# Towards Mapping Timbre to Emotional Affect

Niklas Klügel, Georg Groh
Technische Universität München
Boltzmannstrasse 3
Garching, Germany
{kluegel,grohg}@in.tum.de

## ABSTRACT

Controlling the timbre generated by an audio synthesizer in a goal-oriented way requires a profound understanding of the synthesizer's manifold structural parameters. Especially shaping timbre expressively to communicate emotional affect requires expertise. Therefore, novices in particular may not be able to adequately control timbre in view of articulating the wealth of affects musically. In this context, the focus of this paper is the development of a model that can represent a relationship between timbre and an expected emotional affect[1]. The results of the evaluation of the presented model are encouraging and thus support its use in steering or augmenting the control of the audio synthesis. We explicitly envision this paper as a contribution to the field of Synthesis by Analysis in the broader sense, albeit being potentially suitable to other related domains.

## Keywords

Emotional affect,Timbre, Machine Learning, Deep Belief Networks, Analysis by Synthesis

## 1. INTRODUCTION

In many ways, timbral qualities constitute a central building block of music. From the perspective of a composer, timbre is integral to making voice leading effective [10]. From the perspective of a performer the variation of timbre is "[...] one of the principal ways through which performers communicate musical structure, ideas, emotions and musical personality" [9]. From the perspective of a listener it induces expectations towards the flow of a piece [20]. In this regard timbre communicates musical intend, structural concepts and interpretative imagination. The component of timbre we focus on is its (emotional) affect. As in the contribution by Eerola *et al.* [4], the term affect is preferred here because it is an "[...] umbrella term that covers all evaluative - or valenced (i.e., positive/negative) - states such as emotion, mood, and preference." [11, p.10], and thus keeps the model's complexity manageable. While it has been shown that timbre contributes to emotion judgements in larger structures of music [7], also alterations in musical emotional expression can be detected at the level of single notes [6], such that isolated instrument sounds contain cues that indicate affective expression independently of the presence or absence of other cues such as melodic ones [4, 7]. Because of the high inter-consistency of listeners' affects in judgements for both [4, 7] levels and the property of timbre being a form-bearing dimension for a composer or performer, we also see timbre as "[...] central to interpretative decision-making and projection in performance" as argued by Holmes [9].

However, the issue with shaping sounds generated by an audio synthesizer is that its sound shaping parameters are in most cases not grounded in the domain of *perception* (timbral qualities) and *affect* but in the generating structure and its technical functioning [21, 22]. Therefore, without expert knowledge, making change to timbre in these domains is non-intuitive.

While our long-term goal is to develop a full affect related audio synthesis environment, for this contribution, the primary aim is to model a relationship between timbre and affect. In this way individual sounds that may be part of a larger musical context can be automatically given a label or value of an expected affect for the composer or listener[2]. Such labels can be *soothing, aggressive, happy, sad,...* or, as will be presented, a coordinate in Russel's Valence/Arousal (V/A) space [23]. This space is considered useful for our purposes since it eliminates ambiguity and provides a consistent global model of affect and has been applied to a large body of related work (e.g. [4, 27, 26]). Since our model is able to label data in an unsupervised manner, the labeling process can be applied to large databases of sounds with different timbres. In this way the tedious task of supervised labeling for sets of sounds with various timbres regarding the emotional affect may be avoided, thus providing the opportunity to create new applications based on Synthesis by Analysis methods as will be exemplified in the next section.

The proposed model will be created via training a Deep Belief Network [8] using a large data set created from the Freesound.org [1] database and text-mining. We've chosen to use a Machine Learning (ML) based approach since an analytic understanding of the affect of timbre is still missing as most research in the context of music so far has focused on the affect of tempo, dynamics and mode [4]. It should be emphasized that the difference to the discipline of mood classification of songs in Music Information Retrieval (MIR) is the type of data analyzed and therefore the implied model: here we do not use a mixture of sounds but isolated, singular ones that are predominantly short (compared to a recording of a large musical structure) and detached from a harmonic and rhythmic context. Since songs themselves communicate

[1]The underlying data set will be made available online at `http://bit.ly/15Mi9WP`

[2]To simplify the model, throughout this paper we do not distinguish between the effective perspective of composer and listener.


