A GENRE-AFFECT RELATIONSHIP NETWORK WITH TASK-SPECIFIC UNCERTAINTY WEIGHTING FOR RECOGNIZING INDUCED EMOTION IN MUSIC

Wei-Hao Chang, Jeng-Lin Li, Yun-Shao Lin, Chi-Chun Lee

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan.
davidchang83110@gmail.com, cclee@ee.nthu.edu.tw

ABSTRACT

Emotion is a core fundamental attribute of humans. Using music to induce emotional responses from subjects to better facilitate human behavior shaping have been effective across domains of health, education, and retail. Computationally model the musically-induced emotion provides necessary content-based analytics for large-scale and wide-applicability of such human-centered applications. In this work, we propose a relationship neural network architecture to learn to regress the induced emotion attributes with an auxiliary task of genre classification. Our proposed *Genre-Affect Relationship Network* with homoscedastic uncertainty weighting embeds the relationship between affect and genre as tensor normal prior within task-specific layers; the architecture is optimized further by incorporating task-specific uncertainty. The proposed architecture achieves a state-of-art 0.564 average Pearson correlation computed over nine induced emotion ratings in the Emotify database. Furthermore, we provide an analysis to understand the relationship between the induced emotions of these musical pieces and their associated genres.

***Index Terms*—** multi-task learning, induced emotion, music, homoscedastic uncertainty, relationship network

1. INTRODUCTION

Emotion drives our decision making and further motivates our behaviors and actions [1, 2]. The relationship between music, i.e., a cultural activity using sound as the medium, and emotion, i.e., a fundamental human internal attribute, has been extensively studied by music psychologists - demonstrating that music is indeed capable of evoking human's internal affect states [3, 4] (exemplary illustration is shown in Figure 1). In fact, this effect has already found its use cases in human-centered applications, e.g., in fields of health, marketing, and education. In musical therapy, appropriate exposure to music, which induces positive emotion, has been shown to achieve therapeutic effect in addressing issues of physical, emotional, cognitive, and social for those individuals in need [5]. Marketing researchers have further shown that the environmental background music plays a significant role in triggering purchase intentions [6]. Finally, learning efficiency can also be improved with proper musical exposure in order to relax students in stressful situations [7].

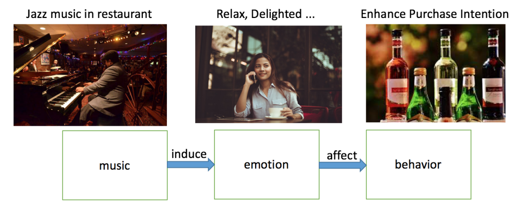

Fig. 1. *A causal chain from music, emotion, to behavior.*

This vast opportunity in using music across a wide range of human-centered applications further emphasizes the importance of developing computational methods to recognize the induced emotion with content-based music analytics. The Geneva Emotional Music Scale (GEMS) [8], i.e., numerical emotion attributes profile for music, is shown to offer a better approximation to humans perception in representing the induced emotion in music as compared to the general valence-activation plane [9, 10, 11]. Several past works have demonstrated that GEMS scales can be automatically computed using features in music. For example, Aljanaki et al. conducted a comprehensive study comparing correlations obtained using different low-level feature sets [12]. Jakubik et al. experimented with different sparse coding approaches to derive feature representations [13] and has recently extended the feature learning approach using a gated recurrent neural network (GRU) to improve accuracies further [14].

Studies have shown that the types of musically-induced emotion felt depend also on multiple other meta factors, e.g., genre of the music, age and gender of the listener, etc [15, 16, 17]. Most of the existing works focus on regressing each of the GEMS' emotion attribute in isolation without considering other joint factors nor the correlated structures of GEMS multi-attribute ratings. Some but limited works has attempted to incorporate genre as a separate module in their emotion recognition algorithm [18, 19, 20]. In this work, we propose a novel *Genre-Affect Relationship Network* to au-

