Application-Specific Music Transcription for Tutoring

Ye Wang and Bingjun Zhang
National University of Singapore

An application-specific, music-transcription approach uses a customized human–computer interface to combine the strengths of humans and computers to enhance music transcription through instrument modeling and multimedia fusion.

Automatic music transcription (AMT) refers to the ability of computers to write note information—such as the pitch, onset time, duration, and source of each sound—after listening to the music. The ability to do AMT effectively is a Holy Grail for researchers working in this field. However, despite decades of research worldwide, a practically applicable, general-purpose transcription system doesn't exist at present.[1,2] It's a problem that has been recognized by other researchers as well.[3,4]

Our application scenario is computer-assisted, musical-instrument tutoring, where the user has to practice most of the time without human coaching. The target users of our initial system are beginning violin students and singing students who are comfortable with computers. After consulting with a professional violin teacher and collaborators from our university's Conservatory of Music, we tried the following three strategies to build a system that can be tested by users.

Application-specific music transcription

AMT is in many ways analogous to automatic speech recognition, which has enjoyed much greater success both academically (in terms of impact) and commercially (in terms of applications). The natural question is how to make AMT more effective. We believe that a use-inspired approach will provide the much-needed steam in the research engine to move AMT closer to practical application.

In contrast to most published work in music transcription, we advocate application-specific music transcription (ASMT), which takes into account the real needs in music education and focuses on the transcribers to satisfy such needs. This approach is different from those that invent a transcription algorithm first without knowing its applications. We believe music transcription, without compelling applications, will probably remain a toy in various research labs and will have little scientific and social impact. If music transcription stays in the research labs, it will affect the academic community adversely in terms of resources (students, research grants, and so forth).

By way of analogy, AMT is like a generalist who can do many things but doesn't have expert-level skill, while ASMT is like a specialist who can do a specific job professionally. We need both generalists and specialists. Given the relatively slow progress in the general-purpose approach of music transcription, we believe it's important to turn to an application-specific approach.

We also believe that necessity is the mother of innovation and music transcription should be inherently application oriented. To facilitate our use-inspired approach, we propose the following three strategies:

- To combine the strengths from both humans and computers for better system performance, a good human–computer interface (HCI) is the key to achieve the synergy. The HCI allows the user interaction and intervention to compensate for the weakness of the computer, although we do envision future AMT will self-adapt operation parameters dynamically for optimized performance and workload balancing.[5] We hope the improved performance and interactivity will help motivate those who need to practice.

- Equipped with a good HCI, we can let the user provide context information, such as which instrument is going to be played. This context information, which is simple for humans to provide but difficult for computers to recognize, allows the system to select a corresponding instrument model

1070-986X/08/$25.00 © 2008 IEEE Published by the IEEE Computer Society
IEEE MultiMedia mult-15-03-wang.3d 17/6/08 02:35:07 0 Cust # wang

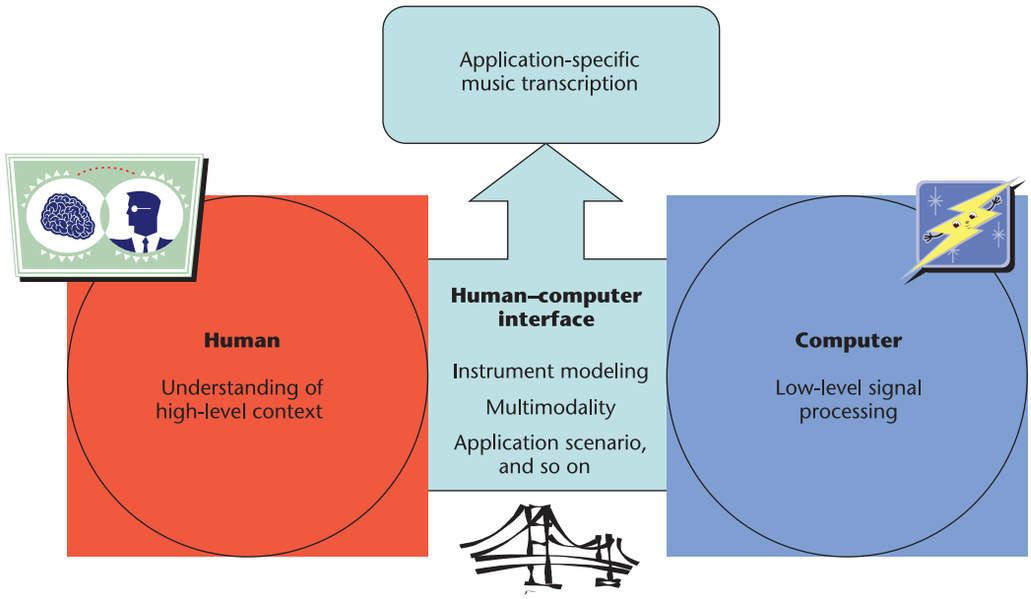

Figure 1. Application-specific music transcription (ASMT) combines the strengths of humans and computers.

and to impose suitable constraints on music transcription, thus simplifying the task significantly.

