PERCEPTUAL VS. AUTOMATED JUDGEMENTS OF MUSIC COPYRIGHT INFRINGEMENT

Yuchen Yuan[1], Sho Oishi[1], Charles Cronin[2], Daniel Müllensiefen[3], Quentin Atkinson[4], Shinya Fujii[1], Patrick E. Savage[1]*

[1]Keio University, Japan; [2]George Washington University Law School, USA; [3]Goldsmiths, University of London, UK; [4]University of Auckland, New Zealand

*psavage@sfc.keio.ac.jp

ABSTRACT

Music copyright lawsuits often result in multimillion dollar damage awards or settlements, yet there are few objective guidelines for applying copyright law in infringement claims involving musical works. Recent research has attempted to develop objective methods based on automated similarity algorithms, but there remains almost no data on the role of perceived similarity in music copyright decisions despite its crucial role in copyright law. We collected perceptual data from 20 participants for 17 adjudicated copyright cases from the USA and Japan after editing the disputed sections to contain either full audio, melody only, or lyrics only. Due to the historical emphasis in legal opinions on melody as the key criterion for deciding infringement, we predicted that listening to melody-only versions would result in perceptual judgements that more closely matched actual past legal decisions. Surprisingly, however, we found no significant differences between the three conditions, with participants matching past decisions in between 50-60% of cases in all three conditions. Automated algorithms designed to calculate melodic and audio similarity produced comparable results: both algorithms were able to match past decisions with identical accuracy of 71% (12/17 cases). Analysis of cases that were difficult to classify suggests that melody, lyrics, and other factors sometimes interact in complex ways difficult to capture using quantitative metrics. We propose directions for further investigation of the role of similarity in music copyright law using larger and more diverse samples of cases and enhanced methods, and adapting our perceptual experiment method to avoid relying for ground truth data only on court decisions (which may be subject to selection bias). Our results contribute to important practical debates, such as whether jury members should be allowed to listen to full audio recordings during copyright cases.

1. INTRODUCTION

Music copyright law protects the lawful rights and interests of music creators and performers, but in some music copyright infringement cases, its application has caused bitter controversy. As litigation becomes more frequent, inappropriate music copyright lawsuits not only inhibit music creativity but also waste millions of taxpayer dollars annually to cover the adjudication of these disputes. The legal system and music industry could both benefit from automated methods that could reduce subjectivity in music copyright decisions, and several recent studies have proposed such automated methods [1-4]. While the accuracy of some algorithms have been tested against previous court decisions, they have not yet been tested against perceptual data to determine how different musical and extra-musical factors interact in copyright law.

"Substantial similarity" and "protectable expression" are central concepts in US copyright law, the understanding of which could potentially be supplemented through automated and/or perceptual analyses. The concept of "substantial similarity" requires not only that the defendant can be shown to have copied musical material, but that this copying of protected musical expression was so extensive that the two works are substantially similar [5]. Data on degrees of computed and/or perceived similarity can help to determine objective standards for how much copying is required to be considered "substantial".

Evaluating what is considered "protectable expression" is more qualitative and complex. Many musical aspects such as scales, certain rhythmic patterns, and timbres are considered to be such basic and commonplace musical ideas as not to be copyrightable. For example, many blues songs all use very similar blues scales, 12-bar harmonic progressions, vocal styles and instrumentation, but copying these aspects is not considered copyright infringement. Instead, melody (i.e., the sequence of pitches) and lyrics have traditionally played predominant roles against other musical factors [6-7]. However, it has been disputed whether jury members should be allowed to listen to full-audio or melody-only versions of musical works because people may perceive and judge differently when comparing pairs of melodies or other musical features [8]. For example, a core issue in the recently concluded case involving the Blurred Lines [9] was whether the jury should be allowed to listen to a full audio recording including lyrics and background instrumentation of the complaining work, or whether it should only be exposed to the sheet music that was deposited with the US Copyright Office [6].

To quantitatively compare the effects of melody, lyrics, and other factors, we designed a controlled experiment where we constructed versions of a disputed musical work containing the full audio (including lyrics, melody, and other factors such as instrumentation), melody only (pitches and rhythms in MIDI representation), and lyrics only (text representation). Because of the historical dominance of melody, we predicted that participants would most accurately match past decisions when presented with melody-only versions, and that automated algorithms based on melodic data would more accurately match past decisions than ones based on full-audio data.

Section 2 discusses related research. Section 3 discusses the data selection for our study. In Section 4, we demonstrate the design and details of the perceptual experiment. In Section 5, we show how the melodic and audio similarity are calculated by automated algorithms (PMI and Musly, respectively). In Sections 6, 7, and 8, we discuss the performance of the two automated methods, compare the automated and perceptual methods, summarize current results, and discuss future directions for improvement.

2. RELATED WORK

2.1 Perceptual Experiments

In one previous experimental study, Lund used two past court cases (*Swirsky v. Carey* and *Gasté v. Kaiserman*) and manipulated MIDI representations of the works to change aspects such as tempo, rhythm, and instrumentation [8]. Lund found that such manipulations reduced the accuracy of participants' judgements of copyright infringements even though it was assumed that such non-melodic features should not play a role in decisions. Lund argued that this demonstrated that the "lay listener test" was flawed because it relies on subjective listening to audio recordings that may differ in non-melodic aspects. However, Lund did not compare full audio recordings with these MIDI representations, so it remains unknown whether listeners are in fact more accurate when listening to MIDI representations than when listening to full audio recordings.

