"You Seem Aggressive!" Monitoring Anger in a Practical Application

Felix Burkhardt

Deutsche Telekom Laboratories, Berlin, Germany
Felix.Burkhardt@telekom.de

Abstract

A monitoring system to detect emotional outbursts in day-to-day communication is presented. The anger monitor was tested in a household and in parallel in an office surrounding. Although the state of the art of emotion recognition seems sufficient for practical applications, the acquisition of good training material remains a difficult task, as cross database performance is too low to be used in this context. A solution will probably consist of the combination of carefully drafted general training databases and the development of usability concepts to (re-) train the monitor in the field.

Keywords: emotion, classification, detection

1. Introduction

We present a monitoring system to detect emotional outbursts in day-to-day communication. The recognition of emotional vocal expression has gained more and more interest in the academic world for the past few decades (Picard, 1997), (Cowie et al., 2001), (Batliner et al., 2003), (Burkhardt et al., 2006). Until today applications in the real world are mainly restricted to detection of frustration in automated voice portal services (Burkhardt et al., 2007). Although little is known about real voice portals working with this technology, the companies Crealog and NICE offer automated emotion recognition services on their website.

In this project we concerned ourselves with the detection of anger or frustration in day-to-day communication. As an application, we envisage a regulative effect in communication situations by emotional monitoring. A person who is signaled that his/her manner of speaking is classified as being aggressive and thus becomes aware of his negative impression on the communication partner might change this behavior in favor of a more regulated way of expressing his thoughts. It is of great advantage in this situation if the corrective comes from a machine that does not play a part in the situation and shows no emotions in itself.

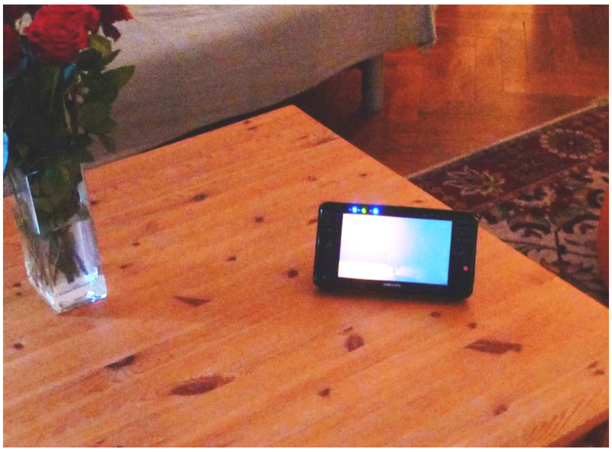


Figure 1: *The emotional monitor in a living surrounding.*

The anger monitor was tested in a household inhabited by two adults and two adolescents for a period of roughly four months in Berlin, Germany. A parallel testing was done in an office surrounding with three colleagues (including the author).

Other possible applications include emotion aware toys and suchlike and are envisaged in (Picard, 2003) or (Burkhardt et al., 2007).

In Figure 1, the anger monitor is shown in the living room. During the testing phase the application, running on a Samsung Q1 Ultra tablet PC, monitored the living room as well as the kitchen during meals. For the office testing phase, the anger monitor run simply on the working laptop with an external PC microphone attached.

This article is structured as follows. Section 2. gives a brief overview on the literature. In Section 3., the application architecture is explained, followed by Section 4., where the features and classifiers are discussed. In Section 5., the Speechalyzer labeling tool gets introduced and Section 6. is about the evluation experiments that we did. Section 7. concludes the article.

2. Literature review

No humans are ever non emotional. We speak emotional, perceive others emotions and communicate emotional. Despite this, contemporary human machine dialog systems always speak with the same unmoved voice and ignore customer's irony, anger or elation. This is partly due to insufficient technological performance with respect to recognition and simulation, and partly to a gap with respect to the necessary artificial intelligence to support emotional behavior.

