3M3-OS-07d-1

クラウドソーシングに基づく 能動的音楽鑑賞サービス Songle と音声情報検索サービス PodCastle

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

後藤 真孝 Masataka Goto　吉井 和佳 Kazuyoshi Yoshii　中野 倫靖 Tomoyasu Nakano　緒方 淳 Jun Ogata

産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

In this invited talk, we describe two crowdsourcing-based web services, Songle (http://songle.jp) and PodCastle (http://podcastle.jp). Songle and PodCastle collect voluntary contributions by anonymous users in order to improve the experiences of users listening to music and speech content available on the web. These services use automatic music-understanding and speech-recognition technologies to provide content analysis results, such as music scene descriptions and full-text speech transcriptions, that let users enjoy content-based multimedia retrieval and active browsing of music and speech signals without relying on metadata. When automatic content analysis is used, however, errors are inevitable. Songle and PodCastle therefore provide an efficient error correction interface that lets 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.

1. はじめに

本招待講演ではメディア処理におけるクラウドソーシング利用の先駆事例として，音楽理解技術によって音楽の聴き方を豊かにする能動的音楽鑑賞サービス「Songle（ソングル）」(http://songle.jp) と，音声認識技術によって動画中の音声を書き起こせる音声情報検索サービス「PodCastle（ポッドキャッスル）」(http://podcastle.jp) を紹介する．いずれも計算機による音楽理解あるいは音声認識の誤りを，ユーザが Web 上で訂正できるインタフェースを備えているところが，クラウドソーシングに関連している．そして，不特定多数のユーザによる自発的な訂正をユーザ体験の向上に結びつけていくことで，さらなる利用を促す仕組みを持っている点が大きな特長である．

デジタル化された音楽・音声コンテンツが持つ潜在的な可能性は，まだ充分には引き出されていない．デジタル化がもたらす価値として，膨大な音楽・音声コンテンツをいつでもどこでも聴くことが可能になるという量的な変化は，日常生活で起きた．本研究ではさらに，音楽・音声コンテンツの聴き方や活用のされ方が，より能動的で豊かで便利になる質的な変化をエンドユーザの日常生活で起こすことを，最終的な目的とする．その変化を起こす鍵となるのが，音楽理解技術（音楽の音響信号中の様々な要素を自動的に理解できる技術）と音声認識技術（音声の音響信号を自動的にテキストで書き起こす技術）である．

インターネット上の動画共有サービスや音楽・音声配信サービスの普及に伴い，誰でも視聴できる音楽・音声コンテンツは日常的に生成・蓄積されて増え続けている．しかし，そうしたコンテンツはテキスト（文字）コンテンツと異なり，コンテンツの中身を直接索引として使えないため，音楽の内容や発言内容などに基づく詳細な情報の検索ができなかった．そのため，人手で付与されたアノテーション（書誌情報等のメタデータやソーシャルタグ）による検索が通常利用されているが，コンテンツの内容を充分に反映できているとは限らず，限界があった．また，もし興味のある音楽・音声コンテンツを見つけても，