2.2 Automatic Analysis

Müllensiefen and Pendzich developed an algorithm for judging melodic similarity that compares the profile of successive pitch intervals in two disputed songs against each other, while weighting them against a database of comparable profiles from 14,063 pop songs using a weighting formula for estimating perceptual salience [1]. When they applied this algorithm to a database of 20 past music copyright decisions focused on melodic similarity, they found the best-performing version of their algorithm was able to accurately identify 90% (18/20) of past cases.

Savage et al. later developed a Percent Melodic Identity (PMI) method for quantifying melodic evolution based on automatic sequence alignment algorithms used in molecular genetics to measure melodic similarity [10]. When they applied this method to the same set of cases as Müllensiefen & Pendzich, it accurately predicted 80% (16/20) of cases, despite being a simpler method that didn't require calibration to an existing database of popular songs [3].

While the related task of cover song detection has a long history of study in music information retrieval [11-12], to our knowledge no audio similarity algorithms have yet been tested for their ability to evaluate copyright infringement. However, many general audio similarity algorithms have been evaluated through the Music Information Retrieval Exchange (MIREX) competition. We thus chose the audio similarity algorithm implemented in Musly, an open-source library of audio music similarity algorithms, because it has consistently performed at or near the top of audio similarity algorithms as evaluated in MIREX [13].

3. DATASET OF MUSIC COPYRIGHT INFRINGEMENT CASES

We chose a set of 17 court decisions whose main copyright issue focused on substantial similarity of the melodies (Table 1). 14 of these 17 cases are from the US, and these 14 represent a subset of 20 cases from the Music Copyright Infringement Resource [14] that were previously analyzed [1, 3] for which full audio recordings were available for both of the disputed musical works (the remaining 6 cases were not included because one or both musical works were represented only by sheet music and/or MIDI files). We also included 3 court decisions from Japan in order to increase cultural diversity in the dataset for further study on adaptability to music other than Western music. Of the 17 cases courts found no infringement in 8 cases, and infringement in 9.

4. PERCEPTUAL EXPERIMENT

4.1 Experiment Design

We conducted an online perceptual experiment where participants were each asked to judge substantial similarity for the 17 cases. The disputed segments of the musical works (mean length: 22s; range: 3-55s) were presented in one of three different versions: full-audio (the recorded versions including all instrumental and/or vocal parts), melody-only (MIDI rendition of the pitches and rhythms of the main melody), and lyrics-only (lyrics shown as visual text, without any accompanying audio). For the melody condition, in order to control for all non-melodic factors including instrumentation, key, and tempo, transcribed melodies from the original audio recordings were edited as necessary to exactly correspond to the audio recordings[1]: these transcribed melodies were transposed to have a tonic of C, and were then recorded using the MIDI piano in MuseScore played back at a tempo that was the average of the tempi from the plaintiff and defendant recordings. For the lyrics condition, the three instrumental works without lyrics (cf. Table 1) simply showed "[no lyrics]". These three types of presentations were repeated twice: once using the originally disputed pair of musical works, and once using the original defendant work but comparing it against a randomly selected plaintiff work from the other 16 cases.

[1] While preparing the audio files for experiments we noticed several minor inconsistencies between the audio files and the transcriptions provided by the authors of [1]. In some cases, these were small errors in pitch/rhythm; in others, only one half of a larger section was transcribed. The original transcriptions were not initially published but have now been uploaded to https://github.com/pesavage/copyright/tree/master/MIDIs_plagiarismcases_MullensiefenPendzich2009 to allow comparison as necessary. The corrected transcriptions are available at https://github.com/compmusiclab/music-copyright.

