1

Introduction

Audio is an integral and ubiquitous aspect of our daily lives: we intentionally produce sound (e.g. when communicating through speech or playing an instrument), we actively listen (e.g. to music or podcasts), can focus on a specific sound source in a mixture of sources, and we (even unconsciously) suppress sound sources internally (e.g. traffic noise). Similar to humans, algorithms can also generate, analyze, and process audio. This book focuses on the algorithmic analysis of audio signals, more specifically the extraction of information from musical audio signals.

Audio signals contain a wealth of information: by simply listening to an audio signal, humans are able to infer a variety of content information. A speech signal, for example obviously transports the textual information, but it also might reveal information about the speaker (gender, age, accent, mood, etc.), the recording environment (e.g., indoors vs. outdoors), and much more. A music signal might allow us to derive melodic and harmonic characteristics, understand the musical structure, identify the instruments playing, perceive the projected emotion, categorize the music genre, and assess characteristics of the performance as well as the proficiency of the performers. An audio signal can contain and transport a wide variety of content beyond these simple examples. This content information is sometimes referred to as *metadata*: data about (audio) data.

The field of *Audio Content Analysis (ACA)* aims at designing and applying algorithms for the automatic extraction of content information from the raw (digital) audio signal. This enables content-driven and content-adaptive services which describe, categorize, sort, retrieve, segment, process, and visualize the signal and its content.

The wide range of possible audio sources and the multi-faceted nature of audio signals results in variety of distinct ACA problems, leading to various areas of research, including

- *speech analysis*, covering topics such as automatic speech recognition [1, 2] or recognizing emotion in speech [3, 4],

An Introduction to Audio Content Analysis: Music Information Retrieval Tasks & Applications, Second Edition. Alexander Lerch.

- *urban sound analysis* with applications in noise pollution monitoring [5] and audio surveillance, i.e. the detection of dangerous events [6],
- *industrial sound analysis* such as monitoring the state of mechanical devices like engines [7] or monitoring the health of livestock [8], and, last but not least,
- *musical audio analysis*, targeting the understanding and extraction of musical parameters and properties from the audio signal [9].

This book focuses on the analysis of musical audio signals and the extraction of musical content from audio. There are many similarities and parallels to the areas above, but there exist also many differences that distinguish musical audio from other signals beyond simple technical properties such as audio bandwidth. Like an urban sound signal, music is a polytimbral mixture of multiple sound sources, but unlike urban sound, its sound sources are clearly related (e.g., melodically, harmonically, or rhythmically). Like a speech signal, a music signal is a sequence in a language with rules and constraints, but unlike speech, the musical language is abstract and has no singular meaning. Like an industrial audio signal, music has both tonal and noise-like components which may repeat themselves, but unlike the industrial signal, it conveys a (musical) form based on hierarchical grouping of elements not only through repetition but also through precise variation of rhythmic, dynamic, tonal, and timbral elements.

As we will see throughout the chapters, the design of systems for the analysis of musical audio often requires knowledge and understanding from multiple disciplines. While this text approaches the topic mostly from an engineering and Digital Signal Processing (DSP) perspective, the proper formulation of research questions and task definitions often require methods or at least synthesis of knowledge from fields as diverse as music theory, music perception, and psychoacoustics. Researchers working on ACA thus come from various backgrounds such as computer science, engineering, psychology, and musicology.

The diversity in musical audio analysis is also exemplified by the wide variety of terms referring to it. Overall, musical ACA is situated in the broader area of *Music Information Retrieval (MIR)*. MIR is a broader field that covers not only the analysis of musical audio but also symbolic (nonaudio) music formats such as musical scores and files or signals compliant to the so-called *Musical Instrument Digital Interface (MIDI)* protocol [10]. MIR also covers the analysis and retrieval of information that is music-related but cannot be (easily) extracted from the audio signal such as the artist names, user ratings, performance instructions in the score, or bibliographical information such as publisher, publishing date, the work's title. Other areas of research, such as music source separation and automatic music generation are often also considered to belong within MIR. Various overview articles clarify how the understanding of the field of MIR has evolved over time [11–16]. Other, related terms are also in use. Audio event

detection, nowadays often related to urban sound analysis, is sometimes described as computational analysis of sound scenes [17]. The analysis of sound scenes from a perceptual point of view has been described as *Computational Auditory Scene Analysis (CASA)* [18]. In the past, other terms have been used more or less synonymously to the term "audio content" analysis. Examples of such synonyms are *machine listening* and *computer audition*. Finally, there is the term *music informatics*, which encompasses essentially any aspect of algorithmic analysis, synthesis, and processing of music (although in some circles, its meaning is restricted to describe the creation of musical artifacts with software).

