Int. J. Human–Computer Studies 173 (2023) 103007

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

"Give me happy pop songs in C major and with a fast tempo": A vocal assistant for content-based queries to online music repositories

Luca Turchet [a,*], Carlo Zanotto [a], Johan Pauwels [b]

[a] Department of Information Engineering and Computer Science, University of Trento, Italy
[b] Center for Digital Music, Queen Mary University of London, United Kingdom

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Music technology
Internet of Musical Things
Voice assistant
Online music repository
Networked interaction

ABSTRACT

This paper presents an Internet of Musical Things system devised to support recreational music-making, improvisation, composition, and music learning via vocal queries to an online music repository. The system involves a commercial voice-based interface and the Jamendo cloud-based repository of Creative Commons music content. Thanks to the system the user can query the Jamendo music repository by six content-based features and each combination thereof: mood, genre, tempo, chords, key and tuning. Such queries differ from the conventional methods for music retrieval, which are based on the piece's title and the artist's name. These features were identified following a survey with 112 musicians, which preliminary validated the concept underlying the proposed system. A user study with 20 musicians showed that the system was deemed usable, able to provide a satisfactory user experience, and useful in a variety of musical activities. Differences in the participants' needs were identified, which highlighted the need for personalization mechanisms based on the expertise level of the user. Importantly, the system was seen as a concrete solution to physical encumbrances that arise from the concurrent use of the instrument and devices providing interactive media resources. Finally, the system offers benefits to visually-impaired musicians.

1. Introduction

In recent years a novel paradigm in music technology has emerged as an extension of the Internet of Things to the musical domain, the so-called Internet of Musical Things (IoMusT) (Turchet et al., 2018). The IoMusT refers to a network of Musical Things, which are devices dedicated to production, interaction with, and reception of musical content. Musical Things embed electronics, sensors, data processing software and network connectivity enabling the collection and exchange of data for musical purposes. Examples of Musical Things include smartphones or devices running voice-based assistants used in musical contexts.

Musical Things may also consist of servers hosting online music databases. Nowadays, several online music databases exist as well as services for them (e.g., streaming). While some of these have a commercial nature (e.g., Spotify, Apple Music), others are freely available (e.g., Freesound[1]). Regarding the latter, one of the most popular and largest online repositories is Jamendo,[2] which provides a collection of thousands of crowd-sourced musical compositions licensed under Creative Commons (Bazen et al., 2015). Jamendo is part of the Audio Commons Initiative (Font et al., 2016), a recent endeavor aiming to bridge the gap between audio content producers, providers and consumers through a web-based ecosystem. The approach combines techniques from Music Information Retrieval (to extract creative metadata by automatically annotating audio content Kostek, 2018; Su et al., 2022) and the Semantic Web (to structure knowledge and enable intelligent searches Turchet and Antoniazzi, 2023). Content aggregators part of the Audio Commons ecosystem, such as Jamendo, provide access to audio data through application programming interfaces (APIs).

Jamendo offers a large amount of musical content that is largely unknown. This has spurred the exploration of novel methods of discovery, which are not based on the currently adopted search methods that are based on textual input (e.g., typing on a browser or a mobile app the title of the song or the name of an artist) (Xambó et al., 2018a; Turchet et al., 2020). Notwithstanding the possibility of exploring and playing the songs from such databases through text-based music search engines and playlists, to the authors' best knowledge there is still a gap in interaction mechanisms making these large collections of music adapted to listeners' and instrument players' needs (e.g., content-based queries for music enjoyment, improvisation, learning, or composition).

* Corresponding author.
E-mail addresses: luca.turchet@unitn.it (L. Turchet), carlo.zanotto@unitn.it (C. Zanotto), j.pauwels@qmul.ac.uk (J. Pauwels).

[1] https://freesound.org/.
[2] https://www.jamendo.com/.

https://doi.org/10.1016/j.ijhcs.2023.103007
Received 19 October 2022; Received in revised form 20 January 2023; Accepted 30 January 2023
Available online 7 February 2023

Devices equipped with voice-based interfaces offer opportunities to bridge this gap. Thanks to their direct connectivity to the Internet and embedded sound delivery system, such devices can send a request for musical content to remote repositories, receive a response in the form of a music piece, and reproduce it. This would allow listeners to simply use their voice to query the repository instead of using a textual search. Instrument players would be enabled to play along with downloaded musical content and simply use their voice to make the search. This would be a more comfortable interaction for musicians, who could interact with the repository while holding their instrument in their hands and, therefore, encounter difficulties when typing at the same time textual input for the search, as highlighted in Martinez-Avila et al. (2019). Furthermore, voice-based interfaces have clear benefits for addressing accessibility issues of the visually-impaired population.

However, to date, the challenge of connecting voice-based interfaces to online music repositories has been only marginally addressed in academic and industrial applications. For instance, in summer 2020 Spotify has been integrated with Amazon Alexa, although the possibilities for interaction are still very limited. Enhancing the integration and the affordances of such kinds of interconnection has the potential to enable novel kinds of Ubiquitous Music activities. Ubiquitous Music (Keller et al., 2014) is a branch of the Sound and Music Computing field which develops and analyzes musical activities supported by ubiquitous computing concepts and technology.

In this paper we explore the use of a voice-based system able to interface with the Jamendo Audio Commons ecosystem for the retrieval of online music content and its repurposing for various musical practices (e.g., listening, recreational music-making, composition). We present a prototype involving a commonly available vocal interface, which was used to query content from Jamendo adopting content-based criteria. These criteria are therefore different from the ones usually involved in queries to online music repositories, such as artist's name or song title. Specifically, we focus on retrieval by genre, mood, chords, tempo, key, and tuning. Thus far, scarce research has been conducted on the task of querying a music repository using such criteria, especially when utilized simultaneously. The repository was indexed along such criteria leading to a database offering a wider range of search possibilities compared to previous efforts. Such a system may foster a variety of Ubiquitous Music applications, such as those involving music listening or music playing, for a wide variety of purposes: music consumption, education, composition, performance, improvisation, self-learning, recreational music-making, and exploration of unknown pieces and artists of the database.

2. Related work

2.1. The Audio Commons initiative and its artistic use

The Audio Commons Initiative (Font et al., 2016) provides an ecosystem through which musicians (performers, sound designers, composers) can access audio content with various tools. These tools include interfaces based on web browsers (e.g., Freesound Fonseca et al., 2017, Jamendo Bazen et al., 2015), audio plugins, or live coding tools (Xambó et al., 2018b). Such a web-based approach provides access to distributed audio content in a user-friendly way and aims to bridge the gap between audio content producers, providers and consumers. This is accomplished in a way different from methods based on traditional digital audio workstations and digital musical interfaces, which were conceived to operate with local audio content (e.g., personal recordings gathered by the musicians themselves). In Audio Commons tools, the available metadata information about the sounds depends on what has been provided by authors during uploads including tags, descriptions or file names. Notably, such tools enable designers to create third-party applications exploiting its audio content in live applications by granting access to the database through dedicated APIs.

