WHICH AUDIO FEATURES CAN PREDICT THE DYNAMIC MUSICAL EMOTIONS OF BOTH COMPOSERS AND LISTENERS?

Eun Ji Oh[1]
[1] Center for Digital Humanities and Computational Social Sciences, KAIST, Republic of Korea
muye35@kaist.ac.kr

Hyunjae Kim[2]
[2] Graduate School of Culture Technology, KAIST, Republic of Korea
present@kaist.ac.kr

Kyung Myun Lee[123]
[3] School of Digital Humanities and Computational Social Sciences KAIST, Republic of Korea
kmlee2@kaist.ac.kr

ABSTRACT

Are composers' emotional intentions conveyed to listeners through audio features? In the field of Music Emotion Recognition (MER), recent efforts have been made to predict listeners' time-varying perceived emotions using machine-learning models. However, interpreting these models has been challenging due to their black-box nature. To increase the explainability of models for subjective emotional experiences, we focus on composers' emotional intentions. Our study aims to determine which audio features effectively predict both composers' time-varying emotions and listeners' perceived emotions. Seven composers performed 18 piano improvisations expressing three types of emotions (*joy/happiness*, *sadness*, and *anger*), which were then listened to by 36 participants in a laboratory setting. Both composers and listeners continuously assessed the emotional valence of the music clips on a 9-point scale (1: '*very negative*' to 9: '*very positive*'). Linear mixed-effect models analysis revealed that listeners significantly perceived the composers' intended emotions. Regarding audio features, the RMS was found to modulate the degree to which the listener's perceived emotion resembled the composer's emotion across all emotions. Moreover, the significant audio features that influenced this relationship varied depending on the emotion type. We propose that audio features related to the emotional responses of composers-listeners can be considered key factors in predicting listeners' emotional responses.

1. INTRODUCTION

Music holds the power to convey emotions and evoke strong emotional responses in its listeners. There is a growing interest in utilizing Music Emotion Recognition (MER) systems for personalized music experiences, such as music recommendations, automated music generation, and diverse multimodal experiences. However, identifying the variables that effectively predict listeners' emotional experiences is a challenging problem due to the complexity of its mechanisms [1]. While recent MER studies employ machine learning techniques to predict emotions based on dynamic listener annotations [2, 3], they often lack an interpretation of the underlying factors driving emotions.

This study aims to explain the prediction of musical emotions by empirically investigating the relationship between the composer's intended emotion, the listener's perceived emotion, and various audio features of music through time-series data. We specifically focus on the composer's emotional intentions during the music creation process, prior to listener exposure.

1.1 Background

MER tasks are inherently user-centered [4], bringing researchers from interdisciplinary fields such as musicology, cognitive science, and computer science. A range of factors, including individual traits (*e.g.*, personality, mood regulation strategies, etc.) and musical elements (*e.g.*, timbre, rhythm, harmony, etc.) [1], can impact the MER systems, posing challenges for enhancing model performance. Many MER studies rely on emotion datasets where listeners annotate their perceived or felt emotions [5, 6]. Given this, the outcome of the study can be significantly influenced by the taxonomy used to define emotions and the methods used to identify listeners' annotations [4, 7]. While previous studies have often relied on discrete emotion ratings [8–11], the latest trends favor continuous assessments that capture emotional fluctuations during music listening, reflecting the nature of music experiences [2, 3].

Recent studies on Music Emotion Recognition (MER) face several limitations. First, they often overlook the potential influence of emotions expressed by composers or performers on listeners' emotional experiences. Second, MER models commonly encounter challenges in accurately predicting valence compared to arousal [3].

The emotional intentions of composers/performers can play an important role in predicting listeners' emotions, but their significance is often underestimated. Composers or performers express their emotions through musical features such as tempo, dynamics, and timbre [12–14]. Listeners then perceive these cues and interpret the emotions conveyed by the music. When the emotions perceived by

the listener align closely with those expressed by the composer or performer, it can foster a strong connection between the listener and the composer/performer, which can positively impact listeners' emotional experiences [15]. This connection can also be observed in physiological responses; a previous study [16] has shown that the similarity of brain activity between audiences and violinists can predict the audiences' fondness for the performance.

Taken together, the relationship between listeners' and composers/performers' emotions is highly correlated with the emotional responses that listeners experience from music, such as music engagement and enjoyment. Therefore, we propose that composers' intentions may play an important role in MER systems that seek to predict listeners' emotions. To determine the impact of composers' intentions on predicting listeners' perceived emotions, we aim to compare prediction outcomes using only audio features against those utilizing both audio features and composers' emotional intentions.