*NIME'13,* May 27 – 30, 2013, KAIST, Daejeon, Korea.

emotion on additional layers (structurally) [7] and since this also implies the utilization of additional feature data for the audio analysis, a solely timbral model, such as ours, is -in comparison- expected to perform *a-priori* worse.
The paper is organized as follows: first potential applications of the proposed model are shown, second related work is referenced and third the acquisition of the training data set is laid out. Then the performance on a simplified training set of various common ML classifiers is evaluated which subsequently lead to the reason to employ a regression method based on Deep Belief Networks. After the presentation of this method, the paper concludes with an outlook of future work.

## 2. POTENTIAL APPLICATIONS

One use-case scenario for our research is to combine such a timbre-affect model with an audio synthesis model, primarily to help novices shape sounds with respect to a desired emotional affect. Therefore, we see music creation rather than pure analysis as the main application field of our contribution, such as controlling the audio synthesis by using the emotional affect as hyper-parameter, created in the following way:

- establish a corpus of sounds generated by an audio synthesizer instantiated with permutations of its structural parameters
- retain the generating parameter settings and label each sound using the proposed model
- use the labels as meta-mapping for the structural parameters (depending on the audio synthesizer these structural parameters can be interpolated)

In this way exemplifying use-cases in regards to control timbre could be:

- present the user(s) with a pre-selection of sounds fitting to the mood they want to create with a song (static mood)
- help user(s) to create transitions from one mood to another (dynamic mood)
- help users create a static mood (e.g. *calm*) but offer a set of different timbres with a similar emotional affect so that the music is still dynamic/varying

## 3. RELATED WORK

There are several timbral qualities that have been studied and are known to relate to or to be fundamental to affect [29, 19]. These can be roughly grouped into spectral energy, -structure and -variation. Depending on the method, they can be quantitatively or qualitatively measured in the spectral, temporal or spectro-temporal domain; examples are the spectral centroid (geometric center of the spectrum; brightness), spectral spread (standard deviation of the spectrum), HF-LF ratio (high to low energy ratio), attack slope or inharmonicity (deviation of partials from the harmonic frequencies). In [4], Eerola *et al.* performed an extensive study to relate such audio features to the affect dimensions. The analyzed sounds were orchestral samples (105) with various articulations. Within this study correlations between affect and various audio features have been found for both affect dimensions. Furthermore it was possible to construct a model to predict the affect from feature data with reasonable results via linear regression. As will become clear in the section Machine Learning, a linear model is not sufficient for our data set. It was also shown in [4] that the two affect dimensions led to the most consistent results when compared to an affect model employing the three dimensions of valence, arousal and tension [25]. Because the aim of their study not to rely on artificial sound generation schemes is antipodal to ours, we decided not to use their provided data set. It is of particular interest for our model not to emphasize a certain set of timbres but to allow the classification of inherently inharmonic or initially "non-musical" sounds. For an in-depth review of the applied audio features, it is recommended to consult the contribution by Eerola *et al.* [4] as we made use of the same toolbox (MIRToolbox) [15] with a congruent feature set.
Similar to our goals, Oliveira *et al.* [17] developed a model that used a ML approach to control the selection of timbres in relation to their affect in an automated music composition system [18]. Their model yields a correlation of 75% between timbre (audio features) and affect labels. The affect labels (discrete V/A space) were acquired in a previous listener study and are based on short orchestral pieces. With respect to our own research, we especially see the derivation of affect from musical pieces as problematic since a model informed by *timbral* qualities will then be biased towards the affect of the pieces themselves. Le Groux *et al.* [16] follow a different path by first designing a physically informed synthesizer (modal synthesis) to generate sounds in accordance to perceptually relevant features and then evaluating their impact on the affect in a listener study. They show that the spectral centroid as well as the spectral flux are directly related to arousal. However, no significant relationship to valence could be established. This and the comparatively simple synthesis model (percussive sounds) may reveal the issue that more variation in timbre is necessary to cover the wealth of affects. The low number of participants (10) may also contribute to this.
In summary, a data set for our purpose would have to consist of a variety of different timbres, musical (acoustic, synthesized) and non-musical ones, as well as covering a large spectrum of emotional affects. Furthermore, as a rule of thumb, it is usually recommended to have at least 10 to 20 times more observations than predictors (e.g. audio features) to be able to perform a meaningful multivariate analysis of the data. Getting hold of such a data set is therefore an important item for our work. There are related data sets available, but to our knowledge they have either limited musical relevance such as e.g. The International Affective Norms for Digitized Sounds (IADS-2) [3] which focuses on affects of real-life stimuli, or do not meet the requirement to have isolated timbres as the audio data consists of a mixture or sequence of several timbres like the large data sets used in the MIR community (e.g. MTurk [30]) for song emotion (and genre) classification. To our knowledge, the only exception is the data set used by Scott *et al.* [27]. It consists of the individual tracks of a song labeled with continuous V/A labels. However, this data set is biased in timbre since it contains only 50 songs from the genre of Rock music. This issue may be prevalent in other related MIR data sets as well (USPop2002; orchestral film scores).

