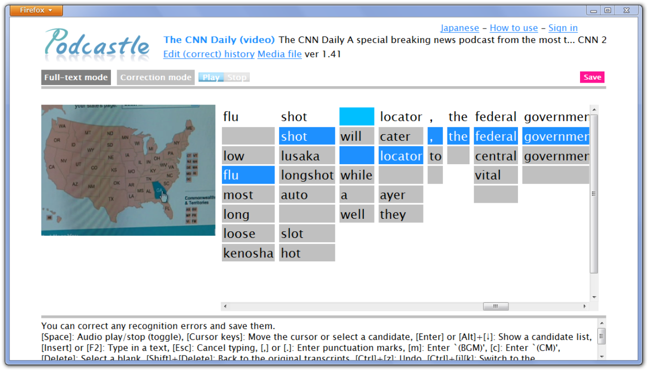
PodCastle and Songle: Crowdsourcing-Based Web Services for Spoken Document Retrieval and Active Music Listening

Masataka Goto, Jun Ogata, Kazuyoshi Yoshii, Hiromasa Fujihara, Matthias Mauch, and Tomoyasu Nakano
National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan



The CNN Daily (video) The CNN Daily A special breaking news podcast from the most t... CNN 2
Edit (correct) history Media file ver 1.41

flu	shot		locator	,	the	federal	governmer
	shot	will	cater	,	the	federal	governmer
low	lusaka		locator	to		central	governmer
flu	longshot	while				vital	
most	auto	a	ayer				
long		well	they				
loose	slot						
kenosha	hot						

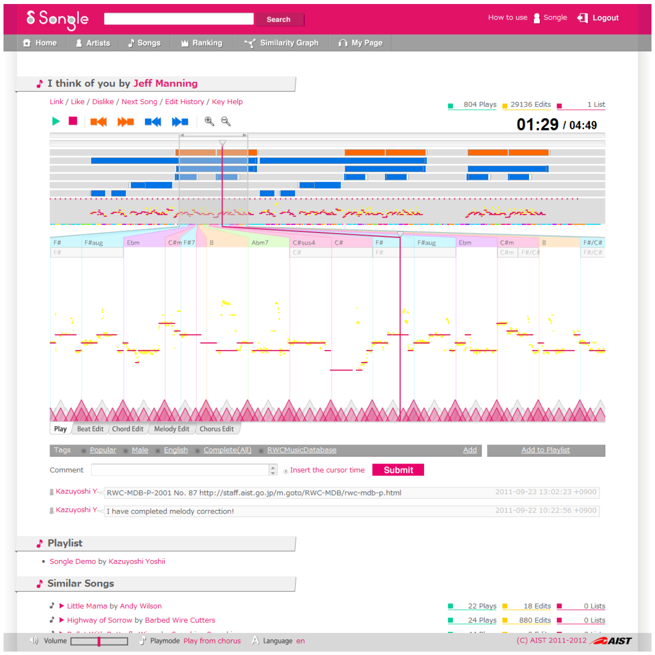
You can correct any recognition errors and save them.
[Space]: Audio play/stop (toggle), [Cursor keys]: Move the cursor or select a candidate, [Enter] or [Alt]+[↓]: Show a candidate list,
[Insert] or [F2]: Type in a text, [Esc]: Cancel typing, [,] or [.]: Enter punctuation marks, [m]: Enter `(BGM)', [c]: Enter `(CM)',
[Delete]: Select a blank, [Shift]+[Delete]: Back to the original transcripts, [Ctrl]+[z]: Undo, [Ctrl]+[i][k]: Switch to the

Fig. 1. PodCastle screen snapshot of an interface for correcting speech recognition errors (competitive candidates are presented underneath the normal recognition results). Three errors in this excerpt were corrected by selecting from the candidates.

Abstract

In this talk, we describe two web services for speech and music, PodCastle (Figure 1) and Songle (Figure 2), that collect and amplify voluntary contributions by anonymous users to improve user experiences. Our goal is to provide users with public web services based on speech recognition, music understanding, signal processing, machine learning, and crowdsourcing so that they can experience state-of-the-art research-level technologies through useful services.

PodCastle (http://en.podcastle.jp) [1]–[7] is a spoken document retrieval service that provides full-text searching of speech data (such as podcasts, individual audio or movie files on the web, video clips on video sharing services YouTube, Nico Nico Douga, and Ustream.tv) on the basis of automatic speech recognition. PodCastle enables users to find speech data (in English or Japanese) that include a search term, read full texts of their recognition results, and easily correct recognition errors by simply selecting from a list of candidates. The resulting corrections were used to improve both the speech retrieval and recognition performances. In our experiences from its practical use over the past five years (since December, 2006), over five hundred eighty thousand recognition errors in about one hundred forty thousand speech data were corrected by anonymous users and we confirmed that the speech recognition performance of PodCastle was actually improved by those corrections.



