Affective Classification of Generic Audio Clips using Regression Models

Nikolaos Malandrakis[1], *Shiva Sundaram*[2], *Alexandros Potamianos*[3]

[1]Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA 90089, USA
[2]Audyssey Laboratories, Los Angeles, CA 90071, USA
[3]Dept. of ECE, Technical Univ. of Crete, 73100 Chania, Greece

malandra@usc.edu, Shiva.Sundaram@IEEE.org, potam@telecom.tuc.gr

Abstract

We investigate acoustic modeling, feature extraction and feature selection for the problem of affective content recognition of generic, non-speech, non-music sounds. We annotate and analyze a database of generic sounds containing a subset of the BBC sound effects library. We use regression models, long-term features and wrapper-based feature selection to model affect in the continuous 3-D (arousal, valence, dominance) emotional space. The frame-level features for modeling are extracted from each audio clip and combined with functionals to estimate long term temporal patterns over the duration of the clip. Experimental results show that the regression models provide similar categorical performance as the more popular Gaussian Mixture Models. They are also capable of predicting accurate affective ratings on continuous scales, achieving 62-67% 3-class accuracy and 0.69-0.75 correlation with human ratings, higher than comparable numbers in literature.

Index Terms: emotion recognition, audio content processing, affective modeling, regression models

1. Introduction

Generic unstructured sound clips are pervasive in multimedia and contribute significantly to the sensory, semantic and affective interpretation of content. Recently, generic audio has received significant research interest, especially for the task of classification to semantic categories [1] and the associated task of audio event detection [2]. Such sound clips can also have significant affective content [3], which can be important for the affective interpretation of audio streams (especially authored multimedia content such as movies and video clips). Ambiances and sound effects can be used by a film director to convey the desired emotions. Using source separation and handling the resulting audio streams individually has also been proposed [4]. Generic sounds provide context that helps better understand the scene. In that regard, knowing and measuring the affective ratings provides valuable information to autonomous robots and content retrieval systems. Despite this potential importance, affective content analysis and modeling of generic audio is a little-researched problem mainly due to the diversity of the content and the lack of comprehensive annotated databases.

Among the main hurdles in the analysis and modeling of generic audio are its inherent diversity both in terms of generation source (nature, city, human, animal, machine etc.) and acoustic characterization (noise, chirps, cries, harmonic etc.), as well as, its lack of structure (unlike music). As a result, a large database is needed to adequately characterize such diverse content. The only affectively annotated generic audio corpus available is IADS [5], however its' limited size (167 clips) fails to capture the richness of generic audio and make it hard to apply machine learning methods. In this paper, we present the affective annotation and analysis of a comprehensive collection of 1472 clips from the BBC sound effects library [6] that can serve as a stepping stone for future research in the field.

Modeling of generic audio for semantic classification and audio event detection usually employs generic features and models, such as Mel Frequency Cepstral Coefficients (MFCCs) and Gaussian Mixture Models (GMMs). There is virtually no research in the area of affective classification of generic audio apart from the exploratory work in [3], which did not focus on classification, so we turn to the affective literature for feature extraction and modeling of speech and music. Speech emotion is the most researched area in the audio emotion field, and a wide variety of features, methods and datasets have been proposed [7]. However, most of the systems participating in INTERSPEECH 2009 emotion challenge [8] seem to prefer generic features and modeling methods. Although, GMMs and MFCCs are also popular for music tagging and affect recognition [9], alternative features and models have proved more successful in music processing. Statistics of the short-time spectrum, chromas, Gaussian super vectors, music key-related features, and spectral novelty features have been successfully combined with MFCCs for music processing tasks as outlined in the MIREX challenges, e.g., [10]. Also regression models have recently emerged as popular alternatives to GMMs for music modeling [11, 12].

In this paper, we investigate a large set of (mostly) frame-based features from the speech and music processing literature and combine them via functionals to model their time dynamics. We use regression models, as well as, GMMs for the problem of affective classification of generic audio. Feature selection algorithms are used to identify a subset of good performing features. Features and models are evaluated on the BBC Affective Database in terms of classification accuracy and correlation with human ratings for each of the arousal, valence and dominance dimensions.