A number of systems have recently exploited the creative opportunities offered by the Audio Commons ecosystem thanks to its audio content search engine informed by semantic metadata and audio content-based features. This search engine enables quick access to thousands of sounds from various online content providers based on requirements matching the user's needs. A web-based tool designed for beginners or advanced musicians willing to explore music composition based on semantic ideation and spectrogram sound inspection was proposed in Stolfi et al. (2018) utilizing content queried from Freesound. Similarly, Xambó et al. (2018a) reports a web-based prototype allowing expert and novice musicians to discover songs in Jamendo by specifying a set of chords. The study reported in Skach et al. (2018) proposed a sonic wearable interface letting users trigger and transform sounds downloaded from Freesound through body-based gestural interactions tracked by e-textile sensors. Along the same lines, the system reported in Turchet et al. (2020) used sounds retrieved from Jamendo onto a smart guitar for music learning, improvisation, composition purposes.

2.2. Voice-based interfaces

Today, speech recognition and synthesis systems widely embedded in mobile and in-home digital assistants (e.g., Amazon's Alexa, Microsoft's Cortana, Apple's Siri) are fostering novel interactive applications to support communication, collaboration, and information seeking. Their increasing availability also provides novel opportunities for broad, accessible interaction by voice. Not requiring the visual and motor skills needed for text input through a keyboard lowers the barriers of entry for usage by the elderly and people with disabilities. Voice-based interfaces are particularly relevant to the visually-impaired, providing a valuable alternative to textual input.

Several studies have investigated the use of voice-based interactions for the control of interactive applications (Igarashi and Hughes, 2001), in particular for accessibility purposes (Brewer et al., 2018). Noticeable examples within this domain include interfaces that (i) substitute input devices (such as the mouse Harada et al., 2006), (ii) address various kinds of applications (e.g., web navigation Christian et al., 2000 or computer games Harada et al., 2011), (iii) target different classes of users (including the visually-impaired Bigham et al., 2009, elderly population Kopp et al., 2018, people with cognitive disabilities Granata et al., 2010) or motor impairments (Pradhan et al., 2018). Overall, these studies show the effectiveness of voice-based interfaces for replacing other kinds of interaction modalities.

In the context of the artistic exploitation of the tools of the Audio Commons Initiative, the study reported in Turchet and Zanetti (2020) presents a voice-based interface for soundscape composition, which enables composers (including those visually-impaired) to create soundscapes by vocally querying online sounds repositories such as Freesound.

3. Requirements and design

We conducted a user study consisting of an online questionnaire in order to understand the needs and expectations of users and derive requirements for the implementation of the vocal assistant. The primary target users of the system were identified as those willing to retrieve content from an online music repository and with a musical knowledge sufficient to enable content-based queries to the repositories. These users included musicians (beginners, intermediates, and experts), sound designers, sound producers, and music technologists. Such users were recruited both from the personal network of the authors and via social media.

A total of 112 participants were recruited (23 females, 80 males, 3 non-binary, 3 preferred not to say), aged between 22 and 58 (mean = 33.5, standard deviation = 10.9). They had 39 different nationalities from all continents and played a wide variety of instruments; 10 of them considered themselves as beginners, 43 as intermediate, and 56

as experts (respectively, with an average years of musical experience of 7, 15, and 23).

The questionnaire was composed of four parts: an information sheet and consent form, three questions, and demographics collection. The questions and the structure of the questionnaire were chosen after a pilot study conducted with three musicians, who were not included in the final user study. The three questions of the questionnaire were:

(1) Imagine you have a vocal assistant (such as Alexa, Siri, or Cortana) that allows you to download music pieces from an online music repository (such as Spotify or YouTube). Which criteria would you use to vocally retrieve the music you want? Select the criteria listed below and assess their importance.

- Absence of instruments (e.g., pieces with no guitar)
- Beats per minute — BPM (e.g., 120 BPM)
- Chords (i.e., one or more chords contained in the piece, e.g., A minor and E7)
- Difficulty level to play with (e.g., easy, very difficult)
- Genre/style (e.g., rock, classical, folk)
- Key (eg., C major, B flat minor)
- Melody (expressed vocally, by singing the melodic line)
- Melodic richness (e.g., pieces that contain or lack a dominant melodic component or pieces that feature several kinds of melodies or polyphony)
- Mood (e.g., happy, sad)
- Music content type (e.g., pieces with full ensemble versus backing tracks, original songs versus cover, or live recordings versus the album version)
- Name of the artist/band
- Playing technique (e.g., pieces that contain arpeggios)
- Popularity level (e.g., very popular songs)
- Presence of instruments (e.g., pieces having the harp)
- Ranking (e.g., the ranking of a song expressed by a community of listeners)
- Rhythm (e.g., a given rhythmic pattern expressed with the voice)
- Scales (e.g., pieces that contain melodies using the G mixolydian scale)
- Tempo signature (e.g., 4/4, 7/8)
- Timbre (e.g., pieces that contain some audio effects applied to the electric guitar)
- Title of the song
- Tuning (e.g., A = 440 Hz or A = 432 Hz)
- other (please explain)

(2) For which musical or non-musical activities would you use a voice-based system with these characteristics? Please list some use cases.

(3) Do you have any comments?

Regarding question 1, participants were asked to rate the 21 items (plus the optional "other") on a 9-point Likert scale (1 = not important at all, 9 = very important). The results were expected to provide insights on which criteria should be selected and prioritized when developing the vocal assistant. On the one hand, the provided items were selected on the basis of the study reported in Turchet et al. (2020). Such study included a set of semi-structured interviews with guitar players, which allowed to identify a set of criteria useful to this category of users to retrieve content from music repositories. On the other hand, the list of items included also the conventional search criteria when exploring a music repository, such as title of the piece and the name of the artist/band. This was due to the fact that we were interested in comparing the extent to which participants were interested in non-conventional criteria with respect to the conventional ones.

The purpose of question 2 was to understand how participants would use the system if it existed, for which musical or music-related activities, and in which contexts. We were also interested in assessing whether any correlation existed between the envisioned use cases and the demographics of participants.

3.1. Results

Results of question 1. Fig. 1 shows the mean and standard error of the 21 questionnaire items. The conventional criteria title and name received the highest scores, followed by genre, melody and mood. The lowest scores were received by the criteria popularity, ranking, and tuning. Responses were not normally distributed (as assessed with a Shapiro–Wilk test), therefore all analyses are henceforth reported using non-parametric statistical tests.

First, a relevant question is whether there are statistically significant differences between such conventional and non-conventional criteria. The absence of differences would indicate that they are essentially equally important, and therefore, that there is an actual need for some content-based criteria in addition to the ones usually offered by today's interfaces. For this purpose, we run a Kruskal–Wallis test to assess differences between responses for each criteria. A significant main effect was found ($\chi^2(20) = 586.71$, $p < 0.001$). Pairwise comparisons using the Wilcoxon Signed-Rank pairwise tests (adjusted using the Benjamini & Hochberg correction, which was selected because it is one of the most conservative correction methods) showed that only responses to genre were not statistically different from those to title and name.

Secondly, the questionnaire data were analyzed using the Kruskal–Wallis test to assess differences between responses to each criterion for each category of participants' expertise level. The analysis showed no significant differences between beginners, intermediates and experts for any of the 21 criteria.

