Computational Similarity of Portuguese Folk Melodies Using Hierarchical Reduction

Nádia Carvalho
University of Porto, Faculty of Engineering and INESC TEC
Porto, Portugal
up201208223@up.pt

Daniel Diogo
University of Porto, Faculty of Engineering
Porto, Portugal
danicafd@sapo.pt

Gilberto Bernardes
University of Porto, Faculty of Engineering and INESC TEC
Porto, Portugal
gba@fe.up.pt

ABSTRACT

We propose a method for computing the similarity of symbolically-encoded Portuguese folk melodies. The main novelty of our method is the use of a preprocessing melodic reduction at multiple hierarchies to filter the surface of folk melodies according to 1) pitch stability, 2) interval salience, 3) beat strength, 4) durational accents, and 5) the linear combination of all former criteria. Based on the salience of each note event per criteria, we create three melodic reductions with three different levels of note retention. We assess the degree to which six folk music similarity measures at multiple reduction hierarchies comply with collected ground truth from experts in Portuguese folk music. The results show that SIAM combined with 75th quantile reduction using the combined or durational accents best models the similarity for a corpus of Portuguese folk melodies by capturing approximately 84-90% of the variance observed in ground truth annotations.

CCS CONCEPTS

• **Applied computing → Sound and music computing**; • **Information systems → Digital libraries and archives**.

KEYWORDS

Folk music, computational similarity, melody, structural reduction, Portuguese.

ACM Reference Format:
Nádia Carvalho, Daniel Diogo, and Gilberto Bernardes. 2023. Computational Similarity of Portuguese Folk Melodies Using Hierarchical Reduction. In *10th International Conference on Digital Libraries for Musicology (DLfM 2023), November 10, 2023, Milan, Italy*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3625135.3625152

1 INTRODUCTION

Folk music similarity has garnered significant interest within the music information research and computational ethnomusicology communities. Existing research has primarily focused on three key areas. Firstly, there is a pursuit to comprehend folk music styles across diverse cultures [11, 27]. Secondly, efforts have been made to identify variations within folk music, particularly those resulting from oral tradition [7, 25, 26, 32]. Lastly, the most active area of research in this field, which is the central focus of this article, revolves around computational modeling of folk song similarity to capture human judgments [1, 6, 14, 15, 23, 37].

The realm of digital libraries in musicology has traditionally overlooked Portuguese folk music due to the scarcity of digitized resources. However, the recent release of the I-Folk dataset, encompassing Iberian (Portuguese and Spanish) folk music [9], has provided a valuable resource. This dataset facilitates empirical analysis. Our research addresses this gap by proposing a method for assessing the similarity of symbolically-encoded Portuguese folk melodies. By evaluating existing similarity measures, as summarized in [26], for their alignment with expert perceptual judgments, we not only contribute to the understanding of Portuguese folk music but also lay the groundwork for leveraging advanced query methods at scale within digital libraries.

The main novelty of this study lies in the utilization of a preprocessing melodic reduction technique at multiple hierarchies, in the line of studies by Orio and Rodà [40] and Simonetta et al. [47]. It aims at grasping the perceptual overarching structures and patterns that emerge over time beyond attending to individual notes or melodic fragments [13]. In detail, the proposed technique aims to unveil the anchor notes within a melody by considering several critical factors, including pitch stability, interval salience, beat strength, durational accents, and their combined influence. By computing the salience of each note event based on these criteria, the melodies can be reduced at multiple levels by gradually filtering the notes with smaller salience.

To evaluate the efficacy of our proposed method, we conduct a comparative analysis using six different folk music similarity measures at four reduction hierarchies. Ground truth data collected from two experts in Portuguese folk music serve as a reference to assess the alignment between our computational results and their perceptions. This evaluation enables us to validate and refine our method in line with the expertise of individuals intimately familiar with Portuguese folk music tradition.

The remainder of this paper is organized as follows. Section 2 summarizes the state-of-the-art in computational folk music similarity and melodic reduction. Section 3 presents the methods for melodic reduction at multiple hierarchies. Section 4 details the evaluation of the proposed melodic reduction on a small Portuguese corpus of folk melodies in assessing the degree of similarity between melodies. Finally, Section 5 presents the conclusions of our study and avenues for future work.

DLfM 2023, November 10, 2023, Milan, Italy

ACM ISBN 979-8-4007-0833-6/23/11.
https://doi.org/10.1145/3625135.3625152

2 RELATED WORK

2.1 Computational folk music similarity

The literature on computational modeling of folk music similarity has predominantly concentrated on three core aspects. Firstly, it emphasizes digitizing and compiling comprehensive folk music datasets, accompanied by relevant annotations. Secondly, researchers propose various representations or features that aim to capture the significant dimensions of the musical structure listeners attend to. Lastly, considerable attention is given to defining measures that capture the perceptual similarity between folk melodies.

The access to digitized music datasets of symbolically-encoded folk music is fundamental to computational folk music studies, notably including similarity. Efforts have been made to collect and digitize data for multiple cultures. Representative examples of these collections are the EsAC [44] and its derivative Essen Folk Song Collection [21], featuring a broad set of Asian, European, and American folk songs, the Meertens Tune Collections of Dutch folk songs [50], or the Corpus of Annotated Irish Traditional Dance Music [3]. Recently, the I-Folk datasets featuring Iberian folk melodies from Portugal and Spain have been made publicly available [9], filling an important gap in the access to digitally-encoded Iberian folk music.

Finally, various measures for comparing and assessing the similarity of folk melodies have been proposed. They can be broadly split into distance metrics, such as correlation distance [46], city block distance, and Euclidean distance [48], and alignment measures, such as the Local Alignment (LA) and Structure induction (SIAM) [51]. The latter two measures have been shown to capture human judgments better when compared to the remaining distance metrics [9, 26].

2.2 Melodic reduction

Musicological approaches that propose melodic reductions have been proposed in both music theory [8, 17, 45] and music cognition [30] and music information research [16, 18, 34]. Their working assumption is that melodies are not immutable objects, and thus have multiple manifestations, namely as the result of listening [33].

