歌詞に基づく歌声アノテーションのためのインタフェース構築

中野 倫靖 * 加藤 淳 * 渡邉 研斗 * 濱崎 雅弘 * 後藤 真孝 *

概要. 本稿では、歌声に対する時間局所的なアノテーションを行う際に、その歌詞を用いるインタラクションを提案する。従来、時系列メディアのアノテーションでは、アノテーション内容に時刻情報を含める強ラベルと、時刻情報が含まれない弱ラベルを基本として、それらの派生や改善が提案されてきた。本研究では、歌詞の文節を選択するだけでその時刻情報を指定できて、簡単にアノテーションできる「歌詞ベース」のアノテーションを提案する。歌詞ベースのアノテーションでは、その音源を再生するプレーヤと、既存のテキストエディタや Excel 等のスプレッドシートがあれば可能であるので、本稿ではまず、Excel をアノテーションエディタとして用いて実際にセマンティックタグをアノテーションした結果を分析することで、実用性を検証する。そしてさらに、その使いやすさを向上するためのインタフェースとして、Lyrics-Based Singing Annotator を提案する。本インタフェースでは、クリック可能な歌詞と音源を同期して再生する機能、付与対象の歌詞をループ再生する機能、特定のタグが付与された歌詞をハイライトする機能を持つ。

1 はじめに

音楽へのアノテーションは、機械学習の学習データとしての利用や、音楽の特性分析などのために重要である。音楽や音声などの時系列メディアへのアノテーションでは、ジャンルや歌手名などの楽曲固有で時不変なラベルだけでなく、音響イベント・歌唱テクニック・セマンティックタグなどの時変なラベルが存在する。これら二種類のラベルは、時刻情報付きラベルが強（Strong）ラベル、時刻情報なしラベルが弱（Weak）ラベルと呼ばれている（図 1）。

機械学習データにおいては弱ラベルが広く利用されることが多い [24, 31, 43]。個人ごとの結果の変動を吸収するためには複数人によるのアノテーションが必要となるが、強ラベルは困難かつ時間消費の大きい高コストな作業であり、多くのデータを収集するのが難しいためである。例えば、音源の一部分（例えば、10 秒間）を切り出し、その区間にラベルの存在のみをアノテーションする弱ラベルは、強ラベルより時刻的な精度は低下するが、相対的に低コストで収集が容易であるために採用される。

しかし、時刻情報を含むラベルは依然として重要である。音響イベント検出等、音の種類と時刻を同時に推定するタスクでは、評価のために時刻情報が必要となるためである [30]。さらに、自動分類のための機械学習モデルにおいて、強ラベルを用いることによる性能向上が報告されている [14]。そこで、弱ラベルのようにアノテーションコストを下げながら、時刻情報の正確性を上げる方法として、点（Point）ラベル [22] と重複弱ラベル（Overlapping weakly-label）[29] が提案された（図 1）。前者の点ラベル

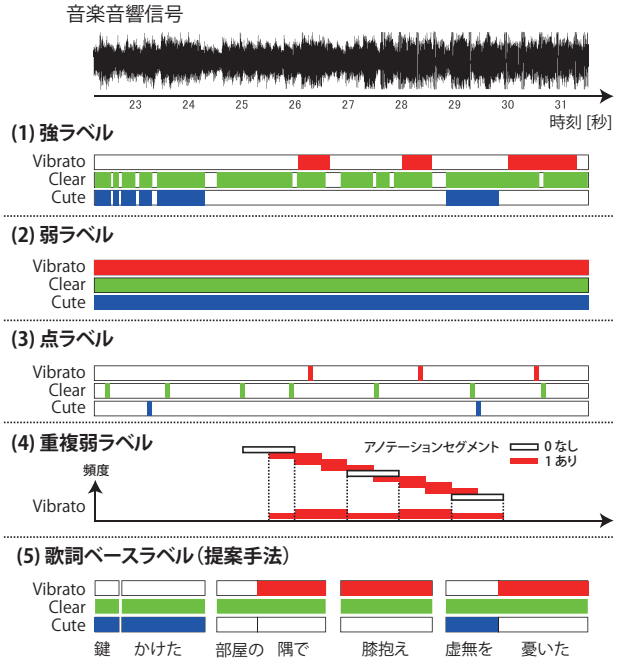

図 1. 5 種類の時系列ラベル例。歌詞ベースラベルの全てと、強ラベルの Vibrato のみが実データに基づいており、それ以外は近似的に図を作成した。Clear と Cute について、点ラベルは歌詞ベースラベルと同様に文節区間ごとに配置し、強ラベルは最も精密なアノテーションを想定して音素区間に配置した。重複弱ラベルは、セグメント長を 1 秒、シフトを 0.5 秒として作図した。

は 1 回の指摘で一つの音響イベントを示すことができ、後者の重複弱ラベルは複数アノテータによる結果の統合が行いやすいという特長がある。

点ラベルでは、複数アノテーション結果を統合す

* 産業技術総合研究所

るためには、それらの対応関係を別途推定する必要がある。また、長く持続するラベルについて、どこの時刻を指摘するのが良いか直感的ではない。一方、重複弱ラベルは、一つの音響イベントに対して複数のアノテーションが必要となって、アノテーションが必要となる回数が増える。また、アノテーション対象（セグメント）の長さやシフト時間を短くすると（時間分解能を上げると）、さらに回数が増えることになる。点ラベルではアノテータが時間を指定する必要があるが、重複弱ラベルでは不要である。

