J-049

映像コンテンツ解析によるＢＧＭサウンドトラックの自動生成
Automatic Synthesis of Background Music Track Data by Analysis of Video Contents

茂出木 敏雄†
Toshio Modegi

1．まえがき

動画像を主体とした番組を制作する際にセリフ・ナレーションといった音声とともに、音響効果としてＢＧＭや効果音が挿入されることが多い。ドラマやドキュメンタリーなどではストーリーに合わせて、映像を盛り上げるために特に楽曲の選定は重要で、基本的にサウンドデザイナの手作業に頼らざるを得ず、ストーリーに合わせて新規に作曲されることもある。しかし、環境映像、教育・情報番組のＣＧアニメーションなどのバックに流す音楽では、ストーリー性が要求されないことがあり、楽曲選定にあまり凝らず、むしろ著作権料や編集コストを安価に抑えることが望まれる。本研究は後者のようなケースを対象として、映像データに相応しい安価なＢＧＭを自動的に選曲し、映像に音を付加する作業を軽減するシステムを構築することを目的としている。

これまで文献[1]のように、音楽素材に相応しい映像を自動的に付加したり、音響信号を解析してＣＧで映像化する試みはなされていたが、映像から音楽を付加するアプローチはあまり試みられなかった。本研究は、文献[2]で提案したＢＧＭを大量に生成する技術を基盤に、音響パラメータで所望のＢＧＭを検索できるようにする感性検索技術と、ソース映像の動画解析パラメータを音響パラメータに変換しＢＧＭを検索できるようにする映像解析技術を組合せて実現したので、その概要を述べる。

2．ＢＧＭを大量に生成する音楽合成技術

音楽は少なくともメロディー、コード、リズムの３要素から構成されるが、一般の楽曲は各要素が更に複数の楽器パートに分割されているため、更に多くのトラックを用いて素材録音や編集が行われる。ここで、複数の楽曲間で互いにパート素材を交換しても、音楽的に整合性がとれるように各素材を制作できると、（パート数）×（楽曲数）のマトリックス状の基本素材を準備するだけで、（楽曲数）の（パート数）乗の種類の音楽を生成可能になる。

図１の右端に示したような、５つのパート素材で構成される楽曲を５種類制作し、25 種類の音楽素材をマトリックス状に準備すると、3125 通りの音楽を生成できる。各音楽素材は総演奏時間長だけ波形データを高精細ステレオでデジタル化しておけば、プレーヤ側で選定された波形データを先頭から合成することにより高精細なＢＧＭを再生できる。また、文献[2]のように各素材にロスレス圧縮を施せば、ＣＤ品質を維持しながら約６分の素材 25 点全て一式をＣＤ１枚に収納することができる。

†大日本印刷株式会社　研究開発センター　企画開発部
Research & Development Center, Dai Nippon Printing Co., Ltd.
(e-mail: Modegi-T@mail.dnp.co.jp)

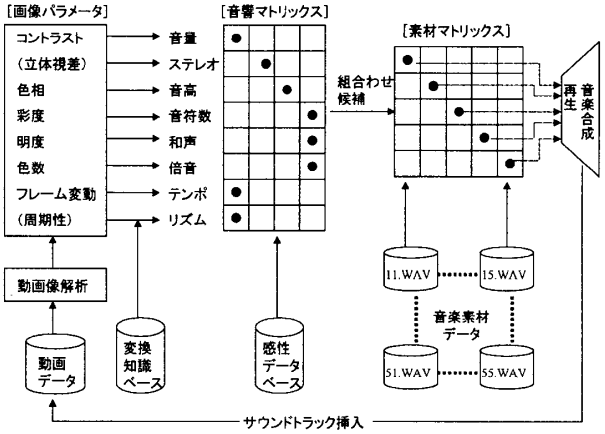


図１　ＢＧＭサウンドトラックの合成システム構成

3．楽曲の感性検索技術

前述の 3125 曲から所望の音楽的特徴をもつ素材合成パターンを指定する方法として、図１のように、音楽を客観的に特徴付けられる音響パラメータの大小を指定できる音響マトリックスを素材マトリックスのフロントエンドに配置させ、音響パラメータとしては、音量、ステレオ、音高、音符数、和声、倍音、テンポ、リズムの８項目とした。テンポについては、音楽的な整合性を維持する都合上、音楽素材レベルでは全て同一に設定しているが、合成波形の見かけ上のテンポは楽器編成により変化する。そのため、3125 通りの全合成波形データに対して以下計算式で音響解析を行い、前記全８項目について定量化した感性データベースを構築しておく。そうすると、音響マトリックスで指定された項目に合致する素材マトリックスを検索できる。また、「元気が出る音楽」といった感性キーワードと音響パラメータへの変換ルールを設定しておけば、感性キーワードでも検索が可能になる。

