems would improve the interaction quality between this agent and a human user. Although emotions can be expressed in several and different ways, most of the work aiming at improving an agent's expressions naturalness, focus on speech (with or without the corresponding facial expressions) (Jia et al., 2011) or isolated facial expressions (Moosaei et al., 2014). In this work we focus on non-verbal audio-visual expressions. Affect bursts are defined in (Krumhuber and Scherer, 2011; Scherer, 1994) as very short non-verbal expressions with facial, vocal, and gestural components that are highly synchronized and triggered by an identifiable event.

Emotional expressions are culturally, socially and even individually variable. This means that a particular emotion will be expressed in different ways and modalities at different times. From this, comes the difficulty of using a naturalistic database with parametric modeling techniques for synthesis purposes. Indeed, the data recorded for a certain emotional expression in a naturalistic database will be different from one another. Therefore, although being adequate for recognition and classification tasks (if the data quantity is enough with respect to the task) this kind of data would most probably not be suitable to create parametric models such as Hidden Markov Models (HMM) (Rabiner and Juang, 1986) or Deep Neural Networks (DNN) (Hinton and Osindero, 2006; Anderson, 1995) with the goal of generating descriptors. This difficulty concerns not only modeling a specific emotion expression without taking the arousal level of the emotion into account, but also doing so on continuous arousal scale. Being able to synthesize emotion expressions on a continuous arousal level scale, will enable more control over the avatar's expressiveness.

In order to do so, we present an Audio-Visual (AV) database collected for the purpose of synthesizing AV expressions at different levels via statistical parametric modeling, i.e. using HMM in our case. Indeed previous work proved HMM-based systems to be efficient for such applications (Cakmak et al., 2015).

We will first discuss our motivations and how do we intend to use this database in Section 2.. The protocol along with the hardware we used to record the database will be exposed in Section 3.. We will present our work on data post-processing, aiming at obtaining suitable data for parametric synthesis (in our case statistical parametric synthesis using HMM) in Section 4.. We will then, in Section 5., give details about the data collected and finally conclude and give our perspectives in Section 8..

2. Database Use and Motivation

As stated in the introduction, previous work related to this topic have already been proposed in (Çakmak et al., 2014) and (Cakmak et al., 2015). In the first, Çakmak et. al. present a database recorded for the purpose of AV laughter synthesis using HMM. In the second, AV laughter has been successfully synthesized using HMM-based systems. Also, an audio laughter synthesis system with a controllable arousal level was also presented in (Urbain et al., 2014). The authors in that article report good results obtained using an HMM-based system.

In order to be able to synthesize emotion expressions on different arousal levels, the database presented here contains facial expression data along with their corresponding audio affect bursts for three different affect states and at three levels each. The three affect states are disgust, startle and surprise. Those three emotions were chosen because they can be expressed relatively easily on different levels of arousal and in both modalities (vocally and through facial expressions). Having three levels of each affect burst, will allow us to not only attempt to synthesize them but also to obtain intermediate levels via interpolation.

Thus, we intend, as a first step, to use the database presented here to model the AV affect bursts present in it. The models will be then used to synthesize AV affect bursts using the HTS software (Tokuda et al., 2008), patch code to the Hidden Markov Model Toolkit (HTK). To fit our purpose of HMM-based synthesis, the data will be processed as described in the remainder of this paper.

3. Recording Protocol

The recordings were made from a single male participant who was instructed to act disgust, surprised and scared using both non-verbal sounds and facial expressions simultaneously. He was asked to produce three levels of each emotional expressions. The experience was repeated several times. The recordings happened in several sessions and in a quiet room. The participant was asked to clap before each session. The claps are used to synchronize the audio and motion capture data (see Section 4.). Indeed, The motion capture and audio data were recorded each on a different machine were therefore asynchronous.

Differences between levels were noted both on the facial expression and on the audio sound. Concerning the facial expression, the higher the arousal level, the more accentuated the expression was. Concerning the audio, a higher arousal level was usually expressed by a higher amplitude and also by a longer duration of the sound in the "disgust" case.