In figure 2 we display some possibilities of emotional processing in human machine interaction. this topic is further discussed in (Batliner et al., 2006). Emotional awareness can be included in several places of an information-processing system (Picard, 1997):

- **a) Broadcast**: Emotional expression is an important channel of information in human communication. In telecommunication it might be desirable to provide for a special channel for emotional communication. A popular example are the so-called '*emoticons*' used in e-mail communication.

- **b) Recognition**: The human emotional expression can be analyzed in different modalities, and this knowledge is used to alter the system reaction.
- **c) Simulation**: Emotional expression can be mimicked by the system in order to enhance a natural interface or to access further channels of communication, like e.g. uttering urgent messages in an agitated speech style.
- **d) Modeling**: Internal models of emotional representations can be used to represent user- or system states or as models for artificial intelligence, e.g. influence decision making.

In cases a), b) and d), emotional speech analysis can be used to recognize and react on emotional states. Thinking of scenarios, the following lists some ideas:

- Fun applications, e.g. "how enthusiastic do I sound"
- Problematic dialog detection
- Alert systems, i.e. analysis of urgency in speaker's voice
- Adapted dialog and/or persona design
- ....
- Believable agents, artificial humans

This list is ordered in an ascending time line when these applications can be expected. Because a technology has to be developed for a long time before it is stable and able to work under pressure, first applications will be about less serious topics like gaming and entertainment or will be adopted by users that have a strong motivation like elderly people that are able to live independently while being monitored by stress detection systems.

The applications further down the list are closely related to the development of artificial intelligence. Because emotions and intelligence are closely mingled (Damasio, 1994), great care is needed when computer systems appear to react emotional without the intelligence to meet the user's expectations with respect to dialog abilities.

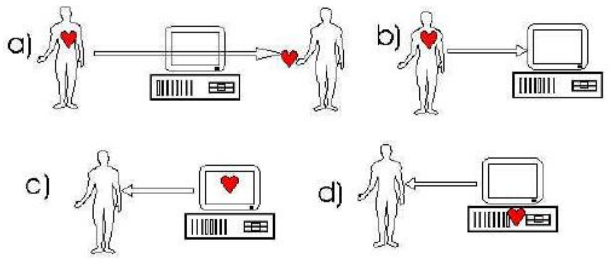


Figure 2: *Possibilites for emotional processing in human machine interaction.*

Clearly, the application described in this article belongs to category b) as human emotion gets detected but also category c), as emotion gets mirrored back. The next paragraphs discuss the technical aspects of feature extraction and classification of audio signals.

Principally most classification algorithms for the detection of anger are based on a three-step approach (Witten and Frank, 2005a): First, a set of acoustic, prosodic, or phonotactic features are calculated from the input speech signal. In a second step different classification algorithms, e.g. Gaussian Mixture Models (GMMs, e.g. (Neiberg et al., 2006), (Burkhardt et al., 2005), (Lee and Narayanan, 2005)), Support Vector Machines (SVMs, e.g. (Shafran and Mohri, 2005), (Shafran and und M. Mohri, 2003)) or other vector clustering algorithms like k-nearest neighbor (KNN, e.g. (Sato and Obuchi, 2007), (Lee and Narayanan, 2005)) or linear discriminant analysis (LDA, e.g. (Blouin and Maffiolo, 2005)) are applied to derive a decision whether the current dialog turn is angry or not angry. Finally, postprocessing technologies can be utilized for consideration of time dependencies of subsequent turns or for combination of the results of different classifiers. All these algorithms heavily depend on the availability of suitable acoustic training data that should be derived from the target application.

With respect to the features that are used to classify the speech data, mainly prosodic features, often in conjunction with lexical based and/or dialog related features, were investigated (e.g. (Burkhardt et al., 2005), (Lee and Narayanan, 2005), (Shafran and Mohri, 2005)), while newer studies also include spectral features derived from Mel Frequency Cepstral Coefficients (MFCCs), e.g. (Shafran and und M. Mohri, 2003), (Blouin and Maffiolo, 2005), (Neiberg et al., 2006) or (Sato and Obuchi, 2007).

