*Acoust. Sci. & Tech.* **25**, 6 (2004)

**INVITED REVIEW**

# Recent studies on music information processing

Masataka Goto[1] and Keiji Hirata[2]

[1]*National Institute of Advanced Industrial Science and Technology (AIST)*
[2]*NTT Communication Science Laboratories*

**Keywords:** Music information processing, Music information retrieval, Music understanding, Metadata of musical pieces, End-user interfaces

**PACS number:** 43.75.−z [DOI: 10.1250/ast.25.419]

## 1. LINKING MUSIC INFORMATION PROCESSING WITH THE REAL WORLD

Music information processing has been widely deployed in music industries over the years. Of course, technologies oriented to musicians have long been studied including sound synthesis on music synthesizers, desktop music production based on MIDI sequencers, and various kinds of support for composing, performing, and recording music. Such tools have already become an essential part of the music-production process. But more recently, focus has shifted from these conventional tools to new technologies that target the direct enjoyment of music by end users who are not musicians. For example, it has become relatively easy to "rip" audio signals from compact discs (CD) and compress them and to deal with many musical pieces on a personal computer. It has also become possible to load a huge number of songs onto a portable music player (e.g., Apple iPod) enabling anyone to carry their personal collection of music anywhere and to listen to it at anytime.

A variety of factors can be given for this trend including advances in computer hardware (high processing speeds and large-capacity/small-size memory and hard disks), spread of the Internet, and provision of low-cost audio input/output devices as standard equipment. The standardization of MPEG Audio Layer 3 (MP3) in 1992 and its spread in the latter half of the 1990s and the establishment of MP3-based businesses in response to end-user demand have also played a role here. This trend is accelerating all the more in the first half of the 2000s with the proposal of Ogg Vorbis, MPEG-4 AAC, Windows Media Audio (WMA), and other compression systems following on the heels of MP3. Enterprises for delivering music via the Internet are also appearing in rapid succession.

End users who are not musicians are not generally proficient in music — their knowledge of notes, harmony, and other elements of music is usually limited. Furthermore, they generally have little desire to create music. They are quite interested, however, in retrieving and listening to their favorite music or a portion of a musical piece in a convenient and flexible way. Recent research themes of music information processing reflect such end-user demand. The target of processing is expanding from the internal content of individual musical pieces (notes, chords, etc.) to entire musical pieces and even sets of musical pieces. Accordingly, research is becoming active in music systems that can be used by people with no musical knowledge. Typical technologies driving this trend are technology for computing similarity between musical pieces and for retrieving and classifying music; technology for referring to what music friends and other people listen to and for selecting music accordingly; and technology for creating advanced music-handling interfaces.

Focusing on this emerging research trend, this paper introduces recent studies on music information processing from a unique perspective.

## 2. HANDLING SETS OF MUSICAL PIECES

In contrast to past research that focused on the internal contents of individual musical pieces, the past ten years have seen the growth of a new research field targeting the retrieval, classification, and management of large sets of music in which a single musical piece is treated as a unit. This field, which is called Music Information Retrieval (MIR), has become quite active, and since 2000, the International Conference on Music Information Retrieval (ISMIR) has been held annually. A variety of topics are being researched in this field, but in the following, we introduce three ways of retrieving music based on audio signals as opposed to text searches based on bibliographic information (such as titles and artist names from CDDB, an online database of CD information).

### 2.1. Research on Melody: Query by Humming

As the name implies, Query by Humming (QBH)

enables one to retrieve the title of a musical piece by humming or singing its melody using sounds like "*la-la-la...*" In other words, humming or singing a melody becomes the search key for finding a musical piece with that melody. The use of such search keys raises some issues, however, such as how to deal with errors when singing off key and how to absorb differences in key and tempo. Specific methods differ in terms of the database used, which may consist of melodies only [1–5], standard MIDI files (SMF) of entire musical pieces [6–8], or audio signals of entire musical pieces [9,10]. If we use a melody-only database, similarity with a search key can be directly computed, but for SMF, the track containing a melody must be identified before computing similarity. In the case of audio signals, similarity with a melody included in a mixture of sounds must be computed, which is even harder to achieve.