There is virtually no prior work on generic audio affect apart from [3]. Unlike that, this paper focuses on the classification task itself. The results achieved are very encouraging and an improvement over the limited prior work. We also believe that the audio database annotation, containing almost 1500 unstructured sound clips from a variety of sources is a significant contribution. Its' size and content variance should enable the use of machine learning methods to the task.

2. Dataset Annotation and Analysis

In order to apply supervised machine learning methods, we manually annotate a generic audio database [6] in accordance with the 3-D affective model of arousal, valence and dominance. This affective model has been shown to offer sufficient descriptive power in a similar context [5] and has been very popular in affective research in recent years.

2.1. BBC Affective Database

The dataset contains 1472 audio clips from the BBC sound effects library. The clips contain generic, non-music, non-speech sounds, including sound effects and ambiances, such as "baby crying", "beach ambiance" and "factory machinery". Reflecting the wide variance in content is the distribution of clip lengths, shown in Fig 3; clips containing sound effects are very short, whereas clips containing ambiance sounds can last for minutes. The clips were annotated by 383 annotators between June 2009 and July 2011. The annotators rated their own genuine emotional response to each clip's emotional content in terms of arousal, valence and dominance in a range of 0 to 8 using self-assessment manikins [5]. The annotators were not informed of the content of the clips, so they did not know what produced the specific sound. Each annotator rated (on average) 124 clips, chosen randomly. Overall, an average of 32 ratings are available for each clip. The listening experiments were performed in an acoustically treated environment using headphones connected to a computer. The audio clips were presented using an automatic web-based software interface running on the same computer. To derive a ground truth from these individual annotations we use the weighting/rejection method proposed in [13], where the final ratings are weighted combinations of the individual users' ratings and the weights are proportional to the Pearson correlation between ratings.

Table 1: Agreement metrics for each dimension

Inter-annotator agreement			
Metric	Arous.	Valen.	Domn.
avg. pairwise correlation	0.52	0.55	0.16
avg. pairwise mean abs. dist.	2.02	1.84	2.32
Krippendorff's alpha (ordinal)	0.39	0.47	0.11
Krippendorff's alpha (interval)	0.39	0.46	0.10
Agreement with the ground truth			
Metric	Arous.	Valen.	Domn.
avg. correlation	0.55	0.60	0.41
avg. mean abs. dist.	1.42	1.18	1.36

2.2. Annotation Results

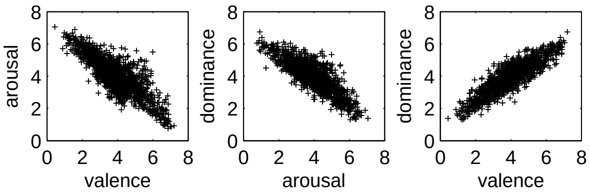

Figure 1: Scatter plots of clip affective ratings.

In addition to the 1472 clips, an extra set of 5 clips were annotated by all users and used to calculate agreement statistics, pairwise correlation, pairwise distance and Krippendorff's alpha. which are presented in Table 1. The agreement ratings for arousal (A) and valence (V) are as expected, perhaps even high given that each user rates his or her own subjective emotional experience when listening to a clip. The higher agreement for V is consistent with [3]. The users however were notably less in agreement with regards to dominance (D). Also shown in Table 1 are the ratings of average user agreement to the ground truth, which are as expected much higher. In Fig. 1 the two dimensional scatter plot of the derived ground truth are shown. The shape of the valence-arousal plot in particular does not match the "V" shape shown in [5] for a similar data set: in our case the positive valence - high arousal quadrant is relatively empty, indicating that very few clips were considered "uplifting", though that may be a result of the random clip selection process, whereas IADS was created so as to elicit specific reactions from the listeners. The three dimensions are weakly correlated when taking into acount each user's ratings (V-A: -0.45, V-D: 0.47, A-D: -0.38), however the three ground truth dimensions are strongly correlated (V-A: -0.82, V-D: 0.88, A-D: -0.84). The high correlation has also been noted in the IADS dataset (V-A: -0.44, V-D: 0.94, A-D: -0.54). Examining the results reveals no particular issues, with samples having affective ratings close to the expected. Some samples with extreme affective values are: burglar alarm ([A,V,D] $= [6.9, 1.6, .19]$), ambulance siren ([A,V,D] $= [6.7, 1.4, 1.7]$), birds and insects ([A,V,D] $= [0.8, 6.8, 6.1]$), blackbird ([A,V,D] $= [0.9, 7.2, 6.7]$), Wembley stadium crowd ([A,V,D] $= [5.6, 5.1, 3.8]$).