Despite the potential importance of this relationship, there is a lack of research exploring the link between listeners and musical intentions. While some studies investigated how accurately emotional intentions were conveyed to listeners through discrete emotion ratings [8, 17, 18], there is a need to investigate the dynamic emotional responses of composer/performer and listener as the music unfolds over time.

1.2 Research Question

To examine the predictive role of audio features and emotional intentions in shaping listeners' perceived emotions, as well as the significance of time-varying emotional data in this context, we set the following research questions:

RQ1. Do the predictors of listeners' perceived emotions (audio features and composer's emotions) vary based on the methodology, discrete vs. dynamic emotional ratings?
RQ2 Which audio features predict the dynamics of the composer/performer's emotional intentions and listener-perceived emotions, respectively?
RQ3 Which audio features reflect both the composer/performer's and listeners' emotions?

To address these questions, we initially recruited composers to create emotionally expressive piano improvisations. We then collected composers' real-time assessments of the emotions they intended to convey during their performances. The emotional valence scale was only used for the assessments to reduce the complexity of predicting emotions. This approach may reduce cognitive overload for lay participants, who might find 2D emotion mapping (*arousal-valence*) unfamiliar.

For listeners' emotional data, we played the composers' music clips and instructed listeners to continuously infer the expressed emotion. Audio features were extracted via the *librosa* library, including root-mean-square (*RMS*), *flatness*, *zero-crossing*, *spectral centroid*, and *roll-off*, chosen based on previous research on audio features and emotions [3, 9, 19].

We employed Linear Mixed-Effects (LME) models for multi-level regression analysis, which are suitable for handling hierarchical, non-independent time-series data. By accounting for variability within and between music clips, we investigated whether listeners effectively captured changes in the composer's intentions, independent of the specific characteristics of individual clips [20].

2. MATERIALS

2.1 Composers' Emotion Data

2.1.1 Participants

We recruited eight composers from various composition departments in the College of Music, Republic of Korea (4 males and 4 females, $M = 26.88$, $SD = 1.73$). These participants were either undergraduate students or recent graduates with a bachelor's degree in Western classical music composition. On average, they had 14.13 years of formal music training ($SD = 5.72$), with an average of 10.25 years of piano experience ($SD = 3.28$). All composers had prior experience in improvised performances.

2.1.2 Music Performance Setting

Composers were instructed to prepare three semi-improvised piano performances, each lasting 1-2 minutes, expressing primary emotions: *joy/happiness*, *sadness*, and *anger*. These emotions were chosen based on prior literature [10, 20–22] for their distinctiveness in conveying or interpreting emotions through music.

Performances took place in a soundproof booth using a Casio Contemporary CDP-120 digital keyboard, with default piano sound and fixed volume settings. Video recordings were made using a Canon EOS 5D Mark IV Full Frame DSLR, capturing audio via the built-in microphone. The recordings were in .mp4 format, with a resolution of 1920 x 1080, 25 fps, and an audio sampling rate of 48 kHz.

2.1.3 Recording Procedure

Composers were briefed about the experiment and provided consent. They had 15 minutes to prepare, followed by a 30-second sample performance for technical setup. The order of recording for the three performances was randomized, with breaks between each to refresh emotions.

After each performance, composers rated their expressed emotions using arousal, valence, and dominance on a 9-point Likert scale (discrete ratings). Following the recording session, they watched videos of themselves in a randomized order, continuously rating the emotions they expressed during the performance on a 9-point valence scale (1: '*very negative*' - 9: '*very positive*') in real-time, mirroring the setup described in Section 3.2.

2.1.4 Music Selection

Twenty-four music clips were initially recorded, featuring performances of three emotions by eight composers. Five authors and colleagues participated in the decision-making process for music selection. The selection criteria ensured

	joy/happiness	sadness	anger
arousal	6.33 (2.16)	3 (1.87)	7.57 (1.27)
valence	8.33 (0.82)	3.6 (0.55)	2.14 (0.69)
dominance	6.17 (1.94)	4.8 (1.64)	7.71 (1.60)

Table 1. The *mean* (*SD*) scores of emotion provided by composers for the final 18 music clips (*n* = 6 for each emotion). They were assessed with a 9-point Likert scale.

that 1) each clip effectively conveyed its intended emotion (*e.g.*, a performance expressing sadness was excluded since some researchers felt it was positive valenced music) and 2) was free from distracting noise (*e.g.*, the sound of fingernails on keyboards). An equal distribution of male and female composers per emotion was maintained resulting in 18 chosen clips (six per emotion) from four males and three females.

