音楽動画への印象評価データセット構築とその特性の調査

大野直紀[†1, †2]　中村聡史[†1, †2]　山本岳洋[†3, †2]　後藤真孝[†4, †2]

音楽に対する印象評価に関する研究は多数なされており，そうした研究を促進するためのデータセットもさまざまなものが構築されている．一方，音楽と映像が同期して提示される音楽動画を対象とした印象評価に関する研究は，ほとんどなされていない．本研究では 500 曲の音楽動画のサビ区間を対象とし，音楽のみ，映像のみ，音楽と映像の組み合わせという 3 つのタイプの評価対象コンテンツを用意する．また，このコンテンツに対して 8 軸の印象評価を行ってもらうことで，メディアの及ぼす影響を明らかにする．さらに，これまで我々が行ってきた，音楽動画全体に対する印象評価と，本研究で収集した音楽動画のサビ区間に対する印象評価とを比較することにより，印象評価において注意すべき点について考察を行う．

1. はじめに

コンテンツ制作支援システムの普及や発展により，だれでも楽曲や動画を創作することが容易になった．また，大規模動画共有サイトの普及により，多くのアマチュア作曲者や動画製作者の創作したコンテンツを容易に閲覧することが可能となった．これにより，人々が音楽動画に接する機会は非常に増加したといえる．なお，本稿では音楽が主としてありながらも，その音楽と時間的に同期して映像が提示されるものを「音楽動画」と呼ぶ．

音楽動画の増加に関して顕著な例が，初音ミクをはじめとする VOCALOID を使用したものである．日本最大の動画共有サイトであるニコニコ動画において膨大な量の動画が存在している（2015 年 7 月 31 日時点で約 38 万件）．また，アマチュア作曲者および動画作成者が投稿した音楽動画が 100 万回以上再生されているものも多数存在している．これはニコニコ動画という動画投稿および共有基盤の存在と，VOCALOID の人々への普及も大きいが，それに加え Miku Miku Dance*[1] などの 3D をベースとした映像コンテンツ制作ツールと，その制作ツールで利用可能な 3D オブジェクト（キャラクタや各種の部品，世界など）の充実が大きく寄与している．

人々がアクセスできる音楽動画の数が増加している一方で，音楽動画を探すための検索手段は多様ではない．たとえば，ニコニコ動画では音楽動画名やアーティスト名，タグといったテキスト情報に対するキーワード検索や，再生数や投稿日による動画のソートなどの方法でしか検索を行うことができない．

音楽動画を検索する際，キーワードを思いつくことは容易ではなく，ユーザは求める音楽動画を雰囲気や印象といった曖昧な情報でしか表現できないことも多いと考えられる．しかし，雰囲気や印象などの情報はテキストとして含まれていることが少なく，検索は難しい．また，ユーザがタグを付与できるサービス上では，楽曲の印象に関するタグを付与することが可能だが，その割合はニコニコ動画では 5%[9]，音楽に関するソーシャルメディアである Last.fm では 14%[11]と少なく，現状では検索に利用するには不十分である．

こうした問題を解決するため，音楽情報検索の分野では，楽曲の視聴を通してユーザが受ける主観的な印象に基づく検索に関する研究が多数行われている[6][9]．ここで，主観的な印象に基づく検索とは，ユーザが楽曲を聴いて受ける印象に合うように，「人気のある切ない音楽」や「元気の出る印象を受ける動画」といった，主観的な印象語をクエリに含んだ楽曲の検索を可能とする手法のことである．

このような主観的な印象に基づく検索が可能となれば，VOCALOID 楽曲のような，比較的新しいドメインであり，ユーザ自身が好むアーティストやジャンルがまだないドメインにおける楽曲を探しているユーザへの検索手段となる．また，これまでにない新しい観点からの検索手段を提供することができる．ここで，楽曲自体については様々な印象評価に関する研究が行われているものの，楽曲と音楽動画から受ける印象は異なると考えられるため，ある楽曲に対する印象をそのままその音楽動画に適用することは難しい．

メディアを融合した際にどのように印象評価が変化するかという点に関する研究としては，静止画と音楽の組み合わせによっておこる印象の変化などが検証されている[4]．また，映像と音楽の組み合わせでは視覚刺激による類似度が高いことなどが明らかにされている[5]．しかし，同一音楽動画内でのメディアの組み合わせによってどのような違いが生じるのかといったことに対する大規模な調査研究は存在していない．また，音楽動画を評価する際に音楽動画全体を対象とした印象評価と，サビ区間のような音楽動画のある特定の部分を対象とした印象評価にはどのよう

†1 明治大学
Meiji University
†2 JST　CREST
JST CREST
†3 京都大学
Kyoto University
†4 産業技術総合研究所
AIST

*[1] http://www.geocities.jp/higuchuu4/

な違いがあるかなどについての調査研究も存在していない.

