社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－AVM－63 (18)
2008 / 12 / 12

人間の知覚特性を考慮した音と映像の特徴検出および調和の許容時間を考慮したマッチング

飯塚　太郎[†]　YueYonghao[†]　土橋　宣典[††]　西田　友是[†]

† 東京大学
〒 277-8561 千葉県 柏市 柏の葉 5-1-5
†† 北海道大学
〒 060-0808 北海道 札幌市 北区 北 8 条 西 5 丁目

あらまし　心理学的知見から、音と映像とのリズムを時間的に同期させると、人間は心地よいと感じると言われており、その同期作業はコンテンツ作成においても頻繁に行われている。本研究では、従来手作業で行っていた同期作業をコンピューターを用いて支援するシステムの構築を目指す。音や映像において急激な変化のある箇所を特徴として検出し、映像の再生速度を調整して音と映像の特徴を同期させる。特徴検出では、まず、音楽や映像の局所的な平均変化率をその区間における信号の分散により測り、特徴検出のための閾値をもとめる。次に閾値を超えた信号を特徴として検出し、その超過度をその特徴の重みとする。音と映像との同期作業では、まず、それぞれの特徴の時間的なずれと、特徴の重みを考慮して、より時間的に接近し、かつ重みの大きいもの同士をマッチングさせる。次に、マッチングされた特徴の各ペアについて、音と映像の特徴が現れる時間の差が、人間が違和感を感じない許容限界よりも小さくなるように、動画の再生速度を調節する。
キーワード　音楽、映像、調和、同期

Detection and matching of the accents in music and animation based on the human perceptual characteristic

Taro IIZUKA[†], Yue YONGHAO[†], Yoshinori DOBASHI[††], and Tomoyuki NIHSITA[†]

† University of Tokyo
5-1-5, Kashiwanoha, Kashiwa-shi, Chiba, 277-8561, Japan
†† Hokkaido university
Kita 8-jyo Nishi 5-chome, Kita-ku Sapporo-shi, Hokkaido, 060-0808, Japan

Abstract　From the psychological point of view, it is said that people feel comfortable if the rhythms of sound and video are synchronized. Therefore, works for the synchronization are performed frequently in the process of content making. We are aimed at developing a computer assisted system for the synchronization, which was done manually. In our system, we detect rapid variations in the sound and the video as accents, and adjust the playback speed of the video to make the accents of the video and the sound matched. To detect the accents, we first calculate the time-varying local variances of the variation in the signals, and determine the local thresholds for the detection. Next, we extract signals which exceed the thresholds as accents and assign the exceeded amount as the weight of each accent. In the synchronization process, we first take into account both the weights and the time difference between the occurrences of the accents of the sound and the video, and search a best matched accent of the video for each accent of the sound. Then, we adjust the playback speed of the video to ensure that the difference between the occurrences of the pair of the accents is small enough that people would not feel uncomfortable.
Key words　Music, Animation, Synchronism, Matching

1. はじめに

映画やテレビなどのマルチメディアコンテンツにおいて、コンテンツをより印象的なものにするため、映像と音楽とが組み合わされている。音楽と映像とが組み合わされたとき、人間は両者の間に何らかの調和感あるいは非調和感を感じる。心理学的見知によると、一般的に音楽と映像の調和は、両者の時間的な変化箇所の一致によると考えられている [1]。音の強弱や高低から感じとれる音楽のリズムに対し、映像が同期しているとき、人間は音楽と映像との間に調和を感じる。

これら音や映像において変化の大きくなる箇所を特徴とし、その時間的な並びをそれぞれの時間的構造と呼ぶ。映像と音楽とを単純に組み合わせるだけでなく、両者が変化する箇所を一致させること、すなわち特徴の時間的構造をマッチングすることで、そのコンテンツの印象をより高めることができる。特にミュージックビデオやアニメのオープニング・エンディングビデオのようなコンテンツでは、音楽に合うように映像が付加されており、両者の相乗効果によって、より強い印象を与えるものになっている。

