TROMPA

TROMPA: Towards Richer Online Music Public-domain Archives

Deliverable 5.4

Music Performance Assessment Mechanisms

Grant Agreement nr	770376
Project runtime	May 2018 – April 2021
Document Reference	TR-D5.4-Music Performance Assessment Mechanisms v1
Work Package	WP5 - TROMPA Contributor Environment
Deliverable Type	Report
Dissemination Level	PU - Public
Document due date	30 April 2019
Date of submission	30 April 2019
Leader	MDW
Contact Person	Werner Goebl (goebl@mdw.ac.at)
Authors	David M. Weigl (MDW), Werner Goebl (MDW), Álvaro Sarasúa (VL), Helena Cuesta (UPF)
Reviewers	Aggelos Gkiokas (UPF), Emilia Gomez (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). Further, TROMPA aims to develop automated models of these kinds of assessments, developed and systematically validated via human feedback.

Automatic assessment of score difficulty is far from trivial. The 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.

Section 2 of this deliverable summarises the lower-level descriptors that in aggregate may be used to approximate notions of quality and difficulty as described above. These descriptors may be i) obtained explicitly from human judgements obtained through crowd-sourcing, user interactions, or from sources in the literature; ii) implicitly, from musicians' behaviours during musical performances; or iii) they are derived algorithmically from performance audio recordings, MIDI streams, and other information modalities, which we briefly summarise in Section 3.

In Section 4, we describe mechanisms envisioned to determine these lower-level descriptors using technologies from TROMPA deliverables D3.2 and D3.5. These comprise:

- User judgements obtained via tools and methods developed in D5.2 (Digital Score Edition) and D5.5 (Annotation tools). These provide individual ratings of difficulty aspects of a score (for individual sections or entire pieces) or on specific aspects of a performance, such as overall performance quality (high, low), expressivity of a performance (expressive, mechanical), tempo and dynamics judgment (too slow, too fast, too loud, too soft) .
- Expert assessments of difficulty harvested from the pedagogical literature, providing reference data ("difficulty indices") for the training of WP3 technologies, as well as metadata available for consumption by end users.
- Measures of performance "errors" (that is deviations from the notated score) identified during performance to score alignment (D3.5) and through characterisation of intonation accuracy and timing deviation (D3.2), 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.

- Quantifications of individual performer (instrumental or singers) output over time, investigating the number of rehearsal repetitions, and the rate of performance *improvement* across rehearsal sessions, as signatures of piece difficulty.
- Quantifications of score difficulty according to *motor-physiological* requirements of its performance (tempo; attack density; hand displacement; fingering; mastery of specialised performance techniques).
- Quantifications of score difficulty according to cognitive-structural requirements (harmonic and rhythmic complexity; deviations from key; information-theoretic *compressibility* of the score).
- Consequently, quantifications of both motoric-physiological and cognitive-structural *fatigue* liable to be produced by the performance of a piece.
- Measures of performance quality determined in aggregate from performances of a particular musical piece (on the notion, grounded in the literature, that *typical* performances along particular musical parameters tend to produce qualitatively *better* performances).

Finally, in section 5 we summarise performance assessment workflows in terms of the TROMPA data infrastructure (D5.1) , first describing interactions with the Contributor Environment via the TROMPA Processing Library (D5.3) before detailing the workflow specific to the instrumental performers and the choral singers use cases.

Version Log		
#	Date	Description
v0.1	24 April 2019	Initial version submitted for internal review
v0.2	30 April 2019	Revised version after internal review
v1.0	30 April 2019	Final version submitted to EU

Table of Contents

1. Introduction

The TROMPA project proposal describes the aims of Task 5.4 (music performance assessment mechanisms) and the associated deliverable, a first 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.

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.

2. Which qualities are being assessed?

2.1. Score-related measures

In pursuing the high-level dual notions of score difficulty and performance quality, we rely on a number of lower-level descriptors and information sources. These are obtained explicitly from human judgements obtained through crowd-sourcing, user interactions, or from sources in the literature; implicitly, from musicians' behaviours during musical performances; or they are derived

algorithmically from performance audio recordings, MIDI streams, and other information modalities summarised in Section 3. The lower-level descriptors used to inform difficulty indices and performance quality measures are summarised in this section, and the mechanisms envisioned to determine them (using technologies from TROMPA deliverables D3.2 and D3.5) are discussed in Section 4.

