EAI Endorsed Transactions
on [illegible]

# Towards Interactive Agents that Infer Emotions from Voice and Context Information

D. Formolo[1,*] and T. Bosse[1]

[1]1Vrije Universiteit Amsterdam - Department of Computer Sciences, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands

## Abstract

Conversational agents are increasingly being used for training of social skills. One of their most important benefits is their ability to understand the user`s emotions, to be able to provide natural interaction with humans. However, to infer a conversation partner's emotional state, humans typically make use of contextual information as well. This work proposes an architecture to extract emotions from human voice in combination with the context imprint of a particular situation. With that information, a computer system can achieve a more human-like type of interaction. The architecture presents satisfactory results. The strategy of combining 2 algorithms, one to cover 'common cases' and another to cover 'borderline cases' significantly reduces the percentage of mistakes in classification. The addition of context information also increases the accuracy in emotion inferences.

**Keywords:** speech analysis, voice, emotions, context, virtual agents, social skills training.

Received on DD MM YYYY, accepted on DD MM YYYY, published on DD MM YYYY

Copyright © YYYY Author *et al.*, licensed to ICST. This is an open access article distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/3.0/), which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

doi: 10.4108/_____

*Corresponding author. Email:d.formolo@vu.nl

## 1. Introduction

Embodied Conversational Agents (ECAs) can be defined as computer-generated characters 'that demonstrate many of the same properties as humans in face-to-face conversation, including the ability to produce and respond to verbal and nonverbal communication' (Justine Cassell et al., 2000). As research into ECAs is becoming more mature, conversations with ECAs are increasingly being perceived as natural, or at least 'believable'. As a result, there is a growing interest in the use of ECAs for training of communicative skills, such as negotiation, conflict management or leadership skills e.g., (Bosse and Provoost, 2014; Bruijnes et al., 2015; Hays et al., 2012; Jeuring et al., 2015; Kim et al., 2009; Vaassen et al., 2012). The main motivation is that a training system based on conversational agents provides a cost-effective method to replace (or at least complement) human actors, as it can be used anytime, anywhere.

Despite this promising prospect, developing effective conversational agents for communication training is far from easy. An important requirement for effective ECAs is their ability to react to behaviour of the trainee in a similar manner as a human interlocutor would do. Otherwise, there is a risk that the system reinforces the wrong behaviour. For instance, a virtual agent that only listens to you if you address it with a submissive attitude is probably not very useful for leadership training. Hence, making an ECA show the appropriate response to the appropriate behaviour of the trainee is crucial. However, this introduces another challenge, namely to define what is 'appropriate behaviour' of the trainee. Obviously, one relevant aspect of behaviour involves the content of what the trainee says. And indeed, most ECA-based training systems have been designed in such a way that the ECA's responses depend on what the user says (e.g., by analysing the user's speech, or by generating appropriate responses based on selected options within a multiple choice menu).

However, although most ECAs respond to *what* the user says, they often do not respond to *how* the user says it. This is a serious limitation, as the style of a person's speech is very important during social interactions: as discussed in

(Patrik N. Juslin and Klaus R. Scherer, 2005), humans heavily rely on vocal cues (such as volume, or speed of talking) to infer other people's emotions. For example, the phrase 'sorry sir, we cannot accept 100 Euro bills' can be perceived as very friendly when it is uttered calmly and gently, but it can be perceived as offensive when it is uttered with a quick and monotone voice. Especially for communication training it is important to take such differences into account, as it allows professionals to learn not only what to say during their job, but also how to say it. Hence, this paper proposes the use of ECAs for social skills training that adjust their behaviour based on vocal signals that are extracted from the user's speech†.

A second element that is addressed in this paper is the context of the interaction with the ECA. Depending on a context, a particular behaviour of the user could be interpreted completely differently. For instance, a person speaking with a loud and aggressive voice would usually be perceived as angry, but in the context of an important sports competition, these cues could also indicate motivation. With respect to systems that infer emotions from vocal signals, context has only recently been taken into account (e.g., (Wöllmer et al., 2010), (Baur et al., 2016)). Following up on this work, the current paper introduces the term 'Context Information' (CI) as the most likely emotions that a person is expected to have given the current context (e.g., in the context of a conversation between friends, 'happiness' would be part of the CI, whereas 'fear' would not). Based on this concept, one of the goals of this paper is to investigate whether the use of CI increases the accuracy of our ECA in interpreting the emotional state of the human conversation partner.

The paper first introduces, in Section 2, the existing theories about the relation between vocal signals and emotions. Section 3 describes the proposed system in detail, while Section 4 presents the methodology used to measure the performance of the system. Section 5 shows the results of the experiments. The paper is concluded with a discussion in Section 6.

## 2. Emotions in Vocal Signals

Many factors influence the generation of emotion in humans. Emotions can remain stable for a long time or may come and go fast, and sometimes various emotions are mixed at same moment. In the literature, roughly three theoretical perspectives may be distinguished. First, categorical theories are based on the assumption that there is a limited set of basic emotions categories such as joy, sadness, fear, anger, surprise and disgust (Ekman, 1992). Second, dimensional theories view emotions as states that can be represented as points within a continuous space defined by two (or three) dimensions, namely valence and arousal (and dominance) (Russell, 1980; Yik et al., 2011). Arousal refers to a general degree of intensity while valence refers to the level of pleasure. An example of a dimensional theory is Russell's Circumplex Model (Russell, 1980). The author has related different positions in the 2-dimensional space to 28 words that give meaning to the main emotions. Third, componential theories highlight the role of different components that play a role in the emotion generation process, such as the desirability and likelihood of the events that trigger the emotion, cf. appraisal theory (Scherer et al., 2001). Due to its practical appeal, in the current paper, we will mainly make use of the categorical perspective.

Emotions arise from brain circuits involving the amygdala, the orbitofrontal cortex, the insula and various other brain areas (El Ayadi et al., 2011). Affect-related activity in those areas can be reflected in the human voice, which in principle makes it possible to recognise such affective features in human speech.

To realise this in the context of virtual agents, the presented approach builds upon a vast body of previous work. For instance, in (Truong et al., 2012) an approach was put forward to detect emotions in speech in terms of arousal and valence. Similarly, (Lefter et al., 2012) has shown that more specific emotions (e.g., aggression) can be identified as well. Moreover, Rodriguez et al. analyse changes of vocal patterns in humans when they interact with ECAs (Rodriguez et al., 2008). Inspired by these developments, a number of recent systems use vocal cues to trigger the behaviour of virtual agents. For example, in (Bevacqua et al., 2010) vocal cues are used to generate backchannels (i.e., non-intrusive signals provided during the speaker's turn). Acosta and Ward proposed a system that uses speech and prosody variation to build rapport between human and agent (Acosta and Ward, 2011), and Cavazza et al. used vocal signals for character-based interactive storytelling (Cavazza et al., 2009). Furthermore, the virtual human SimSensei Kiosk uses voice, speech and other features to analyse user emotions in the context of healthcare decision support (DeVault et al., 2014). More generally, (van der Wal and Kowalczyk, 2013) used Random Forests to classify vocal signals into a set of emotions.

Most of the above papers do not take context into account. However, some recent works do address context to a certain extent. For instance, (Wöllmer et al., 2010) add to their classification algorithm the bidirectional Long Short-Term Memory recurrent neural networks (BLSTM) to model the context of a conversation. The input nodes of BLSTM correspond to a number of different features per utterance whereas the output nodes correspond to a number of target classes. Recently, (Baur et al., 2016) drew attention to the importance of modelling context into the growing number of conversational systems. They propose a probabilistic framework that provides cues of possible user attitudes, in order to promote a more natural communication between humans and machines. The inputs of their framework are voice, posture and facial expression. Another approach is proposed by (Salam and Chetouani, 2015).

† Obviously, vocal signals are not the only aspect of behaviour that is relevant for communication training. Other aspects include facial expression, gestures, and posture, among others. However, these aspects are beyond the scope of this paper.

Based on assumption that the context of a situation guides the behaviour and emotion of participants, they propose a set of contexts described as Competitive, Informative, Educative, Collaborative, Social, Guidance and Negotiation. They also propose a system that produces an expectation of user reactions, based on one of the above contexts.

As can be observed, these works address emotion recognition through either vocal signals or a combination of sensors, including voice to infer the user's attitudes. All these works are closely related to the proposed system, although most of them don't combine voice signals and context and when they do it, the solution is complex, involving many sensors. Also, they focus on different applications than social skills training. In contrast, one recent system that does focus on communication training (in the context of job interviews) is put forward in (Youssef et al., 2015).

As opposed to existing papers, the current system proposes a simple approach to integrate context within emotion recognition: it assumes that the most likely emotions for a particular context (called Context Information) are known beforehand, and uses this information in the emotion classification process.

## 3. The System

The proposed system is expected to be easily integrated within different serious games or other specialised systems. The final module is a library that is available in the Windows platform as a DLL and that may be extended to Linux-like operating systems. Figure 1 shows an overview of the system (i.e., the ECA`s backend). It contains various modules, including 1) an interface to capture the user's speech (e.g., using a microphone), 2) the off-the-shelf openSmile tool to process this speech (Eyben et al., 2013), 3) a module to disambiguate emotion categories based on CI, and 4) a module to generate a response to the user. Below, the last three of these modules are discussed separately.