連絡先: 後藤 真孝，産業技術総合研究所，m.goto [at] aist.go.jp

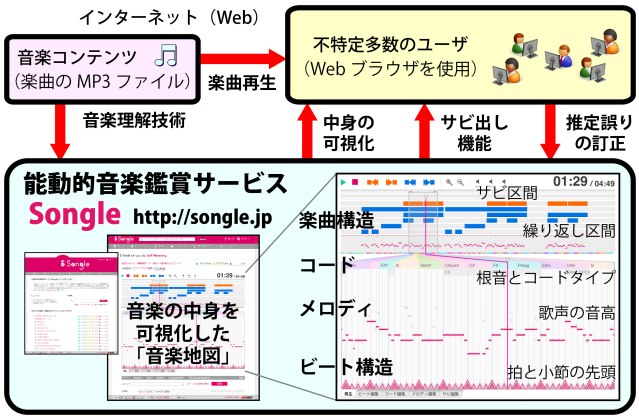

図 1: 音楽理解技術を活用した能動的音楽鑑賞サービス「Songle」

それを実際に再生して聴くのはコンテンツの長さと同じ時間がかかっていた．

そこで本研究では，音楽理解技術や音声認識技術により，人間に代わって計算機が膨大なコンテンツを「聴く」ことで，人間による鑑賞やブラウジングを支援する Web サービスとして，Songle と PodCastle を実現・公開した．これらによりコンテンツの中身（音楽の場合にはサビ，ビート，メロディ，コード，音声の場合には発言内容を書き起こしたテキスト）が可視化されることで，内容を聴く前に興味のある箇所へランダムアクセスしたり，より深くコンテンツを理解したりすることが可能になった．また，コンテンツの中身に基づく検索も可能になった．

こうした音楽理解技術や音声認識技術による自動推定では，誤りが不可避である．そこで効率的な誤り訂正インタフェースを Web 上で提供し，誤りを人手で訂正するというクラウドソーシングにより，金銭的な報酬のない自発的な貢献を促している．それにより，自動推定結果は完全でなくても，その誤り訂正が他のユーザと共有されていくことで，ユーザ体験が向上してユーザが増え，さらに訂正が増えるポジティブスパイラルを回す仕組みが実現できる．

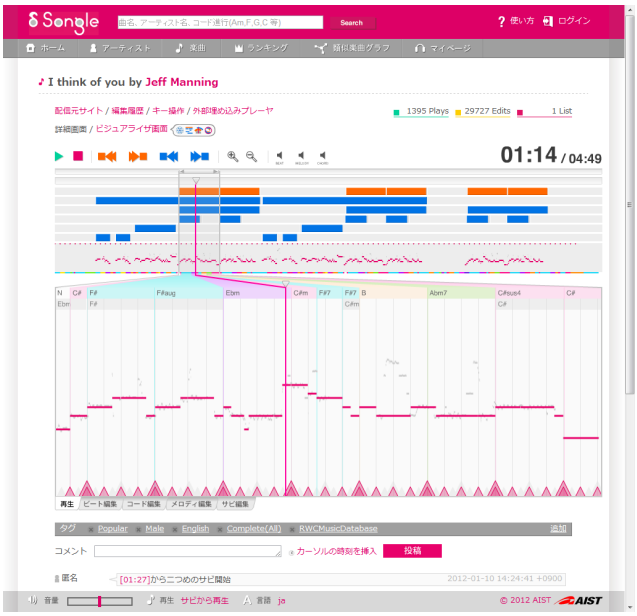

図 2: 楽曲の中身を「音楽地図」として可視化した詳細画面

2. 能動的音楽鑑賞サービス Songle

能動的音楽鑑賞サービス「Songle」[Goto 11b, Goto 12a, Goto 12b, 後藤 13]（図 1）は，音楽理解技術を用いて，Web 上で公開されている任意の音楽コンテンツ（MP3 形式の音楽音響信号ファイルの楽曲，ピアプロおよび SoundCloud 上の楽曲）中の様々な音楽情景記述（音楽的要素）[Goto 03a, Goto 04] を推定する．現在の実装では，歌声を伴うポピュラー音楽を主な対象として，

1. 楽曲構造（サビ区間と繰り返し区間）
2. 階層的なビート構造（拍と小節の先頭）
3. メロディライン（メロディの歌声の基本周波数 (F0)）
4. コード（根音とコードタイプ（構成音））

