(19)日本国特許庁(JP)

(12)特許公報(B1)

(11)特許番号
特許第7635953号
(P7635953)

(45)発行日　令和7年2月26日(2025.2.26)

(24)登録日　令和7年2月17日(2025.2.17)

(51)国際特許分類
Ｇ１０Ｈ　1/00　(2006.01)

ＦＩ
Ｇ１０Ｈ　1/00　１０２Ｚ

請求項の数　8　（全25頁）

(21)出願番号　特願2024-205849(P2024-205849)
(22)出願日　令和6年11月26日(2024.11.26)
審査請求日　令和6年11月27日(2024.11.27)
早期審査対象出願

(73)特許権者　515120143
井澤 佑斗
東京都小金井市貫井北町５－１６－８
(73)特許権者　524350250
加藤 健資
新潟県燕市小高６２４５－１
(72)発明者　加藤 健資
新潟県燕市小高６２４５－１
(72)発明者　井澤 佑斗
東京都小金井市貫井北町５－１６－８
審査官　山下 剛史

最終頁に続く

(54)【発明の名称】　音を作成する方法及び装置

(57)【特許請求の範囲】
【請求項１】
音を作成する方法であって、
装置が、
動画内のオブジェクトの動作の特徴である動作特徴を特定し、
類似する動作特徴間の時間を単位時間として、類似する動作特徴から生ずる２以上の単位時間の中から、相対的に短い単位時間を拍とする音楽を生成することをAIに要求する、
方法。
【請求項２】
音を作成する方法であって、
装置が、
動画内のオブジェクトの動作の特徴である動作特徴を特定し、
類似する動作特徴間の時間を単位時間として、類似する動作特徴から生ずる２以上の単位時間の中から、特定された単位時間の出現頻度が相対的に高い単位時間を拍とする音楽を生成することをAIに要求する、
方法。
【請求項３】
音を作成する方法であって、
装置が、
動画内のオブジェクトの特徴を特定し、

２以上の特徴があった際には、少なくとも、特徴が認められる時間が相対的に長い特徴に基づいて音楽を生成することをAIに要求する、
方法。
【請求項４】
音を作成する方法であって、
装置が、
動画内のオブジェクトの特徴を特定し、
２以上の特徴があった際には、少なくとも、特徴が認められる出現頻度が相対的に高い特徴に基づいて音楽を生成することをAIに要求する、
方法。
【請求項５】
音を作成する装置であって、
動画内のオブジェクトの動作の特徴である動作特徴を特定し、
類似する動作特徴間の時間を単位時間として、類似する動作特徴から生ずる２以上の単位時間の中から、相対的に短い単位時間を拍とする音楽を生成することをAIに要求する、
装置。
【請求項６】
音を作成する装置であって、
動画内のオブジェクトの動作の特徴である動作特徴を特定し、
類似する動作特徴間の時間を単位時間として、類似する動作特徴から生ずる２以上の単位時間の中から、特定された単位時間の出現頻度が相対的に高い単位時間を拍とする音楽を生成することをAIに要求する、
装置。
【請求項７】
音を作成する装置であって、
動画内のオブジェクトの特徴を特定し、
２以上の特徴があった際には、少なくとも、特徴が認められる時間が相対的に長い特徴に基づいて音楽を生成することをAIに要求する、
装置。
【請求項８】
音を作成する装置であって、
動画内のオブジェクトの特徴を特定し、
２以上の特徴があった際には、少なくとも、特徴が認められる出現頻度が相対的に高い特徴に基づいて音楽を生成することをAIに要求する、
装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音を作成する方法に関する。
【背景技術】
【０００２】
この項における供述は、本開示に関する背景の情報を提供するだけであり、先行技術を必ずしも構成しない。
【０００３】
特許文献１には、音楽生成システムが公開されている。
【先行技術文献】
【特許文献】
【０００４】
【文献】特開2006-154777号公報
【発明の概要】

