THE ROLE OF ANNOTATION FUSION METHODS IN THE STUDY OF HUMAN-REPORTED EMOTION EXPERIENCE DURING MUSIC LISTENING

Timothy Greer[§] *Karel Mundnich*[§] *Matthew Sachs*[†*] *Shrikanth Narayanan*[§]

[§] Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA 90089, USA
[†] Center for Science and Society, Columbia University, New York City 10025, New York, USA

ABSTRACT

Music is a universally-enjoyed art form, but listeners often respond to it in tremendously different ways. The same song can bring one person great joy and another deep sorrow. This paper focuses on modeling human music experience at the group level. In this scenario, human annotations serve an important role in computational modeling, especially where the target constructs under study are hidden, such as dimensions of emotion or enjoyment to music listening. In this work, we investigate several ways to represent aggregate human annotations of the complex, subjective emotional experience of listening to music. We show the utility of several methods for fusing self-reported emotion and enjoyment ratings by predicting these responses with auditory features. Using traditional methods such as time alignment with simple averaging and Dynamic Time Warping, as well as state-of-the-art methods based on Expectation Maximization and Triplet Embeddings, we show that it is possible to accurately represent hidden constructs in time under noisy sampling conditions, evidenced by better performance on behavioral response predictions. That subjective responses to complex musical stimuli can be accurately captured using these methods suggests more general applications to research in areas such as affective computing and music perception.

Index Terms— Music perception, music emotion recognition, annotation fusion, Triplet Embeddings, inter-rater agreement

1. INTRODUCTION

Music emotion recognition tasks are of great interest to academic and industrial researchers, due to applications in fields as disparate as music therapy [1] and music recommender systems [2]. The relationship between music and measures of experience, however, has several components (both objective and subjective), which are related to the theory of music, physiological responses of listeners, and the accuracy with which listeners are able to express their experiences.

In previous work, we computationally analyzed the link between how people report happiness, sadness, and enjoyment during music listening as part of an effort to form a more complete picture of music's role in human affective experience [3, 4]. We applied a set of multivariate time series prediction models to subjective responses and used auditory features as predictors, analyzing which elements of the musical stimuli were important for predicting emotion and enjoyment. However, in these studies (and many studies, in fact), the effect of the "ground truth" labels on the task was not studied [5].

A key aspect in the field of supervised learning methods is the need of accurate labels that appropriately represent the construct that one wants to study. In our previous work, we averaged the labels collected from music listeners in real time to create a single label, providing a basic understanding of group-level models of experience. If deeply complex constructs like emotion and enjoyment were more faithfully and/or accurately represented, perhaps avenues would open up for better quantification, analysis, and explanation of results about how music listening affects human experience. Following these ideas, several methods for merging the labels of multiple annotators (usually called *annotation fusion* in the affective computing literature) have been proposed, which aim to create aggregate labels in ways that are more robust to noise and artifacts than simple averaging. These methods are usually designed and employed to estimate a ground truth variable for a subjective construct, such as a dimension of affect that a person is conveying through audio. In this paper, we study the effect of different annotation fusion methods in the studying of two subjective group-level emotional responses to music.

1.1. Related Work

Several widely-used music emotion recognition (MER) tasks involve analyzing reported human responses to music listening [6, 7, 8]. In order to draw meaningful, accurate conclusions using annotations from these datasets, it is necessary to understand and represent the latent states of the annotators. However, as far as the authors know, there has not been research that investigates the role of annotation fusion methods for group-level experiences *during music listening*, such as emotion and enjoyment. In this paper, we explore several annotation fusion methods, using them to generate a single label when subjects have annotated the same target (in this case, a musical stimulus). This label can be a proxy for group-level models of behavioral music experience.

Many researchers have proposed algorithms for fusing annotations to generate a set of labels for use as ground truth in machine learning. A general method is averaging individual annotations after first performing time alignment to remove time lags produced by latencies in annotation that vary from person to person. For example, [9] demonstrates an improvement in regression performance by aligning annotations in time with a uniform shift computed per annotator based on mutual information, then averaging. Dynamic Time Warping (DTW) [10] is another popular time alignment method which warps time to maximize alignment between features and annotations [11]. Some methods, like canonical correlation analysis [12] and correlated spaces regression [13], learn to warp the fused annotation space so the resultant fusion is more correlated with its associated features. Many combined time and space warping methods have also been proposed: canonical time warping [14], gen-

[*] Performed the work while at the Brain and Creativity Institute, University of Southern California.

eralized time warping [15], and deep canonical time warping [16] are some examples. More recent work explores other strategies for computing meaningful annotation fusions. Lopes et al. [17] show that in some cases the gradient in an annotation is more informative than the annotation value, which can be exploited to produce a better ground truth based on sudden changes in annotations. Booth et al. [18] hypothesize that trends in the annotation values contain more meaning than the values themselves; they propose a fusion approach using comparative information collected from humans to produce labels. These methods have been extended to cases where only continuous annotations are present [19] or when an annotation scheme can be discretized [20].