Computational melodic reduction involves identifying structurally relevant notes within a melody to abstract its higher-level structure. This process draws inspiration from notable theories such as Shenkerian analysis [45] and Generative Theory of Tonal Music (GTTM) [30], which provide frameworks for understanding the hierarchical organization of melodies. Researchers have developed various computational techniques that consider different criteria for melodic reduction and have been pursued in a wide range of musical applications [19, 20, 22, 35]. Here, we adopt the GTTM as a baseline framework from which we derive some melodic reduction principles, which we describe next.

Tonal or harmonic relationships play a crucial role in melodic reduction. Note events that strongly support the underlying tonality or harmonic progression are often prioritized, as they are key structural points in the melodic structure. Several strategies have been pursued to capture pitch stability within Western tonal contexts [4, 28, 29, 39].

Interval leaps, particularly larger leaps, loudness changes, registral density (the distribution of pitches across registers), and register changes also influence melodic reduction, as they contribute to the melodic contour and expressive qualities of the melody, and capturing the listener's attention. Furthermore, the metrical strength serves as an important criterion in melodic reduction. Note events that align with strong metrical positions and emphasize the underlying beat pattern contribute to the melody's rhythmic structure and overall coherence. Existing literature refers to the above categories as phenomenal and metrical accents, respectively [30, 43].

Other criteria may include considerations of ornamentation, rhythmic patterns, and stylistic conventions specific to the musical genre or culture. By integrating these criteria, computational approaches aim to identify and retain the most salient and structurally relevant note events within a melody, facilitating a more concise representation of its higher-level structure.

3 HIERARCHICAL MELODIC REDUCTION

A melodic reduction aiming at finding the more salient or structural notes in a melody is pursued according to five quantifiable criteria: 1) pitch stability, 2) interval salience, 3) beat strength, 4) durational accent, and 5) linear combination of all aforementioned criteria. These criteria are detailed in Sections 3.1-3.5. Our approach is inspired by the concepts defining a note's structural importance or perceptual salience within a melody from music theory and cognitive psychology, namely the GTTM [30]. Once we quantify the salience of each note event per criteria, the importance of each note event is ranked and systematically removed from the melody to create the melodic reductions. Figure 1 shows the computed values per criteria for the Portuguese folk song 'Rosa Branca ao Peito' and can be used to illustrate the processes detailed next.

3.1 Pitch stability

We adopt the tonal interval vector $T(k)$ in Eq. 1 to compute a pitch space where vector distances representing pitch structures of any cardinality (e.g., notes, chords, or scales) capture perceptual relationship within the Western tonal music. The smaller the distance, the more related two pitch structures are.[1]

$$\begin{gathered} T(k) = w(k) \sum_{n=0}^{N-1} \bar{c}(n) e^{\frac{-j2\pi kn}{N}}, \\ \text{with} \quad k \in \mathbb{Z} \quad \text{and} \quad \bar{c}(n) = \frac{c(n)}{\sum_{n=0}^{N-1} c(n)} \quad , \end{gathered} \tag{1}$$

where $N = 12$, and $w(k) = \{2, 11, 17, 16, 19, 7\}$, which are weights empirically driven from complementary dyad consonance ratings to regulate the importance of each DFT dimension k [4]. Eq 1 adopts $\bar{c}$ as the L_1 norm of the $c(n)$ vector to represent all pitch structures (i.e., from a note to a chord, a key, or a segment of any length) within the same unit circle space.

We adopt the tonal interval vector space $T(k)$ to compute the pitch stability P of a note event i within a melody from the 'tonal' context of the melody. To this end, we first define the tonal context of the melody $T_m(k)$ as the accumulated pitch class distribution

[1] A potential disconnect between the adopted tonal space and folk music expressions may arise, as many folk music does not adopt the major/minor Western tonal music modes. However, in the context of the Portuguese folk melodies within the I-Folk dataset, this is not the case, as most examples are in either of these modes. Beyond the context of this work, alternative tonal pith spaces must be pursued.

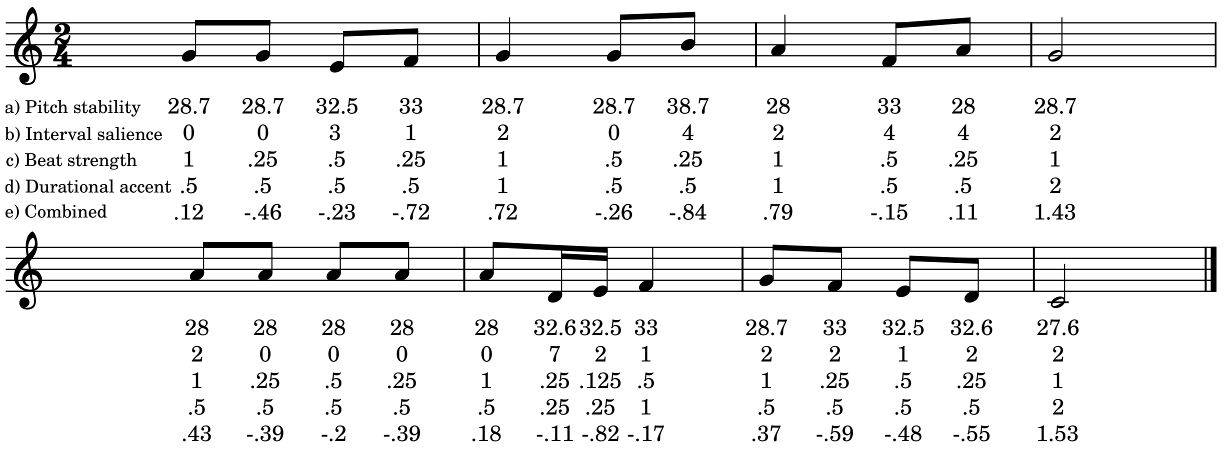
a) Pitch stability 28.7 28.7 32.5 33 28.7 28.7 38.7 28 33 28 28.7
b) Interval salience 0 0 3 1 2 0 4 2 4 4 2
c) Beat strength 1 .25 .5 .25 1 .5 .25 1 .5 .25 1
d) Durational accent .5 .5 .5 .5 1 .5 .5 1 .5 .5 2
e) Combined .12 -.46 -.23 -.72 .72 -.26 -.84 .79 -.15 .11 1.43
28 28 28 28 28 32.6 32.5 33 28.7 33 32.5 32.6 27.6
2 0 0 0 0 7 2 1 2 2 1 2 2
1 .25 .5 .25 1 .25 .125 .5 1 .25 .5 .25 1
.5 .5 .5 .5 .5 .25 .25 1 .5 .5 .5 .5 2
.43 -.39 -.2 -.39 .18 -.11 -.82 -.17 .37 -.59 -.48 -.55 1.53

