Chapter 2
Motivation, Aims, and Solutions

It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment.
—Carl Friedrich Gauss

2.1 Motivation of Intelligent Audio Analysis

There are numerous scenarios and fields for potential application of Intelligent Audio Analysis that are commercially interesting and may help us in our daily lives. These are detailed out in the application part of this book (Part III) that aims to give some practical examples, but a more general perspective on use-cases of the whole field is given for a motivational introduction at this point. Intelligent Audio Analysis is currently used and holds future promises in particular for

Audio Encoding: Obviously, in an acoustic representation, highest bitrates are required, which can be eased step-wise by going to partly or fully parametric representation [1], and partly or fully symbolic representation (cf. Fig. 2.1). As for speech, 'symbolic' could thereby be phones as acoustic realisations of phonemes, which are "the smallest segmental unit of sound employed to form meaningful contrasts between utterances" [2]. In the case of music, 'symbolic' could refer to note events or chords, etc. However, highest bit rate reduction is only reached by semantic encoding—though obviously at the highest loss factor as, rather than preserving the original audio, only its semantics are kept for storage or transmission via highly band limited channels. This then requires to synthesise audio at the moment of decoding based on these semantics. In music, an example would be note events and instrumentation saved in symbolic representation for storage and later synthesis for play-back. However, compromises can be made also at this level by combination with (few) parameters or even highly compressed acoustics—the semantics can then touch certain aspects of the audio signal for good reproduction at the moment of decoding and regeneration.

B. W. Schuller, *Intelligent Audio Analysis*, Signals and Communication Technology,
DOI: 10.1007/978-3-642-36806-6_2, © Springer-Verlag Berlin Heidelberg 2013

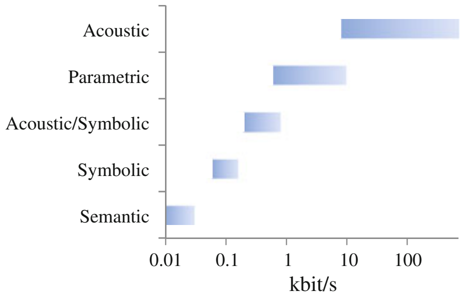

Fig. 2.1 A rough overview on obtainable audio bit-rates by partly lossy compression depending on the representation type.

Audio Alteration: In a chain of analysis, edition, and synthesis, audio can be modified and altered. Examples include voice transformation [3] including for example the change of the emotional tone of a voice, and music alteration such as combining drum tracks from one musical piece with the singer of another, etc.
Audio Retrieval: In audio search, manifold search tags are used today and can be used in the future such as by speaker identity or emotion, music artist or genre and positions of the chorus, sound type, etc. However, such information needs to be provided at first and additionally stored. As this may involve considerable human labelling effort and labelling may easily be erroneous if larger user groups of laymen are involved, Intelligent Audio Analysis may help to assess such information fully automatically off-line or even on-line.
Audio-based Interaction: In Human-Machine and Human-Robot communication, machine listening and understanding capabilities beyond speech and sound recognition and interpretation can allow for injection of 'social competence'. For example, speaker state and trait analysis allows for improved socio-emotional contextual comprehension of a machine. In music analysis, powerful user-interfaces can be provided to musicians, that allow for example for user input by clapping, singing, humming or playing of real musical instruments for interaction with the machine.
Monitoring and Surveillance: In this domain, speaker states can be of interest, such as sleepiness or intoxication of responsible persons in steering and control tasks [4]. Another example in this respect is monitoring of a customer's interest in sales presentations [5]. Also terrorism and vandalism alert systems may be realised by such systems—potentially combining speech and sound analysis [6]. An example of a hardware product is the WhyCry®—a device that aims to indicate a new-born's annoyance, boredom, hunger, sleepiness, and stress to less experienced parents. In music analysis, monitoring can for example be used for on-line auto mixing and balancing. Sound monitoring can for example be used to ensure proper functionality of bearings, pipelines, etc.
Coaching: Voice coaching includes training for public speeches or help in foreign language acquisition [4], but also holds promises for empowerment and inclusion. In the European ASC-Inclusion project,[1] children with autism spectrum condition shall

[1] http://www.asc-inclusion.eu

acquire improved socio-emotional skills by digital gaming including appropriate interpretation and expression of emotion. This example also includes monitoring and alteration, as their vocal expression is monitored in the game and the voice is altered for exemplification. In the music domain, a learning software can notify a student of an instrument if mistakes occur as by Fraunhofer's "Songs2See", or help in training vibrato singing [7], etc.

Entertainment: As the entertainment sector can often be more forgiving if accuracies are not at perfection level, this domain has seen many products make it to the market by now. Such software includes a console game around speech-based deception recognition ("Truth or Lies—Someone Will Get Caught", THQ® Entertainment) already appeared on the market. Software centred around singing intonation in Karaoke-style games such as "SingStar" and "RockBand" by Harmonix or more recently Ubisoft®'s guitar learning game "Rocksmith®" based on real guitar audio analysis are examples of huge market success.

