Music Information Retrieval: An Inspirational Guide to Transfer from Related Disciplines

Felix Weninger[*1], Björn Schuller[1], Cynthia C. S. Liem[†2], Frank Kurth[3], and Alan Hanjalic[2]

1 **Technische Universität München**
Arcisstraße 21, 80333 München, Germany
weninger@tum.de

2 **Delft University of Technology**
Mekelweg 4, 2628 CD Delft, The Netherlands
c.c.s.liem@tudelft.nl

3 **Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie FKIE**
Neuenahrer Straße 20, 53343 Wachtberg, Germany
frank.kurth@fkie.fraunhofer.de

Abstract

The emerging field of Music Information Retrieval (MIR) has been influenced by neighboring domains in signal processing and machine learning, including automatic speech recognition, image processing and text information retrieval. In this contribution, we start with concrete examples for methodology transfer between speech and music processing, oriented on the building blocks of pattern recognition: preprocessing, feature extraction, and classification/decoding. We then assume a higher level viewpoint when describing sources of mutual inspiration derived from text and image information retrieval. We conclude that dealing with the peculiarities of music in MIR research has contributed to advancing the state-of-the-art in other fields, and that many future challenges in MIR are strikingly similar to those that other research areas have been facing.

1998 ACM Subject Classification H.5.5 Sound and Music Computing, J.5 Arts and Humanities–Music, H.5.1 Multimedia Information Systems, I.5 Pattern Recognition

Keywords and phrases Feature extraction, machine learning, multimodal fusion, evaluation, human factors, cross-domain methodology transfer

Digital Object Identifier 10.4230/DFU.Vol3.11041.195

1 Introduction

Music Information Retrieval (MIR) still is a relatively young field: Its first dedicated symposium, ISMIR, was held in 2000, and a formal society for practitioners in the field, taking over the ISMIR acronym, was only established in 2008. This does not mean that all work in MIR needs to be newly invented: Analogous or very similar topics and areas to those currently of interest in MIR research may already have been researched for years, or even decades, in neighboring fields. Reusing and transferring findings from neighboring fields, MIR research can jump-start and stand on the shoulders of giants. At the same time, the

* Felix Weninger is funded by the German Research Foundation through grant no. SCHU 2508/2-1.
† The work of Cynthia Liem is supported in part by the Google European Doctoral Fellowship in Multimedia.

Multimodal Music Processing. *Dagstuhl Follow-Ups, Vol. 3.* ISBN 978-3-939897-37-8.
Editors: Meinard Müller, Masataka Goto, and Markus Schedl; pp. 195–216

Dagstuhl Publishing
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany

nature of music data may pose constraints or peculiarities that press for solutions beyond the trodden paths in MIR, and thus can be of inspiration the other way around too. Such opportunities for methodology transfer, both to and from the MIR field, are the focus of this chapter.

In engineering contexts, audio typically is considered to be the main modality of music. From this perspective, an obvious neighboring field to look at is automatic speech recognition (ASR), which just like MIR strives to extract information from audio signals. Section 2 will discuss several methodology transfers from ASR to MIR, while Section 3 gives a detailed example of one of the first successful transfers from MIR back to ASR. Section 4 focuses on the topic of evaluation, in which current MIR practice has strong connections to classical approaches in Text Information Retrieval (IR). Finally, in Section 5, we consider MIR from a higher-level, more philosophical viewpoint, pointing out similarities in open challenges between MIR and Content-Based Image and Multimedia Retrieval, and arguing that MIR may be the field that can give a considerable push towards addressing these challenges.

2 Synergies between Speech and Music Analysis

As stated above, it is hardly surprising that audio-based MIR has been influenced by ASR research—as obvious opportunities to transfer ASR technologies to MIR, lyrics transcription [38] or keyword spotting in lyrics [17] can be named. Yet, there are more intrinsic synergies between speech and music analysis, where similar methodologies can be applied to seemingly different tasks. These will be the focus of the following section. We point out areas where speech and music analysis have been sources of mutual inspiration in the past, and sketch some opportunities for future methodology transfer.

2.1 Multi-Source Audio Analysis in Speech and Music

Generally, music signals are composed of multiple sources, which can correspond to instruments, singer(s), or the voices in a polyphonic piano piece; thus, aspects of multi-source signal processing can be considered as an integral part of MIR. Similarly, research on speech recognition in the presence of interfering sources (environmental noise, or even other speakers) has a long tradition, resulting in numerous studies on source separation and model-based robust speech recognition. Many approaches for speech source separation deal with multi-channel input from microphone arrays by beamforming, i. e., exploitation of spatial information. An example of such beamforming in music signals is the well-known 'karaoke effect' to remove the singing voice in commercial stereophonic recordings: Many popular songs are mixed with the vocals being equally distributed to the left and right channels, which corresponds to a center position of the the vocalist in the recording/playback environment. In that case, the vocals can be simply eliminated by channel subtraction, which can be regarded as a trivial example of integrating spatial information into source separation. However, to highlight the aspects of methodology transfer, we restrict the following discussion to monaural (single-channel) analysis methods: We argue that the constraints of music signal processing—where usually no more than two input channels are available—have leveraged a great deal of research on monaural source separation, which has been fruitful for speech signal processing in turn. In this section, we attempt a unified view on monaural audio source separation in speech and music, presenting a rough taxonomy of tasks and applications where synergies are evident. This taxonomy is oriented on the general procedure depicted in Figure 1, depending on which of the system components (source models, transcription/alignment, synthesis) are present.

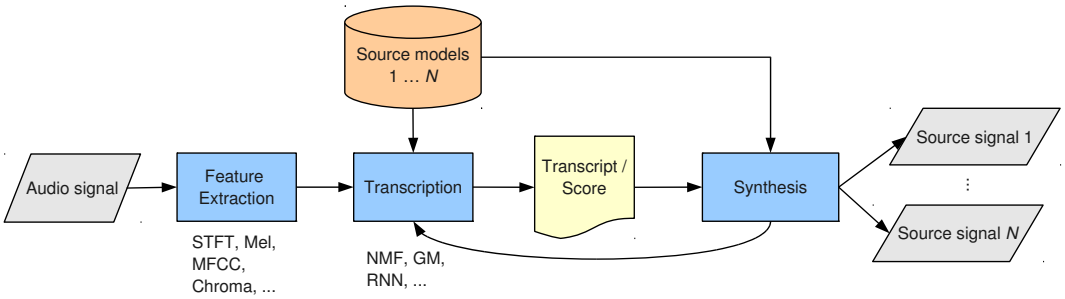

Figure 1 A unified view on monaural multi-source analysis of speech and music. Spectral (short-time Fourier Transform, STFT) or cepstral features (MFCCs) are extracted from the audio signal, yielding a transcription based on non-negative matrix factorization (NMF), graphical models (GM), recurrent neural networks (RNN) or other machine learning algorithms. The transcription can be used to synthesize signals corresponding to the sources or to enable (more robust) transcription in turn.

Polyphonic transcription and multi-source decoding

The goal of these tasks is not primarily the synthesis of each source as a waveform signal, but to gain a higher-level transcription of each source's contributions, e. g., the notes played by different instruments, or the transcription of the utterances by several speakers in a cross-talk scenario (the 'cocktail party problem'). Polyphonic transcription of monaural music signals can be achieved by sparse coding through non-negative matrix factorization (NMF) [64, 68], representing the spectrogram as the product of note spectra and a sparse non-negative activation matrix. These sparse NMF techniques have successfully been ported to the speech domain to reveal the phonetic content of utterances spoken in multi-source environments [18]: Determining the individual notes played by various instruments and their position in the spectrogram can be regarded as analogous to detecting individual phonemes in the presence of interfering talkers or environmental noise. An important common feature of these 'joint decoding' approaches for multi-source speech and music signals is the explicit modeling of parallel occurrence of sources; this can also be done by a graphical model representation of probabilistic dependencies between sources, as demonstrated in [69] for multi-talker ASR. Furthermore, polyphonic transcription approaches that use discriminative models for multiple note targets [46] or one-versus-all classification [50] seem to be partly inspired by 'multi-condition training' in ASR, where speech overlaid with interfering sources is presented to the system in the training stage, to learn to recognize speech in the presence of other sources. Finally, to contrast transcription or joint decoding approaches to the methods presented in the remainder of this section, we note that the former can principally be used to resynthesize signals corresponding to each of the sources [69], yet this is not their primary design goal; results are sometimes inferior to dedicated source separation approaches [19, 73].