- Existing educational tools, such as audio and video recordings of teacher's play, don't provide any explicit feedback. If we can leverage the available multimodal information from both audio and video streams to improve transcription accuracy and speed, it's possible to provide the user with the much-needed feedback almost instantaneously.

In our project at the National University of Singapore, we have been trying these strategies and have produced several encouraging results. AMT performance has been significantly improved using the multimodal approach.[6] We intend to create a fairly intelligent learning companion to increase the effectiveness of practice. In addition, with current network technology, we can even envision a networked learning companion that connects students and teachers who are physically remote. This networked learning companion would be similar to the ideas presented in another article.[7]

Interface to combine the strengths of humans and computers

Many researchers generally assume that AMT should be fully automatic. That is, given the recorded acoustic music signal, the standalone system should transcribe the signal without any human intervention. Despite the fact that there are many papers published on AMT (multipitch estimation, blind source separation, beat and rhythm analysis, and so forth), we believe that neglecting users in the design loop is one reason why most existing AMT systems cannot achieve the necessary performance for real-life applications. Existing AMT systems are not capable enough to extract the note information reliably in a real application scenario. Therefore, as shown in Figure 1, instead of a fully automatic approach, we propose a semiautomatic and interactive approach that has the potential to make instrument practice interesting and effective.

ASMT leverages a suitable HCI to combine the strengths of humans and computers. By focusing on a specific application, we can narrow the problem so that we can develop a transcriber with a performance good enough for the particular application, which, in our case, is music education. For example, if the user simply gives the context information (violin, single instrument, and so forth) to the ASMT, the system can impose meaningful constraints to improve the transcription performance significantly. The initial HCI design of our system is shown in Figures 2a and 2b. The idea behind such HCI design is to allow the user to give context information.

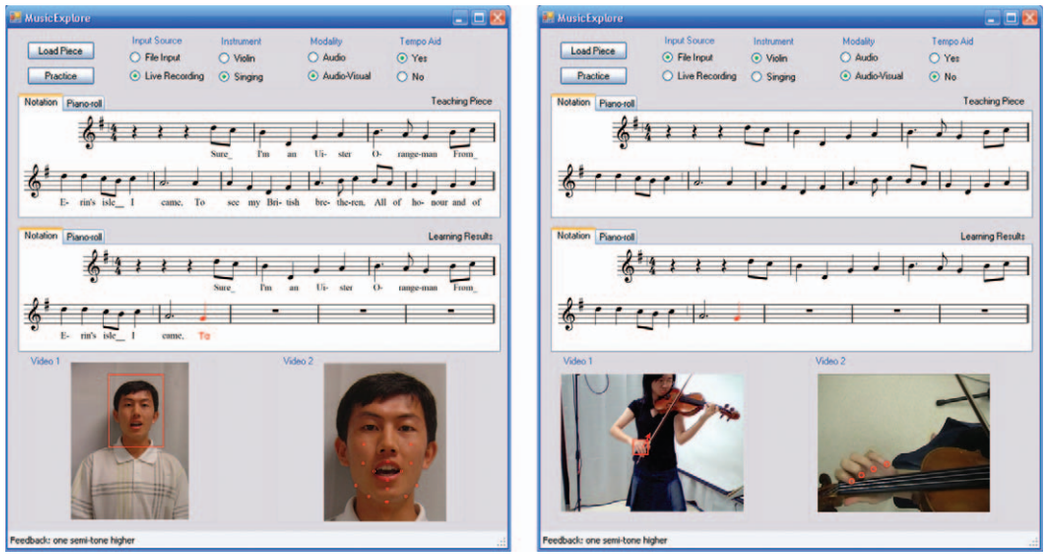

Figure 2. Initial design of the human–computer interface for our system with (a) singing or (b) violin selected.

As Figure 2 shows, our HCI displays the selected score (by selecting load piece) from the database. The system records the user's performance, transcribes it, and aligns it with the score. The user can select (via input source) whether the practice will be transcribed in real time (score following) or offline (score matching).[3] Instrument indicates whether the user is going to practice violin or singing. This information allows the ASMT to select the correct model for the next transcription. In the interface, modality indicates whether to perform audio-only or audiovisual transcription. Tempo aid indicates whether a metronomic sound is generated to assist the user to control the tempo. Transposition allows students to transpose up or down the original melody by a number of semitone levels.