Figure 1: Second and third phrases of the Portuguese folk song 'Rosa Branca ao Peito' with annotated values for each criterion under study in our melodic reduction approach. Please note that the first note in the excerpt is preceded by a G note in the same octave; as such, it has zero interval salience.

vector for the entire melody weighted by duration (quarter note equals unity).[2] The pitch salience of each note event $i \in \mathbb{Z}^+$ is the Euclidean distance of its pitch class vector $T_i(k)$ from $T_m(k)$ subtracted from unity, such that:

$$P_i = 1 - \sqrt{\sum_{k=1}^{M} |T_m(k) - T_i(k)|^2} \tag{2}$$

Higher values of P denote greater proximity to the melodic 'tonal' context as opposed to lower values that are further from the context, with higher and lower pitch stability, respectively.

3.2 Interval salience

Following [30], larger melodic interval leaps create stronger phenomenological accents and enforce the salience of a given note event within a melody. In this context, we define the interval salience I of note event i as the number of semitones from the preceding note event $i-1$, such that:

$$I_i = \begin{cases} \texttt{null} & \text{if i=0} \\ |P_i - P_{i-1}| & \text{otherwise} \end{cases}, \tag{3}$$

P is a integer value in the $[0, 128]$ following the MIDI standard. The interval salience of the first note event is set by default to `null` as no preceding note exists.

3.3 Beat strength

Beat strength weights each note event according to its position within the metrical grid. Stronger positions will have a heavier beat strength as opposed to weaker positions. In greater detail, it assigns values in the $[0, 1]$ range to five metrical pulses: the measure, half-measure, beat, division, and subdivision pulse. These measure positions are assigned to the following beat strength weights: 1, .5, .25, .125, and .0625. The beat strength weights are computed using the `Music21Object.beatStrength` object from `music21` Python library [12]. The current algorithm's interpretation of compound meters, with three notes per beat, may cause inconsistencies in capturing folk music nuances. To resolve this, a future refined algorithm, attuned to musical surface structure, is needed.

3.4 Durational accents

Many music theorists describe durational accents [5, 31, 41], though sometimes under different names. Broadly, their understanding enforces that durational accents are 'caused by relatively long durations following one or shorter durations' [31, p. 18].

The notion of durational accents in music perception is well-supported by empirical evidence. Perceptual experiments have demonstrated that longer inter-onset-intervals in sine tone stimuli are perceived as generating stronger accents [42]. Furthermore, a study by Müllensiefen et al. [38] investigated over 30 accent rules within a corpus of pop music melodies and showed that durational accents significantly impact the perception of melodic accents.

In our work, the durational accents salience per note event equals to the note duration as ratios of the quarter note, represented by unity.

3.5 Combined

A linear combination of the four above criteria–pitch stability P, interval salience I, beat strength B, and durational accents D–is explored as a multidimensional criterion for extrapolating the importance of each note event in a melody. Before the combination, the resulting values are normalized per dimension to zero mean and unit variance. The normalized values per dimension are notated as $\bar{P}$, $\bar{I}$, $\bar{B}$, and $\bar{D}$. The linear combination C of the note event i is given by:

$$C_i = \bar{P}_i + \bar{I}_i + \bar{B}_i + \bar{D}_i \tag{4}$$

3.6 Reducing the melodic surface

We filter or exclude note events based on their quantified criteria to reduce the folk melodies. Our work uses quantiles to split the

[2] The context vector for the excerpt in Figure 1 is $T_m(k) = \{2, 0, .75, 0, 1.25, 2.5, 0, 4, 0, 4, 0, .5\}$.

values into four reduction levels or hierarchies. Quantiles are points dividing the range of a probability distribution into continuous intervals with equal probabilities. We adopt the 75th, 50th, and 25th quantiles as cutting points. The sequence of quantile cuts filters a gradually larger number of notes (e.g., the 75th quantile filters a reduced number of notes compared to the 25th quantile).

Figure 2 shows the melodic reductions at the three above-defined quantile levels per criteria for the third phrase of the Portuguese folk song 'Rosa Branca ao Peito'. The salience values per criteria are shown in Fig. 1.

4 EVALUATION

We conducted an experiment to assess which melodic reduction hierarchy and similarity measure best captures expert annotations assessing the similarity of Portuguese folk melodies. To this end, we followed a threefold procedure. First, we collect perceptual ground truth annotations from experts, detailed in Sections 4.1. Second, we computed the similarity for multiple measures and representations proposed in the literature at four melodic reduction hierarchies, detailed in Section 3.6 – including the three quantile cuts and the original melody without reduction. Third, we compared the computational models and ground truth annotations to assess which model best matches the ground truth.

We adopt two instruments to compute our results: the intraclass correlation coefficient (ICC) and the coefficient of determination, R^2. The ICC aims to assess the reliability of the ground truth across experts' annotations. We expect to unify expert annotations towards a single global assessment per compared folk melody pair. The R^2 indicates how well the computational similarity models predict the perceptual ground truth annotations from experts. Two main variables under study are the similarity measures and the melodic reduction hierarchy, that best model perceptual similarity ratings of Portuguese folk melodies from experts.

