TROMPA: Towards Richer Online Music Public-domain Archives

Deliverable 5.4

Music Performance Assessment v2

Grant Agreement nr	770376
Project runtime	May 2018 – April 2021
Document Reference	TR-D5.4-Music Performance Assessment Mechanisms v2
Work Package	WP5 - TROMPA Contributor Environment
Deliverable Type	Report
Dissemination Level	PU - Public
Document due date	28 February 2021
Date of submission	28 February 2021
Leader	mdw
Contact Person	Werner Goebl (goebl@mdw.ac.at)
Authors	David M. Weigl, Werner Goebl (mdw), Helena Cuesta (UPF)
Reviewers	Cynthia Liem (TUD), David Linssen (VD), Aggelos Gkiokas (UPF)

Executive Summary

One of the aims of TROMPA is to formalise expert (musicologists and educators) and crowd (music enthusiasts) knowledge on various aspects of performances and musical scores in terms of performance quality (such as intonation and voice quality in case of singing or technical brilliance in case of instrumental music) and of piece difficulty (i.e., the difficulty of performing a piece as a singer or instrumental player).

The notion of “score difficulty” is a complex construct, dependent upon (and hence, necessarily to be understood relative to) the expertise and skill level of the performer. Further, motoric-physiological aspects of difficulty depend strongly on the particularities of specific instruments, the tempo of a performance, and on the playing style called for by the piece or preferred by the individual performer. As such, different experts may disagree in their assessment of piece difficulty, depending on how they weigh these diverse aspects in their judgement; and automatic algorithms to determine such a measure must therefore be understood as coarse-grained abstraction of the many facets influencing the performance difficulty of a particular score. Similarly, a performance’s quality (in the sense of “goodness”) is difficult to pin down, representing a highly subjective notion likely to confound consistent ratings even among human judges.

Section 2 of this deliverable details the conceptualisations of performance quality and score difficulty adopted for TROMPA’s performance-based use-case prototypes (D6.4 - Working prototype for instrumental players; D6.5 - Working prototype for choir singers) by virtue of quantifiable measures that arise implicitly from the musical-score-aligned performances recorded with these prototypes. Corresponding features include intonation accuracy (choir singers) and degree of conformity of performances to the score (instrumental players). Performance errors not only provide a cue to the performance quality of individual renditions, they can also be aggregated across renditions to give insights into score difficulty, by observing the incidence of errors at certain score sections over time.

In Section 3, we operationalise these measures in terms of empirically observed performance errors -- inserted and deleted notes for instrumental players, and intonation inaccuracies for choir singers. These may be automatically determined in TROMPA’s prototypes, through the use of digital performance capture, and alignment to semantic score encodings. Further, we identify how interpretive aspects of individual performances -- features such as tempo, rubato, and performance dynamics -- may be aggregated to determine the *typicality* of individual performances, a notion connected to aesthetic quality judgements in the literature.

In Section 4 we outline the technical implementation of workflows developed to determine these measures and present them to the user, including the performance-to-score alignment workflow (detailed further in D3.5 - multimodal music information alignment) used to determine inserted and omitted notes in instrumental performances, and the singing performance assessment workflow yielding intonation scores by the application of an algorithm discussed in section section 3.2.

Finally, we consider directions for future research in Section 5, to investigate hypotheses of score difficulty, including cognitive-structural aspects (“difficulty of understanding”) and motoric-physiological aspects (“difficulty of physically realising”); such investigations depend on the availability of semantic, highly-granular, empirical performance data, of the type generated by the assessment workflows presented here.

Version Log		
#	Date	Description
v1.1	19 February 2021	Initial version submitted for internal review
v1.2	26 February 2021	Revised version after internal review
v2.0	28 February 2021	Final version submitted to EU

Table of Contents

1. Introduction

1.1 Evolution of this deliverable

The TROMPA project proposal describes the aims of Task 5.4 (music performance assessment mechanisms) and the associated deliverable, the final version of which is presented in this document, as follows:

"Task 5.4 aims at investigating ways to formalise expert (musicologists and educators) and crowd (music enthusiasts) knowledge on various aspects of the multiple performances and the scores linked to them (T3.5 and T5.3) in assessing their individual qualities. Strategies will be defined to identify valid representations of overall (per performance or piece) and detailed (section-wise, T3.6) ratings of individual aspects of performance quality (such as intonation and voice quality in case of singing or technical brilliance in case of instrumental music). Additionally, properties of particular pieces (e.g. difficulty to perform a given composition) may be included as well. These subjective representations are combined with automatic descriptors delivered in T3.2 and integrated into the pilot applications of WP6.

Deliverable 5.4 will include automatic models to predict performance difficulty of (solo) instrumental scores and choir scores to rate the quality of their performance. At first, "difficulty indices" to denote performance difficulty of a given piece are specified and validated on preliminary data (e.g., solo lute, piano and choir repertoire). Subsequently, performance quality assessment metrics involving audio and score information are developed and systematically validated via human feedback."

