Crowdsourcing-Based Web Services for Speech and Music

Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Japan
m.goto [at] aist.go.jp

Abstract
This column introduces two crowdsourcing-based multimedia systems, *PodCastle* (http://en.podcastle.jp for the English version and http://podcastle.jp for the Japanese version) and *Songle* (http://songle.jp). PodCastle and Songle collect voluntary contributions by anonymous users in order to improve the experiences of users listening to speech and music content available on the web. These multimedia systems, implemented as public web services, use automatic speech-recognition and music-understanding technologies to provide content analysis results, such as full-text speech transcriptions and music scene descriptions, that let users enjoy content-based multimedia retrieval and active browsing of speech and music signals without relying on metadata.

When automatic content analysis is used, however, errors are inevitable. PodCastle and Songle therefore provide an efficient error correction interface that let users easily correct errors by selecting from a list of candidate alternatives. Through these corrections, users gain a real sense of contributing for their own benefit and that of others and can be further motivated to contribute by seeing corrections made by other users.

Our web services promote the popularization and use of speech-recognition and music-understanding technologies by raising user awareness. Users can grasp the nature of those technologies just by seeing results obtained when the technologies applied to speech data and songs available on the web.

1. Introduction

Our goal is to provide end users with public web services based on speech recognition, music understanding, signal processing, machine learning, and crowdsourcing so that they can experience the benefits of state-of-the-art research-level technologies. Since the amount of speech and music data available on the web is always increasing, there are growing needs for the retrieval of this data. Unlike text data, however, the speech and music data itself cannot be used as an index for information retrieval. Although metadata or social tags are often put on speech and music, annotations such as categories or topics tend to be broad and insufficient for useful content-based information retrieval [1]. Furthermore, even if users can find their favorite content, listening to it takes time. Content-based active browsing that allows random access to a desired part of the content and facilitates deeper understanding of the content is important for improving the experiences of users listening to speech and music. We therefore developed two web services for speech and music, PodCastle (Figures 1 and 2) and Songle (Figures 3 and 4).

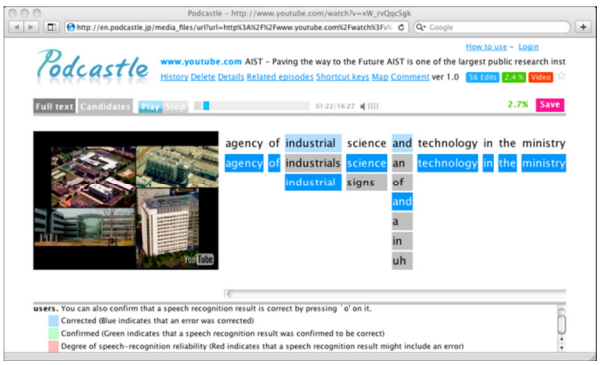


Figure 1. Screen snapshot of PodCastle's interface for correcting speech recognition errors. Competitive candidate alternatives are presented under the recognition results. A user corrected two errors in this excerpt by selecting from the

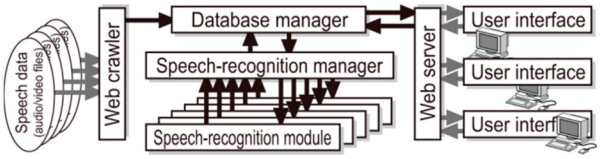


Figure 2. Implementation overview of PodCastle.

2. PodCastle

PodCastle (http://en.podcastle.jp for the English version and http://podcastle.jp for the Japanese version) [3–8, 10, 11] is a spoken document retrieval service that uses automatic speech recognition (ASR) technologies to provide full-text searching of the speech data in podcasts, individual audio or movie files on the web, and the video clips on the video sharing services (*YouTube*, *Nico Nico Douga*, and *Ustream.tv*). PodCastle enables users to find English and Japanese speech data including a search term, read full texts of their recognition results, and easily correct recognition errors by simply selecting from a list of candidate alternatives displayed on an error correction interface. The resulting corrections are used to improve the speech retrieval and recognition performance, and

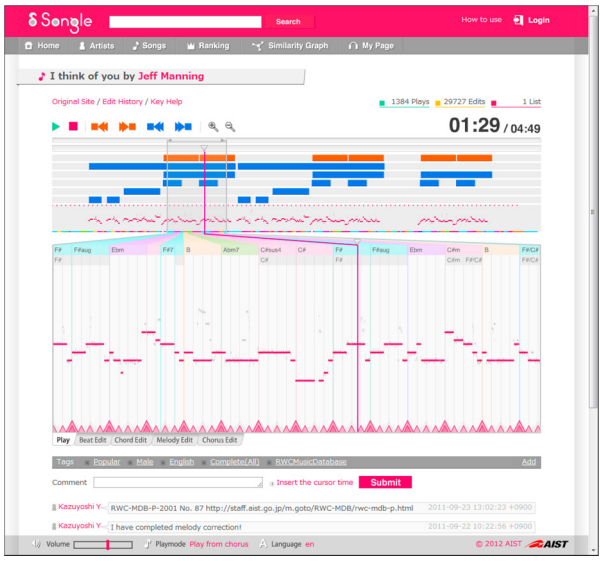


Figure 3. Screen snapshot of Songle's main interface for music playback with the visualization of automatically estimated music scene