1.1 A Short History of Audio Content Analysis

Historically, the first systems analyzing the content of audio signals appeared shortly after technology provided the means of storing and reproducing recordings on media in the twentieth century. One early example is Seashore's Tonoscope, which enabled the pitch analysis of an audio signal by visualizing the fundamental frequency of the incoming audio signal on a rotating drum [19]. However, the more recent evolution of digital storage media, DSP methods, and machine learning during the last decades, along with the growing amount of digital audio data available through downloads and streaming services, has significantly increased both the need and the possibilities of automatic systems for analyzing audio content, resulting in a lively and growing research field.

Early systems for audio analysis were frequently so-called "expert systems" [20], designed by experts who implement their task-specific knowledge into a set of rules. Such systems can be very successful if there is a clear and simple relation between the knowledge and the implemented algorithms. A good example for such systems are some of the pitch-tracking approaches introduced in Section 7.3: as the goal is the detection of periodicity in the signal in a specific range, an approach such as the Autocorrelation Function (ACF), combined with multiple assumptions and constraints, can be used to estimate this periodicity and thus the fundamental frequency.

Later, data-driven systems became increasingly popular and traditional machine learning approaches started to show superior performance on many tasks. These systems extract so-called features from the audio to achieve a task-dependent representation of the signal. Then, training data are used to build a model of the feature space and how it maps to the inferred outcome. The role of the expert becomes less influential in the design of these systems, as they are restricted to selecting or designing a fitting set of features, curating a representative set of data, and choosing and parametrizing the machine learning approach. One prototypical example for these approaches is musical genre classification as introduced in Section 12.2.

Modern machine learning approaches include trainable feature extraction approaches (also referred to as feature learning, see Section 3.6) as they are a part of deep neural networks. These approaches have consistently shown superior performance for nearly all ACA tasks. The researcher seldom imparts much domain knowledge beyond choosing input representation, data, and system architecture of these systems. An example of such an end-to-end system could be a music genre classification system based on a neural network with a convolutional architecture with a Mel Spectrogram input.

It should be pointed out that while modern systems tend to have superior performance, they also tend to be less interpretable and explainable than traditional systems. For example, deducing the reason for a false classification result in a network-based system can be difficult, while the reason is usually easily identifiable in a rule-based system.

1.2 Applications and Use Cases

The content extracted from music signals improves or enables various forms of content-based and content-adaptive services, which allow to sort, categorize, find, segment, process, and visualize the audio signal based on its content.

1.2.1 Music Browsing and Music Discovery

One of the most intuitive applications is the content-based search for music in large databases for which consistent manual annotation by humans is often infeasible. This allows, for example, Disk Jockeys (DJs) to retrieve songs in a specific tempo or key, to find music pieces in a specific genre or mood, or to search for specific musical characteristics by example. The same information can be used in end consumer applications such as audio-based music recommendation and playlist generation systems using an in-depth understanding of the musical content [21].

The content-based annotation and representation of the audio signal can also be used to design new interfaces to access data. Fingerprinting (see Chapter 11), for example allows to identify a song currently played from a large database of songs, while Query-by-Humming systems identify songs through a user-hummed melody. Content can also be utilized for new ways of sound visualization and user interaction; a database of music could, for example be explored by a user navigating a virtual music similarity space [22].

1.2.2 Music Consumption

Audio analysis has already started to transform consumer-facing industries such as streaming services as mentioned above. So far, most services focus on

recommending music to be played back, but in the near future, we might see the rise of creative music listening applications that enable the listener to interact with the content itself instead of being restricted to only choosing content to be played. This could include, for example, the gain adjustment for individual voices, replacing instruments or vocalists, or interactively changing the musical arrangement or even stylistic properties of the music piece.