This document presents the final version of this deliverable, the first version of which[1] was published in M12 of the project, at a very early stage of software implementation. As such, the current version concretises the performance assessment mechanisms that have been implemented in TROMPA's two performance-centric use-case prototypes (**D6.5-Working prototype for instrumental players**[2]; **D6.6-Working prototype for singers**[3]). It has been restructured, summarising the previously more prospective Sections 2 ("Which qualities are being assessed?") and 3 ("What kinds of data could be subject to assessment?") within the current version's more descriptive Section 2 ("Assessment measures"). The previous Section 4 ("Assessment mechanisms") has been rescoped in the present document's Section 3, with sharpened focus on the underpinnings of our implementations; which are then detailed in this document's Section 4 (previously Section 5) on "Infrastructure for performance assessment". Finally, the newly introduced Section 5 presents directions for future investigation opened up by TROMPA's prototypes and associated data infrastructure, which provide the means for generating large quantities of performance data, interlinked at note-level precision with semantic encodings of the score being performed.

This restructuring represents an evolution in approach to performance assessment as development has proceeded within the project: while we had initially considered implementing integrated performance assessment components that would serve both performance-centric use cases (instrumental players and singers), subsequent implementation -- informed by requirements

[1] https://trompamusic.eu/deliverables/TR-D5.4-Music_Performance_Assessment_v1.pdf

[2] https://trompamusic.eu/deliverables/TR-D6.5-Working_Prototype_for_Instrument_Players_v2.pdf

[3] https://trompamusic.eu/deliverables/TR-D6.6-Working_Prototype_for_Singers_v2.pdf

gathering and user studies, see D6.1[4] (Final mockups testing) and D6.8[5] (Mid-term evaluation) -- have clear that separate implementations better serve our users' needs: whereas we have converged on the notion of performance assessment informed by empirical observations of performance errors, the nature of such errors are sufficiently divergent (see Section 3.1) that software implementation has instead proceeded separately.

1.2 Automatic performance assessment

Automatic assessment of score difficulty is far from trivial. The aggregated notion of "difficulty" is in fact a complex construct, involving both cognitive-structural aspects ("difficulty of understanding") and motoric-physiological aspects ("difficulty of physically realising"). The ability to tackle facets of both aspects of difficulty are dependent upon (and hence, must be understood relative to) the expertise and skill level of the performer. Further, the motoric-physiological aspects depend strongly on the particularities of particular instruments, the tempo of a performance, and on the playing style called for by the piece or preferred by the individual performer. As such, different experts may disagree in their assessment of piece difficulty, depending on how they weigh these diverse aspects in their judgement; and automatic algorithms to determine such a measure must therefore be understood as coarse-grained abstraction of the many facets influencing the performance difficulty of a particular score. Similarly, a performance's quality (in the sense of "goodness") is difficult to pin down, representing a highly subjective notion likely to confound consistent ratings even among human judges.

Through its prototype applications for instrumental players (D6.5) and choir singers (D6.6), TROMPA has developed a means of generating substantial corpora combining performance recordings with note-level descriptors of aspects such as intonation accuracy, dynamic variability and temporal deviation. These provide empirical evidence for aggregate measures of performance quality and score difficulty we describe in the remainder of this document. They also provide grounding for future investigations into variegated granular measures that fall outside the scope of the project.

2. Assessment measures

In this section, we provide conceptualisations of performance quality and score difficulty as informed by the data generated in TROMPA's performance-centric use-case prototypes: empirical observations of the musician's adherence to the score during performance. Such observations are made possible by the use of digital music encodings which capture the musical meaning of the score in an algorithmically-accessible way, allowing the score-fidelity of captured performances to be automatically characterised according to music semantics.

In Section 3, we will operationalise the concepts presented here, incorporating aspects of the data models representing relevant aspects of the captured performances; while Section 4 will detail the implementation of these concepts within TROMPA's data infrastructure.

[4] This deliverable is confidential to the consortium only

[5] https://trompamusic.eu/deliverables/TR-D6.8-Mid_Term_Evaluation.pdf

2.1 Data resources subject to assessment

Two of TROMPA's prototypical use-cases -- D6.5 (instrumental players) and D6.6 (choir singers) -- are designed around interactions that involve users' performances of scores represented as digital music encodings, recorded as MIDI streams or audio signals. These performances are aligned with the score encodings on a note level, either by virtue of the performance being recorded against a reference timeline in the case of choir singers, or by use of the automated music alignment workflow described in **D.3.5-Alignment of Musical Resources**[6] in the case of instrumental players.

These alignments allow the granular comparison of different performance recordings, and inject musical semantics into the recorded signals by reference to the aligned music encodings. Conversely, features derived from the recorded signals provide performative context to the static, invariable score encodings. Together, these qualities support empirical assessments of both performance quality and score difficulty.

Assessment measures implemented in TROMPA are based on measurement of salient features of performance recordings, including timing, intonation, performance dynamics, and the incidence of errors, as well as derivative measures that observe patterns in these features across multiple performances.

2.2. Empirical measures of performance quality

The digital music encodings used in TROMPA's applications provide ground-truth music information against which the features of individual performance recordings can be compared in order to empirically identify performance errors. These features include intonation accuracy (choir singers; in the sense of pitch tuning deviations from equal temperament, e.g., Pfordresher et al, 2010, even though other tuning systems such as just intonation or Pythagorean tuning may as well be defined) and degree of conformity to the score in terms of omitted or inserted notes (instrumental players; see Flossmann et al, 2009). Though the quality of a performance is a complex, subjective notion, the presence or absence of performance error provide an approximate measure that can be operationalised through such comparisons.