The audio data were recorded using a rode podcaster microphone sampled at 44.1 kHz and stored in a 16 bit PCM WAV format.

The motion capture data were collected using Naturalpoint's Optitrack system. Infrared sensitive cameras capturing data at 100 fps are used in this system with light reflective markers. For our data collection, a 12-camera setup was used with 37 reflective markers. Among those 37 markers, 33 were glued to the participant's face while 4 were fixed on a headband (see Section 5. for the recorded data characteristics).

Also, a 640x480 grayscale video was recorded for each take and synchronously with the motion capture data using the camera facing the participant which will be used for later data segmentation. This data was stored in an AVI file and recorded also at 100 fps. Thus 11 cameras were used to collect the motion capture data and 1 was used to record the grayscale video. Although the grayscale videos are used for the visual data segmentation in this work, they could provide valuable information on the upper body motion and facial expressions of the participant for later use.

4. Data Post-Processing

4.1. Data segmentation

During the recordings, the sessions were stored in separate files. Each session therefore contains several affect burst utterances. In order to serve our purposes and be used in an HMM-based synthesis system, the data in the sessions were segmented. This means that each affect burst event was stored separately in a different file. The audio and motion capture data were segmented with respect to the visual annotations. This is due to the fact that, in this database, it is noted that the facial expressions of all affect bursts began before and ended after or simultaneously with the corresponding audio. Annotations based on the audio data alone are also available. These could be used in further study for audio affect bursts synthesis.

Therefore, before the segmentation, session files of both modalities had to be synchronized since they were recorded on different machines.

For each segmented file of each modality, a label file was created. Each label file contains the beginning and ending times of the affect bursts and the preceding and following silences, as well as labels describing the content of the corresponding data file. This content is the affect burst and its level.

The label files were created in the HTK label format (Young et al., 2006) since we plan on using the HTK software for the HMM-based modeling and synthesis.

4.2. Synchronization

Synchronizing the data in our case, is important for later modeling and synthesizing. As previously mentioned, each session started with a clap which was capture by the grayscale video and by the microphone. This clap was used to synchronize the data. Indeed, the synchronization was made by aligning the frame at which the hands touch in the video and the onset of the corresponding sound in the audio.

Since the frame rate is 100 fps, the synchronization precision is of 10 ms.

5. Data Description

This section gives some details about the data present in our database. Table1 gives the amount of instances collected for each affect burst and at each level(level 1 being the lowest arousal level and level 3 being the highest). By instance we mean the utterance of a single affect burst.

Affect Burst	Arousal Level	Instances
Disgust	Level 1	40
	Level 2	25
	Level 3	19
Startle	Level 1	37
	Level 2	7
	Level 3	34
Surprise	Level 1	34
	Level 2	11
	Level 3	39

Table 1: Data instances

Visual: As explained earlier, the Optitrack system uses infrared sensitive cameras along with 33 face markers and 4 headband markers. The system tracks the 33 face markers and converts them into 3D coordinates of each marker per frame. Thus giving us 99 degrees of freedom per frame representing the facial expression with respect to time. The 4 headband markers have a fixed distance with respect to one another and their positions are not affected by the facial expressions themselves but rather by the head movement these expressions induce. They are thus used to calculate the global head movements: 3D head position coordinates along with 3D Euler angles. This gives us 6 more degrees of freedom making it a total of 105 degrees of freedom representing the motion capture data.

