NOVEL AFFECTIVE FEATURES FOR MULTISCALE PREDICTION OF EMOTION IN MUSIC

Naveen Kumar[1], *Tanaya Guha*[1,2], *Che-Wei Huang*[1], *Colin Vaz*[1], *Shrikanth S Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory (SAIL), University of Southern California, Los Angeles
[2]Electrical Engineering, Indian Institute of Technology Kanpur, India

ABSTRACT

The majority of computational work on emotion in music concentrates on developing machine learning methodologies to build new, more accurate prediction systems, and usually relies on generic acoustic features. Relatively less effort has been put to the design and analysis of features that are particularly suited for the task. This paper proposes two features (*compressibility* and *sparse spectral components*) that can efficiently capture the emotion-related properties in music. These features capture the overall affective characteristics of music (global features). We demonstrate that they can predict emotional dimensions (arousal and valence) with high accuracy as compared to generic audio features. Secondly, we investigate the relationship between the proposed features and the dynamic variation in the emotion ratings. To this end, we propose a novel Haar transform-based technique to predict dynamic emotion ratings using *only* global features.

Index Terms— affect; features; music; prediction;

1. INTRODUCTION

Human generated signals, such as speech and music, are usually rich in paralinguistic information like emotion. Music, especially, is known to have the ability to evoke emotions in listeners, and also to alter their emotional states [1, 2, 3]. With the readily available and vast digital music libraries and online streaming services, the general approach towards organizing, searching and retrieving music information has grossly changed. It has been shown that emotion is a natural criterion for music search and organization [4, 5]. Consequently, there has been a growing interest in automatic understanding of the emotional content in music in the recent years [6, 7, 8].

The perception of emotion in music is highly subjective, and difficult to quantify. The automatic understanding of music emotion involves predicting either categorical (e.g. happy or sad) [9] or dimensional emotion labels [7]. The dimensional labels are typically measured along the affective dimensions of arousal (A) and valence (V) which intend to correspond to the internal human representations of emotion [10]. These labels can be either computed as dynamic quantities corresponding to the temporal evolution of emotion, or as static measures that quantify the overall representation of emotion.

The majority of computational work on emotion in music concentrates on developing acoustic feature-based machine learning models [6, 11, 12], where the focus is on building better and more suitable algorithms for the prediction phase [6, 11, 12]. Relatively less effort has been put towards designing features that are particularly suited for the task. Most, if not all, studies limit themselves to using general purpose speech and music processing features - both low-level (e.g. chroma, mel frequency cepstrum coefficients (MFCC), statistical spectral descriptors) and high level (e.g. pitch, timbre, harmonicity) [13, 14]. Among these features, only chroma has been shown to directly capture affective information by perceptual experiments on synthesized signals from chromagrams of original music signals [14]. A recent work proposes a deep belief network-based approach to learn emotion features from music [15]. A comprehensive review of the features used in music emotion prediction along with several novel feature representation techniques are discussed in [16].

In this paper, we propose two novel affective features, namely *compressibility* and *sparse spectral component (SSC)*. Both of these features capture the global characteristic of a music signal. The compressibility features is practically parameter-free and fast to compute. The SSC feature captures the emotion-related spectral patterns in music by learning a non-negative matrix factorization (NMF). To establish the usefulness of these features, we compare their ability to predict song-level (static) A-V ratings with other well known acoustic features on the MediaEval 2014 database [17]. We observe that the proposed features are highly useful in emotion prediction as compared to several popular acoustic features.

In addition to the design and evaluation of affective features, we also explore the relationship between global features and emotion ratings at different scales. This is motivated by the observation that the role played by global music features might differ depending on whether the perceived emotion is overall or transient. To this end, we propose a novel Haar transform-based technique to predict the continuous (dynamic) A-V ratings using the compressibility feature (which returns a scalar value per song). This experiment helps in understanding how the emotion annotations at various scales are affected by the features computed at different scales.

The rest of this paper is organized as follows. Section 2 describes the proposed features, Section 3 performs a thorough evaluation of the proposed features. This is followed the proposed dynamic prediction method in Section 4, discussion in Section 5, and conclusion in Section 6.

2. PROPOSED AFFECTIVE FEATURES

In this section, we describe the new affective features - *compressibility* and *sparse spectral components (SSC)*.

