MELODY EXTRACTION IN SYMPHONIC CLASSICAL MUSIC: A COMPARATIVE STUDY OF MUTUAL AGREEMENT BETWEEN HUMANS AND ALGORITHMS

Juan J. Bosch, Emilia Gómez

Music Technology Group, Universitat Pompeu Fabra, Barcelona

Correspondence should be addressed to: juan.bosch@upf.edu

Abstract: **This work deals with the task of melody extraction from symphonic music recordings, where the term 'melody' is understood as 'the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music and that a listener would recognise as being the "essence" of that music when heard in comparison'. Melody extraction algorithms are commonly evaluated by comparing the pitch sequences they estimate against a "ground truth" created by humans. In order to collect evaluation material from our target repertoire, classical music excerpts in large ensemble settings, we collected recordings of people singing along with the music. In this work, we analyse such recordings and the output of state-of-the-art automatic melody extraction methods, in order to study the agreement between humans and algorithms. Agreement is assessed by means of standard measures that compare pitch sequences on a frame basis, mainly chroma accuracy, which ignores octave information. We also study the correlation between this agreement and the properties of the considered musical excerpts (e.g. melodic density, tessitura, complexity) and of the subjects (e.g. musical background, degree of knowledge of each piece). We confirm the challenges of melody extraction for this particular repertoire, and we identify note density and pitch complexity as the melodic features most correlated to the accuracy and mutual agreement for both humans and algorithms.**

1. INTRODUCTION

Melody is one of the most relevant aspects of music. According to Selfridge-Field [1], 'It is melody that enables us to distinguish one work from another. It is melody that human beings are innately able to reproduce by singing, humming, and whistling. It is melody that makes music memorable: we are likely to recall a tune long after we have forgotten its text'. However, the definition of melody as a musicological concept has evolved through time, depending on the context in which it was proposed [2].

The automatic extraction of melodic information has been approached in the Music Information Retrieval (MIR) community for both monophonic [3] and polyphonic [4] music recordings, motivated by the development of systems for automatic transcription [5], melodic retrieval (e.g. query by humming), music classification and transformation. In the Music Information Retrieval (MIR) literature, melody has been defined as *the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognise as being the 'essence' of that music when heard in comparison* [6]. Pitch is considered as a subjective psychoacoustical attribute of sound, but is closely related to the physical concept of fundamental frequency (f_0), and both terms are commonly used as synonymous in the MIR literature, and so will be in this work.

State-of-the-art melody extraction algorithms and evaluation material have mainly focused on repertoires where singing voice is the only predominant instrument. Contrary to previous work, this research addresses classical music, in particular symphonic repertoire, where the melody is played by different instruments or instrument sections in different pitch ranges. To assess the challenges of this particular repertoire, we need to create annotated material for evaluation. According to the previously mentioned definition of melody, we collected in a previous work [7] a series of recordings of people singing along with music. These recordings were manually analysed to identify and transcribe the melody as ground truth for evaluation, as summarised in Section 2. This manual analysis provides an accurate assessment of the agreement between different annotators, considering both pitch and rhythm information from the recorded melodies, and the musical content of the pieces.

Inspired by a related work on beat estimation [8], this work focuses on the automatic analysis of this agreement. We process the obtained voice recordings and compute mutual agreement between pitch sequences that humans and algorithms considered as the melody of a given excerpt. Given the different pitch ranges under comparison (coming from human voices and symphonic orchestra instruments), we select chroma accuracy as the melody extraction metric, since it ignores octave information (see Section 3). In Section 4, we present the results of the analysis, and we study correlations between mutual agreement and different characteristics of music excerpts and subjects. In Section 5, we further discuss the results and present our conclusions and plans for further work.

2. DATA

We collected a total of 1376 pitch sequences which correspond to the extracted melodies from 86 excerpts of a varied and representative set of symphonic music pieces, which were either sung by humans (1032) or automatically estimated (344) by four different algorithms. After manually analysing the agreement between people, we selected 64 excerpts in which subjects agreed on the melody notes, as detailed in section 2.2.

This dataset includes symphonies and symphonic poems, ballets suites and other musical forms interpreted by symphonic orchestras. They mostly belong to the romantic period, including some classical and 20th century pieces. Music recordings were taken from private collections and selected to have an adequate recording audio quality. They were sampled to create a varied set of short excerpts with a potential dominant melody, maximising the existence of voiced segments (containing a melody pitch) per excerpt.

2.1. Human melody extraction

We gathered human melody extraction data by means of recording sessions, where subjects were asked to 'hum or sing the main melody (understood as the sequence of notes that best represent the excerpt)', focusing on pitch information more than timing (onsets and offsets).