No.	Country	Case	Complaining Work	Length (seconds)	Defending Work	Length (seconds)	Court Decision	PMI (cutoff = 46.8%)	Musly-calculated Similarity (cutoff = 32.8%)	Perceptual Accuracy - Full audio	Perceptual Accuracy - Melody only	Perceptual Accuracy - Lyrics only	Perceptual Similarity - Full audio	Perceptual Similarity - Melody only	Perceptual Similarity - Lyrics only
1	JP	Harry vs. Suzuki	“Boulevard of Broken Dreams”	33	“ワン・レイニーナイト・イン・トーキョー” (One Rainy Night in Tokyo)	23	0	25%	25%	65%	80%	100%	3.15	2.9	1.4
2	**US**	**Cottrill vs. Spears**	**“What You See is What You Get”**	**22**	**“What U See is What U Get”**	**24**	**0**	**35%**	**41%**	70%	95%	65%	2.75	1.85	3
3	US	Baxter vs. MCA	“Joy”	7	“Theme from ‘E.T.’”	19	0	37%	12%	85%	90%	N/A	2.7	2	N/A
4	*US*	*Swirsky vs. Carey*	*“One of Those Love Songs”*	*29*	*“Thank God I Found You”*	*32*	*1*	*45%*	76%	60%	35%	0%	3.45	3	1.4
5	US	Repp vs. Lloyd-Webber	“Till You”	27	“Phantom Song”	38	0	45%	15%	50%	35%	100%	3.15	4.35	1.25
6	**JP**	**Kobayashi vs. Hattori**	**“どこまでも行こう” (Dokomademoikō)**	**23**	**“記念樹” (Kinenju)**	**40**	**1**	**47%**	**10%**	55%	45%	10%	3.6	3.2	1.55
7	US	Three Boys Music vs. Michael Bolton	“Love Is A Wonderful Thing”	10	“Love Is A Wonderful Thing”	17	1	47%	63%	70%	30%	50%	3.65	3.25	3.7
8	**US**	**Herald Square Music vs. Living Music**	**“Day By Day”**	**32**	**“Theme N.B.C.’s ‘Today Show’”**	**30**	**1**	**51%**	**5%**	45%	40%	N/A	3.6	2.85	N/A
9	**US**	**Grand Upright vs. Warner**	**“Alone Again (Naturally)”**	**5**	**“Alone Again”**	**6**	**1**	**53%**	**25%**	70%	30%	50%	4.2	2.9	4
10	US	Bright Tunes Music vs. Harrisongs Music	“He’s So Fine”	27	“My Sweet Lord”	55	1	58%	35%	25%	45%	5%	2.5	3.25	1.3
11	*US*	*Selle vs. Gibb*	*“Let It End”*	*21*	*“How Deep Is Your Love”*	*19*	*0*	*63%*	11%	55%	40%	95%	3.25	3.65	1.65
12	US	Louis Gaste vs. Morris Kaiserman	“Pour Toi”	17	“Feelings”	21	1	65%	33%	50%	50%	0%	3.4	3.8	1.35
13	*US*	*Granite Music vs. United Artists*	*“Tiny Bubbles”*	*18*	*“Hiding The Wine”*	*11*	*0*	*67%*	4%	60%	40%	N/A	3.3	3.8	N/A
14	***US***	***Fantasy vs. Fogerty***	***“Run Through The Jungle”***	***21***	***“The Old Man Down The Road”***	***21***	***0***	***67%***	**62%**	40%	45%	100%	3.45	3.3	1.4
15	*US*	*Jean et al. vs. Bug Music*	*“Hand Clapping Song”*	*3*	*“My Love Is Your Love”*	*4*	*0*	*71%*	20%	45%	80%	90%	3.75	2.6	2.8
16	US	Levine vs. McDonald’s	“Life Is A Rock (But The Radio Rolled Me)”	22	“McDonald’s Menu Song”	26	1	80%	63%	65%	45%	10%	4	3.6	1.8
17	JP	HarumakiGohan vs. Mori	“八月のレイニー” (Hachigatsu no reinī)	21	“M.A.K.E”	22	1	100%	54%	75%	85%	10%	4.25	4.35	2

Table 1. The 17 music copyright infringement cases analyzed and respective melodic similarity (PMI), audio music similarity (Musly), and perceptual experiment results. Cases are ordered by increasing PMI values. In “Court Decision” column, “0” represents no infringement, and “1” represents infringement. Cases in italics are those PMI failed to accurately classify, and bold indicates those Musly failed to classify. Columns highlighted in light blue are the accuracy of perceptual judgement for the 17 court cases judged by the 20 participants for full-audio, melody-only, and lyrics-only, corresponding to the data in Figure 1. Columns highlighted in light green are perceptual similarity values used for comparison between automatic methods and perceptual judgement in Section 6, corresponding to the data in Figures S2 and S3. Three defending works marked in orange text are instrumental.

This gave a total of 102 different pairs of musical works to evaluate (17 cases × 3 presentations [full, melody, lyrics] × 2 pairings [original plaintiff vs. random plaintiff]), presented in fully random order (without separate blocks for different conditions; i.e., any given sample might be full-audio, melody-only, or lyrics-only; original case or not). Each experiment took approximately 2 hours for one participant to complete evaluations for these 102 pairs.

For each pair, the participant is given a pair of music excerpts, “A” and “B”. “A” is always a plaintiff’s work while “B” is always a defendant’s work. After listening to the full-audio or MIDI or reading the lyrics of the two music works, the participant needs to answer two questions: 1) How similar are A and B? (5-point Likert scale: “not at all similar”, “a little similar”, “somewhat similar”, “very similar”, and “extremely similar”). 2) Do you think the second music work (“B”) infringed the copyright of the first one’s (“A”)? (Yes/no answer.) The following criteria for infringement were provided, taken from [8] (which was in turn adapted from real instructions given to juries [details of the adaptation were not provided]):

> *To find music copyright infringement between plaintiff’s and defendant’s songs, you must find that the songs are substantially similar. Two works are substantially similar if the original expression of ideas in the plaintiff’s (Song #1) copyrighted work and the expression of ideas in the defendant’s work (Song #2) that are shared are substantially similar. Original expression are those unique aspects of plaintiff’s song that are not common or ordinary to the genre or to music generally. The amount of similarity must be both quantitatively and qualitatively significant, that is the defendant’s song copied either a substantial portion of the original expression of the plaintiff’s song, or copied a smaller but qualitatively important portion of the plaintiff’s song.*

In short, this investigation imitates the lay listener test used to see whether an ordinary observer recognizes that the defendant appropriated something belonging to the plaintiff [8, 15].