All 18 clips were pre-processed using Adobe Premiere Pro, ensuring .wav format, 44.1 kHz sampling rate, 16-bit depth, stereo, and normalization according to ITU BS.1770-3 standards [1]. The mean clip length was 97.5 seconds (*SD* 14.50), ranging from 73 to 125 seconds. The mean scores of discrete emotional ratings provided by composers are shown in Table 1.

2.2 MIR Audio Features

To select the audio features, we reviewed prior research on emotion perception and acoustic features. Studies highlighted the importance of timbre, tempo, mode, harmony, loudness, and pitch in emotional communication [2, 9, 23, 24]. In particular, tonality, pitch, harmony, articulation, and timbre (e.g., brightness, roughness) were crucial for predicting emotion valence [25, 26]. Machine-learning methods have shown that valence emotion prediction models achieve high explanatory power when incorporating spectral [3, 19] and rhythmic features [19] available in the *librosa* package [27]. Based on this, we used *librosa* to extract audio features from 18 music clips, focusing on loudness (root-mean-square; *RMS*), timbre (*flatness*, *zero-crossing*, *spectral centroid*, and *roll-off*), harmony (Mel-Scale Frequency Cepstral Coefficients; *MFCC*, *chroma*, *spectral contrast*), and rhythm (*dynamic tempo*). To compare audio features with 2D data (*time-valence*) of composers' and listeners' emotional ratings, we selected five features: *RMS*, *flatness*, *zero-crossing*, *spectral centroid*, and *roll-off*. These features were computed using non-overlapping 500 ms windows to match the 2 Hz sampling rate of the emotional ratings.

2.3 Linear Mixed-Effect Models

The linear mixed-effects (LME) model to predict the dynamics of listeners' perceived emotions based on composers' emotional intentions is formulated as:

$$y_{ij} = \alpha + \beta x_{ij} + a_i + b_i x_{ij} + \epsilon_{ij} \quad (1)$$

This equation describes how listeners' perceived emotions (y_{ij}) relate to composers' emotional intentions (x_{ij}) for each music clip (i) at each time point (j). α, β, a_i, and b_i represent the intercept, coefficient for composers' emotional intentions, random intercept for each clip, and random slope for composers' emotional intentions within each clip, respectively. Terms (a_i, b_i) follow a bivariate normal distribution, while ϵ_{ij} represents the residual error.

To investigate the influence of a specific audio feature on this relationship, we employed a LME model:

$$\begin{aligned} y_{ij} = \alpha + \beta_1 x_{ij} + \beta_2 \cdot feature_{ij} + \\ \beta_3 x_{ij} \cdot feature_{ij} + a_i + b_i x_{ij} + \epsilon_{ij} \end{aligned} \quad (2)$$

The terms y_{ij} and x_{ij} represent listeners' perceived emotions and composers' emotional intentions, respectively, for each music clip (i) at each time point (j). α, β_1, β_2, and β_3 denote the intercept, composer's emotional intention coefficient, audio feature coefficient, and their interaction coefficient. Random intercept (a_i) and slope (b_i) account for variation within each clip, while ϵ_{ij} represents the residual error.

3. EXPERIMENT

3.1 Participants

We recruited 36 participants (19 males, 17 females; *mean* age 26.06, *SD* 3.56) through campus mail and online bulletin boards. Except for one participant, who held a master's degree in piano performance, all others were non-musicians. On average, they had about 5.81 years of musical training (*SD* 3.91).

To minimize cultural influences on emotional judgments, only native Korean speakers were included in the experiment. Participants had to meet certain criteria: aged 20 or older, normal vision and hearing, no hand movement disabilities, no diagnosed neurological or psychiatric conditions, and no current use of psychiatric medications.

3.2 Emotional Ratings

The experiment was primarily designed to examine the modality effects on musical emotion inference [28, 29]. Using a counterbalanced design, each participant rated six out of 18 music clips per modality (audio-only, video-only, and video-and-audio), with two clips per emotion (joy/happiness, sadness, and anger). This resulted in 216 emotional ratings for each modality and a total of 648 ratings collected across all clips. We used the 216 emotional ratings from the audio-only condition for the analysis to investigate listeners' emotional experiences during music listening in a more ecological setting.