Despite the appearance of first commercial and non-commercial usage of Intelligent Audio Analysis products and solutions, the state-of-the-art today is often not sufficient for the often very high requirements given by several of the above use-cases. According research work is thus still urgently needed. In addition, standard references in the literature that provide a broader perspective are just to appear given the rather young age of the field and its more recent emergence on a broader level.

2.2 Aims of the Book

It is the aim of this book to help allow for improved and extended exploitation of Intelligent Audio Analysis in the illustrated and further application scenarios. In particular and by that, the goals are as follows:

1. To provide a unified perspective on audio analysis tasks and a broad overview on recent advancements in the field exemplified primarily by work of the author and his colleagues. The intention is to stimulate synergies arising from transfer of methods and lead to a holistic audio analysis [8]—audio is usually highly complex and blended in the real world, but research is usually focused on isolated aspects at the present day.

2. To help approach improved robustness and reliability of today's Intelligent Audio Analysis systems by suited and innovative methods.

3. To stimulate extension of the range of Intelligent Audio Analysis applications by showing its potential in new tasks that were not or hardly touched in the literature so far, which, however, can be of broad commercial and technical interest.

4. To provide the reader with benchmark results and standardised test-beds for a broader range of audio analysis tasks. The main focus thereby lies on the parallel advancement of realism in audio analysis, as too often today's results are overly optimistic owing to idealised testing conditions.

5. To show deficiencies in current approaches and future perspectives in and for the field.

2.3 Solutions

From a technical point of view, the discussed solutions to the described ends foremostly consist of the inventory provided by the methods of pattern recognition. This includes advanced and recent methods of signal processing and machine learning. In more detail, these are:

Audio enhancement and source separation as needed for emphasising the characteristics and isolation of the signal part of interest.

Brute-forcing of large heterogenous audio feature spaces to provide a broad feature basis for the space initialisation in the approach of new audio tasks.

Careful design of new audio feature types as systematic brute-forcing may have its limitations.

Combination, adaptation, and application of recent learning methods to profit from synergies and inject new paradigms such as graphical modelling aspects and long short-term memory into the machine learning process and enable partly supervised self-learning.

As for the non-technical side, practical solutions include in the first place:

Establishment of unified test-beds and transparent benchmarks as this invites the research community to compare results in a well-defined way and by that may help to advance on the state-of-the-art. This includes or partly requires the following two aspects worth mentioning in isolation.

Collection and annotation of suited audio data to consider new tasks of Intelligent Audio Analysis or enrich the ever sparse data-base in the field.

Provision of standardised (open-source) software implementations where such are currently missing to allow for comparability of findings and potentially code additions by others.

References

1. Ruske, G.: Automatische Spracherkennung, 2nd edn. Methoden der Klassifikation und Merkmalsextraktion. Oldenbourg, Munich, Germany (1993)
2. I. P. Association: Phonetic Description and the IPA Chart, Chapter 2, pp. 3–17. Cambridge University Press, Cambridge (1999)
3. Stylianou, Y.: Voice transformation: A survey. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp 3585–3588. Taipei, Taiwan (2009)
4. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: Paralinguistics in speech and language–state-of-the-art and the challenge. Comp. Speech Lang. Special Issue Paralinguistics Naturalistic Speech Lang. 27(1), 4–39 (2013)
5. Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H.: Being bored? recognising natural interest by extensive audiovisual integration for real-life application. Image Vis. Comput. Special Issue Vis. Multimodal Anal. Hum. Spontaneous Behavior 27(12), 1760–1774 (2009)
6. Schuller, B., Wimmer, M., Arsić, D., Moosmayr, T., Rigoll, G.: Detection of security related affect and behaviour in passenger transport. In: Proceedings INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, incorporating

12th Australasian International Conference on Speech Science and Technology, SST 2008, pp. 265–268, Brisbane, Australia, ISCA/ASSTA, ISCA (2008)
7. Weninger, F., Amir, N., Amir, O., Ronen, I., Eyben, F., Schuller, B.: Robust feature extraction for automatic recognition of vibrato singing in recorded polyphonic music. In: Proceedings 37th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2012, pp. 85–88, Kyoto, Japan, IEEE, IEEE (2012)
8. Weninger, F., Schuller, B., Liem, C., Kurth, F., Hanjalic, A.: Music information retrieval: An inspirational guide to transfer from related disciplines. In: Müller, M., Goto, M. (eds.) Multimodal Music Processing, Seminar, vol. 69, pp. 195–215. 1041 of Dagstuhl Follow-UpsSchloss Dagstuhl, Germany (2012)

http://www.springer.com/978-3-642-36805-9

Intelligent Audio Analysis
Schuller, B.W.
2013, XXVIII, 345 p., Hardcover
ISBN: 978-3-642-36805-9