4.2 Results

We collected perceptual data from 20 participants from our institution. 9 were male, and 11 were female. 17 were between 17-28 years old, 1 was between 29-50, and 2 were over 50. The native languages of participants were Chinese (13 participants), Japanese (6) and English (1). 11 reported substantial music experience while 9 did not. Table 1 summarizes all results for perceptual and automated experiments.

Figure 1 and S1 show how accurately the 20 participants’ judgement of infringement matched the official court decisions when they were given full-audio, melody-only, or lyrics-only versions of music pieces from the 17 court cases. Note that accuracy is measured as how likely participants were to match court decisions, whether that decision was of infringement or no infringement. Although the perceptual data were collected for the three cases including instrumental works, these cases were omitted from the lyrics-only analyses because infringement of lyrics is clearly impossible for instrumental works. In Figure S1, individual data points represent mean accuracy for individual *participants* (n = 20) across the 17 cases, while in Figure 1 individual data points represent mean accuracy for individual *cases* (n = 17) across the 20 participants.

Surprisingly, the accuracy numbers by participants of the three condition groups distributed quite closely, with mean accuracy of 58%, 54%, and 49% for full-audio, melody-only, and lyrics-only groups respectively. Not only was our predicted difference between melody-only and full-audio not significant (paired t = 1.7, df = 19, one-tailed p = 0.95), but what small difference there was between full-audio and melody-only was in the opposite direction from our predictions (participants were slightly *more* accurate when presented with full-audio than with melody alone). However, randomized control pairs had modal accuracies of 100% for all three conditions (cf. Figure S1), confirming that participants were able to perform all three tasks much more accurately than by chance. In addition, participants who self-reported as musicians showed no significant differences in accuracy compared to non-musicians (full: t = 0.63, df = 11, 1-sided p-value = 0.27; melody: t = 1.20, df = 17, 1-sided p-value = 0.12; lyrics: t = 0.68, df = 14, 1-sided p-value = 0.25).

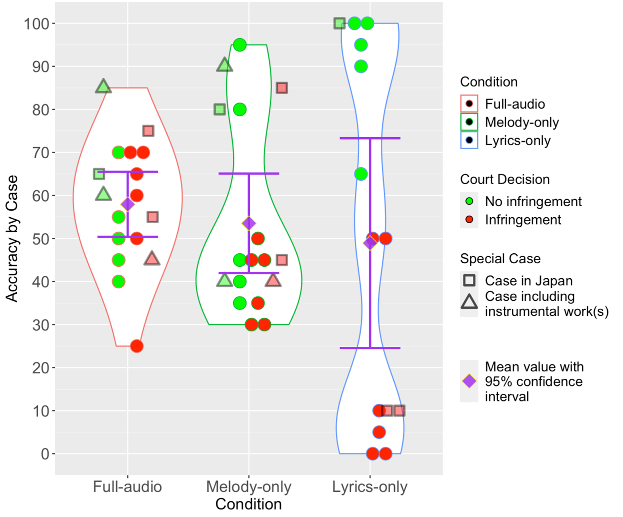

Figure 1. Accuracy of perceptual judgement for each of the 17 court cases, as measured by the percentage of the 20 participants whose judgements of music copyright infringement matched court decisions.

Figure 1 plots the accuracy of perceptual judgement for the 17 court cases judged by the 20 participants. As in Fig. S1, the means of the three conditions are similar. Interestingly, however, the accuracy values are approximately normally distributed for the full-audio condition, while the melody-only and lyrics-only conditions have bimodal, hourglass-shaped distributions. Furthermore, full-audio cases show no major differences in the distribution of infringing vs. non-infringing cases, while melodic cases show some skewing toward higher accuracy for non-infringing cases and lyrics show a strong dichotomy between high accuracy for non-infringing cases and low accuracy for infringing cases. No clear differences are notable for the small subsets of cases from Japan or those involving instrumental works.

5. AUTOMATIC ANALYSIS

We performed automatic similarity analysis of these cases using two different automated algorithms focused on melodic and audio similarity, respectively.

5.1 Melodic Similarity (Percent Melodic Identity [PMI])

We chose the PMI (Percent Melodic Identity) method to calculate melodic similarity because it has been validated in previous research using a similar sample of copyright cases. Like Judge Learned Hand's "comparative method" [6] to test musical similarity, the PMI method begins by transposing two melodies transcribed in staff notation into a same key, eliminating rhythmic information by assigning all notes equal time values, and then aligning and counting the confluence of notes[1]. Following the procedure, we prepared note sequences of disputed melodies all transposed to a C tonic for consistency (just as was done when preparing MIDI files). The PMI algorithm then automatically aligns each sequences pair, and counts the number of identical notes (*ID*). The percentage of identical notes shared between the pair of melodies, named percent melodic identity (*PMI*) [3], is calculated by dividing *ID* by the average length of the melodies pair (L_1 and L_2), as follows:

$$PMI = 100\left(\frac{ID}{\frac{L_1 + L_2}{2}}\right)$$

5.1.1 Melodic Similarity Results

The PMI values computed for all 17 music copyright infringement cases are shown in Table 1. Receiver Operating Characteristic (ROC) analysis was used to assess the prediction given by PMI values. The area under the ROC curve (AUC) is 0.61. The optimal cutoff PMI value is 46.8% with sensitivity = 0.89 and specificity = 0.50. Using this cutoff, PMI method was able to accurately classify 12 out of the 17 cases (71%) to match their court decisions. The five cases highlighted by italic font in Table 1 are those that the PMI method failed to classify correctly, discussed further below.