Leading voice extraction and noise cancellation

For many MIR applications, the leading voice is of particular relevance, e. g., the voice of the singer in a karaoke application. Similarly, in many speech-based human-human and human-computer interaction scenarios, including automatic analysis of meetings, voice search or mobile telephony, the extraction of the primary speech source, which delivers the relevant content, is sufficient. This application requires modeling of the characteristics of the primary source, and speech and music processing considerably differ in this respect; unifying the

approaches will be an interesting question for future research. In music signal processing, main melody extraction is often related to predominance: It is assumed that the singing voice contributes the most to the signal energy[1]. Thus, extraction of the leading voice can be achieved with little explicit knowledge, e. g., by fixing a basis of sung notes and estimating the vocal tract impulse response in an extension of NMF to a source-filter model [14]. In speech processing, one usually does not rely on the assumption that the wanted speech is predominant in a recording, as signal-to-noise ratios can be negative in many realistic scenarios [9]. Hence, one extends the previous approaches by rather precise modeling of speech, often in a speaker-dependent scenario. Still, combining knowledge about the spectral characteristics of the speech with unsupervised estimation of the noise signal, in analogy to the unsupervised estimation of the accompaniment in [14], results in a semi-supervised approach for speech extraction as, e. g., in [48]. In contrast, often a pre-defined model for the background such as in [19, 53, 73] is used in a supervised source separation framework, and this kind of background modeling can be applied to leading voice extraction as well: Assuming the characteristics of the instrumental accompaniment of the singer are similar in vocal and non-vocal parts, a model of the accompaniment can be built; this allows estimating the contribution of the singing voice through semi-supervised NMF [21].

Instrument Separation and the Cocktail Party Problem

As laid out above, leading voice extraction or speech enhancement can be conceived as source separation problems with two sources. A generalization of this problem to extraction of multiple sources, or sources with large spectral similarity such as in instrument separation or the 'cocktail party' scenario, from a monophonic recording generally requires more complex source modeling. This can include temporal dependencies: In [45], NMF is extended to a non-negative Hidden Markov Model for extraction of the individual speakers from a multi-talker recording. Including temporal dependencies appears promising for music contexts as well, e. g., for separation of (repetitive) percussive and (non-repetitive) harmonic sources; furthermore, this approach is purely data-based and generalizes well to multiple sources.

In music signal processing, especially for classical music, higher-level knowledge can be incorporated into signal separation by means of score information (score-informed source separation) [15, 24]. Not only does this allow to cope with large spectral similarity, but it also enables separation by semantic aspects, which would be infeasible from an acoustic feature representation, and/or allows for user guidance; for instance, the passages played by the left and right hand in a piano recording can be retrieved [15]. Transferring this approach to the speech domain, we argue that while in most speech-related applications availability of a 'score' (i. e., a ground truth speaker diarization including overlap and transcription) cannot be assumed, score-informed separation techniques could be an inspiration to built iterative, self-improving methods for cross-talk separation, speech enhancement and ASR, recognizing what has been said by whom and exploiting that higher-level knowledge in the enhancement algorithm.

2.2 Combined Acoustic and Language Modeling

Language modeling techniques are found in MIR, e. g., to model chord progressions [47, 58, 80] or playlists [36]. Conversely, the prevalent usage of language models in ASR is

[1] Other common assumptions are that the singing voice is the highest voice among all instruments, or that it is characterized by vibrato.

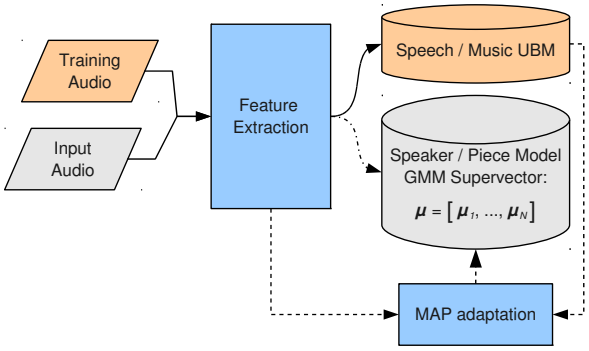

Figure 2 Use of universal background models (UBM) in speech and music processing: A generic speech/music model (UBM) is created from training audio. A speaker/piece model can be generated directly from training audio (dashed-dotted curve) or from the UBM by MAP adaptation (dashed lines). In the latter case, the parameters of the adapted model (e. g., the mean vector μ in case of GM modeling) yield a fingerprint (*supervector*) of the speaker or the music piece.

to calculate combined acoustic-linguistic likelihoods for speech decoding: Informally, the acoustic likelihood of a phoneme in an utterance is multiplied with a language model likelihood of possible words containing the phoneme to integrate knowledge about word usage frequencies (unigram probabilities) and temporal dependencies (n-grams) [82]. This immediately translates to chord recognition: For instance, unigram probabilities can model the fact that major and minor chords are most frequent in Western music, and there exist typical chord progressions that can be modeled by n-grams [56]. Thus, accuracy of chord recognition can be improved by combined acoustic and language modeling in analogy to ASR [8, 29]. A different approach to combined acoustic and language modeling is taken in [30] for genre classification: Music is encoded in a symbolic representation derived from clustered acoustic features, which is then encoded in a language model for different genres.

2.3 Universal Background Models in Speech Analysis and Music Retrieval

Recent developments in content-based music retrieval include methodologies that were introduced for speaker recognition and verification. These include universal background models (UBM)—trained from large amounts of data, and representing generic speech as opposed to the speech characteristics of an individual—and Gaussian Mixture Model (GMM) supervectors [4, 35, 81]. GMM supervectors are equivalent to the parameters of a Gaussian Mixture UBM adapted to the speech of a single speaker (usually only few utterances). Hence, they allow for effective and efficient computation of a person's speech 'fingerprint', i. e., its representation in a concise feature space suitable for a discriminative classifier. The generic approaches incorporating UBMs for speech and music classification are shown in Figure 2: A basic speaker verification algorithm uses a UBM to represent the acoustic parameters of a large set of speakers, while the speaker to be verified is modeled with a specialized GMM. For an utterance to be verified, a likelihood ratio test is conducted to determine whether the speaker model delivers sufficiently higher likelihood than the UBM. Translating this paradigm to music retrieval, one can cope with out-of-set events—i. e., that the user may be querying for a musical piece not contained in the database. Specific pieces in the database are represented ('fingerprinted') by Gaussian mixture modeling of acoustic features, while the UBM is a generic model of music. Then, the likelihoods of the query under the specialized GMMs versus the UBM allow out-of-set classification [39].

On the other hand, adapting the UBM to a specific music piece using maximum-a-posteriori (MAP) adaptation yields an audio fingerprint in shape of the adapted model's mean (and possibly variance) vectors. These fingerprints can be classified by discriminative models such as Support Vector Machines (SVMs), resulting in the GMM-SVM paradigm which has become standard in speaker recognition in the last years. In [5], the GMM-SVM approach was successfully applied to music tagging in the 2009 MIREX evaluation; recent studies [6, 7] underline the suitability of the approach to analyze music similarity for recommender systems.