## 4. DATASET FOR TRAINING

Because of the general lack of a suitable data set, we decided to construct one based on a large set of samples and related meta-data downloaded from the Freesound.org [1] library. The insight into the samples' emotional affect is gained by processing the meta-data. The predictor data for training is generated by audio analysis. Freesound.org is an online collaborative sound database where people can share recorded audio clips (royalty free) and, among other things, tag these sounds. This database focuses to a large degree on the creative use of the material by sound and video artists.

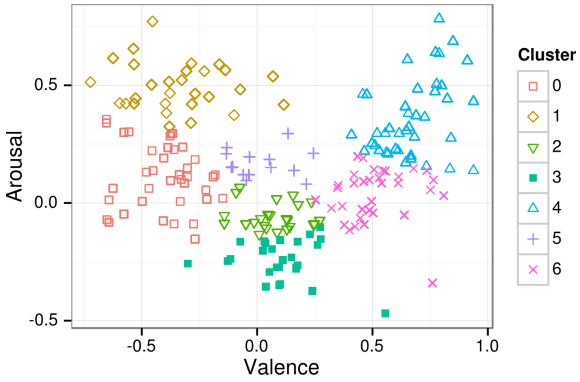


**Figure 1: V/A values of all samples in the data set; shapes/colors specify the cluster membership for the preliminary performance analysis**

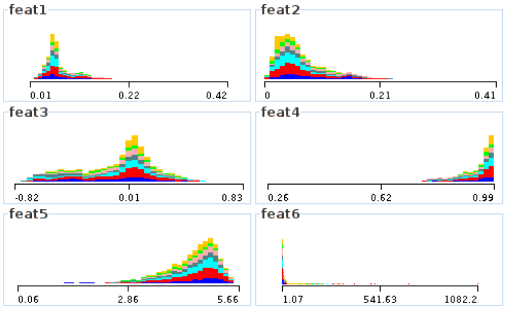


**Figure 2: Histogram of values of top six features according to clusters (cluster membership color encoding as in fig. 1)**

The main idea for the data acquisition is that the audio data can be used as source for audio features and the linked tags as source for a folksonomy discerning the emotional affect of the audio data. Thus, we downloaded 139155 samples (May 2012) with the accompanying meta-data as first step.

## 4.1 Dictionary Analysis

The tag data are based on a narrow folksonomy [5], so a single tag can be assigned only once to each sound. For the 139155 sounds, 39337 unique tags have been used, yielding an average of 6.64 tags per sound. As pointed out by Font and Serra [5], the Freesound.org folksonomy is quite noisy and therefore suffers from inconsistencies such as synonymy and polysemy which complicates the extraction of structural information. An issue of properly deriving V/A values for the tags is that the tags are part of different semantic categories (subjective, context, content,...). The goal is to estimate the affect from all categories (ideally emphasizing the *subjective* category) and disregarding all information that can lead to inconsistencies. In view of that and to relax the previously mentioned issue of noisiness, a combination of finding sentiment related synonyms for each tag and of filtering tags based on the content of the words was employed. Furthermore, tag data and dictionary data were stemmed (using SentiWordNet [1]).