5.2 Audio Similarity (Musly)

Musly currently implements two music similarity algorithms. One implements Mandel-Ellis audio similarity algorithm [16]. The other one, which is the default one, improves Mandel-Ellis algorithm to compute audio similarity for best results. Specifically, it computes a representation of each song's audio signal based on 25 Mel-Frequency Cepstral Coefficients (MFCCs) to estimate a Gaussian model and finally a single timbre model to be compared, computes similarity between each pair of timbre models using Jensen-Shannon approximation, and normalizes the

[1] Note that rhythms are not eliminated for the perceptual stimuli, only for the PMI calculation (see [10] for discussion of treatment of rhythm in the PMI method).

similarities with Mutual Proximity [17-18]. We used the default algorithm because it has been found to have higher accuracy [17].

We prepared the full-audio version of the music excerpts from the dataset of court cases and fed them to the default algorithm of Musly to compute similarity. The output of the Musly algorithm is a distance matrix where distances, i.e. differences, between every two songs are listed. Because the Musly default algorithm normalizes the results, all the distances range between 0 and 1. Consequently, we calculated the audio music similarity by subtracting distance values from 1 and multiplying by 100 to convert the results into percentage terms for consistency with our other methods.

5.2.1 Audio Similarity Results

The results of Musly-calculated audio music similarity values for all 17 tested cases are shown in Table 1, appended next to the column of PMI values. The area under the ROC curve (AUC) is 0.69. The optimal cutoff threshold of Musly-calculated similarity is 32.8% with sensitivity = 0.67 and specificity = 0.75. Using this cutoff, Musly algorithm was also able to accurately classify 12 out of the 17 cases (71%) to match the court's decisions. The five failure cases are highlighted by bold font in Table 1 and briefly analyzed below.

6. AUTOMATED VS. PERCEPTUAL JUDGEMENTS

6.1 PMI vs. Perceptual Data

Mean perceptual similarity of each court case was calculated by averaging participants' individual ratings of similarity. The perceptual similarity values for the 17 court cases are listed in Table 1 and highlighted by light green. Figure S2 shows the relationship between PMI values and perceptual similarity under the three different conditions. Regression analyses show that the PMI melodic similarity is significantly correlated with perceptual similarity for both full-audio and melody-only conditions (full: $R = 0.58$, $p = 0.014$; melody: $R = 0.59$, $p = 0.012$), but not for the lyrics-only condition ($R = -0.058$, $p = 0.84$).

6.2 Musly vs. Perceptual Data

We also compared the Musly-calculated audio music similarity with the perceptual data collected. Figure S3 shows the correlation between Musly similarity and perceptual similarity of the 17 tested court cases under three different conditions for perceptual judgement. Regression analyses indicate that the Musly audio similarity has no significant correlations with perceptual similarity for all three condition groups of "full-audio", "melody-only", and "lyrics-only" (full: $R = 0.26$, $p = 0.32$; melody: $R = 0.082$, $p = 0.76$; lyrics: $R = 0.059$, $p = 0.84$).

7. DISCUSSION

Overall, our analyses showed moderate agreement between automated and perceptual judgements of music copyright infringement. Both automated similarity algorithms – PMI for symbolic data and Musly for audio data – matched past court decisions with relatively high accuracy (both 71%). The fact that PMI was significantly correlated with perceptual similarity for both melody-only and full-audio provides validation for PMI as a perceptually relevant measure of melodic similarity and is consistent with the idea that melodic similarity plays a role in judgements of overall musical similarity [19].

The lack of correlation between Musly's audio similarity algorithm and perceptual similarity was surprising given that Musly's algorithm has previously performed well in evaluations of general musical similarity. This may be partly explained by Musly's reliance on MFCCs to capture timbral and rhythmic similarity, not melodic similarity. Previous studies have shown that limited inter-rater reliability in judgements of musical similarity can limit the performance of automated algorithms [13]. Future analyses using supervised learning or other algorithms for capturing melodic similarity [1] may be able to improve performance, although the subjective nature of musical similarity will still place limits on the ability of any algorithm to match human judgements.

Surprisingly, both automated methods had higher accuracy than that of perceptual judgement, with both automated methods able to accurately predict 71% (12/17) of previous court decisions while perceptual accuracy were 58% and 54% under full-audio and melody-only conditions respectively. We suspect that allowing the algorithms to optimize the similarity threshold via the ROC analysis helped to improve - and probably overfit - the automated analyses[1]. Future analyses with larger data samples should consider calibrating parameters on a training subset before evaluating them on a separate test subset.