2.4 Transfer from Paralinguistic Analysis

To elucidate a further opportunity for methodology transfer from the speech domain, we consider the field of paralinguistic analysis (i. e., retrieving other information from speech beyond the spoken text), which is believed to be important for natural human-machine and computer mediated human-human communication. Particularly, we address synergies between speech emotion recognition and music mood analysis: While relating to different concepts of emotion (or mood), the overlap in the methodologies and the research challenges are striking. At first, we would like to recall the subtle difference between those fields: Speech emotion recognition aims to determine the emotion of the speaker, which is—for most practical applications such as in dialog systems—the emotion perceived by the conversation partner; conversely, music mood analysis does not primarily assess the (perceived) mood of the singer, but rather the overall perceived mood in a musical piece—often, that is the intended mood, i. e., the mood as intended by the composer (or songwriter). Despite these differences, in the result, similar pattern recognition techniques have been proven useful in practice.

For instance, in order to assess the emotion of a speaker, combining 'what' is said with 'how' it is said, i. e., fusing acoustic with linguistic information, has been shown to increase robustness [78]—and similar results have been obtained in music mood analysis when considering lyrics and audio features [26, 57]. Apart from low-level acoustic and linguistic features, specific music features seem to contribute to music mood perception, and hence, recognition performance, including the harmonic 'language' (chord progression) and rhythmic structure [60], which necessitates efficient fusion methods as, e. g., for audio-visual emotion recognition. Besides, similarly to emotion in speech [77], music mood classification is lately often turned into a regression problem [60, 79] in target dimensions such as the arousal-valence plane [55], in order to avoid ambiguities in categorical 'tags' and improve model generalization.

Furthermore, when facing real-life applications, the issue of non-prototypical instances—i. e., musical pieces that are not pre-selected by experts as being representative for a certain mood—has to be addressed: It can be argued that a recommender system based on music mood should retrieve instances associated with high degrees of, e. g., happiness or relaxation from a large music archive. Here, music mood recognition can profit from the speech domain as this task bears some similarity to applications of speech emotion recognition such as anger detection, where emotional utterances have to be discriminated from a vast amount of neutral speech [66]. Relatedly, whenever instances to be annotated with the associated mood are not pre-selected by experts according to their prototypicality, the establishment of a robust ground truth, i. e., consistent assessment of the music mood by multiple human annotators, becomes non-trivial [27]. This might foster the development of quality control and 'noise cancellation' methods for subjective music mood ratings [60], as developed for speech emotion [20], in the future.

Finally, in the future, we might see a shift towards recognizing the affective state of singers themselves: First attempts have been made to estimate the 'enthusiasm' of the singer [10], which is arguably positively correlated with both arousal and valence; hence, the task is somewhat similar to recognition of level of interest from speech as in [78]. Another promising research direction might be to investigate long-term singer traits instead of short-term states such as emotion: Such traits include age, gender [59], body shape and race, all of which are known to be correlated with acoustic parameters, and can be useful in category-based music retrieval or identifying artists from a meta-database [74]. In a similar vein, the analysis of voice quality and 'likability' [72] could be a valuable source of inspiration for research on synthesis of singing voices.

3 From Music IR to Speech IR: An Example

Starting from the general overview above, we now discuss a particular example on how technologies from both domains of music and speech IR interact with each other. In particular, we start with the well known MFCC (Mel Frequency Cepstral Coefficients) features from the speech domain which are used to analyze signals based on an auditory filterbank. This results in representing a speech signal by a temporal feature sequence correlating with certain properties of the speech signal. We then review corresponding music features and their properties, with a particular interest on representing the harmonic progression of a piece of music using chroma-type features. This, in turn, inspires a class of speech features correlating with the phonetic progression of speech.

Concerning possible applications, chroma-type features can be used to identify fragments of audio as being part of a musical work regardless of the particular interpretation. Having sketched a suitable matching technique, we subsequently show how similar techniques can be applied in the speech domain for the task of keyphrase spotting.

Whereas the latter matching techniques focus on local temporal regions of audio, more global properties can be analyzed using self-similarity matrices. In music, such matrices can be used to derive the general repetitive structure (related to the musical form) of an audio recording. When dealing with two different interpretations of a piece of music, such matrices can be used to derive a temporal alignment between the two versions. We discuss possible analogies in speech processing and sketch an alternative approach to text-to-speech alignment.

3.1 Feature Extraction

Many audio features are based on analyzing the spectral contents of subsequent short temporal segments of a target signal by using either a Fourier transform or a filter-bank. The resulting sequence of vectors is then further processed depending on the application. As an example, the popular MFCC features which have been successfully applied in automatic speech recognition (ASR) are obtained by applying an auditory filterbank based on log-scale center frequencies, followed by converting subband energies to a dB- (log-) scale, and applying a discrete cosine transform [51]. The logarithmic compression in both frequency and signal power serves to weight the importance of events in both domains in a way a human perceives them. Because of their ability to describe a short-time spectral envelope of an audio signal in a compact form, MFCCs have been successfully applied to various speech processing problems apart from ASR, such as keyword spotting and speaker recognition [54]. Also in Music IR, MFCCs have been widely used, e. g., for representing the timbre of musical instruments or speech-music discrimination [34].

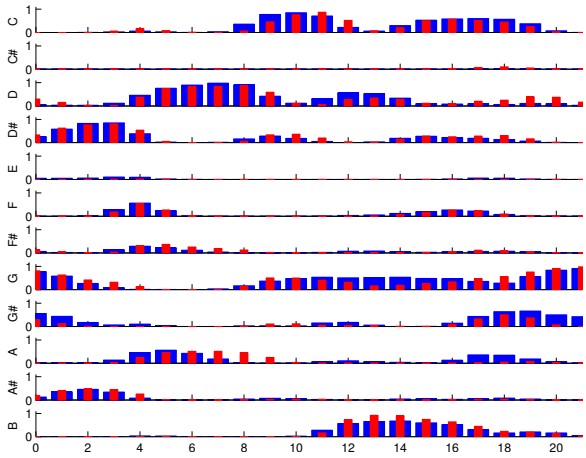

Figure 3 Chroma-based CENS features obtained from the first measures (20 seconds) of Beethoven's 5th Symphony in two interpretations by Bernstein (blue) and Sawallisch (red).

While MFCCs are mainly motivated by auditory perception, music analysis is frequently performed based on features motivated by the process of sound generation. Chroma features for example, which have received an increasing amount of attention during the last ten years [2], rely on the fixed frequency (semitone) scale as used in Western music. To obtain a chroma feature for a short segment of audio, a Fourier transform of that segment is performed. Subsequently, the spectral coefficients corresponding to each of the twelve musical pitch classes (the *chroma*) C, C$^\sharp$, D,..., B are individually summed up to yield a 12-dimensional chroma vector. In terms of a filterbank, this process can be seen as applying octave-spaced comb-filters for each chroma.

From their construction, chroma features do well-represent the local harmonic content of a segment of music. To describe the temporal harmonic progression of a piece of music, it is beneficial to combine sequences of successive chroma features to form a new feature type. CENS-features (chroma energy normalized statistics) [43] follow this approach and involve calculating certain short-time statistics on the chroma features' behaviour in time, frequency, and energy. By adjusting the temporal size of the statistics window, CENS-feature sequences of different temporal resolutions may be derived from an input signal. Figure 3 shows the resulting CENS feature sequences derived from two performances of Beethoven's 5th Symphony.

In the speech domain, a possible analogy to the local harmonic progression of a piece of music is the phonetic progression of a spoken sequence of words (a *phrase*). To model such phonetic progressions, the concept of energy normalized statistics (ENS) has been transferred to speech features [70]. This approach uses a modified version of MFCCs, called HFCCs (human factor cepstral coefficients), where the widths of the mel-spaced filter bands are chosen according to the bark scale of critical bands. After applying the above statistics computations, the resuling features are called HFCC-ENS. Figure 6 (c) and (d) show sequences of HFCC-ENS features for two spoken versions of the same phrase. Experiments show that due to the process of calculating statistics, HFCC-ENS features are better adapted to the phonetic progression in speech than MFCCs [70].