For each audio excerpt, subjects had to carefully listen to it twice and then sing or hum along with the music for three more times. The excerpt length was relatively short (10 to 32 seconds), in order to facilitate subjects to memorise the fragments. During the session, they rated how well they knew each of the excerpts before the experiment (ranking from 1 to 4). After the recordings, they also filled out a survey asking for their age, gender, musical background ("None", "Non formal training", "Formal training less than 5 years" and "Formal training more than 5 years"), dedication to music playing ("No", "Less than 2 hours per week", "More than 2 hours per week", "Professional musician") and confidence rating with their own singing during the experiment, in terms of the percentage of melody notes that they considered they sang correctly ("Less than 30%", "30-60%", "60-90%", "More than 90%").

32 people with a varied musical background (including the authors) participated in the recordings. We discarded 9 subjects which could not properly accomplish the task, based on both their confidence

(those which responded "Less than 30%") and their performance in some excerpts, where the melody was easy to follow. The selected 23 subjects sung a subset of the collection, and were distributed to have three different subjects singing each excerpt. Additionally, the main author sung all excerpts.

The recordings were converted into pitch sequences using probabilistic yin (pYin) [1], a monophonic pitch estimator with temporal smoothing, which has been shown to have higher accuracy than other commonly used algorithms [9]. The step size used corresponds to 5.8 ms. We identified recordings which were not properly converted, due to the fact that the subject was whistling instead of singing. In that case, we employed MELODIA [10][2] in the monophonic setting, with a range between 110 and 1760 Hz to estimate the pitches, which performed better.

2.2. Manual melody annotation

From the previous data, we performed manual melodic transcription of each audio excerpt based on the sung melodies and the underlying music. Inspired by the definitions of melody in [6] and [1], we first selected the excerpts in which human listeners agreed in their 'essence', that is, the sequence of notes that they hum or sing to represent it. In order to do so, we listened to the sung melodies and selected the excerpts with small differences between subjects. This analysis was carried out manually in order to cope with small errors in singing. Given the difficulty of singing some excerpts (fast tempo, pitch range, etc.), the notes sung by the participants were contrasted with the musical score of the piece, mapping them to the played notes, using both pitch and rhythm information from the recorded melodies. The goal was to transcribe the notes that participants intended to sing, allowing some deviations in the sung melodies. Such deviations typically arise from an incorrect singing of some notes (mistuning, timing deviations), notes which were not present in the piece but the participants sung, or from the presence of a chord in the excerpt, in which some subjects sung a different note compared to the rest. Since vocal pitch range may be different to melody instruments range, notes were transposed to match the audio. For excerpts in which melody notes are simultaneously played by several instruments in different octaves, we resolved the ambiguity by maximising the melodic contour smoothness.

In the final selection, we only kept those excerpts in which subjects agreed on most notes of the melody and disagreement was only caused by one participants. In this process, we also considered the reported self-confidence on their singing, giving less importance to notes which disagree with the rest if they were sung by people with less self confidence. From an initial set of 86 excerpts (initial dataset), we selected 64 excerpts (final dataset), and transcribed the notes sang by the annotators. Results presented in Section 4 have been computed in the final dataset, unless otherwise stated.

2.3. Automatic melody extraction

We used melody extraction algorithms to automatically obtain melody pitch sequences in our music collection. We considered relevant state-of-the-art methods based on their performance in the Music Information Retrieval Evaluation eXchange (MIREX) initiative[3] and their availability, ideally as open source software. We used original author implementations of the methods.

There are different strategies for melody extraction, which are divided in the literature into two main approaches: salience-based and separation based methods [4]. The former ones are based on three main steps: pitch salience estimation, melody tracking and voicing detection (whether a melody pitch is present or not). The latter ones perform an initial melody separation stage and then estimate both pitch and voicing. We consider two salience based ([10], and [11]) and two separation based approaches ([12][4], and [13][5]). Salamon and Gómez [10] (SAL) use a harmonic summation based pitch salience function and then create contours, which are used to perform melody tracking and filtering using a set of ad-hoc rules. Dressler [11] (DRE) uses uses a salience function based on pair-wise comparison of spectral peaks, and streaming rules for tracking. The considered method is very similar to [14], except for the frequency range in the selection of pitch candidates, which is narrower in the case of melody extraction. Fuentes et al. [12] (FUE) use Probabilistic Latent Component Analysis on a Constant-Q Transform to build a pitch salience function, and Viterbi smoothing to estimate the melody trajectory. Durrieu et al. [13] (DUR) aim to first model the signal with a source/filter model, then apply Non-negative Matrix Factorisation (NMF) to estimate pitch salience, and a Viterbi algorithm for tracking. Voicing detection (deciding if a particular time frame contains a pitch belonging to the melody or not) is approached differently by the considered algorithms: by means of a dynamic threshold [11], an energy threshold [13, 12], or a strategy based on salience distribution [10].

We adapted the frequency range to the dataset under evaluation (from 103 Hz to 2.33KHz) for all algorithms except those by Salamon and Dressler, which cannot be configured to these values.