そこで本稿では、作業コストが低く、複数アノテーション結果の統合がしやすい新しい弱ラベルとして歌詞ベースラベルを提案する。つまり、歌詞の文節を単位として、それに該当する歌声にラベル — 音域（高音域、中音域、等）、テクニック（ビブラート、しゃくり、等）、意味（かわいい、明るい、等）— をアノテーションするためのインタフェースを構築する。つまり、歌詞という音響イベントに対して、そこにさらに意味的なラベル付けをすることになる。

歌詞ベースラベルには以下の特長がある。まず、歌詞は自然言語であり、波形等の音響特徴量に不慣れな人にもなじみ深い。次に、歌声の発声区間と紐づいており、固定長の時間切り出しよりも詳細な時刻情報が得られる。また、複数人のアノテーション結果を文節ごとに対応付けることも容易である。さらに、歌詞の異なる 2 名以上の歌唱が重複する場合には、そのいずれに対するアノテーションかを明確にできる。歌声音楽音響信号の中ではイントロや間奏など、歌声が含まれていない区間も多いため、そこを効率的に除ける点でも有用である。

例えば、図 1 の「Clear（クリアな、透き通った、澄んだ）」のように、歌詞のあるフレーズが全て同じ意味を持っていた場合、歌のない短い区間を避けて強ラベルをアノテーションすると（このような区間を避けるアノテーションはかなり難しく、かつコストが高い）、歌詞のいずれかの単位（音素や文節など）でラベル付けすることと同じ結果となる。

なお、歌詞の時刻についてもアノテーションが必要となるが、歌声と歌詞の自動対応付け（アラインメント）技術 [8,11,36,38,41,45] を活用することができる。この際、曖昧性は少ないため、複数人によるアノテーション結果の統合は不要である。

2 関連研究

従来、時系列メディアに対するアノテーションインタフェースとして、強ラベルを付与できるインタフェースが提案されてきた。音響信号や動画を対象として、Praat [3]、Wavesurfer [40]、CLAM Annotator [1]、MUCOSA [13]、ELAN [50], Sonic Visualiser [5]、Timeline Annotator [54]、EMU-webApp [49]、BAT [32]、CrowdCurio [7]、VGG Image Annotator (VIA) [9]、GECKO [27]、audino [12] 等がある。楽譜レベルでのアノテーションインタフェース Score Annotator [54] も提案された。これらのうち、複数アノテータによるアノテーションを考慮したインタフェースもある [9,12,32]。さらに、アノテーションコストを下げるために、機械学習モデルによって自動アノテーションする手法として、区間の自動分割と認識を行う EASEL [44]、認識モデルをインタラクティブに更新する SoundScape [25] と I-SED [20, 21] が提案された。以上のインタフェースは、開始時刻・終了時刻・ラベルの 3 種類、もしくはそれに周波数帯域を加えた 4 種類 [25, 54] を入力するインタフェースである。

一方で、弱ラベル付与のためのインタラクション [6, 23, 46] も提案されてきた。Wang *et al.* は、ラベリング効率を上げるために、類似区間をクラスタリングして同時にアノテーションを提案した [46]。Cartwright *et al.* は、同一区間に重複して存在する音源アノテーション（複数ラベル）のために、複数回のバイナリアノテーション（付与するかしないか）、単一パスでの複数ラベルアノテーション、階層的複数パス・複数ラベルアノテーションの結果を比較した [6]。Kim *et al.* は、歌声へのセマンティックタグ付与のためのインタフェースを構築し、曲単位でのアノテーション結果を初期値として、10 秒のセグメント単位でアノテーションさせた [23]。このような弱ラベルの時間長には、MagnaTagATune (MTAT) [26] は 30 秒、Million Song Dataset (MSD) [2] は 30 秒、MTG-Jamendo [4] は 30 秒以上のフル尺（1 曲）、CAL500exp [46] は 3〜16 秒の可変長、歌唱タグ Kpop Vocal Tag (KVT) [23] では 10 秒、音楽における Multiple Instance learning に関する研究 [28] では 10 秒、アマチュア女性歌唱者の意味ラベル [53] では約 9 秒（固定フレーズ）が用いられた。

以上のような強ラベルと弱ラベルの両者の利点を得るために、新しいアノテーション手法が研究されている。Kim *et al.* [22] は、強ラベルよりも簡便で、弱ラベルよりも時間精度を上げる点ラベルを提案した。アノテーターは音の発生時にマウスクリックするなどにより、音響イベント名を、その音が発生したいずれかの時点でアノテーションする。点ラベルが含まれているセグメントに関して、バイナリクロスエントロピーロスを計算する方法、点ラベルを前後（時間）に拡張する方法が提案された。Martín-Morató *et al.* [29, 30] は、10 秒の長さ（音響イベントの長さに基づいて決定）を持つセグメントを 1 秒シフトさせることで、弱ラベルから時刻情報を特定する方法を提案した。アノテータは時刻を指定する必要はないために作業はシンプルであり、クラウドソーシングを用いてアノテーションを収集しやすい。アノテータの能力を推定する手法を提案して、信頼性の高いアノテーション結果のみを用いた。