2.3 Aggregate measures of performance quality

Beyond the identification of performance errors, empirical measurements determined in the TROMPA prototypes also include expressive measures of tempo and timing (rubato) (instrumental players), and of variation in dynamics (choir singers and instrumental players). Such measures are used to provide users with opportunities for reflection on review of individual recorded performances. In aggregate, such measures also make it possible to determine variability across collections of performances. This allows features such as performance tempo (of movements, sections, subsections etc.) to be put in relation to aggregated performance data or tempo indications from the score (see e.g., Kolisch, 1993). It also supports the determination of performances exhibiting features that are typical (most representative of) entire collections, yielding another measure of performance quality, on the assumption that typicality correlates with aesthetic perception, as suggested by the literature (see e.g. Page et al, 2017; Repp 1997; Wöllner 2012; Wolf et al, 2018).

[6] https://trompamusic.eu/deliverables/TR-D3.5-Multimodal_Music_Information_Alignment_v2.pdf

2.4 Aggregate measures of score difficulty

Aggregations of performance measurements across performance collections also yield cues of score difficulty, on the assumption that the number of rehearsal repetitions and performance errors, and derivative error measures (such as the rate of improvement) across performances correlates with the difficulty of the score to be performed.

2.5 Importing assessments from other sources

Due to Web-based nature of TROMPA's data infrastructure (D5.1)[7], expert assessments from other sources can be linked in: e.g., Henle difficulty ratings[8], UK Associated Board of the Royal Schools of Music (ABRSM) classifications[9]. The empirical difficulty measures outlined above provide interesting opportunities for validation of such assessments, which though informed by expertise are necessarily subjective in nature.

3. Assessment mechanisms

In this section, we expand on the conceptualisations of performance quality and score difficulty presented in Section 2, providing operationalisations incorporating the empirical data generated in D6.5 and D6.6. Difficulty and quality assessments are provided by implicit measures derived from musicians' performance behaviour (e.g., by analysing error rates established in performance-score alignment), and by algorithmic processing (feature extraction) of scores and performance recordings applying TROMPA Work Package 3 technologies (particularly D3.2[10] and D3.5).

3.1. Performance errors

For instrumental performances, the process of performance-to-score alignment (D3.5) produces a set of "note-deletion" and "note-insertion" events corresponding to notes that were specified in the score but not played during a performance, or conversely notes that are not specified in the score but nevertheless played (see e.g., Flossmann et al, 2009). Events in the former category may be meaningfully visualised for assessment by the user in a musical score context, as they have a position in the musical notation prescribed by the score. Events in the latter category involve the sounding of notes not prescribed in the score, and so visualising these in a notation context requires some approximation: by virtue of the performance metadata available for validly performed notes as well as for inserted (but not deleted) ones, a score position for inserted notes may be interpolated by reference to the score positions associated with preceding or succeeding validly performed notes. The corresponding visualisations implemented in the working prototype for instrumental players (CLARA; D6.5) are illustrated in Figure 3.1.

[7] https://trompamusic.eu/deliverables/TR-D5.1-Data_Infrastructure_v2.pdf
[8] https://www.henle.de/en/about-us/levels-of-difficulty-piano/ (accessed 24 April 2019)
[9] http://abrsm.org (Retrieved 24 April 2019)
[10] https://trompamusic.eu/deliverables/TR-D3.2-Music_Description_v2.pdf

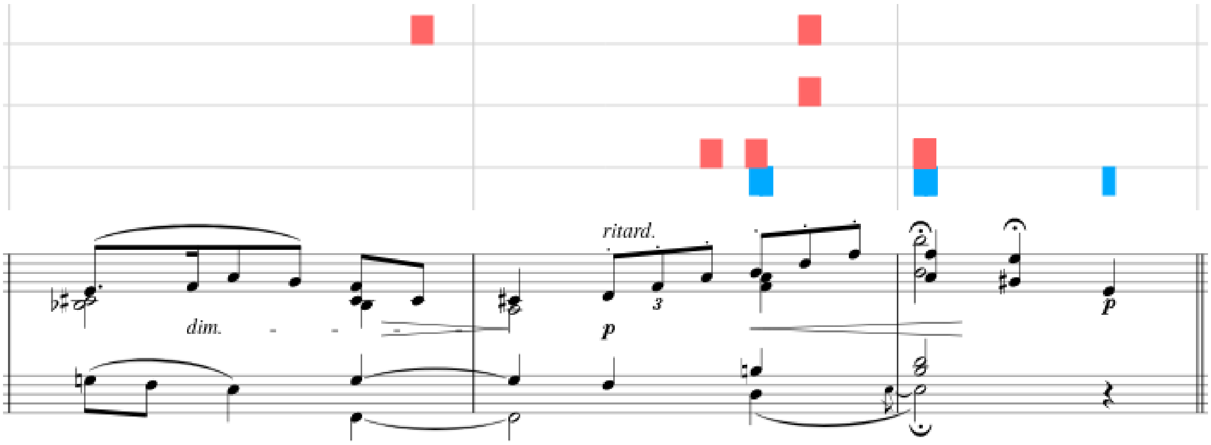

Figure 3.1. Error visualisation in the Working prototype for instrumental players (CLARA; D6.5). Horizontal lines correspond to individual rehearsal recordings; red markers above a line indicate presence of inserted notes; blue markers below a line indicate presence of deleted (omitted) notes. Markers are aligned with the corresponding score position (below the visualisation), and can be clicked to jump audiovisual playback to the appropriate temporal position in the performance timeline.