The dynamic emotional rating task was conducted using *PsychoPy* software, mirroring the dynamic emotional ratings by composers described in Section 2.1.3. Participants,

[1] Sample music clips and supplementary materials are available at `https://osf.io/4dcxu/?view_only=3f1d818e5c4f4e698ebca357daa656cc`.

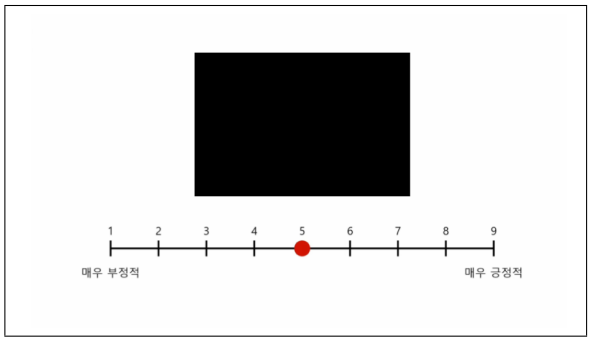

Figure 1. Screenshot of participants' emotional rating task on a valence scale (1: '*very negative*' and 9: '*very positive*', labels in Korean).

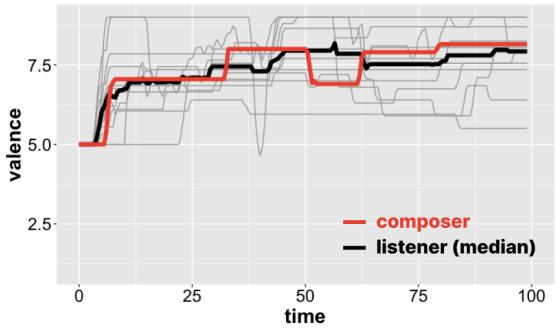

Figure 2. Emotional ratings for a sample joy/happiness music clip: time (x-axis), valence (y-axis, 1: '*very negative*' to 9: '*very positive*'). Red line shows the composer's ratings, black line the median of 12 listeners, and gray lines individual listener ratings.

referred to as *listeners*, were instructed to infer the emotional states of composers expressed in the music. While listening, listeners moved a red dot (initially positioned at 5) along a valence scale (1: '*very negative*' - 9: '*very positive*') whenever they perceived a change in the composer's emotional state [20, 30] (see Figure 1). Ratings were recorded at a sampling rate of 2 Hz, with timestamps every 0.5 seconds.

After each clip, participants evaluated their psychological state using a 9-point Likert scale for arousal, valence, dominance, flow, and empathy. These assessments aimed to minimize the influence of previous emotional experiences on subsequent ratings, and the results were not included in this paper.

3.3 Experiment Procedure

Participants arrived at the lab, completed consent forms, and filled out questionnaires about their music experience. In a soundproof booth, they then performed an emotional inference task. After a practice trial, they listened to six predetermined music clips, inferring the composer's expressed emotion by adjusting a red circle on a scale. Following each clip, they answered five questions about their psychological state and could take breaks. The task was conducted using headphones, with participants adjusting the volume to their preference.

3.4 Data Analysis

All behavior ratings were interpolated using the *scipy* package in Python to maintain consistent time intervals. Silent sections were manually removed from the beginning and end of each audio file before analysis. Audio features were normalized between 0 and 1 at the composer level using min-max normalization. We found strong multicollinearity between spectral centroid and roll-off, so the roll-off feature was excluded from the final analysis to avoid potential overfitting.

For the analysis of listeners' perceived emotions, we selected one representative emotional rating from the responses of 12 listeners for each music clip. The representative value was calculated as the median of 12 ratings for each time point of each clip (see Figure 2). Thus, one time-series data per music stimulus was used for data analysis as *listeners*' emotions. This means that each music clip retained one composer emotion rating, one listener rating, and four audio features. Additionally, the average of listener ratings was used as discrete emotions for comparison with composers' discrete ratings and audio features, as listeners' discrete ratings were not collected (see section 4.2).

Intra-class correlation (ICC) was computed to assess agreement over time among the listener data using the 'ICC' function in the R package *psych*. A two-way mixed, average score ICC was employed for consistency in the 12 valence ratings, following prior research on continuous emotional annotations [2, 31]. The results of this analysis can be found in the supplementary material.