Audio: Concerning the audio data, we expected the intensity of the signal to increase with the level. In order

to quantify this, the mean value of the Root Mean Square (RMS) of the energy computed for each instance of each affect burst was computed. For each instance, this was done using a 10 ms wide window, shifted by 10 ms. The mean RMS energy value was then computed per instance. The density distributions of the mean RMS energy per affect bursts type and per arousal level are shown in Fig. 1.
For disgust, the mean RMS energy values seem to be higher for higher levels. However this is less obvious in the case of startle and surprise. Indeed, for startle, some of L2 and L3 values seem to be higher than L1 values. But the values computed from the two former seem identical. Concerning surprise, the difference between the L2 and L3 levels values and the L1 ones is more obvious than for startle. Even though the values of the two former levels are not identical, the distinction between them is not obvious either since we can see that L3 contains higher and lower values than L2.
Another interesting acoustic aspect of the affect bursts sounds in our database is the fact that some of them are more "voiced" than others. To show this, the probability of voicing was computed using the Snack library (Sjölander, consulted on September 2014) and a 10 ms wide window shifted by 10 ms. The percentage of voicing was then computed per instance. This means that for each affect burst instance, of each level, the sum of the voicing probabilities obtained was divided by the length of the total vector. The density distribution of the voicing percentage per instance is given in Fig. 2 per level and per emotion.
From Fig 2, we can conclude that, in this database, the disgust sounds are mostly voiced since for all levels, generally, more than 55% of each instance is voiced. The startle ones are all unvoiced since 8% is the highest voicing percentage value. The surprise ones are mostly unvoiced since in general less than 50% of the signal is voiced for each instance and most of the voicing percentage values are concentrated around 30% for L2 and L3.

6. PCA for Dimensionality Reduction

The 105 degrees of freedom representing each affect burst per frame seems too much to represent them. Also, from a machine learning point of view, the amount of data representing each level of each affect burst with respect to the dimensionality of the features (degrees of freedom) is relatively large. This may lead to an overfitting the training data by a model. In order to reduce the motion capture features dimensions while also decorrelating the features, a Principal Component Analysis (PCA) is applied.
To choose the order to which the dimension will be reduced, we first compute the reconstruction error for all the possible order values. The reconstruction error is the root mean square error (RMSE) between the original matrix of features from which the reduced feature matrix is created, and the matrix obtained after applying an inverse PCA to the reduced matrix feature.
Fig 3 shows the RMSE evolution with respect to the order to which the dimensions were reduced for each affect and each arousal level.
If we considered these graphs alone, we can see that we can achieve a reconstruction error of around 0.05 for all affect bursts and arousal levels with only 20 components. Also, in Fig. 4, the x axes represent the first n components considered while the y axes represent the percentage of the total variance contained in the n components. For all the affects and arousal levels, 5 components are enough to represent over 90% of the components variability and 11 to represent over 99%.
Table 2 represent the RMSE presented in Fig. 3 for each affect burst at each arousal level. These RMSE values correspond to the order 5 to which the initial features dimensions (allowing 105 degrees of freedom) were reduced.

Affect Burst	Arousal Level	RMSE
Disgust	Level 1	0.056
	Level 2	0.057
	Level 3	0.084
Startle	Level 1	0.080
	Level 2	0.069
	Level 3	0.090
Surprise	Level 1	0.084
	Level 2	0.085
	Level 3	0.222

Table 2: Root Mean Square Error (RMSE) between the initial features and the reconstructed features after a dimensionality reduction via PCA. The initial features were reduced from a dimensionality of 105 to 5 and then reconstructed by multiplying with the inverse transformation matrix.

According to Fig. 4 and Table 2, we can consider that 5 components represent each arousal level for each affect bursts in our database well.

7. Aknowledgements

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.

8. Conclusion and Perspectives

In this work we presented a database suited for audio visual affect burst synthesis for three affects (disgust, startle and surprise) and on three arousal levels each. We exposed our recording protocol and a dimensionality reduction of the motion capture features using PCA. We have shown that a smaller dimension order could be used to model the data since the features with reduced dimensions could be mapped back to its original dimensions quite accurately.
In further work, we intend, on one side, to use this AV Affect-Bursts Database to model and synthesis affect bursts on several arousal levels and ideally on a continuous arousal scale. On the other side we plan on increasing the affect bursts in the database by adding more expressions related to other emotions than those cited here. This database contains two types of segmentation for the same data. One based on the video annotations for bother modalities and the other based on the acoustic annotations for the audio