### 3.1. openSmile

OpenSmile is a toolkit that extracts and analyses features from vocal signals (Eyben et al., 2013). It can be plugged into other systems as a component of it, providing these relevant values extracted from these features.

In the proposed system, OpenSmile was used as a component to extract 6552 features from voice signals. The extracted features are based on the INTERSPEECH 2010 Paralinguistic Challenge feature set (Eyben et al., 2013). Some vocal features used by openSmile and consequently by the system to analyse emotions are Pitch, Formants and Bandwidth, and Temporal characteristics. The openSmile tool allows researchers to attach algorithms to it to explore the values of extracted features. Many algorithms were tested for the task of classifying samples among the set of basic emotions. The best results were achieved by the Support Vector Machine (SVM) algorithm. Moreover, it is one of the most frequently used algorithms when it comes to emotion recognition in vocal signals, see for instance (El Ayadi et al., 2011)

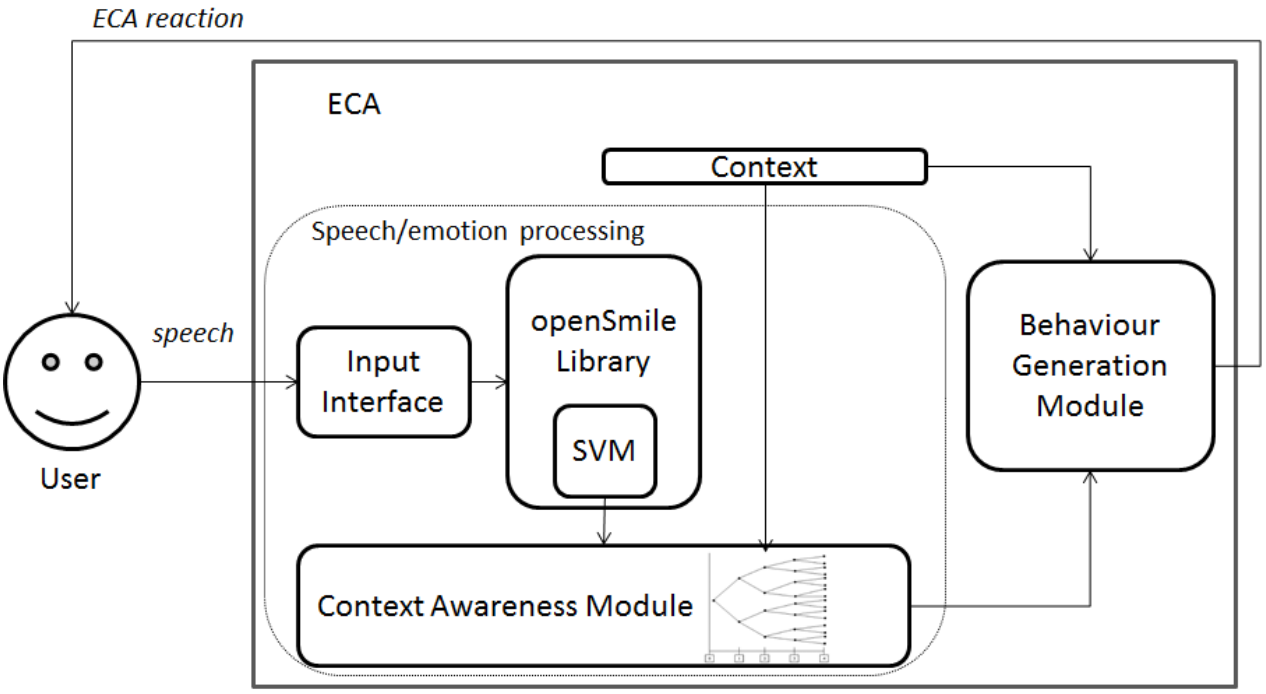



**Figure 1.** Flow diagram of the proposed system

Any model trained by an SVM can easily be updated by another one and even the complete algorithm can be replaced for any other algorithm, because the sub-components are completely independent. One only needs to adjust the connection between the modules.

Currently, this part of the system is responsible for making a preliminary classification of the emotion experienced by the user. It uses the *emodbemotion* model inside the SVM, and the set of emotion categories used is: Anger, Boredom, Disgust, Fear, Happiness, Neutral, and Sadness. For each analysed voice signal, this process assigns a numerical value in the range from 0 to 1 to each emotion. Figure 2 is an example of SVM output. In this case Anger is the emotion that has the biggest value, and is therefore selected as the 'expected emotion' of the user.

```
LibSVM 'emodbEmotion' (#0 = (null)) result [final=0] (@ time: 0.997500, vIdx: 0) : ~~> anger <~~
    prob. class 'anger':       0.960945
    prob. class 'boredom':     0.001419
    prob. class 'disgust':     0.001787
    prob. class 'fear':        0.013305
    prob. class 'happiness':   0.019563
    prob. class 'neutral':     0.002620
    prob. class 'sadness':     0.000362
```

**Figure 2.** Example of SVM output for one voice sample convey Anger.

However, rather than one single emotion, it is expected that voice signals convey a mix of emotions with one prevalent over others. Moreover, the emotion categorisation approach forces the selection of one emotion for each voice signal sample. Although a positive consequence is the simplicity of this approach, it also has a number of drawback, as explained below.

### 3.1.1. Limitations

As mentioned in previous section, approaches that classify emotions according to discrete categories have the advantage that they are simple and pragmatic. Nevertheless, these approaches are susceptible to errors. To illustrate that the SVM models used within OpenSmile sometimes produce wrong or borderline results, consider Figure 3. This figure shows a histogram with values of Boredom measurements to a dataset composed of Boredom samples. Each measured sample was classified into a range of values between 0 and 1 (i.e., like the ones shown in Figure 2), represented by a bin on the x-axis. The y-axis displays the number of occurrences inside each bin. The accuracy to of classifying samples as either Boredom or Not Boredom is 87.65% in this case. An increasing curve trend is expected where bins with higher ranges contain more values than those in lower ranges. The reason why the curve is not higher for the most right bins (0.738 to 1) is because the highest values of boredom collected from this dataset do not exceed 0.738. Notwithstanding, we observe high values in the histogram for the far-right bins (0.492 and 0.615). As expected, wrongly classified samples are concentrated into bins with lower values.

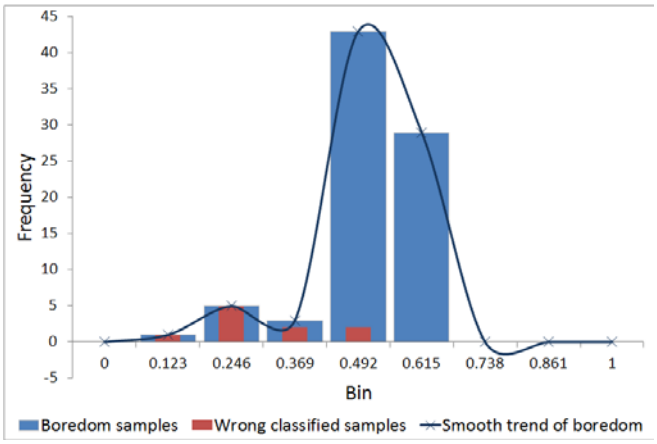



**Figure 3.** Histogram of boredom samples.

Further, when Boredom and Not Boredom samples are put together in the same dataset, and the measurements are plotted in an overlapping histogram, the Not Boredom samples are concentrated into the lowest bins. For these samples with lower Boredom values other emotions scored better, reaching high values, which is reasonable and expected. Figure 4 shows both Boredom samples (in blue) and all samples (Boredom and Not Boredom; in grey) in order to compare the patterns.

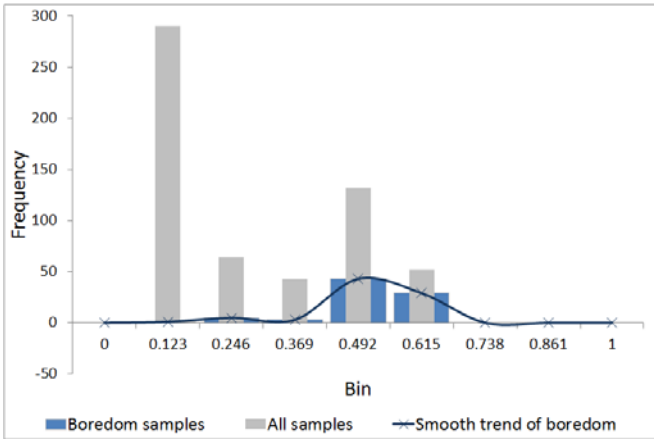



**Figure 4.** Histogram of boredom and all samples.

On the other hand, for the same database, classification of Sadness samples results in a low accuracy, 45%. Plotting the same graphs for Sadness shows distinct results. Figure 5 shows the histogram only for Sadness samples, while Figure 6 compares histograms of Sadness with all samples, including sadness.

