COMPARING AUDIO BOUNDARY ANNOTATION OF VOCAL POLYPHONY: EXPERTS, NON-EXPERTS, AND ALGORITHMS

Mirjam VISSCHER (m.e.visscher@uu.nl) [1] and **Frans WIERING** (f.wiering@uu.nl) [1]

[1] *Department of Information and Computing Sciences*, **Utrecht University**, Netherlands

ABSTRACT

It is a challenging computational problem to perform segmentation on vocal polyphony from the Renaissance and early Baroque. In this genre, boundaries between segments are often hidden by overlapping voices. To test algorithms for segmentation, we need boundary annotations by humans as a ground truth, but experts in this field are rare and short on time.

Our study aims to evaluate the effectiveness of segmentation algorithms on vocal polyphony using both expert and non-expert annotations. For this, we collect boundary annotations by human experts and non-experts on polyphony. Then, we compare the annotations by the two groups to see whether we can use segmentations by non-experts instead of experts. Finally, we use the expert annotations to evaluate different segmentation algorithms from the MSAF library by Nieto and Bello.

The results show that the performance of non-experts comes quite close to that of experts, whereas the tested algorithms are not yet able to perform the task at a similar level. We conclude that non-expert annotations are adequate to act as ground truth for evaluating boundary detectors on vocal polyphony and we present next steps to create a larger dataset for such evaluations.

1. INTRODUCTION

The rich amount of computational methods developed in more than 25 years of MIR research presents a fascinating opportunity for music research. Can these methods be used to gain a deeper understanding of musical structure, music history, and musical cultures around the world? This raises the question of how generic those methods are, since the majority have been developed for and tested on contemporary popular music. How does this affect their performance or even validity when applied to different musics?

One emerging line of computational music research focuses on historical processes in Western music [1–3]. Within this broad area, our work concentrates on late Renaissance and early Baroque polyphony. Polyphony proves to be particularly challenging due to the relative homogeneity of the textures and the overlapping of voices. Also, during this period, music underwent many changes along a number of dimensions, one of which is tonal structure. Many musicologists claim that the modal system (based on melodic relationships) was replaced by modern harmonic tonality (based on chord relationships) during the 16th century. Others have different views [4]. Such conflicting accounts are based on the detailed analysis of music theorists and selected key compositions. It would be interesting to investigate if such a close reading of individual items could be complemented by a distant listening approach [5], and if this would shed a new light on the historical developments of the time.

When conducting research on tonal structure, the most logical approach would seem to be to analyse symbolic encodings. But since these are rather scarce (see e.g. figure 1 in [3]) and moreover employ a variety of encoding formats, it makes sense to turn to audio recordings instead, of which there are a much higher number thanks to the wide interest in early music from the 1960s onward. It is the aim of our CANTOSTREAM project [1] to investigate to what extent meaningful historical patterns can be found in audio recordings with the help of MIR methods.

A first step towards this this long-term aim is to divide each composition automatically into sections at meaningful boundaries. But what are those meaningful boundaries? This research presents a method to collect perceptual data from expert and non-expert listeners, delivering a dataset of boundary annotations, and evaluates six boundary detectors to gain an understanding of their potential for the task. Specifically, the research questions are:

1. Can we use non-expert annotations of boundaries to improve segmentation algorithms for vocal polyphony?
2. How does a selection of current boundary algorithms perform on vocal vocal polyphony?

We will show that it is possible to collect boundary annotations for retrieval experiments, employing a minimalist and economic setup.

2. BACKGROUND

2.1 Boundaries in Vocal Polyphony

Boundaries are points in music that provide a sense of closure to a section of a composition. In vocal polyphony, it is important to make a distinction between boundaries as a section closure on the one hand, and cadences as described

[1] https://www.uu.nl/en/research/interaction/music-information-computing/projects

Algorithm	Full name	Main characteristics
Foote [6]	Foote	Novelty
CNMF [7]	Convex non-negative matrix factorization	Timbre, clustering decomposition matrices
OLDA [8]	Ordinal linear discriminant analysis	Chord, timbre, pitch, timing
Scluster [9]	Laplacian segmentation, spectral clustering	Local timbre and long-term repetition
SF [10]	Structural features	Harmonic pitch class profiles; novelty, homogeneity, repetition
VMO [11]	Variable Markov Oracle	Harmony, timbre