2.4. Statistics of the dataset

The final music collection contains 64 audio excerpts with their corresponding annotations. The length of the excerpts ranges from 10 to 32 seconds (mean = 22.1 s., standard deviation = 6.1 s.). The dataset also includes files with a derived sequence of melody pitches, created by dividing the annotations into 10ms frames. If no melody pitch is annotated at a specific time, the frame is considered as unvoiced, otherwise it is consider as voiced. 93.69% of the frames of the dataset are labelled as voiced while 6.31% are unvoiced (in which case the pitch is set to be 0).

The number of excerpts per composer are: Beethoven (13), Brahms (4), Dvorak (4), Grieg (3), Haydn (3), Holst (4), Mussorgski (9), Prokofiev (2), Ravel (3), Rimski-Kórsakov (10), Schubert (1), Smetana (2), Strauss (3), Tchaikovski (2), Wagner (1). Regarding the instrumentation, only in one of the excerpts there is a single instrument (oboe) playing the melody. In the rest of them, the melody is played by several instruments from an instrument section, or a combination of sections, even alternating within the same excerpts. Figure1 (left) illustrates the statistics of the predominant instrumental sections playing the melody. Figure1 (right) depicts the distribution of pitches of all frames of the dataset, and a gaussian model (mean = 74.1, standard deviation = 12.1). Files were converted to mono combining left and right channels before executing the automatic extraction, in order to ensure that all algorithms work with exactly the same material.

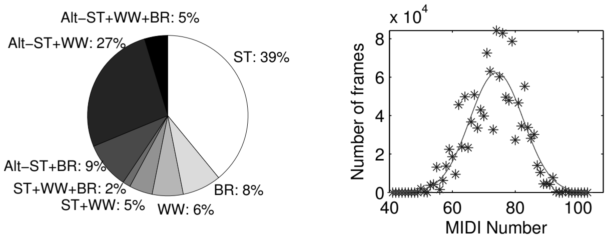

Figure 1: Distribution of the sections of the dominant instruments playing the main melody (left) (ST: Strings, BR: Brass, WW: Woodwinds), where Alt- denotes that the sections alternate within the excerpt. Distribution and gaussian model of the annotated 'melody' pitches (right).

3. Methodology

This section presents the methodology for our study of agreement. To contrast melodies sung by humans and automatically extracted ones, we compare pitch sequences by means of a standard measure

[1] https://code.soundsoftware.ac.uk/projects/pyin

[2] http://mtg.upf.edu/technologies/melodia

[3] http://www.music-ir.org/mirex/wiki/2014:Audio_Melody_Extraction

[4] http://www.benoit-fuentes.fr/articles/Fuentes2012_ICASSP/index.html

[5] https://github.com/wslihgt/separateLeadStereo

used for melody extraction evaluation. One of the challenges for this comparison is that some subjects did not focus on timing, so their sung recordings are not properly aligned to the note onsets in the music. Since evaluation metrics are based on a frame-to-frame comparison, we apply a Dynamic Time Warping (DTW) technique to align both pitch sequences before the metric computation. We then compute Mean Mutual Agreement based on the selected metric, and a set of melodic features to characterise each excerpt.

3.1. Evaluation Measures

Melody extraction algorithms are commonly evaluated by comparing their output against a ground truth. Such pitch sequence is usually created by employing a monophonic pitch estimator on the solo recording of the instrument playing the melody. Pitch estimation errors are then commonly corrected by the annotators. As introduced in Section 2.2, melody annotations are performed on a note level, from which we derive a pitch sequence with time interval of 0.01s. We resampled all pitch sequences to this resolution.

MIREX evaluates separately voicing detection and pitch estimation. Voicing detection is evaluated using two metrics: voicing recall (R_{vx}) and voicing false alarm (FA_{vx}) rates. We define a voicing indicator vector v, whose τ^{th} element (υ_τ) has a value of 1 when the frame contains a melody pitch (voiced), and 0 when it does not (unvoiced). We define the ground truth of such vector as v^*, and $\bar{\upsilon}_\tau = 1 - \upsilon_\tau$ as an unvoicing indicator. Pitch estimation is commonly evaluated using two accuracy metrics: raw pitch (RP) and raw chroma (RC) accuracy.

- Raw Pitch accuracy (RP) is the proportion of melody frames in the ground truth for which the estimation is considered correct (with a certain tolerance (tol)).

$$RP = \frac{\sum_\tau \upsilon^*_\tau \mathscr{T}\left[\mathscr{M}(f_\tau) - \mathscr{M}(f^*_\tau)\right]}{\sum_\tau \upsilon^*_\tau} \qquad (1)$$

$\mathscr{T}$ and $\mathscr{M}$ are defined as:

$$\mathscr{T}[a] = \begin{cases} 1, & \text{if } |a| < tol \\ 0, & \text{else} \end{cases} \qquad (2)$$

$$\mathscr{M}(f) = 12\log_2(f) \qquad (3)$$

where f is a frequency value in Hertz, and tol is a value in semitones, which is commonly set to 0.5 in melody extraction evaluation. In this work, we explore the use of different tolerance values.