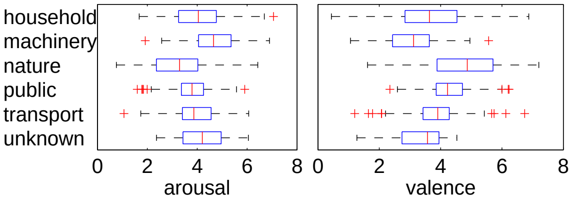

Figure 2: Affective rating distributions per semantic category.

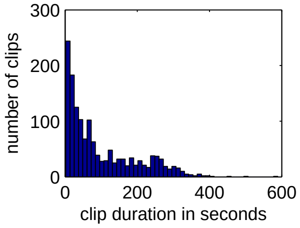

Figure 3: Histogram of audio clip lengths.

Apart from affective ratings, the dataset contains semantic labels (contained in the BBC sound effects library) and onomatopoeia labels produced as described in [1] for most clips, allowing the hierarchical analysis of sounds. A sample of the distributions of affective ratings per semantic category is shown in Fig. 2. The distributions show some expected trends: annotators found nature sounds (containing animal sounds and nature ambiances) particularly positive, whereas machinery sounds were rated as particularly negative, perhaps annoying.

3. Modeling and Feature Extraction

Motivated by recent research in affective modeling for music [11] and text [14] we use regression models for affective classification of generic audio. Specifically, we investigate the use of Multiple Linear Regression (MLR) and Multiple Quadratic Regression without the interaction terms (MQR). Regression models consider the output as the result of a parametric function, with the features taking the role of variables.

Although the valence, arousal and dominance ratings in our database take continuous values, it is not uncommon to use two or three classes (e.g., positive-neutral-negative) to describe each of the three dimensions, since that level of detail is enough for a lot of applications. In order to use the ground truth ratings for a categorical classification task, they were quantized into equiprobable bins using the cumulative distribution function estimated via Parzen windows. Thus we are faced with a 3-class classification problem, where each audio clip has to be categorized in one of the three discrete classes (independently) for each dimension. The results obtained using regression models are continuous values, which can then be quantized to three levels for our task.

Gaussian Mixture Models (GMMs) are also used as baseline classifiers. GMMs are probabilistic models where each category is described by the observation distributions of the features. Since clips contain multiple feature frames, the posterior probabilities estimated in each frame are combined to produce a clip-level score. The clip-level posterior probability is computed as the product of all frame-level posterior probabilities.

3.1. Feature extraction

We take a generic approach to feature extraction: essentially we extract all features that could prove useful, followed by feature selection to identify the best performers. Because the dataset is composed of generic sounds, rather than speech or music, we exclude features that are specific to these audio types, e.g., pitch-related features used for speech. A variety of frame-level descriptors are extracted in the time, frequency or cepstral domains. These features have been used in both speech and music processing and consist of Mel-frequency cepstrum coefficients (MFCCs), chroma coefficients, (log) Mel filter-bank power (log power values of a Mel-scaled bank of 26 filters), energy (RMS ang log), loudness, intensity, spectral rolloff (25%, 50%, 75%), spectral flux, spectral entropy, rhythm irregularity, rhythm fluctuation, spectral brightness, spectral roughness and spectral novelty. All frame-level descriptors were extracted using existing toolkits, namely, the OpenSMILE [15] and MIR toolbox [16], using a hop size of 10ms and a frame size dependent on the feature: 25ms for low-level features like energy, up to a second for music inspired features like rhythm fluctuation. In addition to the base descriptors we also use their first derivatives (deltas) computed over four frames.