一方，我々はこれまでの研究において，ニコニコ動画上の音楽動画 500 曲に対する印象評価データセットを構築してきた[1]．しかし，このデータセットは音楽動画全体（動画の最初から最後まで）に対する評価を行ってもらっているものであり，また音楽と映像がセットになったものを評価してもらっていたため，印象評価のスコアが何を意味するものなのかを明らかにはできていなかった．

そこで本稿では，先述の研究で構築したデータセットで対象とした 500 曲について，音楽動画のサビ部分の 30 秒に限定して印象評価を行う．また，サビ部分を音楽のみ，映像のみ，音楽と映像の組み合わせの 3 種類に分けて 8 つの印象軸に対する印象評価データセットを構築する．さらに，本稿で構築した印象評価データセットを分析するとともに，過去の研究で構築した印象評価データセットとの比較分析を行う．これにより，音楽動画から受ける印象は，音楽動画中のどういった部分のどういったメディアから影響を受けるのかということを明らかにする．

以下, 2 章では関連する研究についてまとめ, 3 章で印象軸データセットの詳しい説明を行う．次に，4 章で構築したデータセットに対する分析を行い，5 章でその結果に対する考察を行う．最後に 6 章で本稿のまとめを行う．

2. 関連研究

音楽情報処理の分野では，ユーザの検索を支援するために，楽曲の印象の推定や印象にまつわる楽曲検索に関する研究が多数行われている．

2.1 楽曲の印象モデル

楽曲の印象の表現方法については，様々なアプローチが提案されている．MIREX では，印象を表す形容詞をクラスタリングすることで，印象を 5 つのクラスに分割し，印象推定のタスクに用いている．また，楽曲のみを対象としたものではないが，楽曲の印象推定にも広く用いられるモデルとして, Russel が提案した Valance-Arousal 空間がある[7]. Valance は快-不快を表す次元，Arousal は覚醒-鎮静を表す次元であり，印象をこの 2 つの軸で表現するという考え方である．

これらの研究のほかにも，印象による検索を行うため，ユーザの検索ニーズに合わせた印象語を選定する手法なども行われている[10].

2.2 楽曲の印象推定

楽曲の印象推定に関する研究は，音楽情報検索の分野において，近年特に取り組まれている．それらの研究では，音響特徴量をベースとした印象の推定が数多くなされている．また，近年では音響特徴量に加え楽曲の歌詞情報を利用した印象推定手法の提案[8]もなされている．

一方，楽曲の音響的特徴に依らない印象推定手法として，楽曲に付与されたタグやコメントによる印象推定[9]も行われている．

このように，楽曲の印象を推定する手法がいくつか提案されているものの，楽曲のアーティスト名やジャンルの推定などと比較すると，印象推定の精度は低い．本稿で作成するデータセットは，音楽動画における印象推定技術のための評価基盤の 1 つとなると考えられる．

2.3 メディア間の印象の差異

音楽動画をはじめとするマルチメディア情報での印象に関する研究として，各メディアから受ける印象の違いに関するものがある．佐藤らの研究[5]では，音楽と静止画では音楽が，音楽と映像では映像から受ける印象が強いことがわかっている．また，長谷川らは静止画と音楽の印象の類似はユーザの好みのジャンルに影響されることを明らかにしている[4].

3. 印象評価データセットの構築

今回構築する音楽動画に対する印象データセットでは，これまでの研究とは異なり，音楽動画のサビ部分のみを対象とする（正確には，サビ開始の 5 秒前から 30 秒間とする）．これは，音楽動画はある程度再生時間長があるため，数分間の視聴の間に評価がぶれると考えたためである．また，メディア間の相違を考慮するため，サビ部分のサビ音楽のみ，映像のみ，音楽と映像の組み合わせについて印象評価データセットを構築する．

評価対象の音楽動画として，過去の研究と同じ音楽動画集合を用いた．この音楽動画集合は，動画共有サイト「ニコニコ動画」上に投稿された音楽動画のうち，タグ「VOCALOID」が付与された動画の 2012 年 8 月時点で再生数が多い動画上位 500 曲を抽出したものである．

以降，音楽動画のサビ区間の検出方法，評価対象とする印象軸，印象評価のインタフェース，印象評価の手続きについて述べる．

3.1 サビ区間の検出

本稿では，音楽動画の全体とサビ部分での印象の相違を分析するため，後述するデータセットの作成時に，サビ区間のみでの印象評価も行う．

しかし，今回評価対象とした音楽動画が投稿されている大規模動画投稿サイト「ニコニコ動画」には，どこからがサビの区間なのかといった情報が付与されているわけではない. そのため, 500 曲のサビ区間を検出する必要がある．そこで本稿では，後藤のサビ区間検出手法 RefraiD[2]を用いる．RefraiD は，サビが楽曲中で最も繰り返されることが多いことに着目した手法である．

RefraiD では, 音響信号の特徴量としてコードとメロディが反映されやすい 12 次元のクロマベクトルを求め，クロマベクトル間の類似度を利用することによって，全体の響きがある程度の区間で類似していれば繰り返し区間であると検出するものである．この RefraiD は，楽曲中のさまざ

まな繰り返し区間をグルーピングすることで繰り返し区間の集合を求め，それぞれの集合ごとに「サビらしさ」を評価し，最終的に「サビらしさ」が高い集合をサビ区間として選択する．RefraiD では，このサビ区間として検出された区間集合中の区間に対して，それがどれくらいほかの区間と似ているかという値を算出できるので，本稿ではそれを各サビ区間の信頼度スコアとみなす．

そして，このスコアを用いて，サビ区間集合の中でもサビらしい区間を求める．そのうえでそのサビらしい区間の開始場所の 5 秒前から 30 秒間を評価対象として抽出する．ここで，サビの開始場所が音楽動画の開始時間 5 秒未満と推定された場合は，音楽動画の開始から 30 秒間を抽出した．なお，ここでサビとして検出されたタイミングの 5 秒前からを抽出対象とした理由は，サビに入る少し前の部分からサビへの変化も重要であると考えたためである．

3.2 印象軸

本稿では，我々の過去の研究と同様に，音楽動画に対する印象として，音楽情報検索ワークショップである MIREX で用いられている 5 つの印象クラスと，Russel らの Valance-Arousal 空間を参考にした．ここで，MIREX では，5 つの印象クラスが用いられているが，これまでの研究[1]により，ニコニコ動画上では「かわいい」と感じる楽曲やそれに関するタグが多く存在することが分かっているため，本稿でもこれまでの研究の評価に則り，MIREX の 5 クラスに加え，可愛らしさを表す印象クラスを加えた 6 軸と，Valance-Arousal に関する 2 軸の合計 8 軸を評価の収集対象とした．