3.2 Matching Techniques

In this section, we describe some matching techniques that use audio features in order to automatically recognize audio signals. Current approaches to ASR or keyword spotting employ suitable HMMs trained to individual words (or subword entities) to be recognized. Usually, speaker-dependent training results in a significant improvement in recognition rates and accuracy. Older approaches used dynamic time warping (DTW) which is simpler to implement and bears the advantage of not requiring prior training. However, as the flexibility of DTW in modeling speech properties is restricted, it is not as widely used in applications as HMMs are [52]. In the context of music retrieval, DTW and variants thereof have, however, regained considerable attention [40].

As particular example, we consider the task of audio matching: Given a short fragment of a piece of audio, the goal is to identify the underlying musical work. A refined task would be to additionally determine the position of the given fragment within the musical work. This task can be cast into a database search: given a short audio fragment (the *query*) and a collection of "known" pieces of music (the *database*), determine the piece in the database the query is contained in (the *match*). Here a restricted task, widely known as *audio identification*, only reports a match if the query and a match correspond to the same audio recording [1, 71]. In general audio matching, however, a match is also reported if a query and the database recording are different performances of the same piece of music. Whereas audio identification can be very efficiently performed using low-level features describing the physical waveform, audio matching has to use more abstract features in order to identify different interpretations of the same musical work. In Western classical music, different interpretations can exhibit significant differences, e. g., regarding tempo and instrumentation. In popular music, different interpretations include cover songs that may exhibit changes in musical style as well as mixing with other audio sources [62].

The introduced CENS features are particularly suitable to perform audio matching for music that possess characteristic harmonic progressions. In a basic approach [43], the query and database signals are converted to feature sequences $q = (q_1, \ldots, q_M)$ and $d = (d_1, \ldots, d_N)$, where each of the q_i and d_j are 12-dimensional CENS vectors. Matching is then performed using a cross-correlation like approach, where a similarity function $\Delta(n) := \frac{1}{M} \sum_{\ell=1}^{M} \langle q_\ell, d_{n-1+\ell} \rangle$ gives the similarity of query and database at position n. Using normalized feature vectors, values of Δ in a range of $[0, 1]$ can be enforced. Figure 4 (top) shows an example of a resulting Δ when using the first 20 seconds of the Bernstein interpretation (see Figure 3) as a query to a database containing, among other material, two different versions of Beethovens Fifth by Bernstein and Sawallisch respectively. Positions corresponding to the seven best matches are indicated in green. The first six matches correspond to the three occurrences of the query (corresponding to the famous theme) within the two performances. Tolerance with respect to different global tempi may be obtained in two ways: On the one hand, one may calculate p time-scaled versions of the feature sequence q by simply changing the statistics parameters (particularly window size and sampling rate) during extraction of the CENS features. This process is then followed by p different evaluations of Δ. On the other hand, the correlation-based approach to calculate a cost function may be replaced by a variant of subsequence DTW. Experiments show that both variants perform comparably.

Coming back to the speech domain, the some audio matching approach can be applied to detect short sequences of words or *phrases* within a speech recording. Compared to classical keyword spotting [28, 76], this kind of *keyphrase* spotting is particularly beneficial when the target phrase consists of at least 3-4 words [70]. Advantages inherited from using the above HFCC-ENS features for this task are speaker and also gender independence. More important,

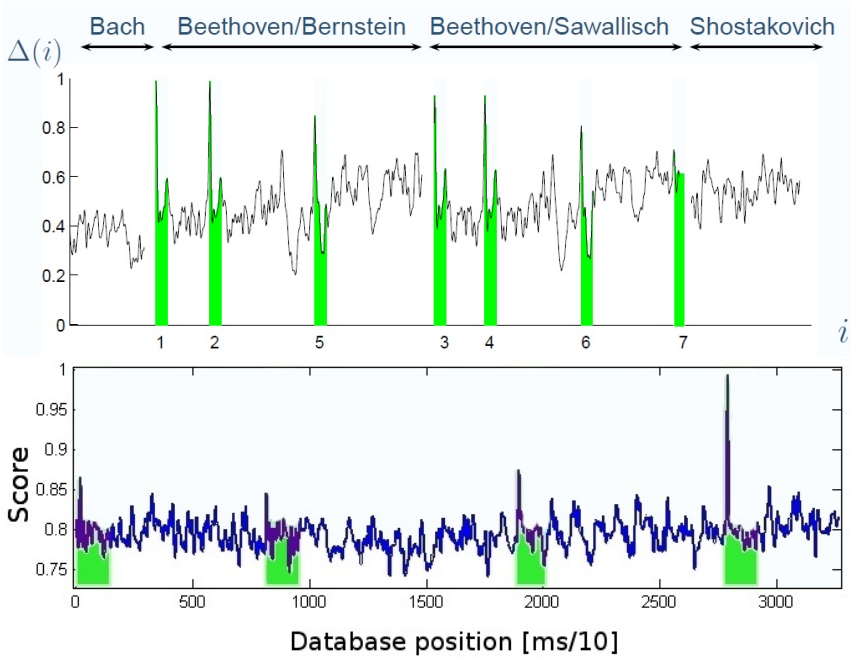

Figure 4 Top: Similarity function Δ obtained in scenarios of audio matching for music. Bottom: Similarity function Δ obtained in keyphrase matching.

no prior training is required which makes this form of keyphrase spotting attractive for scenarios with sparse resources. Figure 4 (bottom) shows an example where the German phrase *"Heute ist schönes Frühlingswetter"* was used as a query to a database containing a total of 40 phrases spoken by different speakers. Among those are four versions of the query phrase each by a different speaker. All of them are identified as matches (indicated in green) by applying a suitable peak picking strategy on the similarity function.

3.3 Similarity Matrices: Synchronization and Structure Extraction

To obtain the similarity of a query q and a particular position of a database document d, a similarity function Δ has been constructed by averaging M local comparisons $\langle q_i, d_j \rangle$ of features vectors q_i and d_j. In general, the similarity between two feature sequences $a = (a_1, \ldots, a_K)$ and $b = (b_1, \ldots, b_L)$ can be characterized by calculating a *similarity matrix* $S_{a,b} := (\langle a_i, b_j \rangle)_{1 \leq i \leq K, 1 \leq j \leq L}$ consisting of all pair-wise comparisons. Figure 5 (left) shows an example of a similarity matrix. Color coding is chosen in a way such that dark regions indicate a high local similarity and light regions correspond to a low local similarity. The diagonal-like trajectory running from the lower left to the upper right thus expresses the difference in the local tempo between the two underlying performances.

Based on such trajectories, similarity matrices can be used to temporally *synchronize* musically corresponding positions of the two different interpretations [25, 44]. Technically, this amounts to finding a *warping path* $p := (x_i, y_i)_{i=1}^{P}$ through the matrix, such that $\delta(p) := \sum_{i=1}^{P} \langle a_{x_i}, b_{y_i} \rangle$ is minimized. Warping paths are restricted to start in the lower left corner, $(x_1, y_1) = (1, 1)$, end in the upper right, $(x_P, y_P) = (K, L)$, and obey certain step conditions, $(x_{i+1}, y_{i+1}) = (x_i, y_i) + \sigma$. Two frequently used step conditions are $\sigma \in \{(0, 1), (1, 0), (1, 1)\}$ and $\sigma \in \{(2, 1), (1, 2), (1, 1)\}$. In Figure 5 (left) a calculated warping path is indicated in red color.