These algorithms approach annotation fusion from many different ways, but the accuracy of any fused annotation fundamentally cannot be measured directly for effectiveness or quality when the underlying construct is a latent mental or behavioral state.

1.2. Contributions

In this paper, we use four methods for annotation fusion—EvalDep followed by simple averaging, DTW followed by simple averaging, EvalDep followed by Expectation Maximization, and EvalDep followed by Triplet Embeddings—to investigate the utility of fused annotations of real-time emotion and enjoyment reports to music. Some anecdotal evidence in [18] suggests that consistency may not be preserved during continuous real-time annotation, but we aim to show that making this simplifying assumption still produces quality fusions for behavioral responses to music listening. Our findings suggest that certain fusion techniques can improve prediction in music emotion recognition, thereby implying its broader utility in annotation tasks where a latent construct must be estimated.

2. ANNOTATION FUSION METHODS

We use several annotation fusion techniques in this paper. These techniques usually contain two key steps: (1) time-aligning the annotations acquired in real-time to account for the reaction lags introduced in the annotation process, and (2) combining the time-aligned annotations to generate a single annotation that faithfully represents all annotators.

2.1. Time Alignment

Annotations that are collected in real-time usually have reaction-time lags with respect to the features being annotated. Therefore, time alignment of these annotations with a set of features is required. We now give a brief review of each one of these methods.

Dynamic time warping (DTW) [10] has been traditionally employed to perform time alignment. DTW aligns two signals to one another by matching similar points in each series, thereby minimizing the distance between the resultant warped signals. In the case of multiple annotation signals, a feature correlated with the responses is often used to warp each annotation.

A different method proposed by Mariooryad et al. called EvalDep [9] finds the lag by maximizing the mutual information between a feature and the annotations. The estimation is done either non-parametrically by using kernels to estimate the density functions, or by assuming Gaussianity of the features and labels, for which the the mutual information is defined by:

$$I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2}\log\left(\frac{\det(\boldsymbol{\Sigma}_{\boldsymbol{XX}})\det(\boldsymbol{\Sigma}_{\boldsymbol{YY}})}{\det(\boldsymbol{\Sigma})}\right),\ \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{XX}}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\\ \boldsymbol{\Sigma}_{\boldsymbol{YX}}\boldsymbol{\Sigma}_{\boldsymbol{YY}}\end{bmatrix}, \quad (1)$$

where $\boldsymbol{X}$ is a feature, $\boldsymbol{Y}$ is a label, and $\boldsymbol{\Sigma}_{\cdot\cdot}$ are the covariance and cross-covariance matrices. Then, the estimated lag is given by:

$$\hat{\tau} = \operatorname*{argmax}_{\tau} I(\boldsymbol{Y};\boldsymbol{X}). \quad (2)$$

This lag is applied to $\boldsymbol{Y}$ to obtain the time-aligned annotations.

To choose the features to which the annotations are aligned, a common empirical approach is to compute the correlation ρ between the annotations and the features, and time-align the annotations with respect to the feature with highest correlation. We employ this method in Section 4.

2.2. Fusion of annotations

Different approaches have been proposed in the literature to combine annotations from different annotators. We now give a brief review of each one of these methods.

2.2.1. An approach based on simple averaging

The simplest and most common approach to annotation fusion is to assume that for a given time step, all annotations have the same distribution. Under this assumption, for each time step, the average of annotations is taken to obtain a single fused annotation. This approach was used in previous work from the authors [3, 4].

2.2.2. An approach based on Expectation Maximization

These approaches go a step further and model each annotator individually [21, 22], and assume that each annotation is a distorted version of the ground truth. These works pose the model as graphical model were parameters can be estimated using a maximum likelihood approach, which is in practice solved using expectation maximization.