There are several possibilities for the low levels of perceptual accuracy. The fact that participants showed very high levels of accuracy (almost 100%) for randomized plaintiff samples suggests that the results were not merely random, but the inclusion of such samples might conceivably have skewed judgements by including levels of dissimilarity rarely included in real court cases. The fact that musicians performed similarly to non-musicians suggests that lack of musical expertise is also unlikely to explain the low performance. Although we cannot rule out effects of participants' familiarity because we failed to collect such data, any familiarity effects when participants were aware of the cases would be predicted to increase, rather than decrease, accuracy.

Instead, some past court decisions (e.g., the cases involving "He's So Fine" and "Blurred Lines") have been so controversial as to be debatable whether they were in fact "correct" [6]. Indeed, it seems likely that the dynamics of copyright lawsuits create a type of selection bias in which cases where infringement or lack of infringement are obvious are more likely to be resolved out of court[2] without a final court decision, while only the most ambiguous

[1] The chance of getting an accuracy of 12 or more correct by chance is actually 26%.

[2] One case (HarumakiGohan v. Mori) was settled out of court, and this case displayed some of the highest levels of participant accuracy.

cases require the court to make a final adjudication. In the future, rather than relying for ground truth only on court decisions and the selection bias they may create, our perceptual experiment may provide an alternative source of ground truth for disputes that were resolved out of court and thus tend to lack objective legal documentation.

Our prediction that listening to melody-only would provide superior accuracy than listening to full-audio was not supported. The fact that our prediction was not only not significant but was in the wrong direction suggests that limited statistical power cannot explain this result. Instead, despite legal arguments suggesting that non-melodic factors should generally be ignored and the sample having been selected based on the criteria of melodic similarity [1], individual cases are always complex and factors such as lyrics, instrumentation, and other non-melodic factors did in fact play roles in past decisions [14]. Overall, participants tended to judge melody-only versions as less similar than full-audio, with accuracy tending to be lower for cases judged as infringing. This suggests that participants have more difficulty detecting infringement using melody only. Since including non-melodic information appears to help (or at least not hurt) improve accuracy even for this sample emphasizing melodic similarity, this may suggest that allowing juries to hear full audio recordings without restricting them to sheet music depositions could actually help improve accuracy in legal cases. However, this hypothesis remains speculative until it can be more rigorously tested at larger scales (and the issue discussed above of determining "correct" decisions more thoroughly addressed).

The average results for each case shown in Figure 1 displayed a normal distribution for full-audio but were hourglass-shaped/bimodal for melody-only and lyrics-only. For the lyrics-only condition, this distribution reflects that most participants judged non-infringement for most cases, which is consistent with the fact that this sample was not selected to include many example of lyrics infringement. The melody-only condition led to higher accuracy for some cases (as predicted), but lower accuracy for others (contra predictions).

The accuracy of the PMI algorithm for the current study of 71% (12/17 cases) was slightly lower than the value of 80% (16/20 cases) reported in a previous study using a similar dataset. There are two reasons for this: 1) The sample was different – this study excluded 6 cases without matching full audio recordings and added 3 new Japanese cases (the new Japanese cases were not selected based on PMI values or any quantitative criteria, but they all were correctly classified by the PMI algorithm). 2) In the process of preparing controlled audio files for the experiment that were exactly matched, we noticed that several of the transcriptions used in [1] and [3] either did not exactly match the audio recordings, or had mismatched lengths.

Compared with the previous published study on PMI [3], the current PMI method successfully classified two case that were not accurately classified in the 2018 testing (*Three Boys Music v. Michael Bolton* and *Grand Upright v. Warner*), but three cases previously classified successfully now failed to be successfully classified (*Swirsky v. Carey*, *Granite Music v. United Artists*, and *Jean et al. v. Bug Music*). Two cases (*Selle v. Gibb* and *Fantasy v. Fogerty*) remained failures in both studies, but these two exceptions were not due primarily to a failure of the melodic similarity algorithm but rather to the complex nature of musical copyright law [3]. These discrepancies show how results from the PMI method can be affected by errors and uncertainties in the transcription process.

While the Musly algorithm resulted in the same overall accuracy as the PMI method (71%), 4 of the 5 mis-classified cases were different between the two methods. Both methods mis-classified *Fantasy v. Fogerty* as infringing when the court decision was non-infringement (see [3] for discussion of legal details). The four cases uniquely misclassified by the Musly but not PMI method largely seemed to be of the type predicted by the melody-centric view of copyright in which non-melodic similarities or differences interfered with assessment of melodic similarity. For instance, *Herald Square Music v. Living Music* showed low audio similarity via Musly despite high melodic similarity and a finding of infringement. In this case, the different timbres where one melody is performed by a saxophone with background noise while the other is sung by a vocalist with piano accompaniment seem to obscure similarities in the two melodies.

The fact that both algorithms failed for different sets of cases, and the fact that participants who made judgements only based on audio similarity without information about the historical/legal context performed even lower than the algorithms, suggests that the complexities of copyright law are difficult to fully capture through objective measurement of similarity alone. The relative emphasis on melody, lyrics, other musical aspects, and extra-musical legal factors changes from case to case, limiting the power of any single objective method. This supports previous caveats that, while objective quantitative methods may help supplement traditional qualitative analysis, "Trial by algorithm will never replace trial by jury, nor should it." [3].