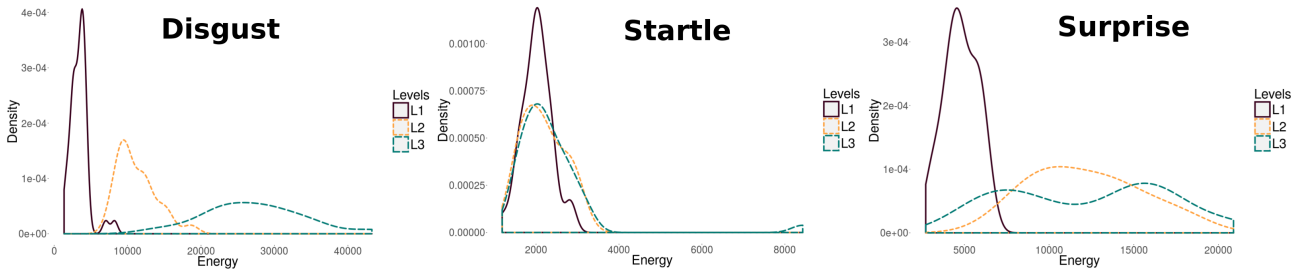

Figure 1: Density distribution of the Root Mean Square energy computed per affect burst and per level.

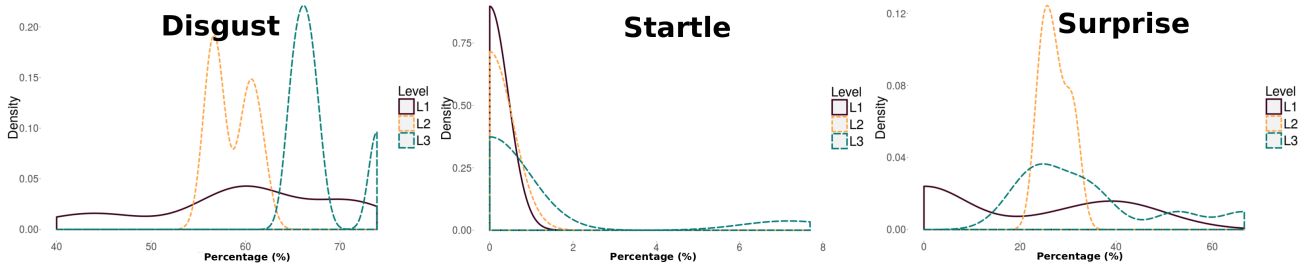

Figure 2: Density distribution of the voicing percentage computed per affect burst and per level.

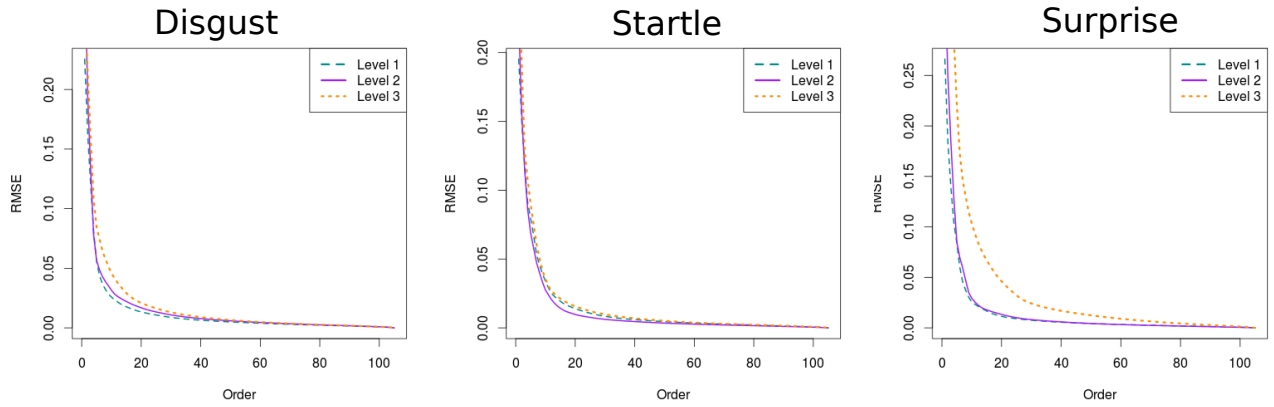

Figure 3: Root Mean Square Error with respect to the dimensionality reduction order.