- **Piece difficulty as per human annotation** (crowd, instrumentalists, teachers, experts)
  - Overall rating combined with explanatory annotations
  - Based on pedagogical literature (such as Klaus Wolters, UK grading system, based on crowd opinion/experience, Henle difficulty ratings[1])
- **Piece difficulty based on empirical performance measures** (automatically extracted)
  - Sections with high error rates (omitted or inserted notes, intonation inaccuracies) are more difficult
  - Sections with many rehearsal repetitions are more difficult
  - Pieces that improve over time (and within a group of similar skill level) faster than others are easier
- **Piece difficulty based on quantitative measures**, automatically determined from the score
  - Motoric-physiological aspects (instrumental players)
    - Tempo signature and markings in score (or in related literature such as Carl Czerny on Beethoven piano works or other scholar pianists)
    - Attack density (event rate) per score unit and per time unit (based on above)
    - Hand displacement (number and size of leaps, number of notes within a hand)
    - Fingering where present (may be difficult to determine)
    - Special performance techniques required (e.g. on lute hammer-ons or pull-offs based on performance directives in score; on piano octave technique, third runs based on interpretation of score)
  - Cognitive-structural aspects
    - Compressibility of a score (pitch sequences, interval sequences, ...) as a measure of the redundancy of information in a score
    - Rhythmic complexity, pulse clarity
    - Number of deviations from key signature (sharps, flats, naturals)
    - Modulation activity (harmonic complexity)
  - Combined aspects
    - Overall length, number of measures, notes (cognitive and physiological fatigue)
    - Key signature. Number of accidentals increase cognitive complexity, but may decrease physiological difficulty

2.2. Performance-related measures

- Overall quality of a performance ("goodness") as per human judgement (crowd, experts, see Wolf et al, 2018)
- Error count (how many notes omitted or inserted, see Flossmann et al, 2009)
- Intonation accuracy (in the sense of pitch tuning, e.g., Pfordresher et al, 2010)

[1] https://www.henle.de/en/about-us/levels-of-difficulty-piano/ (accessed 24 April 2019)

- Overall performance tempo (of movements, sections, subsections etc.), put in relation to aggregated performance data or tempo indications from the score (see e.g., Kolisch, 1993)
- Timing, rubato variability; deviation from prototypical timing as measure of quality (on the assumption that performance *typicality* correlates with aesthetic quality; see e.g. Page et al, 2017; Repp 1997; Wöllner 2012; Wolf et al, 2018)
- Dynamic range (relative and absolute) of audio, symbolic data

3. What kinds of data could be subject to assessment?

In this section we provide a summary of the information streams available for the characterisation of difficulty indices and performance quality measures.

1. Music encodings (e.g. MEI)
2. Human judgements (e.g. TROMPA annotations; user ratings; literature)
3. Audio signal (characterised by D3.2 technologies)
4. Symbolic performance information streams (e.g. flow of MIDI note events)
5. Performance/instrument metadata streams (e.g. CEUS piano key trajectories)
6. Performer metadata streams (e.g. gestural tracking, physiological measures, eye tracking)
7. Video signal (to assess quality of bodily gestures, staging, light, context, other aspects of performer and performance)
8. Audience metadata streams (e.g. physiological measures, facial mood recognition)

Present development work is focused around the first five of these streams. Remaining signals involving performer tracking and biophysiological measurement, video analysis, and audience metadata remain potential sources of information in future development, but require further ethical clearance to be sought if pursued.

4. Assessment mechanisms

Here, we expand on the mechanisms required to determine the measures identified in Section 2. Difficulty and quality assessments will be provided by *explicit* human judgements (cf. D5.5), *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 (Section 2) applying TROMPA Work Package 3 technologies (particularly D3.2 and D3.5).

4.1. Explicit human judgements

Users provide explicit judgements using tools and methods developed in D5.5 (Annotation tools). They provide individual ratings of difficulty aspects of a score (for individual sections or entire pieces) or on specific aspects of a performance, such as overall performance quality (high, low), expressivity of a performance (expressive, mechanical), tempo and dynamics judgment (too slow, too fast, too

loud, too soft). Difficulty assessments will be harvested from the pedagogical literature, to serve as expert-sourced reference data ("difficulty indices") for the training of WP3 technologies, as well as metadata available for consumption by end users. Examples include Klaus Wolters' "Handbuch der Klavierliteratur zu zwei Händen" (1994), which offers classifications and thoughtful descriptions of the difficulty of a comprehensive core of the classical solo piano repertoire; and, the grading classifications offered by the Associated Board of the Royal Schools of Music (ABRSM)[2] in the UK, which offer reference information on the difficulty of musical works spanning voice and a wide variety of instruments.

4.2. Implicit measures derived from performance behaviour

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). 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 choral singers, aggregated measures of intonation accuracies, timing deviations, and singer synchronization at particular sections can fulfil a similar role. These measures are extracted from individual singers' performances aligned with their target scores (D3.2) and if there is sufficient intra-performer consistency, they might provide an indication 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.