- Raw Chroma accuracy (RC) is a measure of pitch accuracy, in which both estimated and ground truth pitches are mapped into one octave

$$RC = \frac{\sum_\tau \upsilon^*_\tau \mathscr{T}\left[\| \mathscr{M}(f_\tau) - \mathscr{M}(f^*_\tau) \|_{12}\right]}{\sum_\tau \upsilon^*_\tau} = \frac{N_{ch}}{\sum_\tau \upsilon^*_\tau} \qquad (4)$$

where $\| a \|_{12} = a - 12\lfloor \frac{a}{12} + 0.5 \rfloor$, and N_{ch} represents the number of chroma matches.

- Overall Accuracy (OAC) represents the proportion of frames that were correctly labelled in terms of both pitch and voicing

$$OAC = \frac{1}{N_{fr}} \sum_\tau \upsilon^*_\tau \mathscr{T}\left[\mathscr{M}(f_\tau) - \mathscr{M}(f^*_\tau)\right] + \upsilon^*_\tau \overline{\upsilon}_\tau \qquad (5)$$

where N_{fr} is the total number of frames

We use the superscripts 'a', and 'h' to denote if the metrics refer to algorithms or human melody extraction respectively, e.g. RC^a refers to the raw chroma accuracy obtained by algorithms.

In this work, we compare pitch sequences sung by humans and those obtained by melody extraction from orchestral music recordings. As they are typically in a different pitch range, we focus in chroma accuracy to compare sequences of pitches, in order to discard octave information.

3.2. Dynamic Time Warping

As mentioned above, pitch sequences of the sung melodies are not properly aligned with the ground truth, the result of the automatic methods, or with other subjects singing, since people commonly sung delayed with respect to the note onsets. In order to minimise this effect, we align both sequences using a Dynamic Time Warping (DTW) technique [6] based on chroma information, which has been extensively used for a number of tasks such as cover version identification, and audio-to audio or audio-score alignment [15].

3.3. Mean Mutual Agreement

In a previous study on beat tracking [8], beat extraction evaluation metrics were used to compute the agreement $A_{i,j}$ between algorithms i and j which aim at automatically identifying the sequence of beats in an audio excerpt. In the present work, we adapt the concept of agreement to the task of melody extraction, and use raw chroma accuracy as it is the most relevant metric for this particular context. The agreement between two sequences i, j, which correspond to the estimated melody of an excerpt k ($A_{i,j}[k]$) is here equal to the raw chroma accuracy when using sequence j as ground truth and i as the estimation, for an excerpt k. With such definition, it is important to note that we obtain different results depending on the sequence used as ground truth, so $A_{j,i}$ is different to $A_{i,j}$. We define Mutual Agreement as:

$$MA_i[k] = \frac{1}{N-1}\dot{\sum}_{j=1,\, j\neq i}^{N} A_{i,j}[k], \quad MA = \frac{1}{N_{exc}}\dot{\sum}_{k=1}^{N_{exc}} MA[k] \qquad (6)$$

where N is the total number of estimators (algorithms or subjects), i is the index of the estimator, k the excerpt number, and N_{exc} the total number of excerpts in the collection.

We define the Mean Mutual Agreement for an excerpt k (MMA[k]) as the average MA[k] for all N estimators, and MMA as the average Mean Mutual Agreement for all excerpts.

$$MMA[k] = \frac{1}{N}\dot{\sum}_{i=1}^{N} MA_i[k], \quad MMA = \frac{1}{N_{exc}}\dot{\sum}_{k=1}^{N_{exc}} MMA[k] \qquad (7)$$

We use MMA^a to denote MMA for algorithms, and MMA^h for humans.

3.4. Melodic features

We extracted a set of melodic descriptors to characterize music excerpts and analyze how different characteristics are related to melody extraction accuracy and MMA for both human and algorithms. We used MIDI Toolbox[7] to extract the following characteristics from the MIDI files of the annotated melodies:

- **Density**: amount of notes per second.
- **Range**: difference in semitones between highest and lowest note pitch.
- **Tessitura**: melodic tessitura based on deviation from median pitch height [16]
- **Complexity (pitch, rhythm, mixed)**: expectancy-based model of melodic complexity [17] based either on pitch or rhythm-related components, or on an optimal combination of them together.
- **Melodiousness**: 'suavitatis gradus' proposed by Euler, which is related to the degree of softness of a melody, and is a function of the prime factors of musical intervals [18].
- **Originality**: Different measurement of melodic complexity, based on tone-transition probabilities [19].

3.5. Data and methodology overview

The collection creation schema, recordings and information used for the computation of the MMA^h and correlation is shown in Fig.2.