Regarding the optional "other" criteria, participants proposed several other options. The most relevant were: queries by period (e.g., release date, year or decade) and location (e.g., country of origin) of a recording (5 participants); part of the lyrics (2 participants); type of license, copyrighted or royalty-free (2 participants); the name of the record label (2 participants); album title (2 participants); presence of a solo section (2 participants); the number of players (e.g., if it is music for a duo, trio, ensemble, orchestra, band, big band) in the recording (2 participants).

Results of question 2. This open-ended question was analyzed with an inductive thematic analysis (Braun and Clarke, 2006). The following themes were identified:

Listening. By far the most recurring theme was listening to music with specific characteristics. 63 participants reported that they would use the system to search for music to listen to (e.g., *"As a system to suggest new music that has a certain set of characteristics"; "For recommendations: I could easily imagine myself singing that melody in my head or tapping a beat to the assistant and then get suggestions back"*). The listening activity would occur in a wide variety of situations such as leisure, relaxation, or parties, and as a background music during work, study or sport (e.g., *"Mainly for activities that see me away from a text keyboard, thus mainly listening while I am doing something else"; "To find a suitable ambient music for a party/dinner"*).

Playing/singing along. For 34 participants the system could be used to play or sing along the downloaded music, in a variety of music-making situations, such as practicing solo improvisation, rehearsing a new piece, or playing over a cover song (e.g., *"For playing along, for instance I want to practice using an instrument and I need a backing track with some specific musical parameters, e.g., key, scale, tempo"; "For choosing some pieces to practice specific techniques, or choosing some pieces to play with based over mood and/or difficulty"*).

Discovery. 20 participants commented that they would use the vocal assistant to discover new music, for different purposes such as listening or playing along (e.g., *"For discovering new songs depending on my requirements"; "I would use the system to survey the musical panorama of certain instruments and to use my curiosity to discover new music"; "To discover songs that use a particular feature (e.g. a scale, an instrument)"*).

Composition and music production. For 16 participants the system would be useful as a source of inspiration to compose music or to

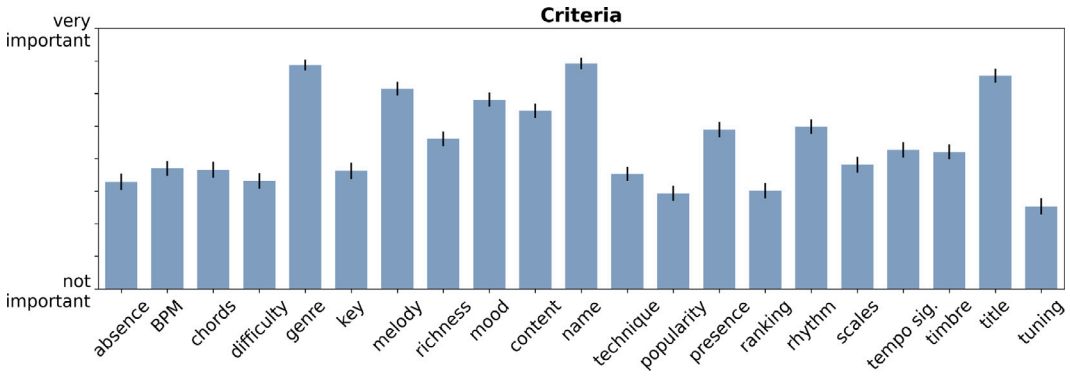

Fig. 1. Mean and standard error of the responses to each questionnaire item filled by the 112 participants.

produce music (e.g., *"To find inspiration to write new music"; "In case I need some reference while producing music"; "For sampling or remixing"*).

Education. 15 participants reported that they would use the system for educational purposes, for teaching or for learning (e.g., *"When teaching lessons to explain tempo, key, etc."; "to make examples to students"; "I will use it for improving my knowledge of music theory"; "To learn how a specific artist uses an instrument, scale, or key in order to emulate them"*).

Retrieval of forgotten pieces' names. 12 participants reported that they would use the system to find music that they can not remember the title of (e.g., *"When I want to listen to a song of which I do not remember the name"; "In case I cannot remember the title or the artist, I expect the system can tell me the song from the melodies I hum"*).

Playlist creation. 11 participants reported that they would use the system to create playlists for different situations (e.g., *"Being able to quickly generate a playlist for a party or similar, based on genre, mood, tempo and style criteria would be useful"; "To have a playlist of songs all in the same key to play along and practice"*).

Performance. 8 participants commented that they would use the vocal assistant in live concerts, for DJ sets or experimental music (e.g., *"If a DJ uses it, it could be an interesting tool to create music lists for DJ sets, thus grouping each song/part by genre, tempo and mood"*).

Results of question 3. An inductive thematic analysis performed on the open comments produced the following themes:

Usefulness and appreciation. 10 participants reported to have appreciated the idea of the system and found it very useful for several situations (e.g., *"It would be cool to be able to tell the assistant how you feel and get suggestions based on that. Because you don't always know what you want."; "This system would save a lot of time. Great idea."; "It would be a quite useful tools especially for music producers and or/sound designers which would allow them to access quicker and more efficiently to the music they are looking for"*).

Concerns. 8 participants reported to have some concerns in using the system, due to privacy reasons or because they do not like to interact with voice-based interfaces (e.g., *"I'm generally cautious when it comes to "open microphones" listening to me for security/integrity reasons"; "I will never use a voice recognition system in my home for privacy concerns"*).

3.2. Identified requirements

Ideally the system should be able to enable vocal queries by using all the 21 criteria above or a combination thereof. Indeed all criteria were ranked above 3 on average, which indicate at least a moderate need for it. However, for the final design of our system we had to take into account several constraints at implementation level. Firstly, for ease of implementation we opted to use the Software Development Kit (SDK) of conventional vocal assistants widely available on the market, which nevertheless limits the range of possible implementations. For instance it is not easy to implement a query-by-humming algorithm using such an SDK. Moreover, it is worth noting that efficient and comprehensive query-by-humming is still today an open research problem (Wang and Jang, 2015).

Relatedly, for our research purposes we could not use copyrighted music repositories (such as that of Spotify) but only deal with creative commons ones. We opted for Jamendo, which relies on a repository indexed by musical features and offers an API for queries serving our purposes. Jamendo, however, does not deal with well-known successes of famous artists. This implied that queries by artist or title of a piece did not make much sense, as artists and their pieces would have not been known by participants of a user test. However, such kind of conventional query is already implemented in some of today's vocal assistants (e.g., Amazon Alexa interacting with Spotify). Our focus was mainly on content-based features, which enable users to explore large music databases on the basis of specific music-related features. This represents a novelty in the realm of voice-based exploration of and retrieval from online music repositories.

For all these reasons, from the identified set of criteria we selected the ones that were most prominent and at the same time more feasible to implement in the context of the voice-based interface at hand: genre, mood, tempo, chords, key, and tuning. Users would have been empowered to use a combination of all such six criteria or a subset of them when vocally querying the Jamendo repository. Such a range of options extends with genre and mood the search criteria utilized in Turchet et al. (2020), which only involved tempo, chords, key, and tuning as content-based features to retrieve music from Jamendo using a guitar. Notably, to our best knowledge the complexity of such kind of queries has not been addressed in any previous content-based music search engines (see e.g., Casey et al., 2008; Jang et al., 2001; Knees et al., 2007; Wold et al., 1996; Tzanetakis and Cook, 2000).

