AUTOMATIC RECOGNITION OF EMOTION EVOKED BY GENERAL SOUND EVENTS

Björn Schuller[1], *Simone Hantke*[1], *Felix Weninger*[1], *Wenjing Han*[1,2], *Zixing Zhang*[1], *Shrikanth Narayanan*[3]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[3]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, USA
schuller@tum.de

ABSTRACT

Without a doubt there is emotion in sound. So far, however, research efforts have focused on emotion in speech and music despite many applications in emotion-sensitive sound retrieval. This paper is an attempt at automatic emotion recognition of general sounds. We selected sound clips from different areas of the daily human environment and model them using the increasingly popular dimensional approach in the emotional arousal and valence space. To establish a reliable ground truth, we compare mean and median of four annotators with their evaluator weighted estimator. We discuss human labelers' consistency, feature relevance, and automatic regression. Results reach correlation coefficients of .61 (arousal) and .49 (valence).

Index Terms— Affective Computing, Sound Event Classification, Sound Emotion Recognition, Feature Relevance, Sound Database

1. INTRODUCTION

When it comes to emotion recognition from the acoustic channel, research investigating speech [1, 2] and music [3] dominate. However, there are a rich variety of sounds besides speech and music in a real acoustic environment, which – without a doubt – also evoke various emotions in a human listener. For instance, the shrill sound of chalk scraping on a blackboard would be unacceptable to most people, while the gentle sound of waves clapping the sand beach would usually make one feel relaxed. Actually, listeners feed back emotion to any sound they are listening to in their daily life, no matter what kind of sound it is and whether the sound itself is subjective or objective. Sound perception is thus wrapped up with emotional response and affect: Infants first attempts to overcome anxiety are centered on sound making [4]. Likewise, it can also be of importance for future intelligent systems to have such a comprehensive sound emotion 'perception' ability. Another obvious field of application is sound design and dubbing of audio plays and movies where one may search for specific emotional sound events such as an angry doorslam vs. a spooky door creek, etc. Yet, relevant research in this direction is quite deficient until now – in fact first steps in this direction have been taken only recently such as annotating 120 clips of the BBC Sound Effects Library in three affective dimensions and retrieval experiments based on mean and standard deviation per second of 12 MFCC features [5]. In this paper, we set our focus on sound emotion recognition of realistic acoustic environment conditions.

The first crucial problem that arises is the lack of specialized sound databases for emotion research. A suitable sound database in this case should cover rich varieties of sounds – especially those commonly encountered in the real world like sounds of human beings, animals, vehicles, musical instruments, etc., and each sound in the database should be annotated with an accurate emotion label. Of course, there already exist some accessible sound databases [6], but they usually do not provide affective labeling. So the first aim of this work is to build up a sound database of annotated sound emotion. In existing work on emotion recognition from speech emphasis is usually put on subject's expressed emotion rather than listeners' emotions evoked by sound. This is more mixed for music emotion recognition. In fact, predicting emotion on the side of the listener is also very important in many cases where it can help identify human reaction ahead. In this paper, when we talk about 'sound emotions', we refer to the listeners' induced emotions.

Another typical problem in general emotion recognition is the selection of a suited emotion representation model. The commonly used representation models can be divided into so-called discrete and dimensional approaches. The former relies on a list of adjectives each describing an emotion tag such as happy, sad or depressed, and has been used in speech and music emotion recognition for long [1, 3]. However, such a discrete approach suffers from two main defects: being too ambiguous for a concise estimation of emotion and being too insufficient for real-life emotion representation if the number of categories stays in reasonable limits and no multiple assignments are allowed. By contrast, the latter describing emotions as points in a multi-dimensional (usually Cartesian) space often offers a more accurate way to represent emotions and becomes increasingly recognized lately [2, 3]. Thus, in this paper, Thayer's frequently encountered 2-D model [7] with VALENCE (i. e., how positive or negative the affect appraisal is) and AROUSAL (i. e., how high or low the physiological reaction is) as dimensions is adopted.

Owing to the novelty of emotion analysis in sound, we also provide an extended discussion on the reliability of our established annotation. Specifically the correlation coefficient and three different Kappa statistics between each labeler's annotation and established ground truth are given. Respecting the divergence between individual labelers, evaluator weighted estimator (EWE) [8] as ground truth is then proposed as – according to our experiments – its usage can improve the robustness of sound emotion recognition (here regression) results significantly.

In the remainder of this paper we introduce the Emotional Sound Database (Section 2), our experiments and results (Section 3) including the used features and regressor before concluding (Section 4).