2.1. Compressibility

We hypothesize that intense emotion like high arousal or valence is evoked by complex interplay of various musical elements resulting in rich spectral information, and therefore, the complexity of a music signal may be closely correlated with its affective content. Although there are numerous ways to estimate data complexity, we propose to use the concept of Kolmogorov complexity [18] as a measure of

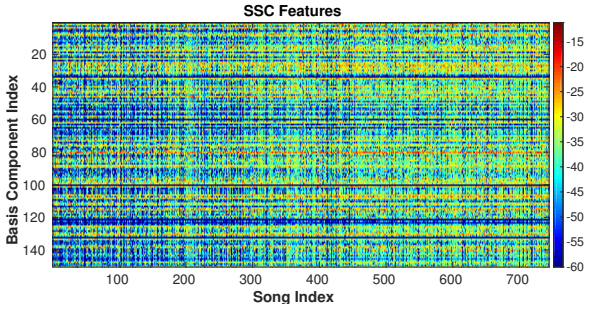

Fig. 1. Visualization of the static NMF features for the songs in training set. The songs are sorted by increasing valence from left to right and the basis components are sorted by increasing spectral center of mass from top to bottom.

complexity of a music signal. Kolmogorov complexity-based measures have been successfully used in image clustering [19, 20].

The notion of complexity of a given data is related to its randomness. For example, the binary string 1101010001 is considered more complex compared to the string 0101010101, because the latter contains a regularity (a repeating pattern), and therefore is less random. The Kolmogorov complexity formalizes this concept. Given a string x, its Kolmogorov complexity $K(x)$ is defined as the length of the shortest program that can effectively produce x on a universal computer, such as a universal Turing machine [18]. However, $K(x)$ is non-computable in general. In practice, it is often approximated by $C(x)$ - the length or the file size of the compressed data. Intuitively, the more a string can be compressed, the lower is its complexity.

Note that $K(x)$ is simply approximated by the compression length of x. Hence, in the context of music, we can take advantage of the various highly efficient music compression algorithms. This leads to a simple implementation and fast computation of this feature. We thus compute the complexity of a music signal x by its *compressibility* $\mathcal{C} = \frac{K(x)}{U(x)} \sim \frac{C(x)}{U(x)}$, where $U(x)$ denotes the uncompressed length of x. Thus, a single scalar value of $\mathcal{C}$ is obtained for each clip.

In order to compute the compressibility feature, we first convert an MP3 music file, x, to the raw audio format which provides the uncompressed data length $U(x)$ (ideally, $U(x)$ should be the raw uncompressed data length. However, for the given database, the raw files were not available). The resulting 'raw' audio file is then compressed using a lossless audio codec (FLAC) to obtain $C(x)$. The ratio of $U(x)$ and $C(x)$ determines the value of $\mathcal{C}$.

2.2. Sparse Spectral Components (SSC)

We observe that songs with similar valence ratings tend to share similar spectral characteristics. This can be thought of as certain spectral components evoking a high or low valence emotional response in listeners. To test this idea, we first construct a dictionary of spectral components for the samples in our training set. We use NMF [21] to learn this dictionary of spectral components for a set of training samples. NMF takes a spectrogram $\mathcal{V}$ and factorizes it into a dictionary $\mathcal{W}$ and time-activation matrix $\mathcal{H}$. Each atom in $\mathcal{W}$ represents a spectral component and $\mathcal{H}$ gives the extent to which each atom in the basis is activated at each time frame. The matrices $\mathcal{W}$ and $\mathcal{H}$ are estimated by minimizing the cost function shown in (1),

$$C = \sum_i \sum_j \mathcal{V}_{ij} \ln \left(\frac{\mathcal{V}_{ij}}{(\mathcal{W}\mathcal{H})_{ij}} \right) + \lambda \|\mathcal{H}\|_1, \qquad (1)$$

with the constraint that all elements in $\mathcal{V}$, $\mathcal{W}$, and $\mathcal{H}$ are non-negative. The subscripts refer to the element in the ith row and jth column of the matrix, and λ denotes a parameter that encourages sparsity in $\mathcal{H}$.

To learn a dictionary for all songs in the training set, we concatenated all the training set songs together and created a spectrogram by taking the short time Fourier transform (STFT). A frame length of 50 ms with a 30 ms shift was used. We perform NMF on this spectrogram to get a dictionary $\mathcal{W}$ that contains the most common spectral components in the training set. In this paper, we set the number of dictionary elements to be 150 and use $\lambda = 80$.

