Chapter 3
Audio Feature Extraction for Similarity Measurement

In this chapter, we will introduce techniques to extract audio features for the purpose of modeling similarity between music pieces. As similarity is a crucial concept for any retrieval and recommendation task, having at hand accurate audio features and corresponding similarity functions is key to building music retrieval systems. Derived from low-level audio features, but not yet representing semantically meaningful music concepts, this information can be categorized as mid-level features, according to Fig. 1.7. This chapter builds upon the previous one in that some of the audio features introduced here rely on the low-level features discussed earlier. In the following section, we will first briefly touch upon psychoacoustic models of music perception, which can be used to ameliorate music feature extraction. Subsequently, we will present the two main categories of audio features used for music similarity: frame-level and block-level features. We will describe state-of-the-art feature extractors in both categories and detail how the resulting feature vectors for a whole piece of music can be aggregated to form a global representation. Techniques covered range from simple statistical summarization to vector quantization to Gaussian mixture models. From the resulting global acoustic feature representation, we show how similarity between music pieces can be computed. The chapter is concluded with a discussion of (and an approach to solve) a problem known as “hubness,” which occurs when computing similarities in high-dimensional feature spaces.

3.1 Psychoacoustic Processing

There is an obvious discrepancy between physical characteristics of sound and its human perception. The frequently used Mel frequency cepstral coefficients, for instance, employ a nonlinear scale to model feature values according to how humans perceive pitch intervals; cf. Sect. 3.2.1. We cannot give a comprehensive discussion
of the field of psychoacoustics here but will outline some of the most important psychoacoustic facts to consider when elaborating audio feature extractors. In particular, we focus on psychoacoustic aspects of loudness and frequency. The reader interested in more details should refer to [124] or [360].

### 3.1.1 Physical Measurement of Sound Intensity

On a physical level, the intensity of a sound is typically described as sound pressure level and measured in decibel (dB). Intensity in dB is defined as shown in Eq. (3.1), where $i$ refers to sound power per unit area, measured in watts per m$^2$, and $i_0$ is the hearing threshold of $10^{-12}$ W/m$^2$.

$$i_{dB} = 10 \cdot \log_{10} \left( \frac{i}{i_0} \right)$$  \hspace{1cm} (3.1)

Decibel is a logarithmic unit. An increase of 10 dB thus corresponds to an increase of ten times in sound pressure level.

### 3.1.2 Perceptual Measurement of Loudness

Human perception of loudness generally varies with sensitivity of the ear, which again varies between different people. In addition, perception of loudness also varies between different frequencies of sound. The human ear is most sensitive to frequencies between 2000 and 5000 Hz. Tones with the same physical sound pressure, measured in decibel (dB), but with frequencies out of this range are thus perceived as softer. To describe sound pressure levels that are perceived as equally loud, irrespective of sound frequency, several listening experiments have been conducted in research groups around the world. The results have eventually been integrated into the ISO standard 226:2003. Figure 3.1 shows the resulting equal loudness curves. Each curve describes the relationship between sound pressure (dB) and a particular level of loudness perception, varying over frequencies. The unit for loudness perception is phon. Phon is defined as having the equal dB value at a frequency of 1000 Hz. Thus, for a pure tone of 1000 Hz, the dB value equals the phon value, which can also be seen in the figure.

To account for human perception of loudness in music signal processing, the transformation from dB to phon is a vital step. As illustrated in Fig. 3.1, the phon scale reflects the nonlinear relationship between sound pressure and human sensation of loudness. Please note, however, that the phon scale is still logarithmic.

For easier interpretation due to its linearity, the sone scale defines 1 sone as being equivalent to 40 phons in terms of loudness, which again amounts to a physical sound pressure level of 40 dB for a signal with 1000 Hz. Doubling the sone value