To assess the *reliability* of the ground truth annotations, namely the degree of agreement between annotators, we computed their intraclass correlation coefficient (ICC), a widely used reliability index in interrater reliability analyses [2]. We adopt the two-way mixed-effects model, as the two raters are the only raters of interest. Results only represent the reliability of the specific raters involved in the reliability experiment. Due to our interest in grouping both raters' annotations into a single similarity value, we adopt the ICC type resulting from their mean. ICC reliability values range between 0 and 1, with values closer to 1 representing stronger reliability. A complementary indicator of ICC reliability (typically a more robust indicator) is its 95% confidence interval around the ICC. Values less than .5 are indicative of *poor* reliability, values between .5 and .75 indicate *moderate* reliability, values between .75 and .9 indicate *good* reliability, and values greater than .9 indicate *excellent* reliability.

The coefficient of determination R^2 is a ratio that shows how dependent the computational models are on the ground truth. It is expressed in the [0, 1] range. A perfect R^2 of one means that our computational similarity models explain 100% of the variance in the ground truth. Combined with the R^2 indicator, the p-value, reporting the F statistical significance of the model, can help draw important conclusions about the models. Statistically significant coefficients p below .001 and .0001 indicate strong and highly strong statistical significance, respectively. To model expert annotation, computational models shall have higher R^2 and lower or significant p value.

4.1 Dataset and expert annotations

From the 59 Portuguese folk melodies in the I-Folk dataset, we selected 20 melodies as a sensible number of examples from which some statistical analysis could be conducted without requiring extensive expert assessment, which could lead to poor results due to fatigue.

Genre and meter were the two criteria for selecting representative melodic with controlled bias. 'Children songs' were the adopted genre due to their predominance in the dataset (80% of the total Portuguese folk melodies).[3] A balanced number of ten folk melodies from binary and ternary meters were randomly selected. Table 1 a) shows the final set of annotated folk melodies.

Based on the folk annotation procedure proposed by Volk and van Kranenburg [52], we asked experts to annotate 30 folk melody pairs. Table 1 b) shows the pairing of melodies, which compares intra- and inter-meter melodies. Ten pairs per binary, ternary, and mixed (binary against ternary) meter melodies were presented—the allocation of pairs aimed at presenting the minimal repetition of the melodies. In total, 30 folk melody pairs were assessed.

The experts were asked to evaluate the similarity of two folk melodies according to four principles on a continuous scale in the [0, 2] range. Zero denotes the least similar, one somewhat similar, and two practically indistinguishable. The four assessed principles are the following: contour, rhythm, motifs, and global. While the global assessment is the only indicator of interest in this study, we adopted and annotated the above principles as proposed in [52] to guide experts in their assessment so that they would direct their listening and analysis towards the same conceptual space and to provide some uniform comparison in future with remaining studies, namely from other folk traditions.

4.2 Computational analysis of folk music similarity

Five measures were adopted to compute the similarity of two given melodies with and without the reduction. Table 2 shows the similarity measures and representations adopted. It includes distance metrics, first proposed in ethnomusicological studies – correlation, city block, and Euclidean distances [46, 48] – and alignment measures from music information research – cardinality score [10, 24, 49], LA [51], and SIAM [9, 36]. Due to space constraints, we cannot fully detail each measure's foundational, mathematical, or implementation details. Please refer to Janssen et al. [25] for thorough details, except for the SIAM [36] measures based on Carvalho et al. [9]. A Python script was utilized to compute the measures and is available online at https://github.com/NadiaCarvalho/iFolkSimilarity.git.

Janssen et al. [26] have shown that the LA and the SIAM measures yield the best results regarding computational similarity. Hence, it was imperative to incorporate these two measures in the present study. Additionally, the inclusion of other measures stemmed from their efficacy or low computational cost. It is worth

[3] "Children's Song" and "Lullabies" were the only two existing genres in the Portuguese set of I-Folk. 'Children's songs' was overly predominant (52 to 7).

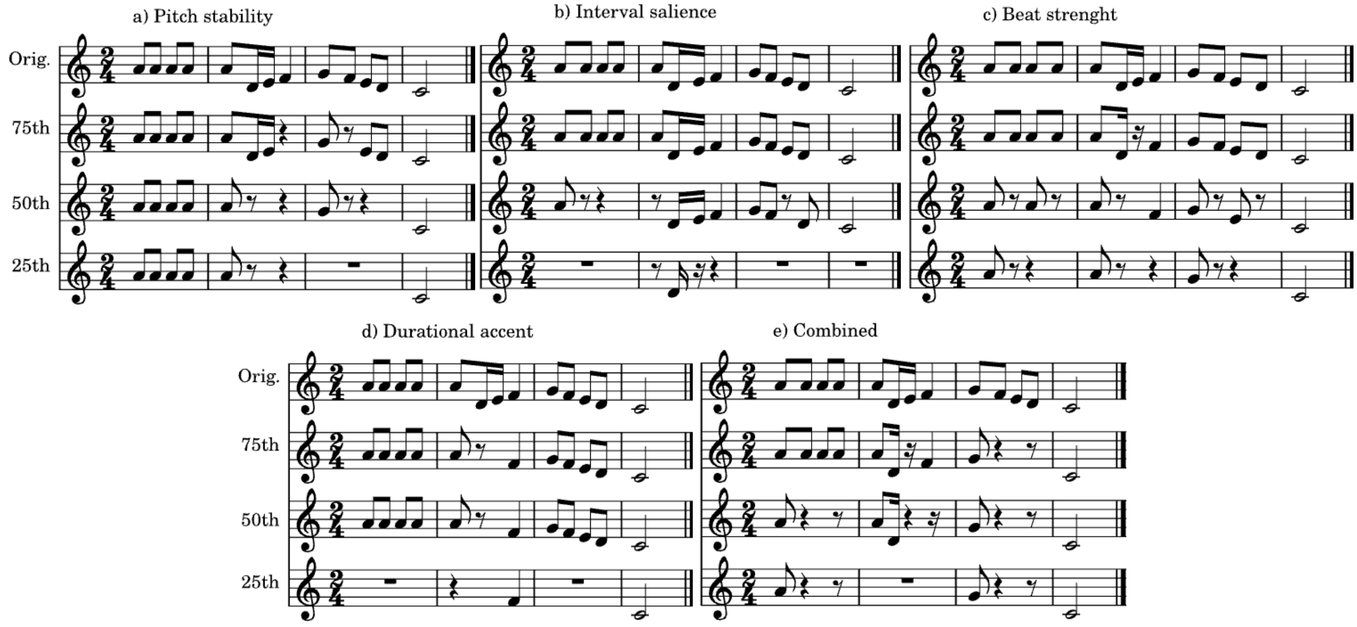

Figure 2: Melodic reductions of the Portuguese folk song 'Rosa Branca ao Peito' for the five criteria adopted in our study using the 25th, 50th, and 75th quantiles as cut points.