The dictionary analysis per sound is roughly accomplished as follows: first sentiment related synonyms for each tag are retrieved, then for each of these synonyms it is evaluated whether a V/A value pair exists. If it exists, the tag and the synonym are compared against a blacklist containing terms of context and a whitelist containing terms of musical affect. The V/A values are kept if the tag or synonym are both not

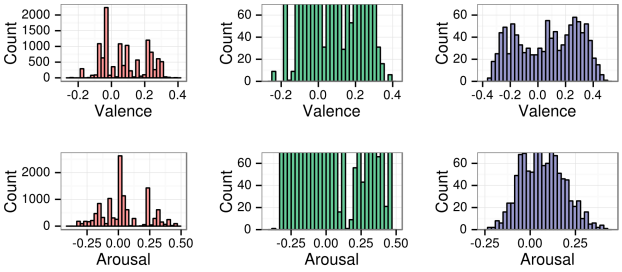


**Figure 3: Histogram of V/A values (left to right) in our data set, the same magnified and in ANEW**

blacklisted or if either one of them is whitelisted, otherwise the V/A value is discarded. Each sound is associated with the mean V/A value of all tags. The sentiment related synonyms are generated with SentiWordNet3.0 [1]. It is based on a dictionary that holds 117684 synonyms. The blacklist dictionary is based on on the General Inquirer Augmented Spreadsheet[3] and was specifically compiled to contain only words of matters that are considered unrelated to music and its affect. Furthermore it includes terms that are deemed to create inconsistencies, such as the word 'piano' which has a V/A value but refers to the content of the sample, therefore it would colorize the valence arousal value for an accordingly tagged sample. The dictionary of the Musical Adjectives Project[4], serves as whitelist. It consists of 690 collaboratively collected adjectives that describe and categorize emotions in music. Finally, The Affective Norms for English Words (ANEW) dictionary [2] was applied to derive V/A values. It holds 1034 English words including verbs, nouns, and adjectives whose emotional affects have been evaluated in a large study. The originally included domain of dominance has been ignored in our musical context (singular timbres).

In the end, 11324 (12.3%) sounds were kept from the original data set for further processing, each having a V/A value associated. Figure 1 shows the V/A values for all sounds in the final data set. When comparing these to the ones in the original ANEW dictionary (cp. figure 3), one can see that the distribution is slightly biased as singular coordinates have a disproportionately high occurrence. In the data set 203 V/A coordinates are unique which contrasts to the 813 in the dictionary. The average number of tags with a V/A value in the data set is 1.32 per sound. The number of unique tags with V/A value is 445 (an excerpt is given in table 1) which were resolved into 405 sentiment related synonyms, thus it may cautiously be concluded that the original tags describe differing concepts at large. Concluding, aliases have been created because different concepts (in the dictionary) have a similar emotional affect. 2.2% (e.g. table 1) of the most frequently used V/A valued tags make up 44.5% of the complete set of V/A valued tags which shows that this transformed folksonomy is less diverse in vocabulary than the original one. Nevertheless, using the data set seems plausible as the *spread* of affect in the data set is comparable to the one of the ANEW dictionary (as indicated by figure 3), so diversity and structure of affects can be considered to be adequately represented. It should be noted that structurally this is also true in comparison to the IADS-2 and the International Affective Picture System [13] data sets.

Given our representation of affect as coordinates in the V/A

[3] `http://www.wjh.harvard.edu/~inquirer/`

[4] `http://themusicaladjectivesproject.wikispaces.com/`

| User tag with V/A value | Count | User tag with V/A value | Count |
|---|---|---|---|
| 'dark' | 1722 | 'dream' | 294 |
| 'spooky' | 832 | 'smooth' | 274 |
| 'evil' | 515 | 'terror' | 261 |
| 'dirty' | 472 | 'cry' | 245 |
| 'happy' | 339 | 'frightening' | 236 |

**Table 1: Excerpt from the 25 most frequently used sentiment related tags with V/A value**

| **Classifier** | 368 feat. **error %** | top 25 feat. **error %** | top 4 feat. **error %** |
|---|---|---|---|
| Naïve Bayes | 70.61 | 67.82 | 74.75 |
| Bayes Net | 65.73 | 65.85 | 71.84 |
| RandomForest | **55.59** | **52.59** | 71.09 |
| LibSVM | 76.37 | 76.20 | **70.97** |

**Table 2: Preliminary performance (classification error) of common ML-methods analysis with simplified data set importance ordered features**

space, we assume that these coordinates can be linearly combined, including even V/A coordinates of *opposite* affects (e.g. happy and sad), leading to a composition of affect coordinates which is representationally unambiguous. On the other hand the combination of affects in a discrete model (e.g. bags of labels) may not be as trivial since a simple union of labels might either lead to ambiguity (with possibly severe impact for the ML model) or sparsity of the affect model (e.g. when using majority voting ).