### 2.2. Research on Music Fragments: Retrieving a Musical Piece Containing a Certain Fragment

For someone who would like to know the title of a musical piece that is currently playing on the street or elsewhere, this retrieval method enables that title to be identified based on a fragment of that piece, which can be recorded on a cellular phone. The fragment is therefore the search key and the method searches for the musical piece containing that fragment. Important issues here are how to achieve efficient searching and how to absorb acoustic fluctuations caused by noise and distortion on the transmission path. Proposed methods include the time-series active search method based on histograms of vector-quantized power-spectrum shapes [11] and a method based on patterns of power-spectrum peaks [12].

### 2.3. Research on Entire Musical Pieces: Retrieval Based on Similarity between Pieces

Given that one likes certain musical pieces, this method searches for another musical piece having a similar feeling. The search key is musical pieces themselves and the method searches for a similar piece. To this end, similarity must be defined based on various features such as timbral texture within a piece (power-spectrum shape) [13,14], rhythm [14–17], modulation spectrum [18], and singer voice [19]. Similarity is also important for purposes other than retrieval. For example, the use of similarity to automatically classify musical pieces (into genres, music styles, etc.) is also being researched [14,17–20]. It is difficult, however, to compute the appropriate similarity between musical pieces considering various factors. This in conjunction with the understanding of musical audio signals as introduced in the following section needs further research in the years to come.

## 3. UNDERSTANDING MUSICAL AUDIO SIGNALS

Research related to the understanding of musical audio signals has developed significantly over the last ten years. Before that, it was common to research the segregation and extraction of individual sound components making up an audio signal (sound source segregation) and to use that information to automatically generate a musical score (automatic transcription). But in 1997, in reconsideration of what it means for human beings to understand music, a new research approach was proposed on music understanding (Sect. 3.2, Music scene description [21–23]) based on the viewpoint that listeners understand music without segregating sound sources and without mentally representing audio signals as musical scores [24]. Research themes conforming to this approach, including beat tracking, melody extraction, and music structure analysis, have also been proposed.

This major development on the understanding of musical audio signals has been supported by advances in hardware and in techniques for processing audio signals. Ten years ago, it was still difficult to calculate a Fast Fourier Transform (FFT) in real time, but nowadays, it can be performed so fast that the time required for its computation can essentially be ignored. This jump in processing performance has let researchers devise computationally intensive approaches that could not be considered in the past, and has also promoted the use of a wide range of statistical techniques. For example, techniques based on probabilistic models such as the Hidden Markov Model (HMM) and various techniques making use of maximum likelihood estimation and Bayes estimation have been proposed.

### 3.1. Sound Source Segregation and Pitch ($F_0$) Estimation

Automatic transcription has a long history as a research theme going back to the 1970s, and has progressed steadily as the difficulty of the target music has increased from monophonic sounds of melodies to polyphonic sounds from a single instrument and a mixture of sounds from several instruments. This progression has been accompanied by a shift toward more specialized research topics, namely, sound source segregation and estimation of fundamental frequency ($F_0$, perceived as pitch).

Because space does not allow an exhaustive introduction to the many studies in this research field, we here focus on new approaches that first appeared in the past ten years. In 1994, Kashino *et al.* introduced a method based on a probabilistic model and implemented as a process model called OPTIMA [25,26]. This method was novel in its use of a graphical model to describe the hierarchical structure

of frequency components, musical notes, and chords and in determining the most likely interpretation based on this hierarchical relationship. Then, in 1999, Goto proposed a predominant-$F_0$ estimation method (PreFEst) that does not assume the number of sound sources [21,23,27]. This method prepares probability distributions that represent the shape of harmonic structures for all possible $F_0$s, and models input frequency components as a mixture (weighted sum) of those distributions. It then estimates the parameters of this model — the amplitude (weight) of each component sound in the input sound mixture and the shape of its harmonic structure — by using Maximum *A Posteriori* Probability (MAP) estimation executed by the Expectation-Maximization (EM) algorithm. This method can be extended, in principle, to an inharmonic structure [23,27], and as such, can be considered a framework for understanding general sound mixtures.