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

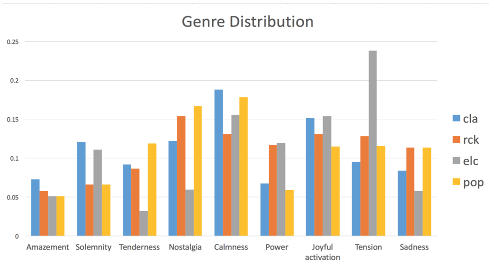

Fig. 2. *Distribution of genre for each GEM emotion attribute*

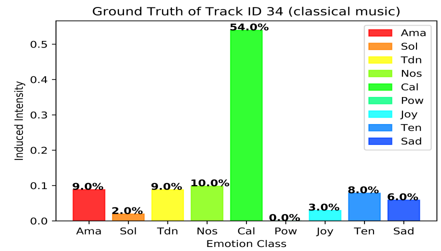

Fig. 3. *Exemplary nine GEMS emotion attribute for a track*

tomatically learn to regress the nine GEMS scales together by leveraging its multi-attribute annotations and the musical genre jointly. The proposed *Genre-Affect Relationship Network* introduces the use of tensor normal prior within a multi-task deep learning framework to model the complex relationship between features, classes, and labels with an additional homoscedastic weighting parameter applied to the main task (emotion regression) and the auxiliary task (genre detection).

The regression experiments are carried out on the Emotify database [8]. We obtain a Pearson correlation of 0.564 (average of nine emotions), which is an absolute improvement of 0.149 over single-task learning. Comparing to other existing state-of-the-art algorithms on the same task, our accuracies is 5.43% higher than the current best result [14]. Finally, we provide an analysis on the relationship between genre and emotion by examining the shared representation.

The rest of the paper is organized as follows: Section 2 will introduce our framework along with the database. Section 3 will summarize our experimental results and discussions. Section 4 is conclusion and future work.

2. RESEARCH METHODOLOGY

2.1. Database

We use the Emotify music database in this paper [8]. Emotify is a well known database collected to study the musically-induced emotion. The annotators are asked explicitly to answer separate questions about the emotions felt and the emotions perceived in music. The exact annotation scheme is based on GEMS, where the database collects over 8407 comments on 400 tracks of music from 138 unique artists. The 400 musical pieces are gathered from four distinct genre classes (classical, rock, electronic, and pop).

In this work, the nine emotions from the middle level of GEMS hierarchy are used. These middle-level categories are: amazement, solemnity, tenderness, nostalgia, calmness, power, joyful activation, tension, and sadness. The final label of a song is a nine-dimensional vector with each dimension has value ranging from 0 to 1 (indicating the relative strength of that particular emotion), and the sum of the nine dimensions equals to 1. An example of the ground-truth label for track ID34 is shown in Figure 3. A plot on the percentages of each of the four different genres against the most prominent emotion of each song is shown in Figure 2.

2.2. Audio Feature Extraction

We extract a total of 6552 audio features for each track using the opensmile toolbox [21]. This particular set of features includes a combination of low-level and supra-segmental features (e.g., chroma features, MFCCs, and energy, etc.) with a variety of statistical functionals applied to them (e.g., mean, standard deviation, inter-quartile range, skewness, kurtosis etc.). It has been shown to provide good performances in tasks of music emotion recognition [12].

2.3. Genre-Affect Relationship Network

The complete architecture of our *Genre-Affect Relationship Network* (GARN) is shown in Figure 4. Overall the architecture is based on a multi-task optimization approach using shared fully-connected layers with task-specific layers. We additionally incorporate homoscedastic uncertainty weighting and tensor normal prior in the task-specific layers to model the relationship between genre and affect. We will briefly describe each of the components below.

2.3.1. Multi-task Deep Neural Network

The GARN is based on multi-task learning structure, where the main task is the nine-attributes emotion regression and the auxiliary task is the four-class genre classification. We learn a shared representation between these two tasks using fully-connected layers (node size: 6552, 4000, 1000) then task-specific layers (node size: 1000, 400), and a final output layer (node size 4 for genre and size 9 for emotion).