表 1　8 つの印象軸

C1（堂々）	堂々とした，どっしりとした 心躍る，にぎやかな
C2（元気が出る）	元気が出る，楽しい気持ちにさせる 陽気な，心地よい
C3（切ない）	切ない，悲痛な，ほろ苦い 気がめいる，哀愁の
C4（激しい）	アグレッシブな，激しい，興奮させる 感情的な，感情あらわな
C5（滑稽）	滑稽な，ユーモラスな，おもしろげな 奇抜な，気まぐれ，いたずらっぽい
C6（かわいい）	可愛らしい，愛くるしげ，愛おしい かわいい
Valance	明るい気持ちになる，楽しい 暗い気持ちになる，悲しい
Arousal	激しい，積極的な，強気な 穏やか，消極的な，弱気な

本稿で用いた 8 つの印象軸は，表 1 に示すとおりである．表中の「印象クラス名」は，著者らが便宜上付与した，印象を表すラベル名である．また，「印象を表す形容詞」は，データセット構築において評価者から評価値を収集する際に，その印象クラスを表現するために用いた表現を表す．C6 については，「かわいい」の類義語を集めた．また Valence-Arousal についても，既存研究を参考に著者らが日本語に直したものを用いた．

3.3 印象評価インタフェース

図 1 に評価データ収集に用いたインタフェースを示す．図にあるように，評価者は音楽動画を視聴し，その音楽動画に対する印象を，以下に示す形で付与する．

- **C1-C6 の印象クラス:** 表 1 に示した形容詞，形容動詞群に対する 1（全くそう思わない）～5（とてもそう思う）の 5 段階のリッカート尺度
- **Valance:** -2（暗い気持ちになる，悲しい）～+2（明るい気持ちになる，楽しい）の 5 段階のリッカート尺度
- **Arousal:** -2（穏やか，消極的な，弱気な）～+2（激しい，積極的な，強気な）の 5 段階のリッカート尺度

図 1　評価用インタフェース

なお，音楽動画を視聴せずに評価してしまうことがないように，音楽動画をすべて視聴し終えるまで，評価ボタンは押すことができないようにした．

3.4 データの収集

2015 年 3 月 26 日から 2015 年 5 月 19 日 にかけて，3.3 節で述べた対象動画の印象に対する評価データを収集した．データセット構築の協力者は明治大学の学部生と著者を含む計 21 人であった．

データセット構築者には，評価対象である「サビ部分の音楽と映像がミックスされているもの」「サビ部分のサビ音楽のみ」「サビ部分の映像のみ」をそれぞれ視聴し，3.3 節で述べた印象評価用のウェブインタフェースを用い，対象コンテンツに対する印象を評価してもらった．これにより，500 曲の音楽動画，3 メディアタイプ（サビ音楽のみ，映像

のみ，音楽と映像の組み合わせ)，8 印象に関して，少なくとも 3 人以上の評価を収集した．つまり，1500 件（500 曲×3 メディアタイプ）の評価データに対して，少なくとも 3 名以上の評価者が 8 印象について評価を行ったデータを実験により構築した．

4. 分析と考察

本章では 3 章で得られた印象評価データセットの分析を行うことで，どういったメディアがどのような影響を及ぼすのかといったことについて明らかにする．

4.1 分析のためのデータの補正

まず C1 から C6 のデータは，1～5 の 5 段階評価から評価を入力してもらうことになっていた．一方 Valance および Arousal では-2 から+2 の 5 段階から評価を入力してもらっていた．分析にあたり，両者のデータの最小値と最大値をそろえるため，C1 から C6 のデータは 1～5 までの数値を，-2～+2 までの数値へと変換してデータ分析に用いた．

次に，各音楽動画に対するデータセット構築者による各軸に対する評価の平均値を，その音楽動画の印象ベクトルとする．つまり，各音楽動画は，8 軸の印象ベクトル値を持つものとなる．

また，メディア間の比較の際，「音楽のみ」「映像のみ」と「音楽と映像」の関係を調べるため，「音楽のみ」「映像のみ」の印象ベクトルの平均を求めた音楽と映像の平均に関する印象ベクトルを作成し，これを「サビ部分」との比較に使用した．

なお，ここでは便宜的に，サビ部分の音楽のみのものを「サビ音楽」，サビ部分の映像のみのものを「サビ映像」，サビ部分を「サビ音楽動画」，サビ音楽とサビ映像のベクトルの平均を「サビ音楽映像平均」，[1]の研究で求められている音楽動画の最初から最後までに対するものを「フル音楽動画」と表記する．

4.2 異なるメディア間での印象の相違

まず，各音楽動画コンテンツに対する，あるメディアタイプでの 8 軸の印象評価が，ほかのメディアタイプでの印象評価と一致しているかどうかを調べるため，メディア間の類似度をコサイン類似度で計算する．なお，コサイン類似度で比較するため，どの印象軸についても特徴が出ていない音楽動画を排除する．ここでは，いずれかの軸の値の絶対値が 1 以上のもののみを比較対象とした．

表 2 は，各メディア間でのコサイン類似度が 0.8 以上の音楽動画の割合を示したものである．また，図 2 は各メディア間での 8 軸のコサイン類似度の音楽動画の割合の分布を示したものである．縦軸は音楽動画の割合を，横軸は類似度をそれぞれ意味している．

