Musicologica Olomucensia 29 – June 2019

Investigating Melodic Annotation Disagreements in String Quartets

Sarah Sauvé

How do we know a melody is a melody? This is probably not a question often contemplated when listening to music, as listeners just know a melody when they hear one. It is however a very relevant question when considering musical auditory scene analysis (ASA), as it can inform the process of melody extraction, or in other words, identifying the foreground of a musical scene. This task, which became very popular in the past few decades in the music information retrieval (MIR) community, consists of identifying the melody in a polyphonic music context, either from audio or symbolic data. The present research deals with symbolic data.

Defining Melody

When discussed in an informal setting, people usually understand each other as to what is meant by melody. It's one of those things that everyone understands without needing to put specifically into words. However, for the purposes of empirical research, a definition is needed. Drawn from music theory[1], melody can be defined as a succession of different pitch sounds brightened up by the rhythm. A more recent music dictionary, the New Grove Dictionary of Music and Musicians[2], similarly defines melody as a combination of a pitch series and a rhythm having a clearly defined shape. These are quite generic; a commonly employed definition of melody in audio MIR is that it is *the sequence of monophonic pitches that a listener might sing or hum when asked to reproduce a polyphonic piece of music, and encompasses the core identity of the piece*[3]. While still generic, this last definition allows for a "correct interpretation" through the identification of

[1] Ernst Toch, *The Shaping Forces in Music: An Inquiry Into the Nature of Harmony, Melody, Counterpoint, Form* (Courier Corporation, 1977), 69.

[2] David K. Rycroft and Stanley Sadie, *The New Grove Dictionary of Music and Musicians* (JSTOR, 1983), http://www.jstor.org/stable/30249775.

[3] Graham E. Poliner et al., "Melody Transcription from Music Audio: Approaches and Evaluation," *IEEE Transactions on Audio, Speech, and Language Processing* 15, no. 4 (2007): 1247.

melody by a listener. Beyond this general definition, melody has been broken down further into different types, as described by Selfridge-Field[4]:

- compound melodies describe melodies where some pitches are melodic and some are either another melody or an accompaniment; this is also called pseudopolyphony and is most common in solo string music
- self-accompanying melodies are melodies where some pitches act as both main theme and harmonic support, also another form of pseudopolyphony
- submerged melodies are melodies in inner voices of a polyphonic work
- roving melodies are melodies that move from part to part, or instrument to instrument, in an ensemble
- distributed melodies are melodies spread across various instruments and the theme cannot be represented by one part alone

Overall, these definitions are heavily biased towards Western ideas of melody in that it is assumed that there is only one such dominant line, characterized by pitch (as opposed to rhythm or timbre), that can be sung to represent a piece of music and that this line is monophonic (though doublings aren't especially rare in Western music, a melody is generally thought of as monophonic). One caveat to keep in mind is that it is not guaranteed that every listener will sing back the same line; currently, when there is disagreement, the most common interpretation is considered correct. Another typical assumption in the MIR field is that the melody cannot change instruments throughout the piece, which is appropriate and performs well for pop music but performs substantially worse for Western classical music, where in instrumental ensembles it is common for the melody to change instruments or rove, as defined by Selfridge-Field. Another related challenge is to identify whether there is a melody present at all, a problem called *voicing*[5]. For this research, the MIR definition of melody will be applied to annotate string quartet movements. This work has two main goals: 1) to create a new melody-annotated dataset for melody extraction evaluation and 2) to investigate melody annotation agreement and disagreement between listeners. Evaluation datasets for melody extraction vary widely, with datasets being relatively small in

[4] Eleanor Selfridge-Field, "Conceptual and Representational Issues in Melodic Comparison," *Computing in Musicology: A Directory of Research*, no. 11 (1998): 9–12.

[5] J. Salamon et al., "Melody Extraction from Polyphonic Music Signals: Approaches, Applications, and Challenges," *IEEE Signal Processing Magazine* 31, no. 2 (March 2014): 120. https://doi.org/10.1109/MSP.2013.2271648.

MIR terms (i.e. a few dozen short excerpts[6]). Datasets containing annotations indicating melody in a Western classical, instrumental context are few and far between due to the extensive time commitment involved in building such a dataset. The majority of melody extraction datasets assume that the melody is contained in one track, or instrument, which may be appropriate and perform well for popular types of music but would perform substantially worse for Western classical music, where in instrumental works it is common for the melody to change instruments or rove. As far as the author is aware, the only two datasets where the annotated melody is allowed to rove are MedleyDB[7] and OrchSET[8].