【発明が解決しようとする課題】
【０００５】
しかし発明者は少なくとも上記実施形態には、動画に基づき音を作成する方法が無いという短所が存在すると認識した。
【課題を解決するための手段】
【０００６】
少なくとも一つの本開示は、
音を作成する方法であって、
装置が、
動画内のオブジェクトの動作の特徴である動作特徴を特定し、
類似する動作特徴間の時間を単位時間として、単位時間を拍とする音楽を生成することをAIに要求する、
方法
を提供する。
【発明の効果】
【０００７】
本構成では少なくとも、産業利用可能な音楽を動画に基づいて生成できるという有用性がある。
【０００８】
本開示のこれらおよび他の態様、特徴、および利点は、以下の図面と併せて取られる好ましい実施形態および態様の以下の詳細な書面の説明から明らかになるが、その変形および修正は、本開示の新規概念の精神および範囲から逸脱せずに実施され得る。本開示におけるある実施形態における態様は、矛盾しない限りにおいて、本開示される別の実施形態における態様の１以上と組み合わせ、又は置き換えることができる。
【発明を実施するための形態】
【０００９】
以下の開示において、提示された主題の異なる特徴を実施するための多くの異なる実施形態や実施例を提供する。本開示を平易にするために、構成部品や配置の具体例を以下に開示する。もちろんこれらは単なる例であり、限定的であることを意図するものではない。例えば、第１の特徴が、続いて開示する第２の特徴に覆われる、あるいはこれと接する構造は、第１の特徴および第２の特徴が直接接触するように形成されている実施形態とともに、 第１の特徴と第２の特徴との間に付加的な特徴を形成して、第１の特徴と第２の特徴とが直接接触しないようになっている実施形態を含んで良い。さらに、本開示では、さまざまな例において参照番号および/または文字を反復している場合がある。このように反復するのは、簡潔明瞭にするためであり、それ自体が、さまざまな実施形態および/または説明されている構成との間に関係があることを必要とするものではない。さらに、第１の要素が第２の要素に「連結されている」または「結合されている」と記述するとき、そのような記述は、第１の要素と第２の要素とが互いに直接的に連結または結合されている実施形態を含むとともに、第１の要素と第２の要素とが、その間に介在する１以上の他の要素を有して互いに間接的に連結または結合されている実施形態も含む。
【００１０】
本明細書中で使用されるように、「少なくとも１の（at least one of）」という記載は、例示するすべての変形例を包合する。例えば、「AとBとCの少なくとも１の（comprises at least one of !, B, or C）」の記載は、「AとBとCと、これらのコンビネーション（consisting of A, B, C and combinations thereof）」と同義である。そして、A、B、C、A+B、A+C、B+C、A+B+Cの考えうる限りの全ての変形例を抱合する。
【００１１】
本開示において、機械、電子オペレータまたはコンピュータを使用する開示は、方法、記録媒体、装置、またはプログラムの実施形態を含むことができる。本明細書で使用される「AはBである」という記述は、矛盾がない限り、あるいは本明細書に別段の記載がない