The 4000-node hidden layer used in the fully-connected layer is learned first using a sparse autoencoder [22] with the following loss function, L'_{Sparse}:

$$L'_{\text{Sparse}} = L_{\text{Auto}} + \beta \sum_i (\rho \log \frac{\rho}{\hat{\rho}_i}) + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_i} \quad (1)$$

where L_{Auto} is the standard mean square error reconstruction loss, ρ is a parameter indicating the desired average activations in the hidden layer, and $\hat{\rho}_i$ is the average activation of i-th neuron in that layer over the whole dataset. This autoencoder encourages neuron activations in hidden layer to be sparse, and β controls the amount of sparsity ($\beta = 10$ and $\rho = 0.05$ are used in this work).

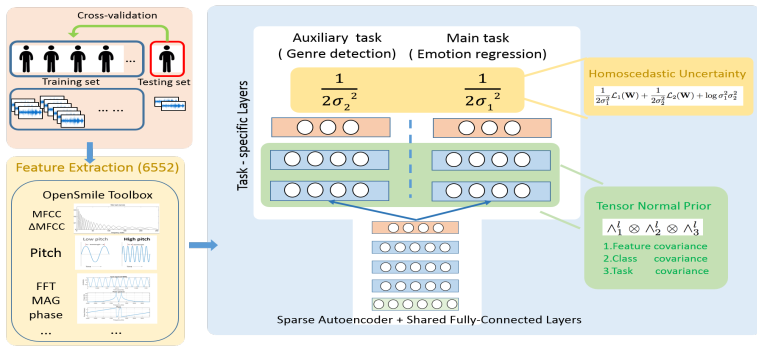

Fig. 4. *A schematic of our proposed Genre-Affect Relationship Network architecture*

Then the fully-connected and task-specific layers are then learned using the following loss function:

$$\lambda^r \sum_{i=1}^{N} \| y_i^r - f(x_i, W^r) \|^2 - \sum_{i=1}^{N} \sum_{a \in A} \lambda^a y_i^a log(p(y_i^a) \mid x_i, W^a) \quad (2)$$

where the first term is the mean square error loss with respect to the nine GEM emotion regression (y_i^r indicates the nine emotion attribute vector for sample i, x_i is the feature inputs, W^r are weights used to regress emotion, N is the number of samples). The second term corresponds to the cross entropy loss used for the four genre class detection (y_i^a indicates one hot encoding of four genre classes for sample i, W^a are weights used to detect genre types). λ^* controls weighted contribution between the main vs. the auxiliary task in this multi-task formulation ($\lambda^r + \lambda^a = 1$).

The choice of λ is mathematically equivalent to controlling the homoscedastic uncertainty in this multi-task prediction problem [23]. The homoscedastic uncertainty in the multi-task predicted outputs can be written as the following overall loss function:

$$L(W, \sigma_1, \sigma_2) = \frac{1}{2\sigma_1^2} L_1(W) + \frac{1}{2\sigma_2^2} L_2(W) + log\sigma_1^2\sigma_2^2 \quad (3)$$

where L_1, L_2 are the losses with respect to the main and the auxiliary tasks. σ_1^2, σ_2^2 are the homoscedastic uncertainty ($\sigma_*^2 = \frac{1}{2\lambda_*}$). By setting a larger λ for our main task, we effectively optimize the learning to reduce the uncertainty in our prediction for the main task as compared to the auxiliary task.

2.3.2. *Relationship Embedding using Tensor Normal Prior*

In order to learn the task relationship in our network parameters for all T tasks, we apply the tensor normal prior in the task-specific layers within our multi-task learning structure [24]. We construct the l-th layer parameter tensor as $W = [W^{1,l}; ...; W^{T,l}] \in R^{D_1^l \times D_2^l \times T}$. The set of parameter tensors of each task-specific layers is indicated as W^l where $W = \{W^l : l \in L\}$. The Maximum a-Posteriori (MAP) estimation of network parameters W given training data X and label Y is the following:

$$p(W|X,Y) \propto p(W) \cdot p(Y|X,W)$$

$$= \prod_{l}^{L} p(W^l) \cdot \prod_{t=1}^{T} \prod_{n=1}^{N_t} p(y_n^t | x_n^t, W^l) \quad (4)$$

the prior, $p(W)$, of the parameter tensor for each layer l is independent of the other layers, i.e., $l_i \neq l_j$. The network parameters when sampled from the prior, tasks become independent of each other. These independent assumptions lead to the factorization of the posteriori in Equation 4.