MedleyDB contains 122 songs, 108 of which contain melody annotations (the remaining songs were not considered melodic by the authors). It contains a wide variety of genres including rock, pop, classical, jazz, fusion, world, musical theater and singer-songwriter and is annotated for melody according to three definitions:

1) The predominant melodic line from one source, or instrument
2) The predominant melodic line from multiple sources, or instruments
3) All melodic lines from multiple sources, or instruments

The first produces an annotated melody that does not rove, while the second definition produces an annotated melody that can and the third produces an annotated melody that may have multiple voices at one time. These annotations were performed by monophonic pitch tracking algorithm pYIN and corrected by human annotators with at least a Bachelor of Music, producing three versions of the melody for each piece of music in the dataset corresponding to each of the three definitions above.

OrchSET is a collection of 64 audio excerpts accompanied by MIDI files containing the melody, as perceived by four listeners. Only excerpts in which all four listeners agreed on the melody were kept in the dataset. This dataset is entirely instrumental, including orchestral music from 15 composers spanning the late Baroque period to the 20th century.

These two datasets present important additions to the set of available evaluation datasets for melody extraction. This paper introduces a new dataset: Melody Annotated String Quartets (MASQ). MASQ is a dataset that aims to continue the expansion of available melody extraction evaluation datasets by providing

[6] Rachel M. Bittner et al., "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research.," in *ISMIR*, vol. 14, 2014, 155.

[7] Bittner et al., "MedleyDB."

[8] Juan J. Bosch, Ricard Marxer, and Emilia Gómez, "Evaluation and Combination of Pitch Estimation Methods for Melody Extraction in Symphonic Classical Music," *Journal of New Music Research* 45, no. 2 (2016): 101–117, https://doi.org/10.1080/09298215.2016.1182191.

melody annotations for string quartets, a genre not yet represented in the existing melody-annotated instrumental MIR datasets. The dataset currently consists of seven Wolfgang Amadeus Mozart and fourteen Franz Joseph Haydn string quartet movements. This paper will present the details of MASQ as well as an analysis of the disagreements between annotators, offering insight into individual differences and commonalities in melody perception.

Method

Data collection

Dataset. All Mozart and Haydn string quartets present on the KernScores website[9] in February 2017 were candidates for this dataset, a total of 372 movements. All movements not beginning with exactly four pitches, one in each instrument, were excluded as this restriction was relevant to the original use of the dataset.[10] Seven Mozart and fourteen Haydn movements were randomly selected from the remaining 97 movements. These are listed in Table 1.

Annotators. Each movement was annotated by three listeners, one of them always being the author (Annotator 2 in the PDF files). The other two annotators were musicians with formal study at the university level, mean age 25 (SD = 1.32), where each annotator had no more than 7 movements to annotate (see Table 1). Each annotator's primary instrument is indicated in Table 1, with 4 pianists, and one each violinist, organist, baritone and trumpet and French horn players.

Table 1. Summary of string quartet movements included in MASQ, along with annotator distribution

Composer	Work	Movement	Annotator 1	Annotator 2	Annotator 3
Mozart	K428	1	Violinist	SS (Pianist)	Pianist
	K428	2			
	K458	3			
	K464	2			
	K499	3			
	K575	2			
	K590	1			

[9] "KernScores," n.d., http://kern.humdrum.org/.

[10] Sarah A. Sauvé, "Prediction in Polyphony: Modelling Musical Auditory Scene Analysis" (Queen Mary University of London, 2018).

Haydn	Op. 1, No. 3	4	French horn	SS (Pianist)	Pianist
	Op. 1, No. 4	4			
	Op. 1, No. 6	2			
	Op. 9, No. 2	2			
	Op. 9, No. 3	3	Pianist		
	Op. 33, No. 3	3			Trumpet
	Op. 33, No. 4	1			
	Op. 50, No. 4	1			
	Op. 64, No. 3	1	Organist		
	Op. 64, No. 6	1			
	Op. 71, No. 2	2			Baritone
	Op. 76, No. 3	2			
	Op. 76, No. 5	4			
	Op. 77, No. 1	3			

Procedure. Each annotator received scores (downloaded from IMSLP[11], an online database of public domain musical scores and audio recordings) for their respective movements and were asked to mark or highlight the melody on the score while they listened to an audio recording of the movement. Scores were either manually annotated on a printed version that was scanned back to the author, or electronically (i.e. directly onto the PDF or similar). One Mozart quartet annotator described their selections in a table detailing the measures, beats, and instruments. Annotators were further instructed to always highlight only one note at any given time, and to leave no mark if they did not feel that there was a melody present. These instructions ensure that the melody is monophonic and that MASQ explicitly includes information about voicing.