の 4 つの代表的な音楽情景記述を自動推定し，「音楽地図」として可視化して音楽内容に基づくブラウジングを可能にした．

ユーザが Songle に登録された楽曲を選ぶと，自動推定結果をさまざまな形式で可視化した画面を見ながら，元の Web サイト上にある楽曲をストリーミング再生して楽しむことができる．可視化画面は，ユーザが音楽的要素を把握しやすい「音楽地図」を表示する詳細画面（図 2）と，再生した楽曲の進行に連動したさまざまなアニメーションを表示するビジュアライザ画面（図 3）の 2 種類がある．これらの可視化により，専門的知識のないユーザでも，各音楽的要素の存在や要素間の関係，楽曲構成上の意図に気づきやすくなる．例えば，サビの繰り返しやイントロとエンディングの繰り返しなどの楽曲全体の構造を把握したり（サビが例外的に多く繰り返す曲や，サビから始まる曲に容易に気づくことができる），同じハーモニー（コード進行）なのにメロディが変化する様子に気づいたり，繰り返すときの歌詞や曲調の変化を聴き比べたりすることもできる．このように，再生に同期して推定結果を「見る」ことで音楽の理解を深めることができる．

さらに Songle では，自動推定結果を利用することで，可視化以外にも音楽鑑賞をより能動的で豊かにする以下のような機能を提供している．

- 楽曲中で一番代表的な盛り上がる主題の部分である「サビ」のように，楽曲中の興味のある箇所を容易に見つけて聴くことができるサビ出し機能を備えている．これは，サビ出し機能付き能動的音楽鑑賞インタフェース「SmartMusicKIOSK」[後藤 03b, Goto 06] の機能を Web 上で実現したものである．通常の再生，停止ボタンだけでなく，楽曲構造に対応した「次・前のサビ区間の頭出し」,「次・前の繰り返し区間の頭出し」ボタンが使用できる．本来音楽は全て聴き終わらなければどんな楽曲でサビはどこに出てくるのかがわからないが,「音楽地図」によって楽曲を聴く前に構造を把握することができ，興味のある区間を直接クリックして再生するようなランダムアクセスが可能となった．

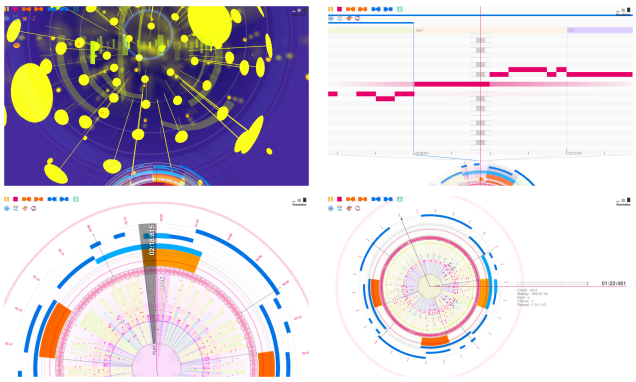

図 3: 楽曲の中身をアニメーション表示するビジュアライザ画面

- 同一のコード進行をもつ複数の楽曲を聴き比べることができるコード進行検索機能を実現した．曲名やアーティスト名といった書誌情報に基づく従来の音楽情報検索に加えて，新たに，コード名の系列を与えるとそれをコード進行として含む楽曲群を検索・列挙する音楽情報検索が可能となった．
- ユーザが自分のホームページやブログなどの外部の Web ページ内に Songle の小型プレーヤを埋め込んで，Songle 上の楽曲を紹介できる外部埋め込みプレーヤ機能を実現した．このプレーヤは楽曲構造の可視化機能と上記のサビ出し機能を備えており，その Web ページを閲覧した人が手軽に試聴しながら Songle を知ることができる．曲名をクリックすれば，Songle 上のその楽曲のページに直接アクセスして利用することができる．

Songle では，音楽理解技術が不充分であっても，ユーザの貢献によってユーザ自身が利便性を感じられる仕組みの実現を目指し，音楽情景記述の推定誤りを容易に訂正して貢献可能なインタフェースを Web 上で提供している．Songle のユーザは推定誤りを見つけたら，自動生成された候補から選んだり，直接編集したりして自発的に訂正する．その結果は他のユーザと共有されて，即座にユーザ体験の向上に資することができる．具体的には，音楽再生に合わせてビートやコード，メロディだけをその場で選択・可聴化する機能を提供し，ユーザが自動推定の誤りに気づきやすくした．誤りが訂正されると元の自動推定結果は違う色で着色され，履歴が残るような機能も付加してある．これには音楽理解技術の性能が過大評価されるのを防ぐ効果もある．