Together, these events identify performance "errors," providing a crude indication of performance "quality," as well as a signature of score difficulty: score sections consistently prone to producing errors across performances presumably exhibiting greater difficulty than sections that tend to be performed more accurately.

For choir singers, aggregated measures of intonation accuracy at particular sections can fulfil a similar role. These measures are extracted from individual singers' performances aligned with their target scores (D3.2). Given sufficient intra-performer consistency, they provide an approximation of score difficulty.

Measured across a user's rehearsal session, sections that are repeated more often may also provide an indication of difficulty, although care must be taken not to confound "difficult" (many rehearsals of a particular passage to overcome its difficulty) and "popular" (many rehearsals because a passage is pleasing to play or to listen to), or any other motivation to repeat particular sections.

Measured longitudinally over rehearsal sessions spanning weeks or months, a derivative measure reporting the degree and speed of improvement of error rate over time can provide further detail, improvement presumably coming more slowly for very difficult pieces.

3.2. Singing assessment

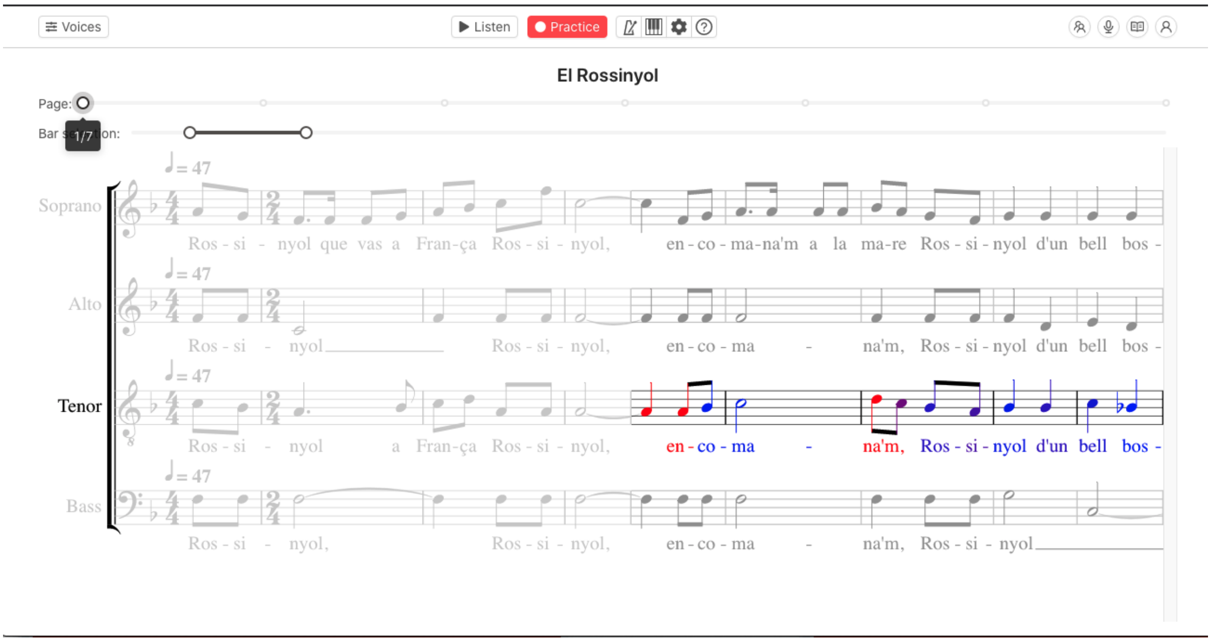
Voices
Listen
Practice
El Rossinyol
Page:
1/7
Bar s on:
♩= 47
Soprano
Ros - si - nyol que vas a Fran-ça Ros - si - nyol, en - co - ma-na'm a la ma-re Ros - si - nyol d'un bell bos -
Alto
Ros - si - nyol Ros - si - nyol, en - co - ma - na'm, Ros - si - nyol d'un bell bos -
Tenor
Ros - si - nyol a Fran-ça Ros - si - nyol, en - co - ma - na'm, Ros - si - nyol d'un bell bos -
Bass
Ros - si - nyol, Ros - si - nyol, en - co - ma - na'm, Ros - si - nyol

Figure 3.2. Singing assessment visualization in the Working prototype for choir singers (D6.6). After recording a performance, the user can see the intonation feedback in the score view of the platform. The color of each note indicates its corresponding intonation rating. In the red-blue color scale, red represents inaccurate intonation, while blue represents accurate intonation.

The assessment algorithm is based on methods presented by Wager et al. (2019) and Cuesta et al. (2018) and it has several input parameters: the pitch curve, the MusicXML score of the song, the start and end bars of the rehearsal, the estimated latency in seconds (see section 4.2.2), and the voice part of the singer.

This method computes an intonation score for each note in the performance - a value between 0 (inaccurate intonation) and 1 (accurate intonation). We explicitly measure this accuracy as the deviation from the score, thus obtaining an objective score. We do not consider any perceptual aspects, nor any temperament other than equal temperament in our implementation. The detailed steps of the algorithm follow:

1. The algorithm selects the excerpt of the score that corresponds to the performance using the start and end bars of the rehearsal. As a result, we obtain a list, `score_excerpt`, of *N* triplets: `[note_onset, note_offset, pitch]`, where *N* is the number of notes in the performance.
2. The `score_excerpt` list is converted into a time series to ease further steps, using the frame rate from the pitch curve. Consequently, we obtain a new list, `pitch`$_{ref}$, of tuples `[timestamp, score_pitch]`, where the pitch from the score is repeated for all frames within the same note.
3. For each frame *i* of the pitch curve, we compute the intonation score, S_i, as the ratio between the performance pitch (from the pitch curve, `pitch`$_{perf}$) and the target pitch (from the score, `pitch`$_{ref}$) using the following equation:

$$S_i = 1200 \cdot log_2 \frac{pitch_{perf_i}}{pitch_{ref_i}}$$

which measures the difference between both values in *cents*. Note that we previously adjusted the pitch curve according to the estimated latency.