a)

ID	Title	
	Binary meter	**T**ernary meter
1	A Barca Virou	A Rolinha Andou, Andou (V1)
2	Lá Vai Uma, Lá Vão Duas	A Rolinha Andou, Andou (V2)
3	Senhora D. Anica	Disse O Galo Prá Galinha
4	Pantaleão	Teresinha De Jesus
5	Passa, Passa, Gabriel	Os Passarinhos (V1)
6	Ó Terrá, Tá, Tá	Os Passarinhos (V2)
7	Que Linda Falua	Lá Vai O Comboio, Lá Vai (V1)
8	Fui Ao Jardim da Celeste	Lá Vai O Comboio, Lá Vai (V2)
9	Marcha Soldado	Sant'António Se Levantou
10	Rosa Branca Ao Peito	Senhores Donos Da Casa

b)

Ground truth pairs					
B vs. B		B vs. T		T vs. T	
1	2	1	2	1	2
3	4	3	4	3	4
5	6	5	6	5	6
7	8	7	8	7	8
9	10	9	10	9	10
1	10	1	10	1	10
2	3	2	3	2	3
4	5	4	5	4	5
6	7	6	7	6	7
8	9	8	9	8	9

Table 1: a) List of Portuguese folk melodies adopted and annotated by expert musicians. A total of 20 melodies were selected, with a balanced number of 10 melodies per binary and ternary meters. b) Pairs of folk melodies adopted in the expert annotation procedure.

Similarity measures	Music representations	Authors
Cardinality score (CS)	pitch and onset	[10, 24, 49]
Correlation distance (CD)	duration weighted pitch sequence	[46]
City block distance (CBD)	pitch sequence	[48]
Euclidean distance (ED)	pitch sequence	[48]
Local alignment (LA)	pitch, onset	[51]
Structure induction (SIAM)	pitch, onset, duration, and beat strength	[36]

Table 2: Similarity measures and music representation used in evaluating our study to compare folk melodies.

exploring whether these measures with lower computational requirements could outperform the purportedly superior measures in the specific context of Portuguese folk music.

4.3 Results

The ground truth was collected from two Portuguese annotators with expert knowledge in music theory and performance experience in Portuguese folk music. The ground truth is available online as supplementary material to this paper at https://github.com/NadiaCarvalho/iFolkSimilarity.git.

Figure 3 shows the ground truth similarity annotation distributions for the 30 melody comparisons per annotator and metric –

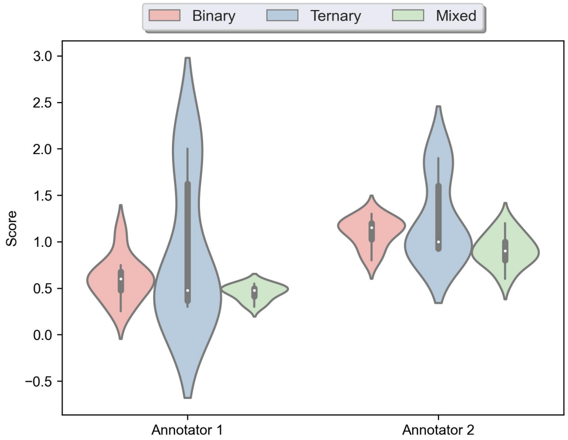

Figure 3: Distribution of the ground truth similarity annotations for the 30 melody comparisons per metric – binary, ternary, and mixed and annotator.

Corpus meter	ICC	F	df1	df2	p	CI95%
Binary	.620	2.632	9	9	≥ .05	[-.53, .91]
Ternary	.922	12.812	9	9	≤ .0001	[.69, .98]
Mixed	.696	3.285	19	19	≤ .01	[.23, .88]

Table 3: The intraclass correlation coefficient (ICC) results for the Portuguese folk melody annotations by two experts. Reported indicators include the ICC, the F statistic, the numerator (df1) and denominator (df2) degrees of freedom, p-value, and the 95% confidence interval around the ICC (CI95%).

binary, ternary, and mixed. The annotations are somehow unbalanced across different meter categories. While the folk melodies annotations with ternary meters are well distributed across the [0,2] range, the remaining binary and mixed categories have a narrower expert annotations' range. These results may be due to the inclusion of multiple voices from the same folk song in the ternary meter category, which led to ratings of 2 (denoting high similarity) from the annotators. Very similar melodies are lacking in the remaining meter categories. This unbalanced distribution does not critically hamper the study; however, a revision of these annotations toward higher balanced across categories shall be examined in future studies.

Table 3 shows the reliability values for the two experts given by the ICC. We assess the binary and ternary meter melodies separately and their combination. Excepting the ternary meter melodies, poor reliability exists between the expert annotations. Due to the subjective nature of the task, the poor ICC reliability value may not be entirely surprising – even when enforcing the sub-evaluation attributes. In this context, we will compare the folk music similarity with the multiple degrees of reduction to each expert score separately. A possible explanation for the ternary meter melodies' moderate to good reliability values is a higher balanced distribution of ratings across the [0,2] range.

Table 4 show the coefficient of determination R^2 assessing the degree to which the similarity measure and representations defined in Section 4.2 model the ground truth from Annotator 1 and 2, respectively. Italic and bold font styles indicate statistical significance for p values < .001 and < .0001, respectively. Underline indicates the best result per measure (i.e., per row).