限り、「AはBを含む」と置き換えることができる。
【００１２】
本開示における用語は、特許請求の範囲に記載された用語を含めて、明細書に記載された記載や図面を考慮して解釈でき、さらには、本開示における示唆と矛盾がない限り、過去もしくは現在、未来において、市民の一人以上がそのように呼称し、表示し、理解しもしくは実行した、またはその可能性のある事柄をもとに解釈できる。
少なくとも１以上の実施形態で用いる作動方法について、以下の実施形態を取ることができる。少なくとも１以上の実施形態をよく説明するJP6456303の記載を引用して説明する（以下、引用開始）。
【００１３】
本明細書中で使用されるように、用語「コンピューター」は、当技術分野で知られているように、プロセッサー、メモリー、例えばハードドライブ、ディスクドライブ又はフラッシュドライブ又はメモリースティック、又は他の非一時的なコンピューター可読媒体又は非一時的記憶装置などの、少なくとも１つの情報記憶／検索装置、例えば、キーボード、マウス、ポイント及びタッチのデバイス、タッチスクリーン、又はマイクなどの、少なくとも一つの入力装置、及び、よく知られたコンピュータースクリーンのようなディスプレイ構造を概略的に含む。加えて、コンピューターは、有線又は無線接続などの、１以上のネットワーク接続を含み得る。当技術分野で知られるように、そのようなコンピューターまたはコンピューターシステムは、上に列挙されるものを多かれ少なかれ含んでもよく、例えばタブレットコンピューターやスマートデバイスに限定されないが、他の電子メディアや電子デバイスを包含する。
【００１４】
本明細書で使用されるように、用語「クラウド」または「クラウドコンピューティング」は、すべてのコンピューティングリソースが共有される集中型の（ｃｅｎｔｒａｌｉｚｅｄ）および仮想型（ｖｉｒｔｕａｌｉｚｅｄ）のコンピューティング設備を指す。 アプリケーションシステムやサブシステムについて、それらはすべて「クラウド」内にあるため、特定のマシンを指すことはもはやできない。
【００１５】
本明細書中で使用されるように、用語「分散型インターネットサービスシステム」は、様々なコンピューティング環境で実行するために、インターネットアプリケーションを変換する分散型インターネットサービスプラットフォームを指す。ＤＩＳシステムは、コンポーネント分散型サーバー（Ｃｏｍｐｏｎｅｎｔ Ｄｉｓｔｒｉｂｕｔｉｏｎ Ｓｅｒｖｅｒ）／資産分散サーバー（Ａｓｓｅｔ Ｄｉｓｔｒｉｂｕｔｉｏｎ Ｓｅｒｖｅｒ）を経由して、コンテンツ、データ及びロジックを含むインターネットアプリケーションを、適切な範囲に（ｔｏ ｗｈａｔｅｖｅｒ ｅｘｔｅｎｔ ａｐｐｒｏｐｒｉａｔｅ）、およびネットワークに沿って、任意の数と任意の種類の機器に対して配信する。ＤＩＳを介して、インターネットアプリケーションを、各ユーザーのニーズに基づいたサービスで、ホストしおよび一元管理し、その完全性を維持しながらユーザーのデバイスまたは近くの場所で局所的にキャッシュし実行することができる。Ｗｅｂ対応のコンピューティングデバイスは、ＤＩＳソフトウェアーでアップグレードして、分散型のインターネットサービスを楽しみ実行するようなＤＩＳ対応になることができる。分散型インターネットサービスシステムは、米国特許第７１３６８５７号、第７１５００１５号、第７１８１７３１号、第７２０９９２１号、第７４３０６１０号、第７６８５１８３号、第７６８５５７７号、第７７５２２１４号、第８３２６８８３号、第８３８６５２５号、第８４４３０３５号、第８４５８１４２号、第８４５８２２２号、第８４７３４６８号、第８５２７５４５号、および第８６５０２２６号、および米国特許公開第２０１２０００５２０５号、および第２０１３００９１２５２号の特許ファミリーの何れか一つに完全に記載されており、これらすべては本発明と同様に、オーピー４０，ホールディングス，インク．によって共有して所有されており、これらすべてが引用により組み入れられる。（以上、引用おわり）
【００１６】

少なくとも１以上の実施形態で用いる作動方法について、分散型インターネットを使用しない従来のインターネットの方式について、以下の実施形態を取ることができる。少なくとも１以上の実施形態をよく説明するJP7113047の記載を引用して説明する（以下、引用開始）。

【００１７】

本明細書で具体的に開示される事項を含む実施形態は、人工知能を基盤として実際に人間と会話するような形態で実現された自動応答システムを提供することができ、これによってユーザとのより自然な通話を実現しながら、問い合わせ、予約、配達注文などを迅速かつ便利に処理することができる。