The prior part $p(w)$ in Equation 4 is the key to enable relationship learning due to its ability to model the relationship across parameter tensors. We assume that the prior distribution for the l-th layer parameter tensor in the task-specific layer as *tensor normal distribution*:

$$p(w^l) = TN_{D_1^l \times D_2^l \times T}(O, \wedge_1^l, \wedge_2^l, \wedge_3^l) \quad (5)$$

where $\wedge_1^l \in R^{D_1^l \times D_1^l}$, $\wedge_2^l \in R^{D_2^l \times D_2^l}$, and $\wedge_3^l \in R^{T \times T}$ are feature covariance, class covariance, task covariance, respectively. By incorporating tensor normal prior into the prior leading to the final MAP estimation of the network, in which the parameters W can be rewritten as a regularized optimization problem with additional terms add to Equation 2:

Equation 2 +

$$\frac{1}{2} \sum_{l \in L} (vec(W^l)^T (\wedge_{1:K}^l)^{-1} vec(W^l) - \sum_{k=1}^{K} \frac{D^l}{D_k^l} ln(| \wedge_K^l |) \quad (6)$$

where $D^l = \prod_{k=1}^{K} D_k^l$ and $K = 3$ is the number of modes in parameter tensor W; $\wedge_{1:3}^l = \wedge_1^l \otimes \wedge_2^l \otimes \wedge_3^l$ is the Kronecker product of the feature covariance of layer l, $\wedge_1^l$, class covariance $\wedge_2^l$,and task covariance $\wedge_3^l$. Equation 6 is the complete loss function used in our proposed *Genre-Affect Relationship Network (GARN)* for recognizing emotion in music.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

We conduct our musically-induced emotion regression task on the Emotify database. The evaluation scheme is based on

Table 1. *A summary of the accuracies (Pearson's r) of our proposed Genre-Affect Relationship Network (GARN) with baseline models and state-of-the-art methods on Emotify dataset*

Method	**STL**	**STL_gnr**	**MTL_1+1**	**GARN**	**MP+harm**	**Autoencoder**	**SEGru+SVR**
Amazement	0.15	0.23	0.26	**0.32**	0.16	0.29	0.29
Solemnity	0.36	0.39	0.52	**0.57**	0.43	0.50	0.53
Tenderness	0.47	0.50	0.54	**0.61**	0.57	0.54	0.53
Nostalgia	0.47	0.50	0.55	**0.60**	0.45	0.50	0.54
Calmness	0.49	0.51	0.59	**0.61**	0.60	0.56	0.56
Power	0.51	0.49	0.59	**0.62**	0.56	0.53	0.56
Joyful activation	0.55	0.56	0.57	**0.62**	0.66	0.53	0.66
Tension	0.37	0.46	0.56	**0.60**	0.46	0.48	0.50
Sadness	0.36	0.38	0.52	**0.53**	0.42	0.33	0.42
Average	0.414	0.447	0.522	**0.564**	0.478	0.473	0.510

leave-one-artist-out cross validation. Each of the validation set includes the amount of songs of each artists. The accuracy measure used is the average of Pearson correlation computed for each of the nine emotion labels. We also construct the following baselines as comparison (a graphical illustration of the three additional baseline structures is also shown in Figure 5).

- **Single-task Learning (STL)**: Each emotion attribute is regressed using a separate neural network (i.e., nine separate models). The core STL network structure is the same as our proposed model without the auxiliary task that produces a single regression output.
- **STL with Auxiliary Task of Genre: STL_gnr**: Each emotion attribute is regressed using a separate neural network. The network structure is the same as STL with an auxiliary task-specific layer responsible for 4-class genre classification.
- **Multi-task Learning 1+1 (MTL_1+1)**: This is the multi-task learning structure (jointly regressed on nine labels together) without applying the homoscedastic uncertainty weighting and tensor normal prior.