There is quite a difference between telephone data as investigated by (Lee and Narayanan, 2005), (Burkhardt et al., 2006) (Shafran and Mohri, 2005) or (Blouin and Maffiolo, 2005) and speech recorded with high quality microphones, as noted e.g. by (Neiberg et al., 2006) in a direct comparison. The difference between real life data and acted speech is so big that a direct comparison does not seem to make sense, e.g. (Sato and Obuchi, 2007) report recognition results for acted emotions far better than those reported on voice portal data.

3. The anger monitor application

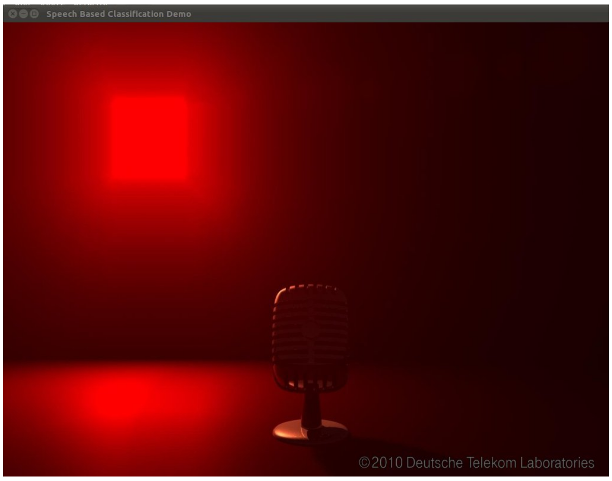

Speech Based Classification Demo
©2010 Deutsche Telekom Laboratories

Figure 3: *Anger monitor interface with aggression detected.*

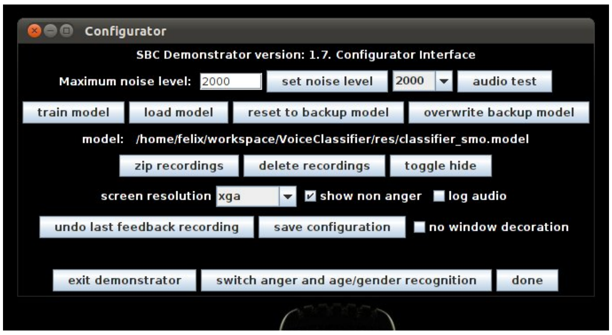
The anger monitor is implemented as a Java application that interfaces C-libraries. We have compiled versions for Windows, Linux and Mac operating systems. The anger monitor interfaces modules to extract the acoustic features and to classify the feature vectors.

Speech is detected by an intensity threshold. Audio gets recorded for analysis while the audio values do not fall below the threshold for a certain time (1 second in the current version) or a pre-defined time has elapsed (6 seconds in the current version). By this, continuing speech gets automatically chunked.

In Figure 5, the overall system architecture is depicted. The audio signal is recorded and, in the case of building a training database, stored on the filesystem. The Speechalyzer labeling tool ((Burkhardt, 2012)) is used to label large amount of emotional speech data in a fast and efficient way, see Section 5. for details.

The audio data gets then feature extracted. In an early version, the Praat system (Boersma, 2001) was used as a feature extractor, but soon replaced by the OpenEAR (Eyben et al., 2009) system because the performance of the processing speed has shown to be much higher. OpenEAR is a toolkit developed by the Technical University of Munich to extract features like FFT-Spectrum, Autocorrelation Function, Mel-Spectrum, Cepstral, Pitch Fundamental Frequency, Probability of Voicing, Harmonics-To-Noise Ratio, Time Signal, Spectral, LPC, Formants and Musical features.

The WEKA (Witten and Frank, 2005b) library is currently used for classification, in a later stage of the project this might be replaced by an own implementation. In Section 4., the choice of features and classifiers is discussed.