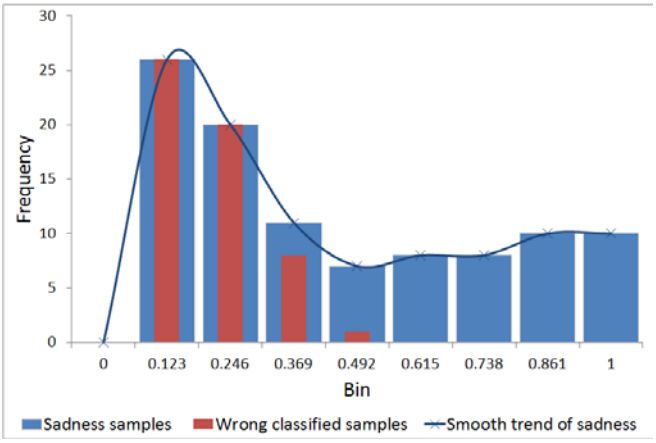



**Figure 5.** Histogram of sadness samples.

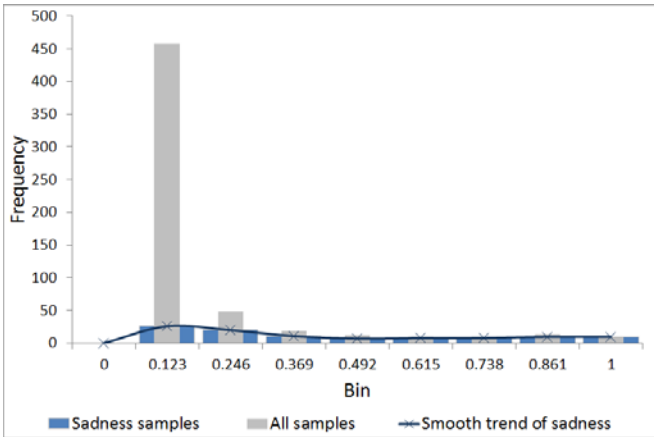



**Figure 6.** Histogram of sadness and all samples.

Boredom and Sadness are two of the most extreme cases. Comparing the shapes of the curves, the peak of Boredom (an emotion that is easy to classify) is for the bin 0.492 and higher, whilst Sadness has its peak for the bin 0.246 and lower. That is an indication of the low accuracy rate of Sadness. Most of the samples tagged as Sadness receive low values for Sadness, what gives the opportunity for other emotions to receive values higher than the ones for Sadness, resulting into wrong classifications. In the case of Boredom, we observe the opposite.

Most classification errors occur to those samples that received a low value for the expected (tagged) emotion. That pattern is also observed in the graphs of Appendix A. Here, similar histograms for all classes of emotions covered in this work are plotted, and analogous to Boredom and Sadness, samples that receive a low value for the target emotion are often wrongly classified.

In order to solve this problem of 'borderline classifications', a new method was introduced into the proposed system. That method, implemented as a Context Awareness Module, still uses the output of the SVM model but combines it with information about the CI in which the conversation takes place.

## 3.2. Context Awareness Module

The Context Awareness Module processes the information received from the SVM. Its task is to deal with ambiguous outputs, which will be done through a decision tree algorithm combined with CI (which is assumed to be present within the ECA's belief base). The motivation for introducing this module is that it is notoriously difficult to distinguish between similar emotions based on vocal signals alone (El Ayadi et al., 2011; Nwe et al., 2003). Besides that, other, environmental factors may contribute to poor measurements of vocal signals, for example noise, distance from the microphone and low emotion expressiveness of users. Those factors may result in many borderline cases, leading to big classification mistakes in some situations, e.g. assuming the emotion anger when in reality the speech expressed a case of fear.

Hence, by using CI, this module intends to minimise problems like that. The Context Awareness Module is fed with the output of openSmile as well as qualitative information about the context. This CI is represented as a prioritised list of emotions that are expected to occur in the current situation. For instance, if the virtual agent has just approached the user with an aggressive attitude, this is represented as a CI with a negative emotion, probably expecting anger as a primary emotion followed by fear as a second option. For training applications, such information is assumed to be available (as it is the application itself that generated the negative stimulus).

### 3.2.1. Decision Trees

This module uses internal decision trees which, in combination with the SVM, are used to decide which emotion category is currently applicable. The crux is that, although SVM produces a set of emotions and their numerical values, the 'correct' (or most applicable) emotion is not necessarily the one with the highest value. For many cases two very different emotions are in the same range of values, varying less than 10%. Some emotions are more uniquely identified than others, what means that the confidence to decide which emotion is applicable is (inversely) linked to the distances between the values of the different emotions. In other words, if there is a large distance between the emotion with the highest value produced by SVM and the one with the second-highest value (e.g., Anger=0.8 and Fear=0.4), then one can be much more certain that this emotion is indeed the 'correct' then when this distance is smaller (e.g., Anger=0.8 and Fear=0.7).

This characteristic is the basis of The Context Awareness Module. In fact, the module consists of a number of decision trees, one for each CI. Depending on the CI, an emotion must have a certain minimum distance to the other emotions to be selected as the applicable emotion.

The module uses a 3-steps process shown in, Figure 7 through a UML activity diagram. The process starts by checking the emotion selected by the SVM. If that emotion is one of the expected emotions according to the Context list

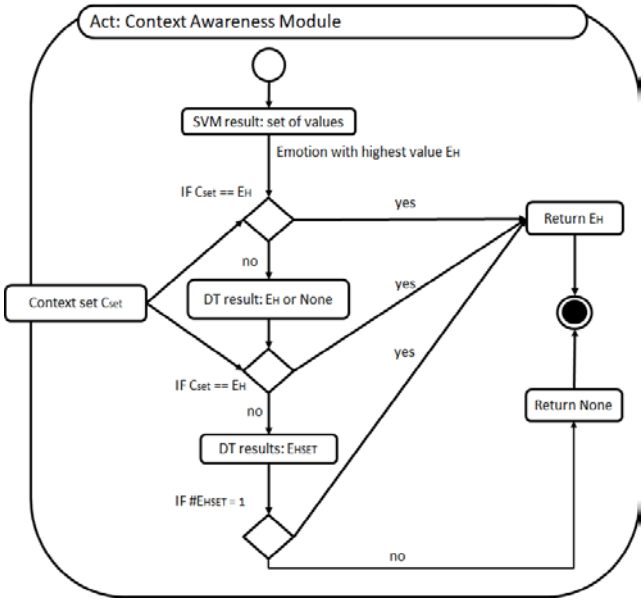



**Figure 7.** Activity diagram of Context Awareness Module.

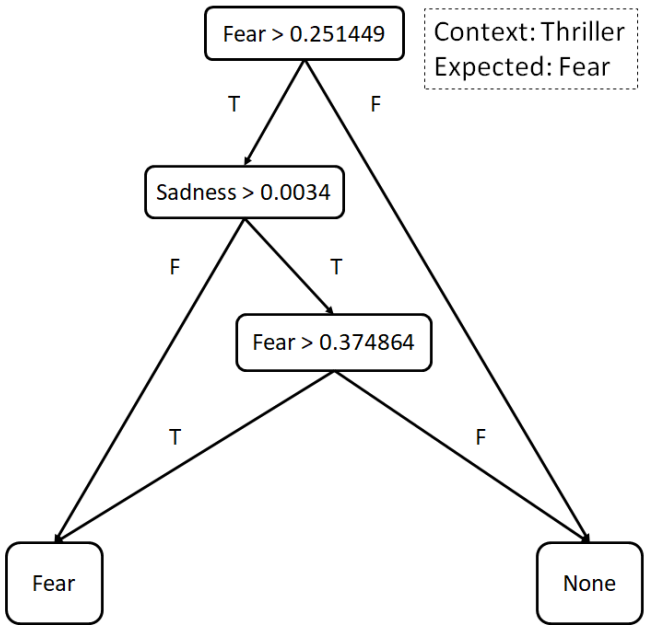



**Figure 8.** Fear decision tree.

$C_{set}$ the module accepts the suggestion of the SVM, returning that emotion. Otherwise, it continues to the second stage, performing the same process using decision trees. In this case, more than one decision tree can match emotions of $C_{set}$, resulting in more than one candidate emotion. If that is the case, the selected emotion $E_H$ will be the one that is ranked as more probable within $C_{set}$.

Even if no emotion was selected, the algorithm moves forward to the third stage, again using decision trees. The third stage opens the possibility to select an emotion outside of $C_{set}$. This is a possibility that might be considered, for example, if the application presents a calm situation, hence it is expected that $C_{set}$ = {happiness, neutral, boredom}, but the user`s reaction can be angry, i.e. outside of $C_{set}$. Angry is unexpected in this example, but still can be inferred by the system if the user`s voice is sufficiently distinctive to express anger. Again, all decision trees are checked, if only one of them returns positively, then the emotion related to that decision tree is selected. For all other cases, when more than 1 decision tree return true or all of them return None, the module takes the most prudent approach and returns None. This should be interpreted as if the collected vocal signal is ambiguous and the best thing to do is to avoid the risk of a wrong classification, hence classifying the emotion as 'unknown'.

To clarify this mechanism, Figure 8 shows an example of a decision tree related to the CI of a Thriller situation. In this case, the most probable emotion is Fear, but openSmile can make a mistake in the classification process, providing high values for Sadness. If that is the case, the decision tree is traversed to find out if the emotion is Fear or not.