【００１８】

複数の電子機器１１０、１２０、１３０、１４０は、コンピュータシステムによって実現される固定端末や移動端末であってよい。複数の電子機器１１０、１２０、１３０、１４０の例としては、ＡＩスピーカ、スマートフォン、携帯電話、ナビゲーション、ＰＣ（ｐｅｒｓｏｎａｌ　ｃｏｍｐｕｔｅｒ）、ノート型ＰＣ、デジタル放送用端末、ＰＤＡ（Ｐｅｒｓｏｎａｌ　Ｄｉｇｉｔａｌ　Ａｓｓｉｓｔａｎｔ）、ＰＭＰ（Ｐｏｒｔａｂｌｅ　Ｍｕｌｔｉｍｅｄｉａ　Ｐｌａｙｅｒ）、タブレット、ゲームコンソール、ウェアラブルデバイス、ＩｏＴ（ｉｎｔｅｒｎｅｔ　ｏｆ　ｔｈｉｎｇｓ）デバイス、ＶＲ（ｖｉｒｔｕａｌ　ｒｅａｌｉｔｙ）デバイス、ＡＲ（ａｕｇｍｅｎｔｅｄ　ｒｅａｌｉｔｙ）デバイスなどがある。一例として、図１では、電子機器１１０としてＡＩスピーカを示しているが、本発明の実施形態において、電子機器１１０は、実質的に無線または有線通信方式を利用し、ネットワーク１７０を介して他の電子機器１２０、１３０、１４０および／またはサーバ１５０、１６０と通信することのできる多様な物理的なコンピュータシステムのうちの１つを意味してよい。

【００１９】

通信方式が限定されることはなく、ネットワーク１７０が含むことのできる通信網（一例として、移動通信網、有線インターネット、無線インターネット、放送網、衛星網など）を利用する通信方式だけではなく、機器間の近距離無線通信が含まれてもよい。例えば、ネットワーク１７０は、ＰＡＮ（ｐｅｒｓｏｎａｌ　ａｒｅａ　ｎｅｔｗｏｒｋ）、ＬＡＮ（ｌｏｃａｌ　ａｒｅａ　ｎｅｔｗｏｒｋ）、ＣＡＮ（ｃａｍｐｕｓ　ａｒｅａ　ｎｅｔｗｏｒｋ）、ＭＡＮ（ｍｅｔｒｏｐｏｌｉｔａｎ　ａｒｅａ　ｎｅｔｗｏｒｋ）、ＷＡＮ（ｗｉｄｅ　ａｒｅａ　ｎｅｔｗｏｒｋ）、ＢＢＮ（ｂｒｏａｄｂａｎｄ　ｎｅｔｗｏｒｋ）、インターネットなどのネットワークのうちの１つ以上の任意のネットワークを含んでよい。さらに、ネットワーク１７０は、バスネットワーク、スターネットワーク、リングネットワーク、メッシュネットワーク、スターバスネットワーク、ツリーまたは階層的ネットワークなどを含むネットワークトポロジのうちの任意の１つ以上を含んでもよいが、これらに限定されることはない。

【００２０】

サーバ１５０、１６０は、それぞれ、複数の電子機器１１０、１２０、１３０、１４０とネットワーク１７０を介して通信して、命令、コード、ファイル、コンテンツ、サービスなどを提供する、１つ以上のコンピュータ装置によって実現されてよい。例えば、サーバ１５０は、ネットワーク１７０を介して接続した複数の電子機器１１０、１２０、１３０、１４０に第１サービスを提供するシステムであってよく、サーバ１６０も、ネットワーク１７０を介して接続した複数の電子機器１１０、１２０、１３０、１４０に第２サービスを提供するシステムであってよい。より具体的な例として、サーバ１５０は、複数の電子機器１１０、１２０、１３０、１４０においてインストールされて実行されるコンピュータプログラムであるアプリケーションを通じ、該当のアプリケーションが目的とするサービス（一例として、自動応答サービスなど）を第１サービスとして複数の電子機器１１０、１２０、１３０、１４０に提供してよい。他の例として、サーバ１６０は、上述したアプリケーションのインストールおよび実行のためのファイルを複数の電子機器１１０、１２０、１３０、１４０に配布するサービスを第２サービスとして提供してよい。