近年、情報機器によるメディア処理技術の向上に伴って、映像と音楽とを用いた作品制作は、一般大衆でも手軽に行えるようになり、同時に、それを公開し閲覧する環境も整ってきた。YouTube(http://jp.youtube.com/) 等に代表されるインターネット上の映像作品投稿サイトの大きな発展は、従来の文字や静止画による情報伝達と比べ、映像や音がより人間の感覚に訴えるメディアであることもあいまって、文化や社会に及ぼす影響も大きくなりつつある。これら映像作品投稿サイトには、連日数多くの映像作品が登録され、アマチュアのコンテンツ制作者による作品がその大半を占めている。

これらコンテンツ作成においては、従来、映像を部分的に追加切除し、特徴同士をマッチングさせる操作を人間が手作業により行っていたため、データを編集・管理する制作者にとって大きな負担となっていた。

本研究では、この負担を軽減するため、コンテンツ作成の素材となる音楽および映像データから、特徴の時間的構造を検出し、映像の追加切除を行わずに、映像の再生速度の調節を行うことによって、映像内容を維持したまま両者を同期させる手法について述べる。特に、人間の視聴覚における時間的分解能を考慮した特徴の検出手法と、映像と音楽の時間的非調和における人間の許容時間を考慮した同期手法を提案する。本研究は、これらの映画やアニメ、ミュージックビデオ等のコンテンツの制作支援に応用が期待される。

2. 音楽・映像の特徴検出

本稿では、ミュージックビデオやアニメのオープニング・エンディングビデオのように、音楽の特徴を基準とし、映像のカットや色調の大きな変化といった特徴を同期させることにより編集を行うようなコンテンツを対象とする。実際、映像作品投稿サイトにおいては、もともと関連性のない映像と音楽とを組み合わせて制作された作品においても、音楽の特徴を基準として映像が編集され、映像のカット等の特徴を同期させた作品が多く見受けられる。これらの特徴を、コンテンツの素材となる任意の映像および音楽データから、それぞれ検出する方法を述べる。

人間の視聴覚は、時間的な分解能に限界がある。視覚においては、ある一定以上の短い時間において完結した映像に関して、人間はその現象を正確に知覚できず、また聴覚においても、時間的に接近しすぎている２つの音が連続した１つの音として知覚されてしまう。そこで、特徴検出では、人間が個々の特徴を認識できるかどうかを考慮するため、知覚の分解能に応じた検出を行う。

2.1 音楽情報からの特徴検出手法

映像に対する音の効果には、物理的効果と構造的効果とがある。物理的効果とは、映像の変化のタイミングに対する、音の鳴った消えた (ON/OFF) であり、構造的効果は、一定時間の範囲内における前後での変化、つまり音量や、和音の色彩の変化である。様々な音楽において、楽器はリズムのタイミングに合わせて発音、または音高を変化させる場面が非常に多い [2]。楽器によるこれら発音や音高の変化のタイミングが、音楽の時間的な特徴を強く表していることを意味している。

ある音が発せられたとき、音が鳴り止むとき、また、音高が変化したとき、その音に対応する周波数のパワースペクトルは変化する。全周波数を同時に考慮したパワースペクトルについて、時間軸方向変化分を求めることにより、特徴の検出を行う。

図１に検出の概要図を示す。図１では紙面奥および手前方向に時刻と周波数の軸を、上方向にパワースペクトルの軸を示している。

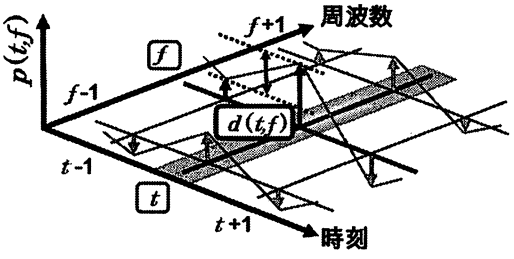

図１　パワースペクトルの変化分の算出

まず、音楽信号を高速フーリエ変換し、周波数軸にて離散化したある周波数 f について、時刻 t および1タイムステップ前の時刻 $t-1$ におけるパワースペクトル $p(t,f)$、$p(t-1,f)$ を算出し、変化分 $d(t,f)$ を式(1)より求める。