Table 1. MSAF algorithms, as named in the library

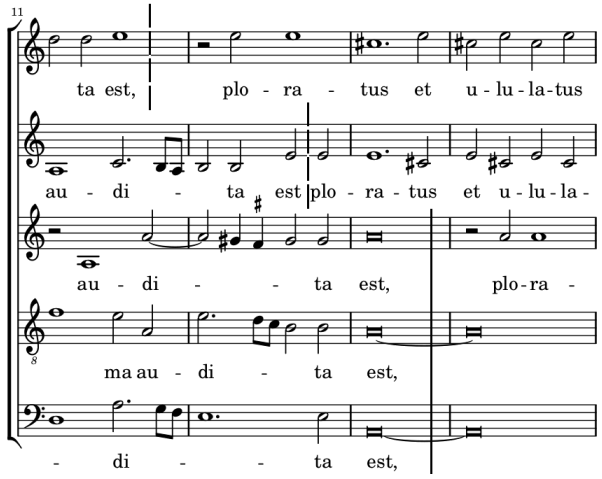

Figure 1. A polyphonic cadence in Giaches de Wert, Vox in Rama, mm 11-14. The vertical line in bar 13 indicates the cadence, the dotted lines indicate the individual endings.

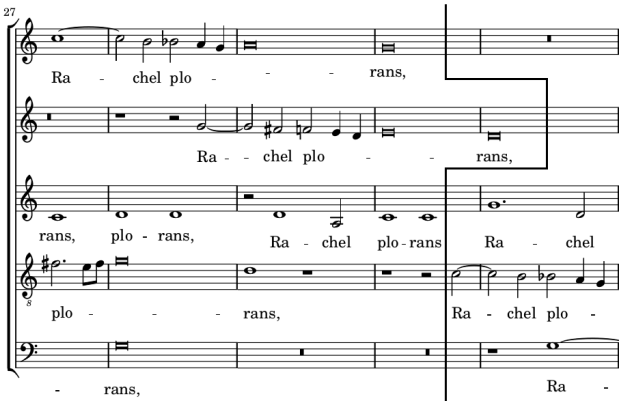

Figure 2. A boundary without cadential patterns in Giaches de Wert, Vox in Rama, mm 27-31.

in music theoretical sources on the other [12–14]. Boundaries can be formed by using a number of compositional techniques, such as change of texture, change in the combination of voices, change of meter, rests and cadences. Figure 1 provides an example of a typical polyphonic cadence in a motet by Giaches de Wert. The three lowest voices form a cadence on A in measure 13, while the two upper voices already have introduced the next phrase. This boundary is easy to recognise by ear, but most algorithms studied in this paper miss this boundary. An important part of the composer's art is to play with cadential structures to manipulate the listeners' expectations of closure. Not all cadence-like patterns are therefore strong boundaries, or even boundaries at all. Conversely, De Wert provides us with an example of a boundary without cadential elements in figure 2. Yet the majority of the participants in our experiment perceive a boundary between the first *Rachel plorans* and its repetition, as indicated by the solid line.

2.2 Automated Boundary Detection in Musical Audio

Boundary detection is usually regarded as the initial stage in audio-based music structure analysis, followed by structural grouping [15]. Three primary approaches were described by [16], reviewing the state of the art up to 2010, repetition-based, novelty-based, and homogeneity-based, to which a recent overview [17] has added regularity-based approaches. Additionally, [17] presents three main challenges in the field, namely subjectivity, ambiguity and hierarchy, affecting the annotator, the annotations and the musical structure respectively. Of these, subjectivity is the most important for our paper.

Both [16] and [17] present a comprehensive overview of segmentation methods. In our research, we focus on the subset of these methods provided by [15] which are implemented in their open-source music structure analysis framework (MSAF). [2] Table 1 displays the six boundary detection algorithms from this library, with their main characteristics.

2.3 Annotation of Musical Boundaries

There is a long history of experimentation with boundary perception in music [18–23]. From these experiments we learn a number of things. All authors recommend the use of surveys to gauge the expertise of their participants, for example the recent Goldsmiths Musical Sophistication Index [24]. However, this survey does not ask about experience with music from a specific period or genre.

Most authors used some form of tool support in their experiments. Recently, Bedoya [23] proposed the CosmoNote platform for the study of musical prosody, one aspect of which is segmentation (scored on a scale of 1-4). This platform is directed at citizen scientists creating their annotations online. They are supported by various visualisations and can review and correct their annotations.