We set other design requirements. Firstly, the system should be highly expressive, namely, it should enable for each search criterion a high number of parameters and a wide range for each of them (e.g., for BPM from 30 to 300). Secondly, the system should allow for both precise parameters (e.g., BPM = 140) and common words (e.g., "fast tempo"). Finally, the system should be equipped with an help functionality to support users in learning the system and recall them about the system's options.

4. Ecosystem architecture

The implemented IoMusT system aimed to enable the recreational or creative use of content retrieved from Jamendo via a voice-based interface. Fig. 2 shows a schematic representation of its main components, user–system vocal interactions, and data flow. The ecosystem technical apparatus comprises the following components:

Database. Jamendo is a digital repository of music content released predominantly under Creative Commons licenses. It provides access to music tracks created by independent artists with high recording standards, which are free for personal use. A random selection of 100k

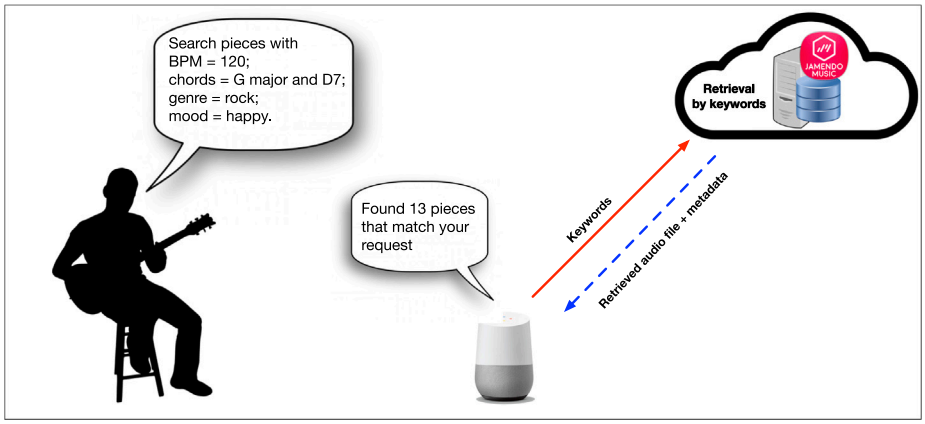

Fig. 2. Schematic representation of the implemented IoMusT ecosystem.

tracks were indexed along the six aforementioned criteria to create a searchable database.

Search service. The search API is a web service consisting of a persistent MongoDB database and short-lived worker processes that are created each time an API call is received, following the function-as-a-service paradigm. The database contains the automatic analysis of the 100k Jamendo tracks according to the six proposed musical criteria. Music pieces get added to the database by means of a one-time API call, which needs to be completed before a piece can show up in the search results. The database and API infrastructure are self-hosted and built upon the open-source OpenFaaS framework and Docker Swarm. Fig. 3 shows a conceptual diagram of the custom API components and interactions with external APIs.

Chord indexing was performed by the same algorithm that was used for the studies reported in Pauwels et al. (2017), Xambó et al. (2018a) and Turchet et al. (2020), genre and mood were scraped from the Jamendo API and the remaining criteria (key, tempo, and tuning) were calculated by the open-source Essentia extractor (Bogdanov et al., 2013).

The worker processes perform a retrieval of the tracks in the database according to the parameters passed through the API call (e.g., tempo, chords, etc.). For each set of search parameters, a number of tracks in the Jamendo catalogue that correspond optimally to the desired criteria is returned. Regarding the tempo, this means that tracks whose BPM falls within 5% of the requested BPM are withheld, with preference given to the ones that are closest to the requested value. Concerning tuning, all tracks whose detected tuning falls within 0.4% of the requested tuning is withheld, again with a preference for the smallest deviation. For the key search, this means that tracks with exactly the same key as the requested one are selected. The same occurs for mood and genre. Finally, a track matches with respect to chords if it is comprised of only the requested chords or a subset thereof (chord roots range from A to G# and the type is one of the following: major, minor, 7, major7, minor7).

The returned tracks are played back through the sound engine on the vocal assistant device by downloading them first using the Jamendo API. The web service returns up to 10 tracks if sufficient tracks are available that correspond to the requested search criteria. They are ordered in descending order according to the fitness of their match. By default the first track is used, the remainder are fallback options.

Voice-based interface. The voice assistant utilized was Hey Google for Google Home devices (which can also be run on any Android-based device), which was connected to the Internet via a Wi-Fi router. The system was implemented using the SDK provided by Google for the development of programs for their vocal assistant. The program was coded in Python and JavaScript and leveraged our custom search API and the Jamendo API.

Regarding the language for the interaction, English was chosen. Besides tracking the six criteria for the queries and combination thereof, the vocal assistant was equipped with a help functionality to support the user in learning how to use the system. When a list of musical pieces matching the parameters was retrieved, the user could instruct the vocal assistant to download and play one of the pieces. The selection of what piece to play was based on different options instructed by the user, such as the piece best matching the criteria or a random selection from the list of up to 10 matches.

5. Evaluation

5.1. Technical evaluation

We considered the total time taken to retrieve the desired musical content as a metric to technically evaluate the system. Such time includes the transmission time of the query, the processing time on the server, and the transmission time of the response. Therefore, we measured this time under different conditions, i.e., from simple queries with just one parameter to complex queries with all six parameters. Measurements were performed with 30 trials for each type of queries. On average for the case of one parameter the time was 1 s, while for six parameters the time taken was 29 s. Notably, most of the time taken to return the list of pieces was due to the search performed on the server hosting the database.

The objective quality of the returned results relies on the extent to which the returned music pieces do have the musical characteristics that were requested. This largely depends on the audio content analysis methods. Despite using state-of-the-art implementations, the extracted metadata stored in the database is inherently noisy due to algorithmic imperfections. For instance, an obviously wrong tempo of zero BPM was found for a non-negligible number of tracks. In order to perform a formal evaluation of the quality of the different content analysis methods, ground truth for the entire database would be required with human annotators. Creating this for all six criteria of all 100k tracks would be prohibitively time-consuming and expensive, which is indeed the reason why automatic analysis has been used in the first place.

No rigorous testing of the audio content analysis algorithms on this specific dataset can therefore be undertaken, but their performance on other data is discussed in their respective references. Informal testing shows that most of the errors are musically related, so the influence on the user experience due to these errors will remain limited. Also the cases where the automatic analysis breaks down completely (such as when zero BPM is detected) are not problematic, since those files will never be returned by the search process. In any case, unsatisfactory search results (due to bad indexing or any other reason) are accounted for as part of the subjective user experience evaluation.