Analysis

Data preparation. Using MuseScore[12] to visualize each MIDI file (downloaded from KernScores), each movement was edited so that only the melody marked by each annotator remained. Therefore, there are three versions of each movement; these MIDI files can be found online[13] alongside the original MIDI files for use by the research community as ground truth for melody extraction. Alongside

[11] "IMSLP," n.d., https://imslp.org/wiki/.

[12] *MuseScore*, n.d., https://musescore.org/en/download.

[13] Sarah A. Sauvé, "MASQ Dataset," April 23, 2019, https://github.com/sarahsauve.

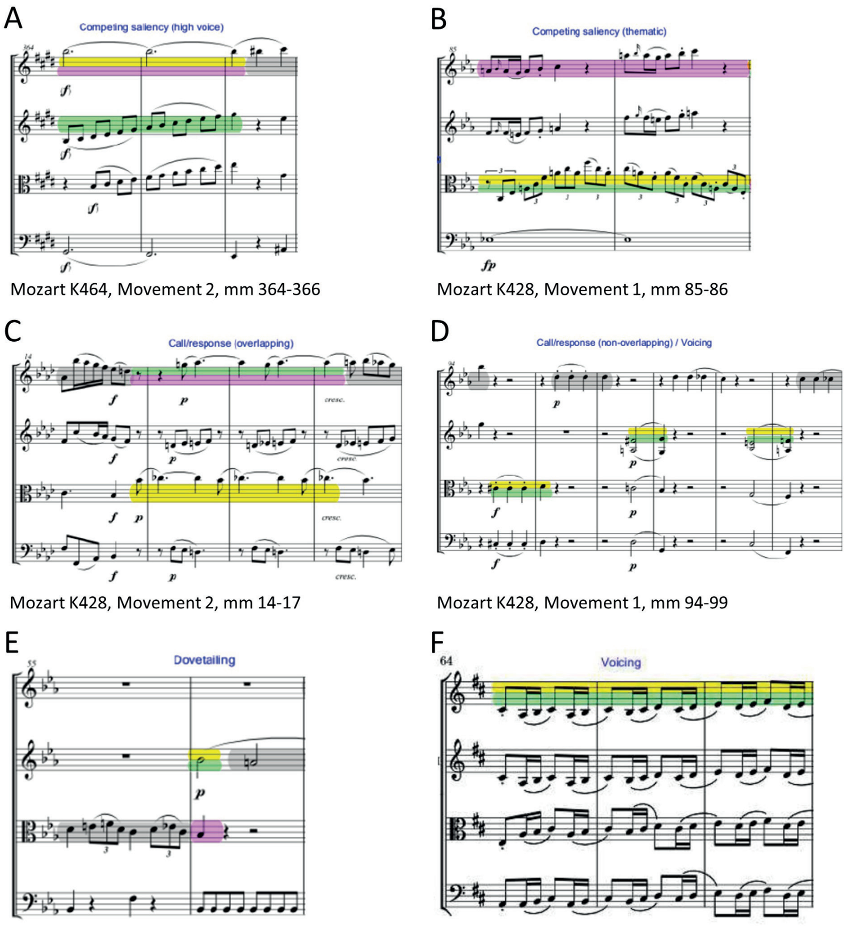
the MIDI files are annotated PDF files of the musical score. The melody is highlighted in light grey where all three annotators agree, and colour-coded by annotator where the annotators disagree (legend included in each score). Each disagreement is labelled with one or more categories defined by the author, as detailed below.

Visual data analysis. Each movement was analysed visually: comparing the three annotations, tabulating disagreements between them and sorting these disagreements into categories, which were determined based on the author's observation. These categories are:

- Competing saliency:
  - High voice: cases when thematic material is presented in a voice other than the highest voice (i.e. annotators may label the thematic material or the highest voice as melodic)
  - Thematic: cases when thematic material is presented simultaneously (i.e. annotators may differ in the material they label as melodic)
- Call/response
  - Overlapping: cases when a pattern is reprised in another instrument and the two iterations of the pattern overlap (i.e. annotators may differ in which iteration they label as melodic)
  - Non-overlapping: cases when patterns that may or may not differ slightly are passed between instruments without overlap (i.e. annotators may label the call or response as melodic but not both; often co-categorized with voicing, though not always)
- Dovetailing: the end of one phrase overlaps with the beginning of a new phrase (i.e. annotators may differ in which phrase they label as melodic)
- Voicing: cases where annotators disagree on whether a melody is present or not

Disagreements that did not fall into any of these categories are labelled 'Other'. Where more than one type of disagreement was present in a measure, both were labelled so that one measure may contain more than one label. Figure 1 illustrates examples of each of these categories.