[2] https://github.com/urinieto/msaf/

Composer	Work	Muziekweb.nl id
Luca Marenzio	Zefiro Torna (1584)	DBX12531-9
Orlande de Lassus	Pater Noster (1573)	DBX10350-13
Giovanni Pierluigi da Palestrina	Vergine tale è terra (1581)	DBX0438-37
Orlande de Lassus	Pauper sum ego (1573)	DBX0961-4
Orlande de Lassus	Justorum animae (1582)	DBX8123-20
Giovanni Pierluigi da Palestrina	Osculetur me osculo oris sui (1584)	DBX0438-1
Claudio Monteverdi	Or che 'l ciel e la terra (1638)	DBX12134-2
Giaches de Wert	Vox in Rama (1581)	DBX12064-4

Table 2. Recordings used for the experiment.

The importance of clear instructions for the participants is emphasised in [21]. It is important to hide visual cues that can influence the placement of boundaries [20, 21].

In such experiments, participants do not indicate a boundary at the exact same time although they respond to the same stimulus. Therefore, all authors use a window of synchrony [25]: "An interval of time over which response events are counted as occurring together", henceforth called window. Most boundary experiments set their window at +/- 3 seconds [15, 23], while [25] sets it at 2 seconds for the task of annotating arousal and valence.

3. METHOD

In our experiment, we collect boundary annotations by letting participants tap. We use these annotations to evaluate selected algorithms. The workflow of the experiment, as visualised in Figure 3, starts with the selection of compositions, continues with collecting annotations, and applying boundary detectors. Then, the workflow divides into two strands. The left strand is a cluster analysis to answer the question whether we can use non-expert annotations of boundaries to improve segmentation algorithms for vocal polyphony. The right strand evaluates the performance of non-experts and the MSAF algorithms respectively.

3.1 Selection of Musical Works

We select eight recorded performances of vocal polyphonic works composed around 1600. Table 2 lists the works, with a total duration of 36 minutes. We chose vocal music to keep timbre similar. There is a variety in polyphonic versus homophonic writing, in ambiguity in the boundaries, in the estimated difficulty to annotate the boundaries, in the tempo, and in the amount of voices. We select the performances for their restraint and professional level of the performers.

3.2 Preliminary Experiments

On three important elements of the experiment, the literature provides no conclusive recommendations. Therefore, we conduct preliminary experiments. First, we test boundary annotation on a printed score. This task took the expert participants about 30 minutes for 5 minutes of music. Some of the boundaries indicated seemed non-intuitive when compared to the perception of the performance, due to a music theoretical approach suggested by the presence of a score.

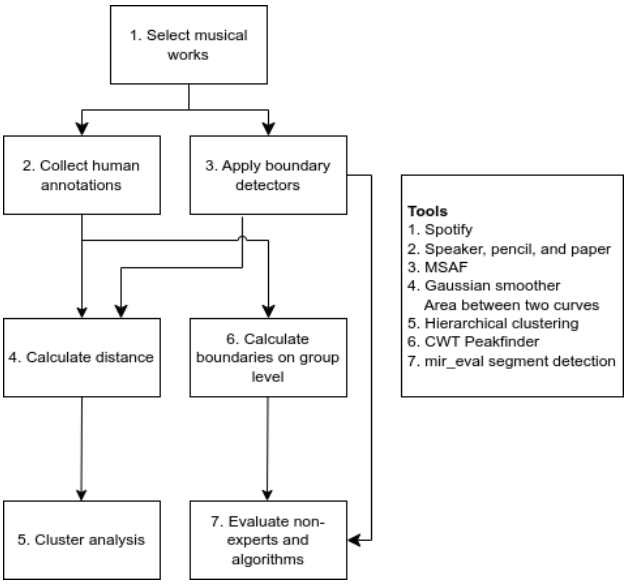

Figure 3. Workflow and tools used for this paper. The order corresponds to the order of the method.

As an alternative, we investigate a combination of audio and touch. Specifically, we test which type of touch would be most intuitive for a listener to use as an indicator of boundary strength in real-time: duration, pressure, amount of fingers pressing, or release of the touch when there is a boundary. Duration was perceived as the most intuitive method.

Finally, we test the process with participants having different levels of musical experience, ranging from no experience at all to several decades of professional experience. The boundaries indicated by all participants suggest a general agreement on the location of the boundaries, with varying preferences for granularity.