Once the dictionary is learned, the feature representation can be learned for each song by obtaining a sparse representation $\mathcal{H}^{(i)}$ from the ith song's spectrogram $\mathcal{V}^{(i)}$ using atoms of the dictionary $\mathcal{W}$. We employ the previously described NMF technique by fixing the dictionary to $\mathcal{W}$ learned on the training set to obtain this sparse time activation matrix $\mathcal{H}^{(i)}$.

Finally, a global feature representation is obtained by averaging each time-activation matrix $\mathcal{H}^{(i)}$ along the temporal direction, giving us a static 150-dimensional feature that represents the average amount of activation of each dictionary atom in the song. Equation 2 shows this procedure:

$$S_i(k) = \frac{1}{T} \sum_{t=1}^{T} \mathcal{H}^{(i)}_{tk}, \forall k = 1 \ldots 150 \qquad (2)$$

where $\mathcal{H}^{(i)}$ contains T temporal frames and the subscript t refers to the tth column of $\mathcal{H}^{(i)}$.

Figure 1 shows the static NMF features for the training set. Each column shows the average activation of the basis components S_i in the log scale for each song, with the songs sorted by increasing valence from left to right. Basis components are sorted by increasing spectral center of mass (higher indices correspond to higher frequency basis components). One can see that the amount of activation for components 65 to 75 and 102 to 150 increases as the song valence increases. This suggests that high valence music contains greater amounts of middle and high frequency content than low valence music. The performance of this feature is evaluated in section 4 in terms of its ability to predict the AV ratings, as direct correlations with the AV ratings are not possible to compute for this high-dimensional feature.

3. PERFORMANCE EVALUATION OF THE PROPOSED FEATURES

In this section we evaluate the two proposed features on the MediaEval2014 database [17], by their ability to predict the *static* emotional rating of the songs. Their performance is also compared with popular features, such as chroma and MFCCs.

3.1. The database

The MediaEval2014 database contains 1744 music clips provided for the *Emotion in Music challenge* at the 2014 Mediaeval Workshop [17]. The annotations for these music clips are collected from multiple annotators on Amazon mechanical turk. Each clip is 30 s

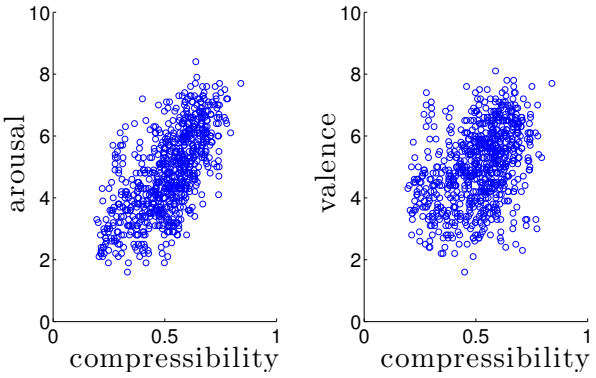

Fig. 2. Scatter plots for 744 songs in the training set showing correlations between compressibility and manual arousal ratings ($\rho = 0.65$), and compressibility and valence ratings ($\rho = 0.39$).

long, which are extended to 45 s during annotation to give annotators additional practice. For each clip, global A and V ratings are obtained on a scale of $1-10$ from each annotator. Mean of these annotator ratings is considered as the ground truth label, and used for experimental validation. Apart from the static emotion ratings, this database also contains dynamic A-V annotations for each song clip at every 0.5 sec that capture the evolution of affect over time. We use 744 songs to construct the training set, and the remaining 1000 songs form the test set for evaluation.

3.2. Evaluation of compressibility

Compressibility is a global feature and returns a single value per clip. To evaluate the performance of this feature directly, correlations between $\mathcal{C}$ and the static A and V ratings are computed for the 744 songs in our training database (see Fig. 2). The compressibility feature shows high correlation with both the manual A and V ratings.

Next, we evaluate the usefulness of this feature in predicting the static A-V ratings. We observe that the A-V ratings in the database are strongly correlated with each other ($\rho > 0.6$). This is not surprising because the songs that are energetic (high arousal), usually induce a positive emotion (high valence) in the listeners. In order to use this correlation to our advantage, we choose to *jointly* predict the A-V ratings, and employ the partial least square regression (PLSR) [22, 23] method for this purpose. PLSR is a bilinear factor model that maps both the predicted and observed variables to a new latent space. It uses ideas from both the principal component analysis (PCA) and the multiple linear regression that help it alleviate multicollinearity issues in the feature and the label sets. In our case, this method allows us to jointly predict the highly correlated A and V ratings. The correlation between the predicted A-V ratings and manual A-V ratings are listed in Table 1).