2.2.3. An approach based on Triplet Embeddings

A different set of approaches uses Triplet Embeddings [18, 19, 20] to generate a fused label. In particular, in this work we use the approach proposed in [19]. The key assumptions here are (1) annotators are better at annotating ordinal relations in time than ratings even when asked to rate, and (2) the ordinal relationships of their ratings across time are kept mostly constant. Therefore, we can answer the following questions from ratings in time:

$$d(\boldsymbol{Y}_i,\boldsymbol{Y}_j) \overset{?}{\lesseqgtr} d(\boldsymbol{Y}_i,\boldsymbol{Y}_k), \quad (3)$$

where i, j and k are time-frames of the annotations $\boldsymbol{Y}$ and $d(\cdot,\cdot)$ is a distance measure. Collecting a set of these comparisons (called triplets) from all annotators and deciding based on majority vote if $\boldsymbol{Y}_i$ is closer to $\boldsymbol{Y}_j$ than $\boldsymbol{Y}_k$ (or the opposite), it is possible to find a 1-dimensional embedding indexed by time that may be used as a fused annotation.

3. DATA COLLECTION

We collected musical stimuli for annotation from social media sites such as Reddit and Twitter, as well as online streaming sites, such as Spotify and Last.fm. We collected songs on these sites with social tags “happy” and its synonyms and “sad” and its synonyms. Collecting social tags provided by users of online music streaming sites is an effective method for classifying music based on their emotional content [23]. To minimize any influence of prior exposure to the songs, we selected a subset of 120 songs with "happy" or "sad" tags and low play counts. Eight human coders listened to 30-second clips from these 120 songs and rated whether the songs conveyed either happiness or sadness. Pieces in which there was at least 75% agreement among these coders on the intended emotion were then included in an online survey that was completed by 82 adult participants via Amazon's Mechanical Turk. The survey, which included 60-second clips from 27 pieces of music, asked participants to rate their enjoyment to the piece, what emotion they felt in response to the piece (calmness, sadness, happiness, boredom, anxiousness), and how familiar they were with the piece.

Due to the potential confounds associated with the semantic information conveyed through the lyrics of a song [24], we then removed all pieces that contained lyrics. We additionally excluded pieces that were rated as highly familiar to prevent bias. Based on these criteria from the survey, we selected three pieces of music to be used for this study: (1) a shorter piece that reliably induces sadness (Ólafur Arnalds's "Fyrsta," the "sad short song," with duration 256 seconds); (2) a longer piece that reliably induces sadness (Michael Kamen's "Discovery of the Camps," the "sad long song," with duration 515 seconds); and (3) a piece that reliably induces happiness (Lullatone's "Race Against The Sunset," the "happy song," with duration 168 seconds). The wav files of these songs were played to participants and used for extracting auditory features.

A group of 60 healthy, right-handed, adult participants was recruited from the greater Los Angeles community based on responses to an online survey in which they listened to a 60-second clip of the final three pieces[1]. All participants in this study were right-handed with normal hearing, normal or corrected-to-normal vision, and no history of neurological or psychiatric disorders.

Participants were then brought into our lab to listen to all three songs twice: during one listen, they were instructed to report changes in their affective experience using a fader with a sliding scale. Participants continuously reported the intensity of felt emotion, from 0 to 10, depending on which piece was being presented. In the sad short song and sad long song, 10 indicated "extremely sad" and 1 indicated "not at all sad." In the happy song, 10 indicated "extremely happy" and 1 indicated "not at all happy."

During the other listen, subjects were instructed to listen attentively to the music and simultaneously report the intensity of their enjoyment of the piece using a fader with a sliding scale. Participants continuously rated their momentary feelings of pleasure from 0 (no pleasure at all) to 10 (extreme pleasure). The order of the pieces and the order of the tasks were counterbalanced across participants.

3.1. Auditory Features

In order to predict response to music, we used auditory features related to dynamics, timbre, harmony, and rhythm, similar to [5]. Dynamics refer to "loudness" and change in "loudness" of music, timbre refers to the "tone color" of the music, harmony refers to the presence of certain musical pitches, and rhythm refers to properties of the musical beat.

Seventy-four features that capture dynamics, timbre, harmony, and rhythm were estimated using Matlab's MIRtoolbox [25] (see Table 1). These features were extracted using a sliding window with a duration of 50 ms and a step size of 25 ms.