3.3 Collection of Annotations

For the main experiment, we ask the participants about their years of musical training, practice, and experience with Renaissance music using a custom survey. [3]

[3] The survey, a summary of the answers, the text used for the instructions, the anonymised annotation data

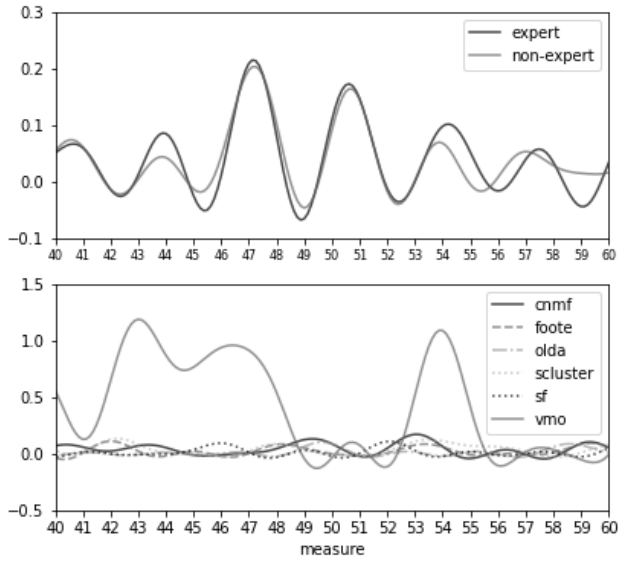

Figure 4. Gaussian curves of boundaries by participants and algorithms for Palestrina, *Vergine tale è terra* mm 40-60.

The participants listen to the music without looking at the score, and they tap the hand of the first author if they perceive a boundary, longer taps for strong boundaries, shorter taps for weak boundaries using a scale ranging from 1 to 4, and a resolution of one quarter note. The researcher records the boundaries and their weights on the score.

Prior to the experiment, we calibrate the strength of the boundaries: the first author asks the participant to tap the weights of 1, 2, 3, and 4 subsequently and provides feedback if the lengths are not clear. We use the first work to test and potentially correct the granularity preferences of the participant. We observed that each participant kept at roughly the same level of granularity and strategy during the session.

We decide not to give the participants the opportunity to re-listen and correct their annotations. First, because we are interested in a immediate, perceptual experience, whereas correction of annotation involves a analytical mode of thinking. Secondly, to keep the duration of the experiment within acceptable limits while still being able to annotate multiple compositions.

One annotation session is observed by the second author, with a focus on validity risks in music information retrieval [26]. The main observations are that the annotator follows the tapping of the participant accurately and without hesitation, and does not appear to steer the participant in a particular direction, whether consciously or unconsciously.

The second author provides reference annotations from a music analytical perspective, based on the score. He focuses on high- and mid-level structures, taking into account compositional techniques as well as the setting of the texts.

and the code used for the analysis are available on https://github.com/MirjamVisscher/cantostream_boundaries.

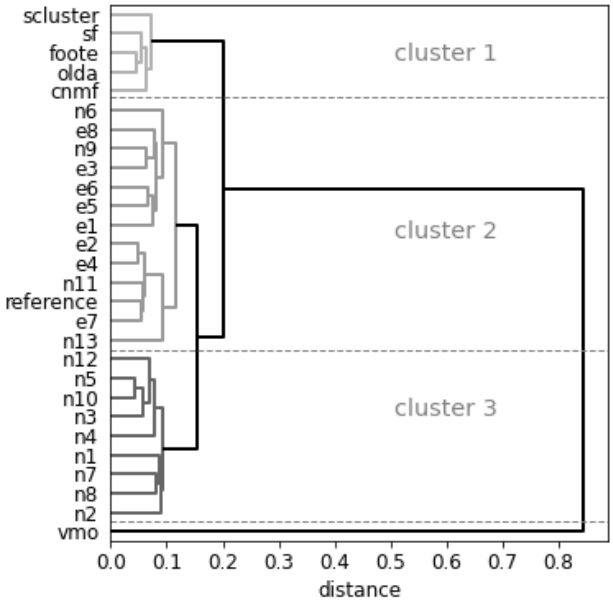

Figure 5. Clustering of distances between Gaussian curves of individual participants non-experts n1 to n13, experts e1 to e8, reference and algorithms.

3.4 Boundary Detectors

We apply the MSAF boundary detectors [15] as provided in the library with standard settings, the input format being wav files. While the algorithms do not give us a weight per boundary, they often provide more than one boundary per quarter note. We do not want to discard this information from the analysis and count number boundaries per quarter note and use this as a weight.