Linear mixed-effects (LME) models were fitted using the *lme4* [32] and the *lmerTest* package [33] in R. Random effects were included in the model structure, and it was found that the random slope of composers' emotions significantly improved the model fit. The random slope was added since the relationship between listeners' perceived emotions and composers' expressed emotions may vary depending on the music clips.

4. RESULTS

4.1 Composers-Listeners Discrete Emotions

To assess the predictability of listeners' perceived emotions using discrete values, we used an LME model analysis. The dependent variable was the average emotional rating from listeners' representative data per music clip. We compared two models: *Model 1* used four audio features (RMS, flatness, zero-crossing, and spectral centroid; the average value of each music clip) as predictors, while *Model 2* added composers' discrete emotional ratings (arousal, valence, and dominance) with four audio features. P-values for fixed effects were obtained using Satterthwaite's approximations, and confidence intervals were computed using the Wald method. Refer to the supplementary material for detailed results of each model.

Model 1 showed that all four audio features sig-

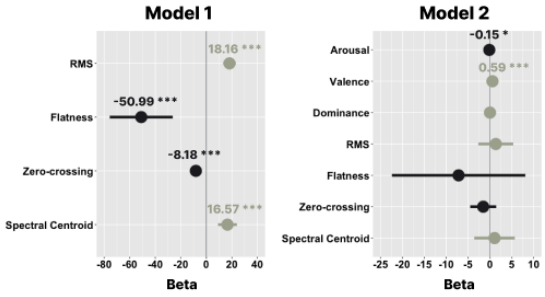

Figure 3. Plot of the effect size of two models.

Metric	Model 1	Model 2	P-value
MAE	2.11	1.17	0.012*
MSE	6.29	2.10	0.009**
RMSE	2.24	1.33	0.009**
MAPE	0.62	0.32	0.012*

Table 2. The *mean* metric values of each model's leave-one-song-out cross-validation for 18 music stimuli. The values were compared using the Wilcox signed-rank test.

nificantly predicted listeners' perceived emotions (see Figure 3). In Model 2, only composers' arousal and valence were significant predictors, with no significant fixed effects for the audio features (see Figure 3). Using leave-one-song-out cross-validation, we confirmed that Model 2 predicted more accurately than Model 1, even for unseen data (see Table 2). This is indicated by its better performance across four metrics: mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and mean percentage error (MPE).

4.2 Composers-Listeners Continuous Emotions

4.2.1 All Emotions

The relationships between dynamic composers' emotional intentions and listeners' perceived emotions were analyzed with an LME model (Equation 1 from Section 2.3). The dependent variable was listeners' emotional representative ratings, and composers' emotion ratings served as the predictor across all music clips (total observations, $N = 3438$; music clips, $N = 18$). The LME analysis revealed that composers' emotional intentions significantly predicted listeners' perceived emotions (beta = 0.26, $p < 0.001$; see Figure 4). Separate analyses for each emotion indicated a significant association between composer-listener emotions except for *joy/happiness* ($p = 0.144$).

4.3 Audio Features & Musical Emotions

LME models were employed to predict composers' and listeners' emotions (see Table 3). For composers' emotions, spectral centroid significantly predicted *all emotions* (beta = 0.74, $p < 0.001$) and for *joy/happiness*, RMS, zero-crossing, and spectral centroid were significant predictors (RMS: beta = 0.60, $p < 0.001$; zero-crossing: beta = -1.11, $p < 0.001$; spectral centroid: beta = 1.09, $p = 0.020$). For

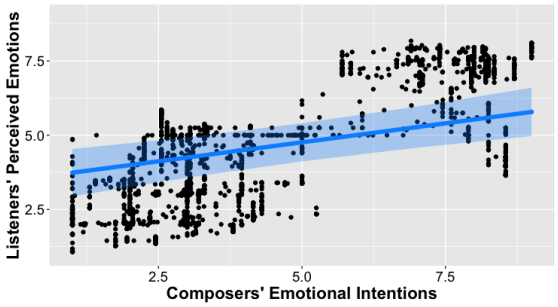

Figure 4. Plot showing listeners' emotions predicted by composers' emotions, with a regression line indicating a slope (beta = 0.26) of the fixed effect for composers.

sadness, RMS was significant (beta = -0.33, $p = 0.016$), and for *anger*, RMS and spectral centroid were significant predictors (RMS: beta = -0.35, $p = 0.042$; spectral centroid (beta = 1.51, $p < 0.001$).