1.2.3 Music Production

Knowledge of the audio content can improve music production tools in various dimensions. On the one hand, content information can enable a more "musical" software interface, e.g. by displaying score-like information synchronized with the audio data, and thus enabling an intuitive approach to editing the audio data. On the other hand, production software usage could be enhanced in terms of productivity and efficiency: the better a system understands the details and properties of incoming audio streams or files, the better it can adapt, for instance, by applying default gain and equalization parameters [23] or by suggesting compatible audio from a sound library. Intelligent music software might also support editors by offering automatic artifact-free splicing of multiple recordings from one session or selecting error-free recordings from a set of recordings.

Modern tools also enhance the creative possibilities in the production process. For example, creating harmonically meaningful background choirs by analyzing the lead vocals and the harmony track is already technically feasible and commercially available today. Knowing and isolating sound sources in a recording could enable new ways of modifying or morphing different sounds for the creation of new soundscapes, effects, and auditory scenes.

1.2.4 Music Education

The potential of utilizing technology to assist music (instrument) education has been recognized as early as the 1930s, when Seashore pointed out the educational value of scientific observation of music performances [24]. The availability of automatic and fast audio analysis can support the creation of artificially intelligent music tutoring software, which aims at supplementing teachers by providing students with insights and interactive feedback by analyzing and assessing the audio of practice sessions. An interactive music tutor with in-depth understanding of the musical score and performance content can highlight problematic parts of the students' performance, provide a concise yet easily understandable analysis, give objective and specific feedback on how to improve, and individualize the curriculum depending on the students' mistakes and general progress. At the same time, music tutoring software can provide an accessible, objective, and reproducible analysis.

1.2.5 Generative Music

Machine-interpretable content information can also feed generative algorithms. The automatic composition and rendition of music is emerging as a challenging yet popular research direction [25], gaining interest from both research institutions and industry. While bigger questions concerning capabilities and restrictions of computational creativity as well as aesthetic evaluation of algorithmically generated music remain largely unanswered, practical applications such as generating background music for user videos and commercial advertisements are currently in the focus of many researchers. The interactive and adaptive generation of sound tracks for video games as well as individualized generation of license-free music content for streaming is additional long-term goals of considerable commercial interest.

References

1 Xuedong Huang, James Baker, and Raj Reddy. A Historical Perspective of Speech Recognition. *Communications of the ACM*, 57(1):94–103, January 2014. ISSN 0001-0782. doi: 10.1145/2500887. URL https://doi.org/10.1145/2500887.

2 Dong Yu and Li Deng. *Automatic Speech Recognition - A Deep Learning Approach. Signals and Communication Technology*. Springer, London, 2015. ISBN 978-1-4471-5779-3. URL https://doi.org/10.1007/978-1-4471-5779-3.

3 Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing Emotion in Speech. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 3, pages 1970–1973, Philadelphia, PA, October 1996. doi: 10.1109/ICSLP.1996.608022.

4 Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognition*, 44(3):572–587, March 2011. ISSN 0031-3203. doi: 10.1016/j.patcog.2010.09.020. URL https://www.sciencedirect.com/science/article/pii/S0031320310004619.

5 Juan Pablo Bello, Charlie Mydlarz, and Justin Salamon. Sound Analysis in Smart Cities. In *Computational Analysis of Sound Scenes and Events*, pages 373–397. Springer, Cham, 2018. ISBN 978-3-319-63449-4. URL https://link.springer.com/chapter/10.1007/978-3-319-63450-0_13.

6 Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. Audio Surveillance: A Systematic Review. *ACM Computing Surveys*, 48(4):52:1–52:46, February 2016. ISSN 0360-0300. doi: 10.1145/2871183. URL https://doi.org/10.1145/2871183.

7 Sascha Grollmisch, Jakob Abeßer, Judith Liebetrau, and Hanna Lukashevich. Sounding Industry: Challenges and Datasets for Industrial Sound Analysis.

In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 1–5, A Coruna, Spain, September 2019. doi: 10.23919/EUSIPCO.2019.8902941. ISSN: 2076-1465.

8 Dries Berckmans, Martijn Hemeryck, Daniel Berckmans, Erik Vranken, and Toon van Waterschoot. Animal Sound... Talks! Real-time Sound Analysis for Health Monitoring in Livestock. In *Proceedings of the International Symposium on Animal Environment and Welfare (ISAEW)*, Chongqing, China, 2015.