表 2，および図 2 より，すべてのメディアタイプの比較において，コサイン類似度での一致率が 0.8 以上の音楽動画の割合が 0.5 を超えていない．このことから，それぞれのメディアから受ける印象は食い違っていることが分かる．特に，「フル音楽動画」と「サビ音楽動画」については，ほかのメディア同士での比較よりも割合が非常に小さい値となっている．これよりメディア同士の比較よりも，音楽動画の部分によって受ける印象が大きく違ってしまっていることが分かる．また，「サビ音楽動画」から受ける印象と「サビ音楽映像平均」から受ける印象は，「サビ音楽動画」と「サビ音楽」，「サビ音楽動画」と「サビ映像」それぞれとの比較に比べ，10%以上も似通った印象となることが分かった．

表 2　コサイン類似度 0.8 以上の音楽動画の割合

比較するメディアタイプ	0.8 以上の割合
サビ音楽動画 vs サビ音楽	0.388
サビ音楽動画 vs サビ映像	0.386
サビ音楽 vs 映像	0.245
サビ音楽動画 vs サビ音楽映像平均	0.496
サビ音楽動画 vs フル音楽動画	0.101

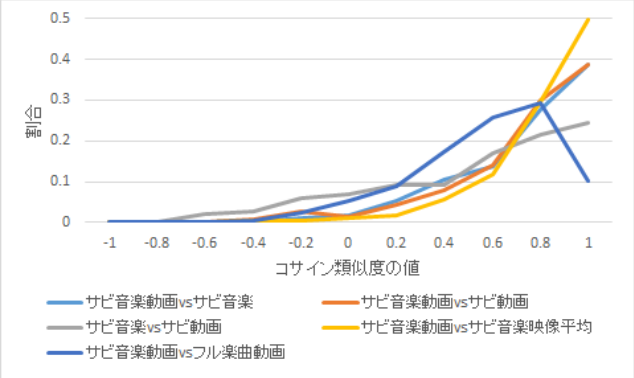

図 2 各メディア 8 軸のコサイン類似度の割合の分布

次に，8 つの印象評価軸がそれぞれどのメディア同士だと類似した印象を与え，どのメディア同士だと違った印象を受けるかを調べるため，8 つの印象評価軸の中からすべての 2 軸ペアについて，各軸での印象評価値の絶対値が各軸，各メディアでどちらも 1 以上の音楽動画のみを用いてコサイン類似度で比較を行った．そのため，比較する音楽動画数は評価対象とする印象軸およびメディアタイプによって異なっていたが，どの組み合わせにおいても音楽動画の数がすべて 150 以上となっていた．

表 3～6 は，「サビ音楽動画」「サビ音楽」「サビ映像」によってできるそれぞれの組と，「サビ音楽動画」と「サビ音楽映像平均」について，コサイン類似度による比較を行った結果のうち，類似度が 0.8 を超えた音楽動画の割合と，それぞれの軸ごとの平均値を表示したものである．また，割合が 0.7 以上のものを太字で表示するとともに，背景色をオレンジ色に設定し，0.8 以上のものは背景色をピンク色で表示した．また，表示の関係上，Valance を「V」，Arousal

を「A」で表示した．

表 3　サビ音楽動画とサビ音楽間の類似度 0.8 以上の割合

	C1	C2	C3	C4	C5	C6	V	A	平均
C1	-	**0.702**	0.675	0.650	0.651	0.678	0.691	**0.740**	0.684
C2	**0.702**	-	**0.768**	0.664	0.669	**0.771**	0.692	**0.841**	**0.730**
C3	0.675	**0.768**	-	0.697	0.633	0.693	**0.761**	**0.861**	**0.727**
C4	0.650	0.664	0.697	-	0.679	0.697	**0.843**	**0.745**	**0.711**
C5	0.651	0.669	0.633	0.679	-	0.658	0.676	0.656	0.660
C6	0.678	**0.771**	0.693	0.697	0.658	-	**0.762**	**0.748**	**0.715**
V	0.691	0.692	**0.761**	**0.843**	0.676	**0.762**	-	**0.700**	**0.732**
A	**0.740**	**0.841**	**0.861**	**0.745**	0.656	**0.748**	**0.700**	-	**0.756**

表 4　サビ音楽動画とサビ映像間の類似度 0.8 以上の割合

	C1	C2	C3	C4	C5	C6	V	A	平均
C1	-	**0.772**	0.645	0.602	0.592	0.668	0.540	0.658	0.640
C2	**0.772**	-	**0.707**	0.636	0.631	**0.797**	0.613	**0.825**	**0.711**
C3	0.645	**0.707**	-	0.668	0.696	**0.750**	0.681	**0.870**	**0.717**
C4	0.602	0.636	0.668	-	0.657	0.695	**0.753**	0.654	0.666
C5	0.592	0.631	0.696	0.657	-	**0.709**	0.646	0.661	0.656
C6	0.668	**0.797**	**0.750**	0.695	**0.709**	-	**0.703**	**0.808**	**0.733**
V	0.540	0.613	0.681	**0.753**	0.646	**0.703**	-	0.650	0.655
A	0.658	**0.825**	**0.870**	0.654	0.661	**0.808**	0.650	-	**0.732**

表 5　サビ音楽とサビ映像間の類似度 0.8 以上の割合

	C1	C2	C3	C4	C5	C6	V	A	平均
C1	-	0.500	0.481	0.450	0.418	0.431	0.394	0.486	0.451
C2	0.500	-	0.585	0.500	0.460	0.579	0.442	0.670	0.534
C3	0.481	0.585	-	0.564	0.531	0.604	0.602	**0.731**	0.586
C4	0.450	0.500	0.564	-	0.599	0.629	0.694	0.584	0.574
C5	0.418	0.460	0.531	0.599	-	0.599	0.536	0.538	0.526
C6	0.431	0.579	0.604	0.629	0.599	-	0.622	0.659	0.589
V	0.394	0.442	0.602	0.694	0.536	0.622	-	0.495	0.541
A	0.486	0.670	**0.731**	0.584	0.538	0.659	0.495	-	0.595