Other proposed methods include a method for sequentially determining the components in a sound mixture by repeatedly estimating the predominant $F_0$ and removing its harmonic components [28]; a method for estimating model parameters such as the number of simultaneous sounds, the number of frequency components making up each sound, $F_0$s, and amplitude by modeling the signal as a weighted sum of sound waveforms in the time domain and applying the Markov Chain Monte Carlo (MCMC) algorithm [29]; a method for estimating notes, tempo, and waveforms by associating them with a graphical model that models the waveform-generation process when performing a musical score at a certain (local) tempo [30]; and a method that formalizes the problem as the clustering of frequency components under harmonic-structure constraints and determines the number of clusters (sound sources) that minimizes the Akaike Information Criterion (AIC) so as to estimate the median ($F_0$) and weight (amplitude) of each cluster [31].

### 3.2. Music Scene Description

Music scene description [21–23] aims to achieve an understanding of musical audio signals at the level of untrained listeners. This contrasts with most studies in the past that aimed to achieve it at the level of trained musicians by identifying all musical notes forming a musical score or obtaining segregated signals from sound mixtures. Music scene description features the description of "scenes" that occur within a musical performance such as melody, bass, beat, chorus and phrase repetition, structure of the musical piece, and timbre of musical instruments. The following introduce methods for obtaining descriptions of such scenes.

**Melody and bass lines** Estimating the fundamental frequency ($F_0$) of melody and bass lines in CD recordings containing sounds of various musical instruments was first achieved in 1999 by applying the previously mentioned PreFEst method [21,23,27] with appropriate frequency-range limitation. Another method for estimating the $F_0$ of the bass line was later proposed by Hainsworth *et al.* [32].

**Beat** While beat tracking and measure (bar line) estimation for obtaining a hierarchical beat structure (including beat and measure levels) were researched in the 1980s using MIDI signals, Goto *et al.* began a series of studies [24,33–36] on real-world sound mixtures in 1994. Then, taking a hint from these studies, Scheirer proposed a beat-tracking method that could accommodate changes in tempo [37]. Still later, a variety of methods having even fewer restrictions were proposed [38–40]. Methods targeting MIDI signals [41,42] have also progressed significantly in recent years.

**Chorus and phrase repetition, music structure** On entering the 2000s, new approaches appeared based on the detection of similar sections that repeat within a musical piece (such as a repeating phrase). These led to methods for extracting the most representative section of a musical piece (usually the chorus) from one location [43–45]; music-summarization methods that shorten a musical piece leaving only main sections [46,47]; and the RefraiD method that exhaustively detects all chorus sections [48]. Among these methods, RefraiD focuses on chorus detection with the capability of determining the start and end points of every chorus section regardless of whether a key-change occurs.

**Timbre of musical instruments** The second half of the 1990s saw the development of sound-source identification methods [49,50] that recognize the musical-instrument name of each component sound in a polyphonic sound mixture simultaneously with $F_0$ estimation. This period also saw the development of methods for estimating the timing of drum sounds in musical performances [33,34,36,51,52]. Sound-source identification methods for isolated monophonic sounds have also been researched from various viewpoints [53–58].

## 4. HANDLING METADATA OF MUSICAL PIECES

To respond directly to end-user demands for flexible retrieval of music for one's listening pleasure, much research has been targeting the extraction and usage of metadata to enhance the listening of musical pieces or to facilitate their retrieval. Such metadata include information on the composers and performers of musical pieces and the listener's preferences with regard to those pieces.

### 4.1. Utilization

End users find it convenient if they can refer to a list of music that other people listened to as a basis for selecting their own music. On the Internet, music-sales sites (e.g., Amazon) and music-review sites (e.g., Allmusic) perform daily collections of metadata including user evaluations and impressions and purchase history. This information can be subjected to collaborative filtering to achieve services that promote the purchasing of music by recommending artists and albums to users [59–61] and proposing playlists [62–65] (The original meaning of "playlist" is a broadcast or concert program of musical pieces but it here means a list of musical pieces to be played back on a media player or other device).

### 4.2. Extraction

Because collaborative filtering by itself cannot easily deal with unknown new musical pieces, it can be reinforced with content-based filtering [61]. This makes the music-understanding methods described in Section 3 all the more important. Using the results of those methods can aid in the generation of more appropriate recommendations and playlists based on the acoustical features and content of musical pieces.