3. 音声情報検索サービス PodCastle

音声情報検索サービス「PodCastle」[Goto 07b, Goto 10a, Goto 10b, 後藤 10c, Goto 11a, Goto 12b]（図 4）は，音声認識技術を

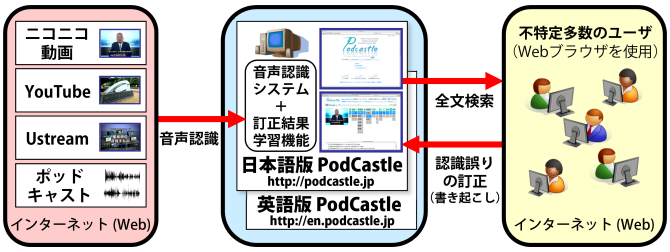

図 4: 音声認識技術を活用した音声情報検索サービス「PodCastle」

用いて，Web 上で公開されている任意の音声コンテンツ（動画共有サービスやポッドキャスト等によって公開されている音声音響信号を伴うコンテンツ）中の音声を認識して自動的にテキストに書き起こす．現在の実装では，代表的な動画共有サービス（ニコニコ動画，YouTube，Ustream）上の音声を含む動画と，RSS で配信されるポッドキャスト，任意の URL にある音声を含む動画ファイルや MP3 形式の音声音響信号ファイルに対応し，日本語と英語を認識可能である．

PodCastle のユーザは，任意の検索語を入力すれば，それを含む音声コンテンツ中の発言を全文検索できる．そして，音声認識結果の書き起こしを Web ブラウザ上で閲覧しながら，元の Web サイト上にある音声コンテンツをストリーミング再生して視聴できる．

PodCastle でも Songle 同様に，音声認識技術が不充分であっても，ユーザの貢献によってユーザ自身が利便性を感じられる仕組みの実現を目指し，音声認識誤りを容易に訂正して貢献可能なインタフェース（図 5）を Web 上で提供している．PodCastle のユーザは認識誤りを見つけたら，自動生成された候補から選んだり，直接テキストを入力して編集したりして自発的に訂正する．その結果は他のユーザと共有されて，即座にユーザ体験の向上に資することができる．また，インターネット上のニュース記事や辞書等から新しい言葉（新語，時事用語，芸能人名等）を自動学習する機能も備えている．

さらに，単なる訂正ではなく，複数のユーザが協調して，読みやすいテキストとして円滑に作成できる書き起こし支援機能を充実させた．例えば，テキスト中の任意の箇所に話者名と改行の入力を可能にし，可読性を向上することができる．同じ音声コンテンツ中の異なる箇所を，複数のユーザが同時に書き起こしていると，お互いの訂正が自動反映されて着色されるので，どこを訂正したかが容易に把握できる．また，ユーザが訂正するだけでなく，音声認識結果の正しい箇所に正解マークを着色することも可能にした．これにより，そうした正しい箇所とまだ訂正されていない箇所を区別することができ，書き起こしの進捗状況を把握しやすい．同一ユーザが後日続きを書き起こす場合にも有用であるし，全体の単語数の何%が書き起こされたか（訂正あるいは正解マーク付与されたか）も達成率として表示できる．

2006 年 12 月から研究者向けに試験公開し，2008 年 6 月に一般公開して実証実験をしてきたが，既に 22 万件以上の音声コンテンツが登録され，その一部の音声認識結果に対して，不特定多数のユーザにより，累計 63 万箇所以上の多数の訂正がなされたことで，音声検索性能が向上した．さらに，訂正結果を言語的・音響的に学習することで，音声認識性能の向上が可能なことも我々は実証した．