4. Using the note boundaries from the score, we compute the median of the frame-wise deviation values within each note. By default, we set a maximum deviation of 100 cents (one semitone), which defines the lowest intonation score.

An example of the intonation assessment feedback is illustrated in Figure 3.2, where the notes in the score view of the platform are color-coded according to the intonation deviations computed between the score and the user's performance.

3.3. Performance features

Performance characteristics informing measures of performance quality, including timing (rubato) variability, intonation, and dynamic range, will be determined by automated music description (D3.2). Here, a performance *typicality* heuristic proposes more typical renditions of a work to be qualitatively "better" than less typical renditions (as suggested by Repp, 1997; Page et al, 2017; Wolf et al, 2018). In considering WP3 feature extraction technologies operating on recordings of performed renditions, it is particularly worth bearing in mind that these will operate not only on high-quality studio-edited recordings, but also on user-provided musical renditions that may exhibit qualities inherent in "live" (or live-recorded) music recordings (Page et al, 2017), including mistakes, noisy signal, or poor recording quality.

4. Infrastructure for performance assessment

We now outline the technical implementation of workflows developed to process performance data required to enact the mechanisms presented in Section 3 and to present the results to the user.

4.1. Instrumental performance assessment

Instrumental performance assessment is enabled through the performance-to-score alignment workflow illustrated in Figure 4.1 (detailed in D3.5). From a user perspective, performance assessments are generated through the following process:

1. The user authenticates with TROMPA's applications using their Web ID and Solid identity provider. The user selects a work using the TROMPA multimodal component, and a reference to an associated MEI music encoding is retrieved from the Contributor Environment; alternatively, a user may bypass the TROMPA CE and specify the URL of a web-hosted MEI file directly.
2. A container of rehearsal recordings for the specified MEI score is selected from the user's Solid Pod - a user-controlled online storage space in which contributions are stored privately by default (see D5.1 Data Infrastructure). Alternatively a new container is created if no matching ones exist.
3. The score is rendered in the CLARA rehearsal companion application's Web interface (see D6.5 Working prototype for instrumental players). The user ensures that a MIDI instrument is plugged into the computer, then starts playing. The rehearsal companion application detects a first MIDI signal, and begins recording note events. Once a pause in incoming MIDI notes

exceeding a specified threshold is detected, the current recording is ended and a new rendition is declared. The collected MIDI notes are sent to the TROMPA Processing Library which initiates the performance-to-score alignment workflow (D3.5 multimodal music information alignment). This workflow produces an RDF structure representing a score-aligned timeline of the rehearsal rendition according to the alignment data model (D3.5).

4. The generated RDF structure is ingested into the selected container in the user's Solid Pod, where it is stored privately (accessible only to the user). The user is able to make choices to control access for each recorded rehearsal through the rehearsal companion application.
5. When the score-aligned timeline arrives in the container of rehearsal recordings, the rehearsal companion application retrieves the new data, providing interactive visualisations of performance errors and other performance aspects (tempo, dynamics) to the user (see D6.5 Working prototype for instrumental players). Each aspect may be visualised in detail for a given performance, or assessed in context with corresponding measures in other performances within the selected container.