$$d(t,f) = p(t,f) - p(t-1,f) \quad (1)$$

次に、式(1)の計算を全ての周波数ついて行い、全周波数領域における時間変化分 $D_{music}(t)$ を式(2)により算出する。

$$D_{music}(t) = \sum_f d(t,f) \quad (2)$$

この値が大きく変化する箇所を特徴として検出するための閾値を求める。この時、$D_{music}(t)$ の時間変化に適応して変化する閾値を設定し、前後時間の $D_{music}(t)$ の値と比較し、大きな変化値を示す箇所を検出するようにする。これを各時間における、$D_{music}(t)$ の平均値から求める。図2に閾値設定の概要図を示す。横軸は時刻、縦軸は $D_{music}(t)$ を表し、図2において、矢印aで示したタイムステップに着目したとき、その前後2タイムステップを含んだ範囲(100ms)が、聴覚の時間的分解能の範囲となる。

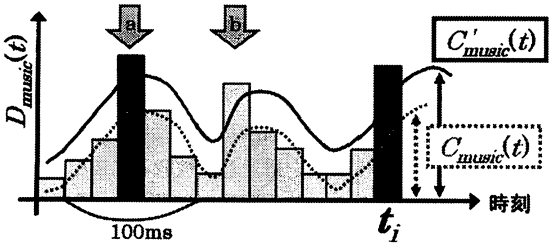

図2 音楽情報からの特徴検出

聴覚における時間的分解能に関しては、時間的なずれを与えた2つの音の場合、約20〜30ms以上の間隔があれば、時間的に独立した音であること、また、その前後関係も認識できることが判っている[3]。しかし、連続して次々に呈示される複数の音に対し、その前後関係を正しく知覚するためには、少なくとも約100msの間隔が必要であるとされている[4]。

独立して認識される1つの特徴に対し、$D_{music}(t)$ の時間変化に適応した閾値を適切に設定するために、あるタイムステップ t について、その前後時間において分解能の時間区間に含まれる合計5タイムステップから、$D_{music}(t)$ の平均値 $C_{music}(t)$ を求め、それを全タイムステップにおいて計算する。また、全タイムステップの合計時間を T としたとき、算出した平均値を元に式(3)より分散 σ^2_{music} を求め、式(4)より閾値 $C'_{music}(t)$ を求める。このとき、Aは音楽の種類により異なる定数であり、実験による経験則から求める。

$$\sigma^2_{music} = \frac{1}{T}\int_0^T (D_{music}(t) - C_{music}(t))^2\, dt \quad (3)$$

$$C'_{music}(t) = C_{music}(t) + A\,\sigma \quad (4)$$

そして、閾値よりも $D_{music}(t)$ が大きくなる時刻を検出する。以下、i 番目に検出された時刻を t_i とする。

特徴の検出処理において、分解能の時間区間内に複数の特徴が検出されることを防ぐため、閾値を超える箇所が検出された際、それ以前の分解能の時間区間に特徴が存在していなければ特徴とし、そうでない場合は取得しない。図2において、矢印bで示されるタイムステップでは、$D_{music}(t)$ は閾値を超える値を示しているが、100ms前に特徴が存在しているため、特徴として検出されない。

そして、検出された特徴に、閾値からの差分を各特徴の持つ重み S_{t_i} として与える。

2.2 映像情報からの特徴検出手法

人間の視覚においては、色や明るさが重要な要素を占めていることから、映像中のカットの切り替わりや、カット以外の色彩や明るさの大きな変化を動画における特徴構造として検出する[5]。

まず、動画中の j および $j-1$ 番目のフレームにおける全ピクセルについて、RGB各要素を調べ、両フレームのヒストグラム(RGB各256階調)を計算する。式(5)より、両ヒストグラムのRGB各要素 $\mathbf{X}_y(j)$、$\mathbf{X}_y(j-1)$ においてその差分から、ヒストグラム変化 $\mathbf{D}_{histo}(j)$ を全フレームについて算出する。そして $\mathbf{D}_{histo}(j)$ について256階調の総和 $D_{video}(j)$ を式(6)より求める。