Statistical data analysis. In addition to summary statistics, the primary statistical test used was the chi-square test, where a significant result is obtained when categories do not contain equal numbers of measures. Chi-square tests were performed to compare: (1) disagreement categories (all sub-categories and general categories); (2) composer; and (3) movement type (i.e. first, second, third or fourth movement), where raw number of disagreements were used in the first case and mean number of disagreements per movement was used in the second and third cases. One-sided binomial tests confirmed whether each individual category was significantly different than expected for a random distribution of disagreement frequencies between categories.


A
Competing saliency (high voice)
364
B
Competing saliency (thematic)
85
fp
C
Call/response (overlapping)
14
cresc.
D
Call/response (non-overlapping) / Voicing
94
E
Dovetailing
55
F
Voicing
64

Mozart K464, Movement 2, mm 364-366

Mozart K428, Movement 1, mm 85-86

Mozart K428, Movement 2, mm 14-17

Mozart K428, Movement 1, mm 94-99

Mozart K428, Movement 1, mm 56

Haydn Op. 76, No. 5, Movement 4, mm 64-66

Figure 1. Illustrative examples of the seven different types of disagreement categories in visual data analysis: competing saliency high voice (A) and thematic (B), call/response overlapping (C) and non-overlapping (D), dovetailing (E) and voicing (F). Each annotator is assigned a colour: A1 = yellow, A2 = green, A3 = pink; grey represents annotator agreement.

Results

Visual analysis

A summary of disagreements is presented in Table 2, where the two most common types of disagreements are *voicing* and *competing saliency – high voice*, representing 47.7% and 34.5% of disagreements respectively. All other disagreement categories each represent less than 10% of disagreements. Further observations will be presented in the Discussion section below.

Statistical analysis

The mean number of measures per movement containing disagreements was 31.1 (SD = 25.6), corresponding to a mean percentage of 25.8% (SD = 11.5) of each movement. Excluding the five movements with the highest percentage of disagreements, where most of the disagreements in each movement are attributable to a single annotator's systematic differences, the mean percentage of disagreements falls slightly to 21.2% (SD = 8.4), just under one quarter of each piece.

As described above, chi-square tests were employed to detect whether the distribution of annotation disagreements amongst categories is different than expected if these were randomly distributed. Annotation disagreement distribution by category (by overall category and including sub-categories) was significantly different than a random distribution, χ^2 (4) = 522.31, p < .0001 and χ^2 (6) = 758.36, p < .0001 respectively. One-sided binomial tests confirm that disagreements in each disagreement category (overall or sub-categories) were significantly different from chance, all p < .0001. There was no significant difference in the number of disagreements between composers, χ^2 (1) = 0.02, p = 0.86 but there was between types of movements, χ^2 (3) = 13.81, p = .003. One-sided binomial tests on the mean number of disagreements for each type of movement reveal that only *second* movements contain significantly less disagreements than chance, p < .0001, while the number of disagreements for all other movement types are not different from chance, p > .05.

Table 2. Summary of annotator discrepancies by disagreement category. Mozart movements are identified by their K catalogue number followed by movement number while Haydn movements are identified by their Opus number, quartet number, and movement number. Note that the *Total* column does not equate to the sum of the row; this is due to some discrepancies belonging to more than one category.

Movement	Frequency by disagreement category (number of measures)							Summary	
	Call/response		Competing saliency		Dovetailing	Voicing	Other	Total	Percent of work
	Overlapping	Non-over-lapping	High voice	Thematic					
K428-1	2	6	18	6	6	14	0	44	26.8
K428-2	10	0	8	0	6	5	0	27	28.1
K458-3	5	0	7	0	4	3	0	13	24.5
K464-2	0	0	11	12	0	7	0	26	25.0
K499-3	3	6	27	0	1	5	1	36	34.2
K575-2	0	0	4	0	1	3	0	7	10.9
K590-1	0	42	11	0	10	53	2	71	35.8
Op. 1, No. 3–4	0	0	24	0	0	0	0	24	46.1
Op. 1, No. 4–4	0	0	3	2	0	0	0	5	8.6
Op. 1, No. 6–2	0	0	0	12	0	1	0	13	23.2
Op. 9, No. 2–2	0	0	3	0	0	0	0	3	7.1
Op. 9, No. 3–3	0	0	17	0	4	18	1	36	49.3
Op. 33, No. 3–3	0	2	7	0	0	25	1	33	36.2
Op. 33, No. 4–1	0	0	24	0	0	2	0	25	28.8
Op. 50, No. 4–1	2	0	15	14	1	16	0	44	23.7
Op. 64, No. 3–1	5	0	9	4	2	5	0	24	14.0
Op. 64, No. 6–1	3	4	15	0	0	5	0	26	18.0
Op. 71, No. 2–2	0	0	6	2	1	16	3	26	33.7
Op. 76, No. 3–2	0	0	4	4	0	8	6	18	15.3
Op. 76, No. 5–4	0	0	14	0	2	106	1	121	35.3
Op. 77, No. 1–3	0	0	3	3	0	21	6	33	18.2
Total measures	30	60	230	59	38	313	21	M = 31.1 (SD = 25.6)	M = 25.8 (SD = 11.5)
Percent of discrepancies	4.5	9.0	34.5	8.8	5.8	47.7	3.2		