表 6　サビ音楽映像平均と
サビ音楽動間の類似度 0.8 以上の割合

	C1	C2	C3	C4	C5	C6	V	A	平均
C1	-	**0.876**	**0.833**	**0.778**	**0.734**	**0.829**	**0.733**	**0.868**	**0.807**
C2	**0.876**	-	**0.835**	**0.744**	**0.761**	**0.916**	**0.816**	**0.935**	**0.840**
C3	**0.833**	**0.835**	-	**0.847**	**0.774**	**0.805**	**0.881**	**0.940**	**0.845**
C4	**0.778**	**0.744**	**0.847**	-	**0.789**	**0.780**	**0.895**	**0.826**	**0.809**
C5	**0.734**	**0.761**	**0.774**	**0.789**	-	**0.790**	**0.812**	**0.794**	**0.770**
C6	**0.829**	**0.916**	**0.805**	**0.780**	**0.790**	-	**0.852**	**0.901**	**0.839**
V	**0.733**	**0.816**	**0.881**	**0.895**	**0.812**	**0.852**	-	**0.859**	**0.836**
A	**0.868**	**0.935**	**0.940**	**0.826**	**0.794**	**0.901**	**0.859**	-	**0.875**

表 3〜6 を通して，C1（堂々とした），C5（滑稽な）の各軸と残りの 7 軸でのコサイン類似度 0.8 以上の音楽動画の割合は，どのメディア同士でも低くなっている．それに対し，C3（切ない）と残りの 7 軸，Arousal と残りの 7 軸，C6（かわいい）と残りの 7 軸での類似度 0.8 以上の音楽動画の割合は，どのメディア同士においても高い割合を出している．これより，C1, C5 は各メディアで違った印象を，C3, Arousal，C6 は，どのメディアでも同じような印象を受けやすいことが分かる．

また，C6（かわいい）と残りの 7 軸で比較した結果は，「サビ音楽動画」と「サビ音楽」で比較した場合より，「サビ音楽動画」と「サビ映像」で比較している場合で高い値が出ている．これより，かわいらしさは各メディアで同じような印象を与えるが，特に映像に影響されやすいことが分かる． さらに，すべてのメディアにおいて，類似度が最大となっているのが C3（切ない）と Arousal の 2 軸で比較したものとなっている．

次に，どのメディアからだと大きな印象を受けるのかを明らかにするために，評価値の分布においても比較を行う．メディアごとの評価値の分布の傾向を見るために，各メディアの評価値平均を図 3 に表示した．

図 3 は，メディアごとに，8 軸の評価値の分布の平均を示したものである．横軸は評価値の大きさ，縦軸は各軸でその評価値を出した評価データの件数の平均である．

なお，ここで横軸が 3 分の 1 刻みなのは，ほぼ全て（97%）の評価データについて評価者数が 3 人であったためである．

図 3 より，「サビ音楽動画」「サビ音楽」に比べ，「サビ映像」では-2 での評価値が大きく出ていることが分かる．これより，「サビ映像」は他のメディアに比べ全体的にマイナスの印象を強く与えることが分かった．

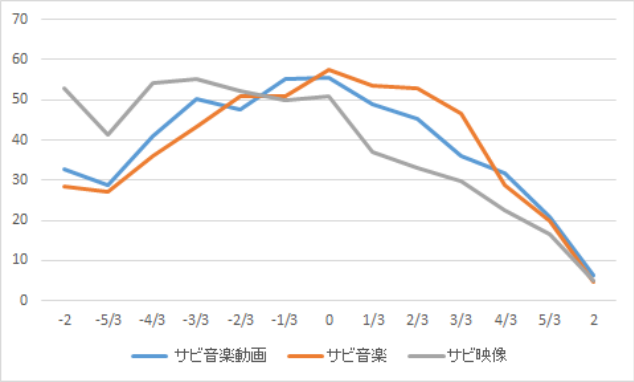

図 3　各メディアでの評価値の分布の平均

それぞれの軸で，どのメディアが大きい印象を与えているのかを調べるため，8 軸それぞれの評価値の分布を表示した．ここでは，特徴が顕著に表れたもののみを表示する．

図 4〜8 は，各メディアの C1，C2，C5，C6，Valance についての評価値の分布を示したものである．図の横軸は評

価値，縦軸はその評価値を出した音楽動画の件数である．

これらの結果より，C1（堂々とした），C2（元気が出る）における「サビ映像」の評価値が，最低値である-2 を示している音楽動画の件数が「サビ音楽動画」「サビ音楽」に比べ，非常に多くなっていることが分かる．つまり，「サビ映像」では特に C1 ，C2 においてマイナス面で強い印象評価となることが分かる．これに対し，C1, C2 では「サビ音楽」からプラスの印象を受けやすいことがわかる．

図 6，図 8 より C5，Valance では 3 つのメディアタイプどれについても評価値の絶対値が 1 に満たないものが多い．これより， C5，Valance では，どのメディアでも強い印象を与えることが困難であることが分かる．

4/3 以上の評価値を出している音楽動画の件数が「サビ音楽」「サビ映像」よりも「サビ音楽動画」において多いのは，C5，C6 のみであった．これは，滑稽さや可愛らしさは「サビ音楽」「サビ映像」がそろって初めて大きく印象を受けるという特徴があるからではないかと考えられる．