$$\mathbf{D}_{histo}(j) = \sum_{(y=R,G,B)} {}_y|\mathbf{X}_y(j) - \mathbf{X}_y(j-1)| \quad (5)$$

$$D_{video}(j) = \sum \mathbf{D}_{histo}(j) \quad (6)$$

この値が大きく変化する箇所を特徴として検出するための閾値を求める。音楽の特徴検出と同様、$D_{video}(j)$ の時間変化に適応した閾値を設定する。これを各時間において、分解能の時間範囲内に含まれる $D_{video}(j)$ の平均値から求める。図3に検出の概要図を示す。横軸は時刻、縦軸は $D_{video}(j)$ を表しており、$D_{video}(j)$ が閾値よりも大きくなるフレームを特徴として検出し、その時刻を取得する。

一般的なテレビや映画などの映像は、1秒間に24〜30フレームであり、人間の残像の視覚効果により滑らかな映像に見えている。視覚の時間的分解能については諸説あるが、外界から入力された視覚的情報が、神経伝導路を経て大脳皮質の視覚野で知覚されるまでの伝導にかかる時間が、約100msであり、また大脳皮質の視覚野の時間的な二点弁別能が100ms程度であることから、それよ

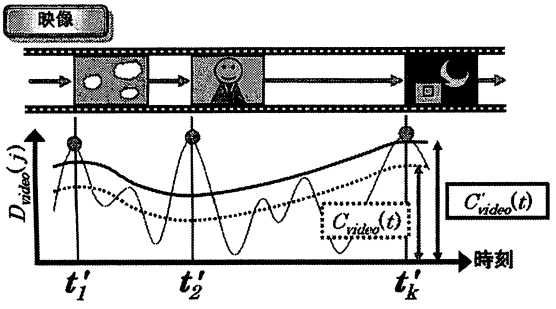

図 3　映像情報からの特徴検出

り短い時間で完結した現象は正確に認識できないとされている。

独立して認識される 1 つの特徴に対し、$D_{video}(j)$ の時間変化に適応した閾値を適切に設定するために、あるフレーム j の前後時間の分解能の時間区間に含まれる前後フレームから、$D_{video}(j)$ の平均値 $C_{video}(j)$ を求め、それを全タイムステップにおいて計算する。また、算出した平均値を元に分散 σ^2_{video} を求め、これを平均値に加え、閾値 $C'_{video}(j)$ とする。そして、$D_{video}(j)$ が閾値よりも大きくなるフレームの検出を行う。

2.1 節で述べた音楽情報からの特徴検出と同様に、分解能の時間区間内に複数の特徴が検出されることを防ぐため、閾値を超えるフレームが検出された際、それ以前の分解能の時間区間に閾値を超えるフレームが存在していない事を条件とし、特徴の検出を行う。そして、k 番目に検出されたフレームの時刻を t'_k とする。また同時に、閾値からの差分を、各特徴のもつ重み S'_{t_k} として求める。

3.　音楽・映像の特徴のマッチング

コンテンツ制作時、映像の一部を追加切除することにより、映像と音楽の時間的構造のマッチングを行う従来の手法では、時間と労力がかかるだけでなく、コンテンツ制作者が用意した映像素材の内容を変更せざるを得ないという制約が生じていた。そこで、映像再生速度の調整により、映像素材全てを用いたまま、映像の特徴を音楽の特徴に一致させたコンテンツの構成を行う。視覚はその時間的分解能と反応速度や残存時間 (残像) の効果により、聴覚よりも広い心理的時間幅をもつことから、音楽に比べ映像の速度変化に対する人間が知覚できる違和感は小さい [6]。よって、音楽の特徴の時間的構造を基準に、映像の速度変化によるマッチングを行う。

このとき、マッチングに用いる特徴を決定する際に、基準となる特徴に対する時間的な近さだけでなく、両特徴に与えられた重みを考慮し、最も適当と思われる特徴を選択する。両者の時間的な接近の度合いのみを考慮した場合、図 4 に示すように、2 つの映像の特徴 S'_{t_1}、S'_{t_2} に