Discussion

In this section, observations related to patterns of disagreements will be elaborated upon. Firstly, while the *voicing* and *competing saliency – high voice* categories remain important, it is important to acknowledge the influence of two annotators on the especially high prevalence of these two types of disagreements. In the first case, the baritone annotator is the source of much of the *voicing* disagreements, particularly in Haydn's Op. 76, No. 5, Movement 4, where they often did not label a melody when the other two annotators did. Removing those 106 disagreements from the voicing category still leaves the category representing 31.6% of total disagreements. In the case of the *competing saliency – high voice* category, the pianist annotator for the middle group of Haydn movements tends to label the lower voices as melodic while other annotators prefer the first violin. This is the case in all movements they annotate. However, if all these disagreements were ignored, *competing saliency – high voice* still accounts for 25.4% of total disagreements.

The form of the movement also has an influence on the number and type of disagreements observed. For example, minuets tend to have low disagreement overall, particularly Haydn's Op. 1, No. 4, Movement 4 and Op. 9, No. 2, Movement 2, where annotators disagree for less than 10% of the movement. One exception to this is Haydn's Op. 1, No. 3, Movement 4, where the trio portion consists almost entirely of disagreements. It is particularly rare in this dataset that all three annotators label a different instrument as melodic; however, this is the case throughout almost this entire trio. While the first annotator almost always labels the most rhythmically active instrument (the second violin or the cello), the second and third annotators label the sustained notes but at different octaves (Figure 2). This is perhaps unsurprising as the minuet form is relatively simple with clear melody and accompaniment parts and sparse texture. Indeed, the lowest rate of disagreement amongst the Mozart movements is found in K575, Movement 2, a slow movement with a high degree of rhythmic synchrony and sparsity. Disagreements are usually restricted to partial measures, with one case where a disagreement spans two measures (mm16–17).

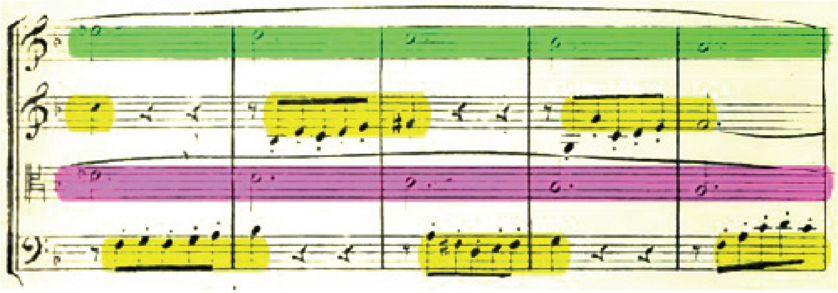


Figure 2. Haydn Op. 1, No. 3, Movement 4, mm 38–42. In this trio, all three annotators label different instruments as melodic: the first label the most rhythmically active voice while the second and third label the sustained notes at different octaves. Yellow = A1; Green = A2; Pink = A3.

One more noteworthy form with high rates of agreement is the theme & variations of Haydn's Op. 76, No. 3, Movement 2. With relatively few disagreements, these are mostly restricted to *voicing* disagreements and segments where the theme is doubled (*competing saliency – thematic*). Still, disagreements are somewhat surprising as in this theme & variations, the theme is explicitly repeated in each variation, with the accompanying material providing variation. Thus, one would expect that annotators would continue to perceive the theme as the primary melodic content, particularly since all annotators agreed that the first presentation of the theme, as well as its first repetition, were melodic. Perhaps annotators were still aware of the theme throughout but grew bored of it and were temporarily drawn to other melodic lines (i.e. Variation III). That being said, the annotators agree that the original theme remains melodic throughout the vast majority of the movement.