4.3. Measures derived from scores

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 (section 2).

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

[2] http://abrsm.org (Retrieved 24 April 2019)

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 score analysis (D3.2): 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, and considered out of scope for this project, we are investigating 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, we can 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, we can 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 can 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, we will investigate information-theoretic (entropy-based) compressibility of scores, hypothesising that highly compressible scores exhibiting low entropy pose lower cognitive overheads to memorise and process, and thus are 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.

4.4. Measures derived from performances

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 will propose 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.

Quality measures informed both by these performance-derived features, and by implicit measures derived from performance behaviour (the error rates described in the earlier subsection) will be used to estimate (“bootstrap”) initial quality ratings that may then be associated with a particular performance within user-facing pilot applications. It is clear that these measures provide mere computational estimates, which may be validated and improved through explicit human judgements (user feedback).

5. Infrastructure for performance assessment

5.1. Instrumental performance assessment

5.1.1. Data processing sites and tasks

1. **In-situ device** at performance (e.g. tablet computer sitting on a piano)
    a. User interaction
    b. Listens to MIDI notes (“start” of rendition) and a threshold-gap in MIDI notes (“end” of rendition”); sends MIDI stream to MAPS server
2. **MAPS feature server**
    a. Performs performance–score alignment, storing performance in performance repository and alignment metadata in CE
    b. Performs feature extraction and assessment activities, accessing CE for feature aggregation (with metadata of other renditions), and storing results in CE
3. **Performance repository**
    a. Stores performance streams (e.g. MIDI streams, audio recordings, ...) pending user permission
    b. Associates these with metadata through interaction with CE
4. **Contributor environment (CE)**
    a. Houses metadata associated with individual performances (including user annotations)
    b. Maintains and dynamically updates session containers (LDP containers)