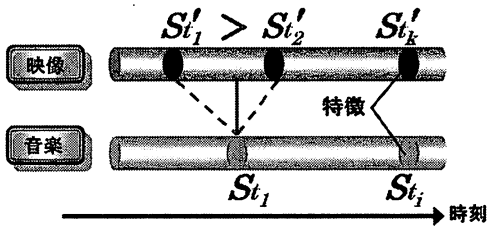

図 4　違和感を感じない許容限界の時間幅

おいて、$S'_{t_1} > S'_{t_2}$ であった際も、基準 S_{t_1} に対し、より時間的に接近している重みの小さい特徴について、マッチングを行ってしまう。

また、両特徴を時間的に完全に一致させるのではなく、音楽と映像との時間的なずれに対し、人間が違和感を感じない許容限界を考慮し、映像の変化を最小にとどめる。

音と映像の同期がどの程度ずれていると違和感を感じるかに関しては、様々な視聴覚刺激を用いた実験が行われており、物を叩く映像とその音という、実際に対応関係にある視聴覚刺激の同期に関しては、音が映像より進んでいる場合には約 50〜100ms、音が映像より遅れている場合には約 100〜180ms がそれぞれ許容限界であるとされている [7]。これに対し、音楽と映像というように、現実の世界での対応関係のない視聴覚刺激の同期に関して、許容限界を求めるため実験を行った。

音楽の特徴が映像の特徴より進んでいる場合、および遅れている場合について、20ms 間隔ずつ時間的なずれを大きくしたサンプル動画を用意し、被験者に視聴してもらい、違和感を感じない許容限界の時間幅の統計をとった。その結果を図 5 に示す。

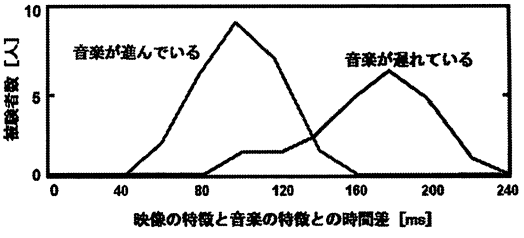

図 5　違和感を感じない許容限界の時間幅

横軸が音楽と映像の時間的なずれ幅、縦軸が許容限界であると感じた被験者の人数であり、音が映像より進んでいる場合には約 60〜140ms、音が映像より遅れている場合には約 100〜220ms がそれぞれ許容限界であり、対応関係にある視聴覚刺激に比べ、同期に対し広い許容範囲を持つことが分かる。マッチング時の計算においては、この値を許容限界 t_l とし、音が映像より進んでいる場合、および遅れている場合において、それぞれ正の値、および負の値を取るとする。

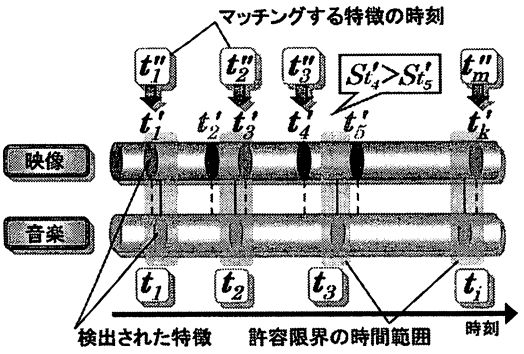

図 6　特徴のマッチング

まず、検出された音楽と映像それぞれの特徴の中から、マッチングさせるものを決定する。図 6 に特徴のマッチングにおける概要図を示す。横軸は時刻を示しており、基準となる音楽の特徴の時刻 t_i について、その前後一定時間内に存在している映像の特徴の中で、式 (7) で表される特徴間のスコア S_{ik} が最も大きな値となる組み合わせを見つけ、マッチングを行う。

$t'_k > t_i + t_l$ の場合、各特徴に与えられた重みが大きいほど、また特徴間の時間差が小さく許容限界に近いほどスコアは大きな値をとる。また、$t'_k \leqq t_i + t_l$ の場合は基準 t_i に対し、許容限界の時間範囲内に位置しており、違和感を感じない程度時間的に接近していることから、時間差による寄与を考慮しないスコアとする。

$$S_{ik} = \frac{1}{\{max(t_l, |t'_k - t_i|)\}^2} S_{t_i} S'_{t_k} \qquad (7)$$

次に、選び出された映像の特徴の時刻 t'_k (図中にて濃い文字で表記) を順に t''_m としたとき、t''_m および t''_{m-1} 間の時間、そして、t_i および t_{i-1} 間の時間から、その区間における映像の再生速度の比率 R を式 (8) より計算する。