3.5 Analysing Distance between Participants and Algorithms

The participant annotations have a high resolution of one quarter note, and annotations based on the same musical cue may be several quarter notes apart. Therefore it does not make much sense to directly compare annotations. Instead, we smooth the time series using a Gaussian kernel with sigma = 0.002. Subsequently, we compute the pairwise distances between the Gaussian curves of all participants and algorithms [27]. Utilizing the resulting distance matrix, we analyze how individual participants and algorithms cluster through hierarchical cluster analysis using the Ward method from the SciPy implementation [28].

3.6 Compute Boundaries on Group Level

The second strand of the workflow involves identifying boundaries on the level of participant groups. For both experts and non-experts, we begin by averaging annotation weights. Then, we utilize a peakfinder with continuous wavelet transformation [29], applying a window of 4 quarter notes to correct for the spread in timing by participants around perceived boundaries.

	window = 8			window = 4			window = 2		
	P	R	F	P	R	F	P	R	F
reference	.92	.71	.80	.88	.68	.76	.80	.62	.70
non-experts	.81	.90	.85	.75	.83	.79	.65	.72	.68
Foote	.42	.39	.41	.23	.22	.23	.17	.16	.17
SF	.40	.34	.37	.19	.16	.17	.14	.12	.13
CNMF	.39	.52	.45	.21	.27	.24	.14	.18	.16
OLDA	.34	.34	.34	.25	.25	.25	.18	.18	.18
Scluster	.31	.40	.35	.17	.22	.19	.11	.14	.12
VMO	.06	.84	.10	.05	.73	.09	.04	.66	.08

Table 3. Evaluation of reference, non-experts, and MSAF algorithms for all Early8 works combined, evaluated against the expert ground truth.

3.7 Evaluate Non-Experts and Algorithms

We consider the boundaries by the expert participants to be the ground truth. We evaluate the group level boundaries by non-experts and the boundaries by the separate algorithms against this ground truth, using precision, recall and F1 score. Next, we compare the algorithm boundaries to the peaks of the experts. For the evaluation, we use the segment detection method [30] from mir_eval [4] with a window of synchrony of 8 quarter notes, comparable to the aforementioned window of 3 seconds. This means that a boundary is considered in agreement when it is within a window of 8 quarter notes from the expert peaks. From a musical standpoint a window of 8 quarter notes or 3 seconds is quite lenient. Therefore, we also evaluate with window sizes of 2 and 4 quarter notes to see how this impacts the outcomes. For the evaluation of the non-experts, we take into account both precision and recall. For the algorithms, we only take into account the precision since we can employ more than one algorithm, each one specialised in its own type of boundaries.

4. RESULTS

4.1 Collection of Annotations

We recruited 22 participants: 8 experts, 13 non-experts, and one reference participant with a music analysis background. On average, the experts have 30 years of musical training and 27 years of Renaissance practice, the non-experts have on average 3 years of musical training and no Renaissance practice. The experiment took place between September 2022 and January 2023. Eleven non-experts and all experts finished the complete experiment. Each session had a duration of 45 to 90 minutes, depending on the concentration and the desired breaks of the participants.

There were no noticeable changes in granularity after calibration of the granularity in the first work. Therefore, we decided to keep the results of the first work in the dataset.

As an example, figure 4 shows the Gaussian curves for Palestrina, *Vergine tale è terra*, the upper pane for the participants, and the lower pane for the algorithms. Experts and non-experts show a clear agreement on where the boundaries are, whereas the algorithms show no agreement at al. A similar picture emerges for all compositions.

[4] https://github.com/craffel/mir_evaluators

4.2 Clustering of Participants and Algorithms

The clustered individual participants and algorithms in Figure 5 show a clear division between humans and computer. Cluster 1 contains all algorithms, except for VMO, which has a large distance to the other algorithms. Cluster 2 contains the reference, all experts and four non-experts; cluster 3 contains non-experts only. Due to its high density of boundaries, VMO is an outlier, as is shown in Figure 4.

4.3 Evaluation of the Humans and the Algorithms

We evaluate the boundaries of the non-experts, the reference, and the algorithms against the expert ground truth, using precision, recall and F1 score, and with 3 window sizes: 8 quarter notes, 4, and 2.

In Table 3, window = 8, we see that the non-experts have a precision of .81 and a recall of 0.90. This suggests that the non-experts generally find the same boundaries as the experts, although on a slightly more granular level. This is confirmed by a visual inspection of the annotations. The reference in turn has a precision of .92 and a somewhat lower recall of .71. This suggests that the reference yields similar boundaries as the experts, but on a higher structural level.

All algorithms have a much lower performance than the non-experts. The best performing algorithms are Foote, SF and CNMF, having a precision around .40.