This research has been supported by the German Research Foundation (DFG) through grant no. SCHU 2508/2 and the Chinese Research Council. The authors would further like to thank Jun Deng (TUM) and Masao Yamagiahi (Tokio Institute of Technology) for their contributions.

Table 1. Details on the Emotional Sound Database. Times in (minutes:)seconds.milliseconds. Human agreement: mean correlation coefficient (CC) and majority kappa values over the labelers.

Class	# Clips	Duration		AROUSAL				VALENCE			
		total	mean	CC	κ	κ^1	κ^2	CC	κ	κ^1	κ^2
All	390	24:53.55	3.50	.584	.386	.411	.436	.796	.490	.601	.699
Animals	90	6:06.53	4.05	.524	.350	.364	.378	.685	.448	.507	.569
Musical Instruments	75	3:41.17	2.57	.659	.392	.458	.529	.712	.435	.505	.592
Nature	30	2:43.65	5.29	.541	.355	.360	.356	.759	.430	.511	.575
Noisemaker	30	1:58.12	3.56	.569	.409	.406	.415	.869	.522	.650	.747
People	60	3:20.55	3.21	.629	.344	.386	.414	.823	.495	.622	.722
Sports	30	1:37.63	3.17	.550	.389	.390	.396	.607	.347	.363	.384
Tools	30	2:09.48	4.20	.621	.435	.454	.474	.738	.480	.543	.607
Vehicles	45	3:16.43	4.22	.473	.357	.322	.281	.688	.414	.459	.518

Table 2. Overview on the labelers' (ID A–D) agreement: correlation coefficient (CC) of the individual labelers with the mean, and Cohen's κ and weighted κ of the labelers with the majority vote for AROUSAL (Aro) and VALENCE (Val).

ID	CC		κ		κ^1		κ^2	
	Aro	**Val**	**Aro**	**Val**	**Aro**	**Val**	**Aro**	**Val**
A	.343	.769	.265	.442	.186	.544	.099	.635
B	.701	.869	.445	.590	.505	.702	.566	.794
C	.542	.744	.399	.477	.435	.582	.474	.683
D	.749	.800	.435	.454	.519	.575	.604	.684

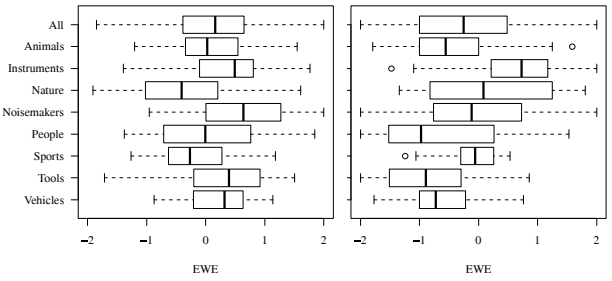

Fig. 1. Boxplots of evaluator weighted estimator (EWE) distribution per sound category: AROUSAL (left) and VALENCE (right).

2. EMOTIONAL SOUND DATABASE

To build our 'Emotional Sound Database'[1] and evaluate our system, we selected the on-line freely available engine FindSounds.com[2] [9]. This huge database hosts sound files manually sorted into 16 main categories and 365 sub-categories. For our experiment, we chose 390 sound files out of more than 250 000 (10 000 different sound clips of which each can be downloaded in 25 different speeds). We decided to use the following eight categories taken from FindSounds.com: *Animals*, *Musical instruments*, *Nature*, *Noisemaker*, *People*, *Sports*, *Tools* and *Vehicles*. With this choice the database represents a broad variety of frequently occurring sounds in everyday environment. The stereo sound files were MPEG-1 Audio Layer III (MP3) or Audio Interchange File Format (AIFF) encoded at differing sample rates and bit rates of at least 128 kBit/s. To work with these files, they were converted to the free lossless audio codec (Flac) and .wav container audio format by changing the sample rate to 44.1 kHz (Flac) and 16 kHz (Wave) and keeping the original bit rate. More details on the used database are given in Table 1. As can be seen, the corpus size is well in line with first datasets for emotion recognition in speech (such as the Berlin or Danish emotional speech databases) or music (such as the first MIREX mood classification task set).