For the Thriller example, Fear is the dominant emotion in the majority of situations. In this case, if the value of Fear is lower than 0.251449 then this decision tree concludes that the input is not Fear. If the value of Fear is bigger than 0.251449 then it is necessary to look into Sadness, because the algorithm that generated the Fear decision tree understands that in many borderline cases Sadness is classified as Fear. To avoid mistakes the decision tree checks if the Sadness measurement is low to guarantee that Fear is the emotion conveyed by the current voice sample.

Hence, low values of Sadness (<= 0.0034) indicate a high chance of Fear to be the right emotion. In that case, the decision tree decides that the emotion is Fear. If this is not the case, the decision tree checks once more the value of Fear. If it is bigger than 0.374864, the measurement of Fear is high enough to select it as the right emotion, even with a significant value for Sadness. If the value for Fear is lower than 0.374864 and Sadness has a value bigger than 0.0034, the decision tree limits its choice to "None".

Besides Fear and Sadness, in principle, other emotions could be applicable as well, but to select one of them with sufficient certainty, they have to reach very high values, not only values bigger than the others. Hence, each CI has different thresholds for each emotion.

In this case, the system can also be programmed to explore another decision tree. For instance, the Thriller CI can have other candidates like Anger, Disgust or Neutral, in a pre-defined order, according to pre-programmed definitions. However, for some cases, the values produced by openSmile and SVM may simply be insufficiently decisive to select an applicable emotion category. For these cases, all decision trees that are applied select "None". These situations could occur due to a noisy environment, low quality of the microphone, the way the speaker talks, and of course in situations where the speech simply does not

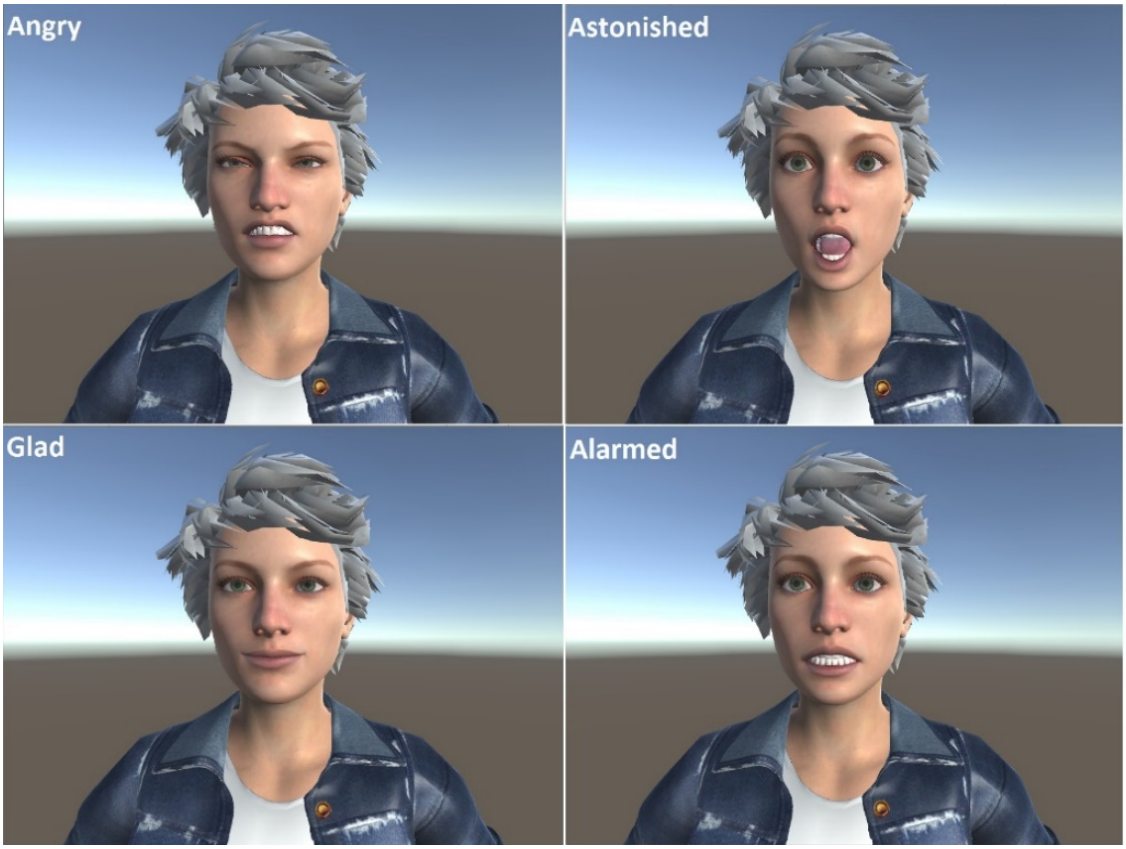



**Figure 9.** Example of recognised emotions visualised through an ECA.

contain very clear affective signals. Also, for long sentences, we can have one emotion for the first part of a sentence and a second emotion for the last part. To deal with such cases where the emotion category is unclear, the Context Awareness module simply does not send any new information to the next module (the Behaviour Generation module).

### 3.2.2. Generating decision trees

Note that the set of decision trees is fixed and, currently, generated in advance. For each CI, we apply the C4.5 decision tree algorithm to a set of samples. The features used for each sample are: expected emotion (target) and values of emotion extracted via openSmile and SVM. OpenSmile provides 7 values, one for each emotion measured by the algorithm implemented in the cLibsvm-LiveSink component of openSmile.

To build the Fear decision tree shown in Figure 8, all voice samples are used. The Fear samples are kept like they are. The remaining samples of the training set are tagged as "None". The C4.5 algorithm produces a pruned decision tree using 2 as minimum number of instances per leaf and a confidence factor of 0.25. The confidence factor is used for pruning, where smaller values incur more pruning. The other parameters are the default values of the J48 Weka classifier (Witten et al., 2016). The same is done for other trees. Appendix B shows all trees that have been built, one per emotion.

## 3.3. Behaviour Generation Module

After the Context Awareness module has selected an applicable emotion, this information is transmitted to the Behaviour Generation Module, which generates an appropriate response to the user. Obviously, this module can be very complex in itself (e.g., including modules for dialogue management and speech generation), but this is outside the scope of this paper. As a simple proof of concept, the Behaviour Generation Module currently just makes the ECA show a facial expression that is similar to the category into which the human emotion is classified. However, in other situations, it might be more effective to respond in a different way to the perceived emotion (see the more extensive discussion below).

To illustrate the working of the system, one prototype application was developed that consists of an ECA that responds to the emotion inferred from the user's voice. Figure 9 shows 4 different emotions expressed by the ECA, which reflect the emotion in the voice of the user. However, the system could also be applied in many different situations in which the ECA not only mirrors the emotions of the user but also shows variations of those, like in negotiations, where a happy emotion from the user could produce an angry reaction from the ECA.

## 4 Methodology

In order to measure the impact of the new Context Awareness module on the performance of emotion classification, 4 types of experiments were performed. They were divided into online and offline experiments. Offline experiments analyse vocal signals from recorded files in perfect conditions, without environmental noise and knowing the duration of utterances in advance. Those conditions allow openSmile to obtain more precise measurements. As a consequence, SVM and Decision Trees perform well in the classification task. In contrast, online samples are susceptible to interferences from a noisy environment, microphone and speaker quality and uncertainty of when the utterance stops, which leads to worse results than offline.

All experiments used the Berlin Database of Emotional Speech (El Ayadi et al., 2011). For online measurements a Lenovo Yoga 2 Pro, model 20226 was used to play and record the samples. During the measurements, the average noise in the environment was: 26dB, with a minimum value of 16dB and the maximum value of 58dB. The modal class interval is (20dB to 30dB) with occasionally peaks out of that range.

Experiment 1 measures the capacity of decision trees separately in solving borderline measurements. It compares the hit rates of SVM with that of decision trees for each emotion, like it was done in Figure 1 to Figure 4. Both SVM and decision trees are evaluated in offline measurements. Experiment 2 compares the results of SVM and the proposed method (which uses a combination of SVM and decision trees) configured to only one CI emotion a time. It uses offline measurements and all samples are applied to each emotion; Experiment 3 follows the same procedure as Experiment 2, but using online measurements; In Experiment 4, we investigate the performance of the proposed system using offline measurements and by adding new emotions in the CI one by one.

Accuracy and Cohen Kappa coefficient are used to measure the performance of SVM and the proposed system in Experiment 2, 3 and 4. Among other applications, Cohen Kappa coefficient removes distortions of performance results to multi-class classifications problems. More about it can be found in (Vieira et al., 2010).

Table 1 describes the concepts of Hits and Fails to calculate accuracy. Ex, Ey and Ez are distinct emotions, while None means no emotion was selected because the sample is considered ambiguous. For instance, in Case 1, the sample is Ex, and the CI also gives Ex as input to the algorithm. Therefore, the expected result of the classification is Ex. Since the (predicted) output is also Ex, this situation can be considered a Hit. However, when the CI is different from the Sample we also consider something a hit if the algorithm`s output is None, because it is an unexpected emotion (e.g., Case 5). In this case it is better if the algorithm does not assume a position than to select a wrong emotion. Nevertheless, if the CI is identical to the Sample, it is considered a fail to predict None (e.g., Case 3).

