
Music Scene Description Project: Toward Audio-based Real-time Music Understanding

Masataka Goto

“Information and Human Activity,” PRESTO, JST / National Institute of Advanced Industrial Science and Technology (AIST)
IT, AIST, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
m.goto@aist.go.jp

Abstract

This paper reports a research project intended to build a real-time music-understanding system producing intuitively meaningful descriptions of real-world musical audio signals, such as the melody lines and chorus sections. This paper also introduces our efforts to add correct descriptions (metadata) to the pieces in a music database.

1 Introduction

Our goal is to build a computer system that can understand musical audio signals in a human-like fashion. People listening to music (especially popular music) can easily hum the melody, notice a phrase being repeated, and find chorus sections. The brain mechanisms underlying these abilities, however, have not been understood. It has also been difficult to implement these abilities on a computer system, although a system with them would be useful in various applications such as music information retrieval and music production/editing. We therefore want to build a real-time system that can understand complex real-world monaural music signals like those recorded on commercially distributed compact discs (CDs).

Two popular approaches are to build a sound source separation system (Casey and Westner, 2000) or an automatic music transcription system (Katayose and Inokuchi, 1989; Klapuri et al., 2001). Although these technologies are valuable from an engineering viewpoint, neither separation nor transcription is necessary or sufficient for understanding music. The fact that human listeners understand various properties of audio signals is not necessarily evidence that the human auditory system extracts each individual audio signal: even if a mixture of two components cannot be separated, that the mixture includes them can be understood from their salient features. Indeed, as pointed out by Goto and Muraoka (1999) and Scheirer (2000), untrained listeners understand music to some extent without mentally representing audio signals as musical scores: music transcription is a skill mastered only by trained musicians. Furthermore, even if we could derive separated signals and musical notes, it would still not be easy to obtain high-level music descriptions like

melody lines and chorus sections.

We have therefore been trying to construct a *real-time music-scene-description system* that obtains descriptions intuitively meaningful to untrained listeners. By considering what is to be achieved to understand music, we have proposed the following descriptions: hierarchical beat structure, melody line, bass line, repeated sections, and chorus sections (Figure 1). The following sections introduce our methods for producing these descriptions and report our efforts to provide the songs in a music database with correct descriptions by using a metadata editor we developed.

2 Real-time Methods for Obtaining Music Scene Descriptions Automatically

We have proposed and implemented the following methods that produce, in real time, the descriptions shown in Figure 1 of real-world audio signals containing simultaneous sounds of various instruments (with or without drum-sounds).

- (1) An audio-based real-time beat-tracking method (Goto and Muraoka, 1999; Goto, 2001a)

This method recognizes, in audio signals sampled from popular-music CDs, a hierarchical beat structure comprising the quarter-note and measure levels. Its main advantage is that it can track beats above the quarter-note level by using three kinds of musical knowledge: onset times, chord changes, and drum patterns.

- (2) A predominant-F0 estimation method for detecting melody and bass lines (*PreFEst*) (Goto, 2001b, 2003b)

The *PreFEst* (PreFundamental-F0 Estimation Method) estimates the fundamental frequency (F0) of the melody and bass lines. To do this without assuming the number of sound sources, it considers every possible F0 at the same time and estimates a probability density function of the F0 (relative dominance of each possible F0) by using the MAP (maximum *a posteriori* probability) estimation and the EM (expectation-maximization) algorithm.

- (3) A chorus-section detection method (*RefraiD*) (Goto, 2003a,c)

The *RefraiD* (Refrain Detecting Method) detects sections being repeated and identifies the chorus (refrain) sections of songs in popular-music CDs. Most previous methods detected as a chorus a repeated section of a given length (Logan and Chu, 2000; Cooper and Foote, 2002) and had difficulty identifying both ends of a chorus section and

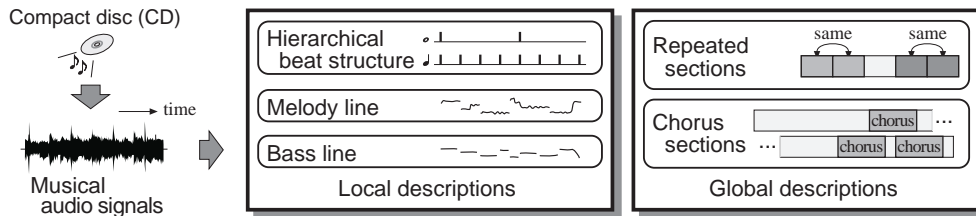


Figure 1: Descriptions in our music-scene-description system.

dealing with modulations (key changes) (Peeters et al., 2002; Dannenberg and Hu, 2002). By analyzing relationships between various repeated sections, RefraiD can detect all the chorus sections in a song and identify both ends of each section. It can also detect modulated chorus sections by introducing a similarity measure that enables modulated repetition to be judged correctly.