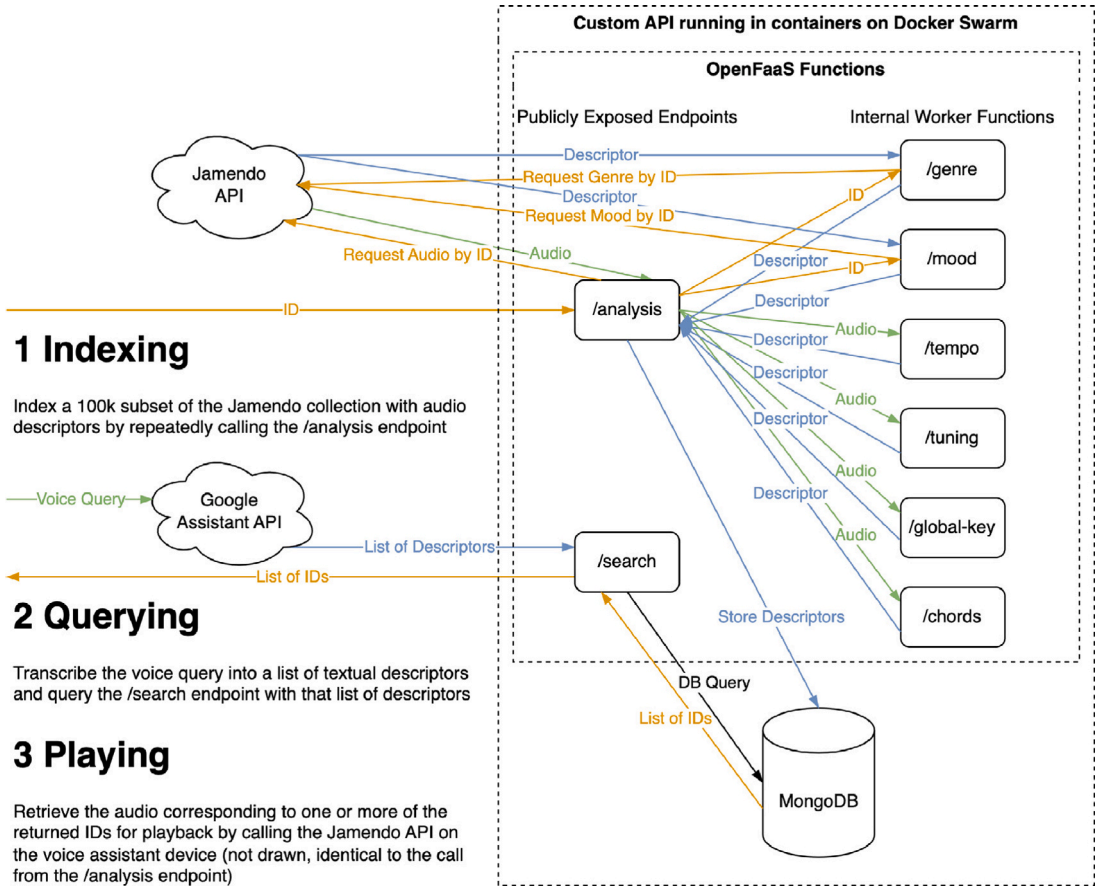

Fig. 3. Conceptual diagram of the custom API components and interactions with external APIs. Certain implementation details have been omitted. In practice, the setup also includes an API gateway and audio caching storage running in additional containers.

The underlying algorithms are designed based on their performance on a limited development set, comprising between hundreds and the low thousands of songs. Even for these sets, the extraction problems are considered unsolved (Pauwels et al., 2017; Faraldo et al., 2017). These shortcomings are multiplied by the application of these algorithms to a dataset that is multiple orders of magnitude larger. The sheer scale makes that the algorithms encounter songs that are far more varied in terms of musical content, musical culture and production value than what was in their development sets. Since there is no ground-truth available for this large set, it is hard to quantify how strong the performance deteriorates between development sets and our set. It is likely to be significant though, as can be seen in the large difference in key extraction performance between variants tailored to a specific genre or not (Faraldo et al., 2017).

The usage of confidence measures to promote high precision over a high recall can partially overcome these algorithmic imperfections, as shown in Pauwels et al. (2017). However, the pruning associated with such filtering by confidence has a negative effect on another scaling issue: the compound query. The distribution of musical characteristics is generally not uniform, but concentrated in a small number of values. Therefore, when combining multiple musical characteristics in a query, the number of songs satisfying such a query will decrease quickly for all but the most common query values. Not only does a song with the specific combination of query values need to exist in the dataset, all its musical characteristics also need to be recognized correctly (and with high enough confidence if considered). A dataset of 100k tracks quickly becomes not that large anymore.

5.2. User experience evaluation

Twenty participants were recruited (5 females, 15 males) aged between 20 and 37 (mean = 25.2, standard deviation = 3.77), and belonging to different nationalities (Italian, Dutch, Sri Lankan, British). They all played a variety of instruments (guitar, piano, ukulele, drums, synthesizers, tuba). Ten of them considered themselves beginner players, six were intermediate and four expert. All of them reported listening to or even playing along with songs streamed from services such as YouTube, Spotify, or Soundcloud. The experiments were conducted in part in a laboratory of University of Trento, in part in schools of music and in part at the home of participants.

Each participant was instructed by the experimenter about the system functionalities and the interactions modalities. After a familiarization phase of about 10 min, participants were invited to try the system by retrieving 5 musical pieces using the combination of keywords they preferred. For each downloaded music piece, participants were also asked to play along with it using their instrument. After having tried the system, participants were asked to fill an ad-hoc questionnaire. Such a questionnaire comprised the System Usability Scale questionnaire (SUS) (Brooke, 1996), the questionnaire to calculate the creativity support index (CSI) (Cherry and Latulipe, 2014) and a 9-point Self-Assessment Manikin (SAM).

In addition, participants were asked to answer the following open-ended questions devised to further assess the usability of the system, investigate whether the system was more suitable for composition, practice, or casual enjoyment purposes, and understand the hedonic qualities of the systems (Wechsung and Naumann, 2008).

- *What did you like the most in the system?*

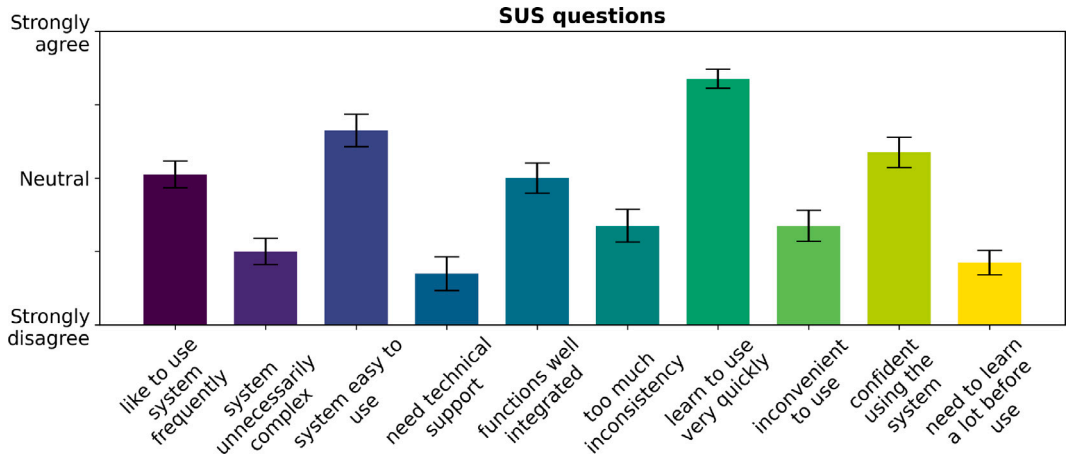

Fig. 4. Mean and standard error of the SUS questionnaire completed by the 20 subjects.