【００２１】
図２は、本発明の一実施形態における、電子機器およびサーバの内部構成を説明するためのブロック図である。図２では、電子機器に対する例として電子機器１１０の内部構成およびサーバ１５０の内部構成について説明する。また、他の電子機器１２０、１３０、１４０やサーバ１６０も、上述した電子機器１１０またはサーバ１５０と同一または類似の内部構成を有してよい。
【００２２】
電子機器１１０およびサーバ１５０は、メモリ２１１、２２１、プロセッサ２１２、２２２、通信モジュール２１３、２２３、および入力／出力インタフェース２１４、２２４を含んでよい。メモリ２１１、２２１は、非一時的なコンピュータ読み取り可能な記録媒体であってよく、ＲＡＭ（ｒａｎｄｏｍ ａｃｃｅｓｓ ｍｅｍｏｒｙ）、ＲＯＭ（ｒｅａｄ ｏｎｌｙ ｍｅｍｏｒｙ）、ディスクドライブ、ＳＳＤ（ｓｏｌｉｄ ｓｔａｔｅ ｄｒｉｖｅ）、フラッシュメモリ（ｆｌａｓｈ ｍｅｍｏｒｙ）などのような非一時的な大容量記録装置を含んでよい。ここで、ＲＯＭ、ＳＳＤ、フラッシュメモリ、ディスクドライブのような非一時的な大容量記録装置は、メモリ２１１、２２１とは区分される別の非一時的な記録装置として電子機器１１０やサーバ１５０に含まれてもよい。また、メモリ２１１、２２１には、オペレーティングシステムと、少なくとも１つのプログラムコード（一例として、電子機器１１０においてインストールされて実行されるブラウザや、特定のサービスの提供のために電子機器１１０にインストールされたアプリケーションなどのためのコード）が記録されてよい。このようなソフトウェア構成要素は、メモリ２１１、２２１とは別のコンピュータ読み取り可能な記録媒体からロードされてよい。このような別のコンピュータ読み取り可能な記録媒体は、フロッピー（登録商標）ドライブ、ディスク、テープ、ＤＶＤ／ＣＤ－ＲＯＭドライブ、メモリカードなどのコンピュータ読み取り可能な記録媒体を含んでよい。他の実施形態において、ソフトウェア構成要素は、コンピュータ読み取り可能な記録媒体ではない通信モジュール２１３、２２３を通じてメモリ２１１、２２１にロードされてもよい。例えば、少なくとも１つのプログラムは、開発者またはアプリケーションのインストールファイルを配布するファイル配布システム（一例として、上述したサーバ１６０）がネットワーク１７０を介して提供するファイルによってインストールされるコンピュータプログラム（一例として、上述したアプリケーション）に基づいてメモリ２１１、２２１にロードされてよい。
【００２３】
プロセッサ２１２、２２２は、基本的な算術、ロジック、および入出力演算を実行することにより、コンピュータプログラムの命令を処理するように構成されてよい。命令は、メモリ２１１、２２１または通信モジュール２１３、２２３によって、プロセッサ２１２、２２２に提供されてよい。例えば、プロセッサ２１２、２２２は、メモリ２１１、２２１のような記録装置に記録されたプログラムコードにしたがって受信される命令を実行するように構成されてよい。
【００２４】
通信モジュール２１３、２２３は、ネットワーク１７０を介して電子機器１１０とサーバ１５０とが互いに通信するための機能を提供してもよいし、電子機器１１０および／またはサーバ１５０が他の電子機器（一例として、電子機器１２０）または他のサーバ（一例として、サーバ１６０）と通信するための機能を提供してもよい。一例として、電子機器１１０のプロセッサ２１２がメモリ２１１のような記録装置に記録されたプログラムコードにしたがって生成した要求が、通信モジュール２１３の制御にしたがってネットワーク１７０を介してサーバ１５０に伝達されてよい。これとは逆に、サーバ１５０のプロセッサ２２２の制御にしたがって提供される制御信号や命令、コンテンツ、ファイルなどが、通信モジュール２２３とネットワーク１７０を経て電子機器１１０の通信モジュール２１３を通じて電子機器１１０に受信されてよい。例えば、通信モジュール２１３を通じて受信されたサーバ１５０の制御信号や命令、コンテンツ、ファイルなどは、プロセッサ２１２やメモリ２１１に伝達されてよく、コンテンツやファイルなどは、電子機器１１０が