## 4.2 Audio Analysis

First the samples were normalized, the preceding and subsequent silence was trimmed, and finally they were resampled to 22kHz sample rate. Then all audio files were analyzed using the MIRToolbox [15] to generate audio features (30) in accordance with the presented focus, namely features that relate to timbral and dynamic qualities of these sounds only. This resulted in 684GB of feature data, most of them being time-series data. The size of the data had to be reduced in order to circumvent memory constraints for the applied ML-algorithms. A further justification to reduce the features to a single descriptor per sample is that the available V/A data is static. Statistical properties such as variance, kurtosis, entropy, or periodicity were calculated for each time-series feature, thus the time series data was collapsed onto a single feature descriptor. This transformation reduced the data set to 48MB. All feature vectors where then concatenated to create the observation/feature matrix. Vectors containing undefined values were removed while the remaining ones were standardized by z-score transformation as suggested in [24]. The final observation/feature matrix represented 11324 observations with 368 features each. The downside of the noisiness of our data (e.g. due to the empirical nature of the applied dictionaries, variations in recording quality, etc. ) is that training a representative model may be difficult or impossible to achieve, the benefit however may be the applicability of such a model to real world data. In conclusion, the data set contains 30.8 times more observations than predictors, more importantly the requirements postulated beforehand have been met inasmuch as the wealth of emotional affects and a variety of timbres from unrelated sound sources is represented.

# 5. MACHINE LEARNING

To establish a model, the task is to find a mapping between audio features and emotional affect values. For a *preliminary* overview of how various established ML algorithms perform, the aspired mapping is *simplified*: the former continuous V/A coordinates were clustered into 7 distinct emotional affect classes which roughly resemble the discrete emotional affect classes when basic affects are represented on the V/A plane [13, 3, 2]. Hence, the ML task was relaxed to creating a mapping from audio features to these 7 classes. Figure 1 shows the color encoded cluster correspondence of each sample on the V/A plane. To overcome the issues of many ML classification algorithms (such as SVM) of being sensitive to ambiguous or inconsistent features [24, 29] a feature selection is usually applied. Although generally suggested [24], we did not perform a wrapper-based feature selection since that would have involved testing all $2^{368}$ combinations of feature sets for each classifier. Instead, feature performance was evaluated with three sets of importance ordered features where the importance of the features was estimated using ReliefF [28]. Table 2 shows the results of this performance evaluation for various standard classifiers. The measure was the misclassification rate of the predicted class and the observed one. In spite of the fact that some classifiers (e.g. Naïve Bayes) make the assumption that the features are independent which is certainly not true for audio features, the initial idea is to see how, in general, prevalent classifiers perform. The SVM has been applied using Radial Basis Functions with a grid search to find the optimal parameters for the kernel. One central issue regarding the feature data is shown in figure 2, namely that, on a per feature basis, the distribution of the features in relation to cluster membership is not discriminative. This is in accordance with observations made in other works as pointed out in [12]. Although the presented algorithms have already been successfully applied in similar contexts [12], for our data set the results are disappointing as all algorithms show unreliable classification performance. The results also suggest that the data set may be very noisy and may contain inconsistent labeling. Otherwise, from the empirical evidence as shown in [12], the SVM based approach should have performed better. Another indicator is that RandomForest, which is generally considered to be less susceptible to these issues, performed best. It is also possible that the tested ML algorithms are, from a computational point of view, not able to find a generalizing pattern in the feature space for discrimination, or that the feature data actually originate from a higher dimensional manifold. In particular, the measurement of error may be inappropriate for the task as an error of absolute misclassification introduces an artificial discrete segmentation of the V/A space without accounting for "near misses". The latter part of this section introduces the *final* ML method and an improved measure of error, and concludes with a performance evaluation.