The first observation of interest is the SIAM measure and its underlying melodic representations as the most aligned with the perceived similarity of both annotators by an important margin (best R^2 values for annotators 1 and 2 are .895 and .838, respectively). The second best measure is the correlation distance, whose best R^2 values for annotators 1 and 2 drops to .773 and .643. Furthermore, the SIAM measure is highly statistically significant across all possible reduction levels under study with $p < .0001$ in both annotators.

The melodic reductions improve the R^2 results marginally, except for the correlation distance measure. All remaining measures of melodic reductions have marginal improvements. Furthermore, irrespective of the adopted dimension, the 75th percentile of our melodic reductions show the best results overall. Based on the assumption that the 75th percentile primarily excludes passing (potentially, outside of key) notes in weaker metrically positions, leading to the concept of embellishment notes, these results suggest that excluding this layer of ornamentation, notes can lead to better computational similarity models. Interestingly, the measure in which our reduction method best improves the R^2 results in the cardinality score, which compares how many equal events are between two melodies. In improving this measure with the reduced melodies, the results suggest that the reductions capture the most fundamental notes in the melodic structure.

Examining the R^2 values for each structural dimension in the reduced melodies, we can observe that the durational accents and beat strength algorithms provide better structural reductions. The best R^2 results for annotators 1 and 2 are durational accents D are the combined criteria C, respectively.

Table 5 shows the results of the SIAM similarity measure in capturing the experts' annotations as the coefficient of determination R^2 per meter: binary, ternary, and mixed. To simplify the results report, we only expose this tripartite meter results for the best-performing measure, SIAM. Detailed metrics results are available online at: https://github.com/NadiaCarvalho/iFolkSimilarity.git. While the results show that SIAM models the similarity of ternary melodies to a higher degree, these results can be skewed by the biased ground truth data with ratings across a wider range for the ternary (shown in Fig. 3), particularly featuring very similar folk songs, which are undoubtedly easier to identify by experts and computer models alike.

5 CONCLUSIONS AND FUTURE WORK

In the present study, we propose a hierarchical melodic reduction that involves assigning salient values to individual notes within a melody. This assignment is based on five distinct criteria: pitch stability, interval salience, beat strength, durational accents, and their linear combination. These computed values indicate the relative significance of each note within the overall melodic structure.

Annotator 1

Simil measure	Original melody	Pitch stability *P*			Interval salience *I*			Beat strength *B*			Duration accent *D*			Combined criteria *C*		
Red (%)	100	75	50	25	75	50	25	75	50	25	75	50	25	75	50	25
CS	.171	.232	.191	.218	.191	.234	.306	.152	.141	.093	.132	.132	.134	.171	.145	.091
CD	**.773**	.144	*.422*	.035	**.489**	.000	.002	**.703**	**.694**	**.728**	**.695**	**.631**	**.670**	*.364*	.227	.103
CBD	.202	.119	.220	.179	.253	.182	.136	.215	.205	.318	.230	.224	.310	.197	.158	.136
ED	.286	.024	.124	.003	.297	.024	.113	.318	.111	.302	*.404*	*.371*	.144	.181	.050	.000
LA	**.439**	**.432**	.241	.305	**.476**	**.456**	**.461**	**.425**	*.345*	.305	*.406*	*.412*	*.351*	*.403*	.219	.157
SIAM	**.877**	**.867**	**.803**	**.606**	**.878**	**.846**	**.848**	**.877**	**.860**	**.851**	**.895**	**.872**	**.835**	**.880**	**.836**	**.665**
Annotator 2																
CS	.192	.264	.194	.175	.205	.254	.301	.176	.158	.180	.160	.159	.208	.215	.195	.173
CD	**.643**	.095	*.368*	.001	**.438**	.002	.004	**.560**	**.544**	**.573**	**.542**	**.504**	**.526**	.280	.177	.042
CBD	.188	.112	.222	.179	.237	.134	.123	.197	.167	.219	.199	.204	*.371*	.183	.149	.125
ED	.260	.023	.138	.001	*.329*	.003	.142	.289	.087	.186	*.340*	*.328*	.177	.170	.057	.000
LA	*.390*	*.385*	.253	.274	*.396*	*.375*	*.395*	*.387*	.296	.301	*.375*	*.384*	*.361*	*.353*	.212	.233
SIAM	**.833**	**.828**	**.780**	**.614**	**.813**	**.774**	**.776**	**.837**	**.806**	**.809**	**.838**	**.832**	**.771**	**.842**	**.792**	**.690**

Table 4: Coefficient of determination R^2 assessing how well the similarity measures defined in Section 4.2 model the ground truth from Annotators 1 and 2. *Italic* and bold font styles indicate statistical significance for p values $< .001$ and $< .0001$, respectively. Underline indicates the best result per measure (i.e., per row).

Annotator 1

Simil measure	Original melody	Pitch stability *P*			Interval salience *I*			Beat strength *B*			Duration accent *D*			Combined criteria *C*		
Red (%)	100	75	50	25	75	50	25	75	50	25	75	50	25	75	50	25
Binary	.136	.144	.037	.063	.053	.103	.006	.069	.046	.012	.056	.061	.000	.102	.104	.126
Ternary	.957	.941	.954	.856	.956	.911	.920	.972	.968	.967	.973	.973	.930	.973	.946	.868
Mixed	.500	.530	.402	.284	.387	.301	.356	.478	.390	.136	.467	.452	.192	.471	.174	.146
Annotator 2																
Binary	.136	.144	.037	.063	.053	.103	.006	.069	.046	.012	.056	.061	.000	.102	.104	.126
Ternary	.957	.941	.954	.856	.956	.911	.920	.972	.968	.967	.973	.973	.930	.973	.946	.868
Mixed	.500	.530	.402	.284	.387	.301	.356	.478	.390	.136	.467	.452	.192	.471	.174	.146

Table 5: Coefficient of determination R^2 assessing how well the SIAM similarity measure defined in Section 4.2 model the ground truth from Annotators 1 and 2 per binary, ternary, or mixed metrics. *Italic* and bold font styles indicate statistical significance for p values $< .001$ and $< .0001$, respectively. Underline indicates the best result per measure (i.e., per row).