映像の再生速度の変化を最小限に抑えるため、$t''_m > t_i + t_l$ の場合、違和感を感じない許容限界の時刻 $t_i + t_l$ に対し同期を行う。逆に、$t''_m \leqq t_i + t_l$ の場合は基準 t_i に対し、違和感を感じない程度に、時間的に充分接近していると判断し、位置を変化させない。

式 (8) における t_I は、$t''_m > t_i + t_l$ の場合、許容限界の時刻を基準とするため $t_i + t_l$ となり、$t''_m \leqq t_i + t_l$ の場合は同期を行わないため $t_I = t_m$ となる。また、t_M は、t''_{m-1} にて同期が行われた際に変化した後の時刻 $t_M = t_{i-1} + t_l$ であり、同期が行われず、変化しなかった場合は、$t_M = t''_{m-1}$ のままである。よって、$t''_{m-1} \leqq t_{i-1} + t_l$ かつ $t''_m \leqq t_i + t_l$ のとき、$R = 1$ であり、再生速度は変化させずに済むことになる。

$$R = \frac{t_m - t_M}{t_I - t_M} \qquad (8)$$

こうして算出した速度にて映像を再生し、音楽と合成することによりコンテンツを構成した。

4. 評価実験および結果

提案手法の有効性を確認するために、提案手法の適用前後のコンテンツについて、音楽と映像の特徴が同期し、調和度が高く、印象の強いコンテンツになっているか、被験者による評価実験を行った。

表 1 に示すように、3 種類の異なるコンテンツ A,B,C それぞれについて、(a) 入力音楽と映像を単純に重ね合わせた場合、(b) 時間的に最も接近している両特徴についてその時刻を完全に一致させるマッチングを行った場合、(c) 特徴に与えられた重みも考慮したスコアを計算してマッチングを行った場合、および、(d) 許容限界を考慮してマッチングを行った場合の、4 つの条件にて組み合わせた、サンプルコンテンツを用意した。

表 1　実験に用いるサンプルコンテンツの種類

	条件 a	条件 b	条件 c	条件 d
コンテンツ A	Aa	Ab	Ac	Ad
コンテンツ B	Ba	Bb	Bc	Bd
コンテンツ C	Ca	Cb	Cc	Cd

これを被験者に視聴してもらい、図 7 に示すような 2 つの評価基準に対する 7 段階の両極尺度を用意し、各サンプルコンテンツに対する評価を数字で答える、SD (Semantic Differential) 法 [8] によって評価実験を行った。また、表 2 に示すように、6 つの各評価尺度を変量として調和およびインパクトの因子に集約する、因子分析による評価を行った。表内の各数字は評価尺度と因子との関係の強さを表した因子負荷であり、値が大きいものほど、両者の関係が強いことを示す。

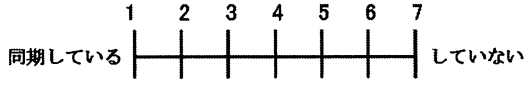

図 7　SD法による評価実験における評価尺度

表 2　評価尺度と各調和因子における因子負荷

評価尺度	調和因子	インパクト因子
同期している-していない	0.854	0.146
同期が妥当-妥当でない	0.844	0.156
違和感がない-ある	0.816	0.184
印象深い-薄い	0.072	0.928
迫力のある-ない	0.163	0.837
力強い-弱々しい	0.369	0.631

それぞれのサンプルコンテンツについて、各評価尺度ごとの評価値を図 8 にまとめた。同期については、マッ

チング操作により、マッチング前では平均評価値が 1.2～1.5 だったのに対し、マッチング後では約 6.0～6.8 と約 75 %の向上が見られる。また、同期させる特徴の組み合わせにおける妥当性については、スコア計算の導入により、導入前に比べ約 25 %の向上が見らる。また、許容限界を考慮したマッチングを行うことにより、違和感に関しての評価値が、約 20 %向上している。全体として提案手法により約 65～75 %の向上が見られ、それぞれの有効性が確認できる。

