PodCastle and Songle: Crowdsourcing-Based Web Services for Spoken Content Retrieval and Active Music Listening

Masataka Goto Jun Ogata Kazuyoshi Yoshii
Hiromasa Fujihara Matthias Mauch Tomoyasu Nakano

National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
m.goto [at] aist.go.jp

ABSTRACT

In this keynote talk, we describe two crowdsourcing-based web services, 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 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 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.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services; H.5.5 [**Sound and Music Computing**]: Signal analysis, synthesis, and processing; Systems

Keywords

Multimedia retrieval, web services, spoken content retrieval, active music listening, wisdom of crowds, crowdsourcing

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 (Figure 1) and Songle (Figure 2).

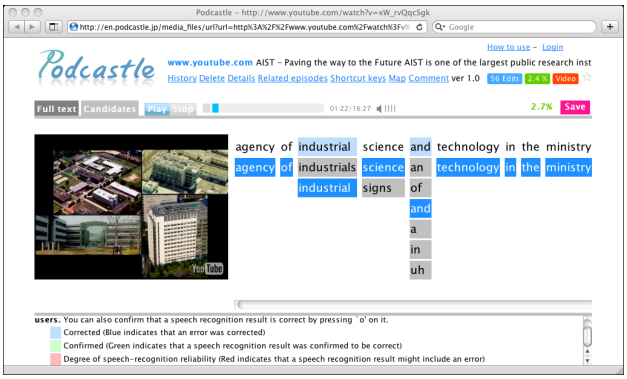

Figure 1: Screen snapshot of PodCastle's interface for correcting speech recognition errors. Competitive candidate alternatives are presented under the recognition results. A user corrected three errors in this excerpt by selecting from the candidates.

2. PODCASTLE

PodCastle (http://en.podcastle.jp for the English version and http://podcastle.jp for the Japanese version) [3–7, 9, 10] 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

CrowdMM'12, October 29, 2012, Nara, Japan.
ACM 978-1-4503-1589-0/12/10.

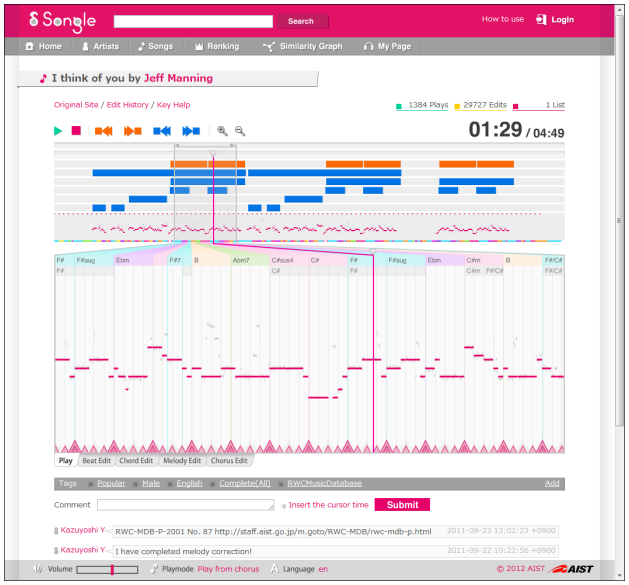

Figure 2: Screen snapshot of Songle's main interface for music playback with the visualization of automatically estimated music scene descriptions.

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 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 five 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, 8], 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 2). 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.

4. CONCLUSION

PodCastle and Songle make 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) [12] like the ESP Game [13] 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* [11] to achieve a better user experience.

Acknowledgments: We 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 also thank anonymous users of PodCastle and Songle for correcting errors. This work was supported in part by CREST, JST.

5. 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 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*, 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. 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.

[8] 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.

[9] 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.

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

[11] 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.

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

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