Subsequently, these salience values filter the melodic surface hierarchically, with three different reduction levels. Specifically, quantiles at the 75th, 50th, and 25th percentiles were adopted as cutting points for filtering the salience of notes, enabling a refined melodic representation.

To evaluate the potential effectiveness of our melodic reduction approach, we conducted an assessment using five existing computational similarity measures proposed in the literature. These measures were employed to capture similarity judgments provided by two expert musicians. For our evaluation, we collected data on 30 pairs of melodies from the Portuguese subset of the I-Folk dataset. The specific computational similarity measures utilized in our study are listed in Table 2.

The results of our evaluation indicate that the SIAM measure emerges as the most effective similarity measure for the I-Folk Portuguese folk music corpus. This finding aligns with previous studies conducted in different cultural contexts [25, 26]. The SIAM model accounts for approximately 84 − 90% of the variance observed in the ground truth data. Moreover, we observed that incorporating our melodic reduction strategies resulted in only marginal improvements when a 75th quantile cut was applied. Specifically, our findings highlight durational accents and the combined linear criteria as the most successful reduction strategies in enhancing the computational similarity measures.

Applying our methods to structurally segmented phrases or motifs within folk songs could yield valuable insights. Existing evidence suggests that human listeners better recognize the presence or absence of short melodic segments in folk songs rather than the overall global structure [52]. Moreover, while the combined model demonstrates enhanced results compared to the individual criteria, the outcomes obtained from each criterion indicate varying degrees of contribution towards an informed reduction. Consequently, assigning appropriate weights to the constituent criteria within the linear combination C holds promise for further refining and improving the effectiveness of our method.

ACKNOWLEDGMENTS

This research has been funded by the Portuguese National Funding Agency for Science, Research, and Technology [2021.05132.BD and UIDB/50014/2020] and the European Union's Horizon Europe research and innovation programme under the Marie Sklodowska-Curie grant agreement no 101086338.

REFERENCES

[1] Korinna Bade, Andreas Nürnberger, Sebastian Stober, Jörg Garbers, and Frans Wiering. 2009. Supporting Folk-Song Research by Automatic Metric Learning and Ranking.. In *ISMIR*. 741–746.

[2] John J Bartko. 1966. The intraclass correlation coefficient as a measure of reliability. *Psychological reports* 19, 1 (1966), 3–11.

[3] Pierre Beauguitte, Bryan Duggan, and John D Kelleher. 2016. A Corpus of Annotated Irish Traditional Dance Music Recordings: Design and Benchmark Evaluations. In *ISMIR*. 53–59.

[4] G. Bernardes, D. Cocharro, M. Caetano, C. Guedes, and M. Davies. 2016. A multi-level tonal interval space for modelling pitch relatedness and musical consonance. *Journal of New Music Research* 45, 4 (2016), 281–294.

[5] Wallace Berry. 1987. *Structural functions in music.* Courier Corporation.

[6] Peter Boot, Anja Volk, and W Bas de Haas. 2016. Evaluating the role of repeated patterns in folk song classification and compression. *Journal of New Music Research* 45, 3 (2016), 223–238.

[7] Bertrand Harris Bronson. 1950. Some observations about melodic variation in British-American folk tunes. *Journal of the American Musicological Society* 3, 2 (1950), 120–134.

[8] Aaron Carter-Ényì and Gilad Rabinovitch. 2021. Onset and contiguity: Melodic feature reduction and pattern discovery. *Music Theory Online* 27, 4 (2021).

[9] Nádia Carvalho, Sara Gonzalez-Gutierrez, Javier Merchan Sanchez-Jara, Gilberto Bernardes, and Maria Navarro-Cáceres. 2021. Encoding, analysing and modeling i-folk: A new database of iberian folk music. In *8th International Conference on Digital Libraries for Musicology.* 75–83.

[10] Tom Collins, Sebastian Böck, Florian Krebs, and Gerhard Widmer. 2014. Bridging the audio-symbolic gap: The discovery of repeated note content directly from polyphonic music audio. In *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio.* Audio Engineering Society.

[11] Darrell Conklin and Christina Anagnostopoulou. 2011. Comparative Pattern Analysis of Cretan Folk Songs. *Journal of New Music Research* 40, 2 (2011), 119–125. https://doi.org/10.1080/09298215.2011.573562

[12] Michael Scott Cuthbert and Christopher Ariza. 2010. music21: A toolkit for computer-aided musicology and symbolic music data. (2010).

[13] Irène Deliège. 2001. Similarity perception ↔ categorization ↔ cue abstraction. *Music Percept.* 18, 3 (March 2001), 233–243.

[14] Tuomas Eerola, Topi Järvinen, Jukka Louhivuori, and Petri Toiviainen. 2001. Statistical features and perceived similarity of folk melodies. *Music Perception* 18, 3 (2001), 275–296.

[15] Jörg Garbers, Anja Volk, Peter van Kranenburg, Frans Wiering, Louis P Grijp, and Remco C Veltkamp. 2009. On pitch and chord stability in folk song variation retrieval. In *Mathematics and Computation in Music: First International Conference, MCM 2007 Berlin, Germany, May 18–20, 2007 Revised Selected Papers 1.* Springer, 97–106.

[16] Édouard Gilbert and Darrell Conklin. 2007. A probabilistic context-free grammar for melodic reduction. In *Proceedings of the International Workshop on Artificial Intelligence and Music, 20th International Joint Conference on Artificial Intelligence.* Citeseer, 83–94.

[17] Robert Gjerdingen and Janet Bourne. 2015. Schema theory as a construction grammar. *Music Theory Online* 21, 2 (2015).

[18] Ryan Groves. 2016. Automatic Melodic Reduction Using a Supervised Probabilistic Context-Free Grammar.. In *ISMIR*. 775–781.

[19] Masatoshi Hamanaka, Keiji Hirata, and Satoshi Tojo. 2016. Implementing methods for analysing music based on lerdahl and jackendoff's generative theory of tonal music. *Computational music analysis* (2016), 221–249.