As an example of using both metadata and acoustical features, Whitman *et al.* have developed methods for detecting artist styles [20] and identifying artists [66] by combining acoustical features of artist's musical pieces with statistical data on words or phrases on WWW pages that include that artist's name. Ellis *et al.*, moreover, have investigated an automatic measure of the similarity between artists by extracting metadata from lists of similar artists on a music-review site, from end-user music collections, and from statistical data on words or phrases on WWW pages that include artist names [67]. In addition, Berenzweig *et al.* have compared similarity based on such metadata with similarity based on audio signals [68].

### 4.3. Description and Standardization

While methods for encoding only musical-score information have already reached a sufficiently practical level [69], there are currently several XML-based proposals for music-description methods (including metadata) and their standardization. For example, MusicXML [70] and WEDELMUSIC [71] have been proposed for describing music at a symbol level including musical-score information. Likewise, MPEG-7 Audio [72] has been standardized for describing metadata related to musical audio signals such as melody contour and statistical data on the power spectrum. We expect various research and development activities conforming to these proposals to appear.

## 5. CONSIDERING END-USER INTERFACES

To enable end users without detailed knowledge of music to deal with music on their own terms, it is important that new types of interfaces be developed since existing tools designed for musicians are not sufficient for this purpose.

### 5.1. Real-World Oriented Approach

To achieve interfaces that can be used naturally without hassle, we can consider the use of real-world objects themselves as interfaces. Here, it is important that such an approach conforms to conventional usage. The musicBottles [73] and FieldMouse [74] are two examples of this real-world approach. The musicBottles is a music-playback interface that associates each musical-instrument part in a musical piece with different glass bottles and then enables a user to play the sound of any part only when the cap of that bottle is in the open position. The FieldMouse is an input device that combines an ID-tag detector such as a barcode reader with a relative-position detector such as a mouse. A user can then select musical pieces and adjust the playback volume, for example, by moving the FieldMouse through space to read ID-tags that correspond to those operations.

In conventional listening stations and media players, an end user who would like to listen to only the chorus section of a song must search for it himself by pressing the fast-forward button repeatedly. To make this task easier for an end user, SmartMusicKIOSK [75] adds a "NEXT CHORUS" button by employing the RefraiD method [48] described earlier. Pressing this button forces a jump to the next chorus section in that song through automatic chorus detection. This makes it easy for a user to immediately skip a section (part) of no interest within a song much like the "NEXT TRACK" button on a CD player enables him to skip a song (track) of no interest.

### 5.2. Onomatopoeia

Considering that end users will be faced with various types of music systems in the future, they should be able to input musical information (such as melody and rhythm) associated with specific musical pieces. Effective means to this end include humming (Sect. 2.1) and onomatopoeia as introduced in the following. A music notation system "Sutoton Music" [76] enables a melody to be described in text form in the manner of "*do re mii so-mmi re do*" for playback on a computer. A drum-pattern retrieval method by voice percussion (beatboxing) [77] aims to recognize the drum part of a musical piece that the user utters using natural sounds like "*dum ta dum-dum ta*" and to search for that piece on the basis of that drum part.

### 5.3. Communication Tool

It was mentioned earlier that end users do not generally have much interest in creating music. Nevertheless, if easy-to-use support functions for creating music could be embedded in social networking tools (e.g., Orkut), music might also become a means of communication for end users.

One example of music systems that place particular emphasis on inter-user communication is CosTune [78]. In this system, pads that control different sounds are attached to a user's jacket or pants and touching these pads in a rhythmical manner enables the user to jam with other nearby users via a wireless network. Another example is Music Resonator [79], which enables a user to process and edit annotated fragments of musical pieces and to share that music with other users for collaborative music productions. In addition, RemoteGIG [80] enables remotely located users to jam together along a repetitive chord progression like 12-bar blues in real time over the Internet despite its relatively large latency. This system overcomes the network latency by having users listen to each other's performance delayed by just one cycle of the chord progression (several tens of seconds).