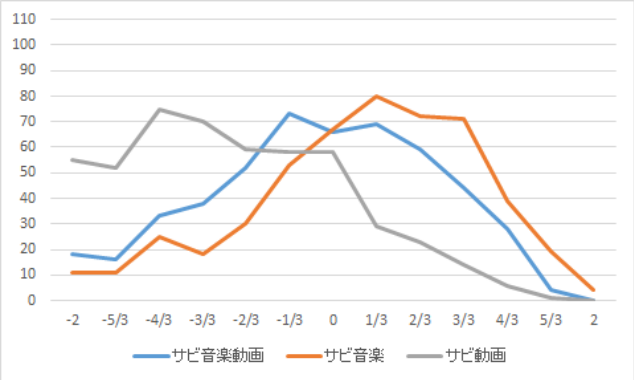

図 4　C1 における評価値の分布

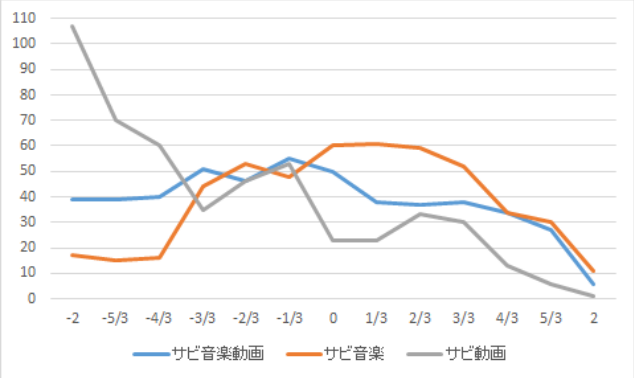

図 5　C2 における評価値の分布

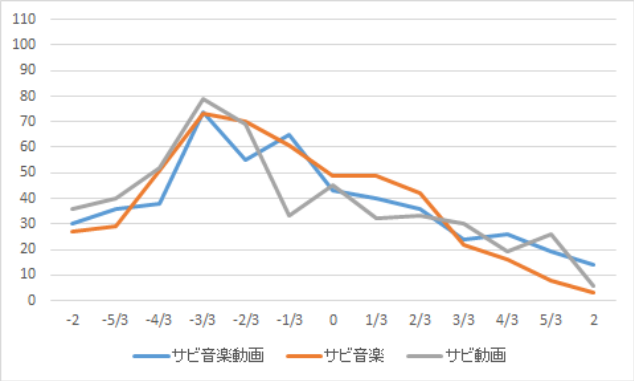

図 6　C5 における評価値の分布

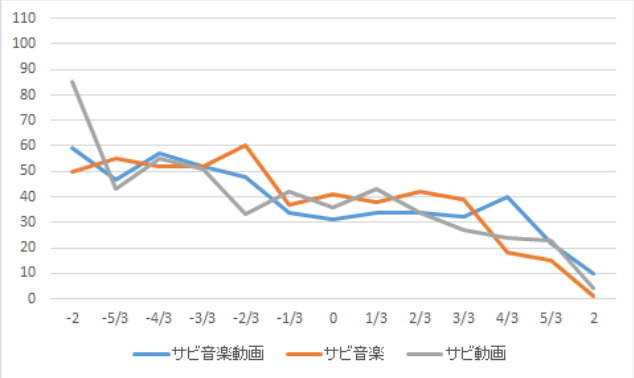

図 7　C6 における評価値の分布

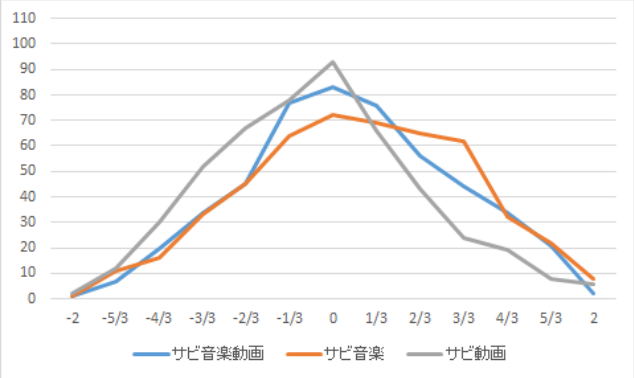

図 8　Valance における評価値の分布

4.3　音楽動画全体と音楽動画のサビ部分での印象の相違

本節では，我々の過去の研究[1]で作成したフル音楽動画への印象評価データセットと，本稿で作成した印象評価データセットについての比較および分析を行う．

比較では，コサイン類似度を使用するが，その際，4.1 節と同様の補正を行ったデータで比較を行う．

ここでは，「サビ音楽動画」と「フル音楽動画」について，8 つの印象評価軸の中からすべての 2 軸ペアについて，各軸での評価値の絶対値が 1 以上の音楽動画を用いてコサイン類似度で比較を行う．

表 7 は，こうして得られた結果のうち，類似度が 0.8 を超えた音楽動画の割合を表示したものである．なお，値が

0.7 を超えるセルを太字にするとともに背景色をオレンジ色で表示している．

表 7　フル音楽動画とサビ音楽動画における類似度 0.8 以上の割合

	C1	C2	C3	C4	C5	C6	V	A	平均
C1	-	**0.723**	0.674	0.622	0.520	0.668	0.338	0.415	0.566
C2	**0.723**	-	**0.719**	0.599	0.606	**0.797**	0.333	0.388	0.595
C3	0.674	**0.719**	-	0.599	0.556	0.668	0.379	0.464	0.588
C4	0.622	0.599	0.599	-	0.566	0.606	0.302	0.349	0.522
C5	0.520	0.606	0.566	0.566	-	0.619	0.353	0.361	0.511
C6	0.668	**0.797**	0.668	0.606	0.619	-	0.392	0.457	0.601
V	0.338	0.333	0.379	0.302	0.353	0.392	-	0.124	0.317
A	0.415	0.388	0.464	0.349	0.361	0.457	0.124	-	0.366