Frame-level features are combined into long-term descriptors using a set of 51 functionals to the frame level descriptors, including simple statistics like arithmetic, quadratic and geometric mean, standard deviation, variance, skewness and kurtosis, extrema, ranges, quartiles, inter-quartile ranges, linear and quadratic regression coefficients (where linear coefficient 1 is the slope) and regression errors (metrics of how much the frame-level descriptors deviate from the ideal estimated form), curvature statistics (% of time with left of right curvature) and histogram descriptors (% of samples in 4 equally spaced bins). All functionals are applied for the length of a clip, so a single value is extracted per clip for each frame-level feature. Extraction of all functionals was done using the OpenSMILE toolkit. Overall the feature pool contains 7140 long-term features (the cartesian product of functionals and frame-level features).

3.2. Feature Selection and Experimental Procedure

Due to the large number of resulting features, it is imperative to use a feature selection algorithm to choose the top performers. To do so we use wrappers [17], that is we use the performance of the models themselves while running cross-validation experiments to evaluate each candidate feature set. Due to the large number of available features and the limited dataset size running a backwards selection strategy is not possible (in some cases we have more features than training samples). The strategy we use is one of *best-first forward selection*: starting from an empty feature set we iteratively add more features without deletions, e.g., when choosing the second feature we do not evaluate all pairs but only those that include the best performing feature selected during the first iteration.

The feature selection criterion used for the GMM model is 3-class accuracy, while for the regression models Pearson correlation (with human ratings) is used[1]. For both GMMs and regression models, features are selected and performance is evaluated by conducting 10-fold cross-validation experiments. Specifically, using wrappers we select the first one hundred best performing features for each model and affective dimension. The output of the MLR model is a continuous value for each sample; to convert to discrete category labels we use the same quantization boundaries used to convert the continuous ground truth to discrete values. This makes the results of GMMs and regression models directly comparable.

4. Experimental Results

Next, we report affective classification results for the arousal, valence and dominance dimensions. Performance is reported in terms of classification accuracy and Pearson correlation (pooled) between the estimated and hand-labeled ratings. Results from three experiments are reported in order to demonstrate: (i) the performance of the short-term (frame-level) vs long-term (functionals) features (Table 2), (ii) the relative performance of the regression and GMM models in terms of classification accuracy shown in Fig. 4, and (iii) the performance of the regression models in terms of correlation with human ratings (Table 3).

Table 2: GMM classification accuracy for LLDs and functionals

Scope	Low Level. Descr.	Arous.	Valen.	Domn.
frame level	chroma + Δ	0.41	0.45	**0.43**
	log Mel power + Δ	0.44	0.48	0.44
	MFCC + Δ	0.45	0.44	0.43
long term	chroma + Δ	0.41	**0.46**	0.42
	log Mel power + Δ	**0.46**	**0.49**	**0.46**
	MFCC + Δ	**0.48**	**0.48**	**0.45**

In Table 2, we compare the classification accuracy (3-class) for the frame-level vs the long-term features. Results are reported for the following frame-level descriptors: chroma, log Mel power and MFCCs (along with their first times derivative). For the computation of long-term features, we select *a single functional*, the best performing one for each dimension, applied to all frame-level descriptors (LLDs). Therefore the same number of features is used for both frame-level and long-term features. The results in Table 2 show a clear benefit when moving from frame-level to long-term features in almost all cases. This trend can be attributed to the better representation of audio dynamics when functionals are used (rather than simply multiplying frame-level posteriors). It should be noted that only a single

[1] Note that using classification accuracy (instead of correlation) as a feature selection criterion gives a slight advantage to GMMs.