9 Daniel P W Ellis. Extracting Information from Music Audio. *Communications of the ACM*, 49(8):32–37, August 2006. ISSN 00010782. doi: 10.1145/1145287.1145310. URL http://portal.acm.org/citation.cfm?doid=1145287.1145310.

10 MIDI Manufacturers Association. Complete MIDI 1.0 Detailed Specification V96.1, 2nd edition. Standard, MMA, 2001.

11 J Stephen Downie. Music Information Retrieval. *Annual Review of Information Science and Technology*, 37:295–340, 2003.

12 Nicola Orio. Music Retrieval: A Tutorial and Review. *Foundations and Trends in Information Retrieval*, 1(1):1–90, 2006.

13 Michael A Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. Content-based Music Information Retrieval: Current Directions and Future Challenges. *Proceedings of the IEEE*, 96(4):668–696, April 2008. ISSN 1558-2256. doi: 10.1109/JPROC.2008.916370. Conference Name: Proceedings of the IEEE.

14 Markus Schedl, Emilia Gómez, and Julián Urbano. Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends® in Information Retrieval*, 8(2-3):127–261, September 2014. ISSN 1554-0669, 1554-0677. doi: 10.1561/1500000042. URL http://www.nowpublishers.com/article/Details/INR-042.

15 Alexander Lerch. Music Information Retrieval. In Stefan Weinzierl, editor, *Akustische Grundlagen der Musik*, number 5 in Handbuch der Systematischen Musikwissenschaft, pages 79–102. Laaber, 2014. ISBN 978-3-89007-699-7.

16 John Ashley Burgoyne, Ichiro Fujinaga, and J Stephen Downie. Music Information Retrieval. In Susan Schreibman, Ray Siemens, and John Unsworth, editors, *A New Companion to Digital Humanities*, pages 213–228. John Wiley & Sons, Ltd, 2015. ISBN 978-1-118-68060-5. doi: 10.1002/9781118680605.ch15. URL http://onlinelibrary.wiley.com/doi/10.1002/9781118680605.ch15/summary.

17 Tuomas Virtanen, Mark D Plumbley, and Dan Ellis, editors. *Computational Analysis of Sound Scenes and Events*. Springer International Publishing, Cham, 2018. ISBN 978-3-319-63449-4 978-3-319-63450-0. doi: 10.1007/978-3-319-63450-0. URL http://link.springer.com/10.1007/978-3-319-63450-0.

18 DeLiang Wang and Guy J Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006. ISBN 978-0-470-04338-7. URL https://ieeexplore.ieee.org/book/5769523.

19 Carl E Seashore. A Voice Tonoscope. *Studies in Psychology*, 3:18–28, 1902.

20 Peter J F Lucas and Linda C van der Gaag. *Principles of Expert Systems*. Addison-Wesley, 1991.

21 Peter Knees, Markus Schedl, and Masataka Goto. Intelligent User Interfaces for Music Discovery: The Past 20 Years and What's to Come. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 44–53, Delft, Netherlands, 2019. URL http://archives.ismir.net/ismir2019/paper/000003.pdf.

22 Masahiro Hamasaki, Masataka Goto, and Tomoyasu Nakano. Songrium: Browsing and Listening Environment for Music Content Creation Community. In *Proceedings of the Sound and Music Computing Conference (SMC)*, page 8, Maynooth, Ireland, 2015.

23 Joshua D Reiss and Øyvind Brandtsegg. Applications of Cross-Adaptive Audio Effects: Automatic Mixing, Live Performance and Everything in Between. *Frontiers in Digital Humanities*, 5, 2018. ISSN 2297-2668. doi: 10.3389/fdigh.2018.00017. URL https://www.frontiersin.org/articles/10.3389/fdigh.2018.00017/full.

24 Carl E Seashore. *Psychology of Music*. McGraw-Hill, New York, 1938.

25 Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. *Deep Learning Techniques for Music Generation*. Computational Synthesis and Creative Systems. Springer International Publishing, Cham, 2020. ISBN 978-3-319-70162-2 978-3-319-70163-9. doi: 10.1007/978-3-319-70163-9. URL http://link.springer.com/10.1007/978-3-319-70163-9.