SBC Demonstrator version: 1.7. Configurator Interface

Maximum noise level: 2000

2000

model: /home/felix/workspace/VoiceClassifier/res/classifier_smo.model

screen resolution xga

☑ show non anger

□ log audio

□ no window decoration

Figure 4: *The configuration interface.*

The GUI of the anger monitor is displayed in Figure 3. It consists simply of a picture of a microphone that gets highlighted when recording is active and two light bulbs that glow, if either anger or non-anger is detected. We envisage GUIs more suitable for a home surrounding, e.g. a digital picture frame that glows red when aggression is detected, but haven't realized them yet.

As can be seen in Figure 4, a configurator can be opened by the user in order to train the acoustic model based on the last recordings. This enables the user mainly to adapt the audio recording settings to the environment. Furthermore, some train- and test runs can be done directly from the configurator, in order to adapt the acoustic aggression model to persons that will use the monitor often.

A tap on the microphone (de)activates the monitor and for training purposes all recordings might be stored for later labeling. In order to train the monitor on the fly, the user can simply tap the lower left or right corner to categorize the last recorded utterance as either angry or not.

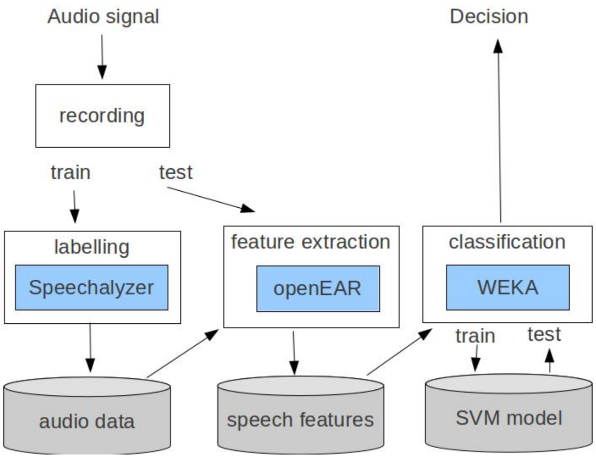


Figure 5: *Overview of the system architecture.*

4. Features and classifiers

As stated in Section 3., we use the openEAR toolkit from the Technical University of Munich for audion feature extraction. The set of extracted features we currently use is the same reduced set of 450 features that was used in the 2010 Paralinguistic Challenge (Schuller et al., 2010) and was achieved by a feature reduction process done on a brute-force analysis including Signal energy, The set is shown in Table 1.

Table 1: *The feature set resulting from the combination of low level descriptors and functionals extracted by openEAR that are currently used. LSP: line spectral pairs*

Descriptors	Functionals
MFCC [0-14] LSP Frequency [0-7] F0 by Sub-Harmonic Sum. F0 Envelope Voicing Prob. Jitter local Jitter consec. frame pairs Shimmer local	arith. mean, std. deviation skewness, kurtosis percentile 1/99 percentile range 99–1

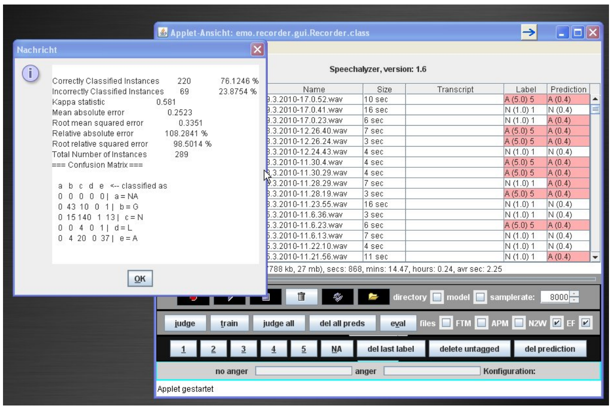
The WEKA (Witten and Frank, 2005b) library is used for classification. We've run some tests with Naive Bayes, CART based j48 classifier and Support Vector Machines. The best results were achieved with the SMO (sequential minimal optimization) SVM classifier using the polynomial kernel.