Let us now turn to look more closely at movements with particularly high rates of disagreement. The three movements with the most disagreements are Haydn's Op. 9, No. 3, Movement 3, Op. 1, No. 3, Movement 4 and Op. 33, No. 3, Movement 3. The source of the majority of disagreement in these have already been discussed: all these movements were annotated by one of the pianists, where in those annotations the lower voices were labelled as melodic where other annotators preferred higher voices. The movement with the next highest rate of disagreement is Mozart's K590, Movement 1. Here, the majority of disagreements fall under the *voicing* and *call/response – non-overlapping* categories. In these disagreements, the third annotator systematically labels only the 'call' portion of the call-response figure illustrated in Figure 3 while the other two annotators label both the 'call' and the 'response'.

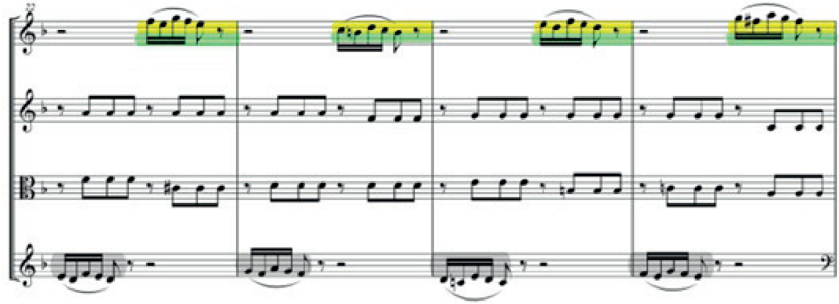


Figure 3. Mozart's K590, Movement 1, mm 22–25. Call/response pattern between the cello and the first violin, where third annotator TB marks only the call portion of the figure (cello). This disagreement accounts for half the disagreements in this movement. Yellow = A1; Green = A2; Grey = all annotators.

The only movement to contain all types of disagreements is Mozart's K428, Movement 1. With lots of interesting thematic material throughout, areas of sparsity and overlapping phrase beginnings and endings, it offers plenty of opportunity to hear different lines as melodic, particularly when restricted to a monophony definition. An area of concentrated disagreement is mm 123–132, a part of the recapitulation (the last portion of classic sonata form: *exposition*, where thematic material is presented; *development*, where this material is manipulated and developed; and *recapitulation*, where the theme is reprised). In this excerpt (Figure 4), a descending 8th-note pattern is repeated in the viola and the cello, but not all annotators followed this pattern, hearing instead the highest voice as melodic. In mm 127, perhaps it is the opening rhythmic figure of the measure that attracted the attention before continuing with the descending 8th-note pattern and becoming the highest voice in the ensemble until mm 129. Measures 130–132 present a pattern seen elsewhere in the movement, where two annotators label the descending 8th-note pattern as melodic while the other labels the highest voice.

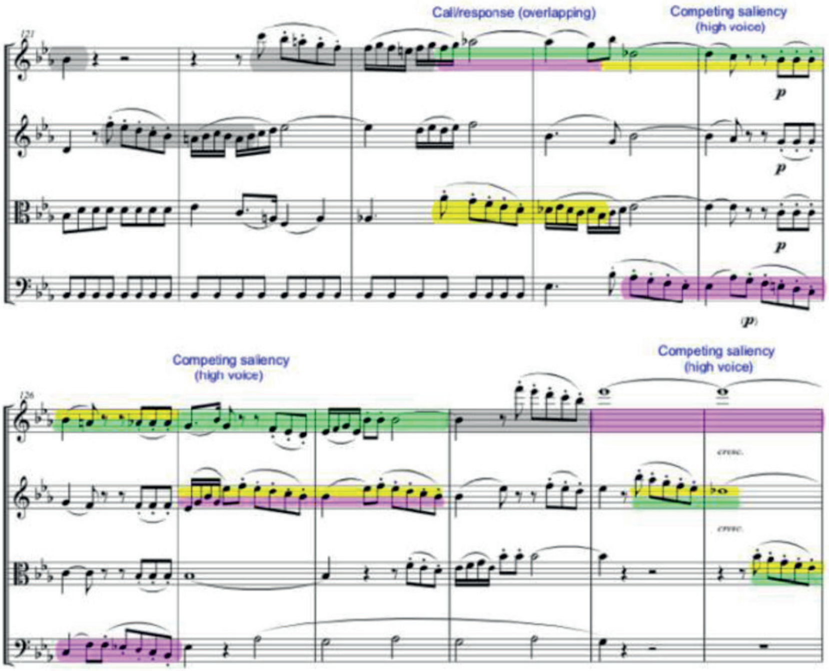

Call/response (overlapping)
Competing saliency (high voice)
121
p
p
p
(p)
Competing saliency (high voice)
Competing saliency (high voice)
126
cresc.
cresc.