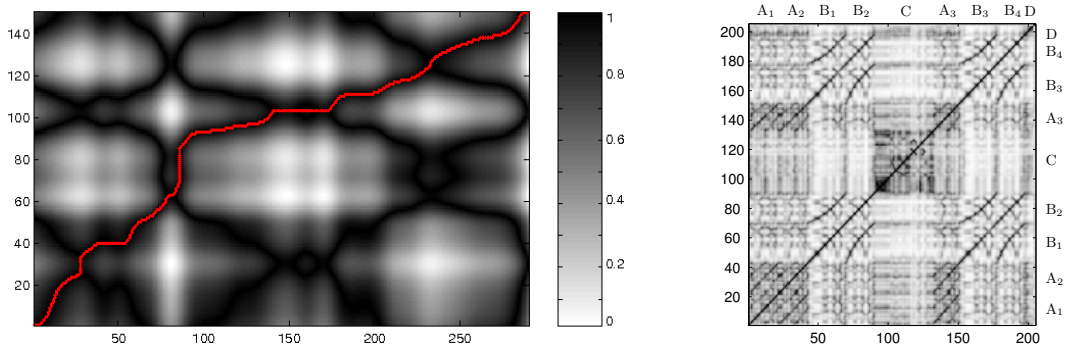

Figure 5 Left: Example of a similarity matrix with warping path indicated in red color. **Right:** Self-similarity matrix for a version of Brahms Hungarian Dances no. 5. The extracted musical structure $A_1A_2B_2CA_3B_3B_4D$ is indicated. (Figures from [40].)

Besides synchronizing two audio recordings of the same piece, the latter methods can be used to time-align musically corresponding events *across* different representations. As a first example, consider a (symbolic) MIDI representations of the piece of music. In a straightforward approach, an audio version of the MIDI can be created using a synthesizer. Then, CENS features are obtained from the synthesized signal, thus allowing a subsequent synchronization with another audio recording (in this context an audio recording obtained from a real performance). Alternatively, CENS features may be generated *directly* from the MIDI [25]. In a second example, scanned sheets of music (i. e., digital images) can be synchronized to audio recordings, by first performing optical music recognition (OMR) on the scanned images, producing a symbolic, MIDI-like, representation. In a second step, the symbolic representation is then synchronized to the audio recording as described before [16]. This process is illustrated in Figure 6 (left). Besides the illustrated task of audio synchronization, the automatic alignment of audio and lyrics has also been studied [37], suggesting the usability of synchronization techniques for human speech.

Transfered to the speech domain, such synchronization techniques can be used to time-align speech signals with a corresponding textual transcript. Similarly to using a music synthesizer on MIDI input to generate a music signal, a text-to-speech (TTS) system can be used to create a speech signal. Subsequently, DTW-based synchronization can be performed on HFCC-ENS feature sequences extracted from both speech signals [11], see Figure 6 (right).

Text-to-speech synchronization as decribed here may be applied for example to political speeches or audio books. We note that a more classical way of performing this synchronization consists of first performing ASR on the speech signal, resulting in an approximate textual transcript. In a second step, both transcripts can then by synchronized by suitable text-based DTW techniques [23].

ASR-based synchronization is advantageous in case of relatively good speech quality or when a prior training to the speaker is possible. In this case, the textual transcript will be of sufficiently high quality and a precise synchronization is possible. Due to the smoothing process involved in the ENS calculation, TTS-based synchronization typically has a lower temporal resolution which has an impact on the synchronization accuracy. However, in scenarios with a high likelihood of ASR-errors, TTS-based synchronization can be beneficial.

Variants of the DTW-based music synchronization perform well if the musical structure underlying a and b are the same. In case of structural differences, advanced synchronization methods have to be used [41]. To analyze the structure of a music signal, the *self-similarity matrix* $S_a := S_{a,a}$ of the corresponding feature sequence a can be employed. As an example,

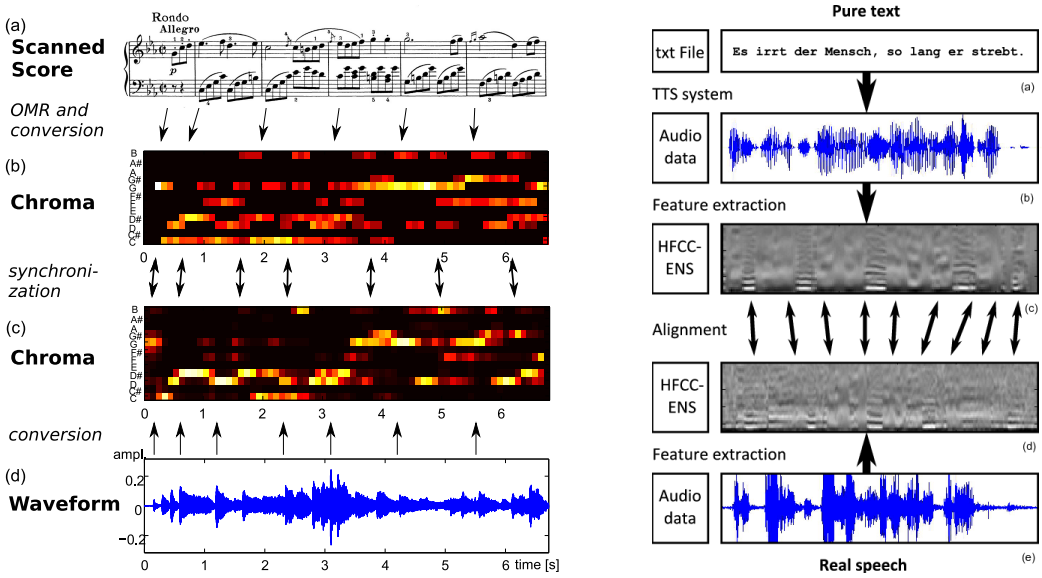

Figure 6 Left: Score-Sheet to audio synchronization—(a) Score fragment, (b) Synthesized Chroma features, (c) Chroma obtained from audio recording (d). **Right:** Text to audio synchronization—(a) Text, (b) Synthesized speech, (c) HFCC-ENS features of synthesized speech, (d) HFCC-ENS features of natural speech (e).

Figure 5 depicts the self-similarity matrix of an interpretation of Brahms Hungarian Dances no. 5 by Ormandy. Darker trajectories on the side diagonals indicate repeating music passages. Extraction of all such repetitions and systematic structuring can be used to deduce the underlying musical form. In our example, the musical form $A_1A_2B_2CA_3B_3B_4D$ is obtained by following an approach to calculate a complete list of all repetitions [42].

Concluding, we discuss possible applications of structure analysis in the speech domain, where one first has to ask for suitable analogies of *structured speech*. In contrast to music analysis, where the target signal to be analyzed frequently corresponds to a complete piece of music, in speech one frequently analyses unstructured speech fragments such as isolated sequences of sentences or a dialog between two persons. Lower-level examples of speech structure relevant for unstructured speech could be repeated words, phrases, or sentences. More structure on a higher level could be expected from speech recorded in special contexts such as TV shows, news, phone calls, or radio communication. An even closer analogy to music analysis could be the analysis of recited poetry.

4 Evaluation: The Information Retrieval Legacy

We now move on to another field with considerable influences on MIR research: Information Retrieval (IR). This field, after which the MIR field was named, deals with storing, extracting and retrieving information from text documents. The information can be both syntactic and semantic, and topics of interest cover a wide range, involving feature representations, full database systems, and information-seeking behavior of users.

Evaluation in MIR work, especially in retrieval settings, has largely been influenced by IR evaluation, with *Precision*, *Recall* and the *F-measure* as most stereotypical evaluation criteria. However, already in the first years of the MIR community benchmark evaluation endeavor, the Music Information Retrieval EXchange (MIREX), the need arose to find

significance levels for system results. Earlier findings from the Text REtrieval Conference (TREC) benchmarking efforts led to the adoption of Friedman's ANOVA with Tukey-Kramer "Honestly Significant Difference" post-hoc correction [13], which subsequently were widely adopted in the presentation of MIREX results.

Not all of the IR practices were immediately transferable to MIR evaluation: many MIREX tasks turned out to be specialized enough to a degree that they require task-specific evaluation criteria. In addition, precision and recall have frequently been challenged for their appropriateness. In cover song retrieval and audio matching settings, recall may be the most appropriate, since the goal would be to retrieve as many matching items or fragments as possible [61]. On the other hand, in web-scale environments, the amount of data will be so huge that striving for recall will not make sense anymore. In addition, in multimedia settings one can wonder if precision would be an appropriate measure at all, since user data suggests that multimedia search is more of an entertaining browsing activity rather than a focused information need with a concrete query and an establishable ground truth [63]. Exactly the same will hold for music search.