Informally, Deep Belief Networks (DBNs) [8] share architectural similarities with neuronal networks. They belong to the class of deep architectures, hence their output is produced by consecutive layers of computational units. As a result, they allow for hierarchical learning such that a model is created progressively from a low to a high level structure throughout the layers. With this, a deep architecture may be able to model highly complex functions with only a limited number of parameters whereas the few layers of shallow models (e.g. linear models, single layer neural networks, kernel SVMs) may require an exponential amount of computational units [14]. A DBN [8] is a probabilistic generative model whose building blocks (layers) are Restricted Boltzmann Machines (RBMs) which themselves model stochastic latent variables as part of a two-layer neuronal network. Learning a DBN is performed in two steps. First, the layers are trained consecutively in a greedy fashion so as to initial-

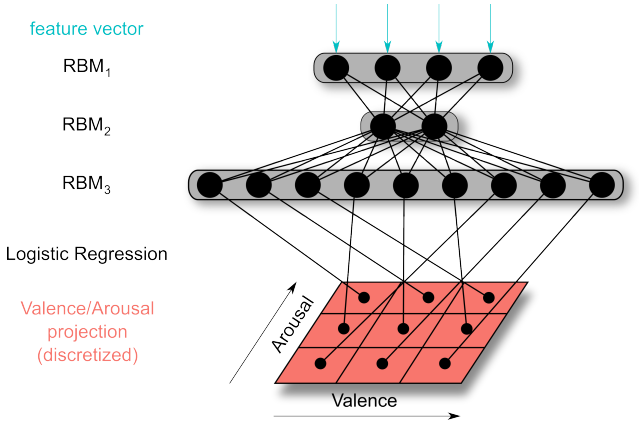


**Figure 4: Schematic of the applied model (number of layers/nodes is exemplary)**

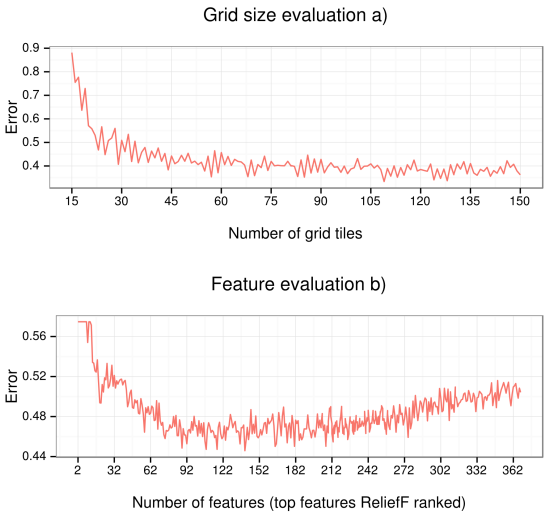


**Figure 5: Performance evaluation w.r.t. to a) grid size and b) number of importance ordered features (fixed grid size: 25)**

ize the RBM weights (generalization phase, *unsupervised*), followed by fine-tuning the network's weights with respect to a chosen learning criterion (specialization phase, *supervised*). The generalization phase is performed by training the first RBM layer to generate the input data and then by successively feeding the output of each layer as input to the next. Thus, with each layer the original input is abstracted further. In our work, the learning objective of the specialization phase is to minimize the error between observation and prediction of V/A coordinates given the feature vector. For the mapping from the last RBM layer onto the V/A plane, a logistic regression was employed whose output logit variables were interpreted as encoding of grid tiles of a uniformly discretized grid of the V/A plane. In this 1-of-$K$ encoding each output variable corresponds to exactly one tile of the grid. Note that this differs from the previous representation in that it is more general than the limited number of generated affect clusters and hence embodies a more detailed representation of affect. Employing logistic regression allows for fine-tuning the whole network via supervised gradient descent on the negative log-likelihood cost function. Furthermore, the combination of 1-of-$K$ coding and the softmax saturation in the logistic regression proved to be beneficial as early attempts to use a linear regression method similar to [26] led to disappointing results. This means, instead of a continuous representation of the V/A space, the model uses a discretized one. However, with increasing grid resolution it converges towards a continuous solution. As will be shown, it is possible to create a model that performs well while making rather fine granular predictions. Figure 4 schematically shows the architecture of the presented model.