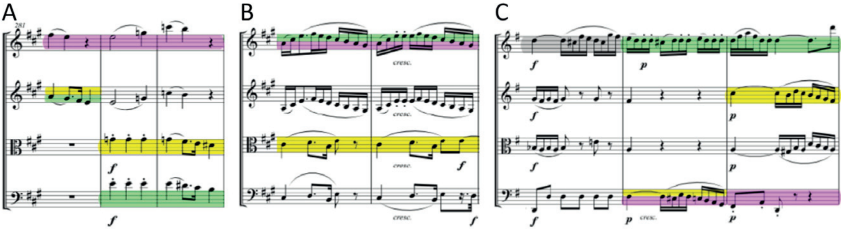
Figure 4. Mozart's K428, Movement 1, mm 123–131. The largest concentration of disagreements in the movement, this excerpt is part of the recapitulation of this piece. Yellow = A1; Green = A2; Pink = A3; Grey = all annotators.

Finally, it is worth noting that a number of disagreements may have been directly caused by the instructions given to the annotators, notably to only highlight one note at any time. Cases of a canon (overlapping instances of thematic material, e.g. Figure 5C) or thematic material presented in thirds or octaves (e.g. Figure 5A) sometimes leads to annotator disagreements when it is entirely possible that both parts of the canon were perceived simultaneously and that the harmonized voices were perceived as one. This provides evidence that a monophonic definition of melody does not suit all types of music, particularly in Western classical styles. Future annotation projects should allow listeners to mark an unrestricted number of simultaneous notes in order to capture a more sophisticated definition of melody. It would be expected that the percentage of disagreements between listeners might decrease and it would be interesting to see where remaining disagreements exist, offering additional insight into the perception of melody in polyphonic musical contexts. A melody annotated dataset with multiple versions of melody could also be useful to develop an algorithm that potentially extracts more than one possible melody, ranked in order of likelihood or preference. The regular presence of disagreements throughout the present dataset support this type of melody extraction approach, as it would be uncommon for all listeners to perceive the melody identically in instrumental music such as the string quartet.

High-voice superiority

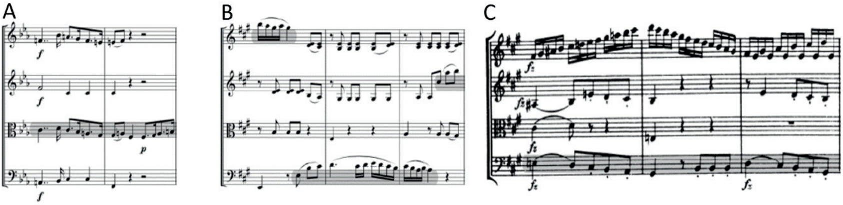
On the subject of melody extraction, it is worth mentioning the important role of high-voice superiority in human perception. Indeed, the earliest method of melody extraction involved systematically selecting the highest pitch throughout the piece and labelling it as the melody. This is known as the skyline algorithm[14]. Though this method works well for popular and folk music, it will reach a ceiling performance for Western classical instrumental music, as the melody does not always correspond to the highest pitch at any given time. The MASQ dataset provides examples of the high-voice superiority effect in action, where thematic material is overlooked in favour of the highest voice; and of where the skyline algorithm would fail. Figure 5 illustrates a few examples of the former and Figure 6 illustrates a few examples of the latter. In the three excerpts seen in Figure 5, thematic material is located in the lower voices (cello in 5A and 5C and viola in 5A and 5B). While some annotators label this thematic material as melody, some annotators instead label the highest voice as melodic, thus demonstrating the high-voice superiority effect. In Figure 6, examples of instances where the perceived melody does not correspond to the highest pitch are given, where in all of these cases the melody is played by the two lower voices in the ensemble, the viola (6A) and the cello (6B and 6C). Here all three annotators agree, demonstrating that there is a need to refine the skyline algorithm to allow the opportunity to fully reflect melody extraction as it is perceived by human listeners.

[14] Alexandra L. Uitdenbogerd and Justin Zobel, "Manipulation of Music for Melody Matching," in *Proceedings of the Sixth ACM International Conference on Multimedia* (ACM, 1998), 237.