[6] `http://www.ee.columbia.edu/ln/LabROSA/matlab/dtw/`

[7] `https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/miditoolbox`

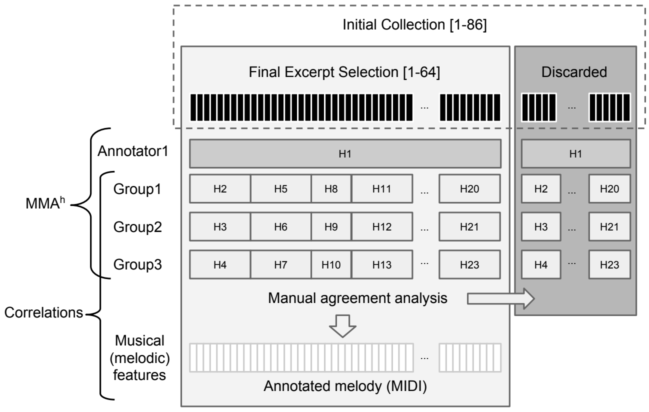

Figure 2: Final collection creation and data analysis schema. H1, H2, etc. refer to the recordings of each of the subjects, which correspond to several excerpts

As mentioned in section 2, we collected singing data from 4 different subjects for each excerpt, one of them being always the main author (Annotator1). Since the amount of excerpts sung by Annotator1 was thus much higher than for the rest of subjects, we did not include his data in the computation of the RC^h, nor in the statistical analyses, so as not to bias them. In the case of Mean Mutual Agreement, we used his recordings since we are only analyzing musical factors.

4. Results

In this section we analyze mutual agreement and its correlation with the mentioned melodic descriptors and subject related factors. We first analyse mutual agreement between humans vs melody annotations and algorithms vs melody annotations. We then study mutual agreement between different humans and different algorithms. Raw chroma accuracy and mutual agreement results are provided as the average value for all the excerpts in the database.

4.1. Agreement between humans and symbolic annotations

We first compare sung melodies against annotated melodies in symbolic format (ground truth). For each of the three takes we recorded, we compute the average RC^h (μ_{RC^h}) for all excerpts and the three subjects. In order to understand the influence of mistunings, we increased the tolerance in the evaluation measure from 0.5 to 1.5 semitones in steps of 0.25 semitones, as shown in Table 1.

Table 1: μ_{RC^h} in different takes, and with different tolerances in semitones

	0.5	**0.75**	**1**	**1.25**	**1.5**
Take1	37.9	47.5	53.7	58.5	61.9
Take2	40.3	50.5	56.7	61.4	64.7
Take3	43.4	53.5	59.6	64.3	67.4

Results show that accuracy values are relatively small in general, due to differences in timing and tuning between the compared pitch sequences. In fact, the tolerance has a very clear impact in the results, indicating the presence of small mistunings between both pitch sequences. In addition, we observe that subjects moderately increase their accuracy with each new take, with a relatively constant increase when evaluating with different interval tolerances. This suggests that this is not only due to a correction of mistunings but to the refinement of the selection of notes belonging to the melody. The results reported in the rest of the paper have been computed using the third take, with *tol* = 1 semitone, in order to allow some flexibility with human errors in tuning. In the case of algorithms we keep the standard value of *tol* = 0.5 semitones, since they are less affected than humans by tuning.

As mentioned, we identified timing deviations in subjects' singing in relation to notes onsets and offsets, which ideally should not be considered when measuring the overall agreement. We apply the DTW algorithm introduced before, allowing only temporal deviations between -0.5 to 0.25 seconds, since subjects were typically slightly delayed. After the alignment, RC^h increases from 59.6 to 67.4% for the considered take (3) and tolerance (1 semitone).

Table 2: Correlation of melodic features with raw chroma accuracy obtained by humans in original and aligned singing, and algorithms

	RC^h	$RC^{h_{al}}$	RC^a
excerpt knowledge	0.16	0.11	NA
age	-0.17	-0.16	NA
range	-0.37	-0.37	-0.13
density	-0.43	-0.35	-0.44
tessitura	0.06	0.06	0.06
pitch complexity	-0.45	-0.38	-0.33
rhythm complexity	-0.26	-0.21	-0.11
mixed complexity	-0.37	-0.32	-0.22
melodiousness	-0.08	-0.03	-0.05
originality	-0.01	-0.01	-0.12

Table 2 shows the Pearson correlation between RC^h and the melodic features of the considered music excerpts, for both original and aligned pitch sequences. According to Table 2, there is a strong correlation of raw chroma accuracy with several musical parameters such as melodic range, density and melodic complexity (in pitch, rhythm and mixed). Correlation is negative with all of the mentioned factors, and the strongest one is mixed complexity. There is no strong correlation with melodiousness, originality and tessitura.