In the LME model predicting listeners' emotions, adding composers' emotional ratings with audio features significantly improved the model fit across all 18 music clips and for each emotion-specific model. RMS consistently predicted listeners' emotions, and zero-crossing emerged as a significant predictor for *anger* music (beta = 0.38, $p = 0.045$).

4.3.1 All Emotions

Building on previous findings, we explored whether audio features that significantly predicted composers' and listeners' emotions could simultaneously predict both subjects' emotions (Equation 2 from Section 2.3). An LME model with listeners' ratings as the dependent variable, and composers ratings, RMS, and their interaction term for predictors (N total observations = 3438; AIC = 5970.1, LogLik = -2977.0), outperformed the model in Section 4.2.1 ($X^2 = 57.24$, $p < 0.001$).

All fixed effects terms were statistically significant in predicting listeners' emotions, particularly the interaction term between composer ratings and RMS (beta = 0.15, $p < 0.001$). This suggests that the relationship between composer and listener emotions varied significantly with changes in RMS levels (see Figure 5), highlighting RMS's role in modulating both composer and listener emotions. Conversely, the interaction term with spectral centroid was not statistically significant (beta = -0.06, $p = 0.055$).

4.3.2 Joy/Happiness

As in Section 4.3.1, we assessed interaction terms between audio features and composers in LME models for each emotion, focusing on *joy/happiness*. Using RMS, zero-crossing, and spectral centroid as predictors, each including an interaction term with composer ratings. Results indicated statistically significant interaction effects across all models: *composer x RMS* (beta = -0.55, $p < 0.001$), *composer x zero-crossing* (beta = 0.58, $p < 0.001$), and *composer x spectral centroid* (beta = 0.74, $p < 0.001$).

	Composer	Listener	Composer & Listener
All Emotions	Spectral centroid	Composer, & RMS	RMS
Joy/Happiness	RMS, Zero-crossing, & Spectral centroid	Composer, & RMS	RMS, Zero-crossing, & Spectral centroid
Sadness	RMS	Composer, & RMS	-
Anger	RMS, & Spectral centroid	Composer, RMS, & Zero-crossing	RMS, & Zero-crossing

Table 3. Significant predictors included four audio features for composers' emotional intentions and listeners' perceived emotions. Composers' emotions were added as predictors in the models predicting listeners' emotions.

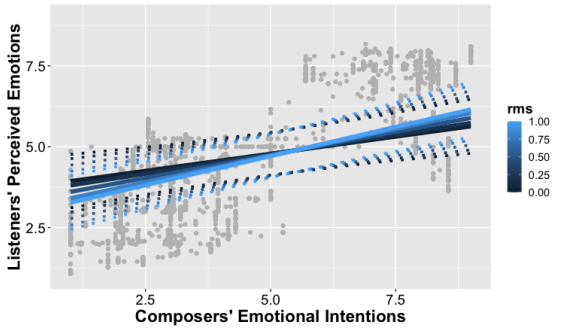

Figure 5. An interaction plot illustrating the model predicting listeners' emotions with composers' emotions and RMS. The solid line depicts the slope of the fixed effect of *composer*, varying with changes in RMS values.

4.3.3 Sadness

For *sadness*, an LME model was fitted with the RMS feature and the interaction term between RMS and composer as predictors. The results showed that neither the interaction term (p = 0.409) nor the fixed effect of the RMS feature (p = 0.365) were statistically significant.

4.3.4 Anger

The LME model for *anger* music included fixed effects and interaction terms for RMS, zero-crossing, and spectral centroid features. Results showed that the interaction terms for RMS (beta = -0.32, $p < 0.001$) and zero-crossing (beta = 0.53, $p < 0.001$) were statistically significant. However, the spectral centroid model did not show significant results upon model comparison (p = 0.222).

5. DISCUSSION

In this study, we employed linear mixed-effects (LME) models to explore how spectral features of music predict both the composer's real-time intended emotional expression during piano improvisations and the listener's perceived emotion. This included gathering emotional ratings empirically from composers and also listeners on a valence scale. We then examined the relationship between these ratings and the features extracted from the music clips.

We found that composers' emotional intentions were conveyed to listeners' perceptions of musical emotions. Discrete emotional ratings showed that composers' intentions were stronger predictors of listeners' perceived emotions than other audio features. Conversely, continuous emotional data emphasized the importance of both composers' intentions and RMS features. These results underscored the impact of emotional assessment methodologies, suggesting that discrete emotion ratings may overlook acoustic cues conveying composers' intentions.