5. The "Speechalyzer" labeling tool

In addition, we developed a speech analysis and annotation tool named "Speechalyzer" that is shown in Figure 6 ((Burkhardt, 2012)). It can be used for fast labeling of collected acoustic data, training of models with several classifiers, again by interfacing openEAR (Eyben et al., 2009) and WEKA (Witten and Frank, 2005b), and to evaluate the different data sets and models.

One problem of labeling anger in speech is that it normally appears rarely and large amounts of data must be listened into in order to detect relatively small amounts of anger. Therefore, the data was classified in a first step with previously trained models and then only the recordings surrounding detected anger were looked into.

As can be seen in the figure, for each recording the predicted as well as the annotated anger is marked by highlighted table cells, so the performance can be visualized easily and problematic utterances analyzed manually.


Applet-Ansicht: emo.recorder.gui.Recorder.class
Nachricht
Speechalyzer, version: 1.6

```
Correctly Classified Instances       220          76.1246 %
Incorrectly Classified Instances      69          23.8754 %
Kappa statistic                    0.581
Mean absolute error                    0.2523
Root mean squared error                    0.3351
Relative absolute error             108.2841 %
Root relative squared error            98.5014 %
Total Number of Instances             289
=== Confusion Matrix ===

  a   b   c   d   e   <-- classified as
  0   0   0   0   0 |   a = NA
  0  43  10   0   1 |   b = G
  0  15 140   1  13 |   c = N
  0   0   4   0   1 |   d = L
  0   4  20   0  37 |   e = A
```

Name	Size	Transcript	Label	Prediction
9.3.2010-17.0.52.wav	10 sec		A (5.0) 5	A (0.4)
9.3.2010-17.0.41.wav	16 sec		N (1.0) 1	N (0.4)
9.3.2010-17.0.23.wav	6 sec		N (1.0) 1	A (0.4)
8.3.2010-12.26.40.wav	7 sec		A (5.0) 5	A (0.4)
8.3.2010-12.26.24.wav	3 sec		A (5.0) 5	A (0.4)
8.3.2010-12.24.43.wav	4 sec		N (1.0) 1	N (0.4)
8.3.2010-11.30.4.wav	4 sec		A (5.0) 5	A (0.4)
8.3.2010-11.30.29.wav	4 sec		A (5.0) 5	A (0.4)
9.3.2010-11.28.29.wav	7 sec		N (1.0) 1	A (0.4)
8.3.2010-11.28.19.wav	3 sec		A (5.0) 5	A (0.4)
8.3.2010-11.23.55.wav	16 sec		N (1.0) 1	N (0.4)
5.3.2010-11.6.36.wav	3 sec		N (1.0) 1	N (0.4)
5.3.2010-11.6.23.wav	6 sec		A (5.0) 5	A (0.4)
5.3.2010-11.6.13.wav	7 sec		N (1.0) 1	N (0.4)
5.3.2010-11.22.10.wav	4 sec		N (1.0) 1	N (0.4)
5.3.2010-11.21.56.wav	11 sec		N (1.0) 1	A (0.4)

788 kb, 27 mb), secs: 868, mins: 14.47, hours: 0.24, avr sec: 2.25

directory □ model □ samplerate: 8000

files □ FTM □ APM □ N2W ☑ EF ☑

no anger anger Konfiguration:

Applet gestartet

Figure 6: *The Speechalyzer graphical interface.*

6. Evaluation

The anger monitor was tested in a household inhabited by two adults and two adolescents over a period of roughly four months in Berlin, Germany, with the Q1 Ultra tablet. A parallel testing was done in office surroundings using a laptop PC with external microphone with three colleagues (including the author). During a training phase, all detected speech chunks were recorded and labeled with the aforementioned Speechalyzer by the author. We distinguished between three classes: angry speech, non-angry speech and non speech.