Our Emotional Sound Database was annotated by four labelers (by ID: A: male, 25 years; B: female, 28 years; C: male, 27 years, plays guitar; D: male, 26 years, plays Chinese DiZi flute). They were all post graduate students working in the field of audio processing. All labelers are of Southeast-Asian origin (Chinese and Japanese) in order not to introduce strong cross-cultural effects – such questions need to be left for further investigations. For the annotation these four listeners were asked to make a decision according to the two dimensions in the emotion plane assigning values on a five-point scale in $\{-2, -1, 0, 1, 2\}$ for AROUSAL and VALENCE. They were instructed to annotate the perceived emotion and could repeatedly listen to the sounds that were presented in random order across categories. Annotation was carried out individually and independently by each of the labelers. The annotation procedure is described in detail in [10] and the tool can be downloaded as *Foobar2000* plugin[3].

Due to the novelty of the regression task defined for this study, it has to be investigated whether it is well-defined, or, how to deduce a 'gold standard' ground truth from the individual human labels that is to be used as target for learning algorithms. Taking into account the ordinal scale nature of the dimensional emotion ratings, we calculate weighted Kappa (κ) statistics; in order to provide a rough comparison to the performance metric for the automatic regression, we consider correlation coefficients (CC) as well. Weighted κ (κ^1, κ^2) statistics use the absolute value of disagreement or its square, respectively, to quantify the amount of disagreement on an ordinal scale; for reference, we also provide unweighted (Cohen's) κ. Inter-labeler agreement in terms of CC is determined by first computing the mean rating for each instance, then calculating the CC of each labeler with the mean. Inter-labeler agreement in terms of κ is calculated for each labeler with the majority vote of the labelers.

Results of agreement analysis are shown in Table 2. Interestingly, the agreement is much higher ($\kappa^2 = .699$) for VALENCE than for AROUSAL ($\kappa^2 = .436$). Furthermore, a more detailed analysis by sound category reveals that the human agreement – particularly, on VALENCE – is strongly dependent on the sound cat-

[1] Our labels and partitions for exact recreation are available at http://www.openaudio.eu.

[2] http://www.findsounds.com – accessed 25 July 2011.

[3] http://www.openaudio.eu/wsh_mood_annotation.zip

Table 3. Set of 31 low-level descriptors and 42 functionals. [1]Not applied to delta coefficient contours. [2]For delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied. [3]Not applied to voicing related LLD.

Energy & spectral low-level descriptors (25)
loudness (auditory model based), zero crossing rate, energy in bands from 250 – 650 Hz, 1 kHz – 4 kHz, 25 %, 50 %, 75 %, and 90 % spectral roll-off points, spectral flux, entropy, variance, skewness, kurtosis, psychoacoustic sharpness, harmonicity, MFCC 1–10
Voicing related low-level descriptors (6)
F_0 (sub-harmonic summation (SHS) followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: 'jitter of jitter'), logarithmic Harmonics-to-Noise Ratio (logHNR)
Statistical functionals (23)
(positive[2]) arithmetic mean, root quadratic mean, standard deviation, flatness, skewness, kurtosis, quartiles, inter-quartile ranges, 1 %, 99 % percentile, percentile range 1 %–99 %, percentage of frames contour is above: minimum + 25%, 50%, and 90 % of the range, percentage of frames contour is rising, maximum, mean, minimum segment length[3], standard deviation of segment length[3]
Regression functionals[1] (4)
linear regression slope, and corresponding approximation error (linear), quadratic regression coefficient a, and approximation error (linear)
Local minima/maxima related functionals[1] (9)
mean and standard deviation of rising and falling slopes (minimum to maximum), mean and standard deviation of inter maxima distances, amplitude mean of maxima, amplitude mean of minima, amplitude range of maxima
Other[1,3] (6)
Linear Prediction (LP) gain, LP Coefficients 1 – 5

egory. For instance, VALENCE of noisemakers are highly agreed upon ($\kappa^2 = .747$) while sounds from sports are not ($\kappa^2 = .384$); for AROUSAL, strongest agreement is found for musical instruments ($\kappa^2 = .529$), and vehicles ($\kappa^2 = .281$) are observed on the other end of the scale. Self agreement in a complete repetition (in shuffled order) of the labeler's original annotation after one full week of pause (only used for self conformity analysis) was highest for labeler B ($\kappa^2 = .554$ for AROUSAL, $\kappa^2 = .772$ for VALENCE) who also found highest agreement with the ground truth (cf. table 2). Considering the 'reliability' of individual labelers, i. e., their agreement with the 'consensus', we observe striking differences especially for AROUSAL: Here, CC ranges from .343 (labeler A) to .749 (labeler D), which is also reflected in the κ statistics ($\kappa^2 = .099$ for labeler A, $\kappa^2 = .604$ for labeler D). For VALENCE, differences are visible but less strong, with labeler B showing the strongest agreement with the 'consensus'. These observations motivate the use of the EWE as a robust estimate of the desired labeler-independent emotion rating in addition to the arithmetic mean. For each instance n, the EWE r_n^d is defined as:

$$r_n^d = \frac{1}{\sum_{k=1}^{K} \mathrm{CC}_k} \sum_{k=1}^{K} \mathrm{CC}_k r_{n,k}^d. \quad (1)$$

where $K = 4$ is the number of labelers, d is the dimension (AROUSAL or VALENCE) and $r_{n,k}^d$ is the rating of instance n by labeler k in dimension d. Informally, the EWE is a weighted mean rating, with CCs of the labelers as weights. The distribution of the EWE for each sound category is shown in Figure 1 as a box-and-whisker plot. Boxes range from the first to the third quartile and all values that exceed that range by more than 1.5 times the width of the box are considered outliers, depicted by circles. In our following experiments with automatic emotion regression, we will evaluate whether usage of the EWE instead of the arithmetic mean (or median) can improve robustness of the results.

Table 4. Automatic regression results by correlation coefficient (CC) with different types of ground truth: evaluator weighted estimator (EWE), median, and mean in three and ten-fold stratified cross-validation. Number of trees varied for the regressor.

CC		3 folds			10 folds		
# tree		100	200	500	100	200	500
ARO	**EWE**	.605	.603	.607	.611	.608	.606
	median	.547	.554	.557	.553	.555	.548
	mean	.569	.569	.571	.558	.563	.559
VAL	**EWE**	.436	.446	.441	.458	.469	.473
	median	.399	.432	.436	.446	.449	.454
	mean	.429	.443	.428	.467	.484	.485

3. EXPERIMENTS AND RESULTS

The audio feature set used is our openSMILE toolkit's AVEC set with 1 941 features brute forced by functional application to low-level descriptors (LLD). Details for the LLD and functionals are given in Table 3. The set of LLD covers a standard range of commonly used features in audio signal analysis and emotion recognition. We avoid LLD/functional combinations that produce values which are constant, contain (very) little information, and/or high amount of noise. Features are computed per whole sound clips. For recognition, we consider a generalization of the random forest principle: We use Random Subspace meta-learning, which can provide very good generalization properties, in combination with REPTree – a fast decision tree learner. Stemming from our experience, we decided for no pruning of the trees, a subspace size of 0.05 (i. e., 97 features are randomly chosen out of the 1 941 per tree), and experiment with three different numbers of trees in $\{100, 200, 500\}$ to grow a forest. As our labeling and the feature extractor and its configuration are available for reproduction, we decided to further research reproducibility by choosing the implementations in the broadly used free and open Weka toolkit. The experiments were performed in stratified three-fold and ten-fold cross-validation[4]. Table 4 shows the correlation coefficients for AROUSAL and VALENCE employing EWE, median, and mean as methods to merge the evaluation results of the four evaluators. We can see that the regression of sound emotion performs well with CC of around .61 (AROUSAL) and up to .49 (VALENCE) when evaluating on EWE. The tendency that AROUSAL is better assessed by acoustics is well in line with experience from speech and music emotion analysis [1, 10]. It can also be seen clearly that the performance evaluated on EWE mostly (except for VALENCE in the case of ten folds) outperforms the other two methods, mean and median. Median on the other end always performs worst for its instability when evaluators show huge disagreement. Besides that, as expected, CCs by ten-fold cross validation are

[4]Partitioning by default random seed in Weka.

Table 5. Automatic regression results by correlation coefficient (CC) per sound category for one example (EWE, 10 folds, 500 trees).

Class	CC	
	AROUSAL	VALENCE
Animals	.643	.448
Musical Instruments	.516	.217
Nature	.688	.589
Noisemaker	.579	.778
People	.604	.048
Sports	.682	.198
Tools	.590	-.057
Vehicles	.579	.279

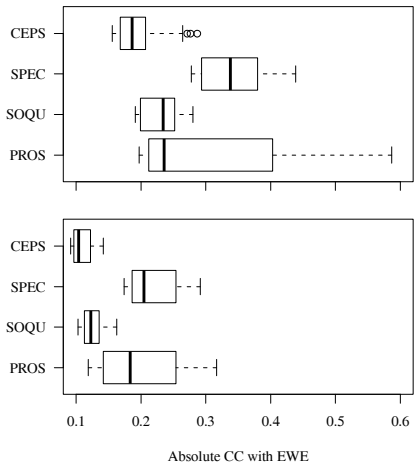

Fig. 2. Boxplots of the 30 highest absolute correlation coefficients (CC) with the evaluator weighted estimator (EWE) per feature group: cepstral (CEPS), spectral (SPEC), sound quality (SOQU: voicing probability, logHNR, jitter, and shimmer), and prosodic features (PROS: loudness, zero-crossing rate, and F0). AROUSAL (top) and VALENCE (bottom).