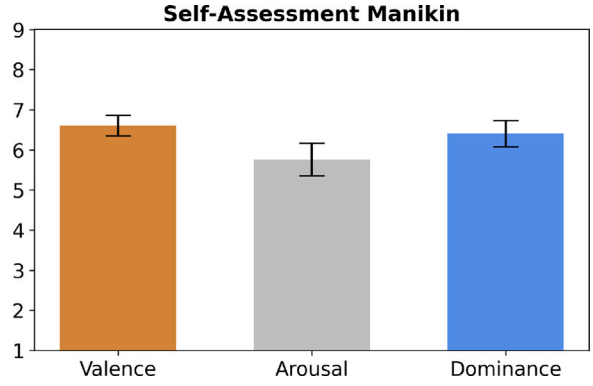

Fig. 5. Mean and standard error of the SAM questionnaire completed by the 20 subjects.

- *What did you like the least in the system?*
- *How would you improve the system?*
- *What is the added value of the system?*
- *Do you think that your way of searching the musical content would improve compared to the use of streaming services and involving the computer?*

Finally, participants were also allowed to leave an open comment about their experience. On average, participants took 1 h to complete the experiment.

5.3. Results

5.3.1. System usability scale

The SUS metric assesses the usability of a system on a scale from 0 to 100. As a point of comparison, an average SUS score of about 68 was obtained from over 500 studies. The system obtained a mean SUS score of 68.47 (95% confidence interval: [62.05; 74.88]), which is around average. Fig. 4 shows the breakdown of the result across the SUS topics. The results indicate that on average, participants found the system easy to use, quick to learn and to use without technical support.

5.3.2. Self-assessment manikin

Results of SAM are illustrated in Fig. 5. The figure shows that participants deemed that their experience was on average pleasant and sufficiently stimulating, as well as that they felt in control of the system.

Table 1
Average CSI results (SD reported in brackets). The highest average value is reported in bold in each column. The mean CSI score is 67.6 (SD = 12.3). Ranges: Avg. factor counts (0 to 5), avg. factor score (0 to 20), avg. weighted factor score (0 to 100).

Creativity factor	Avg. factor counts	Avg. factor score	Avg. weighted factor score
Enjoyment	2.7 (1.4)	13 (2.5)	34 (14.5)
Exploration	**3.9** (1.1)	**15.2** (2.4)	**59.9** (19.2)
Expressiveness	3.4 (1.1)	14.3 (3.9)	50.8 (25)
Immersion	2.2 (1.1)	8.6 (3.7)	18.2 (13.1)
Results Worth Effort	2.7 (1.1)	14.3 (2.2)	39.7 (18.2)

5.3.3. Creativity support index.

The CSI metric, ranging in [0, 100], enables to assess the ability of a tool to support the open-ended creation of new artifacts. The CSI section comprises 15 paired comparisons to determine the relative importance of the six creativity factors in musical practice tasks (Collaboration, Enjoyment, Exploration, Expressiveness, Immersion, Results Worth Effort). The scale related to collaboration, which is an aspect not present in the systems under study, was defaulted to 0 as indicated by the authors of CSI (Cherry and Latulipe, 2014). The system obtained a mean CSI of 67.7 (SD = 12.3). Table 1 presents the average CSI results broken down into factor counts (the number of times a creativity factor was judged more important than another for the task, as based on paired comparisons), factor scores (the ratings of the various factors irrespective of their importance for the task), and the weighted factor scores, which combine the factor counts and scores to make it more sensitive to the factors that are the most important to the given task. The achieved result is an indication of a relatively good creativity support.

The creativity factor which was judged the most important for the task of retrieving music from the repository was Exploration, closely followed by Expressiveness. This suggests that such factors are important to users engaged in the task, which is plausible as participants had to search for music according to a diverse set of criteria. The lowest average weighted factor score was reported for Immersion, which evidences that some participants could be distracted from the task. A plausible explanation for this result can be attributed to the fact that a waiting time to return the results is present. The Results Worth Effort scores indicate that the task was not perceived as too tiring.

5.3.4. Inductive thematic analysis

The following themes were identified:

Concept and novelty. Seven participants commented to have greatly appreciated the idea behind the system, i.e., the possibility of performing content-based queries via their voice. The system was deemed innovative for its ability to allow musicians to retrieve music that

corresponded more directly to their actual needs, especially for improvisation purposes (e.g., *"It's an upgrade to what I use currently (youtube)"; "Quite nice, it seems understanding me fairly good"*). For some of those participants the system was perceived even more novel because they were not very acquainted with the use of vocal assistants (e.g., *"I enjoyed it, it was kind of a first time since I don't usually use voice assistants"*). Participants also commented to prefer the style of interaction based on voice compared to the search with textual input on YouTube, given the fact that it was perceived as more natural and immediate (e.g., *"It's definitely easier to set up than going to youtube, searching music and manually filtering the results"*).

Usefulness. Five participants commented that the system would be very useful in their daily practice because it allows to avoid the use of the hands for inputting text on a PC or smartphone while handling the instrument (e.g., *"It can help to gather suggestions for practicing and improvising without the need of taking the instrument off the hands and without watching and interacting with a screen."; "It's certainly easier to operate with an instrument in hand than using YouTube."; "Definitely yes for searching music if my hands are already busy"; "It allows me to search even with my hands busy so it's a plus"*). This was deemed to have the potential to drastically improve their workflow.

Surprise and exploration. Four participants reported to have appreciated the possibility of discovering musical content that was unknown. For them such an element of surprise fostered the exploration of new content available on the database (e.g., *"I appreciated the randomness of the returned songs, I would not have had the possibility to discover new songs with those features with the standard searches on Spotify"*). Moreover, they commented that the retrieval of unknown songs was useful to them to go beyond their own comfort zones (e.g., *"With the system I could explore music that was unknown to me, so I could challenge myself in improvising with new styles knowing in advance the tonality"*). This was also reported to foster their musical creativity.

Jamendo's song variety was commented upon by seven participants. Four expert musicians found annoying the contrast between the songs quality, which ranges from very amateur to professional. On the contrary, three beginners found this variety as interesting and sometimes fun. One participant stated that *"When they [amateur quality songs] come up it puts me in a more lighthearted mood and I really feel like I'm having fun"*.

Features requests. Ten among experts and intermediate suggested to extend the search criteria, including the presence or absence of instruments, the use of complex chords (such as the suspended or the diminished ones), and the use of humming or singing to search for melodies. These criteria parallel those emerged during the requirements gathering phase. In particular nine participants requested the possibility to filter the songs by the presence of lyrics or by instrument. The main concern in these cases was that the lyrics or a given instrument can overshadow some of the practice activities, such as improvising. A participant verbalized this concern as *"I want to be protagonist of the improvisation session, not a supporting member"*.