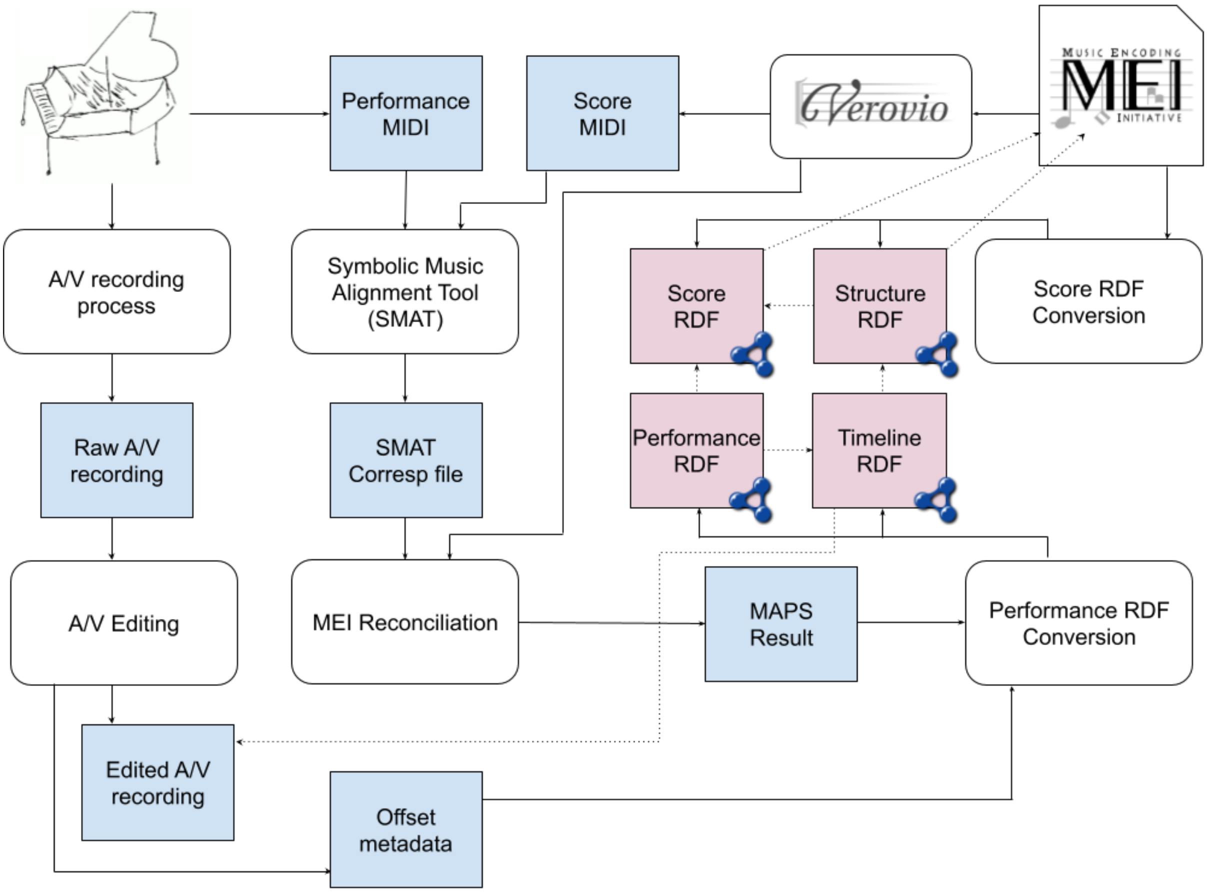

Figure 4.1 Instrumental players: Performance-to-score alignment workflow - ref. D3.5.

4.2. Choir singing performance assessment

4.2.1. Performance assessment workflow

The choir performance assessment involves several components in the Choir Singers Pilot and in the TROMPA Processing Library:

1. The user (singer) selects a piece from their choir's repertoire.
2. The user prepares the performance by adjusting the volume of synthetic voices as desired and choosing a section to practice. The user records their singing while wearing headphones, with the synthesized voices playing in the background. Alignment is not necessary in this use case since the tempo is imposed by the background voices, and the user's recording is automatically adjusted to compensate for the latency of the audio recording process.
3. Once a recording is finished, the performance is saved to the Voiceful Cloud servers. Subsequently, the VoDesc API is used to compute the pitch curve (fundamental frequency) of the user's recording.
4. The performance assessment is then computed using the following steps in a backend API function of the Choir Singers Pilot:
    a. Create two DigitalDocument objects in the TROMPA CE:
        i. the score associated with the performance
        ii. the pitch curve of the user's performance (computed by VoDesc), in JSON format, encrypted using a secret key
    b. The API creates a ControlAction request in the CE to request the performance assessment algorithm to run, with the following parameters: start measure, end measure, recording latency offset, id of the score DigitalDocument, id of the pitch curve DigitalDocument
5. The performance assessment algorithm, part of the TROMPA Processing Library (TPL), receives the newly created ControlAction request and processes it as follows:
    a. Reads parameters and downloads associated score and pitch curve from the CE, and decrypts the encrypted pitch curve using the shared secret key.
    b. Executes the actual singing assessment algorithm (see Section 3.2.).
    c. Stores a link to the result in the ControlAction on the CE.
6. The Choir Singers Pilot API waits for the result to appear in the ControlAction on the CE, and once it does, it is displayed to the user on the My Rehearsals page, where the user can view the assessment results indicated with colors on the score view of the piece.

The following figure illustrates this workflow:

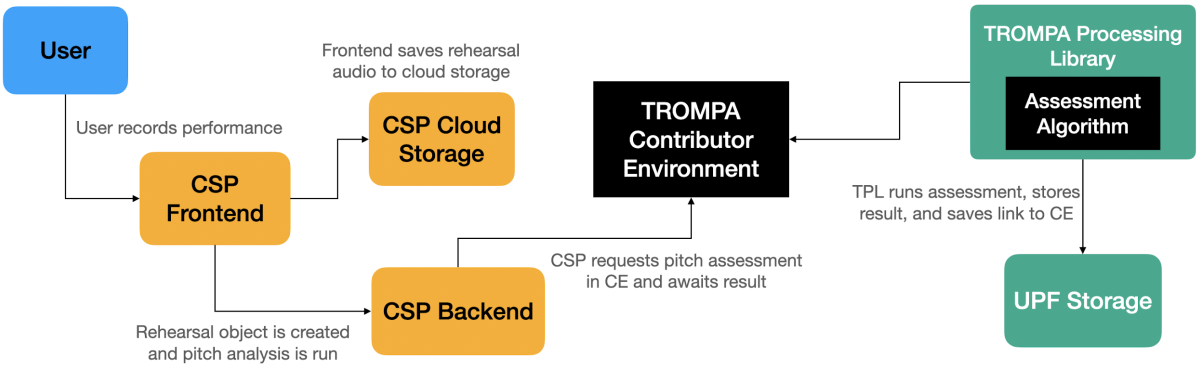
User
User records performance
CSP Frontend
Frontend saves rehearsal audio to cloud storage
CSP Cloud Storage
TROMPA Contributor Environment
TROMPA Processing Library
Assessment Algorithm
TPL runs assessment, stores result, and saves link to CE
UPF Storage
CSP requests pitch assessment in CE and awaits result
CSP Backend
Rehearsal object is created and pitch analysis is run

Figure 4.2. Choir singers: Performance assessment workflow. When a user records a performance, the pitch is extracted on the CSP backend and sent to the TROMPA CE. The TPL runs the performance assessment algorithm and stores the results on the CE. Finally, the results are displayed in the CSP Frontend.