mostly slightly higher as compared to three-fold cross-validating due to more data used for training – this holds especially for VALENCE. In Table 5 we look at the CC and its relation to sound category for one exemplary configuration. As can be seen, AROUSAL prediction is somewhat balanced across sound categories. However, as for VALENCE, especially *Noisemakers* and *Nature* can be identified well above others. Yet, comparing this with Figure 1, it seems safe to argue that the regressor does not merely implicitly recognize the sound category, since VALENCE for *Noisemakers* is quite widespread despite the fact that there are considerable differences in the mean VALENCE, as one can expect ('nature is more pleasant than vehicles').

Finally, we investigate the 30 best features ranked by CC with the EWE as ground truth per each of four groups: cepstral, spectral, 'sound quality' (in analogy to voice quality), and prosody (longer term signal properties). The result is shown as boxplots per dimension in Figure 2. There we can see that independent of the dimension, spectral features perform best on average, but the best individual feature is of prosodic nature, in each case. The following can be observed from the full list of top 30 features: AROUSAL is highly correlated with loudness. In fact, loudness features have almost as strong a correlation with the EWE as the regressor prediction. Highest CC is observed for the root quadratic mean of loudness (.587). Since the correlation is measured with the EWE, this seems to be the consensus; however, it should be pointed out that the first labeler strongly disagreed with the others on AROUSAL. Next, VALENCE is correlated with loudness as well, but not as strongly, and the correlation is negative: Loud sounds are unpleasant. Highest absolute CC with VALENCE EWE is observed for the third quartile of loudness (-.316). VALENCE is also negatively correlated with spectral flux, i. e., large spectral variations are perceived as unpleasant. The CC of the inter quartile range 1–2 of spectral flux is -.292. VALENCE is further negatively correlated with spectral harmonicity: The phenomenon here may be that quasi-sinusoidal sounds are unpleasant. The CC of 50 % up-level time of harmonicity is -.241.

4. CONCLUSION

We investigated the automatic recognition of emotion evoked by general sound events. We observed good agreement of independent labelers in this respect and were able to demonstrate feasibility of automatic assessment of emotion in sound in two dimensions. In fact, results were found in the rough range of typical dimensional speech and music emotion recognition when operating in high realism [1, 2, 3, 10]. And indeed, the sound events considered here were completely independent and often of lower acoustic quality. We further found spectral features to be most important as a group after individual prosodic features for this task. Future efforts need to be invested into creation of larger sound emotion resources, e. g., by shared community efforts or Amazon Mechanical Turk or similar. Naturally, deeper analysis of feature relevance including individual analysis per sound category is another interesting research question. Finally, multi-task learning of the sound category and the evoked emotion seems a promising avenue to improve both tasks in a synergistic way.

5. REFERENCES

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9/10, pp. 1062–1087, 2011.

[2] H. Gunes, B Schuller, M Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. EmoSPACE*, Santa Barbara, CA, 2011, IEEE, pp. 827–834.

[3] Y.E. Kim, E.M. Schmidt, R. Migneco, B.G. Morton, P. Richardson, J. Scott, J.A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. ISMIR*, Utrecht, The Netherlands, 2010, pp. 255–266.

[4] M.A. Forrester, "Auditory perception and sound as event: theorising sound imagery in psychology," *Sound Journal*, 2000.

[5] S. Sundaram and R. Schleicher, "Towards evaluation of example-based audio retrieval system using affective dimensions," in *Proc. ICME*, Singapore, Singapore, 2010, pp. 573–577.

[6] B. Gygi and V. Shafiro, "Development of the database for environmental sound research and application (DESRA): Design, functionality, and retrieval considerations," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, Article ID: 654914, 12 pages, 2010.

[7] R. E. Thayer, *The Biopsychology of Mood and Arousal*, Oxford University Press, Boston, Mass, USA, 1990.

[8] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[9] S. V. Rice and S. M. Bailey, "A web search engine for sound effects," in *Proc. 119th AES*, New York, 2005.

[10] B. Schuller, J. Dorfner, and G. Rigoll, "Determination of non-prototypical valence and arousal in popular music: Features and performances," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, Article ID: 735854, 19 pages, 2010.