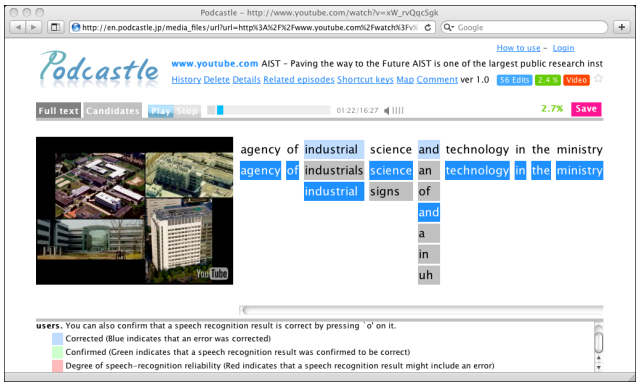

図 5: 英語の音声コンテンツに対応した PodCastle の訂正インタフェース画面

4. おわりに

本研究は，音楽理解技術に基づいて能動的音楽鑑賞インタフェースを楽しむための Web 上のサービスと，音声認識技術に基づいて音声コンテンツの全文検索・書き起こしが可能な Web 上のサービスを公開・運用して，エンドユーザの役に立つという社会的意義を持っている．音楽理解技術（あるいは能動的音楽鑑賞インタフェース [Goto 07a]）は，既に普及しつつある音声認識技術（あるいは音声コンテンツ検索技術 [Larson 12]）や画像理解技術と違い，そもそもそういう技術があるということ自体の認知度が高くなく，Songle によって音楽理解技術の潜在的な可能性が認知され，他の応用事例開拓に波及する効果も期待できる．また，音楽理解技術で Web 上の様々な楽曲に対して推定した結果をユーザが見れば，どのような箇所で推定が難しいかがわかる．音声認識技術でも同様に，背景音のない丁寧な発声の音声に対する高い性能を確認できる一方で，どのような音声だと認識が難しいかがわかる．そこで推定結果に誤りが多い場合には，批判を受ける可能性はあるが，そうした現状をユーザと共有してはじめて，音楽理解技術および音声認識技術の真の普及と発展があると我々は考える．

本研究の学術的意義は，不特定多数のエンドユーザに誤り訂正の協力をしてもらうことで，サービスの利便性とユーザによる利用率をどこまで向上できるかを探求することにある．こうした発想は，従来の音楽理解研究・音声認識研究にはなかった．この新たな研究アプローチでは，

(i) ユーザが音楽理解技術・音声認識技術に基づくサービスを利用することでその性能を理解する

(ii) そのサービス改善にユーザが貢献する

(iii) その改善がより良いユーザ体験に結びつく

という三段階から成る「ポジティブスパイラル」を回すことができる点が重要である．(iii) のユーザ体験の向上が，(i) のサービス利用を促進するからである．従来の GWAP (game with a purpose) や人間計算 (human computation) [Ahn 06]（ESP Game [Ahn 04] も含む）といったゲームの楽しさをインセンティブとしたクラウドソーシングのアプローチでは，この (iii) という重要な段階が欠けていた．金銭的な報酬を伴う多くのクラウドソーシングのアプローチでも，同様である．

Songle と PodCastle は，多数のユーザの訂正結果を Web サービス上で共有して性能改善を図る「社会的訂正」の枠組みであり，貢献するとサービスが改善して自分を含む他のユーザの役に立てるということを明確に意識できる上に，他のユーザが訂

正している活動を見ることで，訂正の意欲も高まる点が優れている．このように Songle と PodCastle では，集合知 (wisdom of crowds) やクラウドソーシング [Parent 11, Baeza-Yates 12, Jones 13] を活用しつつ，ユーザ体験向上を実現していく点が重要である．ただし，訂正は利他的に貢献したいという動機からだけではなく，訂正されている状態の方が単に自分が便利だという動機，好きなコンテンツや自作のコンテンツを訂正されている状態にしたいという動機，訂正操作自体が面白いという動機等，様々な理由で訂正がなされていると考えられる．