8. FUTURE DIRECTIONS

The primary limitation of our study at present is its limited size and scope, with a dataset of only 17 court decisions (17 from USA) and perceptual ratings from only 20 participants. Furthermore, some of the cases include non-musical aspects that make it difficult for current automated methods focusing on musical similarity to identify those exceptions. Thus, we plan to expand the testing data by including more usable cases which have court decisions and have no non-musical factors that have affected the court decisions. Preliminary screening of the 238 cases at the Music Copyright Infringement Resource [14], we found 50 potentially usable court cases we plan to investigate in future studies. To increase diversity and cross-cultural generalizability, we also plan to identify more non-US cases, particularly from Japan and China where music industry revenues are substantial.

One promising direction may be to expand from a focus purely on music copyright infringement to also include the related domain of cover-song detection. Because there are larger databases and more sophisticated algorithms being developed for cover-song detection, these may provide more powerful methods that could be adapted to copyright infringement in future research [11-12].

9. DATA/CODE AVAILABILITY

Musical stimuli, data and analysis code are available at https://github.com/compmusiclab/music-copyright. The full experiment can be accessed at https://music.keio.moe/experiments/copyright/full.

10. AUTHOR CONTRIBUTIONS

Conceptualization: PES, CC, QDA, DM, SF, SO, YY; Methodology/ Analysis/ Investigation/ Visualization: YY, SO, PES; Project administration/ Supervision/ Funding acquisition: PES; Writing – original draft: YY, PES, SO; Writing – review & editing: CC, DM, QDA, SF.

11. ACKNOWLEDGMENTS

We thank our experiment participants; Delton Ding and Rei Konno for creating the experimental interface; Shoichiro Sato for assistance correcting the transcribed melodies; and John McBride for feedback on the manuscript. This work was supported by a Grant-In-Aid from the Japan Society for the Promotion of Science (#19KK0064) and by grants from Keio University (Keio Global Research Institute, Keio Research Institute at SFC, and Keio Gijuku Academic Development Fund) to PES.

12. REFERENCES

[1] D. Müllensiefen, and M. Pendzich, "Court decisions on music plagiarism and the predictive value of similarity algorithms," *Musicae Scientiae*, 13(1 Suppl), pp. 257-295, 2009.

[2] M. Robine, P. Hanna, P. Ferraro, and J. Allali, "Adaptation of string matching algorithms for identification of near-duplicate music documents," in *Proc. of the International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 2007)*, Amsterdam, Netherlands, 2007, pp. 37-43.

[3] P. E. Savage, C. Cronin, D. Müllensiefen, Q. D. Atkinson, "Quantitative evaluation of music copyright infringement," in *Proc. of the 8th International Workshop on Folk Music Analysis (FMA 2018)*, Thessaloniki, Greece, 2018, pp. 61-66.

[4] E. Selfridge-Field, "Substantial Musical Similarity in Sound and Notation: Perspectives from Digital Musicology," *Colorado Technology Law Journal*, vol. 16, pp. 249-284, 2018.

[5] *Lotus Development Corp. v. Borland Intern., Inc.*, 49 F.3d 807, 813 (1st Cir. 1995).

[6] J. P. Fishman, "Music as a Matter of Law," *Harvard Law Review*, vol. 131, no. 7, pp. 1861-1923, 2018.

[7] E. Selfridge-Field, "Conceptual and Representational Issues in Melodic Comparison," in *Melodic Similarity: Concepts, Procedures, and Applications*, Cambridge, MA, USA: MIT Press, 1998.

[8] J. Lund, "An empirical examination of the lay listener test in music composition copyright infringement," *Virginia Sports & Entertainment Law Journal*, vol. 11, pp. 137-177, 2011.

[9] *Williams v. Gaye*, 885 F.3d 1150, 1160 (9th Cir. 2018).

[10] P. E. Savage and Q. D. Atkinson, "Automatic tune family identification by musical sequence alignment," in *Proc. of the 16th International Society for Music Information Retrieval Conf. (ISMIR 2015)*, Malaga, Spain, 2015, pp. 162-168.

[11] J. Serrà, E. Gómez, and P. Herrera, "Audio cover song identification and similarity: background, approaches, evaluation, and beyond," in *Advances in Music Information Retrieval (Studies in Computational Intelligence, vol. 274)*, Z. W. Raś and A. A. Wieczorkowska (eds), Springer-Verlag Berlin Heidelberg, 2010, pp. 307-332.

[12] F. Yesiler, C. Tralie, A. A. Correya, D. F. Silva, P. Tovstogan, E. G. Gutiérrez, and X. Serra, "DaTACOS: A dataset for cover song identification and understanding," in *Proc. of the 20th International Society for Music Information Retrieval Conf. (ISMIR 2019)*, Delft, Netherlands, 2019, pp. 327-334.

[13] A. Flexer and T. Grill, "The problem of limited interrater agreement in modelling music similarity," *Journal of New Music Research*, vol. 45, no. 3, pp. 239-251, 2016.

[14] C. Cronin, *Music Copyright Infringement Resource*, Retrieved from https://blogs.law.gwu.edu/mcir/, 2020.

[15] *Arnstein v. Porter*, 154 F.2d 473 (2d Cir. 1946).

[16] M. Mandel and D. Ellis, "Song-level features and support vector machines for music classification," in *Proc. of the 6th International Society for Music Information Retrieval Conf. (ISMIR 2005)*, London, UK, 2005.