The compressibility feature was calculated by computing the ratio between the file size of each window's wav format and that same window's mp3 format, after conversion using ffmpeg [26]. Key strength was computed as the maximum value of the 24-dimensional output vector from MIRtoolbox's `key_strength` function. Mel frequency cepstral coefficients (MFCCs) were calculated with Matlab's `mfcc` function [27] using a Hamming window, pre-emphasis coefficient of .97, frequency range of 100-6400 Hz, 20 filterbank channels, and 22 liftering parameters, similar to [28, 4]. All other features were extracted using MIRtoolbox's eponymous functions and default parameters.

Table 1. Auditory features used and feature type. If there are parentheses after a feature, the number indicates how many features are in that set. If there are no parentheses, the feature is a scalar

Feature Type	Feature (#)
Timbre	MFCCs (13), ΔMFCCs (13), $\Delta\Delta$MFCCs (13), Brightness, HCDF, Skewness, Kurtosis, Spread, LPCs (11), Spectral Flux
Harmony	Centroid, Chroma (12), Key Strength, Key Mode
Dynamics	RMS, Compressibility
Rhythm	Pulse Clarity [29]

4. METHODS

4.1. Parameter Selection for Annotation Fusion

In order to time align the annotations, a time series that is correlated with the annotations must be chosen. To find a time series like this, we averaged the raw annotations and ran a correlation of this signal with each of the 74 musical features. The correlation coefficient ρ between the average annotations and the RMS of the audio signal (computed at 1 Hz with no overlap) was found to be between .29 and .38 for every response: the highest of any auditory feature. We time-aligned each annotation with this feature for the "DTW" representation and the EvalDep representation, and used the latter as input for the "Expectation Maximization" and "Triplet Embedding" representations, as described in Section 2.

EvalDep [30] requires a maximum lag parameter. Previous work has shown that mood classification has a high latency [31], and the additional task of reporting judged emotion adds to that latency, so we decided to use 10 seconds (10 times the sampling rate) for the maximum lag parameter in this method. We then average the ratings at each time step across individuals to create the "EvalDep" response.

Lastly, all response signals were upsampled from 1 Hz to 40 Hz using linear interpolation in order to match the sampling rate of the auditory features.

4.2. Prediction Model

For every experiment, the first 30 seconds of each song were removed, ensuring that responses would be stable. Then, data was split into ten folds, with the first tenth of the song as the first fold, the second tenth of the second tenth of the song as the second fold, and so on. We used wen-fold cross-validation and report cross-validation error.

In [3] and [4], it was found that a distributed lag model using ℓ_1 (which we call LASSO-DL) performed best in predicting responses

[1] This group was 60% female, with an average age of 19.5 and a standard deviation of age of 2.86

we use this model for experiments here[2].

The regularization parameters tried for this regression were 10^i and 5×10^i for $i \in [-7, 7]$ and the parameter with lowest ten-fold cross-validation root mean squared error (RMSE) was chosen for each experiment. Response variables were scaled by their maximum value and features were ℓ_2-normalized.

Additionally, two distributed lag parameters were set: the time horizon and the autoregression length. A model with time horizon h and autoregression length a predicts the response $\hat{y}_t$ at timestep t according to

$$\hat{y}_t = \beta_0 x_{t-h} + \beta_1 x_{t-h-1} + ...\beta_a x_{t-h-a}, \tag{4}$$

where β is a coefficient vector for a particular timestep and x_k is a vector of musical features at timestep k.

For all fused annotations—which were created by warping with RMS—we omit RMS and the 0th MFCC from the feature set. RMS was used in the process of creating the fused annotations, and a model predicting these annotations would inherently be biased.

We assume that emotion stabilization will be shorter than six seconds, as the songs we selected are repetitive and contain more context than the clips used in MER studies. We use $h = 80$ (two seconds), $a = 80$ (two seconds), to capture latency in feeling and subsequently reporting felt emotion.

5. RESULTS

Table 2 shows the cross-validation error for self-reported emotion in the sad short song, sad long song, and happy song. The fusions using Triplet Embeddings result in the lowest RMSE in the happy song and the sad long song, while the fusions made using EM result in the lowest error in the sad short song (lowest RMSEs are bolded).

Table 2 also shows the cross-validation error for self-reported enjoyment in the sad short song, sad long song, and happy song, respectively. The fusions using Triplet Embeddings result in the lowest RMSE in the sad long song and the happy song, while the fusions made using EM again result in the lowest error in the sad short song.