Nonetheless, there still are existing IR evaluation findings that provide useful opportunities for strengthening evaluation in MIR, an important area being that of *meta-evaluation* [67]. Through meta-evaluation, the experimental validity of (M)IR experiments can be assessed. This validity can be assessed according to different subcategories, which are listed below together with reflections on the way in which they are applicable to the MIR domain:

Construct validity

The extent to which the variables of an experiment correspond to the theoretical meaning of the concept they are intended to measure. To give an example for MIR, it is tempting to try to infer music 'mood' from features present in musical audio (e.g. presence of major/minor chords and tonalities); however, the situation is often more complicated. Most importantly, mood implies a human property, and is usually experienced due to a certain (multimodal) context. Thus, in order to truly address mood, work related to music and mood should not only look at audio features and take the user and this context into account.

Content validity

The extent to which the experimental units reflect and represent the elements of the domain under study. For example, an experiment aimed at measuring 'audio similarity' between songs cannot be (solely) based on item co-occurrences of these songs in a social network.

Convergent validity

The extent to which the results of an experiment agree with other results they should be related with (both theoretical and experimental). As an example from the MIR domain, a good tempo estimator should involve a good beat estimating component. Thus, this beat estimating component would be expected to perform well on beat extraction tasks.

Criterion validity

The extent to which the results of an experiment are correlated with those of other experiments already known to be valid. In the case of e.g. relevance assessments, if results from crowdsourced ground truth turn out to correlate well with results from earlier expert-established ground truth, the suitability of the corresponding crowdsourcing platform as a

scalable and less time-consuming ground truthing platform is strenghtened. An investigation like this has e.g. been done in [31] for the MIREX Audio Music Similarity and Retrieval task.

Internal validity

The extent to which the conclusions of an experiment can be rigorously drawn from the experimental design followed, and not from other factors unaccounted for. An optimal combination of musical attributes (e.g. good voice, catchy tune) will only partially explain high sales numbers for an artist; next to this, contextual aspects (such as recent high-profile appearances) will also play a role.

External validity

The extent to which the results of an experiment can be generalized to other populations and experimental settings. Of all the validity types mentioned here, issues with external validity may be the most concretely recognized in the MIR community at this moment. For example, many mid-level feature representations and assumptions in the MIR field have been modeled for Western popular music, but turn out not to be a good fit for other types of music: e.g. many classical music pieces do not have a constant tempo or steady beat, and an equal-tempered 12-tone chroma representation is not very well suited to capture the traditional music of other cultures.

Conclusion validity

The extent to which the conclusions drawn from the results of an experiment are justified. A notorious example is the claim that successful published work 'closed or bridged the *semantic gap*' (which will be discussed in more detail in the following section) — while indeed, low-level features often do not match high-level concepts, cases in which a better correspondence between these two levels is found frequently deal with domain-specific cases, and do not address any fundamental and generalized 'understanding' problems that a 'semantic gap' would imply. In addition, the whole metaphor of a semantic gap may not be appropriate; this will be addressed in the following section as well.

As we showed, meta-evaluation principles can readily be applied to many realistic MIR cases. By applying meta-evaluation principles, more insight can be gained into the scientific solidness of evaluation results, and because of this, the true intricacies of proposed systems will become clearer. This is very useful, since music data often is intangible data that is difficult to be understood, as we will discuss in the following section.

5 Opportunities for MIR: Universal Open Challenges

So far, we discussed transfer opportunities for two domains that are closely connected to the field of MIR. In this section, we will zoom out and take a higher-level perspective on open issues in the MIR field, and demonstrate that these are very similar to open fundamental issues as identified in the Content-Based Image Retrieval (CBIR) and Multimedia Information Retrieval (MMIR) communities, suggesting bridging opportunities for these fields and MIR.

5.1 The Nature of Music Data is Multifaceted and Intangible

Music is a peculiar data type. While it has communicative properties, it is not a natural language with referential semantics that indicate physically tangible objects in the world. One can argue that lyrics can contain such information, but these will not constitute music when considered in isolation.

The typical main representation of music is usually assumed to be audio or symbolic score notation. However, even such a representation in itself will not embody music as a whole, but rather should be considered a 'projection' of a musical object [75]. The composer Milton Babbitt proposed to categorize different music representations in three domains: 1) the *acoustic* or physical domain, (2) the *auditory* or perceived domain, and (3) the *graphemic* or notated domain. In [75], different transformations between these domains are mentioned: for example, a *transcription* will transform a mental image of music in the auditory domain to a notated representation in the graphemic domain, while a *performance* will transform the same mental image into an acoustic domain representation. The interplay between the three domains, in the presence of a human spectator, will establish experiences of the musical object, but that musical object itself remains an intangible, abstract concept.

Due to the multifaceted nature of music, and the strong dependence of experiences of music on largely black-boxed processes in the human auditory domain with strongly affective reactions, it is a very hard data type to grasp from a fundamental point of view. In an increasing amount of Music-IR tasks, we are typically not interested in precise (symbolic or digital) music encoding, nor in its sound wave dispersion behavior, but exactly in this difficult area of the effect music has on human beings, or the way humans interact with music. This poses challenges to the evaluation of automated methods: a universal, uncompromising and objective ground truth is often nonexistent, and if it is there, there still are no obvious one-to-one mappings between signal aspects and perceived musical aspects. The best ground truth one can get is literally grounded: established from empirical observations and somehow agreed upon by multiple individuals.

Issues with nonexistent ground truth, multifaceted representations and subjective and affective human responses are not new at all. In fact, they have been frequently mentioned in the CBIR and MMIR communities — although no clear and satisfying solution to them has been found yet.

5.2 Open Challenges are Shared Across Domains

In 2000 (incidentally, the year in which the first ISMIR conference was held), a seminal review [65] on content-based image retrieval (CBIR) was published, touching upon the state-of-the-art and outlining future directions. In this review, several trends and open issues were mentioned by the authors. It is striking to see how natural the following phrases read if transferred from the image to music processing domain, substituting 'CBIR' with 'MIR' and 'computer vision' with 'signal processing':

- The wide availability of digital sensors, the Internet, and the falling price of storage devices were considered as the *driving forces* for rapid developments in CBIR. However, more precise foundations would be desired, indicating what problem exactly is to be solved, and whether proposed methods would perform better than alternatives. A call was made for classification of usage-types, aims and purposes for the man-machine interface, domain knowledge, and database technology alike.
- *The heritage of computer vision*, from which CBIR developed, was considered to be an obstacle. CBIR is stronger about solving a general 'image understanding' problem and

evaluating results in terms of a user-defined ground truth than about providing algorithms with 100% segmentation accuracy according to a fully objective measure, which would be more typical of fundamental computer vision. Thus, in certain cases, goals could not exactly be taken over between these two related domains.

- Different goals and requirements in CBIR actually had *influence on computer vision* and (re)kindled interest in larger, dedicated datasets, weak segmentation and saliency, color image processing, and attention for invariance.
- It was argued that the notion of *similarity* should be considered from a human perspective. In addition, *learning* would be necessary to extend knowledge from partially labeled data to larger datasets.
- *Interaction* was mentioned as a major difference between CBIR and computer vision. Interaction and feedback mechanisms have been explored for a longer time in IR, but there are some fundamental differences between the two retrieval areas, especially in terms of query vs. result modalities. Visualization, so a move towards multimodal interfaces, was suggested as an important means to deal with this.
- Larger amounts of data increase the need for solid underlying *database* technology. Database research and CBIR traditionally have been separate fields, but were suggested to work together in this.
- *Evaluation* is a major issue. Results can be biased towards dataset composition, and it is hard to assess the 'difficulty' of a dataset. A call was made for reference standards such as TREC in Text IR. Furthermore, it was suggested to borrow concepts from the fields of psychological and social sciences.
- This review became particularly famous for coining the term *semantic gap* to indicate the mismatch between signal representations and analyses and the human assessments of their success. The authors wrote about resolving the gap by including *additional sources* of information. Here, insights from natural language processing and computer vision could be beneficial.