話声・歌声信号の代わりに、歌詞など、対応する

テキストを扱うインタラクションもこれまで提案されてきた。Fujihara *et al.* [11] は、歌詞の自動アラインメント結果を活用して、歌詞に基づく音楽シーク機能を提案した。最近では、Amazon Music や Apple Music などの音楽再生プレーヤにも、歌詞のテキスト（行）で再生時刻をシークできる機能がある。また、音声収録結果を編集するために、音声波形を音声認識結果のテキストで編集できるインタラクションが提案されてきた [15,33,37,39]。その他、歌詞の自動同期結果を用いて、同じ歌詞やそのトピックが歌われている別の曲を検索できる Hyperlinking Lyrics [10] や LyricListPlayer [34,57]、より一般的な歌詞駆動型のインタラクティブな視覚表現として「リリックアプリ」[18] が提案されてきた。

以上、アノテーションインタフェースや、歌詞を活用するインタラクションは提案されてきたが、歌詞ベースでのアノテーションに着目するインタラクションは研究されてこなかった。

3 歌詞ベースラベル

歌詞ベースラベルは、手動で付与した歌詞ラベルや自動歌詞アラインメント結果を活用し、時刻情報を持つアノテーションを間接的に可能とする。

3.1 単位

歌詞をアノテーションする単位としては、ユーザが指定するか（可変長）、事前に分割しておくことが考えられる。ユーザが単位を指定する場合、強ラベルと同様、開始時刻と終了時刻を指定する必要があって、ユーザに要求する行動が複雑になり、また作業コストが増える。したがって、本稿では、事前に歌詞を分割することとした。歌詞の分割が共通だと、複数アノテーション結果を統合しやすい利点もある。

事前の分割の単位としては、日本語であれば、段落、文、文節、形態素、文字、音素などが考えられるが、本稿では文節を対象として、CaboCha/南瓜[1] を用いて自動分割（係り受け解析まではせずに、文節解析のみ）した。さらに、本来の歌詞に存在する空白と空行でも分割した。

3.2 歌声ラベル

歌声アノテーションに用いるラベル（歌声記述子）は、歌唱や声質に関する先行研究 [17, 19, 23, 53, 55] に基づいて決定した。これらの研究では、アノテーター間の合意、了解性、または同義性が考慮された。まず、Kim *et al.* による KVT データセット [23] で使用されたセマンティックタグと歌唱タグを含む CAL500exp [46] を参考にした。音域（高域、中域、低域）については、男女合わせて 6 つの記述子とした。また、音色に関する 30 の記述子（Husky/Throaty, Thick, Thin, Warm, Bright, Clear, Relaxed, Dark, Energetic, Mild/Soft, Sharp, Rich, Rounded, Stable, Breathy, Lonely, Sad, Passion, Charismatic, Pretty, Cute, Delicate, Emotional, Pure, Robotic/Artificial, Embellishing, Sweet, Young, Compressed, Dynamic）を用いた。

次に、歌唱印象 [17, 53] と声質 [19, 55] に関する研究で共通に用いられていた以下の 7 つの記述子（Powerful, Nasal, Calm, Weak, Sexy, Resonant, Dosu (Threatening/Frightening)）を追加した。さらに、歌唱印象に関する研究 [17,53] から、上記に関連する Beautiful, Cool の 2 つの記述子を追加した。

また、歌唱テクニックは重要なため、KVT データセットと、歌唱タグ・歌唱テクニックに関する先行研究 [16,46,48,51,56] を参考に、12 の記述子（Whisper/Quiet, Shout, Vibrato, Falsetto, Spoken/Speech-like, Fall, Growl/Scream, Kobushi, Mix voice, Rap, Shakuri, Vocalfry）を選択した。

以上、合計 57 種類の歌声ラベルを用いた。

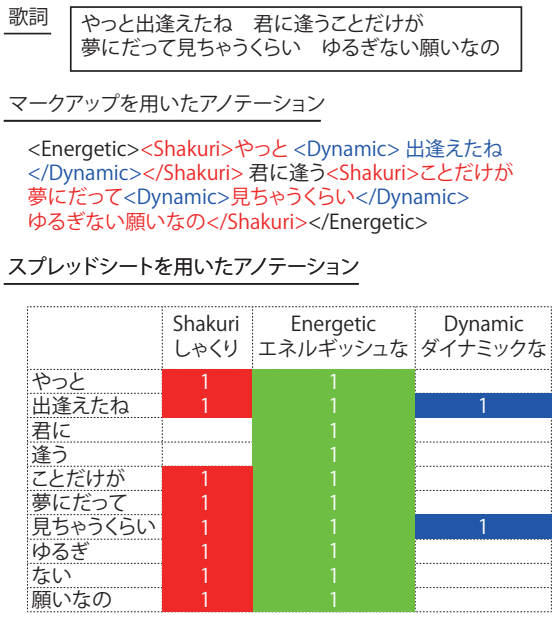

図 2. シンプルなアノテーションインタフェース例

3.3 シンプルなアノテーションインタフェース

歌詞ベースラベルは、仮に歌詞テキストに対する時刻情報がなくても、ユーザ（アノテータ）は歌詞と音楽を対応づけて聴取できる特長がある。つまり、既存の音楽プレーヤとエディタさえあればアノテーションできる。例えば、図 2 に示すように、歌唱ラベルをマークアップとしてテキストに付与したり、スプレッドシートの縦軸に歌詞、横軸にラベルを配置してアノテーションできる。

そこで本研究ではまず、歌詞ベースラベルの特性を分析するために、シンプルなアノテーションで歌

[1] https://taku910.github.io/cabocha/

詞ベースラベルを収集した。歌声ラベルの数が 57 と多くてマークアップは不適と考え、スプレッドシートとして Microsoft Excel を用いた。Excel では文字の入力に応じてハイライトすることも可能で、各歌詞に付与されたラベル数もカウントできる（図 4）。