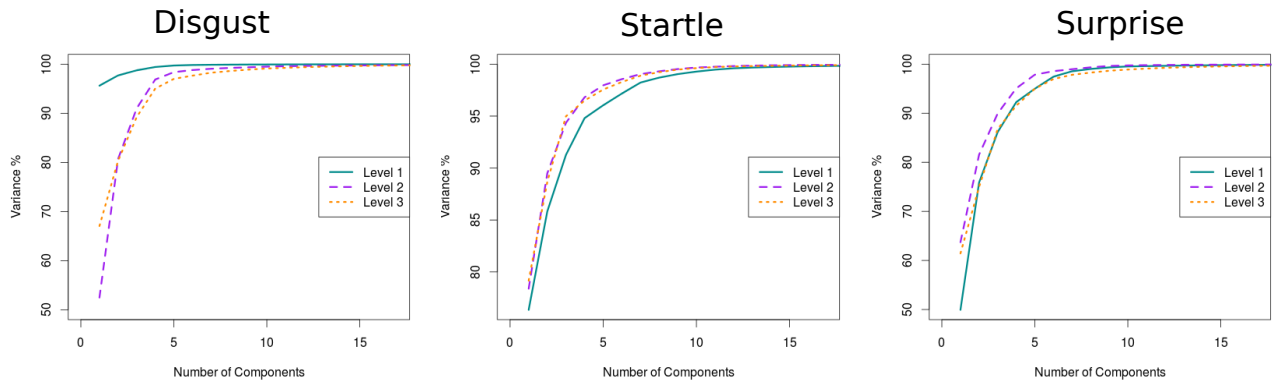

Figure 4: First n components considered with respect to the percentage of the total variance contained in them for each affect bursts and at each level.

modality only. This was done to allow a study on the efficiency of each segmentation method when used for HMM-based audio synthesis (after training the HMMs with the segmented data) and choose the best one for synthesis.

9. Bibliographical References

Anderson, J. A. (1995). *An introduction to neural networks*. MIT press.

Cakmak, H., El Haddad, K., and Dutoit, T. (2015). Audio-visual laughter synthesis system. In *4th Interdisciplinary Workshop on Laughter and Other Non-Verbal Vocalisations in Speech*, pages 11–14, Enschede, Netherlands, 14-15 May.

Hinton, G. E. and Osindero, S. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006.

Jia, J., Zhang, S., Meng, F., Wang, Y., and Cai, L. (2011). Emotional audio-visual speech synthesis based on pad. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(3):570–582, March.

Krumhuber, E. G. and Scherer, K. R. (2011). Affect bursts: dynamic patterns of facial expression. *Emotion*, 11(4):825.

Moosaei, M., Gonzales, M., and Riek, L. (2014). Naturalistic pain synthesis for virtual patients. In Timothy Bickmore, et al., editors, *Intelligent Virtual Agents*, volume 8637 of *Lecture Notes in Computer Science*, pages 295–309. Springer International Publishing.

Rabiner, L. R. and Juang, B.-H. (1986). An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16.

Scherer, K. R. (1994). *Affect bursts*.

Sjölander, K. (consulted on September, 2014). The Snack Sound Toolkit [computer program webpage].

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

Urbain, J., Cakmak, H., Charlier, A., Denti, M., Dutoit, T., and Dupont, S. (2014). Arousal-driven synthesis of laughter. *Selected Topics in Signal Processing, IEEE Journal of*, 8(2):273–284, April.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2006). The htk book (for htk version 3.4).

10. Language Resource References

Çakmak, H., Urbain, J., Dutoit, T., and Tilmanne, J. (2014). The AV-LASYN database : A synchronous corpus of audio and 3d facial marker data for audio-visual laughter synthesis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 3398–3403.