We investigate the impact of window size on the performance of non-experts and algorithms by rerunning the evaluation with window sizes of 4 and 2. The non-experts show a moderate decrease, whereas the algorithms drop by a much larger degree. This suggests that, even if algorithms are able to find a boundary, they tend to place it on a different location than humans.

5. DISCUSSION

5.1 Not Every Boundary is a Cadence

As discussed in section 2.1, not every boundary in vocal polyphony is marked by a formal cadence: there are other

Work	Cadences	Plagal	Non-cadences	Total
1	14	2	2	18
2	13	3	2	18
3	15	2	0	17
4	6	1	2	9
5	5	3	1	9
6	8	1	1	10
7	17	3	4	24
8	9	3	3	15
Total	87	18	15	120

Table 4. Expert boundaries classified by the second author into cadence and non-cadence types.

types of boundaries, too. In Table 4 we count the different types of boundaries that occur in the expert annotations (as assessed by the second author). The majority of boundaries is indeed marked by a cadence that follows period theory to a larger or smaller extent. Fifteen percent of the boundaries are plagal cadences (which are not described in the theoretical sources), and another 13 percent of the boundaries is not marked by a cadence at all. This finding has an important implication: if we were to segment compositions based on cadences alone, we would miss out on a significant part of the boundaries that expert listeners deem to be relevant.

5.2 Segmenting Homophonic Compositions

In complex vocal polyphony, cadences and other structural markers are overlapped by voices that follow their own course uninterrupted. This may be one important reason for the low performance of the MSAF algorithms. To put this in perspective, we tested the performance of the algorithms on two works of a more homophonic nature, *Innsbruck, ich muss dich lassen* by Heinrich Isaac, and *Ave verum corpus* by William Byrd. The reference annotations are provided by the first author, by means of a structural analysis as described in section 4.1. The evaluation results are shown in Table 5.

All algorithms in this additional experiment, except for VMO, show a precision between 0.77 and 0.94, and a recall between 0.43 and 0.69. As explained in section 3.8, the high precision scores are particularly interesting since this would potentially allow the combination of multiple detectors combined to increase recall. The outcomes also show that these detectors can be generalised beyond the repertoire they were trained on, namely popular and some classical music. Even though the timbral and tonal features of the homophonic compositions differ substantially from the training data, the algorithms appear able to deal with these. The most important challenge for improving the algorithms for vocal polyphony therefore seems to be the presence of overlapping voices in more complex compositions.

	P	R	F
SF	.94	.43	.59
OLDA	.87	.57	.69
Scluster	.86	.69	.76
Foote	.83	.54	.66
CNMF	.77	.69	.73
VMO	.06	1.00	.12

Table 5. Evaluation the MSAF algorithms for Isaac and Byrd using a window size of 8.

5.3 Experimental Design

Our experiment is primarily designed to efficiently capture perceived musical boundaries. To promote perception, we provide no score or other visual cues that can prompt analytical instead of perceptual annotations. We found that revision of the annotations by participants is not needed to obtain consistent results.

The choice for the task of tapping to music was deliberate, as it falls within the range of spontaneous gestures induced by music, whereas traditional computer interfaces can pose a barrier between perception and task execution. However, the quality of the data depends on the skills of the experimenter, and the setup of this experiment imposes limitations on the number of annotations possible.

Fortunately, [22] offers a model of how to scale up and collect consistent annotations in an online setting. For our purposes, it would be crucial to disable annotation review functionalities and not to use any score visualisation. Also we would recommend a simplified input mode that allows the annotation of a boundary through a single gesture, for example by tapping a touchpad and recording the duration of the tap.

5.4 Limitations

The number of participants in this experiment is limited to 8 experts and 13 non-experts, which invites further investigation on the optimum number of annotators. In addition, annotation of a wider range of works is important for gaining a better insight in early music segmentation. The eight selected musical works are all vocal: we have no information on how instrumental works would impact the results. The works are from a limited time span in the late Renaissance and early Baroque, we do not know how participants and algorithms would perform on works selected from a wider time span.

During the the experiment, we initially notated strong boundaries that were perceived on long notes, at the end of that specific note. Later, based on our first experiences, we decided to annotate boundaries on the exact location where the participant tapped. As a consequence, the strong boundaries annotated by the non-experts may appear simultaneously, while in reality they might have been tapped at slightly different moments.

For the MSAF algorithms, we have used the standard, out-of-the-box settings, using a fixed seed for stochastic

elements in the algorithms, to ensure that the results are reproducible. The algorithms might perform better with custom settings or after being retrained on a more fitting corpus.