さらに含むことのできる記録媒体（上述した非一時的な記録装置）に記録されてよい。
【００２５】
入力／出力インタフェース２１４は、入力／出力装置２１５とのインタフェースのための手段であってよい。例えば、入力装置は、キーボード、マウス、マイクロフォン、カメラなどの装置を、出力装置は、ディスプレイ、スピーカ、触覚フィードバックデバイスなどのような装置を含んでよい。他の例として、入力／出力インタフェース２１４は、タッチスクリーンのように入力と出力のための機能が１つに統合された装置とのインタフェースのための手段であってもよい。入力／出力装置２１５は、電子機器１１０と１つの装置で構成されてもよい。また、サーバ１５０の入力／出力インタフェース２２４は、サーバ１５０に接続するかサーバ１５０が含むことのできる入力または出力のための装置（図示せず）とのインタフェースのための手段であってよい。より具体的な例として、電子機器１１０のプロセッサ２１２がメモリ２１１にロードされたコンピュータプログラムの命令を処理するにあたり、サーバ１５０や電子機器１２０が提供するデータを利用して構成されるサービス画面やコンテンツが、入力／出力インタフェース２１４を通じてディスプレイに表示されてよい。
【００２６】
また、他の実施形態において、電子機器１１０およびサーバ１５０は、図２の構成要素よりも多くの構成要素を含んでもよい。しかし、大部分の従来技術的構成要素を明確に図に示す必要はない。例えば、電子機器１１０は、上述した入力／出力装置２１５のうちの少なくとも一部を含むように実現されてもよいし、トランシーバ、カメラ、各種センサ、データベースなどのような他の構成要素をさらに含んでもよい。より具体的な例として、電子機器１１０がＡＩスピーカである場合、一般的にＡＩスピーカが含んでいる各種センサ、カメラモジュール、物理的な各種ボタン、タッチパネルを利用したボタン、入力／出力ポート、振動のための振動器などのような多様な構成要素が、電子機器１１０にさらに含まれるように実現されてよい。（以上、引用おわり）
【００２７】
機械について開示する。少なくとも１つの実施形態によれば、ユーザー端末は、制御部、ＲＡＭ、ストレージ部、グラフィックス処理部、通信インタフェース、インタフェース部からなり、それぞれ内部バスにより接続されている。
【００２８】
少なくとも１つの実施形態によれば、制御部は、ＣＰＵやＲＯＭから構成される。制御部は、ストレージ部に格納されたプログラムを実行し、ユーザ端末の制御を行なう。ＲＡＭは、制御部のワークエリアである。ストレージ部は、プログラムやデータを保存するための記憶領域である。制御部は、プログラム及びデータをＲＡＭから読み出して処理を行なう。制御部は、ＲＡＭにロードされたプログラム及びデータを処理することで、描画命令をグラフィックス処理部に出力する。
【００２９】
少なくとも１つの実施形態によれば、グラフィックス処理部は表示部に接続されている。表示部は表示画面を有している。制御部が描画命令をグラフィックス処理部に出力すると、グラフィックス処理部は、表示画面上に画像を表示するためのビデオ信号を出力する。ここで、表示部はタッチセンサを備えるタッチパネルであってもよい。この表示部のタッチパネルが入力部として機能する。
【００３０】
少なくとも１つの実施形態によれば、通信インタフェースは無線又は有線により通信ネットワークに接続が可能であり、通信ネットワークを介して、サーバ装置とデータを送受信することが可能である。通信インタフェースを介して受信したデータは、ＲＡＭにロードされ、制御部により演算処理が行われる。インタフェース部には外部メモリ（例えば、ＳＤカード等）が接続されている。
【００３１】
少なくとも１つの実施形態によれば、ユーザー端末は、表示画面と入力部を有するコン

ピュータ装置であれば特に限定されない。ユーザー端末としては、例えば、従来型の携帯電話、タブレット型端末、スマートフォン、デスクトップ型・ノート型のパーソナルコンピュータなどが挙げられる。VRゴーグル、すなわち頭にストラップで固定または取り付けられたフレーム（またはヘッドセット）に取り付けられた画面（または2つのディスプレイパネル、各目に対し1つずつ）で構成されてもよい。ユーザー端末は、音声の出力部を有する。

【００３２】

少なくとも１つの実施形態によれば、ユーザー端末は、通信ネットワークを介してサーバ装置と通信接続が可能である。通信ネットワークを介して通信接続をして、情報を送信し、若しくは情報を受信することができる。