[20] Masatoshi Hamanaka, Keiji Hirata, and Satoshi Tojo. 2018. deepgttm-iii: Multi-task learning with grouping and metrical structures. In *Music Technology with Swing: 13th International Symposium, CMMR 2017, Matosinhos, Portugal, September 25-28, 2017, Revised Selected Papers 13.* Springer, 238–251.

[21] Helmut. 1995. The Essen folksong collection in the Humdrum Kern Format (D. Huron, Ed.). *Menlo Park, CA: Center for Computer Assisted Research in the Humanities* (1995).

[22] Dorien Herremans and Elaine Chew. 2016. MorpheuS: automatic music generation with recurrent pattern constraints and tension profiles. In *Proceedings of IEEE TENCON,-2016 IEEE Region 10 Conference.* IEEE, 282–285.

[23] Ruben Hillewaere, Bernard Manderick, and Darrell Conklin. 2009. Global Feature Versus Event Models for Folk Song Classification.. In *ISMIR*, Vol. 2009. 10th.

[24] Berit Janssen, Tom Collins, and Iris Yuping Ren. 2019. Algorithmic Ability to Predict the Musical Future: Datasets and Evaluation.. In *ISMIR*. 208–215.

[25] Berit Janssen, Peter Van Kranenburg, and Anja Volk. 2017. Finding occurrences of melodic segments in folk songs employing symbolic similarity measures. *Journal of New Music Research* 46, 2 (2017), 118–134.

[26] Berit Janssen, Peter van Kranenburg, Anja Volk, et al. 2015. A comparison of symbolic similarity measures for finding occurrences of melodic segments. In *Proceedings of the 16th ISMIR Conference, Málaga, Spain, October 26-30, 2015.* 659–665.

[27] Zoltán Juhász. 2006. A systematic comparison of different European folk music traditions using self-organizing maps. *Journal of New Music Research* 35, 2 (2006), 95–112.

[28] Carol L Krumhansl. 2001. *Cognitive foundations of musical pitch.* Vol. 17. Oxford University Press.

[29] Fred Lerdahl et al. 2001. *Tonal pitch space.* Oxford University Press, USA.

[30] Fred Lerdahl and Ray Jackendoff. 1983. A Generative Theory of Tonal Music.

[31] Joel Lester. 1986. *The rhythms of tonal music.* Pendragon Press.

[32] Jukka Louhivuori. 1990. Computer aided analysis of Finnish spiritual folk melodies. In *Probleme der Volksmusikforschung*, H. Braun (Ed.). Peter Lang, Bern, 312–323.

[33] Alan Marsden. 2005. Generative structural representation of tonal music. *Journal of New Music Research* 34, 4 (2005), 409–428.

[34] Alan Marsden. 2010. Schenkerian analysis by computer: A proof of concept. *Journal of New Music Research* 39, 3 (2010), 269–289.

[35] Alan Marsden, Keiji Hirata, and Satoshi Tojo. 2013. Towards computable procedures for deriving tree structures in music: Context dependency in GTTM and Schenkerian theory. (2013).

[36] David Meredith, Kjell Lemström, and Geraint A Wiggins. 2002. Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research* 31, 4 (2002), 321–345.

[37] Daniel Müllensiefen and Klaus Frieler. 2007. Modelling experts' notions of melodic similarity. *Musicae Scientiae* 11, 1_suppl (2007), 183–210.

[38] Daniel Müllensiefen, Martin Pfleiderer, and Klaus Frieler. 2009. The perception of accents in pop music melodies. *Journal of New Music Research* 38, 1 (2009), 19–44.

[39] María Navarro-Cáceres, Marcelo Caetano, Gilberto Bernardes, Mercedes Sánchez-Barba, and Javier Merchán Sánchez-Jara. 2020. A computational model of tonal tension profile of chord progressions in the tonal interval space. *Entropy* 22, 11 (2020), 1291.

[40] Nicola Orio and Antonio Rodà. 2009. A Measure of Melodic Similarity based on a Graph Representation of the Music Structure.. In *ISMIR*. Citeseer, 543–548.

[41] Peter Petersen. 2013. Music and Rhythm: Fundamentals. *History, Analysis* (2013).

[42] Dirk-Jan Povel and Hans Okkerman. 1981. Accents in equitone sequences. *Perception & Psychophysics* 30 (1981), 565–572.

[43] Bruno H Repp. 2010. Do metrical accents create illusory phenomenal accents? *Attention, Perception, & Psychophysics* 72, 5 (2010), 1390–1403.

[44] Helmut Schaffrath. 1997. The Essen Associative Code: A Code for Folksong Analysis. *Beyond MIDI: The handbook of musical codes* (1997), 343.

[45] Heinrich Schenker. 1935. *Der Freie Satz.* Universal Editions, Vienna. Trans. E. Oster (1979) Free Composition, New York: Longman..

[46] Deborah K Scherrer and Philip H Scherrer. 1971. An experiment in the computer measurement of melodic variation in folksong. *The Journal of American Folklore* 84, 332 (1971), 230–241.

[47] Federico Simonetta, Filippo Carnovalini, Nicola Orio, and Antonio Rodà. 2018. Symbolic music similarity through a graph-based representation. In *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion.* 1–7.

[48] Wolfram Steinbeck. 1982. *Struktur und Ähnlichkeit: Methoden automatisierter Melodienanalyse.* Vol. 25. Bärenreiter.

[49] Esko Ukkonen, Kjell Lemström, and Veli Mäkinen. 2003. Geometric algorithms for transposition invariant content-based music retrieval. (2003).

[50] Peter van Kranenburg, Berit Janssen, Anja Volk, et al. 2016. The Meertens tune collections: The annotated corpus (mtc-ann) versions 1.1 and 2.0. 1. *Meertens Online Reports* 2016, 1 (2016).

[51] Peter Van Kranenburg, Anja Volk, and Frans Wiering. 2013. A comparison between global and local features for computational classification of folk song melodies. *Journal of New Music Research* 42, 1 (2013), 1–18.

[52] Anja Volk and Peter Van Kranenburg. 2012. Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae* 16, 3 (2012), 317–339.