Next we performed a variance decomposition analysis to study the individual contribution of each factor to the observed variance in the responses. We started with the saturated random-effects linear model containing all main factors, and iteratively simplified it by removing factors whose effect was not statistically significant (alpha=0.05). Once the model was simplified, we ran an ANOVA analysis and computed the individual contributions to total variance. Table 3 shows the percentage of variance in raw chroma accuracy.

Table 3: % of variance in human raw chroma accuracy, for both original and aligned pitch sequences

	% Var	**% Var. (aligned)**
density	26.69	16.03
range	15.66	19.35
excerpt knowledge	0.12	0.00
time playing	1.19	1.65
self confidence	1.18	0.90
musical background	24.99	26.54
sex	4.18	4.16
residual	25.98	31.34

We observe that most variance in RC^h is due to note density, melodic range, and also the musical background. People without musical training obtained lower accuracies than those with formal background or with non formal training. There is also some residual variance that could be either related to subjects' or excerpts' characteristics which we are not considering, or from interactions between the considered factors. In both previous tables we observe that the effect of note density decreases in the case of aligned performances, since we allow temporal deviations.

An important result of Table 2 and Table 3 is that the methodology followed in the recording sessions (subjects listened to four repetitions of the excerpt before the considered take) was enough to ensure that the degree of knowledge of the excerpt before the recording session would not affect the extracted melody. Table 2 shows that there is only a very small correlation of RC^h with user knowledge of the excerpt for non aligned pitch sequences, and Table 3 shows that there is no variance due to the excerpt knowledge.

4.2. Agreement between algorithms and manual annotations

The results obtained by comparing the output of melody extraction algorithms against the annotated melodies are shown in Table 4.

Table 4: Values for Raw Pitch (RC^a), Raw Chroma (RC^a) and Overall Accuracy (OAC^a) obtained by algorithms

	RP	**RC**	**OAC**
DRE	49.4	66.5	46.0
DUR	66.9	80.6	62.6
FUE	27.8	60.2	24.1
SAL	28.5	57.0	23.5

According to these results, the highest accuracies are obtained by the algorithm proposed by Durrieu, specially in the case of RP and OAC. For RC, Durrieu achieves 80.6% accuracy, 14.1% above the method by Dressler, which is the following one in terms of RC.

Third column in Table 2 shows the correlation of RC^a with musical properties of the excerpts. We observe that the highest (negative) correlation found belongs to melodic density, followed by pitch complexity. Originality and range have a very small negative correlation. In the case of range, we observe that there is a difference between algorithms and humans, since humans are more (negatively) influenced by the melodic range. A variance components analysis (see Table 5) shows that most variance comes from algorithm ID, followed by melodic density, pitch complexity, and tessitura, and a residual variance of 23.7%.

Table 5: Percent of variance in RC^a due to different factors

	% Var.
mixed complexity	1.03
rhythm complexity	9.10
pitch complexity	16.59
tessitura	12.45
range	0.64
algorithm ID	21.22
residual	23.70

4.3. Mutual agreement between humans

We computed the Mean Mutual Agreement (MMA) between all subjects (MMA^h), with three different tolerances (tol) for raw chroma accuracy computation. With a small tolerance (tol=0.5 semitones), Mean Mutual Agreement in humans only achieves 37.71%. We would expect the agreement to be higher than this value, since MMA was computed for the final excerpt selection, in which the manually analysed agreement between subjects was very high. However, by increasing the tolerance to 1 semitone, MMA^h increases up to 57.47%, and if we allow a deviation in chroma of 1.5 semitones, we achieve a mutual agreement of 68.16%, as we do not penalise possible mistunings.

The identified temporal deviations in subjects' singing also affect MMA. We now perform an alignment using DTW, as previously presented, but allowing temporal deviations between -1 and 1 second, since we need to align sequences of pitches produced by 2 subjects and need thus to consider higher differences in timing, e.g. when one subject is singing too early, and the second one is delayed. After the alignment, we increase from MMA^h = 57.47% to a $MMA^{h_{al}}$ = 76.07 % for a tolerance of 1 semitone.

Table6 shows the correlation of MMA^h and $MMA^{h_{al}}$ with musical properties of the annotated melody. Similarly to the case of humans, density, range, pitch complexity and mixed complexity are the factors more strongly (negatively) correlated with MMA.

We now investigate if the manual selection of excerpts explained in Section 2.2 could have been automatically performed by selecting those excerpts with high $MMA^{h_{al}}$. The mean (μ) and standard deviation (σ) of $MMA^{h_{al}}$ for discarded excerpts is: μ = 62.53, σ = 9.51. In the case of the selected excerpts: μ = 76.07, σ = 10.14. While the mean MMA in the selected excerpts is higher, some manually discarded excerpts have higher MMA than others which had been selected. As previously introduced in section 2.2, this is due to the fact that in the manual selection process we considered not just pitch but also rhythm information from the singing which is not considered in the automatic analysis of agreement, as well as the musical content of the piece. An example of this fact is shown in Fig.3, where we observe differences in the pitch sequences sung by the subjects (ignoring octave information). The agreement between humans in this excerpt is quite low (MMA^h = 34.41%, $MMA^{h_{al}}$ = 56.73%), and the agreement between algorithms is higher (MMA^a = 63.44%). However, a manual analysis of the recordings (contrasting subjects singing to the musical content of the piece) revealed that participants agreed in most of the notes (corresponding to the annotated melody). Disagreement was due to the large melodic range, and difficulties in singing some notes.