Table 1: Accuracy results with and without Context Information Definition of Hits and Fails for sample measurements.

| Case | Context Inf. | Sample | Expect | Predict | Situation |
|---|---|---|---|---|---|
| 1 | Ex | Ex | Ex | Ex | Hit |
| 2 | Ex | Ex | Ex | Ey | Fail |
| 3 | Ex | Ex | Ex | None | Fail |
| 4 | Ex | Ey | Ey or None | Ey | Hit |
| 5 | Ex | Ey | Ey or None | None | Hit |
| 6 | Ex | Ey | Ey or None | Ex | Fail |
| 7 | Ex | Ey | Ey or None | Ez | Fail |

## 5 Experiments

Regarding Experiment 1, a comparison for each emotion was performed. Figure 10 shows the results for Fear. This figure shows a histogram with values of Fear measurements for a dataset composed of Fear samples. Each measured sample was classified into a range of values, represented by a bin in x-axis. The y-axis displays the number of occurrences inside each bin. Overlapping with the total number of Fear samples in each bin, also the number of samples wrongly classified by the decision tree and SVM algorithm are shown.

In the bin with the lowest values (0.123 – 0.246), both SVM and decision trees perform equally bad failing in all samples that returned in that range value to Fear. However, they diverge in the next bins with the decision tree approach being much more open to accept Fear when the sample is Fear. In bin (0.246 – 0.369) the decision tree approach still makes some mistakes but less than SVM. For the other bins no more failure occurs while SVM still misses some samples in a lower scale. Most of the emotions follow the same pattern, which is exactly what is expected from the approach based on decision trees. They cover the border line situations, hence solving the misses of SVM.

The only unexpected result came from the decision tree for the ‘Neutral’ category. As shown in Figure 11, it performed equal or worse than SVM. The reason is that, since the neutral category is in between all the others, the measured values are close to each other, which is why the decision tree returns ‘None’ in most cases.

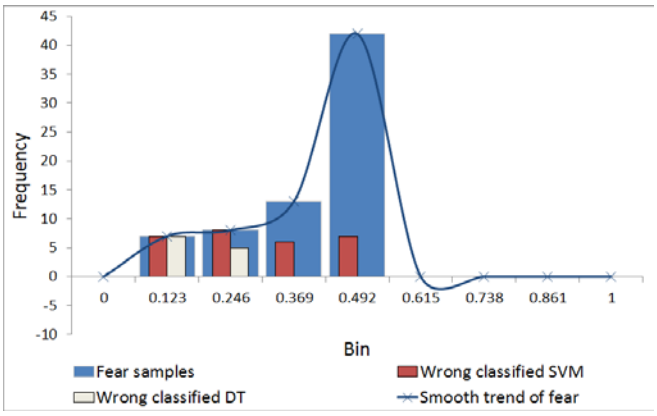



**Figure 10.** Histogram of Fear samples.

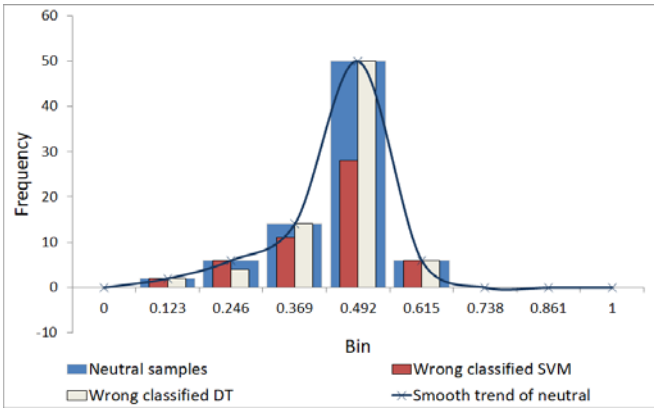



**Figure 11.** Histogram of Neutral samples.

The side effect of a very good borderline classification for samples that are equal to the CI is many False Positives, represented in Table 1 by case 6. Despite the individual weaknesses of SVM and decision trees separately, the combination of both approaches covers the entire spectrum of values for each emotion. Hence, the two approaches complement each other. For this reason, The next experiments compare the proposed hybrid approach with SVM.

Experiment 2 compared the performance of SVM with that of the proposed system, using offline measurements. Table 2 shows the results for SVM per emotion. Each line corresponds to one of possible cases described in Table 1. For instance, the first cell indicates that there were 103 cases in which the CI was Anger, the sample was some other emotion Ey, and the predicted emotion was Ey. Similarly, the fourth cell indicates that there were 241 cases in which the CI was Anger, the sample was Anger, and the predicted emotion was Anger.

Table 2: Performance of SVM for offline samples.

| Case | SVM | | | | | | |
|---|---|---|---|---|---|---|---|
| | Anger | Boredom | Disgust | Fear | Happiness | Neutral | Sadness |
| **1** | **103** | **72** | **30** | **42** | **45** | **6** | **32** |
| 2 | 24 | 14 | 17 | 31 | 27 | 74 | 33 |
| 3 | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| **4** | **241** | **278** | **319** | **276** | **303** | **322** | **325** |
| **5** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** |
| 6 | 17 | 87 | 4 | 70 | 18 | 24 | 12 |
| 7 | 150 | 121 | 193 | 113 | 149 | 116 | 157 |
| | Total of samples: | | 3845 | | | | |

In bold are the hits, whereas the other cells represent fails. For SVM, the overall accuracy is 62.26% and the Cohen Kappa Coefficient is 0.56. Table 3 present the results of the proposed system. For this system, the accuracy is 72.01% and the Cohen Kappa Coefficient 0.67. Hence, for both evaluation measurements, the proposed system increased the performance.

Assuming that the emotion classification process is performed in the CI of a practical human-agent interaction

Table 3: Performance of proposed system for offline samples.

| Case | SVM | | | | | | |
|---|---|---|---|---|---|---|---|
| | Anger | Boredom | Disgust | Fear | Happiness | Neutral | Sadness |
| **1** | **103** | **78** | **36** | **59** | **64** | **24** | **54** |
| 2 | 9 | 3 | 4 | 6 | 6 | 17 | 5 |
| 3 | 15 | 5 | 7 | 8 | 2 | 39 | 6 |
| **4** | **108** | **121** | **144** | **151** | **122** | **156** | **138** |
| **5** | **178** | **148** | **269** | **179** | **130** | **240** | **267** |
| 6 | 17 | 112 | 39 | 134 | 182 | 32 | 34 |
| 7 | 105 | 105 | 64 | -5 | 36 | 34 | 55 |
| | Total of samples: | | 3980 | | | | |

application, we are particularly interested in cases 1 to 3, when samples match with the CI. By far, these situations are the most likely to occur, therefore are the most important and should present good hit rates. Analysing the results, in all cases, the proposed system is equal or better in hits and has lower values to fails compared to SVM alone.

Experiment 3 was equal to experiment 2, but this time the measurements were done online. Not surprisingly, the results of this experiment are worse than experiment 2 (both SVM and the proposed Context Awareness module) due to the external factors that affect the measurements. Nevertheless, the proposed system performed better than SVM again, as shown in Table 4 and Table 5. The accuracy for SVM is 0.28, while the Cohen Kappa coefficient is 0.16. Both are very low, mainly due to many instances of case 7, which is consequence of the more realistic noisy environment. The accuracy for the proposed system is 0.52, and its Cohen Kappa coefficient is 0.44. Both are much higher than the results for SVM.

Table 4: Performance of SVM for online samples.

| Case | Proposed System | | | | | | |
|---|---|---|---|---|---|---|---|
| | Anger | Boredom | Disgust | Fear | Happiness | Neutral | Sadness |
| **1** | **74** | **63** | **5** | **27** | **20** | **5** | **12** |
| 2 | 58 | 29 | 47 | 44 | 51 | 75 | 55 |
| 3 | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| **4** | **141** | **141** | **130** | **105** | **125** | **149** | **134** |
| **5** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** | **N/A** |
| 6 | 128 | 113 | 0 | 87 | 38 | 21 | 9 |
| 7 | 159 | 218 | 386 | 288 | 345 | 324 | 374 |
| | Total of samples: | | 3845 | | | | |

Table 5: Performance of proposed system for online samples.

| Case | Proposed System | | | | | | |
|---|---|---|---|---|---|---|---|
| | Anger | Boredom | Disgust | Fear | Happiness | Neutral | Sadness |
| **1** | **78** | **68** | **5** | **33** | **22** | **8** | **14** |
| 2 | 17 | 1 | 24 | 3 | 26 | 33 | 42 |
| 3 | 37 | 23 | 23 | 35 | 23 | 39 | 11 |
| **4** | **64** | **45** | **58** | **51** | **73** | **77** | **68** |
| **5** | **186** | **186** | **187** | **290** | **178** | **168** | **212** |
| 6 | 127 | 111 | 0 | 119 | 45 | 22 | 22 |
| 7 | 51 | 130 | 271 | 20 | 212 | 227 | 215 |
| | Total of samples: | | 3980 | | | | |

It was anticipated that the proposed system performs better than SVM because it uses an additional source of information (context information). Nevertheless, the results are better than expected, which is not only due to the new CI layer but also to the specific interplay between SVM and decision trees, which balances strengths and reduces the weaknesses of both. In previous experiments, other combinations of the two approaches were tested, which presented poor results, sometimes even worse than using SVM alone.