4.2.2. Data processing sites and tasks

The workflow below is explained with more detail in the 1st version of **D5.3- TROMPA Processing Library**, Section 5[11].

1. **In-situ device** at performance (e.g. tablet computer held by the singer)
    - User interaction
    - Records audio as the user sings along (using headphones) with the synthetic voices.
2. **Processing servers**

    These analyses are performed *after* singing, on the whole audio file and also take care of linking analysis results with relevant metadata in the CE.

    a. **Voiceful Cloud**

    Computes the *VoDesc* analysis on the singing voice audio.

    b. **UPF dedicated server**

    Runs the algorithms developed in Task 3.2 for choir singing analysis.
3. **Analysis results repositories**

    a. **Voiceful Cloud**

    Stores the *VoDesc* analysis, accessible (in encrypted form) through public URIs.

    b. **UPF dedicated server**

    Stores the choir singing-specific analysis, accessible through public URIs.
4. **Contributor environment (CE)**
    - Houses metadata associated with individual performances.

5. Directions for future investigation

TROMPA's performance prototypes provide a platform to generate empirical data relating to choir and instrumental music rehearsal practice, aligned with semantic music score encodings. Such data can serve as ground-truth in investigations of score difficulty, to validate externally-determined difficulty ratings. While undertaking these investigations is out of scope for the current project, here

[11] https://trompamusic.eu/deliverables/TR-D5.3-TROMPA_Processing_Library_v1.pdf

we propose possible avenues for future work in this direction, building directly on TROMPA project outcomes.

The literature on score derived difficulty for piano performances (Sébastien, Ralambondrainy, Sébastien, O., & Conruyt, 2012; Song & Lee, 2016) suggest a number of different informative measures, which we have grouped into motoric-physiological and cognitive-structural categories. The first category relates to the difficulty of realising a score as a physical performance: requirements of performance "speed" (tempo and attack density); hand displacement (e.g., leaps on a keyboard or along a fretboard); fingering (e.g., awkwardness of chord transitions, or polyphonic requirements of sounding multiple notes at once, see Parncutt et al, 1997); and special requirements for the mastery of individual performance techniques (e.g., hammer-on / pull-offs on fretted instruments; octave techniques and third-runs on piano).

Several of these measures can be extracted from suitably rich score encoding metadata: tempo indications for tempo, performance directives for certain performance techniques. These metadata relating to performance tempo may be combined with expert recommendations on "recommended" metronomic markings such as provided by Carl Czerny on Beethoven's piano works (Badura-Skoda, 1994) to compute attack density or similar. Other measures require analyses of the score: attack density, and performance techniques not explicitly described as directives in the score (e.g., third-runs).

Yet others require analysis of physiology in combination with score analysis (e.g., leap size, fingering). Depending on the instrument, such metrics necessarily make assumptions on hand placement (which notes are played by which hand) and fingering, typically encoded implicitly in the score and requiring expert tacit knowledge to translate into a complete motoric realisation. While the implementation of accurate physiological models of performance is daunting, some simple heuristics that provide workable approximations. For instance, note-stem directions in a piano score may provide cues as to which hand is playing which note: where notes are densely present in both the upper and lower staff, one may conclude with reasonable confidence that the in the upper staff's notes are to be played by the right hand and the lower staff's notes by the left; where one of the staves is sparsely populated with notes, and the other exhibits both up- and down-facing note stems, one may expect that the upward-stemmed notes are likely to be played by the right hand, and the downward-stemmed notes by the left. This information could be used to estimate the motoric difficulty of polyphony within chords (how many notes must be played by each hand at once).

The second category relates to the cognitive-structural difficulty posed to the musician of establishing and maintaining a mental model of the score during its performance. Relevant measures here include the score's harmonic and rhythmic complexity; and, the number of deviations from key (crudely, the number of accidentals attached to individual notes). Further, information-theoretic (entropy-based) compressibility of scores, could be investigated, on the hypothesis that highly compressible scores exhibiting low entropy pose lower cognitive overheads to memorise and process, and thus may be perceived as "easier" (in this sense) than more complex scores exhibiting high entropy.

Finally, measures that factor into both the motoric-physiological and cognitive-structural categories include the overall length of a score (potentially contributing to both cognitive and physiological fatigue), and (depending on the instrument), its key signature: according to Chopin, C major is easy to read, but difficult to play on piano, because all notes in the scale map to white keys (e.g. as in his Étude Op. 10 No. 1; see Eigeldinger & Shohet, 1986, p. 34). On the contrary, B major is more difficult to read, but easy to play, because it requires long fingers to be placed on the black keys while the shorter thumb and pinky finger play mostly on white keys.