6. DISCUSSION

We see that Triplet Embeddings outperforms the baseline DTW fusions on each of the six responses, and fusions made using either Triplet Embeddings or EM methods result in the lowest RMSEs for every response. These findings suggest that more involved methods of annotation fusion may be of greater use for tasks in MER than the "bare minimum" time alignment method.

While DTW outperforms the EvalDep fusions in emotion tasks, it is outperformed by all other models in the enjoyment tasks. This suggests that although emotion and enjoyment are correlated (at least in this study), the underlying latent states are likely very different. This result also further demonstrates the utility of using Triplet Embeddings for aggregating human annotations of emotion and enjoyment: in both annotation tasks, Triplet Embeddings outperforms the baseline DTW model, even when DTW performs second-best (as in the case of predicting emotion in the sad long song).

In [19], the authors warped the response signal to the highest-correlated feature and found that this method was the best approach. We apply the same approach here, but generalized time warping or deep canonical time warping may prove to be effective methods for these tasks. We will explore these methods in future work.

Table 2. Reported emotion and enjoyment cross-validation RMSE.

Task	**Song**	**DTW**	**EvalDep**	**EM**	**Triplet**
Emotion	Sad Short	.0502	.0590	**.0273**	.0347
	Sad Long	.0378	.0532	.0548	**.0374**
	Happy	.0140	.0208	.0181	**.0014**
Enjoyment	Sad Short	.0562	.0187	**.0157**	.0177
	Sad Long	.0424	.0137	.0321	**.0126**
	Happy	.0240	.0169	.0234	**.0152**

Emotion and enjoyment were best-predicted in the happy song by using fusions from Triplet Embeddings. Indeed, Triplet Embeddings outperformed other models by a wide margin in tasks related to this stimulus (especially in the emotion prediction task). This may be because emotion and enjoyment in happy music may be easier to agree upon by annotators than those same latent states in sad music. That is, reporting judged emotion and enjoyment in happy music may be a more straightforward task than reporting judged enjoyment and emotion in sad music. Triplet Embeddings are constructed to be more reliable with higher annotator agreement, so this result is feasible.

This study uses sixty participants, which is higher than is likely feasible in most annotation tasks. Given that this is the case, we would like to investigate the effectiveness of these different fusion methods given a varying number of annotators in future work.

7. CONCLUSION

Responses to music can vary greatly from person to person and group to group. Aggregating human annotations, therefore, can be a very difficult undertaking, yet it serves a paramount role in studying tasks in music emotion recognition. In this work, we use several methods to fuse annotations of human-reported emotion and enjoyment responses to music listening, and show the utility of using more involved methods for annotation fusion by predicting affective responses using auditory features. This finding suggests that methods for annotation fusion, especially those that use Triplet Embeddings and expectation maximization, may be a boon to research in music emotion recognition.

8. ACKNOWLEDGMENTS

The support for USC SAIL from NSF and NIH, and from Google for its Center for Computational Media Intelligence, is gratefully acknowledged.

9. REFERENCES

[1] Olga Sourina, Yisi Liu, and Minh Khoa Nguyen, "Real-time eeg-based emotion recognition for music therapy," *Journal on Multimodal User Interfaces*, vol. 5, no. 1-2, pp. 27–35, 2012.

[2] Bruce Ferwerda and Markus Schedl, "Enhancing music recommender systems with personality information and emotional states: A proposal.," in *Umap workshops*, 2014.

[3] Timothy Greer, Matthew Sachs, Benjamin Ma, Assal Habibi, and Shrikanth Narayanan, "A multimodal view into music's

[2] For the sake of thoroughness, we also tested a linear regression model that predicted the response at time t using musical features at time t. However, because this model was (again) consistently outperformed by the distributed lag model, it was omitted from our results.

effect on human neural, physiological, and emotional experience," in *ACM Transactions on Multimedia Computing, Communications, and Applications*. ACM, 2019, In press.

[4] Benjamin Ma, Timothy Greer, Matthew Sachs, Jonas Kaplan, and Shri Narayanan, "Predicting human-reported enjoyment responses in happy and sad music," in *International Conference on Affective Computing and intelligent interaction*, 2019, In press.

[5] 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*, 2010, vol. 86, pp. 937–952.