Experiment 2 and 3 were conducted with only one expected emotion in the CI vector (e.g., a CI where one would only expect 'Fear' but no other emotions). When we add more expected emotions to the CI, the entropy increases, affecting the performance of the proposed system. Experiment 4 adds emotions incrementally to the CI list and compares the performance for each new situation. The results depicted in the graph of Figure 12 indicate that the proposed system has better results than SVM until the point where the CI list contains 2 emotions. After that, the performance declines to similar patterns as for SVM. Curiously, what pushes down the performance of the proposed system are the emotions that are not included in the CI list. For those cases, the Context Awareness module performs much worse than SVM. A third graph in Figure 12 plots only the performance of the emotions that are present in the CI list. As can be observed, for these cases the accuracy is remarkably superior to the others. That result is important because in real situations, emotions that are present in the CI list tend to appear much more frequently than others.


**Figure 12.** Performance comparison – adding emotions step by step in the context information list.

## 6 Discussion

This paper proposes the use of vocal signals that are extracted from the user's speech as one additional component to adjust the behaviour of ECAs. To achieve this goal, we developed an adaptable system that processes human voice and returns a set of emotions and their intensity levels. The system can be easily plugged in into ECAs or other specialised systems that can enrich user experience. Especially for ECAs, the emotional information of a person's voice provides a new element to model their internal behaviour, which may make the interaction between ECAs and humans more natural and effective for training applications.

One specific application of the proposed system is aggression de-escalation training. In this domain, there is an interesting difference between so-called emotional aggression and instrumental aggression. The main difference is that emotional (hot-blooded) aggression is caused by an agent's goals being frustrated, whereas instrumental (cold-blooded) aggression is caused by an agent using intimidation as a means to achieve its goals (Bosse and Provoost, 2014). This distinction is interesting for our system because an emotionally aggressive agent will calm down if the user approaches it empathically. Concretely, this means that the ECA first identifies the emotion conveyed in the user's voice, and if it recognises this as an empathic reaction it will become less aggressive. Similarly, if it interprets the user's utterance as non-cooperative, it will become more aggressive. Instead, for instrumentally aggressive agents this will be the other way around: if such an agent identifies the user's behaviour as empathic, it will become more aggressive, and if it interprets it as non-cooperative, it will calm down. Based on such an application, users could train to take the more suitable conversation style in the appropriate situation. This is very relevant, e.g., for employees in domains such as law enforcement and public transport (Bosse and Provoost, 2014).

The second innovation is the use of CI to extract emotions from human speech more accurately. Often, context conveys crucial information that is neglected by systems and HCI applications. In this paper, context was incorporated in the form of a Context Awareness module. The proposed approach combines 2 algorithms in a complementary way: SVM takes care of most of the common cases, while decision trees are used to solve borderline situations, hence reducing the possibility to make mistakes. The combination of both algorithms is balanced to extract the best from both of them, minimising mistakes and increasing the accuracy, especially for the most important cases, i.e. when the actual emotions are included in the CI list. Besides proposing a specific solution, the presented work can be used as a generic framework, inspiring other algorithms.

Nevertheless, there are circumstances that might limit the use of the proposed system; for example, when the user's environment is noisy or has more than one person speaking at the same time, the system cannot provide precise information. In other cases, the user might not interact much with the system, which could also limit the emotional information extracted by the system. Besides this, it is important to combine the emotional information provided by the user's voice with other sources like facial expressions, gestures and text. In addition, for future work it is necessary to refine the system, develop a method to identify the CI, and test the system in real world applications.

## Acknowledgments

This research was supported by the Brazilian scholarship program Science without Borders - CNPq {scholarship reference: 233883/2014-2}.

## References

Acosta, J.C., Ward, N.G., 2011. Achieving rapport with turn-by-turn, user-responsive emotional coloring. Speech Commun. 53, 1137–1148.

Baur, T., Schiller, D., André, E., 2016. Modeling User's Social Attitude in a Conversational System, in: Emotions and Personality in Personalized Services. Springer, pp. 181–199.

Bevacqua, E., Pammi, S., Hyniewska, S., Schröder, M., Pelachaud, C., 2010. Multimodal backchannels for embodied conversational agents, in: Intelligent Virtual Agents. Springer, pp. 194–200.

Bosse, T., Provoost, S., 2014. Towards aggression de-escalation training with virtual agents: A computational model, in: International Conference on Learning and Collaboration Technologies. Springer, Cham, pp. 375–387.

Bruijnes, M., Linssen, J.M., Akker, H.J.A. op den, Theune, M., Wapperom, S., Broekema, C., Heylen, D.K.J., 2015. Social Behaviour in Police Interviews: Relating Data to Theories, in: D'Errico, F., Poggi, I., Vinciarelli, A., Vincze, L. (Eds.), Conflict and Multimodal Communication, Computational Social Sciences. Springer Verlag, London, pp. 317–347.

Cavazza, M., Pizzi, D., Charles, F., Vogt, T., André, E., 2009. Emotional input for character-based interactive storytelling, in: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1. International Foundation for Autonomous Agents and Multiagent Systems, pp. 313–320.

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., others, 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support, in: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems. International Foundation for Autonomous Agents and Multiagent Systems, pp. 1061–1068.

Ekman, P., 1992. An argument for basic emotions. Cogn. Emot. 6, 169–200.

El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognit. 44, 572–587. doi:10.1016/j.patcog.2010.09.020

Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor, in: Proceedings of the 21st ACM International Conference on Multimedia. ACM, pp. 835–838.

Hays, M.J., Campbell, J.C., Trimmer, M.A., Poore, J.C., Webb, A.K., King, T.K., 2012. Can Role-Play with Virtual Humans Teach Interpersonal Skills? Presented at the Simulation and Education Conference (I/ITSEC).

Jeuring, J., Grosfeld, F., Heeren, B., Hulsbergen, M., IJntema, R., Jonker, V., Mastenbroek, N., van der Smagt, M., Wijmans, F., Wolters, M., others, 2015. Communicate!—a serious game for communication skills—, in: Design for Teaching and Learning in a Networked World. Springer, pp. 513–517.

Justine Cassell, Joseph Sullivan, Scott Prevost, Elizabeth F. Churchill, 2000. Embodied Conversational Agents. MIT Press.

Kim, J.M., Hill Jr, R.W., Durlach, P.J., Lane, H.C., Forbell, E., Core, M., Marsella, S., Pynadath, D., Hart, J., 2009. BiLAT: A game-based environment for practicing negotiation in a cultural context. Int. J. Artif. Intell. Educ. 19, 289–308.

Lefter, I., Rothkrantz, L., Burghouts, G., 2012. Aggression detection in speech using sensor and semantic information, in: Text, Speech and Dialogue. Springer, pp. 665–672.

Nwe, T.L., Foo, S.W., De Silva, L.C., 2003. Speech emotion recognition using hidden Markov models. Speech Commun. 41, 603–623.

Patrik N. Juslin, Klaus R. Scherer, 2005. Vocal expression of affect, in: The New Handbook of Methods in Nonverbal Behavior Research. Oxford University Press, Oxford, pp. 65–123.

Rodriguez, H., Beck, D., Lind, D., Lok, B., 2008. Audio analysis of human/virtual-human interaction, in: Intelligent Virtual Agents. Springer, pp. 154–161.

Russell, J.A., 1980. A circumplex model of affect. J. Pers. Soc. Psychol. 39, 1161–1178. doi:10.1037/h0077714

Salam, H., Chetouani, M., 2015. A multi-level context-based modeling of engagement in human-robot interaction, in: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops On. IEEE, pp. 1–6.

Scherer, K.R., Schorr, A., Johnstone, T., 2001. Appraisal processes in emotion: Theory, methods, research. Oxford University Press.

Truong, K.P., Van Leeuwen, D.A., De Jong, F.M., 2012. Speech-based recognition of self-reported and observed emotion in a dimensional space. Speech Commun. 54, 1049–1063.

Vaassen, F., Wauters, J., Van Broeckhoven, F., Van Overveldt, M., Daelemans, W., Eneman, K., 2012. deLearyous: Training interpersonal communication skills using unconstrained text input, in: European Conference on Games Based Learning. Academic Conferences International Limited, p. 505.

van der Wal, C.N., Kowalczyk, W., 2013. Detecting changing emotions in human speech by machine and humans. Appl. Intell. 39, 675–691.

Vieira, S.M., Kaymak, U., Sousa, J.M., 2010. Cohen's kappa coefficient as a performance measure for feature selection, in: Fuzzy Systems (FUZZ), 2010 IEEE International Conference On. IEEE, pp. 1–8.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques."

Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B., Narayanan, S.S., 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling, in: Eleventh Annual Conference of the International Speech Communication Association.

Yik, M., Russell, J.A., Steiger, J.H., 2011. A 12-point circumplex structure of core affect. Emotion 11, 705.