A feature request to hear snippets of each song before choosing which one to play was very popular, thirteen participants suggested it. The proposed duration of the snippets ranges from 5 to 30 s, with a propensity for the lower end. Most participants desired the snippets in order to choose which song to play afterwards. An exception was made by a participant stating that *"a search with only snippets would be very useful for electronic dance music producers"*, effectively using the snippets as samples for music creation.

Parameters not needed. Three beginners reported that the range of options made available by the system were too many for them, and that they would not use all the expressive power afforded by the system (e.g., *"I'm very much a beginner so many of the system's functions are out of my skill level but I really like the search for genre and tonality"*). Two other participants commented that the tuning criteria was the least relevant, and would be used very rarely and only by someone really interested in playing with a non conventional tuning (i.e., A = 440 Hz).

Suggestions for improvements. The most relevant comment concerning improvements, expressed by seven musicians, was that the parameters of the retrieved songs did not perfectly match the search criteria expressed (e.g., *"An unexpected result was that in a few cases the requested chords and key were not correct"*). This in some cases led to dissatisfaction with the returned results.

6. Discussion

In general, during both the requirement gathering and evaluation phases participants reported to appreciate the idea of a vocal system allowing them to retrieve musical pieces with musical criteria. The developed system implemented only a restricted number of all the possible search criteria. Nonetheless, participants appreciated the six search possibilities offered by the system and positively assessed its usefulness within their creative practices. This was highlighted both by quantitative and qualitative results, which assessed both the system usability and user experience.

As evidenced by some participants, one of the strengths of the developed system is the novel kinds of interaction with online music repositories it affords while musicians are actually practicing. This is a known issue highlighted by different authors (see e.g., MacConnell et al. (2013), Martinez-Avila et al. (2019) and Wang et al. (2021)), which is referred in the literature to with the term "encumbered interaction". The study described in Martinez-Avila et al. (2019) identified that when a musician interacts with multiple support resources and devices, whilst having the instrument at hand, this results in physical impediments that emerge from multi-object manual (and one-handed) interactions. Our system effectively allows musicians to overcome the typical issues occurring while conducting the search of the musical content they want to play along with, such as the difficulty and the reduction of the freedom of movements due to the use of the computer/mobiles while holding the instrument. Notably, our approach to solve such encumbrances problem differs from that proposed by other authors. For instance, the study reported in Avila et al. (2019) focused on the possibility of augmenting the musical instrument with physical interventions (e.g., sensors to track gestures or a touchscreen to be placed onto the instrument body). Our approach does not entail any transformation of the instrument or invasive technique as it leverages the possibilities offered by vocal interfaces.

There were some contrasting comments made by participants. On the one hand some of the intermediate and experts requested new features as they felt limited in the expressiveness of the search, on the other hand some of the beginners suggested to provide less features as the range of options afforded by the system was overwhelming for them. A similar contrast was identified for the variety of the Jamendo's songs, where experts aimed to retrieve only songs with a high level of professionalism, while beginners preferred to retrieve songs recorded by amateur musicians. These differences in the participants' needs point to the need for personalization mechanisms based on the expertise level of the user.

The evaluation of the system also highlighted the technical limits of current algorithms for retrieving information from large datasets of musical content. Some participants noticed that the returned songs did not entirely match their query, especially for mood and chords. This calls for new research on the improvement of state of the art music information retrieval methods. This is even more relevant when considering more complex criteria desired by musicians to perform the search, as highlighted in the responses to the survey during the requirements gathering phase (such as melodies or rhythmic patterns expressed with the voice).

The time taken by the retrieval process is a crucial aspect for the overall user experience, as shown for a system similar to that reported here (Turchet et al., 2020). Only one participant reported a negative comment about the system's latency in returning the results, which in the case of the most complex queries, amounted to less than half a

minute. The greatest contribution to the latency was not attributable to the network but to the computations performed on the server. Therefore, to improve latency in the most demanding cases, there is the need to develop more performing cloud computing methods.

It is worth noticing that the general structure of the proposed system could be applied in domains different from the musical one, such as repositories of other audio content such as speech, non-musical sounds and the combination thereof Kotsakis and Dimoulas (2022).

7. Conclusion

This paper presented an Internet of Musical Things system devised to support recreational music-making, improvisation, composition, and music learning via vocal queries to an online music repository. The system involved a commercial voice-based interface and the Jamendo cloud-based repository of Creative Commons music content. The system enables the user to query the Jamendo music repository with six content-based features and each combination thereof: mood, genre, tempo, chords, key and tuning. Such queries differ from the conventional methods for music retrieval, which are based on the piece's title and the artist's name.

The system's requirements were identified by means an online survey with 112 musicians. This encompassed a variety of diversity factors, including nationality, expertise level, instrument played, age and gender. Most of musicians reported that they would be willing to use the system mostly for listening to music that has given musical characteristics. Another significant portion of musicians indicated that the proposed method of retrieval would be very useful for finding music to play along, especially for practicing purposes. Other participants suggested that the utility of the system lies in other musical activities such as composition, performance and education.

The actual evaluation of the system showed that the system is usable and generally provides a satisfactory user experience. Musicians appreciated the concept and found the system useful in a variety of musical activities. Importantly, the system was seen as a concrete solution to physical encumbrances that arise from the introduction of additional resources beyond musicians' instrument, such as interactive media resources like YouTube to search for music to play along. Finally, it is worth noticing that our system has also implications for accessibility, as it can support visually-impaired musicians in retrieving the wanted music by the sole use of their voice. Future works will focus on testing the system with such a category of users.

To date, networked voice-based interactions represent a scarcely explored line of research within the emerging field of the Internet of Musical Things. The authors hope that this work could inspire other practitioners to investigate future applications of vocal assistants interacting with online music collections to support musicians in their activities.

CRediT authorship contribution statement

Luca Turchet: Conceptualization, Methodology, Formal analysis, Investigation, Software, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Carlo Zanotto:** Software, Investigation, Writing – original draft. **Johan Pauwels:** Software, Validation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Avila, J.P.M., Hazzard, A., Greenhalgh, C., Benford, S., 2019. Augmenting guitars for performance preparation. In: Proceedings of the International Audio Mostly Conference. pp. 69–75.

Bazen, S., Bouvard, L., Zimmermann, J., 2015. Musicians and the creative commons: A survey of artists on jamendo. Inf. Econ. Policy 32, 65–76.

Bigham, J., Lau, T., Nichols, J., 2009. Trailblazer: enabling blind users to blaze trails through the web. In: Proceedings of the 14th International Conference on Intelligent User Interfaces. pp. 177–186.

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Herrera Boyer, P., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J., Serra, X., 2013. Essentia: An audio analysis library for music information retrieval. In: Proceedings of the International Society for Music Information Retrieval Conference. pp. 493–498.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3 (2), 77–101.

Brewer, R., Findlater, L., Kaye, J., Lasecki, W., Munteanu, C., Weber, A., 2018. Accessible voice interfaces. In: Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing. pp. 441–446.

Brooke, J., 1996. SUS–A quick and dirty usability scale. Usability Eval. Ind. 189 (194), 4–7.

Casey, M.A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M., 2008. Content-based music information retrieval: Current directions and future challenges. Proc. IEEE 96 96 (4), 668–696.