3.3. Evaluation of SSC

Unlike compressibility, SSC encodes the musical components into a high-dimensional sparse vector. Due to the high-dimensionality of the feature, it is not possible to measure its direct correlations with the AV ratings. Similar to the case of compressibility, our proposed method for SSC also attempts to exploit the high correlation between the AV ratings.

Table 1. Performance evaluation of the proposed features in terms of their ability to predict static A-V ratings

Feature	ρ_{aro}	ρ_{val}
global openSMILE	0.44	0.36
chroma	0.54	0.42
MFCC	0.66	0.52
chroma, MFCC	0.69	0.52
compressibility	0.73	0.53
SSC	0.68	0.47

3.4. Comparison with other features

We compare the performance of the two proposed features with the global openSMILE feature set described in [24], and with popular features, such as, chroma and MFCCs. Chroma features consist of local twelve dimensional vectors with each dimension corresponding to the intensity associated with a particular semitone. Higher notes are wrapped back onto a single lower octave in accordance with the concept of relative pitch in music perception. Chroma features have been shown to be effective in studying emotion in music [6]. The very popular MFCCs are believed to mimic human audio perception by measuring energy in specific filterbanks on the Mel Scale [25].

To predict the static A-V ratings from the Opensmile feature set, simple linear regression is used. We use Chroma and MFCCs as local descriptors of the music signal, by extracting them at the local frame level. In order to predict the static A-V values from the feature sequence (chroma or MFCCs), we parameterize the feature sequence for each song using the autoregressive (AR) model. An AR model of orders p is denoted by AR(p) and is defined as follows.

$$\sum_{k=0}^{p} a[k]y[n-k] = x[n], \tag{3}$$

where $y[n]$ is the current value of the output and $x[n]$ is a zero-mean white noise input; $a[k]$, $k = 0, \cdots, p$, are the parameters of AR(p). Essentially, an AR model is an all-pole filter. First, PCA is performed on the raw local features, and the first q components are retained. PCA reduces the feature dimensionality as well as decorrelates each dimension. An AR model on each dimension is then estimated via the standard Yule-Walker method. Taking the parameters of all AR models on each dimension as the dynamic feature set, we train a support vector regression (SVR) model with a radial basis function kernel for predicting the static A-V ratings for each song. The hyperparameters for this system include the number of PCA components q, and the order of the AR model p. We tuned these parameters using the leave-one-out cross validation on the training set. We observed that a higher order AR model is needed for predicting arousal as compared to valence prediction.

The comparison in terms of prediction results (see Table 1) show that compressibility outperforms other features while SSC is comparable to MFCC.

4. DYNAMIC PREDICTION FROM GLOBAL FEATURES

In this section, we investigate the relationship between the proposed features and the time evolution of dynamic emotion annotations. Our hypothesis is that feature information relevant to music emotion perception is shared across different scales. To investigate this, we decompose the time series of dynamic annotations for a song in the training set into its Haar transform coefficients. Projecting into the

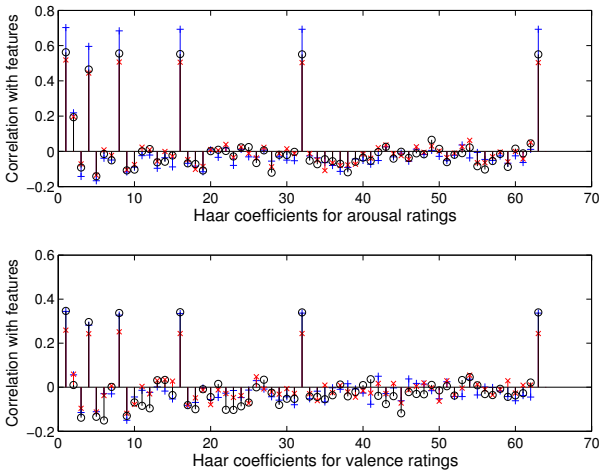

Fig. 3. Correlation of Haar coefficients with different static features. Different static features are shown using different stem plots.