Overall, RMS was identified as a primary predictor for conveying composers' intentions and also served as an indicator of listeners' emotional perceptions. While RMS was the key feature for predicting listeners' emotions, the features that indicated composers' emotions varied with different emotional categories. For *joy/happiness* and *anger*, the spectral centroid emerged as the main predictor of the composers' intentions, likely due to its association with timbral brightness, which helps detect changes in the valence [26].

Our findings highlight RMS as a crucial audio feature for predicting the emotions of both composers and listeners. RMS was strongly associated with emotional dynamics in *joy/happiness* and *anger*, but not in *sadness*. This is consistent with prior research [9], which also found RMS to be an effective predictor of *happiness* and *anger*, but not *sadness*. Additionally, zero-crossing emerged as a significant predictor of the emotional relationship between composers and listeners for both *joy/happiness* and *anger*, further aligning with the findings of previous studies on speech emotion recognition [34].

However, we found no audio features capable of predicting composer-listener emotions for *sadness*, which typically involves lower arousal compared to *joy/happiness* and *anger*. In music with low arousal, features related to valence may not be as prominent. For instance, in *sad* music, changes in loudness (i.e., RMS) may not be as pronounced as in *joy/happiness* or *anger*, thus potentially not serving as cues for both composers' emotional intentions and listeners' perceptions of valence.

Future research should further explore more audio features related to other musical factors (*e.g.*, tempo, pitch, harmony, etc.) that are known to be associated with emotional experiences in music. Additionally, it is needed to determine whether the features identified in this study enhance the performance of MER models. Expanding the sample size of participants and utilizing a larger pool of music stimuli, while considering individual and cultural variations, will also be essential to enhance generalizability and gain more comprehensive insights.

This study offers insights into factors influencing MER system predictions of emotional valence, enhancing machine-learning models' ability to predict listeners' emotions by considering feature importance across different emotions. Additionally, incorporating time-series emotional data, including composers' intentions, adds further significance to the research.

6. ACKNOWLEDGMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023R1A2C100475512).

7. ETHICS STATEMENT

IRB approval was obtained by the Korea Advanced Institute of Science and Technology (KH2023-070).

8. REFERENCES

[1] M. Schedl, A. Flexer, and J. Urbano, "The neglected user in music information retrieval research," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 523–539, 2013.

[2] S. Yang, C. N. Reed, E. Chew, and M. Barthet, "Examining emotion perception agreement in live music performance," *IEEE transactions on affective computing*, vol. 14, no. 2, pp. 1442–1460, 2021.

[3] S. Chaki, P. Doshi, S. Bhattacharya, and P. Patnaik, "Explaining perceived emotion predictions in music: An attentive approach." in *ISMIR*, 2020, pp. 150–156.

[4] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021.

[5] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[6] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PloS one*, vol. 12, no. 3, p. e0173392, 2017.

[7] T. Eerola and J. K. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models, and stimuli," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 3, pp. 307–340, 2012.

[8] L. Turchet and J. Pauwels, "Music emotion recognition: intention of composers-performers versus perception of musicians, non-musicians, and listening machines," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 305–316, 2021.

[9] E. B. Lange and K. Frieler, "Challenges and opportunities of predicting musical emotions with perceptual and automatized features," *Music Perception: An Interdisciplinary Journal*, vol. 36, no. 2, pp. 217–242, 2018.

[10] J. Akkermans, R. Schapiro, D. Müllensiefen, K. Jakubowski, D. Shanahan, D. Baker, V. Busch, K. Lothwesen, P. Elvers, T. Fischinger *et al.*, "Decoding emotions in expressive music performances: A multi-lab replication and extension study," *Cognition and Emotion*, vol. 33, no. 6, pp. 1099–1118, 2019.

[11] F. Pan, L. Zhang, Y. Ou, and X. Zhang, "The audio-visual integration effect on music emotion: Behavioral and physiological evidence," *PloS one*, vol. 14, no. 5, p. e0217040, 2019.

[12] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological bulletin*, vol. 129, no. 5, pp. 770–814, 2003.

[13] P. E. Keller, "Ensemble performance: Interpersonal alignment of musical expression," *Expressiveness in music performance: Empirical approaches across styles and cultures*, vol. 1, pp. 260–282, 2014.

[14] P. A. Holmes, "An exploration of musical communication through expressive use of timbre: The performer's perspective," *Psychology of Music*, vol. 40, no. 3, pp. 301–323, 2012.