If visual information is used, we analyze facial features—the mouth in particular—to assist singing voice transcription employing a single webcam. For the violin students, we employ two simple webcams to track the right-hand and left-hand fingers for assisting violin transcription.[6,8] We initially employed markers in the hand and finger tracking.[6] However, the test users preferred the convenience of a markerless solution.[8] Although the whole tutoring system is still under research and development, our initial audio-only transcription component[3,9] and audiovisual transcription component for hand tracking and fingering analysis have already produced promising results. A singing tutoring component that includes both audio-only transcription and fusing facial motion information is under development.

Instrument modeling to enhance music transcription

For the projected application, our ASMT is designed to detect pitch values and onset and offset times of pitched nonpercussive sounds, such as from the violin and singing, quickly and accurately. Assuming the user has selected the sound source from a single instrument (for example, violin), the system can use an instrument-modeling approach to exploit the unique characteristics of that instrument and to impose meaningful constraints (frequency range, timbre structure, and so forth). Our preliminary work has shown the potential of this approach in terms of transcription accuracy and speed.[10,11]

We can constrain and train the instrument model beforehand, and can further exploit contextual information that the solo violin sound is mostly a monophonic signal where, at most, one note is sounding at a time with occasional polyphonic signals (a chord) where several notes are played simultaneously. A single instrument is the most common scenario in a daily practice setup. This kind of information is extremely helpful for the computer to perform the transcription task. Furthermore, the fact that we can know beforehand which notes musicians are attempting to sing or play can be employed to simplify the task at hand.

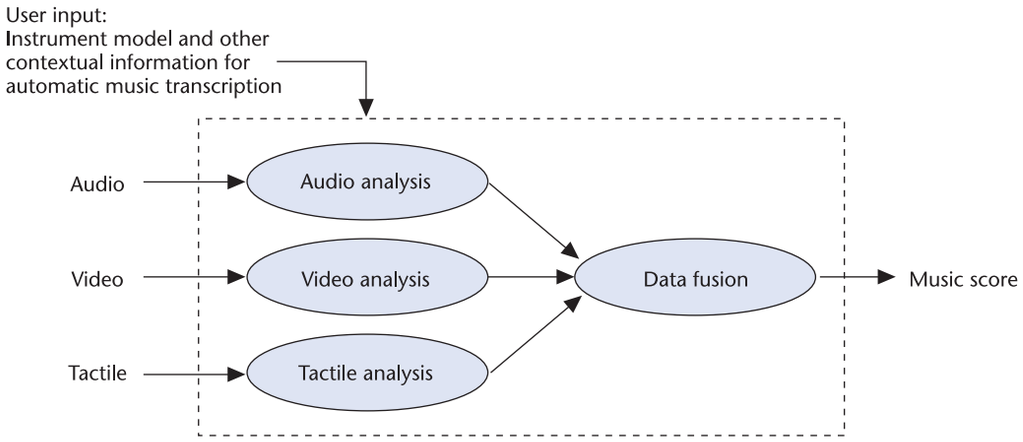

Figure 3. Music transcription leveraging instrument modeling as prior knowledge.

Multimedia fusion to enhance music transcription

Music transcription is traditionally a research topic in the audio domain. However, video (such as that captured by a webcam) is becoming inexpensive and ubiquitous. Researchers have started to leverage such data for music transcription.[6,12,13] Our preliminary work has shown that a multimedia fusion approach can improve onset and offset detection and pitch estimation significantly in comparison with an audio-only approach. We are exploring how to make the system easy to set up by ordinary home users.

We assume that the system must be able to process audio input recorded with low-quality microphones in home environments. Such audio input is typically noisy, which presents one problem. Furthermore, beginners tend to make various mistakes, resulting in an audio signal with irregular patterns. These aspects make it challenging to develop a fast and accurate transcriber. For our application scenario, we focused on three design objectives for the transcriber: accuracy, robustness against noise, and speed. Accuracy is important because an inaccurate transcriber cannot be effective in providing feedback to students. Robustness against noise is important because sound recorded with low-quality microphones in a home environment is usually noisy. And speed is important because students are unlikely to be willing to wait long for feedback. To be useful, the feedback must be almost instantaneous.[14]

To achieve these design objectives, we believe that a multimedia fusion approach, as illustrated in Figure 3, is an attractive alternative to yield satisfactory results. We have successfully shown that by fusing bowing and fingering information into an audio-only approach, we can significantly improve the transcription accuracy in violin tutoring.[6,8,9] In addition, we are researching how to fuse facial motion information into n audio-only singing transcription system. Generally speaking, the facial motion, especially the mouth movement, is indicative of underlying onsets and tempo events during singing. However, we need further research to justify the effectiveness of facial motion in helping singing transcription for our intended tutoring application.