## 6. EXPANDING MUSIC INFORMATION PROCESSING

There are other interesting themes not introduced in this paper that are being actively researched in the field of music information processing. While we have here focused on topics related to audio signal processing, research into symbol-level processing including musical scores and MIDI has also been progressing. For example, there has been much work on computing similarity between melodies at a symbol level [81] and research on musical structure and expression in musical performances [82]. Fusion of symbol processing and audio signal processing is still far from sufficient and will be targeted as an important issue in the future. Such fusion will bridge a gap between them and enable symbol processing to be based on proper symbol grounding and audio signal processing to cover abstract semantic computing, eventually achieving music computing that reflects the manifold meaning of music.

The research environment for music information processing is also expanding. The years 2000 and 2001 saw the construction of the world's first copyright-cleared music database "RWC Music Database" that can be used in common for research purposes [83]. This database makes it easier to use music for comparing and evaluating various methods, for corpus-based machine learning, and for publishing research and making presentations without conventional copyright restrictions. Considering that shared databases of various kinds have long been constructed in other research fields and have made significant contributions to their advancement, we anticipate the RWC Music Database to contribute to the advancement of music information processing in a similar way.

Ten years ago, it was necessary for us to explain that music information processing was not "amusement" but a real research topic. Today, it is common sense to treat it as an important research field. This field is now experiencing the birth of large-scale projects one after another, an increase in international conferences year by year, and an ever increasing number of researchers. We look forward to further advances in music information processing research.

## REFERENCES


[1] T. Kageyama, K. Mochizuki and Y. Takashima, "Melody retrieval with humming," *Proc. ICMC 1993*, pp. 349–351 (1993).

[2] A. Ghias, J. Logan, D. Chamberlin and B. C. Smith, "Query by humming: Musical information retrieval in an audio database," *Proc. ACM Multimedia 95*, pp. 231–236 (1995).

[3] T. Sonoda, M. Goto and Y. Muraoka, "A WWW-based melody retrieval system," *Proc. ICMC 1998*, pp. 349–352 (1998).

[4] S. Pauws, "CubyHum: A fully operational query by humming system," *Proc. ISMIR 2002*, pp. 187–196 (2002).

[5] T. Sonoda, T. Ikenaga, K. Shimizu and Y. Muraoka, "The design method of a melody retrieval system on parallelized computers," *Proc. WEDELMUSIC 2002*, pp. 66–73 (2002).

[6] J. Shifrin, B. Pardo, C. Meek and W. Birmingham, "HMM-based musical query retrieval," *Proc. JCDL 2002*, pp. 295–300 (2002).

[7] N. Hu and R. B. Dannenberg, "A comparison of melodic database retrieval techniques using sung queries," *Proc. JCDL 2002*, pp. 301–307 (2002).

[8] R. B. Dannenberg, W. P. Birmingham, G. Tzanetakis, C. Meek, N. Hu and B. Pardo, "The MUSART testbed for query-by-humming evaluation," *Proc. ISMIR 2003*, pp. 41–47 (2003).

[9] T. Nishimura, H. Hashiguchi, J. Takita, J. X. Zhang, M. Goto and R. Oka, "Music signal spotting retrieval by a humming query using start frame feature dependent continuous dynamic programming," *Proc. ISMIR 2001*, pp. 211–218 (2001).

[10] J. Song, S. Y. Bae and K. Yoon, "Mid-level music melody representation of polyphonic audio for query-by-humming system," *Proc. ISMIR 2002*, pp. 133–139 (2002).

[11] K. Kashino, T. Kurozumi and H. Murase, "A quick search method for audio and video signals based on histogram pruning," *IEEE Trans. Multimedia*, **5**, 348–357 (2003).

[12] A. Wang, "An industrial-strength audio search algorithm," *Proc. ISMIR 2003*, pp. 7–13 (2003).

[13] J.-J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?" *Proc. ISMIR 2002*, pp. 157–163 (2002).

[14] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Proc.*, **10**, 293–302 (2002).

[15] J. Paulus and A. Klapuri, "Measuring the similarity of rhythmic patterns," *Proc. ISMIR 2002*, pp. 150–156 (2002).