【００３３】

少なくとも１つの実施形態によれば、サーバ装置は、制御部、ＲＡＭ、ストレージ部及び通信インタフェースを少なくとも備え、それぞれ内部バスにより接続されている。

【００３４】

少なくとも１つの実施形態によれば、制御部は、ＣＰＵやＲＯＭから構成され、ストレージ部に格納されたプログラムを実行し、サーバ装置の制御を行う。また、制御部は時間を計時する内部タイマを備えている。ＲＡＭは、制御部のワークエリアである。ストレージ部は、プログラムやデータを保存するための記憶領域である。制御部は、プログラム及びデータをＲＡＭから読み出し、ユーザー端末から受信した情報等をもとに、プログラム実行処理を行う。

【００３５】

AIについて開示する。少なくとも１つの実施形態によれば、人工知能は、機械学習、深層学習、生成AI、大規模言語モデル、LLM、基盤モデル、生成AIを含む。生成AIはトランスフォーマーを使い、アテンションという機構を多数用いる。自己教師あり学習、Extract Predictionを用いる。この場合、AIは次の単語を当てることができる。文を与えられると、途中までの文章から次の単語を当てる。教師あり学習の問題を大量に作り出す。これらにより、次の単語を当てられるAIができる。生成AIは文法構造、トピックのつながり、こういう文体の人はこういう文を書きそうという予測をすることができる。さらに、生成AIは、次の文を当てるということだけで、その背後にある構造、因果関係、知識を学習できる。生成AIはスケール速があり、パラメーターの数が大きいほど精度が上がる。普通の統計、機械学習は、データのサンプルサイズに比べてモデルのパラメーターを大きくしすぎるとオーバーフィットする。LLMは、パラメーターの数を大きくすればするほど精度が上がる。ある生成AIは1750億パラメータを持っている、生成AIは、対話をスムーズにやるように教師あり学習を被せる。変なことを言わないように教師づけられている。感想文を書いたり、コールセンターのオペレーターを演じる。

【００３６】

少なくとも１つの実施形態によれば、大規模言語モデル（Large Language Models/LLM）とは、非包括的に、大量のデータセットとディープラーニング技術を用いて構築された、機械学習の自然言語処理モデルのことである。一般的には、特定のタスクでトレーニングする「ファインチューニング」と呼ばれる手法を用いて、テキスト分類・生成や感情分析、文章要約、質問応答といったさまざまな自然言語処理（NLP）タスクに適応させる。少なくとも１つの実施形態によれば、自己教師あり学習は、人の本質的な知能に近い。人は、行動をする時に常に次に起こる事象を予測しており、次の入力を予測している。その過程で外界の構造を学んでいける。次の単語を予測するのは、本質的な知能であり、大脳皮質でやることに近いと考える。少なくとも１つの実施形態によれば、大規模言語モデルは、入力される情報をまる覚えするが、次の単語を予測するのに必要な程度で汎化する。最初から全ての情報を汎化しにいかない。大規模言語モデルは、情報を覚えておくために容量が必要である。また、そのためにパラメータが必要である。少なくとも１つの実施形態によれば、大規模言語モデルは、1750億パラメータや、2200 億パラメーターのモデルを８つ搭載する。