6. CONCLUSION

This paper presents a new annotation dataset and answers two main questions. First, it examines whether non-expert annotations of boundaries are fit to improve segmentation algorithms for vocal polyphony. And secondly, it evaluates the performance of a selection of current boundary algorithms on vocal polyphony.

6.1 Agreement Between Non-Experts and Experts

When clustering, we expected three clearly separated clusters of experts versus non-experts versus algorithms. The participants do indeed form two clusters, but interestingly in one, experts and non-experts are mixed, while the other consists of non-experts only. This suggests that differences between experts and non-experts are not particularly strong.

In the evaluation, we treated experts and non-experts as two separate groups, and considered the expert annotations as the ground truth. We tested the precision and recall of the non-expert boundaries measured against the expert boundaries. The non-experts have a precision of 0.81 and a recall of 0.90. This is a noticeable but not a very large difference.

We conclude that non-expert annotations are close to expert annotations but not fully interchangeable. These findings are in line with [31], where experimental outcomes suggest that musically untrained listeners have the ability to process musical structures in a similar way as musical experts. Philips et al. [32] found supporting results, after comparing expert and non-expert segmentation in contemporary music, concluding that the performance of both groups was very similar.

6.2 Algorithms Boundaries are Far Removed from those by Experts

In our analysis, VMO forms its own cluster due to an unusual high amount of boundaries indicated. The other algorithms form another cluster, clearly separated from the human annotations.

Since the performance of non-experts is much higher than the performance of algorithms and fairly close to the expert ground truth, we conclude that non-expert annotations of boundaries in vocal polyphony are suitable to evaluate the performance of boundary detectors until one or more detectors reach the level of humans.

7. FUTURE WORK

Now we have established that non-expert annotations of boundaries are similar enough to expert annotations to be used in the evaluation of segmentation algorithms, we can define a number of follow-up steps to collect a larger dataset. Preferably, this set is annotated by early music enthusiasts, as they have been exposed to many hours of listening to the genre and could be regarded as lay experts in the area.

We calculated boundaries using annotations by 22 participants. To make the annotation process more efficient, we need to investigate what the optimum number of annotations per composition is to get a reliable set of boundaries. The activity analysis approach described in [25] could supply further statistical underpinning of our conclusions, provided the approach is adapted to suit the scoring method used in our experiment.

The works need to be selected from a wider time span to make sure that the results are more widely generalisable. For the same reason, the selected works need to include instrumental music. To understand the impact of the performance, we need to extend the experiment with different performances of the same works.

The algorithms were tested with out-of-the-box settings: optimising the settings to late Renaissance polyphony could yield better results. Concerning algorithm design, it seems that much can be gained by taking the polyphonic structures such as overlapping voices into account. Whether existing algorithms can be adapted for this or new algorithms need to be developed is an open question.

Once the performance of boundary detection algorithms has reached an acceptable level, we will be able to create more high-level analytical procedures that support our goal of understanding tonal development in early music.

Acknowledgments

The authors would like to thank the participants of the experiment, Libio Goncalves Braz for reviewing the code, and the anonymous reviewers for their feedback.

8. REFERENCES

[1] C. Weiß, "Investigating style evolution of Western classical music: A computational approach," *Musicae Scientiae*, vol. 23, pp. 486–507, 2019.

[2] J. Yust, "Stylistic information in pitch-class distributions," *J. of New Music Research*, vol. 48, no. 3, pp. 217–231, 2019.

[3] D. Harasim, F. C. Moss, M. Ramirez, and M. Rohrmeier, "Exploring the foundations of tonality: statistical cognitive modeling of modes in the history of Western classical music," *Humanities and Social Sciences Communications*, vol. 8, pp. 1–11, 2021.

[4] M. K. Long, *Hearing Homophony: Tonal Expectation at the Turn of the Seventeenth Century*. Oxford University Press, USA, 2020.

[5] N. Cook, *Beyond the Score: Music as Performance*. London, U.K.: Oxford University Press, 2013.

[6] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *2000 Proc. of the IEEE Int. Conf. on Multimedia and Expo*, New York, NY, USA, 2000, pp. 452–455.

[7] O. Nieto and T. Jehan, "Convex non-negative matrix factorization for automatic music structure identification," in *2013 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Vancouver, BC, Canada, pp. 236–240.

[8] B. McFee and D. P. Ellis, "Learning to segment songs with ordinal linear discriminant analysis," in *2014 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Florence, Italy, pp. 5197–5201.