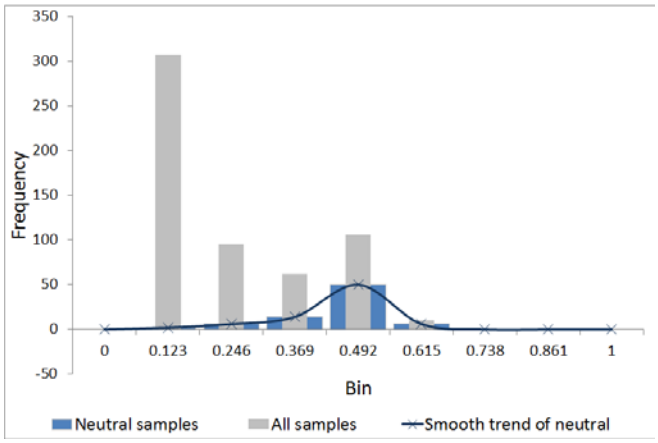
Youssef, A.B., Chollet, M., Jones, H., Sabouret, N., Pelachaud, C., Ochs, M., 2015. Towards a socially adaptive virtual agent, in: International Conference on Intelligent Virtual Agents. Springer, pp. 3–16.

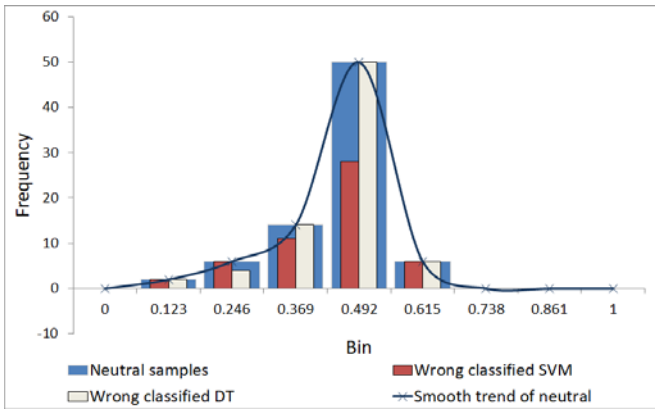
## Appendix A. Histograms

This appendix includes histograms of values measured using SVM for the set of emotions. They are sorted from the emotion that reached the lowest accuracy to the highest accuracy. For each emotion type E, 2 histograms are shown. The first histogram shows the samples tagged with E (in blue) as well as *all* samples (in grey).

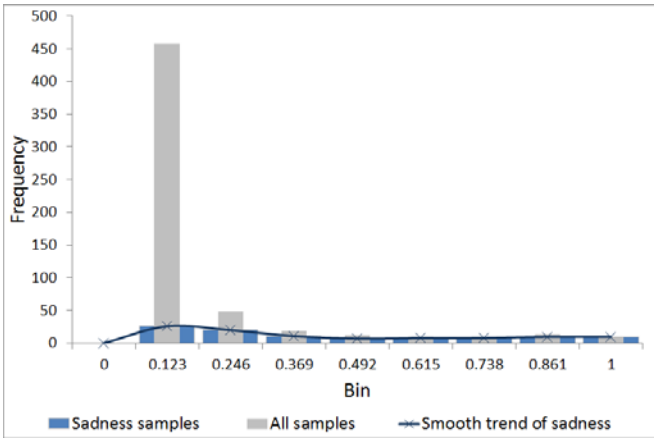
The second histogram shows again the measured values for all samples of E (in blue). It also shows the histograms of wrongly classified samples by the decision tree approach (in grey) and by SVM (in red) when they perform the classification task isolated from each other.