3 Hand-Labeling the Music Scene Descriptions on a Metadata Editor

To evaluate automatic music-scene-description methods, we have been working on labeling the pieces in a music database with their correct descriptions (metadata). We have therefore developed a multipurpose music-scene labeling editor (metadata editor) that enables a user to hand-label a musical piece with descriptions such as hierarchical beat structure, melody and bass lines, and chorus sections. It can deal with both audio files and standard MIDI files and it supports interactive audio/MIDI playback while editing. Along a wave or MIDI-piano-roll display it shows subwindows in which any selected descriptions can be displayed and edited. It also supports practical editing aids such as a magnifying-glass function, a region-based cut-and-paste operation, and cursor movement between context-dependent grid points.

The editor has been ported on several operating systems (Linux, SGI IRIX, and Microsoft Windows). To facilitate the support of other descriptions in the future, its architecture is based on a plug-in system in which an external module for editing each description is installed as plug-in software.

Using the editor, we hand-labeled the chorus sections of all 100 songs of the *RWC Music Database: Popular Music* (Goto et al., 2002) and evaluated the RefraiD. By comparing the output of the RefraiD with those hand-labeled chorus sections, we found that the correct chorus sections had been detected in 80 of the 100 songs. We are also working on labeling the songs in the database with other descriptions.

4 Conclusion

We have described the *Music Scene Description Project* in which we are building a music-scene-description system that understands real-world musical audio signals without deriving musical scores or separating signals and are also developing a metadata editor that enables a user to hand-label audio files and standard MIDI files with descriptions of the music in those files. We have already implemented real-time methods for tracking beats, detecting melody and bass lines, and finding chorus sections. We have also hand-labeled the popular songs in the RWC Music Database with their chorus sections and used them to evaluate our chorus-section detection method.

Because our automatic description methods are useful for obtaining various metadata for music information retrieval, we plan to build a retrieval system based on such metadata. Other future work will include improving the performance of the automatic description methods, supporting other music-scene descriptions, and hand-labeling the RWC Music Database with other kinds of metadata.

Acknowledgments

This project has been funded by “Information and Human Activity,” Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Corporation (JST) during October 2000 and September 2003.

References

- Casey, M. A. and Westner, A. (2000). Separation of mixed audio sources by independent subspace analysis. In *Proc. of ICMC 2000*, pp. 154–161.
- Cooper, M. and Foote, J. (2002). Automatic music summarization via similarity analysis. In *Proc. of ISMIR 2002*, pp. 81–85.
- Dannenberg, R. B. and Hu, N. (2002). Pattern discovery techniques for music audio. In *Proc. of ISMIR 2002*, pp. 63–70.
- Goto, M. (2001a). An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, 30(2):159–171.
- Goto, M. (2001b). A predominant-F0 estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models. In *Proc. of ICASSP 2001*, pp. V–3365–3368.
- Goto, M. (2003a). A chorus-section detecting method for musical audio signals. In *Proc. of ICASSP 2003*, pp. V–437–440.
- Goto, M. (2003b). A real-time music scene description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*. (accepted)
- Goto, M. (2003c). SmartMusicKIOSK: Music listening station with chorus-search function. In *Proc. of UIST 2003*. (accepted)
- Goto, M., Hashiguchi, H., Nishimura, T., and Oka, R. (2002). RWC music database: Popular, classical, and jazz music databases. In *Proc. of ISMIR 2002*, pp. 287–288.
- Goto, M. and Muraoka, Y. (1999). Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions. *Speech Communication*, 27(3–4):311–335.
- Katayose, H. and Inokuchi, S. (1989). The Kansei music system. *Computer Music Journal*, 13(4):72–77.
- Klapuri, A., Virtanen, T., Eronen, A., and Seppänen, J. (2001). Automatic transcription of musical recordings. In *Proc. of CRAC-2001*.
- Logan, B. and Chu, S. (2000). Music summarization using key phrases. In *Proc. of ICASSP 2000*, pp. II–749–752.
- Peeters, G., Burthe, A. L., and Rodet, X. (2002). Toward automatic music audio summary generation from signal analysis. In *Proc. of ISMIR 2002*, pp. 94–100.
- Scheirer, E. D. (2000). *Music-Listening Systems*. PhD thesis, MIT.