[9] B. McFee and D. Ellis, "Analyzing song structure with spectral clustering." in *Proc. of the Int. Soc. for Music Information Retrieval (ISMIR)*, Taipei, Taiwan, 2014, pp. 405–410.

[10] J. Serra, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised detection of music boundaries by time series structure features," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 26, no. 1, 2012, pp. 1613–1619.

[11] C. Wang and G. J. Mysore, "Structural segmentation with the variable markov oracle and boundary adjustment," in *2016 IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, Shanghai, China, pp. 291–295.

[12] F. Fitch, *Renaissance Polyphony*. New York, NY, USA: Cambridge University Press, 2020.

[13] G. Zarlino, *The Art of Counterpoint: Part Three of Le Institutioni Harmoniche, 1558*. Translated by Guy A. Marco and Clade V. Palisca. Music Translation Series. New Haven, CT: Yale University Press, 1968.

[14] H. H. Eggebrecht and F. Reckow, "Kadenz," in *Das Handwörterbuch der Musikalischen Terminologie*. JSTOR, 1968.

[15] O. Nieto and J. P. Bello, "Systematic exploration of computational music structure research." in *Proc. of the Int. Soc. for Music Information Retrieval (ISMIR)*, New York, NY, USA, 2016, pp. 547–553.

[16] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis." in *Proc. of the Int. Soc. for Music Information Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 625–636.

[17] O. Nieto, G. J. Mysore, C. Wang, J. B. Smith, J. Schlüter, T. Grill, and B. McFee, "Audio-based music structure analysis: Current trends, open challenges, and applications," *Transactions of the Int. Soc. for Music Information Retrieval (ISMIR)*, vol. 3, no. 1, 2020.

[18] I. Deliege, "Grouping conditions in listening to music: An approach to Lerdahl & Jackendoff's grouping preference rules," *Music Perception*, vol. 4, no. 4, pp. 325–359, 1987.

[19] E. F. Clarke and C. L. Krumhansl, "Perceiving musical time," *Music Perception*, vol. 7, no. 3, pp. 213–251, 1990.

[20] F. Wiering, J. de Nooijer, A. Volk, and H. J. Tabachneck-Schijf, "Cognition-based segmentation for music information retrieval systems," *J. of New Music Research*, vol. 38, no. 2, pp. 139–154, 2009.

[21] D. Tomašević, S. Wells, I. Y. Ren, A. Volk, and M. Pesek, "Exploring annotations for musical pattern discovery gathered with digital annotation tools," *J. of Math. and Music*, vol. 15, no. 2, pp. 194–207, 2021.

[22] D. Bedoya, "Capturing musical prosody through interactive audio/visual annotations," Ph.D. dissertation, Institut de Recherche et Coordination Acoustique / Musique (IRCAM), Paris, France, 2023.

[23] D. Bedoya, L. Fyfe, and E. Chew, "A perceiver-centered approach for representing and annotating prosodic functions in performed music," *Frontiers in Psychology*, vol. 13, 2022.

[24] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, "The musicality of non-musicians: An index for assessing musical sophistication in the general population," *PloS ONE*, vol. 9, no. 2, p. e89642, 2014.

[25] F. Upham and S. McAdams, "Activity analysis and coordination in continuous responses to music," *Music Perception: An Interdisciplinary Journal*, vol. 35, no. 3, pp. 253–294, 2018.

[26] B. L. Sturm and A. Flexer, "Validity in music information research experiments," *arXiv preprint arXiv:2301.01578*, 2023.

[27] C. F. Jekel, G. Venter, M. P. Venter, N. Stander, and R. T. Haftka, "Similarity measures for identifying material parameters from hysteresis loops using inverse analysis," *Int. J. of Material Forming*, vol. 12, pp. 355–378, 2019.

[28] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.

[29] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, 2006.

[30] D. Turnbull, G. R. Lanckriet, E. Pampalk, and M. Goto, "A supervised approach for detecting boundaries in music using difference features and boosting." in *Proc. of the Int. Soc. for Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007, pp. 51–54.

[31] E. Bigand, "More about the musical expertise of musically untrained listeners," *Ann. of the New York Academy of Sciences*, vol. 999, no. 1, pp. 304–312, 2003.

[32] M. Phillips, A. J. Stewart, J. M. Wilcoxson, L. A. Jones, E. Howard, P. Willcox, M. du Sautoy, and D. De Roure, "What determines the perception of segmentation in contemporary music?" *Frontiers in Psychology*, vol. 11, p. 1001, 2020.