[17] D. Schnitzer, *Musly - an open-source audio music similarity library*, https://www.musly.org, 2014.

[18] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer, "Using Mutual Proximity to Improve Content-Based Audio Similarity," in *Proc. of the 12th International Society for Music Information Retrieval Conf. (ISMIR 2011)*, Miami, Florida, USA, vol. 11, 2011, pp. 79-84.

[19] H. Allan, D. Müllensiefen, and G. A. Wiggins, "Methodological Considerations in Studies of Musical Similarity," in *Proc. of the 8th International Society for Music Information Retrieval Conf. (ISMIR 2007)*, Vienna, Austria, vol. 6, no. 1, 2007, pp. 463-466.

Supplementary Materials

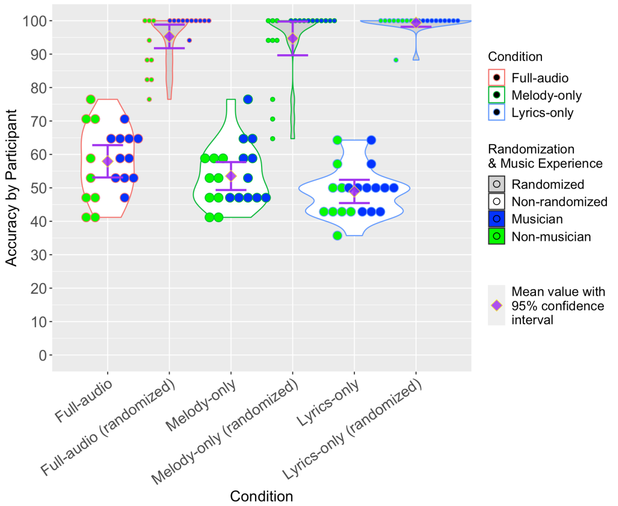

Figure S1. Accuracy for each of the 20 participants, as measured by the percentage of the 17 cases in which participant judgements of music copyright infringement matched court decisions. Dot plots and violin plots are drawn for 3 condition groups separately, where red outline represents full-audio condition, green represents melody-only, and blue represents lyrics-only. Randomized controls are shown beside the non-randomized ones in gray shading. Music experience of the participants is indicated by the filling color of dots; blue = self-reported musician; green = non-musician. Purple diamond dots and error bars represent means and 95% confidence intervals for each condition.

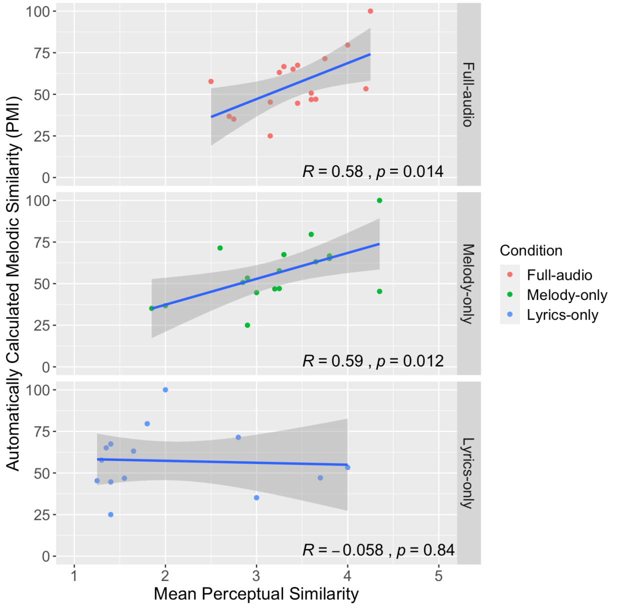

Figure S2. Mean perceptual similarity vs. automatically calculated melodic similarity (PMI) for full-audio, melody-only, and lyrics-only conditions for the 17 cases.

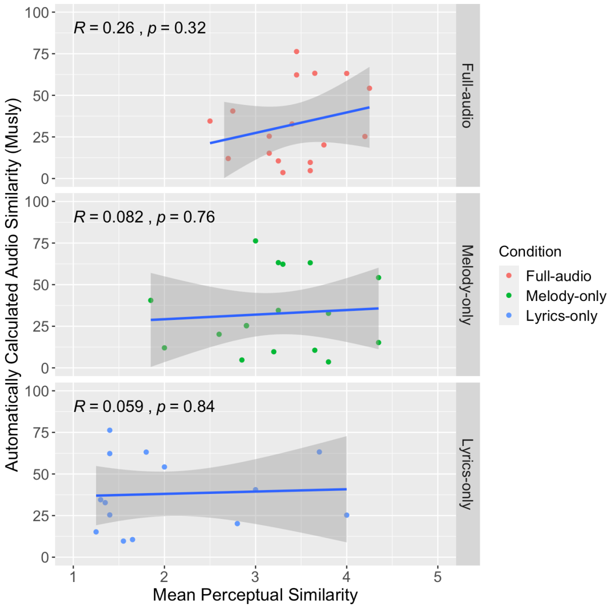

Figure S3. Mean perceptual similarity vs. automatically calculated audio similarity (Musly) for full-audio, melody-only, and lyrics-only conditions for the 17 cases.