I think of you by Jeff Manning

Kazuyoshi Y RWC-MDB-P-2001 No. 87 http://staff.aist.go.jp/m.goto/RWC-MDB/rwc-mdb-p.html 2011-09-23 13:02:23 +0900

Kazuyoshi Y I have completed melody correction! 2011-09-22 10:22:56 +0900

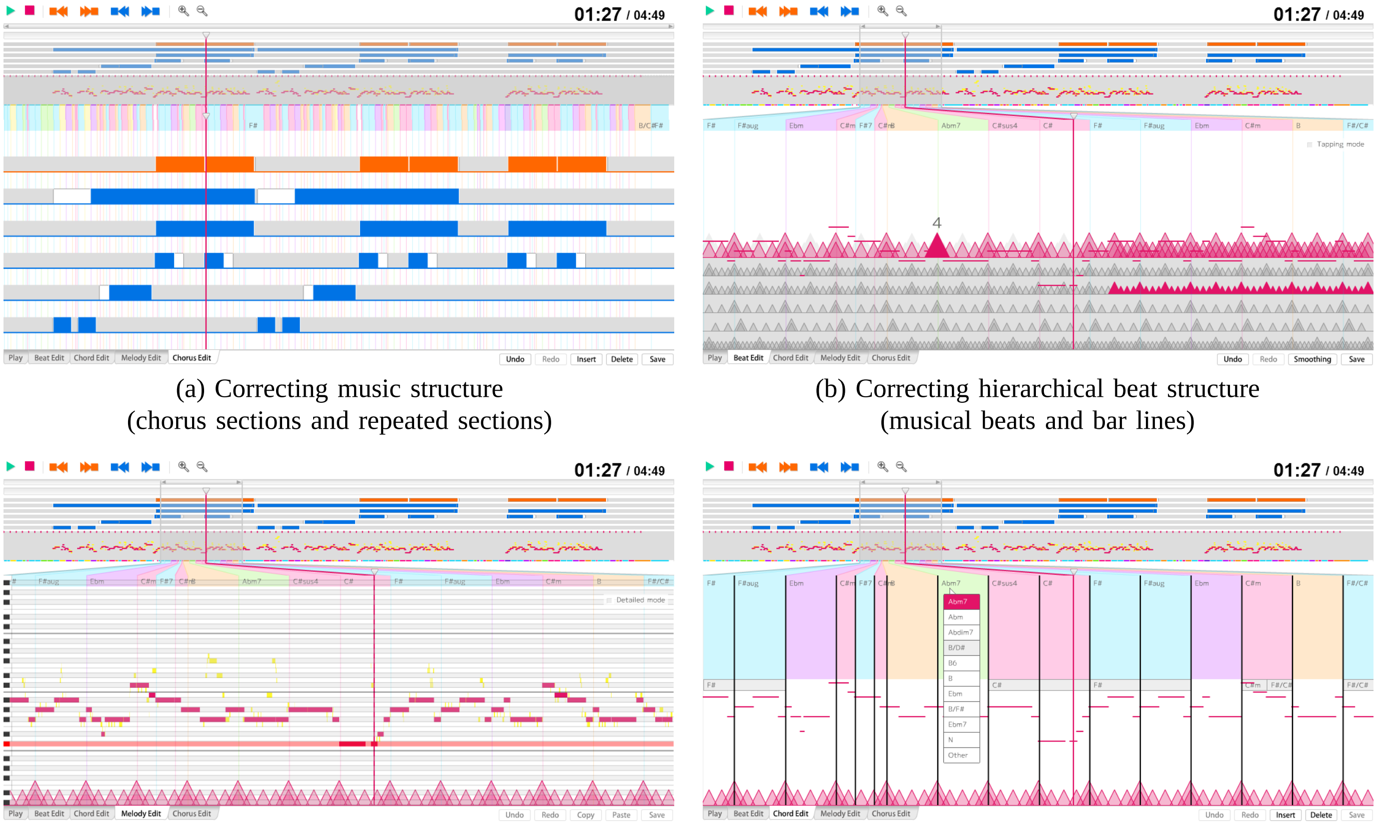
Songle Demo by Kazuyoshi Yoshii

Little Mama by Andy Wilson

Highway of Sorrow by Barbed Wire Cutters

Fig. 2. Songle screen snapshot of the main interface for music playback with the visualization of music scene descriptions estimated automatically.

Following the success of PodCastle, we recently launched Songle (http://songle.jp) [8] that is a public web service for active music listening that enriches music listening experiences by using music-understanding technologies based on signal processing. Songle serves as a showcase to demonstrate how people can benefit from music-understanding technologies by enabling people to experience active music listening interfaces [9] on the web. Songle facilitates deeper understanding of music by visualizing music scene descriptions estimated automatically, such as music structure, hierarchical beat structure, melody line, and chords. When using music-understanding technologies, however, estimation errors are inevitable. Songle therefore features an efficient error correction interface (Figure 3) that encourages people to contribute by correcting those errors to improve the web service. We also propose a mechanism of collaborative training for music-understanding technologies, in which corrected errors will be used to improve the music-understanding performance through machine learning techniques.



(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)

Fig. 3. Songle screen snapshots of the annotation function for correcting music scene descriptions.

The contribution of these studies is to propose a new research approach to speech recognition and music understanding based on signal processing; this approach aims at improving both the speech-recognition and music-understanding performances and 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 improved 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) [10] like the ESP Game [11] 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 crowds or crowdsourcing to achieve a better user experience.

Acknowledgment

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

References

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

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

[3] 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*, pp. 1491–1494, 2009.

[4] J. Ogata and M. Goto, “PodCastle: A spoken document retrieval system for podcasts and its performance improvement by anonymous user contributions,” in *Proc. of SSCS 2009*, pp. 37–38, 2009.

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

[6] M. Goto and J. Ogata, “**[Invited talk]** PodCastle: A spoken document retrieval service improved by anonymous user contributions,” in *Proc. of PACLIC 24*, pp. 3–11, 2010.

[7] 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.

[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*, pp. 311–316, 2011.

[9] M. Goto, “Active music listening interfaces based on signal processing,” in *Proc. of ICASSP 2007*, 2007.

[10] L. von Ahn, “Games with a purpose,” *IEEE Computer Magazine*, vol. 39, pp. 92–94, June 2006.

[11] L. von Ahn and L. Dabbish, “Labeling images with a computer game,” in *Proc. of CHI 2004*, pp. 319–326, 2004.