3.2. Experimental Results

Table 1 summarizes our results. The overall best performance obtained is our proposed GARN, which achieves an average Pearson correlation of 0.564, and by comparing GARN to MTL_1+1, GARN improves 4.22% on average. This indicates the importance of relationship learning in the task-layers using tensor normal prior and task-specific uncertainty weighting between affect and genre. Furthermore, our experiments show that simultaneously regressing on nine attributes help leverage the correlated structures between the emotion annotations; for example by comparing MTL_1+1 to STL_gnr, MTL_1+1 outperforms single-task based STL_gnr by 11.7% on average. This demonstrates the efficacy in learning multiple emotion attributes by exploiting their correlated structures. The improvement in the use of genre as an auxiliary task is also evident by simply comparing STL with STL_gnr, where STL_gnr improves the results by 3.3%.

3.2.1. Comparing to Other Methods

We further compare our proposed *Genre-Affect Relationship Network* with various other methods recently proposed on the same regression tasks for the Emotify database. In specifics, we compare with three different methods: MP + harm, Autoencoder, SEGru+SVR. MP + harm is a recent work on regressing the same nine attributes by combining a comprehensive set of low-level features from several state-of-the-art audio feature extraction toolbox with a new set of harmonically motivated features [12]. Autoencoder is another work on comparing five sparse coding methods and demonstrates that sparse autoencoder achieves the best performance for the same regression task [13]. SEGru+SVR introduces a new feature representation learning approach based on semantic embedding using GRU, and they demonstrate the current state-of-art performance by using SEGru-features with support vector regression (SVR) [14]. In each of these works, we present their best correlation obtained in Table 1.

Comparing to these recent works on the same nine emotion attributes regression task, our proposed method outperforms all of these methods testing on the same dataset (i.e., improvement of 8.56%, 9.11%, 5.44% comared to MP+harm, Autoencoder, and SEGru+SVR, respectively). Overall, the only class of emotion that GARN is lower than those of other methods is "joyful activation", which is mostly distinctively different all other more complexly-intertwined labels. By leveraging relationship learning between affect and genre and jointly learn to regress and classify the nine emotion attributes and the four genre classes, our proposed GARN architecture can better handle the complex emotion attributes and, hence, achieves the current better performances over the previously proposed methods.

3.2.2. Analysis of Genre and Affect using GARN

In order to further visualize the relationship between genre and affect, we extract the representation of our GARN at

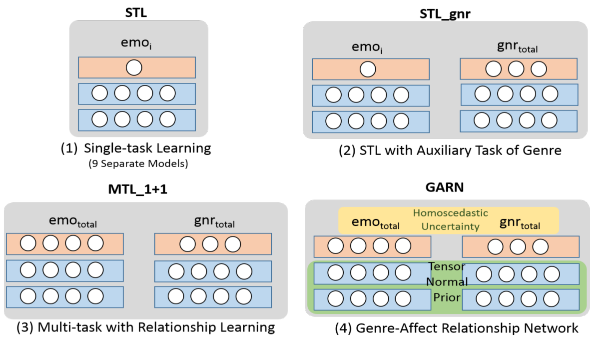

Fig. 5. *(1) STL: single-task learning with nine separate emotion models (2) STL_gnr: STL with auxiliary task of genre detection (3) MTL_1+1: multi-task learning structure without embedding tensor normal prior and uncertainty weighting (4) Genre-Affect Relationship Network: our proposed model*

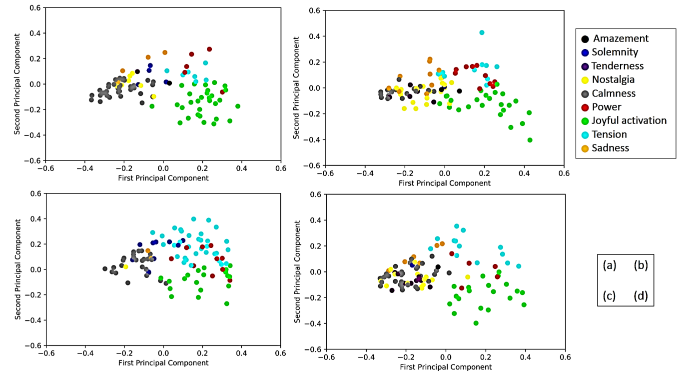

Fig. 6. *This figure shows the 2-D projection (PCA) of each sample's representation derived from GARN with respect to the 4 genre classes: (a) is classical music, (b) is rock music, (c) is electronic music, (d) is pop music, respectively.*