[16] J. Foote, M. Cooper and U. Nam, "Audio retrieval by rhythmic similarity," *Proc. ISMIR 2002*, pp. 265–266 (2002).

[17] S. Dixon, E. Pampalk and G. Widmer, "Classification of dance music by periodicity patterns," *Proc. ISMIR 2003*, pp. 159–165 (2003).

[18] M. F. McKinney and J. Breebaart, "Features for audio and music classification," *Proc. ISMIR 2003*, pp. 151–158 (2003).

[19] W.-H. Tsai, H.-M. Wang, D. Rodgers, S.-S. Cheng and H.-M. Yu, "Blind clustering of popular music recordings based on singer voice characteristics," *Proc. ISMIR 2003*, pp. 167–173 (2003).
[20] B. Whitman and P. Smaragdis, "Combining musical and cultural features for intelligent style detection," *Proc. ISMIR 2002*, pp. 47–52 (2002).
[21] M. Goto, "A real-time music scene description system: Detecting melody and bass lines in audio signals," *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pp. 31–40 (1999).
[22] M. Goto, "Music scene description project: Toward audio-based real-time music understanding," *Proc. ISMIR 2003*, pp. 231–232 (2003).
[23] M. Goto, "A real-time music scene description system: Predominant-$F_0$ estimation for detecting melody and bass lines in real-world audio signals," *Speech Commun.* (2004).
[24] M. Goto and Y. Muraoka, "Real-time rhythm tracking for drumless audio signals — chord change detection for musical decisions —," *Working Notes of the IJCAI-97 Workshop on Computational Auditory Scene Analysis*, pp. 135–144 (1997).
[25] K. Kashino, *Computational Auditory Scene Analysis for Music Signals*, PhD thesis, University of Tokyo (1994).
[26] K. Kashino, K. Nakadai, T. Kinoshita and H. Tanaka, "Organization of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitative information integration mechanism," *Proc. IJCAI-95*, pp. 158–164 (1995).
[27] M. Goto, "A predominant-$F_0$ estimation method for polyphonic musical audio signals," *Proc. ICA 2004*, pp. II–1085–1088 (2004).
[28] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, **11**, 804–816 (2003).
[29] M. Davy and S. J. Godsill, "Bayesian harmonic models for musical signal analysis," *Bayesian Stat.*, **7**, 105–124 (2003).
[30] A. T. Cemgil, B. Kappen and D. Barber, "Generative model based polyphonic music transcription," *Proc. WASPAA 2003*, pp. 181–184 (2003).
[31] H. Kameoka, T. Nishimoto and S. Sagayama, "Extraction of multiple fundamental frequencies from polyphonic music using harmonic clustering," *Proc. ICA 2004*, pp. I–59–62 (2004).
[32] S. W. Hainsworth and M. D. Macleod, "Automatic bass line transcription from polyphonic music," *Proc. ICMC 2001*, pp. 431–434 (2001).
[33] M. Goto and Y. Muraoka, "A beat tracking system for acoustic signals of music," *Proc. ACM Multimedia '94*, pp. 365–372 (1994).
[34] M. Goto, *A Study of Real-time Beat Tracking for Musical Audio Signals*, PhD thesis, Waseda University (1998).
[35] M. Goto and Y. Muraoka, "Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions," *Speech Commun.*, **27**, 311–335 (1999).
[36] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *J. New Music Res.*, **30**, 159–171 (2001).
[37] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, **103**, 588–601 (1998).
[38] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *J. New Music Res.*, **30**, 39–58 (2001).
[39] S. Hainsworth and M. Macleod, "Beat tracking with particle filtering algorithms," *Proc. WASPAA 2003*, pp. 91–94 (2003).
[40] A. P. Klapuri, A. J. Eronen and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Speech Audio Process.* (2004).
[41] A. T. Cemgil and B. Kappen, "Monte carlo methods for tempo tracking and rhythm quantization," *J. Artif. Intell. Res.*, **18**, 45–81 (2003).
[42] H. Takeda, T. Nishimoto and S. Sagayama, "Automatic rhythm transcription of multiphonic MIDI signals," *Proc. ISMIR 2003*, pp. 263–264 (2003).
[43] B. Logan and S. Chu, "Music summarization using key phrases," *Proc. ICASSP 2000*, pp. II-749–752 (2000).
[44] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," *Proc. WASPAA '01*, pp. 15–18 (2001).
[45] M. Cooper and J. Foote, "Automatic music summarization via similarity analysis," *Proc. ISMIR 2002*, pp. 81–85 (2002).
[46] G. Peeters, A. L. Burthe and X. Rodet, "Toward automatic music audio summary generation from signal analysis," *Proc. ISMIR 2002*, pp. 94–100 (2002).
[47] R. B. Dannenberg and N. Hu, "Pattern discovery techniques for music audio," *Proc. ISMIR 2002*, pp. 63–70 (2002).
[48] M. Goto, "A chorus-section detecting method for musical audio signals," *Proc. ICASSP 2003*, pp. V–437–440 (2003).
[49] K. Kashino and H. Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Commun.*, **27**, 337–349 (1999).
[50] J. Eggink and G. J. Brown, "A missing feature approach to instrument recognition in polyphonic music," *Proc. ICASSP 2003*, pp. V–553–556 (2003).
[51] A. Zils, F. Pachet, O. Delerue and F. Gouyon, "Automatic extraction of drum tracks from polyphonic music signals," *Proc. WEDELMUSIC 2002*, pp. 179–183 (2002).
[52] K. Yoshii, M. Goto and H. G. Okuno, "Automatic drum sound description for real-world music using template adaptation and matching methods," *Proc. ISMIR 2004* (2004).
[53] K. D. Martin, *Sound-Source Recognition: A Theory and Computational Model*, PhD thesis, MIT (1999).
[54] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coeddicients and temporal features," *Proc. ICASSP 2000*, pp. II–753–756 (2000).
[55] J. C. Brown, O. Houix and S. McAdams, "Feature dependence in the automatic identification of musical woodwind instruments," *J. Acoust. Soc. Am.*, **109**, 1064–1072 (2001).
[56] M. A. Casey, "Reduced-rank spectra and minimum-entropy priors as consistent and reliable cues for generalized sound recognition," *Proc. CRAC 2001* (2001).
[57] P. Herrera, A. Yeterian and F. Gouyon, "Automatic classification of drum sounds: A comparison of feature selection methods and classification techniques," *Proc. ICMAI 2002*, pp. 69–80 (2002).
[58] T. Kitahara, M. Goto and H. G. Okuno, "Musical instrument identification based on $F_0$-dependent multivariate normal distribution," *Proc. ICASSP 2003*, pp. V–421–424 (2003).
[59] U. Shardanand and P. Maes, "Social information filtering: Algorithms for automating "word of mouth"," *Proc. CHI '95*, pp. 210–217 (1995).
[60] W. W. Cohen and W. Fan, "Web-collaborative filtering: Recommending music by crawling the Web," *Proc. WWW9* (2000).
[61] A. Uitdenbogerd and R. van Schyndel, "A review of factors affecting music recommender success," *Proc. ISMIR 2002*, pp. 204–208 (2002).
[62] M. Alghoniemy and A. H. Tewfik, "A network flow model for playlist generation," *Proc. ICME 2001* (2001).
[63] S. Pauws and B. Eggen, "PATS: Realization and user