6. Conclusion

In this deliverable we have reflected on score difficulty and performance quality, both complex notions liable to evoke differing judgements even among expert human judges, and have outlined how the approaches taken in TROMPA's performance-based use-case prototypes for instrumental players (D6.5) and choir singers (D6.6) produce data to support implicit, empirical assessments. In section 2, we have described how TROMPA's use of digital music encodings and score-aligned performances produce measures of performance quality by quantifying performance error, and how such measures can in aggregate (across many performance recordings) provide empirical indications of score difficulty. In section 3, we operationalise such measures in terms of note insertions and deletions (in the case of instrumental players), and intonation inaccuracy (in the case of choir singers), and describe how further interpretational aspects such as performance tempo and dynamics can provide hints for the assessment of performance quality - beyond the absence of error - based on indications in the literature that *typical* performances tend to be deemed subjectively "better" by listeners. In section 4, we have provided a description of the technical infrastructure used to enact these mechanisms in our instrumental performer and choir singer use cases, summarising interactions with the TROMPA Contributor Environment (D5.1) and Processing Library (D5.3). Finally, in section 5 we point to directions for future investigation of motoric-physiological and cognitive-structural hypotheses of score difficulty, opened up through the score-aligned, note-level rehearsal data generated by the users of TROMPA's performance prototypes.

7. References

7.1. Written references

Badura-Skoda, Paul (Ed.) (1994). Carl Czerny: Über den richtigen Vortrag der sämtlichen Beethoven'schen Klavierwerke: Nebst Czerny's Erinnerungen an Beethoven (Carl Czerny: On the Proper Performance of all Beethoven´s Works for the Piano. In addition to Czerny's Memories of Beethoven), Universal Edition, Vienna.

Cuesta, H., Gómez, E., Martorell, A., & Loáiciga, F. (2018). Analysis of Intonation in Unison Choir Singing. In *Proceedings of the 15th International Conference of Music Perception and Cognition (ICMPC) and 10th Triennial Conference of the European Society for the Cognitive Sciences of Music (ESCOM)*. (pp. 125–130). Graz, Austria.

Eigeldinger, Jean-Jacques and Shohet, Naomi (1986). *Chopin: Pianist and Teacher: As Seen by His Pupils.* Cambridge University Press, Cambridge, U.K.

Flossmann, S., Goebl, W., & Widmer, G. (2009). Maintaining skill across the life span: Magaloff's entire Chopin at age 77. Paper presented at the Proceedings of the International Symposium on Performance Science 2009 (15–18 December 2009), Auckland, New Zealand, Utrecht, The Netherlands.

Kolisch, Rudolf (1993). Tempo and character in Beethoven's music. The Musical Quarterly, 77(1), 90–131. Retrieved from http://www.jstor.org/stable/742431

Page, K. R., Bechhofer, S., Fazekas, G., Weigl, D. M., & Wilmering, T. (2017). Realising a layered digital library: Exploration and analysis of the Live Music Archive through linked data. In *Proceedings of the ACM/IEEE 2017 Joint Conference on Digital Libraries*. doi:10.1109/JCDL.2017.7991563

Parncutt, R, Sloboda, J. A., Clarke, E. F., Raekallio, M., & Desain, P. (1997). An ergonomic model of keyboard fingering for melodic fragments. Music Perception, 14(4), 341–382. doi:10.2307/40285730

Pfordresher, Peter Q., Brown, S., Meier, K. M., Belyk, M., & Liotti, M. (2010). Imprecise singing is widespread. *The Journal of the Acoustical Society of America*, *128*(4), 2182–2190. doi:10.1121/1.3478782

Repp, Bruno H. (1997). The Aesthetic Quality of a Quantitatively Average Music Performance: Two Preliminary Experiments. Music Perception, 14(4), 419–444. doi:10.2307/40285732

Sébastien, V., Ralambondrainy, H., Sébastien, O., & Conruyt, N. (2012, October). Score analyzer: Automatically determining scores difficulty level for instrumental e-learning. In *13th International Society for Music Information Retrieval Conference (ISMIR 2012)* (pp. 571–576).

Song, Y. E., & Lee, Y. K. (2016). A Method for Measuring the Difficulty of Music Scores. *Journal of the Korea Society of Computer and Information*, *21*(4), 39–46.

Wager, S., Tzanetakis, G., Sullivan, S., Wang, C., Shimmin, J., Kim, M., & Cook, P. (2019). Intonation: A Dataset of Quality Vocal Performances Refined by Spectral Clustering on Pitch Congruence. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (pp. 476–480).

Wolf, A., Kopiez, R., Platz, F., Lin, H.-R., & Mütze, H. (2018). Tendency towards the average? The aesthetic evaluation of a quantitatively average music performance: A successful replication of Repp's (1997) study. Music Perception, 36(1), 98–108. doi:10.1525/MP.2018.36.1.98

Wolters, Klaus. (1994). Handbuch der Klavierliteratur. Klaviermusik zu zwei Händen (Handbook of the Two-handed Piano Repertoire), 4th edition, Atlantis Musikbuch-Verlag, Mainz, Germany.

7.2. List of abbreviations

Abbreviation	Description
ABRSM	Associated Board of the Royal Schools of Music (UK)
CE	TROMPA Contributor Environment
CEUS	Bösendorfer computerised reproducing piano system (acronym expansion unclear)
LDP	Linked Data Platform
MAPS	Matcher for Alignment of Performance and Score
mdw	University of Music and Performing Arts Vienna
MEI	Music Encoding Initiative
MIDI	Musical Instrument Digital Interface

UPF	University Pompeu Fabra
VoDesc	Voice Description analysis tool by Voctro Labs