歌声アノテーションでは、歌声と背景音楽との関係性が重要な場合があり、また歌声分離によるアーティファクトの影響を避けるため、従来、背景音楽がミックスされた歌唱（オリジナル音源）でアノテーションされており [23]、本研究でも同様とした。

3.4 アノテーション

楽曲には産業技術総合研究所が学術目的で構築した非公開の音楽データベースを用いた。具体的には、2010 年代後半のポピュラー音楽シーンを考慮して新規に作詞、作曲、編曲をして制作した多様な日本語楽曲 120 曲に対して、アノテーションした。アノテータは、母国語を日本語とする音楽のエキスパート（歌唱に関する知識、音楽的な知識及びそれらの評価に関する経験が十分あり、ポピュラー音楽の歌声を客観的に聴いて評価・タグ付けできる）の 6 名（M1〜M3 の男性 3 名、F1〜F3 の女性 3 名）で、性別が同一とならないように、3 名ずつ 2 組「M1, F1, F3」「M2, M3, F2」に分けた。つまりアノテーションとしては、1 曲毎に、いずれかの組のアノテータ 3 名がアノテーションした。

各組は、120 曲の半分の 60 曲が割り当てられて、各文節において、カテゴリ「ピッチレンジ」、「音色（Low, 抽象度が低い）」、「音色（High, 抽象度が高い）」のそれぞれに分類されたラベルに対して、カテゴリ毎に必ず 1 つ以上のタグを、かつ、当てはまるラベルを漏れなく入力することとした。一方、カテゴリ「テクニック」に関しては、該当するラベルが存在しない場合は入力しなかった。歌唱としてハモリやコーラスが含まれる場合は、歌詞に対応するメインボーカルのみを対象とした。また、歌唱において、メインボーカルの人数が複数となる場合も、歌詞毎にアノテーションした。

3.5 結果

図 3 に、歌詞の行毎及び文節毎の文字数分布を示す。歌詞に最適化された文節推定ではないこと等が原因で 10 文字以上となることがあったが、平均的には 3.71 文字となった。

アノテータによる、文節単位のアノテーションの良かった点ややりやすかった点の回答を以下に示す。

- 各テクニックや音色などが様々な表現で詳細に分類されており、タグ付けの際に当てはまる項目が見つけやすかった。色分けされ視覚的にも見易く工夫されていた。
- 付与する箇所のみに「1」を入力するだけだったので、集中して作業できたのは良かった。

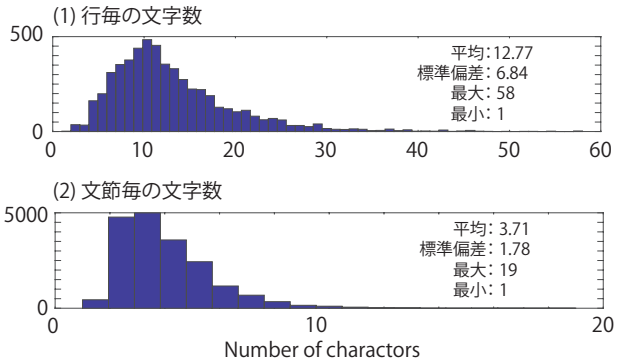

図 3. 歌詞の行及び文節の文字数分布

- 細かい表情にまで対応出来る単位だった。
- 詳細な分類は分析において良い点だと思った。
- 普段ここまで聞き込まないので勉強になった。
- 「1」のみの入力だったので使いやすかった。

つまりアノテータは、「詳細な歌声ラベル」「時間方向の単位の細かさ」「文節選択のみ（1 を入力）の簡易さ」に魅力を感じていた。

また、逆に悪かった点や、改善可能性のある点の回答を以下に示す。

- イメージが近いものなら、1 つのセルの中に 2 つの項目をまとめるのも良いと思う。
- 単位が細かすぎる部分があり、その部分のニュアンス等判断に困ることがあった。
- 単位に統一感が無い部分は作業がし辛かった。
- 単位が一部統一されていない部分があり、その点やりづらさを感じた。
- 単位が細かすぎてタグ付けに困った。区切られ方が曖昧で、もう少し統一されていると良かった。Excel が細かいので目が疲れた。
- 時々単位の長い所や短い所があったので、少し気になった。

特に「文節長さの統一感」「文節という単位が細かすぎる場合がある点」に問題を感じていた。文字数は図 3 のような分布であり、文字数への制限や歌詞への最適化により解決できる可能性がある。その他、「類似した歌声ラベルの扱い」に改善可能性がある。

以上をまとめると、時に歌詞の区切り単位に問題があることもあったが、歌詞ベースのアノテーションは有効であることが示唆された。

3.6 ユーザインタフェースの改善可能性

3.3 節で述べたシンプルな Excel に基づくアノテーションは、新規開発なしに使える利点はあるが、4 章で新たなアノテーション用インタフェースを実装する上で、改善するための追加機能を以下で考察する。