Table 6: Correlation between musical factors and MMA

	MMA^h	$MMA^{h_{al}}$	MMA^a
range	-0.49	-0.46	-0.14
density	-0.58	-0.48	-0.38
tessitura	0.10	0.09	0.12
pitch complexity	-0.62	-0.55	-0.33
rhythm complexity	-0.34	-0.25	-0.05
mixed complexity	-0.53	-0.44	-0.21
melodiousness	-0.035	-0.08	-0.06
originality	0.01	0.05	-0.22

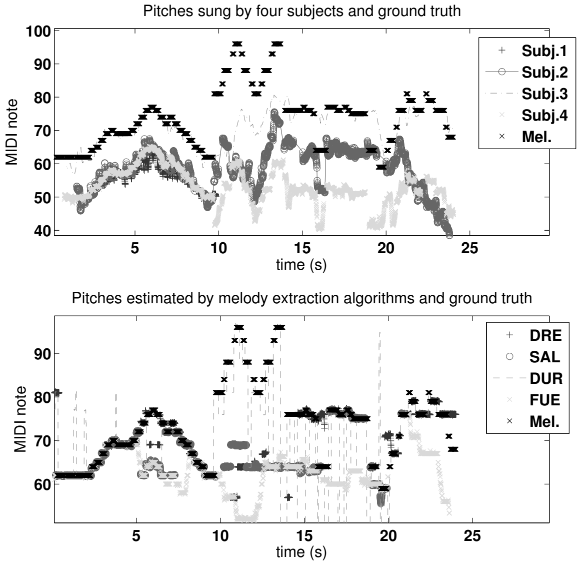

Figure 3: Sequences of pitches sung by the four subjects, four algorithms and the ground truth annotation for the melody.

4.4. Mutual agreement between algorithms

We computed the agreement between algorithms by comparing pairs of automatically extracted pitch sequences. In this case, there is no need to perform any temporal alignment. The highest agreement according to Table 7 is 69.7% obtained between DRE (ground truth) and DUR (estimator). The lowest agreement is 57.1%, between FUE (ground truth) and SAL (estimator). The Mean Mutual Agreement between algorithms is MMA^a = 61.71%.

The correlation analysis of MMA^a is provided in the third column of Table 6. Note density, pitch complexity, originality and mixed complexity are the factors more correlated to the MMA. It is interesting to note that range and rhythm complexity have much lower (negative) correlation than in the case of humans, meaning that algorithms are more robust to them. Originality has a medium positive correlation with MMA^a but practically no correlation with MMA^h, which suggests that the higher the melodic originality the

Table 7: Agreement between algorithms $A^a_{i,j}$, where the names of the rows represent the ground truth sequence, and the column names represent the estimation

	DRE	**SAL**	**DUR**	**FUE**
DRE	100.0	66.1	69.7	62.4
SAL	64.6	100.0	57.5	59.0
DUR	68.1	57.4	100.0	60.0
FUE	59.2	57.1	58.2	100.0

least algorithms will agree.

4.5. *Mutual agreement between subjects and algorithms*

We compute the correlation between subjects and algorithms MMA, which is quite strong (0.3) as shown in Table 8. We also computed the correlations between MMA and RC, which are strong when they both refer to humans or algorithms, but correlation between MMA^a and RC^h or between MMA^h and RC^a is weaker.

Table 8: Correlation between Mean Mutual Agreements and with raw chroma accuracies

	MMA^h	$MMA^{h_{al}}$	MMA^a
MMA^h	1	0.93	0.3
$MMA^{h_{al}}$	0.93	1	0.3
MMA^a	0.3	0.3	1
RC^h	0.58	0.53	0.21
$RC^{h_{al}}$	0.48	0.46	0.15
RC^a	0.24	0.18	0.63

5. Conclusions and further work

After analysing several kinds of agreements between both humans and algorithms in the task of melody extraction in symphonic classical music, we observed that some melodic features of the excerpts are correlated to accuracy results. The analysis of agreement between melody extraction and manual melody annotations showed that melodic range and note density have a clear negative correlation with accuracy results obtained by people. In the case of algorithms, the highest (negative) correlation is with note density, and results suggests that algorithms are less affected by melodic range than humans, as long as pitches are kept within their limits of operation. With regard to subject-related factors, we found out that previous knowledge of an excerpt had almost no correlation with the accuracy obtained by humans, and no contribution to total variance, which validates the proposed design for the recording data gathering in our dataset. With regard to automatic melody extraction, the mean raw chroma accuracy of the four algorithms is 66.1%, but with important differences between them. Durrieu's approach obtains the highest scores, reaching 80.6% raw chroma accuracy.