evaluation of an automatic playlist generator," *Proc. ISMIR 2002*, pp. 222–230 (2002).
[64] B. Logan, "Content-based playlist generation: Exploratory experiments," *Proc. ISMIR 2002*, pp. 295–296 (2002).
[65] J.-J. Aucouturier and F. Pachet, "Scaling up music playlist generation," *Proc. ICME 2002* (2002).
[66] B. Whitman, "Semantic rank reduction of music audio," *Proc. WASPAA 2003*, pp. 135–138 (2003).
[67] D. P. Ellis, B. Whitman, A. Berenzweig and S. Lawrence, "The quest for ground truth in musical artist similarity," *Proc. ISMIR 2002*, pp. 170–177 (2002).
[68] A. Berenzweig, B. Logan, D. P. Ellis and B. Whitman, "A large-scale evaluation of acoustic and subjective music similarity measure," *Proc. ISMIR 2003*, pp. 99–105 (2003).
[69] E. Selfridge-Field, Ed., *Beyond MIDI* (The MIT Press, Cambridge, Mass., 1997).
[70] M. Good, "Representing music using XML," *Proc. ISMIR 2000* (2000).
[71] P. Bellini and P. Nesi, "WEDELMUSIC format: An XML music notation format for emerging applications," *Proc. WEDELMUSIC 2001*, pp. 79–86 (2001).
[72] ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group, Information technology — multimedia content description interface — part 4: Audio, 15938-4:2002 (2002).
[73] H. Ishii, A. Mazalek and J. Lee, "Bottles as a minimal interface to access digital information," *Proc. CHI 2001*, pp. 187–188 (2001).
[74] T. Masui and I. Siio, "Real-world graphical user interfaces," *Proc. HUC 2000*, pp. 72–84 (2000).
[75] M. Goto, "SmartMusicKIOSK: Music listening station with chorus-search function," *Proc. UIST 2003*, pp. 31–40 (2003).
[76] T. Masui, "Music composition by onomatopoeia," *Proc. IWEC 2002*, pp. 297–304 (2002).
[77] T. Nakano, J. Ogata, M. Goto and Y. Hiraga, "A drum pattern retrieval method by voice percussion," *Proc. ISMIR 2004* (2004).
[78] K. Nishimoto, T. Maekawa, Y. Tada, K. Mase and R. Nakatsu, "Networked wearable musical instruments will bring a new musical culture," *Proc. ISWC 2001*, pp. 55–62 (2001).
[79] K. Hirata, S. Matsuda, K. Kaji and K. Nagao, "Annotated music for retrieval, reproduction and exchange," *Proc. ICMC 2004* (2004).
[80] M. Goto, R. Neyama and Y. Muraoka, "RMCP: Remote music control protocol — design and applications —," *Proc. ICMC 1997*, pp. 446–449 (1997).
[81] W. B. Hewlett and E. Selfridge-Field, Ed., *Melodic Similarity: Concepts, Procedures, and Applications*, (The MIT Press, Cambridge, Mass., 1998).
[82] *Proc. ICAD 2002 Rencon Workshop* (2002).
[83] M. Goto, "Development of the RWC music database," *Proc. ICA 2004*, pp. I–553–556 (2004).