(1) 音楽プレーヤとアノテーションエディタを一体化させ、歌詞と時刻が同期する機能。

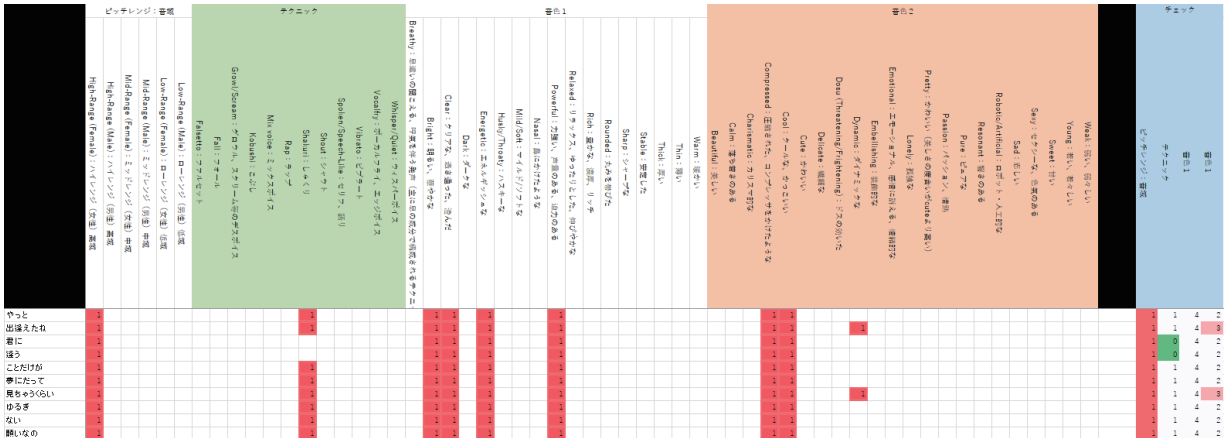

図 4. Excel によるアノテーション結果の例。

(2) Wavesurfer [40] や Sonic Visualiser [5] などの音アノテーションインタフェースと同様、指定した区間の繰り返し（ループ）再生機能。

また、可視化方法に関しても改善可能性がある。スプレッドシートを用いたアノテーションでは、縦方向（歌詞）と横方向（ラベル）を同時に可視化するので、対応関係が常時見える特長がある。しかし、スパースな可視化となって（図 4）、全体像を画面に収めようとすると表示が小さくなることがあり、実際、「目が疲れた」というアノテータがいた。また、歌詞はディスコース（談話）の一種であり、無作為な順序のつながりのない行の集まりではなく、関連する行の連続として構成されている [47] が、歌詞の文節ごとに文節で区切って縦方向に配置したことで、もともとの行や段落の情報が欠落してしまっていた。もともとの行や段落の情報を保持しておくことは、アノテーションに有益な可能性がある。

そこで、歌詞と歌声ラベルをそれぞれ独立に配置することで、視覚的なスペースを節約して全体表示する方針を取る。つまり、歌詞は元々の行・段落の構造を保持し、歌詞と歌声ラベルの対応付けは、歌詞が選択された場合に表示させる。したがって、一覧性のために以下の機能も追加する。

(3) それぞれのラベルが、どの歌詞に付与されているかを可視化できる機能。

4 Lyrics-Based Singing Annotator

3.6 節での考察に基づいて、歌詞ベースラベルのアノテーションに特化した、インタラクティブシステムとして Lyrics-Based Singing Annotator を実装した。歌詞の単位としては、3.5 節のアノテーション結果を用いるためにも、同じ文節区切りを用いる。歌詞の時刻について、本稿ではアノテーションインタラクションに焦点をあてるため、手作業で決定した。

インタフェースを図 5 に示す。画面は以下の通り、大きく四つの機能に分かれている。

音楽再生プレーヤ（左上） 音楽の再生と一時停止ができる。また、再生中の文節をループ再生するボタンがある。それぞれキーボードショートカットが設定されていて、再生・停止はスペースキー、ループ再生は TAB キーとした。機能 (1)(2) に該当する。

歌詞ベース音楽再生プレーヤ・アノテータ（左下） 再生中の歌詞の文節が黄色くハイライトされ、また特定の文節をクリックするとそこから再生される。上下左右キーがショートカットとして設定されていて、左右キーで前後の文節、上下キーで上下の行に移動してシークできる。機能 (1) に該当する。

歌声ラベルの読み込み・保存・破棄・検索（右上） 既にアノテーションされたラベルを読み込んだり、アノテーション結果を保存できる。また全てを破棄してやり直すこともできる。アノテーションが網羅的に実施されたかを確認するために、ラベルを選択してそれが付与された文節が緑にハイライトされる機能も持つ。機能 (3) に該当する。

再生中の歌詞に付与されたラベル（右下） 再生中の文節（黄色にハイライト）に付与されたラベルの選択（オン）／非選択（オフ）をマウスクリックで切り替えられる。このようなラベル表示は、従来のアノテーションインタフェースでも採用されてきた [7, 23, 32]。