the last shared layer before the task-specific layer. We visually represent each music's representation by projecting onto a 2-dimensional plane using Principal Component Analysis (PCA). Figure 6 shows a projection result for each of the four genre classes, where the color indicates the most prominent emotion class for that music sample.

There are several interesting observations to be noted. For example, Figure 6-a shows a spread of emotion for classical music. Classical music is a genre that has been developed for hundreds of years. Every era present its own unique characteristics, e.g., romantic period pays attention to personal emotion expression, which often leads to the induced emotion of "nostalgia" or "sadness" (most known in the compositions of Chopin). Due to the longer history in its genre development, we also observe a relatively richer and complete set of emotion in classical music. As for rock music, this genre has been known to be good at inducing "joyful activation". This effect is also evident in Figure 6-b. As for electronic music, more samples concentrate around the top-right region, indicating a "tension" emotion. Lastly, "tenderness" and "sadness" are depicted more in the pop music genre.

Another thing to note is from Figure 6 (especially in pop music genre). The samples with most prominent emotion classes of "tenderness", "nostalgia", and "calmness" are heavily clustered with each other. These are also the classes where simple multi-task learning (MTL_1+1) does not outperform method such as SEGru+SVR (Table 1). The relationship learning utilized in our proposed GARN model the subtly connected relationship between these classes providing additional discriminability in the feature spaces - achieving about 10% higher correlation than previous methods in these highly-overlapping emotion attributes.

4. CONCLUSION

Emotion can be induced properly with music and translates into effect of human behavior shaping. Computationally represent the induced emotion characteristics in music provides key enabler to achieve a wider-applicability of using music as medium in a variety of human-centered applications. Due the complex manifestation of music as driven by its genre and intended expression, we utilize a relationship learning mechanism to propose a *Genre-Affect Relationship Network (GARN)* to learn to regress the nine GEMS music emotion scales. In specifics, GARN is a multi-task learning architecture with shared representation and task-specific layers, which we embed tensor normal prior and homoscedastic uncertainty weighting between tasks of emotion and genre. We experiment our framework on the Emotify database reaching an average correlation of 0.564. The framework significantly outperforms other recent works on the same tasks. Furthermore, we demonstrate differences in the induced emotion between the four different music genres.

As for future work, firstly, since the complexity in inducing a particular human emotional states varies between different GEMS emotion attributes [13], these dynamics in their complexity may be better captured by introducing a varied-length network structure in the task-specific layers. Secondly, we plan to expand the diversity and the scale of the database by using this model as a seed to include a wider range of factors, e.g., listeners profiles, languages and cultural backgrounds of the music, expended list of genres, etc. Finally, our goal is understand the mechanistic pathway from musical exposure, emotional responses, to behavior changes leading to a better design of human-centered application [25] and advancing knowledge in mental health and wellbeings [26].

5. REFERENCES

[1] David S. Chester. C. Nathan DeWall, Roy F. Baumeister, "How often does currently felt emotion predict social behavior and judgment? a meta-analytic test of two theories.," *SAGE journals*, March 2015.

[2] C. Nathan DeWall. Roy F. Baumeister, Kathleen D. Vohs, “How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation.,” *SAGE journals*, May 2007.

[3] William Forde Thompson, “Music, thought, and feeling : understanding the psychology of music,” chapter CH.7. New York : Oxford University Press, 2014.

[4] E. Glenn Schellenberg Swathi Swaminathan, “Current emotion research in music psychology,” *SAGE journals*, 2015.

[5] Michael J. Silverman, “Music therapy in mental health for illness management and recovery,” New York : Oxford University Press, 2015.