The application showed a parallel entertaining effect to the adolescents who tried to provoke anger detection by the machine, so of course this data doesn't only contain real but also acted anger. Both databases contained about 25 % anger, 15 % non speech and 60 % non-angry speech.

In Table 2, the number of collected chunks as well as the accuracy for the two databases for two conditions is shown. Firstly, the mean accuracy for a tenfold cross validation and secondly, the accuracy when one database is used as a test set and the other for training. As can be seen, the accuracy within one database is sufficiently high for a non-critical application like this, whereas the performance drops critically when the classifier gets trained with data recorded with different recording equipment (Q1 Ultra internal microphone vs. external PC microphone), under different acoustic conditions and containing different speakers.

7. Conclusions

Table 2: *Overview of the two collected data sets.*

Location	# chunks	acc. 10 fold	acc. other train
Home	289	80.27	53.59
Office	385	86.23	53.14

We investigated the practical application of anger monitoring in private households and office surroundings. Although the state of the art of emotion recognition seems sufficient for practical applications, the acquisition of good training material remains a difficult task, as cross database performance is too low to be used in this context. A solution will probably consist of the combination of carefully drafted general training databases and the development of usability concepts to (re-) train the monitor "in the field". Of course, these kind of applications have serious ethical and security risks, because speech data gets recorded without explicit awareness of the users. The adolescents expressed the wish to turn the anger monitor recording off while in training mode as often as possible in order not to be bugged. Furthermore it must be clear to the users that the application can only indicate the possibility of existent aggression and is not to be taken too seriously as it's simply based on the statistical analysis of acoustic data.

8. References

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nth. 2003. How to find trouble in communication.

Anton Batliner, Felix Burkhardt, Markus van Ballegooy, and Elmar Nöth. 2006. A taxonomy of applications that utilize emotional awareness. In *Proc. of the Fifth Slovenian and First International Language Technologies Conference Ljubljana*, pages 246–250.

C. Blouin and V. Maffiolo. 2005. A study on the automatic detection and characterization of emotion in a voice service context. *Proc. Interspeech, Lisbon*.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber. 2005. An emotion-aware voice portal. In *Proc. Electronic Speech Signal Processing ESSP, Prague*.

F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson. 2006. Detecting anger in automated voice portal dialogs. *Proc. ICSLP, Pittsburgh*.

Felix Burkhardt, Richard Huber, and Batliner Anton. 2007. Application of speaker classification in human machine dialog systems. In Christian Müller, editor, *Speaker Classification I: Fundamentals, Features, and Methods*, pages 174–179. Springer.

Felix Burkhardt. 2012. Fast labeling and transcription with the speechalyzer toolkit. *Proc. LREC (Language Resources Evaluation Conference), Istanbul*.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18:3280.

Antonio R. Damasio. 1994. *Descartes' error: emotion, reason, and the human brain.* Avon Books.

Florian Eyben, Martin Wllmer, and Bjrn Schuller. 2009. openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII)*.

C. M. Lee and S. S. Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*.

D. Neiberg, K. Elenius, and K. Laskowski. 2006. Emotion recognition in spontaneous speech using gmms. *Proc. ICSLP, Pittsburgh*.

R. Picard. 1997. *Affective computing*. MIT Press.

R. Picard. 2003. Affective Computing: Challenges. *Journal of Human-Computer Studies*, 59:55–64.

N. Sato and Y. Obuchi. 2007. Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, 2(2):835–848.

Bjorn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2010. The INTERSPEECH 2010 Paralinguistic Challenge. In *Proc. Interspeech*, Makuhari, Japan.

I. Shafran and M. Mohri. 2005. A comparison of classifiers for detecting emotion from speech. In *Proc. ICASSP, Philadelphia*.

I. Shafran and M. Riley und M. Mohri. 2003. Voice signatures. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

I. H. Witten and E. Frank. 2005a. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

Ian H. Witten and Eibe Frank. 2005b. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, second edition, June.