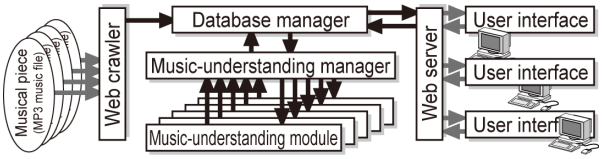


Figure 4. Implementation overview of Songle.

users can actively browse speech data by jumping to any word in the recognition results during playback. In our experience with its use over the past six years (since December 2006), over five hundred ninety thousand recognition errors were corrected by anonymous users and we confirmed that PodCastle's speech recognition performance was improved by those corrections.

3. Songle

Following the success of PodCastle, we launched Songle (http://songle.jp) [7–9], an active music listening service that enriches music listening experiences by using music-understanding technologies based on signal processing. Songle serves as a showcase, demonstrating how people can benefit from music-understanding technologies, by enabling people to experience active music listening interfaces [2] on the web. Songle facilitates deeper understanding of music by visualizing automatically estimated music scene descriptions such as music structure, hierarchical beat structure, melody line, and chords (Figure 3). Users can actively browse music data by jumping to a chorus or repeated section during playback and can use a content-based retrieval function to find music with similar vocal timbres. Songle also features an efficient error correction interface that encourages people to help improve Songle by correcting estimation errors (Figure 5).

4. Conclusion

PodCastle and Songle made academic contributions by demonstrating a new research approach to speech recognition and music understanding based on signal processing; this approach aims to improve the speech-recognition and music-understanding performances as well as the usage rates while benefiting from the cooperation of anonymous end users. This approach is designed to set into motion a *positive spiral* where (1) we enable users to experience a service based on speech recognition or music understanding to let them better understand its performance, (2) users contribute to improving performance, and (3) the improved performance leads to a better user experience, which encourages further use of the service at step (1) of this spiral. This is a *social correction* framework, where users can improve the performance by sharing their correction results over a web service. The game-based approach of Human Computation or GWAPs (games with a purpose) [13] like the ESP Game [14] often lacks step (3) and depends on the feeling of fun. In this framework, users gain a real sense of contributing for their own benefit and that of others and can be further motivated to contribute by seeing corrections made by other users. In this way, we can use the *wisdom of the crowd* or *crowdsourcing* [12] to achieve a better user experience.

Acknowledgments

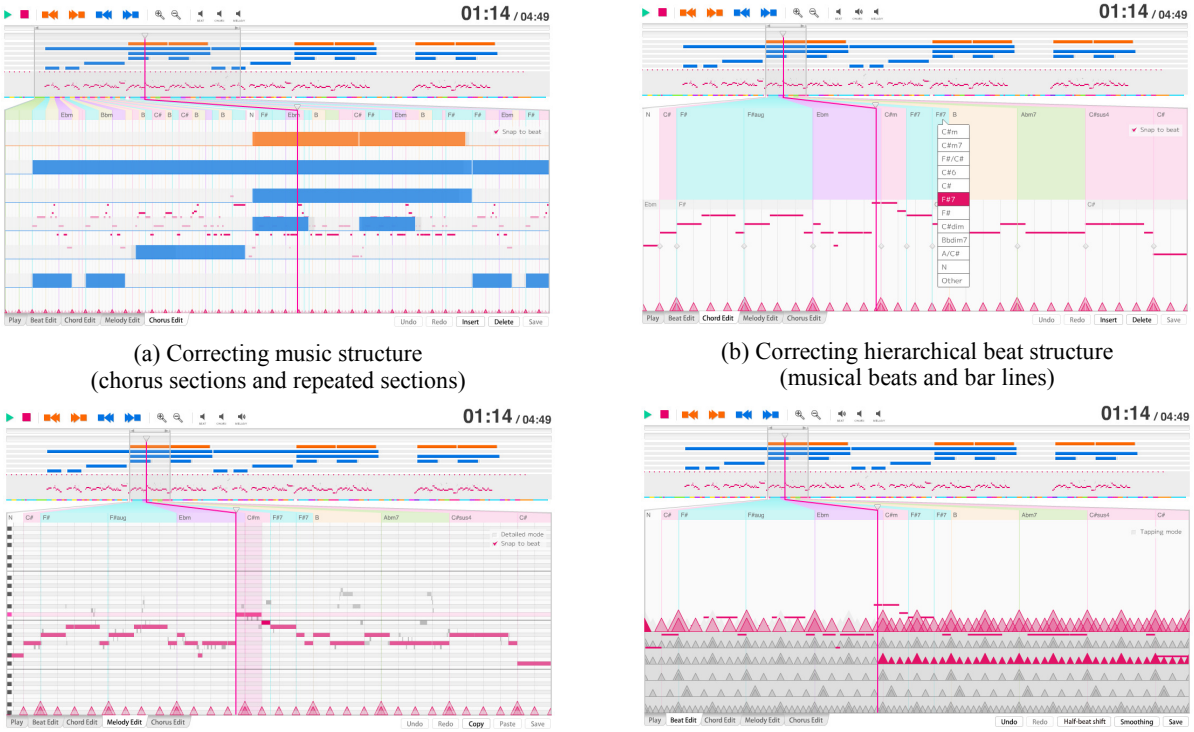
We thank Jun Ogata who collaborates with me for PodCastle, and Kazuyoshi Yoshii, Hiromasa Fujihara, Matthias Mauch, and Tomoyasu Nakano who collaborate with me for Songle. We also thank Youhei Sawada, Shunichi Arai, Kouichirou Eto, and Ryutaro Kamitsu for their web service implementation of PodCastle, Utah Kawasaki for the web service implementation of Songle, and Minoru Sakurai for the web design of PodCastle and Songle. We thank anonymous users of PodCastle and Songle for correcting errors. This work was supported in part by CREST, JST.