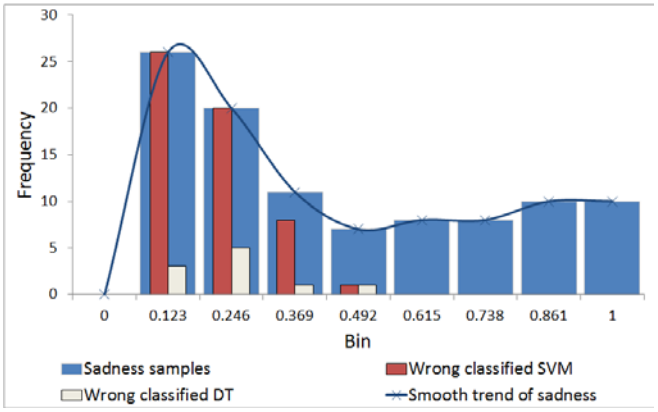
### A.1. Neutral emotion



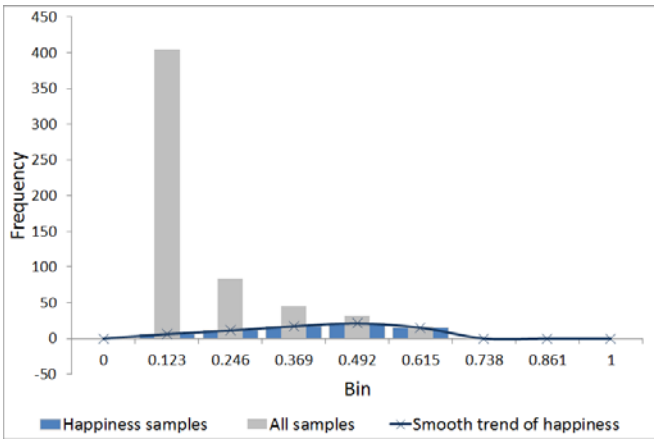



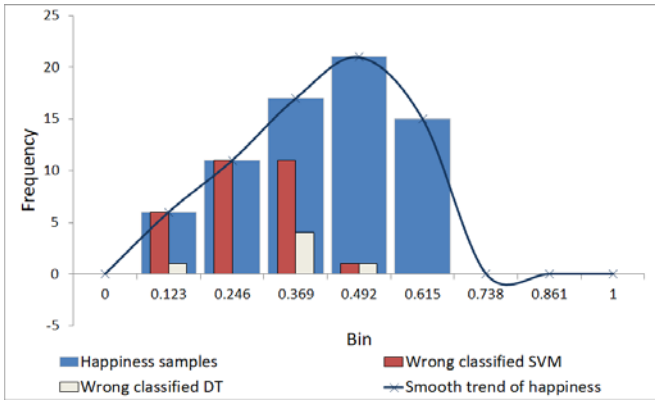

### A.2. Sadness emotion



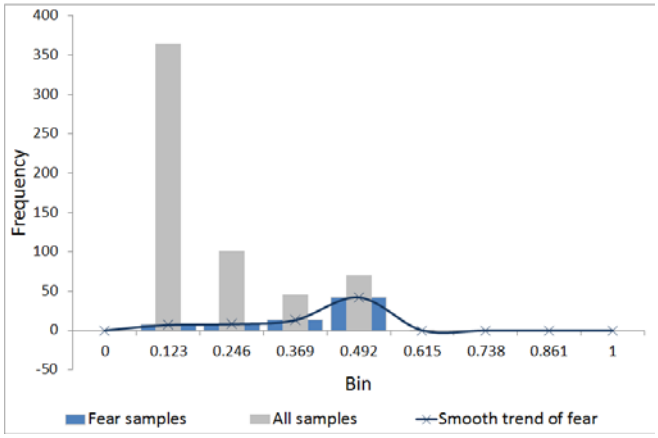



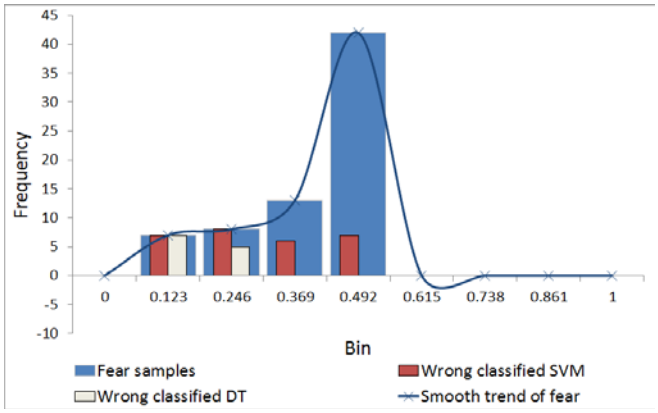

### A.3. Happiness emotion



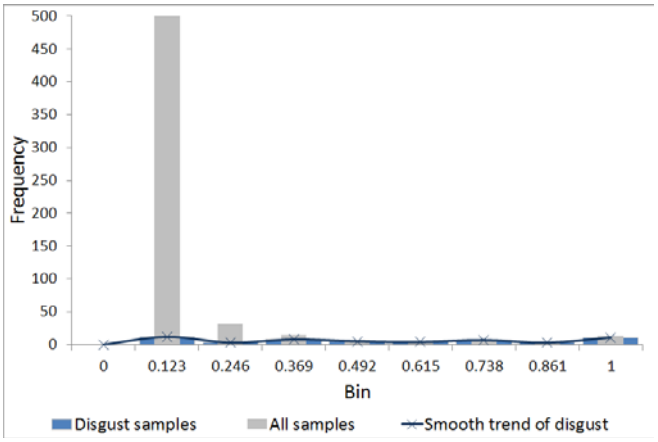


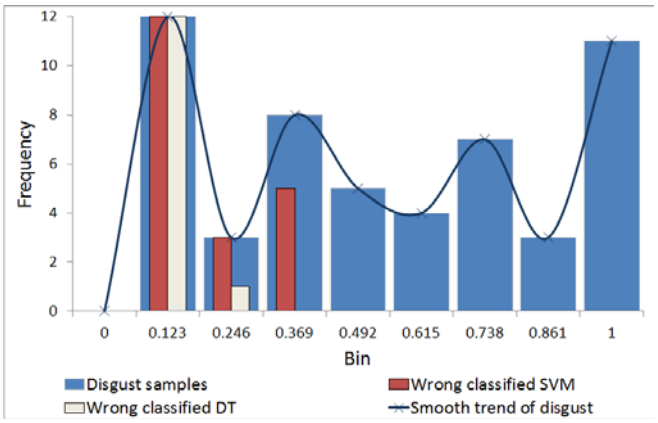

## A.4. Fear emotion



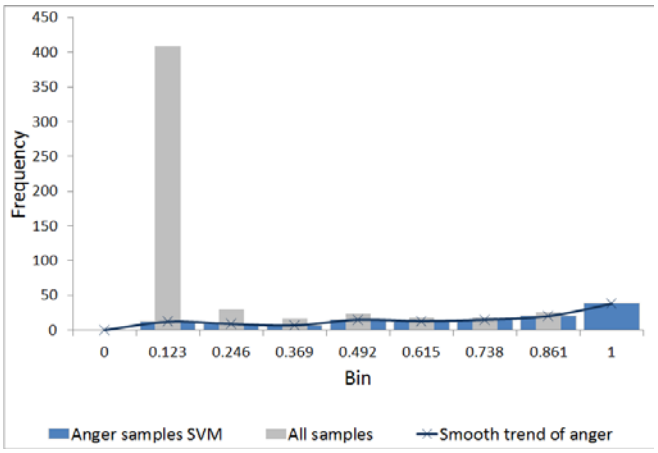



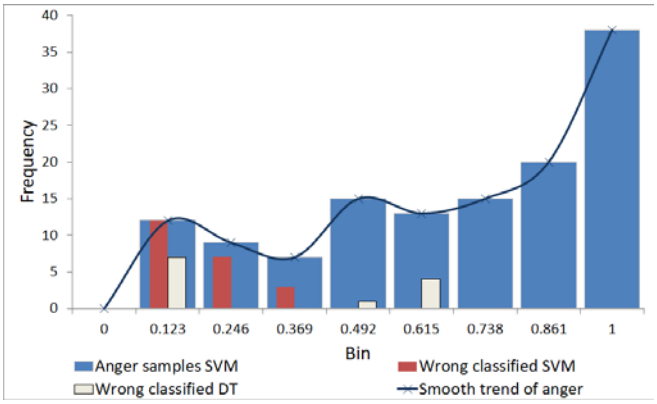

## A.5. Disgust emotion



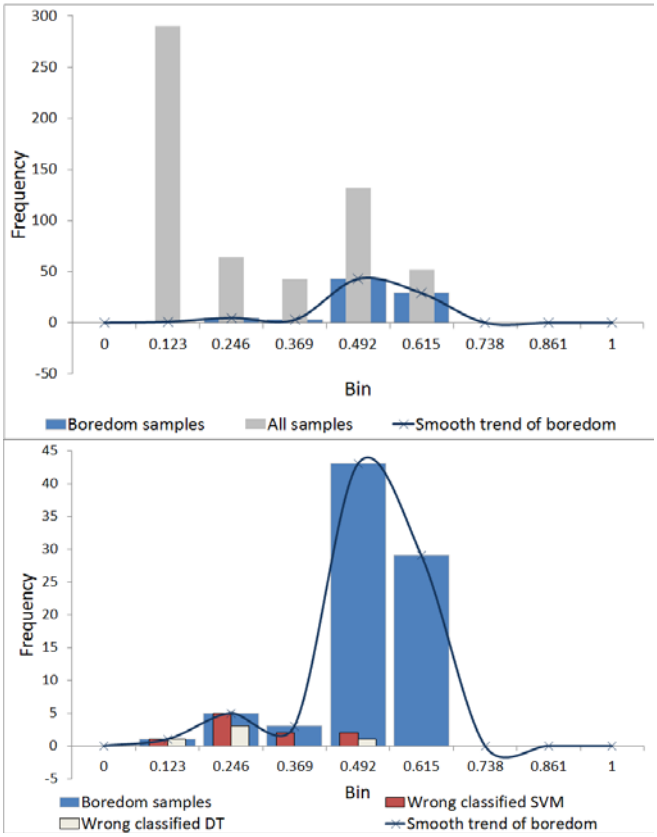




## A.6. Anger emotion






## A.7. Boredom emotion




# Appendix B. Decision Trees

The decision trees used in the validation and currently implemented in the system are shown below.

## B.1. Neutral decision tree

```
neutral <= 0.233205: other
neutral > 0.233205
|   neutral <= 0.451403: other
|   neutral > 0.451403
|   |   neutral <= 0.493349: neutral
|   |   neutral > 0.493349: other
```

## B.2. Sadness decision tree

```
sadness <= 0.03808: other
sadness > 0.03808
|   anger <= 0.021607
|   |   boredom <= 0.4824
|   |   |   disgust <= 0.041588: sadness
|   |   |   disgust > 0.041588
|   |   |   |   boredom <= 0.43149
|   |   |   |   |   disgust <= 0.191108
|   |   |   |   |   |   boredom <= 0.183441: sadness
|   |   |   |   |   |   boredom > 0.183441
|   |   |   |   |   |   |   boredom <= 0.275294: other
|   |   |   |   |   |   |   boredom > 0.275294: sadness
|   |   |   |   |   disgust > 0.191108: other
|   |   |   |   boredom > 0.43149: other
|   |   boredom > 0.4824: other
|   anger > 0.021607: other
```

## B.3. Happiness decision tree

```
happiness <= 0.236942: other
happiness > 0.236942
|   happiness <= 0.410245
|   |   boredom <= 0.047849
|   |   |   happiness <= 0.274921
|   |   |   |   disgust <= 0.050286
|   |   |   |   |   disgust <= 0.021683
|   |   |   |   |   |   anger <= 0.610479: other
|   |   |   |   |   |   anger > 0.610479: happiness
|   |   |   |   |   disgust > 0.021683: happiness
|   |   |   |   disgust > 0.050286: other
|   |   |   happiness > 0.274921: other
|   |   boredom > 0.047849
|   |   |   happiness <= 0.299543
|   |   |   |   happiness <= 0.24349: happiness
|   |   |   |   happiness > 0.24349
|   |   |   |   |   boredom <= 0.078471: happiness
|   |   |   |   |   boredom > 0.078471: other
|   |   |   happiness > 0.299543: happiness
|   happiness > 0.410245
|   |   fear <= 0.396742: happiness
|   |   fear > 0.396742: other
```

## B.4. Fear decision tree

```
fear <= 0.251449: other
fear > 0.251449
|   sadness <= 0.0034: fear
|   sadness > 0.0034
|   |   fear <= 0.374864: other
|   |   fear > 0.374864: fear
```

## B.5. Disgust decision tree

```
disgust <= 0.191108: other
disgust > 0.191108
|  disgust <= 0.384354
|  |  neutral <= 0.032621: other
|  |  neutral > 0.032621
|  |  |  sadness <= 0.031867
|  |  |  |  boredom <= 0.03743: disgust
|  |  |  |  boredom > 0.03743: other
|  |  |  sadness > 0.031867: disgust
|  disgust > 0.384354
|  |  anger <= 0.056122: disgust
|  |  anger > 0.056122
|  |  |  happiness <= 0.091351: other
|  |  |  happiness > 0.091351: disgust
```

## B.6. Anger decision tree

```
anger <= 0.091833: other
anger > 0.091833
|  happiness <= 0.201812
|  |  sadness <= 0.000579: anger
|  |  sadness > 0.000579
|  |  |  neutral <= 0.010465
|  |  |  |  sadness <= 0.003147
|  |  |  |  |  boredom <= 0.00081: other
|  |  |  |  |  boredom > 0.00081: anger
|  |  |  |  sadness > 0.003147: other
|  |  |  neutral > 0.010465: anger
|  happiness > 0.201812
|  |  happiness <= 0.411046
|  |  |  anger <= 0.21423: other
|  |  |  anger > 0.21423
|  |  |  |  happiness <= 0.35847
|  |  |  |  |  boredom <= 0.003078: anger
|  |  |  |  |  boredom > 0.003078
|  |  |  |  |  |  sadness <= 0.011076: other
|  |  |  |  |  |  sadness > 0.011076: anger
|  |  |  |  happiness > 0.35847: anger
|  |  happiness > 0.411046: other
```

## B.7. Boredom decision tree

```
boredom <= 0.433591: other
boredom > 0.433591
|  disgust <= 0.02822
|  |  fear <= 0.004523
|  |  |  disgust <= 0.000727: other
|  |  |  disgust > 0.000727: boredom
|  |  fear > 0.004523
|  |  |  sadness <= 0.033021: other
|  |  |  sadness > 0.033021
|  |  |  |  disgust <= 0.007865: other
|  |  |  |  disgust > 0.007865
|  |  |  |  |  happiness <= 0.011754: boredom
|  |  |  |  |  happiness > 0.011754
|  |  |  |  |  |  sadness <= 0.18092: other
|  |  |  |  |  |  sadness > 0.18092
|  |  |  |  |  |  |  anger <= 0.01183: other
|  |  |  |  |  |  |  anger > 0.01183: boredom
|  disgust > 0.02822
|  |  anger <= 0.075145: boredom
|  |  anger > 0.075145: other
```