Many of these points still have largely remained unsolved. Eight years later, a survey in [12] still mentions user-focused (benchmark) evaluation as a future design goal, and application-oriented, domain-specific solutions as necessary ways to go in order to serve real-world needs.

With an increased interest in video data and multimodal approaches, part of the CBIR field merged into the MMIR field, where once again similar fundamental questions are mentioned. In [32], human-centered methods, multimedia-supported user-to-user collaboration, interactive search and agent interfaces, neuroscience and new learning models and folksonomies are pointed out as open future directions to study. The 'Holy Grail of Multimedia Information Retrieval', *getting the access to the content we like quickly and easily whenever we like it and wherever we are* [22], has not been found yet.

It is very striking to consider the open challenges mentioned above alongside the open challenges as identified at the occasion of the 10th anniversary of the ISMIR conference:

- Increased involvement of real end-users;
- Deeper understanding of the music data and employment of musically motivated approaches;
- Perspective broadening beyond 20th century Western popular music;
- The investigation of musical information outside of the audio domain;
- The creation of full-featured, multifaceted, robust and scalable Music-IR systems with helpful user interfaces.

In all cases, we identify a need for increased user involvement and interaction, understanding of the data while avoiding dataset bias, and the inclusion of multiple available information sources as main open challenges to pay attention to. Actually, even in the well-established IR field, involving the user is no common practice yet [3].

In both MMIR and MIR, it already has been hypothesized [49, 75] that a true semantic gap that can be 'crossed' through rigid algorithmic approaches is an unrealistic metaphor, and that human and cognitive approaches are necessary in any solution that is to be successful. The intangible and abstract nature of music data has strong potential to urgently push research into user-centered and multimodal approaches going towards this direction [33]. Thus, also in this area, we see opportunities, and even a potential flagship role, for MIR work to become of inspirational value to work in neighboring domains.

6 Conclusions

In this chapter, we discussed several methodology transfer opportunities for MIR. We first gave examples of MIR analogues to existing ASR tasks and discussed how MIR findings have benefited ASR the other way around. Subsequently, we mentioned current and promising influences from IR to MIR. Finally, we compared fundamental open challenges within MIR to those that have been mentioned, but never satisfyingly solved yet, in the CBIR and MMIR fields. Here, we argued that music data can be the key to finally address these challenges.

It is our intention that this chapter can serve as an inspirational guide, especially to researchers that are situated on the interfaces between different domains. We hope that increased bridge-building and knowledge exchanging between the domains will be capable of pushing research within these domains beyond limits and boundaries encountered so far.

References

1 E. Allamanche, J. Herre, B. Fröba, and M. Cremer. AudioID: Towards Content-Based Identification of Audio Material. In *Proc. 110th AES Convention, Amsterdam, NL*, 2001.

2 M. A. Bartsch and G. H. Wakefield. Audio Thumbnailing of Popular Music Using Chroma-based Representations. *IEEE Trans. on Multimedia*, 7(1):96–104, Feb. 2005.

3 N. J. Belkin. Some(what) Grand Challenges for Information Retrieval. *SIGIR Forum*, 42(1), June 2008.

4 W. Campbell, D. E. Sturim, D. Reynolds, and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Proc. of ICASSP*, pages 97–100, 2006.

5 C. Cao and M. Li. ThinkIT Submissions for MIREX2009 Audio Music Classification and Similarity Tasks. In *Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.

6 C. Charbuillet, D. Tardieu, and G. Peeters. GMM Supervector for Content Based Music Similarity. In *Proc. of DAFx*, pages 1–4, 2011.

7 Z.-S. Chen, J.-S. Jang, and C.-H. Lee. A kernel framework for content-based artist recommendation system in music. *IEEE Transactions on Multimedia*, 2011. to appear.

8 H.-T. Cheng, Y.-H. Yang, Y.-C. Lin, I.-B. Liao, and H. Chen. Automatic chord recognition for music classification and retrieval. In *Proc. of ICME*, pages 1505–1508, 2008.

9 H. Christensen, J. Barker, N. Ma, and P. Green. The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments. In *Proc. of Interspeech*, pages 1918–1921, Makuhari, Japan, 2010.

10 R. Daido, S.-J. Hahm, M. Ito, S. Makino, and A. Ito. A System for Evaluating Singing Enthusiasm for Karaoke. In *Proc. of ISMIR*, pages 31–36, Miami, FL, USA, 2011.

11 D. Damm, H. Grohganz, F. Kurth, S. Ewert, and M. Clausen. SyncTS: Automatic synchronization of speech and text documents. In *Proceedings of the AES 42nd International Conference Semantic Audio*, Ilmenau, Germany, 2011.

12 R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys*, 40(2), 2008.

13 J. S. Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoust. Sci. & Tech.*, 29(4):247–255, 2008.

14 J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564–575, 2010.

15 S. Ewert and M. Müller. Score-Informed Voice Separation For Piano Recordings. In *Proc. of ISMIR*, pages 245–250, Miami, FL, USA, 2011.

16 C. Fremerey, M. Müller, F. Kurth, and M. Clausen. Automatic mapping of scanned sheet music to audio recordings. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, September 2008.

17 H. Fujihara, M. Goto, and J. Ogata. Linking Lyrics: A Method for Creating Hyperlinks Between Phrases in Song Lyrics. In *Proc. of ISMIR*, pages 281–286, 2008.

18 J. Gemmeke, T. Virtanen, and A. Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2067–2080, 2011.

19 J. F. Gemmeke, T. Virtanen, and A. Hurmalainen. Exemplar-Based Speech Enhancement and its Application to Noise-Robust Automatic Speech Recognition. In *Proc. of CHiME Workshop*, pages 53–57, Florence, Italy, 2011.

20 M. Grimm and K. Kroschel. Evaluation of natural emotions using self assessment manikins. In *Proc. of ASRU*, pages 381–385. IEEE, 2005.

21 J. Han and C.-W. Chen. Improving melody extraction using probabilistic latent component analysis. In *Proc. of ICASSP*, pages 33–36, 2011.

22 A. Hanjalic, R. Lienhart, W.-Y. Ma, and J. R. Smith. The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away? *Proc. IEEE*, 96(4):541–547, April 2008.

23 A. Haubold and J. R. Kender. Alignment of speech to highly imperfect text transcriptions. In *ICME*, pages 224–227, 2007.

24 R. Hennequin, B. David, and R. Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Proc. of ICASSP*, pages 45–48, Prague, Czech Republic, 2011.

25 N. Hu, R. B. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *in Proc. IEEE WASPAA*, pages 185–188, 2003.

26 X. Hu and J. S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proc. Joint Conference on Digital Libraries (JCDL)*, pages 159–168, Gold Coast, Queensland, Australia, 2010.

27 X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann. The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In *Proc. of ISMIR*, pages 462–467, Philadelphia, USA, 2008.

28 J. Keshet, D. Grangier, and S. Bengio. Discriminative keyword spotting. *Speech Communication*, 51:317–329, 2009.

29 M. Khadkevich and M. Omologo. Use of hidden Markov models and factored language models for automatic chord recognition. In *Proc. of ISMIR*, pages 561–566, 2009.

30 T. Langlois and G. Marques. Automatic Music Genre Classification Using a Hierarchical Clustering and a Language Model Approach. In *Proc. of First International Conference on Advances in Multimedia*, pages 188–193, 2009.

31 J. H. Lee. Crowdsourcing Music Similarity Judgments using Mechanical Turk. In *Proc. ISMIR*, pages 183–188, August 2010.