Conclusion and future work

So far, we have only conducted research with the three proposed strategies in a violin-tutoring scenario and obtained some initial results.[6,8,9] Research for singing tutoring, with the same design philosophy, is still under development. There is plenty of room for innovation in designing a simple and pleasant HCI, instrument modeling, and multimedia fusion transcription. Our initial timbre model seems to work well with keyboard instruments, but fails with bowed-string instruments.[11] Specifically, note segmentation in violin and singing sounds seems to be the most challenging task due to the pitched nonpercussive characteristic of those sounds. More sophisticated modeling is needed. Our multimedia fusion presented elsewhere is a simple early approach.[6] It would be interesting to examine the performance difference if we employed model-based or more current fusion strategies.

Considering the availability of multimedia and computer technologies at home, we

believe ASMT represents an exciting new research direction. We are attempting to integrate the three proposed strategies into a Bayesian framework, which is inherently capable of encompassing multimodality and context information naturally and effectively, to yield a high-performance system.[9] We envision that ASMT performed at home will provide music students an entirely new learning experience. It might even help them to compose their own music by simply humming or whistling into their mobile phones. **MM**

Acknowledgments

We thank the guest editors and three anonymous reviewers for their critical comments, which have helped to improve the quality of this article. We also thank our former and current project team members, David Hsu, Terence Sim, Wee Kheng Leow, Jun Yin, Jonathan Boo, Alex Loscos, Olaf Schleusing, Kathleen Koh, and Joyce Quek for their contributions to the project. We thank Steven Halim for his assistance in the initial design of Figure 1. The Singaporean Ministry of Education grant (Workfare Bonus Scheme no. R-252-000-267-112) funded this project.

References

1. A. Klapuri and M. Davy. *Signal Processing Methods for Music Transcription*, Springer, 2006.
2. A. Klapuri, ''Automatic Music Transcription as We Know it Today,'' *J. New Music Research*, vol. 33, no. 3, 2004, pp. 269-282.
3. R. Dannenberg and C. Raphael, ''Music Score Alignment and Computer Accompaniment,'' *Comm. ACM*, vol. 49, no. 8, 2006, pp. 39-43.
4. D. Ellis, ''Extracting Information from Music Audio,'' *Comm. ACM*, vol. 49, no. 8, 2006, pp. 32-37.
5. A. Loscos, Y. Wang, and J. Boo, ''Low Level Descriptors for Automatic Violin Transcription,'' *Proc. Int'l Conf. Music Information Retrieval*, MIT Press, 2006, pp. 164-167.
6. Y. Wang, B. Zhang, and O. Schleusing, ''Educational Violin Transcription by Fusing Multimedia Streams,'' *Proc. ACM Workshop Educational Multimedia and Multimedia Education*, ACM Press, 2007, pp. 57-66.
7. ''How Skype, Podcasts and Broadband are Transforming Language Teaching,'' *The Economist*, June 2007; http://www.economist.com/business/displaystory.cfm?story_id=9304272.
8. B. Zhang et al., ''Visual Analysis of Fingering for Pedagogical Violin Transcription,'' *Proc. ACM Int'l Conf. Multimedia*, ACM Press, 2007, pp. 521-524.
9. B. Zhang, O. Schleusing, and Y. Wang, ''Multimedia Onset Detection within Bayesian Framework for Pitched Non-Percussive Sounds,'' to be published in *Proc. IEEE Int'l Conf. Multimedia and Expo*, IEEE Press, 2008.
10. J. Boo, Y. Wang, and A. Loscos, ''A Violin Music Transcriber for Personalized Learning,'' *Proc. IEEE Int'l Conf. Multimedia and Expo*, IEEE Press, 2006, pp. 2081-2084.
11. J. Yin et al., ''Music Transcription Using an Instrument Model,'' *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, vol. 3, IEEE Press, 2005, pp. 217-210.
12. O. Gillet and G. Richard, ''Automatic Transcription of Drum Sequences Using Audiovisual Features,'' *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, vol. 3, IEEE Press, 2005, pp. 205-208.
13. A. Kapur et al., ''Pedagogical Transcription for Multimodel Sitar Performance,'' *Proc. Int'l Conf. Music Information Retrieval*, MIT Press, 2007, pp. 351-352.
14. J. Yin, Y. Wang, and D. Hsu, ''Digital Violin Tutor: An Integrated System for Beginning Violin Learners,'' *Proc. ACM Int'l Conf. Multimedia*, ACM Press, 2005, pp. 976-985.

Ye Wang is an assistant professor in the department of computer science at the National University of Singapore. His research interests include music transcription and its applications to music education (see http://www.comp.nus.edu.sg/~wangye/). Wang has a PhD in information technology from Tampere University of Technology. Contact him at wangye@comp.nus.edu.sg.

Bingjun Zhang is a PhD candidate at School of Computing, National University of Singapore. His research interests include music transcription, audio feature extraction and modeling, visual motion tracking, and multimodal data fusion. Zhang has a BA in computer science from Tsinghua University. Contact him at bingjun@comp.nus.edu.sg.