Haar space allows us to analyze the emotion predictability at different scales. For instance certain features such as compressibility might be better at predicting overall annotation levels (seeing as they are correlated well with static annotations), whereas other features might be more useful in predicting local variations in emotion. Few previous works have tried to address this problem indirectly by using techniques, such as gradient boosting accompanied with feature selection [26]. Here, we explicitly model this information by using the Haar transform. Our approach also has the advantage that we can inherently incorporate the smoothness in human-annotated emotion labels, by using a basis with respect to which they are expected to be sparse.

To provide an intution as to why this approach, might be useful we first compute correlations between each of the Haar coefficients and global features as shown in Fig.3. We observe that the emotion annotations are sparse in the Haar basis owing to their smoothness in time and can be predicted reliably using only a few highly correlated coefficients. This suggests that we should be able to predict dynamic emotion ratings by predicting only a few coefficients reliably. To test this idea, we perform an experiment to predict dynamic emotion labels indirectly by first predicting their Haar coefficients, which are then transformed via an inverse Haar transform to dynamic emotion annotations.

Table 2. Correlation and RMSE results for submitted predictions on the test set

Features	System	Arousal		Valence	
		ρ	rmse	ρ	rmse
local openSMILE	Lin.Reg.	0.18	0.15	0.11	0.12
global	Haar prediction	0.22	0.12	0.11	0.09
local openSMILE	Lin.Reg. + smoothing	**0.28**	**0.13**	**0.14**	**0.10**

Features	Arousal		Valence	
	Static	Dynamic	Static	Dynamic
Global	0.73	0.22	0.53	0.22
Local	0.69	0.28	0.52	0.28

Table 3. Summary of best prediction results using features and emotion labels at different scales

5. DISCUSSION

We evaluate each method by computing the correlation between the predicted A-V labels (for all 1000 clips in the test set) and the reference ground truth obtained as the mean annotator ratings.

Results in Table 3 clearly show information relevant to music emotion labels is present in both global and local features. We further establish that this is true for both static and time-varying dynamic emotion labels. In general, global features, such as compressibility or SSE, are better at predicting static emotion ratings than the local frame-level features like Chroma. Each global feature individually yields similar performance ($\rho_A \approx 0.73, \rho_V \approx 0.52$). Regarding the local features, we note that MFCC performs better than Chroma features, but combining them together also does not improve the prediction result significantly.

We also note that the improvement is more pronounced in the case of valence prediction which is usually harder to predict. This is because valence is more subjective compared to arousal, and often requires a longer context before annotation. Our findings suggest that unlike arousal, valence depends more on the temporal interplay of local characteristics in a song.

Results for dynamic emotion prediction are presented in Table 2. We compared our proposed approach which uses global features to predict dynamic emotion series via a Haar transform to the direct prediction of emotion labels at the frame level using local openSMILE features [24]. We note that the amount of information contain at the global and local scales are in fact comparable. This underscores the importance of combining multiscale information when trying to predict dynamic emotion labels from music.

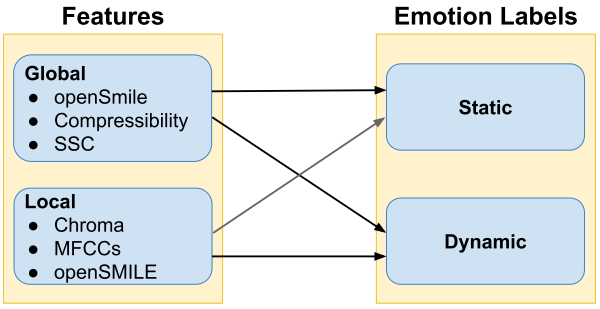

Fig. 4. Emotion information shared across multiple scales.

6. CONCLUSION

In this paper, we experimentally showed that the affective information in music signals is shared across different scales. Such information could be complementary to each other, and combining them could significantly improve the prediction performance. This contrasts with most traditional approaches used in emotion prediction

where features are extracted only at the scale at which the emotion labels are available. The intuition behind our approach stems from the observations that structure in music, unlike other human generated signals, is not completely spontaneous. Hence, emotional information in music is usually conveyed using a combination of global and local characteristics.

We have verified this hypothesis experimentally. We proposed methods to predict static ratings from local features as well as from combined information from features extracted at both the global and local scales. Our results show the merit of fusing information at multiple scales for predicting emotion ratings. In the future, we will extend this approach to modeling of dynamic emotion ratings for music. Other sophisticated dynamic modeling methods can also be employed or designed.