Cherry, E., Latulipe, C., 2014. Quantifying the creativity support of digital tools through the creativity support index. ACM Trans. Comput.-Hum. Interact. 21 (4), 21.

Christian, K., Kules, B., Shneiderman, B., Youssef, A., 2000. A comparison of voice controlled and mouse controlled web browsing. In: Proceedings of the Fourth International ACM Conference on Assistive Technologies. pp. 72–79.

Faraldo, A., Jorda, S., Herrera, P., 2017. A multi-profile method for key estimation in EDM. In: Audio Engineering Society Conference.

Fonseca, E., Pons Puig, J., Favory, X., Font Corbera, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., Serra, X., 2017. Freesound datasets: a platform for the creation of open audio datasets. In: Proceedings of the International Society for Music Information Retrieval Conference. International Society for Music Information Retrieval, pp. 486–493.

Font, F., Brookes, T., Fazekas, G., Guerber, M., La Burthe, A., Plans, D., Plumbley, M., Shaashua, M., Wang, W., Serra, X., 2016. Audio Commons: bringing Creative Commons audio content to the creative industries. In: Audio Engineering Society Conference: 61st International Conference: Audio for Games. Audio Engineering Society.

Granata, C., Chetouani, M., Tapus, A., Bidaud, P., Dupourqué, V., 2010. Voice and graphical-based interfaces for interaction with a robot dedicated to elderly and people with cognitive disorders. In: 19th International Symposium in Robot and Human Interactive Communication. IEEE, pp. 785–790.

Harada, S., Landay, J., Malkin, J., Li, X., Bilmes, J., 2006. The vocal joystick: evaluation of voice-based cursor control techniques. In: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility. pp. 197–204.

Harada, S., Wobbrock, J., Landay, J., 2011. Voice games: investigation into the use of non-speech voice input for making computer games more accessible. In: IFIP Conference on Human-Computer Interaction. Springer, pp. 11–29.

Igarashi, T., Hughes, J., 2001. Voice as sound: using non-verbal voice input for interactive control. In: Proceedings of the Annual ACM Symposium on User Interface Software and Technology. pp. 155–156.

Jang, J.-S.R., Lee, H.-R., Chen, J.-C., 2001. Super MBox: An efficient/effective content-based music retrieval system. In: Proceedings of the Ninth ACM International Conference on Multimedia. pp. 636–637.

Keller, D., Lazzarini, V., Pimenta, M., 2014. Ubiquitous Music. Springer.

Knees, P., Pohle, T., Schedl, M., Widmer, G., 2007. A music search engine built upon audio-based and web-based similarity measures. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 447–454.

Kopp, S., Brandt, M., Buschmeier, H., Cyra, K., Freigang, F., Krämer, N., Kummert, F., Opfermann, C., Pitsch, K., Schillingmann, L., et al., 2018. Conversational assistants for elderly users–the importance of socially cooperative dialogue. In: Proceedings of the AAMAS Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications Co-Located with the Federated AI Meeting, Vol. 2338.

Kostek, B., 2018. Listening to live music: life beyond music recommendation systems. In: 2018 Joint Conference-Acoustics. IEEE, pp. 1–5.

Kotsakis, R., Dimoulas, C., 2022. Extending radio broadcasting semantics through adaptive audio segmentation automations. Knowledge 2 (3), 347–364.

MacConnell, D., Trail, S., Tzanetakis, G., Driessen, P., Page, W., Wellington, N., 2013. Reconfigurable autonomous novel guitar effects (RANGE). In: Proceedings of the International Conference on Sound and Music Computing.

Martinez-Avila, J., Greenhalgh, C., Hazzard, A., Benford, S., Chamberlain, A., 2019. Encumbered interaction: a study of musicians preparing to perform. In: Proceedings of the Conference on Human Factors in Computing Systems. ACM, pp. 1–13, no. 476.

Pauwels, J., O'Hanlon, K., Fazekas, G., Sandler, M.B., 2017. Confidence measures and their applications in music labelling systems based on hidden Markov models. In: Proceedings of the Conference of the International Society for Music Information Retrieval. pp. 279–285.

Pradhan, A., Mehta, K., Findlater, L., 2018. "Accessibility came by accident" use of voice-controlled intelligent personal assistants by people with disabilities. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–13.

Skach, S., Xambó, A., Turchet, L., Stolfi, A., Stewart, R., Barthet, M., 2018. Embodied interactions with E-textiles and the internet of sounds for performing arts. In: Proceedings of the International Conference on Tangible, Embedded, and Embodied Interaction. ACM, pp. 80–87.

Stolfi, A., Ceriani, M., Turchet, L., Barthet, M., 2018. Playsound.space: Inclusive free music improvisations using audio commons. In: Proceedings of the Conference on New Interfaces for Musical Expression. pp. 228–233.

Su, J.-H., Hong, T.-P., Chen, Y.-T., Chin, C.-Y., 2022. High-performance content-based music retrieval via automated navigation and semantic features. Eng. Appl. Artif. Intell. 115, 105267.

Turchet, L., Antoniazzi, F., 2023. Semantic Web of Musical Things: achieving interoperability in the Internet of Musical Things. J. Web Semant. 75, 100758.

Turchet, L., Fischione, C., Essl, G., Keller, D., Barthet, M., 2018. Internet of Musical Things: vision and challenges. IEEE Access 6, 61994–62017.

Turchet, L., Pauwels, J., Fischione, C., Fazekas, G., 2020. Cloud-smart musical instrument interactions: Querying a large music collection with a smart guitar. ACM Trans. Internet Things 1 (3), 1–29.

Turchet, L., Zanetti, A., 2020. Voice-based interface for accessible soundscape composition: composing soundscapes by vocally querying online sounds repositories. In: Proceedings of Audio Mostly Conference. pp. 160–167.

Tzanetakis, G., Cook, P., 2000. 3D graphics tools for sound collections. In: Proceedings of the Conference on Digital Audio Effects.

Wang, C.-C., Jang, J.-S.R., 2015. Improving query-by-singing/humming by combining melody and lyric information. IEEE/ACM Trans. Audio, Speech, Lang. Process. 23 (4), 798–806.

Wang, B., Yang, M.Y., Grossman, T., 2021. Soloist: Generating mixed-initiative tutorials from existing guitar instructional videos through audio processing. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 1–14.

Wechsung, I., Naumann, A.B., 2008. Evaluation methods for multimodal systems: A comparison of standardized usability questionnaires. In: International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems. Springer, pp. 276–284.

Wold, E., Blum, T., Keislar, D., Wheaten, J., 1996. Content-based classification, search, and retrieval of audio. IEEE Multimedia 3 (3), 27–36.

Xambó, A., Pauwels, J., Roma, G., Barthet, M., Fazekas, G., 2018a. Jam with Jamendo: Querying a large music collection by chords from a learner's perspective. In: Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion. ACM, pp. 30:1–30:7. http://dx.doi.org/10.1145/3243274.3243291.

Xambó, A., Roma, G., Lerch, A., Barthet, M., Fazekas, G., 2018b. Live repurposing of sounds: MIR explorations with personal and crowdsourced databases. In: Proceedings of the International Conference on New Interfaces for Musical Expression.