［音響パラメータの計算方法］

波形データ X(i), i=0,S−1 に対して、以下２項目を算出する。

(1) 音量パラメータ（音楽のダイナミックレンジを示す）

$$Pv=20\bullet\log_{10}\{\Sigma_{i=0,S-1}|X(i)|\}/S \quad (1)$$

(2) ステレオパラメータ（左右空間的な広がりを示す）

$$Ps=20\bullet\log_{10}\{\Sigma_{i=0,S2-1}|R(i)|\bullet 2/S \quad (2)$$

If $|X(i*2)|\geq|X(i*2+1)|$ then $R(i)=X(i*2)/X(i*2+1)$

If $|X(i*2)|<|X(i*2+1)|$ then $R(i)=X(i*2+1)/X(i*2)$

次に、波形を解析フレーム k（ステレオ 8192 サンプル）単位で周波数解析を行ったスペクトル Zk(n), n=0,...,N−1 (127), k=0,...,K−1 に対して、以下４項目を算出する。

(3) 音高パラメータ（音楽の平均的な音域を示す）

$$Pp=[\Sigma_{k=0,K-1}\{\Sigma_{n=0,N-1}\, n\bullet Zk(n)/\Sigma_{n=0,N-1}\, Zk(n)\}]/K \quad (3)$$

(4) 音符数パラメータ（楽器数、音色の豊かさを示す）

$$Pn=\{\Sigma_{k=0,K-1}\, C(k)\}/K \quad (4)$$

C(k)はフレームｋにおいて、Zk(n)>設定しきい値となるノートナンバーの総数である。

(5) 和声パラメータ（音楽が短調系・長調系かを示す）

$$Ph=[\Sigma_{k=0,K-1}\{Zk(m+4)-Zk(m+3)+Zk(m+16)-Zk(m+15)+Zk(m-8)-Zk(m-9)\}/6]/K \quad (5)$$

Zk(m)が最大値となるノートナンバーm より上下オクターブ音を含めて、長三度の音程（+4 半音）の成分を加算し、短三度の音程（+3 半音）の成分を減算する。

(6) 倍音パラメータ（倍音の豊かさを示す）

$$Po=[\Sigma_{k=0,K-1}\{\Sigma_{n=0,N-1}(Zk(n)+Zk(n+12)+Zk(n+19)+Zk(n+24)/4)\}]/K \quad (6)$$

更に、波形データを時間軸にステレオ各 60 サンプル単位に間引き、解析フレームｋ単位で周波数解析を行ったスペクトル Zk(n), n=0,...N-1 (127),k=0,...,L-1 に対して、Zk(n)が大きい最上位２つのノートナンバーを M1. M2 (M1<M2)とするとき、以下２項目を算出する。

(7) テンポパラメータ（平均的な基本ビートを示す）

$$Pt=\{\Sigma_{k=0,L-1}440\bullet 2^{(M2-64)/12}\}/L \quad （単位はＢＰＭ） \quad (7)$$

(8) リズムパラメータ（平均的な拍子を示す）

$$Pr=\{\Sigma_{k=0,L-1}100\bullet 2^{(M1-M2)/12}\}/L \quad (8)$$

4. 映像の動画像解析技術

与えられた動画像データに対して画像解析を行い、コントラスト、立体視差、色相、彩度、明度、色数、フレーム変動、周期性の８つの画像パラメータを算出し、図１に示すように８つの音響パラメータに変換し感性検索を行う。画像パラメータと音響パラメータとの対応付けは、画像の色相（光の波長）と音響の振動数（音高）と、画像の明度と音響の和声（短調・長調の調性）いう具合に、ヒトの主観が入らないように物理的根拠に基づいて決めている。ただし、立体視差と周期性については、ステレオ画像や、ループ動画像のような特殊な動画像に対してのみ使用する。そして、検索された素材マトリックスに従ってマルチトラック合成された音楽データを動画像ファイルのサウンドトラック部分に挿入する。