本研究のさらなる意義は，ユーザによる誤り訂正の協力で，音楽理解技術・音声認識技術の性能をどこまで高くできるかを探求していくことにある．PodCastle では，日々の訂正結果を機械学習して音声認識性能も向上させていくことに成功し，「ユーザの貢献を増幅」する新たな枠組みを実現した点が，通常の Web 2.0 にはない「PodCastle ならでは」の大きな特長となっている．例えば，Wikipedia 等の典型的な Web 2.0 の Web サービスでは，通常，ユーザの貢献は編集した項目に限定され，自動的に他の項目へ波及して改善されることはない．それに対して PodCastle では，その訂正内容を学習することで，まだ訂正していない部分や他のコンテンツに対する認識結果が改善されるという技術を初めて実現した．この「ユーザの貢献を増幅して性能向上へ繋げる技術」（ユーザ貢献増幅技術）こそが，PodCastle 以前の Web 2.0 や人間計算にはなかった特長であり，ユーザが貢献（訂正）していない箇所へ波及して改善される点が重要である．これは，ユーザに「音声認識を育ててもらう」アプローチと位置づけることもできる．Songle においても，一部の音楽的要素については，訂正結果の機械学習により自動的に性能が向上する機能に取り組んでおり，「ユーザの貢献を増幅」する新たな音楽情報処理の枠組みとして，その可能性を実証していく予定である．

我々は「ユーザを信頼する」立場から，基本的にはユーザによる訂正の質は高いものと考えている．仮にユーザが故意に不適切な訂正（いたずら）をした場合でも，その信頼性（訂正が楽曲や音声の内容と合致するか）を音響的に検証する方法が実現できる可能性があり，新たな研究課題として興味深い．また，Songle や PodCastle 上で不適切な訂正に誰かが気づけば，誰でもその前の状態に戻すことが可能な機能も提供している．

現在の音楽理解技術や音声認識技術による推定結果には誤りが含まれるが，人間が一生かけても聴ききれないような多量のコンテンツを処理できる利点を持つ．一方，人間はコンテンツの内容をより深く理解・認識して記述でき，推定誤りにも気づくことができるが，何もないところからすべてを記述するのは長時間を要し限界がある．そこで両者が相補的に力を合わせることで，より的確にコンテンツの中身を記述できるようにした．このようにユーザ貢献を積極的に取り込んでユーザ体験を向上させるアプローチは，大規模なコンテンツを扱う上で本質的であり，多くの研究者が取り組むことで，その重要性と将来性がさらに明らかになり，今後の音楽理解・音声認識の研究分野に新たな展開を引き起こすことができればと願っている．

謝辞: Songle の Web サービスの実装に協力頂いた川崎 裕太氏，PodCastle の Web サービスの実装に協力頂いた沢田 洋平氏，新井 俊一 氏，江渡 浩一郎 氏，上津 竜太郎 氏，両サービスの Web デザインに協力頂いた櫻井 稔 氏に感謝する．英語版 PodCastle では，エジンバラ大学音声技術研究所 (CSTR) が実施した EU FP6 AMI および FP6 AMIDA で開発された英語用の音声認識システムを，同研究所が PodCastle 用に運用している．本研究の一部は JST CREST の支援を受けた．

参考文献

[Ahn 04] Ahn, von L. and Dabbish, L.: Labeling Images with a Computer Game, in *Proc. of CHI 2004*, pp. 319–326 (2004)

[Ahn 06] Ahn, von L.: Games With A Purpose, *IEEE Computer Magazine*, Vol. 39, No. 6, pp. 92–94 (2006)

[Baeza-Yates 12] Baeza-Yates, R., Ceri, S., Fraternali, P., and Giunchiglia, F. eds.: *Proc. of the First International Workshop on Crowdsourcing Web Search (CrowdSearch 2012)* (2012)

[Goto 03a] Goto, M.: Music Scene Description Project: Toward Audio-based Real-time Music Understanding, in *Proc. of ISMIR 2003*, pp. 231–232 (2003)