K464, Movement 2, mm 10–12 K575, Movement 2, mm 16–17 K499, Movement 3, mm 72–74

Figure 5. Examples of the high-voice superiority effect in Mozart string quartet movements. In each of these excerpts, thematic material can be found in a voice other than the highest (cello in A and C and viola in A and B), but some annotators still label the highest note as melodic (violin I in all three excerpts). Excerpt C also features a canon, where thematic material overlaps in mm 73–74, also causing some disagreement between annotators. Yellow = A1; Green = A2; Pink = A3; Grey = all annotators.



Mozart K428, Movement 1, mm 51–52 Mozart K575, Movement 2, mm 21–23 Haydn Op. 50, No. 4, Movement 1, mm 124–126

Figure 6. Examples of where the skyline algorithm would fail from Mozart string quartets K428, Movement 1, mm 51–52 (A) and K575, Movement 2, mm 21–23 (B) and Haydn's Op. 50, No. 4, Movement 1, mm 124–126 (C). In each case, the melody is perceived as not being the highest pitch and is agreed by all annotators. Grey = all annotators.

Conclusions

This paper has achieved two goals: 1) presented a new dataset for use as ground truth in the melody extraction task and 2) presented an analysis of melody annotation disagreements in this dataset. MASQ contains twenty-one string quartet movements by W. A. Mozart and F. J. Haydn, where each movement has been annotated by three musician listeners. MIDI files containing the melody of each movement as annotated by each listener and accompanying PDF files highlighting disagreements can be found online to be used freely. This dataset adds to the existing body of annotated datasets for Western classical music that allow the melody to rove between voices throughout a piece of music, with the ultimate aim being to provide such annotations for all string quartets found on the KernScores website.

This paper additionally shared an analysis of annotator disagreements across movements, finding that average disagreement rate was 25.8%, around a quarter of each piece of music annotated. Disagreements were sorted into seven categories: *competing saliency – high voice*, *competing saliency – thematic*, *call/response – overlapping*, *call/response – non-overlapping*, *dovetailing*, *voicing* and *other*. The two most common categories were *voicing* and *competing saliency – high voice*, accounting for 47.7% and 34.5% of disagreements respectively. Though a single annotator sometimes explained up to half of these disagreements, these two categories remain dominant overall. The prominence of voicing disagreements in these annotations highlight the voicing sub-component of melody extraction as important in identifying the correct, or most commonly perceived, melody in any piece of music. On the other hand, the prominence of the *competing saliency – high voice* category demonstrates the high-voice superiority effect, where perception is drawn to the highest voice, regardless of whether or not it contains thematic material. That being said, the skyline algorithm cannot be relied upon entirely and it is important to consider cases where the melody is not located in the highest voice (Figure 6), especially in orchestral and ensemble Western classical music. Finally, the frequency of disagreements suggests a more fine-grained approach to melody extraction and to the empirical definition of melody, potentially allowing for multiple versions of the melody, ranked by preference.

Investigating melodic annotation disagreements in string quartets

Abstract

The article presents an analysis of melody annotation disagreements in a novel dataset containing annotations of a selection of Haydn and Mozart string quartet movements. For this purpose the following definition of melody from the music information retrieval (MIR) community is applied: the sequence of monophonic pitches that a listener might sing or hum when asked to reproduce a polyphonic piece of music, and encompasses the core identity of the piece. The resulting collection of annotations makes up the new Melody Annotated String Quartets (MASQ) dataset, available online. The rates and types of disagreements between annotators are discussed, as well as the influence of musical form and style on melody perception and the suitability of the given definition of melody.

Výzkum nesouladu v melodickém anotování smyčcových kvartet

Abstrakt

Článek prezentuje analýzu nesouladu v anotování melodií v nové sadě dat obsahující anotace výběru vět z Haydnových a Mozartových smyčcových kvartetů. Pro tento účel byla aplikována následující definice z komunity MIR (music information retrieval): sled monofonních výšek, které může posluchač zpívat nebo vokalizovat na brumendo, je-li požádán o reprodukování polyfonní skladby, a zahrnuje jádrovou identitu skladby. Výsledná sbírka anotací tvoří novou databanku melodicky anotovaných smyčcových kvartetů (Melody Annotated String Quartets (MASQ) dataset), jež je dostupná online. Předmětem diskuse je míra nesouladu mezi anotujícími a jeho typy, stejně jako vliv hudební formy a stylu na vnímání melodie a vhodnost dané definice melodie.

Keywords

annotation; melody; perception; string quartet

Klíčová slova

anotace; melodie; vnímání; smyčcový kvartet

Sarah Sauvé
Department of Community Health, Memorial University of Newfoundland
St. John's, NL A1C 5S7
Canada
sarah.sauve@mun.ca