The question remains what a suitable error measure for supervised training may be and, respectively, how the overall performance ought to be evaluated. As previously mentioned, a distance-based measure would be preferable, yet a suitable baseline is missing due to the generated nature of the training data. Hence we use the expected value of the distance of two random points *on the grid* based on the following reasoning: Given a feature vector and its corresponding observation on the V/A grid, the worst prediction of the model is a randomly chosen coordinate on the grid. As the error measure is the Euclidean distance between these points, namely observation and prediction, this *expected* random distance serves well as baseline. All distance measures are themselves normalized against the chosen grid size to overcome to a large degree the quantization error. Based on our initial experiments we chose to use $N$ input nodes ($N$ features) and three layers with $N$, $N \cdot 10$ and $N \cdot 10$ nodes, respectively. This expanding topology performed better than a reducing one suggesting that the feature data may live on an even higher dimensional manifold. All performance evaluations were done with 10-fold cross-validation using stratification w.r.t. the V/A coordinates (which was not applicable in all cases due to the uneven distribution of samples). To evaluate the impact of the grid size we computed the model for varying grid sizes given the top 200 of the ReliefF importance ordered features. Due to the complexity of the model and the DBN training algorithm itself, this calculation took 83 days on two NVIDIA GTX 580 GPUs with 3GiB memory.

Figure 5a) shows the regression error in relationship to the grid size; for the examined range, it can be seen that the error decreases with increasing resolution (increasing number of grid tiles ; the minimum is 124). The comparatively large oscillations are attributed to the uneven distribution of V/A points in combination with the uniform grid. Here, small clusters of V/A points may be merged on a single grid tile or spread across a small neighborhood of grid tiles according to the boundaries of a grid tile. Furthermore, we evaluated the impact of the selection of ReliefF ranked features for an exemplary chosen fixed 25×25 grid (cmp. 5b). In our case we see how the error decreases with the number of features until the tipping point (the top 158 features) of model complexity versus training time is reached. Nevertheless, with $124^2$ grid tiles and 158 features, the error approaches 31%, which is, given the potential issues of the data set and learning algorithms as pointed out before, very reasonable. This corresponds to a distance of ∼18 tiles (15% of the grid in one dimension), therefore we regard this estimated performance as satisfactory for the use-case as introduced at the beginning.

## 6. CONCLUSION

We argued that from the performative and compositional point of view, insight into the emotional affect of timbre would be beneficial to music creation, especially for novices. We pointed out that in order to gain this insight and for practical applications, a model representing this relationship between timbre and affect would be necessary. The largest part of this contribution dealt with the development

of such a model using a ML based approach. For this, we established a suitable new data set based on real world data and applied recent developments in ML to this task of finding a mapping from audio feature data to affect. This included the development of a respective error measure so as to estimate the quality of the model, yielding promising results. For future work, we aim at improving the model by re-assessing the various parameters of the involved ML algorithm and the feature data themselves. For example, instead of the current high level features, the model can be trained to extract relevant features implicitly, e.g. from the magnitude spectrum of the audio data as shown in [26]. We are also currently developing prototypical music applications with the aim to evaluate the model qualitatively.

# 7. REFERENCES


[1] V. Akkermans, F. Font, and J. Funollet. Freesound 2: An Improved Platform for Sharing Audio Clips. *Late-breaking demo*, pages 4–6, 2011.

[2] M. M. Bradley and P. J. Lang. Affective Norms for English Words ( ANEW ): Instruction Manual and Affective Ratings. *Psychology*, Technical(C-1):0, 1999.

[3] M. M. Bradley and P. J. Lang. The Int. Affective Digitized Sounds Affective Ratings of Sounds and Instruction Manual. *Emotion*, pages 29–46, 2007.

[4] T. Eerola, R. Ferrer, and V. Alluri. Timbre and affect dimensions: Evidence from affect and similarity ratings and acoustic correlates of isolated instrument sounds. *Music Perception: An Interdisciplinary Journal*, 30(1):49–70, 2012.

[5] F. Font, G. Roma, P. Herrera, and X. Serra. Characterization of the Freesound Online Community. In *Third Int. Workshop on Cognitive Information Processing*, 2012.