32 M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-Based Multimedia Information Retrieval: State of the Art and Challenges. *ACM Trans. Multimedia Computing, Communications and Applications*, 2(1):1–19, 2006.

33 C. C. S. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic. The Need for Music Information Retrieval with User-Centered and Multimodal Strategies. In *Proc. 1st Int. ACM workshop on MIR with User-Centered and Multimodal Strategies (MIRUM)*, pages 1–6, November 2011.

34 B. Logan. Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*, 2000.

35 B. Ma and H. Li. Text-Independent Speaker Recognition. In H. Li, K.-A. Toh, and L. Li, editors, *Advanced Topics in Biometrics*. World Scientific Publishing Co., 2011.

36 B. McFee and G. Lanckriet. The Natural Language of Playlists. In *Proc. of ISMIR*, pages 537–542, Miami, FL, USA, 2011.

37 A. Mesaros and T. Virtanen. Automatic alignment of music audio and lyrics. In *DAFX08*, 2008.

38 A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009. Article ID 546047.

39 M. Mohri, P. Moreno, and E. Weinstein. Robust Music Identification, Detection, and Analysis. In *Proc. of ISMIR*, Vienna, Austria, 2007.

40 M. Müller. *Information Retrieval for Music and Motion*. Springer Verlag, 2007.

41 M. Müller and D. Appelt. Path-constrained partial music synchronization. In *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, 2008.

42 M. Müller and F. Kurth. Towards Structural Analysis of Audio Recordings in the Presence of Musical Variations. *EURASIP Journal on Applied Signal Processing*, 2007(Article ID 89686):18 pages, January 2007.

43 M. Müller, F. Kurth, and M. Clausen. Audio matching via chroma-based statistical features. In *Proc. ISMIR, London, GB*, 2005.

44 M. Müller, H. Mattes, and F. Kurth. An Efficient Multiscale Approach to Audio Synchronization, 2006.

45 G. J. Mysore and P. Smaragdis. A Non-Negative Approach to Semi-Supervised Separation of Speech from Noise with the Use of Temporal Dynamics. In *Proc. of ICASSP*, pages 17–20, Prague, Czech Republic, 2011.

46 J. Nam, J. Ngiam, H. Lee, and M. Slaney. A Classification-Based Polyphonic Piano Transcription Approach Using Learned Feature Representations. In *Proc. of ISMIR*, pages 175–180, Miami, FL, USA, 2011.

47 M. Ogihara and T. Li. N-gram chord profiles for composer style representation. In *Proc. of ISMIR*, pages 671–676, 2008.

48 A. Ozerov, C. Févotte, and M. Charbit. Factorial scaled hidden Markov model for polyphonic audio representation and source separation. In *Proc. of WASPAA*, pages 121–124, Mohonk, NY, United States, 2009.

49 T. Pavlidis. The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away? - An Answer. `http://www.theopavlidis.com/technology/CBIR/summaryB.htm`, accessed September 2011, 2008.

50 G. Poliner and D. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

51 L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, united states ed edition, Apr. 1993.

52 L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

53 B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh. Non-negative matrix factorization based compensation of music for automatic speech recognition. In *Proc. of Interspeech*, pages 717–720, Makuhari, Japan, 2010.

54 D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3(1):72–83, January 1995.

55 J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

56 R. Scholz, E. Vincent, and F. Bimbot. Robust modeling of musical chord sequences using probabilistic N-grams. In *Proc. of ICASSP*, pages 53–56, 2009.

57 B. Schuller, C. Hage, D. Schuller, and G. Rigoll. "Mister D.J., Cheer Me Up!": Musical and Textual Features for Automatic Mood Classification. *Journal of New Music Research*, 39(1):13–34, 2010.

58 B. Schuller, B. Hörnler, D. Arsić, and G. Rigoll. Audio Chord Labeling by Musiological Modeling and Beat-Synchronization. In *Proc. of ICME*, pages 526–529, New York, NY, July 2009. IEEE, IEEE.

59 B. Schuller, C. Kozielski, F. Weninger, F. Eyben, and G. Rigoll. Vocalist Gender Recognition in Recorded Popular Music. In *Proc. of ISMIR*, pages 613–618, Utrecht, The Netherlands, October 2010. ISMIR, ISMIR.

60 B. Schuller, F. Weninger, and J. Dorfner. Multi-Modal Non-Prototypical Music Mood Analysis in Continuous Space: Reliability and Performances. In *Proc. of ISMIR*, pages 759–764, Miami, FL, USA, 2011.

61 J. Serrà. A Qualitative Assessment of Measures for the Evaluation of a Cover Song Identification System. In *Proc. ISMIR*, pages 319–322, September 2007.

62 J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Trans. on Audio Speech and Language Processing*, 16(6):1138–1152, August 2008.

63 M. Slaney. Precision-Recall is Wrong for Multimedia. *IEEE Multimedia*, 18(3):4–7, 2011.

64 P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180, New Paltz, NY, USA, 2003.

65 A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(23):1349–1380, December 2000.

66 S. Steidl, B. Schuller, A. Batliner, and D. Seppi. The Hinterland of Emotions: Facing the Open-Microphone Challenge. In *Proc. of ACII*, pages 690–697, Amsterdam, The Netherlands, 2009.

67 J. Urbano. Information Retrieval Meta-Evaluation: Challenges and Opportunities in the Music Domain. In *Proc. ISMIR*, pages 609–614, October 2011.

68 E. Vincent, N. Bertin, and R. Badeau. Harmonic and Inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch Transcription. In *Proc. of ICASSP*, pages 109–112, 2008.

69 T. Virtanen. Speech Recognition Using Factorial Hidden Markov Models for Separation in the Feature Space. In *Proc. of INTERSPEECH*, pages 1–4, Pittsburgh, PA, USA, 2006.

70 D. von Zeddelmann, F. Kurth, and M. Müller. Perceptual Audio Features for Unsupervised Key-Phrase Detection. In *Proc. IEEE ICASSP*, Dallas, TX, USA, Mar. 2010.

71 A. Wang. An Industrial Strength Audio Search Algorithm. In *International Conference on Music Information Retrieval, Baltimore*, 2003.

72 B. Weiss and F. Burkhardt. Voice attributes affecting likability perception. In *Proc. of INTERSPEECH*, pages 2014–2017, 2010.

73 F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll. The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments. In *Proc. of CHiME Workshop*, pages 24–29, Florence, Italy, 2011.

74 F. Weninger, M. Wöllmer, and B. Schuller. Automatic Assessment of Singer Traits in Popular Music: Gender, Age, Height and Race. In *Proc. of ISMIR*, pages 37–42, Miami, FL, USA, 2011.

75 G. A. Wiggins, D. Müllensiefen, and M. T. Pearce. On the non-existence of Music: Why Music Theory is a figment of the imagination. *Musicae Scientiae*, Discussion Forum 5:231–255, 2010.

76 J. Wilpon, L. Rabiner, C.-H. Lee, and E. Goldman. Automatic recognition of keywords in unconstrained speech using hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(11):1870–1878, 1990.

77 M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. Interspeech*, pages 597–600, Brisbane, 2008.

78 M. Wöllmer, F. Weninger, F. Eyben, and B. Schuller. Acoustic-Linguistic Recognition of Interest in Speech with Bottleneck-BLSTM Nets. In *Proc. of INTERSPEECH*, pages 77–80, Florence, Italy, 2011.

79 Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):448–457, 2008.

80 K. Yoshii and M. Goto. A Vocabulary-Free Infinity-Gram Model for Nonparametric Bayesian Chord Progression Analysis. In *Proc. of ISMIR*, pages 645–650, Miami, FL, USA, 2011.

81 C. H. You, K.-A. Lee, and H. Li. An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition. *IEEE Signal Processing Letters*, 16(1):49–52, 2009.

82 S. J. Young, G. Evermann, M. J. F. Gales, D. Kershaw, G. Moore, J. J. Odell, D. G. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK book version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.