In the case of Mean Mutual Agreement, we observed a negative correlation with melodic density and complexity (specially pitch complexity), in both humans and algorithms. Comparing the agreement between humans and algorithms, we observed that excerpts with a higher melodic originality make algorithms differ more in their estimations than in the case of humans. Finally, we identified a strong positive correlation between raw chroma accuracy and Mean Mutual Agreement, for both humans and algorithms. However, there is a lack of a strong correlation between the raw chroma accuracy obtained by humans and the Mean Mutual Agreement obtained by algorithms, and vice versa.

Future work deals with an analysis considering more musical factors, related to other characteristics of the excerpt. These could be automatically computed, such as the ratio of energy of the melody in relation to the accompaniment, or manually annotated by a musicologist. Such analysis will give further insights of the way musical factors affect humans and algorithms, and such knowledge will be used for the creation of an automatic melody extraction method, which is able to deal with the challenging characteristics of symphonic music.

6. Acknowledgements

We would like to thank volunteers for participating in the recordings, and authors of melody extraction algorithms for giving us access to their code or execution results. Thanks to Julian Urbano for his comments and suggestions. This work is supported by the European Union Seventh Framework Programme FP7 / 2007-2013 through the PHENICX project (grant agreement no. 601166).

References

[1] E. Selfridge-Field: *Conceptual and representational issues in melodic comparison*. In *Computing in musicology: a directory of research*, (11):3–64, 1998.

[2] A. L. Ringer: *Melody. Grove Music Online, Oxford Music Online. [Online]. Available: http://www. oxfordmusiconline.com/subscriber/article/grove/music/18357*, 2014.

[3] E. Gómez, A. Klapuri, and B. Meudic: *Melody description and extraction in the context of music content processing*. In *Journal of New Music Research*, volume 32(1):23–40, 2003.

[4] J. Salamon, E. Gómez, D. Ellis, and G. Richard: *Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges*. In *IEEE Signal Process. Mag.*, volume 31:118–134, 2014.

[5] A. Klapuri, M. Davy, et al.: *Signal processing methods for music transcription*, volume 1. Springer, 2006.

[6] G. Poliner, D. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. Ong: *Melody transcription from music audio: Approaches and evaluation*. In *Audio, Speech, Lang. Process. IEEE Trans.*, volume 15(4):1247–1256, 2007.

[7] J. Bosch, R. Marxer, and E. Gómez: *Evaluation and Combination of Pitch Estimation Methods for Melody Extraction in Symphonic Classical Music*. (In Preparation).

[8] J. R. Zapata, M. Davies, and E. Gómez: *Multi-feature beat tracking*. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 22:816 – 825, 2014.

[9] M. Mauch and S. Dixon: *PYIN: A fundamental frequency estimator using probabilistic threshold distributions*. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 659–663. 2014.

[10] J. Salamon and E. Gómez: *Melody extraction from polyphonic music signals using pitch contour characteristics*. In *IEEE Trans. Audio. Speech. Lang. Processing*, volume 20(6):1759–1770, 2012.

[11] K. Dressler: *Towards Computational Auditory Scene Analysis: Melody Extraction from Polyphonic Music*. In *Proc. 9th CMMR*. 2012.

[12] B. Fuentes, A. Liutkus, R. Badeau, and G. Richard: *Probabilistic model for main melody extraction using constant-Q transform*. In *IEEE ICASSP*, pages 5357–5360. IEEE, 2012.

[13] J. Durrieu, G. Richard, B. David, and C. Févotte: *Source/filter model for unsupervised main melody extraction from polyphonic audio signals*. In *Audio, Speech, Lang. Process. IEEE Trans.*, volume 18(3):564–575, 2010.

[14] K. Dressler: *Multiple fundamental frequency extraction for MIREX 2012*. In *Music Inf. Retr. Eval. Exch.*, 2012.

[15] J. Bosch, K. Kondo, R. Marxer, and J. Janer: *Score-informed and timbre independent lead instrument separation in real-world scenarios*. In *Proc. Signal Processing Conference (EUSIPCO), 2012*, pages 2417–2421. 2012.

[16] P. Von Hippel: *Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals*. In *Music Perception*, pages 315–327, 2000.

[17] T. Eerola and A. C. North: *Expectancy-based model of melodic complexity*. In *Proc. Int. Conf. Music Perception and Cognition*. 2000.

[18] M. Leman: *Music and schema theory: cognitive foundations of systematic musicology*. Springer, 1995.

[19] D. K. Simonton: *Melodic structure and note transition probabilities: A content analysis of 15,618 classical themes*. In *Psychology of Music*, volume 12(1):3–16, 1984.