[6] Anna Alajanki, Yi-Hsuan Yang, and Mohammad Soleymani, "Benchmarking music emotion recognition systems," *PLOS ONE*, pp. 835–838, 2016.

[7] Yu-An Chen, Yi-Hsuan Yang, Ju-Chiang Wang, and Homer Chen, "The amg1608 dataset for music emotion recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 693–697.

[8] Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun, "The pmemo dataset for music emotion recognition," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ACM, 2018, pp. 135–142.

[9] Mariooryad Soroosh and Busso Carlos, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. affective computing," *IEEE Transactions on*, vol. 6, no. 2, pp. 97–108, 2015.

[10] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[11] Meinard Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[12] Harold Hotelling, "Relation between two sets of variates," *Biometrica*, 1936.

[13] Mihalis A Nicolaou, Stefanos Zafeiriou, and Maja Pantic, "Correlated-spaces regression for learning continuous emotion dimensions," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 773–776.

[14] Feng Zhou and Fernando Torre, "Canonical time warping for alignment of human behavior," in *Advances in neural information processing systems*, 2009, pp. 2286–2294.

[15] Feng Zhou and Fernando De la Torre, "Generalized time warping for multi-modal alignment of human motion," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1282–1289.

[16] George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou, "Deep canonical time warping for simultaneous alignment and representation learning of sequences," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1128–1138, 2017.

[17] Phil Lopes, Georgios N Yannakakis, and Antonios Liapis, "Ranktrace: Relative and unbounded affect annotation," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 158–163.

[18] Brandon M Booth, Karel Mundnich, and Shrikanth S Narayanan, "A novel method for human bias correction of continuous-time annotations," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3091–3095.

[19] Brandon M Booth, Karel Mundnich, and Shrikanth Narayanan, "Fusing annotations with majority vote triplet embeddings," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 83–89.

[20] Karel Mundnich, Brandon M. Booth, Benjamin Girault, and Shrikanth Narayanan, "Generating Labels for Regression of Subjective Constructs using Triplet Embeddings," *Pattern Recognition Letters*, vol. 128, pp. 385–392, 2019.

[21] Rahul Gupta, Kartik Audhkhasi, Zach Jacokes, Agata Rozga, and Shrikanth Shri Narayanan, "Modeling multiple time series annotations as noisy distortions of the ground truth: An expectation-maximization approach," *IEEE transactions on affective computing*, vol. 9, no. 1, pp. 76–89, 2016.

[22] Anil Ramakrishna, Rahul Gupta, Ruth B Grossman, and Shrikanth S Narayanan, "An expectation maximization approach to joint modeling of multidimensional ratings derived from multiple annotators.," in *Interspeech*, 2016, pp. 1555–1559.

[23] Yading Song, Simon Dixon, and Marcus Pearce, "A survey of music recommendation systems and future perspectives," in *9th International Symposium on Computer Music Modeling and Retrieval*, 2012, vol. 4.

[24] Elvira Brattico, Vinoo Alluri, Brigitte Bogert, Thomas Jacobsen, Nuutti Vartiainen, Sirke Katriina Nieminen, and Mari Tervaniemi, "A functional mri study of happy and sad emotions in music with and without lyrics," *Frontiers in psychology*, vol. 2, pp. 308, 2011.

[25] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola, "A matlab toolbox for music information retrieval," in *Data analysis, machine learning and applications*, pp. 261–268. Springer, 2008.

[26] FFmpeg Developers, "ffmpeg tool," http://ffmpeg.org/, 2019.

[27] "Matlab audio toolbox," 2019, The MathWorks, Natick, MA, USA.

[28] Timothy Greer, Benjamin Ma, Matthew Sachs, Assal Habibi, and Shrikanth Narayanan, "A multimodal view into music's effect on human neural, physiological, and emotional experience," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 167–175.

[29] Olivier Lartillot, Tuomas Eerola, Petri Toiviainen, and Jose Fornari, "Multi-feature modeling of pulse clarity: Design, validation and optimization.," in *ISMIR*. Citeseer, 2008, pp. 521–526.

[30] Soroosh Mariooryad and Carlos Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2014.

[31] Cyril Laurier, Jens Grivolla, and Perfecto Herrera, "Multimodal music mood classification using audio and lyrics," in *2008 Seventh International Conference on Machine Learning and Applications*. IEEE, 2008, pp. 688–693.