表 7 より，Valance，Arousal の軸と任意の軸の 2 軸を比較した結果を見ると，すべての結果が低い値を出している，これより，Valance-Arousal は，部分と全体で大きく印象が変わってしまうことが分かる．また，C6 と C1～C5 の 5 軸との比較の結果を見ると，ほかの結果よりも比較的高い値を出していることが分かる．C6（かわいい）はフル音楽動画でもサビ部分でも一貫して同じような印象を受けることが分かった．また，4.2 節で行ったサビ部分での各メディアでの比較では低く出ていた C1（堂々とした）と，C2～C6 の 5 軸との比較の結果の値は決して低いものではない．これより，C1（堂々とした）はサビの部分でのメディアでは類似が少ないが，フル音楽動画を通した評価では類似が多いことが分かる．

また，図 9，図 10 は我々の過去の研究で印象評価を行った 500 曲について，「フル音楽動画」に対する各 8 軸に対する評価値の評価値の分布を，C1～C6 と Valance-Arousal で分けて表示したものである．図の横軸は評価値，縦軸はその評価値を出した音楽動画の件数を表している．

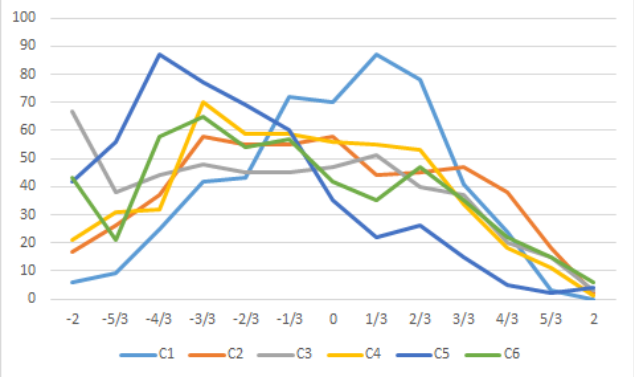

図 9　C1～C6 のフル音楽動画の評価値の分布

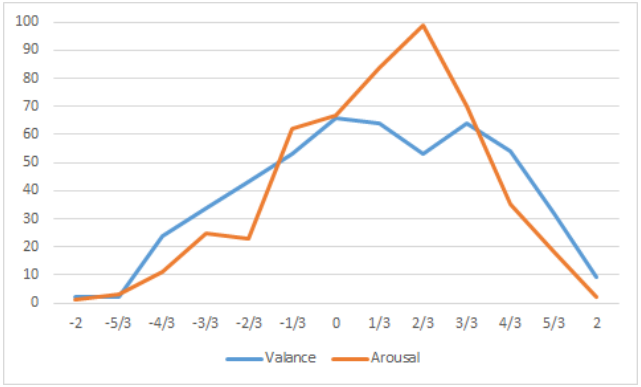

図 10　Valance-Arousal のフル音楽動画の評価値の分布

図 9，図 10 を見ると，C5（滑稽な）では，-4/3 以下の値を出しているフル音楽動画の件数が非常に多いことが分かる．また，図 6 の結果と比較すると，「フル音楽動画」で+1 以上の評価値を出している音楽動画の件数が，「音楽動画」よりも少なくなっている．これより，C5（滑稽さ）は「フル音楽動画」では伝わりづらいことが分かった．

5.　考察

サビ音楽とサビ映像が似ていないにも関わらず，サビ音楽動画とサビ音楽映像平均の評価値が似通っている点はとても興味深いものである．つまり図 11 のように，音楽と映像がベクトルとして違う方向を指しているが，ベクトルの和は，そのメディアを融合したものと同じになるというものである．これは，音楽動画が音楽および映像の双方から同程度の影響を受けていることを示唆している．つまり，音楽のみ，映像のみを評価するだけでは音楽動画の評価はできず，音楽動画を評価するだけでも音楽の評価や映像の評価を行えないことが分かる．一方，この音楽動画が音楽と映像のベクトルの合成で表現できるのであれば，音響分析技術によって音楽の印象推定が可能となり，映像分析技術によって映像の印象推定が可能となると，このそれぞれの印象評価ベクトルを合成することによって音楽動画の印相が推定可能になると期待される．

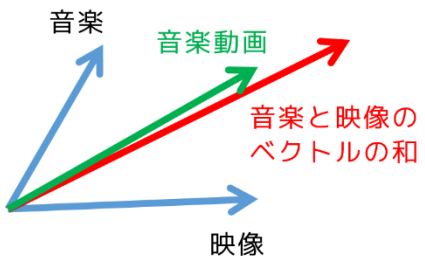

図 11　音楽と映像を合成すると音楽動画の評価に近づく

音楽と映像間の印象の相違としては，C1（堂々とした）と C5（滑稽な）では類似度が低く，C3（切ない），C6（かわいい），Arousal では類似度が高いという結果もまた興味

深い．このことから，切なさやかわいらしさ，激しさにおいてはメディアタイプによらず評価は似通う傾向にあり，それぞれの影響が低いことが分かる．一方，堂々としている感じや滑稽さについては，それぞれのメディアの影響が大きいと考えられる．また，C1，C5 における評価値の分布をみると，評価値の大きさが 1 に満たない音楽動画の件数が多いため，そもそも音楽動画から C1（堂々とした）や C5（滑稽な）を強い印象として感じさせることが難しいということも考えられる．