[6] Gunter Kreutz Raymond MacDonald and Laura Mitchell, “Music, health, and wellbeing,” New York : Oxford University Press, 2013.

[7] Yim-Chi Ho Agnes S. Chan and Mei-Chun Cheung, “Music training improves verbal memory,” *nature journals*, 1998.

[8] D. Grandjean M. Zentner and K. R. Scherer, “Emotions evoked by the sound of music: characterization, classification, and measurement,” *American Psychological Association*, 2008, Emotion, Vol. 8, No. 4, pp. 494521.

[9] Felicia Rodica Balteş, Julia Avram, Mircea Miclea, and Andrei C Miu, “Emotions induced by operatic music: Psychophysiological effects of music, plot, and acting: A scientists tribute to maria callas,” *Brain and cognition*, vol. 76, no. 1, pp. 146–157, 2011.

[10] Wiebke Trost, Thomas Ethofer, Marcel Zentner, and Patrik Vuilleumier, “Mapping aesthetic musical emotions in the brain,” *Cerebral Cortex*, vol. 22, no. 12, pp. 2769–2783, 2011.

[11] Jonna K Vuoskoski and Tuomas Eerola, “Domain-specific or not? the applicability of different emotion models in the assessment of music-induced emotions,” in *Proceedings of the 10th international conference on music perception and cognition*, 2010, pp. 196–199.

[12] Frans Wiering Anna Aljanaki, “Computational modeling of induced emotion using gems,” *ISMIR*, 2014.

[13] Halina Kwas nicka Jan Jakubik, “Sparse coding methods for music induced emotion recognition,” *Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2016.

[14] Halina Kwas nicka Jan Jakubik, “Music emotion analysis using semantic embedding recurrent neural networks,” *INnovations in Intelligent SysTems and Applications (INISTA)*, 2017.

[15] Kathrin Koch, Katharina Pauly, Thilo Kellermann, Nina Y Seiferth, Martina Reske, Volker Backes, Tony Stöcker, N Jon Shah, Katrin Amunts, Tilo Kircher, et al., “Gender differences in the cognitive control of emotion: An fmri study,” *Neuropsychologia*, vol. 45, no. 12, pp. 2744–2754, 2007.

[16] Geoff Luck, Suvi Saarikallio, Birgitta Burger, Marc R Thompson, and Petri Toiviainen, “Effects of the big five and musical genre on music-induced movement,” *Journal of Research in Personality*, vol. 44, no. 6, pp. 714–720, 2010.

[17] Claudio Robazza, Cristina Macaluso, and Valentina D’Urso, “Emotional reactions to music by gender, age, and expertise,” *Perceptual and Motor skills*, vol. 79, no. 2, pp. 939–944, 1994.

[18] Tomoko Matsui Konstantin Markov, “Music genre and emotion recognition using gaussian processes,” *IEEE*, 2014.

[19] Homer H. Chen I-Bin Liao Yeh-Chin Ho Yu-Ching Lin, Yi-Hsuan Yang, “Exploiting genre for music emotion classification,” *Multimedia and Expo, ICME.*, 2009.

[20] Bob L. Sturm, “Evaluating music emotion recognition: Lessons from music genre recognition?,” *Multimedia and Expo Workshops (ICMEW)*, 2013.

[21] Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer, “On the acoustics of emotion in audio: what speech, music, and sound have in common,” *Frontiers in psychology*, vol. 4, 2013.

[22] Y. Bengio, “Foundations and trends in machine learning,” in *Learning Deep Architectures for AI*, 2009, vol. 2, no. 1.

[23] Alex Kendall, Yarin Gal, and Roberto Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” *arXiv preprint arXiv:1705.07115*, 2017.

[24] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and S Yu Philip, “Learning multiple tasks with multilinear relationship networks,” in *Advances in Neural Information Processing Systems*, 2017, pp. 1593–1602.

[25] Shrikanth Narayanan and Panayiotis G Georgiou, “Behavioral signal processing: Deriving human behavioral informatics from speech and language,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[26] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan, “Signal processing and machine learning for mental health research and clinical applications [perspectives],” *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 196–195, 2017.