**Masataka Goto** received his Doctor of Engineering degree in Electronics, Information and Communication Engineering from Waseda University, Japan, in 1998. He then joined the Electrotechnical Laboratory (ETL; reorganized as the National Institute of Advanced Industrial Science and Technology (AIST) in 2001), where he has been engaged as a researcher ever since. He served concurrently as a researcher in Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Corporation (JST) from 2000 to 2003. His research interests include music information processing and spoken language processing. Dr. Goto received the IPSJ Yamashita SIG Research Awards (MUS and SLP) from the Information Processing Society of Japan (IPSJ), Best Paper Award for Young Researchers from the Kansai-Section Joint Convention of Institutes of Electrical Engineering, WISS 2000 Best Paper Award and Best Presentation Award, Awaya Prize for Outstanding Presentation and Award for Outstanding Poster Presentation from the Acoustical Society of Japan (ASJ), Award for Best Presentation from the Japanese Society for Music Perception and Cognition (JSMPC), and Interaction 2003 Best Paper Award. He is a member of the ASJ, IPSJ, JSMPC, Institute of Electronics, Information and Communication Engineers (IEICE), and International Speech Communication Association (ISCA).

**Keiji Hirata** received his Doctor of Engineering degree in Information Engineering from University of Tokyo, Japan, in 1987. He then joined NTT Basic Research Laboratories. He spent 1990 to 1993 at the Institute for New Generation Computer Technology (ICOT), where he was engaged in the research and development of parallel inference machines. In 1999, he joined NTT Communication Science Laboratories, where he has been engaged as a researcher ever since. His research interests include musical knowledge programming and interaction. Dr. Hirata received the IPSJ Best Paper Award from the Information Processing Society of Japan (IPSJ) in 2001, and is a member of the IPSJ, Japanese Society for Artificial Intelligence (JSAI), and Japan Society for Software Science and Technology.