C1（堂々とした），C2（元気が出る）では，メディア間での印象の相違は大きくみられたが，「サビ映像のみ」での評価値の分布において，ノルムが 1 以上を示している音楽動画の数が非常に多くなっている．それに対し，「サビ音楽のみ」では「サビ映像のみ」に対し，プラスの評価値を示したものが多く見られた．これより，C1，C2 は，マイナスの印象を特に映像からの影響を大きく受けやすく，音楽からはプラスの印象を受けるといえる．

「サビ音楽動画」と「フル音楽動画」の比較により，特に Valance と Arousal に関しては，音楽動画の一部分と全体とで大きく評価が異なることが分かる．つまり，Valance や Arousal などの印象評価は音楽動画の部分によって大きく変化していくため，こうした印象評価は音楽動画全体に対して行うのは適していないと考えられる．一方，サビ音楽動画とフル音楽動画の比較において，C3（切ない）と C6（かわいい）の類似度は大きく出ていた．これより，C3，C6 は部分によって影響されない，同じような印象を受けやすい軸であると考えられる．

以上のことより，Valance や Arousal を印象評価に使用する際には，評価してもらう部分を分けてしまうと評価が大幅に変わってしまうため，注意が必要である．また，C1（堂々とした）や，C5（滑稽な）を伝えたい音楽動画を作成する場合は，動画と音楽の印象を合わせるように注意するべきであるといえる．

6. まとめ

本稿では，500 曲の音楽動画のサビの部分に対して，「サビ部分」「サビ音楽のみ」「サビ映像のみ」の 3 タイプのメディア分離を行ったものに対し，印象評価データセットを作成し，それについての分析，考察を行った．また，我々の過去の研究で作成した，本稿で作成したデータセットと同一の評価対象である 500 曲の音楽動画全体に対しての印象評価データセットとの比較，考察を行った．その結果，メディアによって受けやすい印象に差があること，印象は映像に影響されやすいこと，「サビ部分」と「音楽動画全体」での印象は大きく食い違うことなどといった様々なことを明らかにした．なお，今回得られた知見をまとめると表 8 のようになる．

また，今回の実験で用いた音楽動画群では，特定の印象を大きく持つ音楽動画が少ないことが見受けられた．これは，そもそもそうした音楽動画が少ないのか，音楽動画としての印象評価が出にくいものなのかは不明である．今後はより多くの音楽動画に対して調査を行うことにより，こういった点を検証予定である．

今回実現した音楽動画データセットを利用することにより，音響分析による印象推定，映像分析による印象推定技術を構築し，その印象ベクトルの合成によって音楽動画の印象を推定する技術もまた実現する予定である．

表 8　各軸の特徴

C1（堂々）	音楽からプラスの印象 映像からマイナスの印象 メディアによる印象の差が大きい
C2（元気が出る）	音楽からプラスの印象 映像からマイナスの印象 メディアによる印象の差が大きい
C3（切ない）	メディアによる印象の差がやや小さい 音楽動画のサビ部分と全体で差が小さい
C5（滑稽）	メディアによる印象の差が大きい
C6（かわいい）	メディアによる印象の差がやや小さい 音楽動画のサビ部分と全体で差が小さい
Valance	音楽動画のサビ部分と全体で差が大きい
Arousal	メディアによらず評価が似通う傾向がある 音楽動画のサビ部分と全体で差が大きい

参考文献

1)　山本岳洋，中村聡史: 楽曲動画印象データセットの作成とその分析，ARG 第 2 回 Web インテリジェンスとインタラクション研究会 (2013)
2)　後藤真孝: SmartMusicKIOSK: サビ出し機能付き音楽試聴機，情報処理学会論文誌，Vol.44，No.11，pp.2737-2747 (2003)
3)　大出訓史，今井篤，安藤彰男，谷口高士: 音楽聴取における"感動"の評価要因～感動の種類と音楽の感情価の関係，情報処理学会論文誌 Vol. 50, No. 3, pp.1111-1121 (2009).
4)　長谷川優，武田昌一: 好みの音楽ジャンルに着目した静止画と音楽の組み合わせに関する考察: －個人の属性に着目した静止画と音楽に対する印象度の相互比較－，日本感性工学会論文誌，Vol.11，No.3，pp.435-442 (2012).
5)　佐藤淳也, 佐川雄二, 杉江昇: 音と映像の組み合わせによる主観的印象の変化，映像情報メディア学会誌，Vol.55，No.7，pp.1053-1057 (2001).
6)　熊本忠彦, 太田公子: 印象に基づく楽曲検索システムの設計・構築・公開，人工知能学会論文，Vol.21, pp.310-318 (2006).
7)　Russell, James A.: A Circumplex Model of Affect，Journal of Personality and Social Psychology，39(6)，pp.1161-1178(1980)
8)　舟澤慎太郎, 北市健太郎, 甲藤二郎: 楽曲推薦システムのための楽曲波形と歌詞情報を考慮した類似楽曲検索に関する一検討，情報処理学会研究報告オーディオビジュアル複合情報処理，pp.1-5 (2013).
9)　山本岳洋，中村聡史: 視聴者の時刻同期コメントを用いた楽曲動画の印象分類，情報処理学会論文誌，Vol.6，No.3，pp.66-72 (2013).
10)　熊本忠彦, 太田公子:印象に基づく検索のための印象語選定法

の提案，情報処理学会論文誌，Vol.44，No.7, pp.1808-1811 (2003).
11) Hu, X., Bay, M. and Downie, J.: Creating a Simplified Music Mood Classification Ground-Truth Set, Proceedings of the 8th International Conference on Music Information Retrieval, pp.309-310 (2007).