[6] K. N. Goydke, E. Altenmüller, J. Möller, and T. F. Münte. Changes in emotional tone and instrumental timbre are reflected by the mismatch negativity. *Brain Research*, 21(3):351–359, 2004.

[7] J. C. Hailstone, R. Omar, S. M. D. Henley, C. Frost, M. G. Kenward, and J. D. Warren. It's not what you play, it's how you play it: timbre affects perception of emotion in music. *Quarterly journal of experimental psychology (2006)*, 62(11):2141–55, Nov. 2009.

[8] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[9] P. a. Holmes. An exploration of musical communication through expressive use of timbre: The performer's perspective. *Psychology of Music*, Mar. 2011.

[10] D. Huron. Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles. *Music Perception*, 19(1):1–64, 2001.

[11] P. N. Juslin and D. Västfjäll. Emotional responses to music: the need to consider underlying mechanisms. *Behavioral and Brain Sciences*, 31(5):559–575; discussion 575–621, 2008.

[12] Y. Kim, E. Schmidt, and R. Migneco. Music emotion recognition: A state of the art review. *Proc. ISMIR*, (Ismir):255–266, 2010.

[13] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. *Int. Affective Picture System (IAPS): Technical Manual and Affective Ratings*, volume 77. The Center for Research in Psychophysiology, University of Florida, 1997.

[14] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. *Proc. of the 24th int. conf. on Machine Learning (2007)*, (2006):473–480, 2007.

[15] O. Lartillot, P. Toiviainen, and T. Eerola. A Matlab Toolbox for Music Information Retrieval. *Data Analysis Machine Learning and Applications*, 35(M):261–268, 2008.

[16] S. Le Groux and P. F. M. J. Verschure. Emotional Responses to the Perceptual Dimensions of Timbre: A Pilot Study Using Physically Informed Sound Synthesis. *ccrmastanfordedu*, pages 1–15, 2010.

[17] A. Oliveira and A. Cardoso. Emotionally-controlled music synthesis. *Encontro de Engenharia de Áudio da AES*, pages 10–14, 2008.

[18] A. P. Oliveira and A. Cardoso. A musical system for emotional expression. *Knowledge-Based Systems*, 23(8):901–913, Dec. 2010.

[19] A. Padova, L. Bianchini, M. Lupone, and M. O. Belardinelli. Influence of Specific Spectral Variations of Musical Timbre on Emotions in the Listeners. In R. Kopiez, A. C. Lehmann, I. Wolther, and C. Wolf, editors, *Proc. of the 5th ESCOM Conf.*, number September, pages 227–230, 2003.

[20] A. D. Patel. *Music, Language, and the Brain*, volume 66. Oxford University Press, 2008.

[21] J.-C. Risset. and Perception of Musical Sounds The Sound of Music. *Perception*, pages 1–12, 2003.

[22] A. Röbel. Between Physics and Perceptions: Signal Models for High Level Audio Processing. *DAFX - Int. Conf. of Digital Audio Effects*, 2010.

[23] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[24] P. Saari, T. Eerola, and O. Lartillot. Generalizability and Simplicity as Criteria in Feature Selection: Application to Mood Classification in Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1802–1812, Aug. 2011.

[25] U. Schimmack and A. Grob. Dimensional models of core affect: a quantitative comparison by means of structural equation modeling. *European Journal of Personality*, 14(4):325–345, 2000.

[26] E. Schmidt, J. Scott, and Y. Kim. Features Learning in Dynamic Environments: Modeling the Acoustic Structure of Musical Emotion. *Proc. ISMIR*, (Ismir):325–330, 2012.

[27] J. Scott, E. Schmidt, M. Prockup, B. Morton, and Y. Kim. Predicting Time-Varying Musical Emotion Distributions from Multi-Track Audio. *Proc. of the 9th int. sym. on Computer Music Modelling and Retrieval, London, UK*, (June):19–22, 2012.

[28] M. R. Sikonja. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, 53(1-2):23–69, 2003.

[29] Y. Song, S. Dixon, and M. Pearce. Evaluation of Musical Features for Emotion Classification. In *Int. Society for Music Information Retrieval Conf.*, number Ismir, pages 523–528, 2012.

[30] J. Speck and E. Schmidt. A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation. *Int. Society for Music Information Retrieval Conf.*, (Ismir):549–554, 2011.