[後藤 03b] 後藤 真孝 : SmartMusicKIOSK: サビ出し機能付き音楽試聴機, 情報処理学会インタラクション 2003 論文集, pp. 9–16 (2003)

[Goto 04] Goto, M.: A Real-time Music Scene Description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-world Audio Signals, *Speech Communication*, Vol. 43, No. 4, pp. 311–329 (2004)

[Goto 06] Goto, M.: A Chorus-Section Detection Method for Musical Audio Signals and Its Application to a Music Listening Station, *IEEE Trans. on ASLP*, Vol. 14, No. 5, pp. 1783–1794 (2006)

[Goto 07a] Goto, M.: Active Music Listening Interfaces Based on Signal Processing, in *Proc. of ICASSP 2007* (2007)

[Goto 07b] Goto, M., Ogata, J., and Eto, K.: PodCastle: A Web 2.0 Approach to Speech Recognition Research, in *Proc. of Interspeech 2007* (2007)

[Goto 10a] Goto, M. and Ogata, J.: **[Invited talk]** PodCastle: A Spoken Document Retrieval Service Improved by Anonymous User Contributions, in *Proc. of PACLIC 24*, pp. 3–11 (2010)

[Goto 10b] Goto, M. and Ogata, J.: **[Invited talk]** PodCastle: A Spoken Document Retrieval Service Improved by User Contributions, in *Proc. of KJDB 2010* (2010)

[後藤 10c] 後藤 真孝, 緒方 淳, 江渡 浩一郎 : PodCastle: ユーザ貢献により性能が向上する音声情報検索システム, 人工知能学会誌, Vol. 25, No. 1, pp. 104–113 (2010)

[Goto 11a] Goto, M. and Ogata, J.: PodCastle: Recent Advances of a Spoken Document Retrieval Service Improved by Anonymous User Contributions, in *Proc. of Interspeech 2011* (2011)

[Goto 11b] Goto, M., Yoshii, K., Fujihara, H., Mauch, M., and Nakano, T.: Songle: A Web Service for Active Music Listening Improved by User Contributions, in *Proc. of ISMIR 2011*, pp. 311–316 (2011)

[Goto 12a] Goto, M., Ogata, J., Yoshii, K., Fujihara, H., Mauch, M., and Nakano, T.: **[Keynote talk]** PodCastle and Songle: Crowdsourcing-Based Web Services for Spoken Content Retrieval and Active Music Listening, in *Proc. of the 2012 ACM Workshop on Crowdsourcing for Multimedia (CrowdMM 2012)*, pp. 1–2 (2012)

[Goto 12b] Goto, M., Ogata, J., Yoshii, K., Fujihara, H., Mauch, M., and Nakano, T.: PodCastle and Songle: Crowdsourcing-Based Web Services for Retrieval and Browsing of Speech and Music Content, in *Proc. of the First International Workshop on Crowdsourcing Web Search (CrowdSearch 2012)*, pp. 36–41 (2012)

[後藤 13] 後藤 真孝, 吉井 和佳, 藤原 弘将, Mauch, M., 中野 倫靖 : Songle: 音楽音響信号理解技術とユーザによる誤り訂正に基づく能動的音楽鑑賞サービス, 情報処理学会論文誌, Vol. 54, No. 4 (2013)

[Jones 13] Jones, G. J. F.: An Introduction to Crowdsourcing for Language and Multimedia Technology Research, in *Information Retrieval Meets Information Visualization*, Vol. 7757 of *Lecture Notes in Computer Science*, pp. 132–154, Springer Berlin Heidelberg (2013)

[Larson 12] Larson, M. and Jones, G. J. F.: Spoken Content Retrieval: A Survey of Techniques and Technologies, *Foundations and Trends in Information Retrieval*, Vol. 5, No. 4–5, pp. 235–422 (2012)

[Parent 11] Parent, G. and Eskenazi, M.: Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges, in *Proc. of Interspeech 2011* (2011)