同様に、各サンプルコンテンツについて、調和およびインパクトそれぞれの因子得点を図 9 にまとめた。調和因子得点において、マッチング操作を行ったものは、調和度が大きく向上している。音楽と映像の時間的構造の同期が、調和感をもたらすことが確認できる。インパクト因子においても、音楽と映像の時間的構造の同期の効果が認められる。さらに、マッチングを行ったコンテンツの中でも、スコア計算を導入したもの、および、許容限界を考慮したものは、相対的に因子得点が高く、妥当と感じられる特徴同士を同期させ、また映像の速度変化が小さく、違和感が小さいほど、調和感をもたらし、強い印象を与えることが分かる。

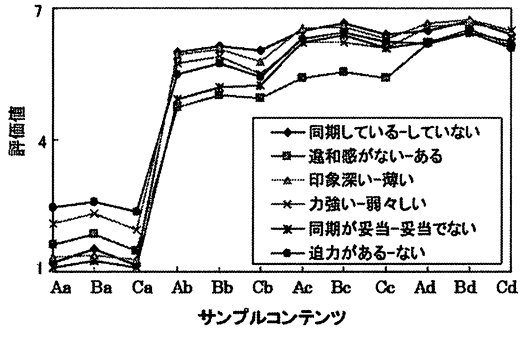

図 8　各評価尺度ごとの評価値

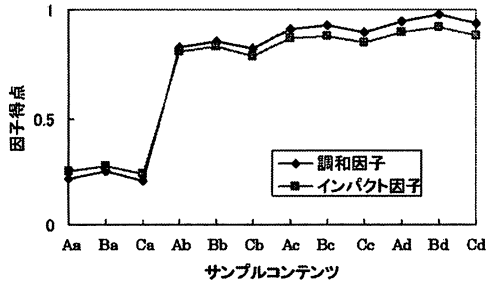

図 9　調和およびインパクトそれぞれの因子得点

5.　まとめと今後の課題

動画および音楽における特徴の時間的構造を、人間の視聴覚における時間的分解能を考慮して、時間変化する閾値を設定し検出した。また、両特徴の時間的な近さに加え、各特徴に与えた重みを考慮したスコア計算と、音楽と映像の時間的なずれに対する人間の許容限界を考慮したマッチングを行った。そして、提案手法により、コンテンツの調和感および印象深さが向上することが確認された。

本稿では視覚が聴覚よりも広い心理的時間幅をもつことから、映像を変化させ音楽に同期させる手法を採用した。今後は、音楽の速度変化において人間が違和感を感じる許容限界に着目し、音楽を変化させることで同期を行う手法について検討したい。

文　献

[1] 岩宮眞一郎 “音楽と映像のマルチモーダル・コミュニケーション” 九州大学出版会, 2000

[2] M. Goto “An audio-based real-time beat tracking system for music with or without drum-sounds” Journal of New Music Research, 30, 2, pp. 159-171, 2001

[3] IJ Hirsh and CE Sherrick “Perceived order in different sense modalities” Journal of Experimental Psychology, 62, 423-432, 1961

[4] ten Hoopen G. and Vos J. “Attention switching and patterns of sound locations in counting clicks” Journal of Experimental Psychology, Human Perception and Performance, 7, 342-355, 1981

[5] Olivier Gillet, Gael Richard “Comparing Audio and Video Segmentations for Music Video Indexing” International Conference on Acoutsics, Speech, and Signal Processing, vol. 5, pp. 21-24, 2006

[6] 長嶋洋一 “音楽的ビートが映像的ビートの知覚に及ぼす引き込み効果” 芸術科学会論文誌, vol.3 No.1 芸術科学会, 2003

[7] 黒住幸一, 岡田清孝 “テレビの映像と音声の相対時間差に関する検討” 日本音響学会講演論文集 (春期), 461-452, 1996

[8] 岩下豊彦 “SD 法によるイメージの測定 -その理解と実施の手引-” 川島書店, 1983