[15] A. C. Miu and F. R. Balteş, "Empathy manipulation impacts music-induced emotions: A psychophysiological study on opera," *PloS one*, vol. 7, no. 1, p. e30618, 2012.

[16] Y. Hou, B. Song, Y. Hu, Y. Pan, and Y. Hu, "The averaged inter-brain coherence between the audience and a violinist predicts the popularity of violin performance," *Neuroimage*, vol. 211, p. 116655, 2020.

[17] S. Vieillard, I. Peretz, N. Gosselin, S. Khalfa, L. Gagnon, and B. Bouchard, "Happy, sad, scary and peaceful musical excerpts for research on emotions," *Cognition & Emotion*, vol. 22, no. 4, pp. 720–752, 2008.

[18] P. N. Juslin, "Cue utilization in communication of emotion in music performance: Relating performance to perception." *Journal of Experimental Psychology: Human perception and performance*, vol. 26, no. 6, pp. 1797–1812, 2000.

[19] L. Xu, X. Wen, J. Shi, S. Li, Y. Xiao, Q. Wan, and X. Qian, "Effects of individual factors on perceived emotion and felt emotion of music: based on machine learning methods," *Psychology of Music*, vol. 49, no. 5, pp. 1069–1087, 2021.

[20] B. A. Tabak, Z. Wallmark, L. H. Nghiem, T. Alvi, C. S. Sunahara, J. Lee, and J. Cao, "Initial evidence for a relation between behaviorally assessed empathic accuracy and affect sharing for people and music." *Emotion*, vol. 23, no. 2, pp. 437–449, 2023.

[21] A. S. Cowen, X. Fang, D. Sauter, and D. Keltner, "What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures," *Proceedings of the National*

Academy of Sciences, vol. 117, no. 4, pp. 1924–1934, 2020.

[22] C. MacGregor, N. Ruth, and D. Müllensiefen, "Development and validation of the first adaptive test of emotion perception in music," *Cognition and Emotion*, vol. 37, no. 2, pp. 284–302, 2023.

[23] A. Gabrielsson and E. Lindström, "The influence of musical structure on emotional expression," in *Music and emotion: Theory and research*, P. N. Juslin and J. A. Sloboda, Eds. Oxford University Press, 2001, pp. 223–248.

[24] T. Eerola, A. Friberg, and R. Bresin, "Emotional expression in music: contribution, linearity, and additivity of primary musical cues," *Frontiers in psychology*, vol. 4, p. 487, 2013.

[25] C. Plut, P. Pasquier, J. Ens, and R. Tchemeube, "The isovat corpus: Parameterization of musical features for affective composition," *Transactions of the International Society for Music Information Retrieval*, vol. 5, no. 1, pp. 173–189, 2022.

[26] I. Salakka, A. Pitkäniemi, E. Pentikäinen, K. Mikkonen, P. Saari, P. Toiviainen, and T. Särkämö, "What makes music memorable? relationships between acoustic musical features and music-evoked emotions and memories in older adults," *PloS one*, vol. 16, no. 5, p. e0251692, 2021.

[27] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." in *SciPy*, 2015, pp. 18–24.

[28] E. J. Oh, "Shared empathic process in music and social contexts: exploring empathic accuracy and physiological responses across modalities and valence," Master's thesis, KAIST, 2024.

[29] E. J. Oh and K. M. Lee, "Intermodal analysis of emotion inference: Examining shared processes in music and social contexts." in *Society for Music Perception and Cognition (SMPC)*, 2024, p. 68.

[30] J. Zaki, N. Bolger, and K. Ochsner, "It takes two: The interpersonal nature of empathic accuracy," *Psychological science*, vol. 19, no. 4, pp. 399–404, 2008.

[31] N. Dibben, E. Coutinho, J. A. Vilar, and G. Estévez-Pérez, "Do individual differences influence moment-by-moment reports of emotion perceived in music and speech prosody?" *Frontiers in behavioral neuroscience*, vol. 12, p. 184, 2018.

[32] D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, G. Grothendieck, P. Green, and M. B. Bolker, "Package 'lme4'," *convergence*, vol. 12, no. 1, p. 2, 2015.

[33] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmertest package: tests in linear mixed effects models," *Journal of statistical software*, vol. 82, no. 13, 2017.

[34] E. Ramdinmawii, A. Mohanta, and V. K. Mittal, "Emotion recognition from speech signal," in *TENCON 2017 - 2017 IEEE Region 10 Conference*, 2017, pp. 1562–1567.