5.1.2. Performance assessment workflow

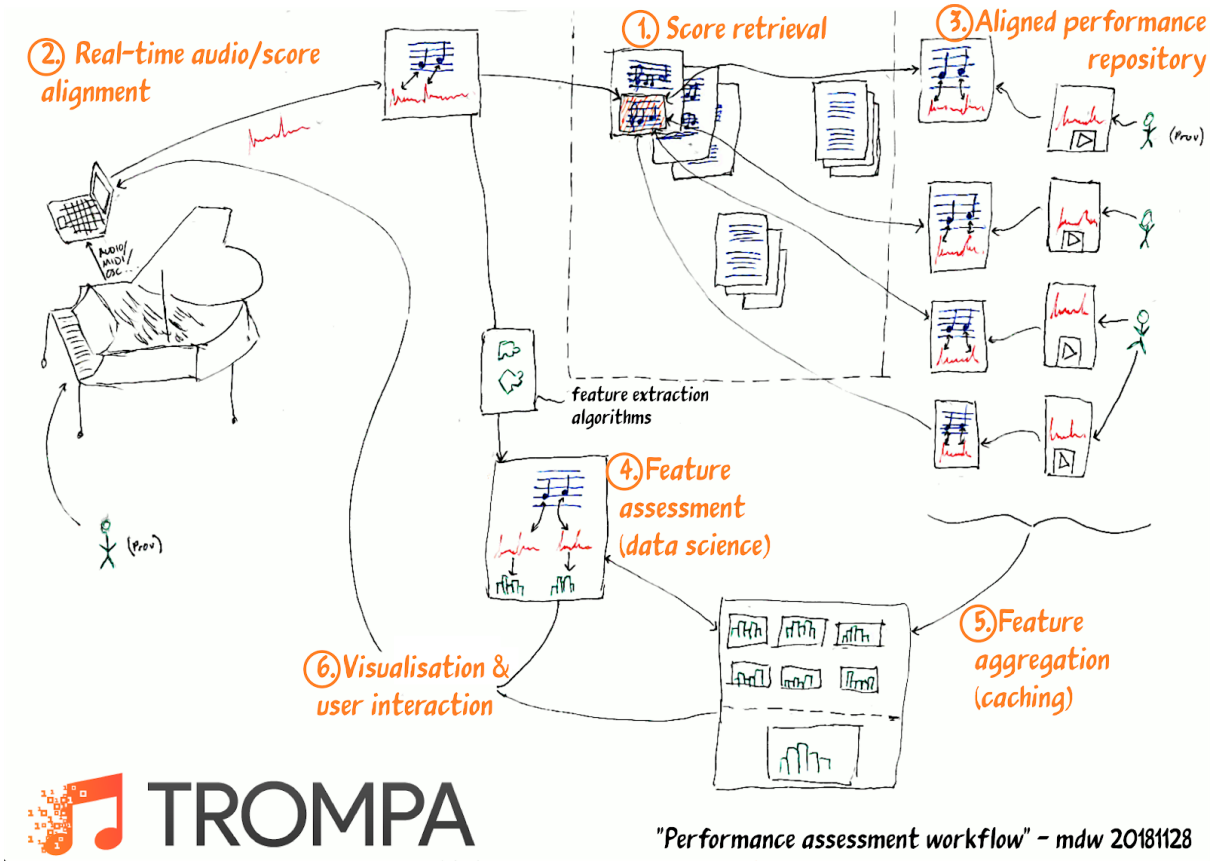

Figure 5.1. Performance assessment workflow

The instrumental performance assessment workflow is illustrated in Figure 5.1, and is described as follows:

1. A work is selected by the user (performer). Metadata, including a reference to an MEI encoding and an LDP container acting as a "playlist" of renditions for comparison, is retrieved from the Contributor Environment. These containers are dynamic (e.g. "my last 5 renditions"; "all my renditions in this session"; "my renditions from the last two weeks"; "my favourite renditions"), consisting of references (links) to individual rendition timelines also maintained by the CE.
2. The user starts playing. The rehearsal companion application detects a first MIDI signal, and begins recording note events. Once a threshold-gap in MIDI notes is detected, the current rendition ends. The MIDI notes are streamed to a MAPS server[3] which generates an alignment data object and converts this to RDF (score segmentation and performance timeline) according to the alignment data model (D3.5).
3. The performance stream is stored within a performance repository, and the alignment metadata associated with the performance is ingested into the Contributor Environment.
4. The performance stream is also fed into a feature assessment pipeline (see sec. 4 for options in terms of technological choices). Generated features and performance alignment metadata

[3] Matcher for Alignment of Performance and Score – other alignment tools remain under consideration

are stored in a repository and ingested by reference into the Contributor Environment (as per D5.3 Section 2.2).

5. Features of related performances (e.g. other performances in session container) are retrieved via the Contributor Environment, in order to determine aggregate measures (e.g., degree of typicality in timing profile)
6. User application retrieves new data and displays them to user for further interaction.

5.2. Choir singing performance assessment

5.2.1. Data processing sites and tasks

The workflow below is explained with more detail in Deliverable 5.3, Section 5[4].

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 analysis 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**
       Computes low level descriptors on the audio with Essentia and runs the algorithms developed in Task 3.2 for choir singing analysis.
3. **Analysis results repositories**
   Audio files for the recorded performances could be stored in any of the two servers listed below. This is an open technical question that will be discussed in the short term.
    a. **Voiceful Cloud**
       Stores the *VoDesc* analysis, accessible through public URIs.
    b. **UPF dedicated server**
       Stores the results of low level descriptors computed by Essentia and the choir singing-specific analysis, both accessible through public URIs.
4. **Contributor environment (CE)**
    - Houses metadata associated with individual performances (including user annotations).

5.2.2. Performance assessment workflow

1. A piece is selected by the user (singer) from a list of available ones for the user retrieved from the Contributor Environment. This list of available pieces for the user is handled directly through the CE.
2. The adjusts the parameters for the synthetic voices, selects a tempo (if desired to sing in a different one from the default one), a section to practice and starts singing along while

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

wearing headphones. Alignment is not necessary in this use case since the tempo is imposed..

3. Once the trial is finished, the performance stream is uploaded to the Voiceful Cloud servers and ingested into the CE.
4. The analysis is requested for this new node. This analysis includes 3 steps:
    a. Compute *VoDesc* analysis using the Voiceful Cloud service.
    b. Compute other low level descriptors using Essentia.
    c. Compute choir-specific analysis from the previous results and the musical score.
5. The application gets notified when the results are ready and the web app shows the results to the user (singer).

6. Conclusion

In this deliverable we have reflected on score difficulty and performance quality, both complex notions liable to evoke differing assessments even among expert human judges. We have provided an outline of various facets whose interplay gives rise to these notions in Section 2; summarised the information streams available for assessment in Section 3; before providing a more detailed discussion of the mechanisms available to TROMPA to both source human judgements and determine automatic measures of score difficulty and performance quality 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).

At the current stage of the TROMPA project, we are thus equipped to begin providing indications of performance quality and score difficulty for multimedia resources ingested into the Contributor Environment, which we can bootstrap and augment with human assessments obtained from the literature. The degree of conformity among our generated metrics, between automated and human-sourced judgements, and indeed the extent of inter-rater agreement on these notions among individual TROMPA users providing feedback remain open questions at this stage. As we gather experience and empirical observations over the coming months, we expect that subsequent iterations of this deliverable will be able to delve into and clarify these issues in greater detail.

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.

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.

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