4.1 実装に基づく議論

Lyrics-Based Singing Annotator を実装して気づいた点、さらなる改良について議論する。

まず、歌唱者が一名の場合、大きな問題なくラベル付けできた。キーボードショートカットを用いて、再生・一時停止・歌詞のシークをしながら、マウス

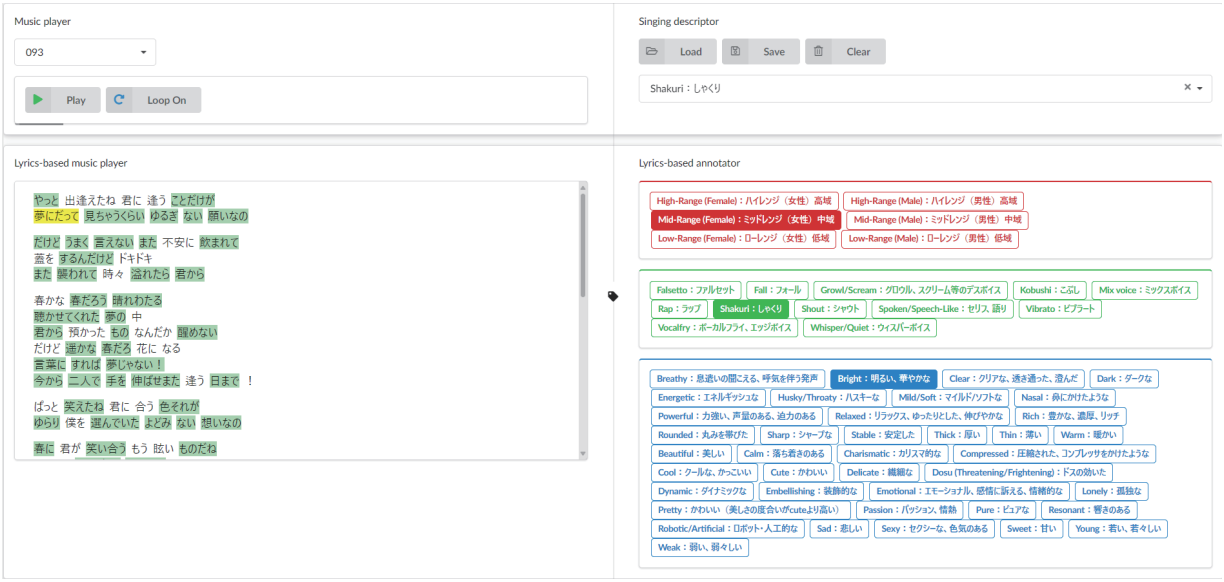

図 5. Lyrics-Based Singing Annotator のスクリーンショット。左上：音楽再生プレーヤ、左下：クリックして再生可能な（クリッカブルな）歌詞、右上：歌声ラベルの読み込み・保存・破棄・検索、右下：再生中の歌詞のラベル。黄色くハイライトされている歌詞の文節は再生中（アノテーション中）であることを意味し、三つのラベルが付与されていることが分かる。また、検索ラベル（Shakuri：しゃくり）が付与された歌詞の文節は緑にハイライトされる。

を用いてラベルを付与できる。ループ再生機能は、特定の歌詞の文節に集中してラベル付けできて、効果的と感じた。強ラベルと違って時刻指定が不要である点に加え、意味的なまとまり（文節）単位でのアノテーションには、やりやすさを感じた。

ただし、複数の文節にまたがるラベルにおいては、工夫が必要である。例えば、音域は一つの歌唱全体で同じラベルになることが多いので、一括で設定できると効率的な可能性は高い。実際、KVT データセットの作成においては、楽曲単位で初期ラベルを付与することが、セグメント単位に 0 からラベル付与するよりも効率的だと主張されている [23]。また、3.5 節でも同様の指摘があった。上記の問題に対する対処法として、歌詞の異なる単位（段落、行、文節、形態素）でのアノテーション機能が効果的な可能性があるので、今後の課題である。この際、楽曲の繰り返し構造や、音響的な類似度に基づくインテリジェントな支援機能はありうるが、初期値がアノテーションに影響を与える可能性もあるので、影響されないインタラクションデザインが必要である。

次に、デュエットなど、同じ時刻に複数の歌唱があると、同時にアノテーションできなかった。画面右下の歌声ラベルセットが、一つの文節にのみ対応しているのが原因である。もし現状のインタフェースを用いる場合は、重複しないようにパートを分けてアノテーションする必要がある。例えば、男性パートと女性パートのある歌声の場合、男性パートだけ、女性パートだけの歌詞を用意して、別々にアノテーションすることは問題なくできるし、このようなパートを分けたアノテーションは不便ではない。

最後に、その他の改善機能として音響特徴量の可視化機能や活用がありうるが、今後の課題である。例えば、ビブラートのアノテーションでは、声の高さが可視化されると効果的であるし、自動認識 [58] も活用できる。ただし、それぞれのラベルに効果的な音響特徴量が何であるかなど、検討が必要である。また、その可視化方法も課題である。従来、行間に音響特徴量を挿入する可視化 [52] や、Tufte による sparklines [42]、musical sparklines [35] などがあり、関連する可能性がある。

5 おわりに

本稿では、歌詞を単位とした新しい弱ラベルである「歌詞ベースラベル」と、そのためのインタラクティブシステムを提案・構築して議論した。歌詞ベースラベルにより、人に理解しやすく、かつ効果的に時間局所性を扱うアノテーションを可能とすることを目指している。また、弱ラベルの一種であることから、複数アノテータによるアノテーション結果を活用しやすい利点もある。 今回は歌詞と歌声を対象としたが、話声にも活用できる。

しかし、さらなる改善の必要性も明らかになった。例えば、音響特徴量の活用、歌詞の分割単位をインタラクティブに決定する機能、アノテーションを支援する自動認識モデルの導入、複数歌唱への対応、等である。今後は、そのようにインタフェースを改善したり、アノテーション結果を実際の機械学習（歌声記述子の推定、等）で用いたりする予定である。

謝辞

本研究の一部は JST CREST JPMJCR20D4 と JSPS 科研費 JP21H04917 の支援を受けた。

参考文献

[1] X. Amatriain, J. Massaguer, D. García, and I. Mosquera. The CLAM Annotator: A Cross-Platform Audio Descriptors Editing Tool. In *Proc. ISMIR 2005*, pp. 426–429, 2005.

[2] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proc. ISMIR 2011*, pp. 591–596, 2011.