References

[1] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. *Content-based music information retrieval: Current directions and future challenges*. Proceedings of the IEEE, 96(4):668-696, 2008.

[2] M. Goto. *Active music listening interfaces based on signal processing*. In Proc. of IEEE ICASSP 2007, 2007.

[3] M. Goto and J. Ogata. **[Invited talk]** *PodCastle: A spoken document retrieval service improved by anonymous user contributions*. In Proc. of PACLIC 24,



(a) Correcting music structure
(chorus sections and repeated sections)

(b) Correcting hierarchical beat structure
(musical beats and bar lines)

(c) Correcting melody line (F0 of the vocal melody)

(d) Correcting chords (root note and chord type)

Figure 5. Screen snapshots of Songle's error correction interface for correcting music scene descriptions.

pages 3-11, 2010.

[4] M. Goto and J. Ogata. **[Invited talk]** *PodCastle: A spoken document retrieval service improved by user contributions*. In Proc. of KJDB 2010, 2010.

[5] M. Goto and J. Ogata. *PodCastle: Recent advances of a spoken document retrieval service improved by anonymous user contributions*. In Proc. of Interspeech 2011, 2011.

[6] M. Goto, J. Ogata, and K. Eto. *PodCastle: A Web 2.0 approach to speech recognition research*. In Proc. of Interspeech 2007, 2007.

[7] M. Goto, J. Ogata, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano. **[Keynote talk]** *PodCastle and Songle: Crowdsourcing-based web services for spoken content retrieval and active music listening*. In Proc. of ACM CrowdMM 2012, pages 1-2, 2012.

[8] M. Goto, J. Ogata, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano. *PodCastle and Songle: Crowdsourcing-based web services for retrieval and browsing of speech and music content*. In Proc. of CrowdSearch 2012, pages 36-41, 2012.

[9] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano. *Songle: A web service for active music listening improved by user contributions*. In Proc. of ISMIR 2011, pages 311-316, 2011.

[10] J. Ogata and M. Goto. *PodCastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription*. In Proc. of Interspeech 2009, pages 1491-1494, 2009.

[11] J. Ogata, M. Goto, and K. Eto. *Automatic transcription for a Web 2.0 service to search podcasts*. In Proc. of Interspeech 2007, 2007.

[12] G. Parent and M. Eskenazi. *Speaking to the Crowd: Looking at past achievements in using crowdsourcing for speech and predicting future challenges*. In Proc. of Interspeech 2011, 2011.

[13] L. von Ahn. *Games with a purpose*. IEEE Computer Magazine, 39(6):92-94, June 2006.

[14] L. von Ahn and L. Dabbish. *Labeling images with a computer game*. In Proc. of ACM CHI 2004, pages 319-326, 2004.

Masataka Goto received the Doctor of Engineering degree from Waseda University in 1998. He is currently a Prime Senior Researcher and the Leader of the Media Interaction Group at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. He serves concurrently as a Visiting Professor at the Institute of Statistical Mathematics, an Associate Professor (Cooperative Graduate School Program) in the Graduate School of Systems and Information Engineering, University of Tsukuba, and a Project Manager of the Exploratory IT Human Resources Project run by the Information Technology Promotion Agency (IPA), Japan. Over the past 20 years, Masataka Goto has published more than 190 papers in refereed

journals and international conferences and has received 31 awards, including several best paper awards, best presentation awards, and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (Young Scientists' Prize).