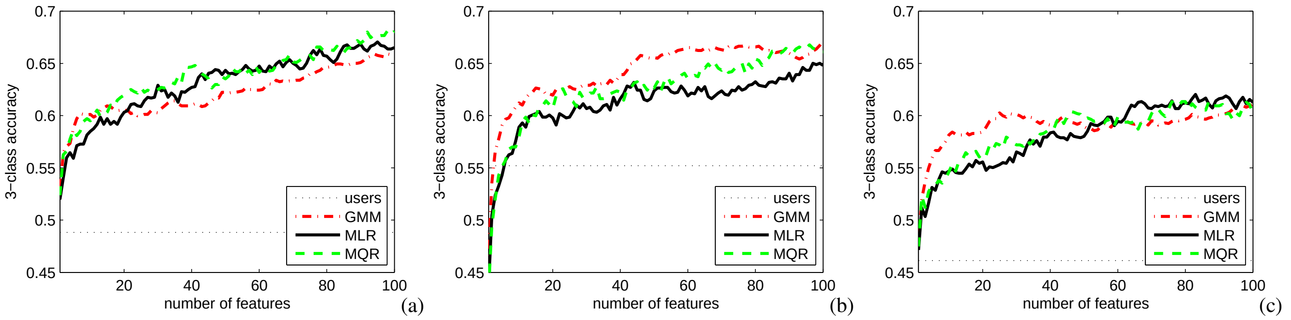

Figure 4: 3-class accuracy achieved by GMM (red dashed-dotted line) and regression models (black solid, green dashed) as a function of the number of features for (a) arousal, (b) valence and (c) dominance. Human annotator performance is shown as "users" (dotted).

functional is used for all 26 frame-level descriptors: by adding more functionals or using different functionals for each LLD performance improves further. Similar results have been obtained for regression models (not reported due to lack of space).

In Fig. 4 (a),(b),(c), we show 3-class classification accuracy for each dimension as a function of the number of (long-term) features used for the GMM and regression (MLR, MQR) models. We also report the average performance of human annotators on the same task[2] shown as the dotted "users" line in the figure. Results are reported using 10-fold cross-validation and feature selection as described in Section 3.2. The following are the main conclusions from these experiments: (i) Human "classifier" performance can be beaten by both GMM and regression models using only a handful of selected features. (ii) Both humans and machines have a harder time estimating dominance scores than arousal and valence. (iii) Performance improves significantly with the number of features and levels off around 67% for arousal and valence, and 62% for dominance. (iv) Looking at the relative performance of the models, we see that GMMs perform best when it comes to predicting valence and when predicting dominance using a small number of features, while regression models perform better at predicting arousal and predicting dominance with a large number of features[3]. (v) Regression models seem to scale better with increased number of features, i.e., the performance of regression models improves faster with increasing number of features. The improved scaling capability of regression models is probably due to the relatively small number of parameters: MLR and MQR models have only one parameter per feature space dimension. (vi) There is a small difference in terms of performance between the linear and quadratic regression models in terms of performance, with the MQR performing somewhat better. Overall, both GMM and regression models perform very well for the problem of 3-class emotion classification surpassing the performance of an average human annotator, reaching accuracies up to 67%. These results are very encouraging given that our dataset contains very diverse audio content that is hard even for human listeners to characterize.

For some application continuous affective ratings are needed [18]. Regression models have the advantage of producing such continuous ratings. In Table 3, we report Pearson correlation between the ratings produced by the MLR model and the ground truth as a function of the number of features. Results were obtained via a double loop 10-fold cross-validation, with the internal loop used for feature selection and the external loop used for evaluation. Correlation performance for a typical human labeler is reported as "Users". As is the case with classification accuracy: (i) the regression model easily beats human performance, (ii) correlation improves with increased number of features and (iii) dominance is harder to predict than arousal and valence. In terms of absolute numbers, high correlation of [0.75, 0.73, 0.69] is achieved for the 3 dimensions.