7. REFERENCES

[1] Lars-Olov Lundqvist, Fredrik Carlsson, Per Hilmersson, and Patrik Juslin, "Emotional responses to music: experience, expression, and physiology," *Psychology of Music*, 2008.

[2] Patrik N Juslin and John A Sloboda, *Music and emotion: Theory and research.*, Oxford University Press, 2001.

[3] Jaak Panksepp, "The emotional sources of" chills" induced by music," *Music perception*, pp. 171–207, 1995.

[4] Fabio Vignoli, "Digital music interaction concepts: A user study.," in *ISMIR*. Citeseer, 2004.

[5] Jin Ha Lee and J Stephen Downie, "Survey of music information needs, uses, and seeking behaviours: Preliminary findings.," in *ISMIR*. Citeseer, 2004, vol. 2004, p. 5th.

[6] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull, "Music emotion recognition: A state of the art review," in *Proc. ISMIR*. Citeseer, 2010, pp. 255–266.

[7] Mohammad Soleymani, Anna Aljanaki, Yi-Hsuan Yang, Michael N Caro, Florian Eyben, Konstantin Markov, Björn W Schuller, Remco Veltkamp, Felix Weninger, and Frans Wiering, "Emotional analysis of music: A comparison of methods," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 1161–1164.

[8] Yi-Hsuan Yang and Homer H Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 40, 2012.

[9] Kerstin Bischoff, Claudiu S Firan, Raluca Paiu, Wolfgang Nejdl, Cyril Laurier, and Mohamed Sordo, "Music mood and theme classification-a hybrid approach.," in *ISMIR*, 2009, pp. 657–662.

[10] Mathieu Barthet, Gyorgy Fazekas, and Mark Sandler, "Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models," in *CMMIR*, 2012, pp. 492–507.

[11] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen, "A regression approach to music emotion recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 448–457, 2008.

[12] Erik M Schmidt and Youngmoo E Kim, "Modeling musical emotion dynamics with conditional random fields.," in *ISMIR*, 2011, pp. 777–782.

[13] JA Speck, EM Schmidt, BG Morton, and YE Kim, "A comparative study of collaborative vs. traditional annotation methods," *ISMIR, Miami, Florida*, 2011.

[14] Erik M Schmidt, Matthew Prockup, Jeffery Scott, Brian Dolhansky, B Morton, and Youngmoo E Kim, "Relating perceptual and feature space invariances in music emotion recognition," in *9th Int. Symp. Computer Music Modeling and Retrieval, London, UK*, 2012.

[15] Erik M Schmidt and Youngmoo E Kim, "Learning emotion-based acoustic features with deep belief networks," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. IEEE, 2011, pp. 65–68.

[16] V. Imbrasaite, *Continuous dimensional emotion tracking in music*, Ph.D. thesis, Univ. of Cambridge, UK, Apr. 2015.

[17] Anna Aljanaki, Y.H. Yang, and M. Soleymani, "Emotion in music task at mediaeval 2014," in *Mediaeval 2014 Workshop, Barcelona, Spain, October 16-17*, 2014.

[18] Ming Li and Paul MB Vitányi, *An introduction to Kolmogorov complexity and its applications*, Springer Science & Business Media, 2009.

[19] Rudi Cilibrasi and Paul MB Vitányi, "Clustering by compression," *Information Theory, IEEE Transactions on*, vol. 51, no. 4, pp. 1523–1545, 2005.

[20] Tanaya Guha and Rabab K Ward, "Image similarity using sparse representation and compression distance," *Multimedia, IEEE Transactions on*, vol. 16, no. 4, pp. 980–987, 2014.

[21] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Neural Information Processing Systems*. 2000, pp. 556–562, MIT Press.

[22] Hervé Abdi, "Partial least squares regression (pls-regression)," 2003.

[23] Paul Geladi and Bruce R Kowalski, "Partial least-squares regression: a tutorial," *Analytica chimica acta*, vol. 185, pp. 1–17, 1986.

[24] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[25] Nelson Morgan, Hervé Bourlard, and Hynek Hermansky, "Automatic speech recognition: An auditory perspective," *SPRINGER HANDBOOK OF AUDITORY RESEARCH*, vol. 18, pp. 309–338, 2004.

[26] Rahul Gupta, Naveen Kumar, and S.S. Narayanan, "Affect prediction in music using boosted ensemble of filters," in *Signal Processing Conference (EUSIPCO), 2015 Proceedings of the 23rd European*, 2015.