[3] P. Boersma. PRAAT, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.

[4] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Proc. ICML 2019*, 2019.

[5] C. Cannam, C. Landone, M. B. Sandler, and J. P. Bello. The Sonic Visualiser: A Visualisation Platform for Semantic Descriptors from Musical Signals. In *Proc. ISMIR 2006*, pp. 324–327, 2006.

[6] M. Cartwright, G. Dove, A. E. M. Méndez, J. P. Bello, and O. Nov. Crowdsourcing Multi-label Audio Annotation Tasks with Citizen Scientists. In *Proc. ACM CHI 2019*, pp. 1–11, 2019.

[7] M. Cartwright, A. Seals, J. Salamon, A. Williams, S. Mikloska, D. MacConnell, E. Law, J. P. Bello, and O. Nov. Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. In *Proc. ACM Human-Computer Interaction*, pp. 1–23, 2017.

[8] S. Choi and J. Nam. A Melody-Unsupervision Model for Singing Voice Synthesis. In *Proc. IEEE ICASSP 2022*, pp. 7242–7246, 2022.

[9] A. Dutta and A. Zisserman. The VIA Annotation Software for Images, Audio and Video. In *Proc. ACM Multimedia 2019*, pp. 2276–2279, 2019.

[10] H. Fujihara, M. Goto, and J. Ogata. Hyperlinking Lyrics: A Method for Creating Hyperlinks Between Phrases in Song Lyrics. In *Proc. ISMIR 2008*, pp. 281–286, 2008.

[11] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno. LyricSynchronizer: Automatic Synchronization System Between Musical Audio Signals and Lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1252–1261, 2011.

[12] M. S. Grover, P. Bamdev, Y. Kumar, M. Hama, and R. R. Shah. audino: A Modern Annotation Tool for Audio and Speech. *CoRR*, abs/2006.05236, 2020.

[13] P. Herrera, Òscar Celma, J. Massaguer, P. Cano, E. Gómez, F. Gouyon, and M. Koppenberger. MUCOSA: A Music Content Semantic Annotator. In *Proc. ISMIR 2005*, pp. 77–83, 2005.

[14] S. Hershey, D. P. W. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal. The Benefit of Temporally-Strong Labels in Audio Event Classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 366–370, 2021.

[15] Z. Jin, G. J. Mysore, S. DiVerdi, J. Lu, and A. Finkelstein. VoCo: text-based insertion and replacement in audio narration. *ACM Transactions on Graphics*, 36(4):96:1–96:13, 2017.

[16] V. Kalbag and A. Lerch. Scream Detection in Heavy Metal Music. In *Proc. SMC 2022*, 2022.

[17] A. Kanato, T. Nakano, M. Goto, and H. Kikuchi. An Automatic Singing Impression Estimation Method Using Factor Analysis and Multiple Regression. In *Proc. Joint ICMC SMC 2014*, pp. 1244–1251, 2014.

[18] J. Kato and M. Goto. Lyric App Framework: A Web-based Framework for Developing Interactive Lyric-driven Musical Applications. In *Proc. ACM CHI 2023*, pp. 1–18, 2023.

[19] H. Kido and H. Kasuya. Representation of Voice Quality Features Associated with Talker Individuality. In *Proc. ICSLP 1998*, pp. 1–4, 1998.

[20] B. Kim and B. Pardo. I-SED: an Interactive Sound Event Detector. In *Proc. IUI 2017*, pp. 553–557, 2017.

[21] B. Kim and B. Pardo. A Human-in-the-Loop System for Sound Event Detection and Annotation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):13:1–13:23, 2018.

[22] B. Kim and B. Pardo. Sound Event Detection Using Point-Labeled Data. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2019.

[23] K. L. Kim, J. Lee, S. Kum, C. L. Park, and J. Nam. Semantic Tagging of Singing Voices in Popular Music Recordings. *IEEE/ACM TASLP*, 28:1656—1668, 2020.

[24] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley. Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2450–2460, 2020.

[25] D. Krijnders and T. Andringa. Soundscape Annotation and Environmental Source Recognition Experiments in Assen (NL). In *Proc. Internoise 2009*, 2009.

[26] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie. Evaluation of Algorithms Using Games: The Case of Music Tagging. In *Proc. ISMIR 2009*, pp. 387–392, 2009.

[27] G. Levy, R. Sitman, I. Amir, E. Golshtein, R. Mochary, E. Reshef, R. Reichart, and O. Allouche. GECKO – A Tool for Effective Annotation of Human Conversations. In *Proc. Interspeech 2019*, pp. 3677–3678, 2019.

[28] M. I. Mandel and D. P. W. Ellis. Multiple-Instance Learning for Music Information Retrieval. In *Proc. ISMIR 2008*, pp. 577–582, 2008.

[29] I. Martín-Morató, M. Harju, and A. Mesaros. Crowdsourcing Strong Labels for Sound Event Detection. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 246–250, 2021.

[30] I. Martín-Morató and A. Mesaros. Strong Labeling of Sound Events Using Crowdsourced Weak Labels and Annotator Competence Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:902–914, 2023.

[31] B. McFee, J. Salamon, and J. P. Bello. Adaptive Pooling Operators for Weakly Labeled Sound Event Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2180–2193, 2018.

[32] B. Meléndez-Catalán, E. Molina, and E. Gómez. BAT: An open-source, web-based audio events annotation tool. In *Web Audio Conference*, 2017.