［動画像の解析方法］

フレーム f の x, y 座標における RGB 画素値を R(f,x,y)、G(f,x,y)、B(f,x,y)とし、以下式で HSV 色空間に変換を行った結果を H(f,x,y)、S(f,x,y)、V(f,x,y)とする。立体視差と周期性を除く各パラメータについて、以下算出方法を示す。

(1) 色数

R(f,x,y)、G(f,x,y、B(f,x,y)の階調を 16 段階に変換し、ｆフレーム内全画素の RGB 組合せの画素数をカウントし、０以外のカウント数になる RGB 組合せ数を求める。全フレームに対して色数の平均値を求め、３段階の倍音パラメータに変換する。

(2) フレーム変動

ｆフレームの R(f,x,y)、G(f,x,y)、B(f,x,y)と、f-1 フレームの R(f-1,x,y)、G(f-1,x,y)、B(f-1,x,y)とのフレーム差分を以下のように、全画素において計算し、平均値を求める。

$$D(x,y)=[\{R(f,x,y)-R(f-1,x,y)\}^2+\{G(f,x,y)-G(f-1,x,y)\}^2+\{B(f,x,y)-B(f-1,x,y)\}^2]^{1/2} \quad (9)$$

先頭フレームを除く全フレームに対して、フレーム差分の平均値を求め、３段階のテンポパラメータに変換する。

(3) 色相、彩度、明度

ｆフレーム内の全画素において、H(f,x,y)、S(f,x,y)、V(f,x,y)の平均値を求める。全フレームに対して、色相Ｈ、彩度Ｓ、明度Ｖの平均値を求め、３段階の音高、音符数、和声の各パラメータに変換する。

(4) コントラスト

ｆフレーム内の全画素において、V(f,x,y)の最大値と最小値を求め、その差分Ｃ＝最大値－最小値をコントラストとする。全フレームに対して、コントラストＣの平均値を求め、３段階の音量パラメータに変換する。

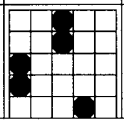
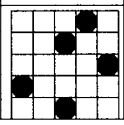
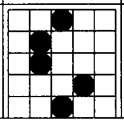
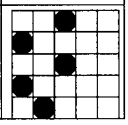
	Video frame				
Image Parameters	contrast	168	132	260	11
	hue	85	88	25	118
	saturation	161	147	175	48
	brightness	252	236	229	254
	color variation	149	150	105	150
	frame change	56	56	20	34
Acoustic Parameters	volume	3	3	3	3
	pitch	2	2	3	1
	notes	2	2	1	2
	harmony	3	2	3	1
	overtone	2	2	2	2
	tempo	2	2	1	1
BGM Matrix	track1				
	track2				
	track3				
	track4				
	track5				

図２　動画クリップに対する BGM 自動挿入実験例

5. あとがき

図１の構成に基づいて BGM 合成ソフトウェアを開発し、20 秒程度の動画クリップに、図２のように BGM を自動挿入することができた。選択された BGM の候補と、それ以外の BGM パターンを幾つかピックアップして、映像制作担当のサウンドデザイナに聞かせると、少なくとも後者の方が前者より映像に適合するというケースは全く無く、感性的にも良くマッチしていた。今後は、映像中のシーン変化を検出して、BGM コンテンツをスムーズに切り替えられる技術を開発し、ストリーミング映像に対してリアルタイムに適切な BGM を選択し付加できるシステムを開発する予定である。

本研究のＢＧＭを合成する技術については、財団法人デジタルコンテンツ協会の平成 15 年度「ブロードバンドコンテンツのブレークスルー技術等開発支援事業」の一環として、経済産業省より助成を受け、（株）マトリックスミュージックと共同で推進したものである。本事業にご尽力いただいた方々、共同開発先、ＤＮＰ情報システム（株）、当社Ｃ＆Ｉ事業部の皆様に謝意を示す。

参考文献

[1] Jonathan Foote, Matthew Cooper and Andreas Girgensohn: “Creating music videos using automatic media analysis”, Proceedings of the tenth ACM international conference on Multimedia, New York, USA, pp.553-560, (Nov. 2002).

[2] 茂出木敏雄：「ロスレス圧縮技術のマルチトラック型環境音楽再生装置への応用」情報処理学会　音楽情報科学　研究報告，Vol.55，No.6, pp.33-38 (May 2004).