Table 3: Pearson correlation performance for the MLR model

Model	# of features	Arous.	Valen.	Domn.
Users	-	0.55	0.60	0.41
MLR Regression Model	10	0.70	0.67	0.63
	20	0.72	0.70	0.65
	30	0.74	0.71	0.67
	40	0.75	0.72	0.68
	50	**0.75**	**0.73**	**0.69**

5. Conclusions

We have shown that regression models and long-term features (estimated using functionals over frame-level features) perform well for estimating continuous affective ratings of generic audio. In addition, feature selection over a family of frame-level features and functionals significantly improves results reaching 62-67% 3-class accuracy and 0.69-0.75 correlation, which are significantly higher than those reported in literature. These are very encouraging results given the increased difficulty compared to music and speech. In the future, we will investigate in more detail how long-term features can better capture the dynamics of audio clips, analyze the output of the feature selection process as a function of audio clip type and length, as well as, improve the modeling and feature extraction process. The annotated database of generic audio will be published for the scientific community in the near future.

6. Acknowledgments

This work was performed while Nikolaos Malandrakis and Shiva Sundaram were at Deutsche Telekom Laboratories, Berlin.

The authors would like to thank Dr. Robert Schleicher of the Quality and Usability Lab, TU Berlin for his input during the early stages of this work.

[2] We assume that the annotation performed by each user is a classification result and compare it to the ground truth. This human annotator classification accuracy is then averaged over all users.

[3] One should keep in mind that feature selection is better tuned to GMMs (where classification accuracy is the selection criterion).

7. References

[1] S. Sundaram and S. Narayanan, “Classification of sound clips by two schemes: Using onomatopoeia and semantic labels,” in *Proc. ICME*, 2008, pp. 1341–1344.

[2] M. Xu, L.T. Chia, and J. Jin, “Affective content analysis in comedy and horror videos by audio emotional event detection,” in *Proc. ICME*, 2005, pp. 622–625.

[3] B. Schuller, S. Hantke, F. Weninger, Wenjing Han, Zixing Zhang, and S. Narayanan, “Automatic recognition of emotion evoked by general sound events,” in *Proc. ICASSP*, 2012, pp. 341–344.

[4] H. L. Wang and L. F. Cheong, “Affective understanding in film,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 689–704, June 2006.

[5] M. M. Bradley and P. J. Lang, “International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings,” Tech. Rep. B-2, The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, 1999.

[6] “BBC sound effects library,” http://www.sound-ideas.com/bbc.html.

[7] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[8] 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.

[9] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, “Semantic annotation and retrieval of music and sound effects,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 467–476, feb 2008.

[10] J.-C. Wang, H.-Y. Lo, S.-K. Jeng, and H.-M. Wang, “MIREX 2010: Audio classification using semantic transformation and classifier ensemble,” in *MIREX*, 2010.

[11] 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*, 2010.

[12] E. M. Schmidt, D. Turnbull, and Y. E. Kim, “Feature selection for content-based, time-varying musical emotion regression,” in *Proc. ISMIR*, 2010, pp. 267–274.

[13] M. Grimm and K. Kroschel, “Evaluation of natural emotions using self assessment manikins,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 381–385.

[14] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, “Kernel models for affective lexicon creation,” in *Proc. Interspeech*, 2011, pp. 2977–2980.

[15] F. Eyben, M. Wollmer, and B. Schuller, “Openear – introducing the munich open-source emotion and affect recognition toolkit,” in *Proc. ACII*, 2009, pp. 1–6.

[16] O. Lartillot, P. Toiviainen, and T. Eerola, “A matlab toolbox for music information retrieval,” in *Data Analysis, Machine Learning and Applications*, pp. 261–268. Springer Berlin Heidelberg, 2008.

[17] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[18] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, “A supervised approach to movie emotion tracking,” in *Proc. ICASSP*, 2011, pp. 2376–2379.