[33] M. Morrison, L. Rencker, Z. Jin, N. J. Bryan, J. P. Cáceres, and B. Pardo. Context-Aware Prosody Correction for Text-Based Speech Editing. In *Proc. IEEE ICASSP 2021*, pp. 7038–7042, 2021.

[34] T. Nakano and M. Goto. LyricListPlayer: A Consecutive-Query-by-Playback Interface for Retrieving Similar Word Sequences from Different Song Lyrics. In *Proc. SMC 2016*, pp. 344–349, 2016.

[35] J. Oh. Text Visualization of Song Lyrics. Technical report, Center for Computer Research in Music and Acoustics, 2010.

[36] J. Park, S. Yong, T. Kwon, and J. Nam. A Real-Time Lyrics Alignment System Using Chroma and Phonetic Features for Classical Vocal Performance. In *Proc. IEEE ICASSP 2024*, pp. 1371–1375, 2024.

[37] S. Rubin, F. Berthouzoz, G. J. Mysore, W. Li, and M. Agrawala. Content-based tools for editing audio stories. In *Proc. UIST 2013*, pp. 113–122, 2013.

[38] K. Schulze-Forster, C. Doire, G. Richard, and R. Badeau. Phoneme Level Lyrics Alignment and Text-Informed Singing Voice Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[39] V. Sivaraman, D. Yoon, and P. Mitros. Simplified Audio Production in Asynchronous Voice-Based Discussions. In *Proc. ACM CHI 2106*, pp. 1045–1054, 2016.

[40] K. Sjölander and J. Beskow. Wavesurfer - an open source speech tool. In *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH 2000)*, pp. 464–467, 2000.

[41] Y. Teytaut and A. Roebel. Phoneme-to-Audio Alignment with Recurrent Neural Networks for Speaking and Singing Voice. In *Proc. Interspeech 2021*, pp. 61–65, 2021.

[42] E. R. Tufte. *Beautiful Evidence*. Graphics Press, 2006.

[43] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah. Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis. In *Proc. DCASE 2019*, pp. 253–257, 2019.

[44] I. Wang, P. Narayana, J. Smith, B. A. Draper, J. R. Beveridge, and J. Ruiz. EASEL: Easy Automatic Segmentation Event Labeler. In *Proc. IUI 2018*, pp. 595–599, 2018.

[45] J.-Y. Wang, C.-I. Leong, Y.-C. Lin, L. Su, and J.-S. R. Jang. Adapting Pretrained Speech Model for Mandarin Lyrics Transcription and Alignment. In *Proc. IEEE ASRU 2023*, pp. 1–8, 2023.

[46] S.-Y. Wang, J.-C. Wang, Y.-H. Yang, and H.-M. Wang. Towards time-varying music auto-tagging based on CAL500 expansion. In *Proc. IEEE ICME 2014*, pp. 1–6, 2014.

[47] K. Watanabe, Y. Matsubayashi, K. Inui, S. Fukayama, T. Nakano, and M. Goto. Modeling Storylines in Lyrics. *IEICE Transactions on Information and Systems*, E101-D(4):1167–1179, 2018.

[48] J. Wilkins, P. Seetharaman, A. Wahl, and B. A. Pardo. VocalSet: A Singing Voice Dataset. In *Proc. ISMIR 2018*, pp. 468–474, 2018.

[49] R. Winkelmann and G. Raess. Introducing a Web Application for Labeling, Visualizing Speech and Correcting Derived Speech Signals. In *Proc. LREC 2014*, pp. 4129–4133, 2014.

[50] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. ELAN: A Professional Framework for Multimodality Research. In *Proc. LREC 2006*, pp. 1556–1559, 2006.

[51] Y. Yamamoto, J. Nam, and H. Terasawa. Analysis and Detection of Singing Techniques in Repertoires of J-POP Solo Singers. In *Proc. ISMIR 2022*, pp. 384–391, 2022.

[52] D. Yoon, N. Chen, B. Randles, A. Cheatle, S. J. Jackson, C. E. Löckenhoff, A. Sellen, and F. Guimbretière. RichReview++: Deployment of a Collaborative Multi-modal Annotation System for Instructor Feedback and Peer Discussion. In *Proc. ACM CSCW 2016*, pp. 194–204, 2016.

[53] 金礪 愛, 中野 倫靖, 後藤 真孝, 菊池 英明. 歌声の印象評価尺度の構築に基づく多様な印象の自動推定手法. 情報処理学会論文誌, 57(5):1375–1388, 2016.

[54] 梶 克彦, 長尾 確. 楽曲に対する多様な解釈を扱う音楽アノテーションシステム. 情報処理学会論文誌, 48(1):258–273, 2007.

[55] 木戸 博, 粕谷 英樹. 通常発話の声質に関連した日常表現語の抽出. 日本音響学会誌, 55(6):405–411, 1999.

[56] 山本 雄也, 中野 倫靖, 後藤 真孝, 寺澤 洋子. ポピュラー音楽の模倣歌唱における歌唱テクニック分析と楽譜情報との対応付け. 情報処理学会論文誌, 64(10):1423–1437, 2023.

[57] 中野 倫靖, 後藤 真孝. LyricListPlayer: 歌詞でリアルタイムにザッピングできる音楽再生インタフェース. 日本ソフトウェア科学会第 23 回インタラクティブシステムとソフトウェアに関するワークショップ (WISS 2015) 論文集, pp. 1–6, 2015.

[58] 中野 倫靖, 後藤 真孝, 平賀 譲. 楽譜情報を用いない歌唱力自動評価手法. 情報処理学会論文誌, 48(1):227–236.