【００３７】
少なくとも１つの実施形態によれば、ビデオや画像をLLMのテキストトークンに似た小さなデータ単位であるビジュアルパッチの集合として表現する。パッチは、視覚データのモデルを効果的に表現し、さまざまな種類のビデオや画像で生成モデルをトレーニングするための非常にスケーラブルで効果的な表現として用いる。まず動画を低次元の潜在空間に圧縮し、次に表現を時空間パッチに分解することで、動画をパッチに変換する。
【００３８】
少なくとも１つの実施形態によれば、Video compression network（ビデオ圧縮ネットワーク）は、視覚データの次元を削減するネットワークで、生の動画を入力として受け取り、時間的および空間的に圧縮された潜在表現を出力する。AIは、この圧縮された潜在空間でトレーニングされ、その後、この圧縮された潜在空間内で動画を生成する。
【００３９】
少なくとも１つの実施形態によれば、Spacetime Latent Patches（時空潜在パッチ）は、圧縮された入力動画が与えられると、トランスフォーマートークンとして機能する一連の時空パッチを抽出する。パッチベースの表現により、Sora はさまざまな解像度、長さ、アスペクト比のビデオや画像でトレーニングでき、推論時にランダムに初期化されたパッチを適切なサイズのグリッドに配置することで、生成されるビデオのサイズを制御する。
【００４０】
少なくとも１つの実施形態によれば、AIは、ディフュージョンモデルであり、ノイズの多いパッチ (およびテキスト プロンプトなどの条件付け情報) が入力されると、元の「きれいな」パッチを予測するようにトレーニングされる。AIはディフュージョントランスフォーマーであり、これは言語モデリング、コンピュータビジョン、画像生成など、様々な領域で顕著なスケーリング特性を示す。ディフュージョントランスフォーマーは動画生成モデルとしても効果的である。AIは、トレーニングの計算量が増えるにつれて、サンプルの品質は著しく向上する。
【００４１】
少なくとも１つの実施形態によれば、AIは、キャプション再生成技術を応用し、非常に説明的なキャプションモデルをトレーニングし、次にそれを使用してトレーニングセット内のすべての動画のテキストキャプションを生成する。高度に説明的なキャプションの訓練は、生成動画の全体的な品質だけでなく、テキストの忠実度を向上させる。GPTを活用して短いユーザープロンプトを長い詳細なキャプションに変換し、モデルに送信する。これにより、AIはユーザーのプロンプトに正確に従った高品質の動画を生成することができる。
【００４２】
少なくとも１の実施形態によれば、AIは、自然言語処理において、以下の流れに沿ってベクトル化が実施できる。まずはじめは、前処理として与えられた文章のクリーニング処理を実施する。クリーニング処理では、テキスト内に含まれるJavaScriptのコードやHTMLタグなどの不要な単語を削除する。これらのコードは、インターネット上に表示させるために利用されているコードであるため、自然言語処理では一般には利用されない情報である。続いて、形態素解析によって文章を単語レベルに分割していく。形態素解析とは、文字で表記された自然言語の文において、意味を持つ最小の言語単位に分類することである。形態素解析ツールとしては、「MeCab」「JUMAN」「JANOME」を使うことができる。正規化では、表記ゆれのような同じ意味の言葉を一つの単語に統一する。ストップワードは、自然言語処理で活用できないなどの理由で処理対象外とする単語のことである。ストップワードの例としては、単語の中でも助詞や助動詞といった単体で意味を持たないものがあげられる。ベクトルの計算に際し、これらを除去し、意味のある単語のみを対象とすることもある。これらのストップワードの除去は行わずにベクトル化を行うこともある。ベクトル化は、文字列である単語をベクトルに変換する処理である。ベクトル化により、単語データから数値データに変換する。単語をベクトルに変換するときには、Bag of

Wordsや分散表現と呼ばれる方法で実施していく。Bag of Wordsとは、与えられた文章の中に出現する単語の出現数を用いて、文章をベクトル化する方法である。文章の中にどれだけ単語が出現したのかに着目するため、単語や文章の並びは考慮していない。分散表現とは、単語が持つ意味に着目してベクトル化する方法である。単語の意味をベクトル化することで、同じような意味や使われ方をする単語に近しいベクトルを与えることができる上に、単語同士の関係性もまたベクトルで表現できる。ベクトルでの表現により、単語の意味同士の加算や減算が可能である。応用処理は、数値データに変換した自然言語を機械学習の入力に活用できる。具体的には、ベクトル化した自然言語を分類器に投入し、文章の分類を実施する。ここで活用されるツールとしては「TensorFlow」、「scikit-learn」、「PyTorch」などがあげられる。

【００４３】

音について開示する。少なくとも１の実施形態において、音は、音楽又は効果音を含む。音楽は、音の長さ、高さ、強さ、音色などを組み合わせて、様々な感情や物語を表現するものであって、歌や楽器の演奏、自然の音を含む。音楽の作曲は、スケール（調）、コード（和音）又はメロディ（旋律）の決定によって、行うことができる。少なくとも１の実施形態において、スケールが決まればその構成音に合わせてコードも決まる。コード進行はセオリーがあり、これまでの曲の中で良い進行とされるパターンがある。コードは曲のムードを大きく左右し、人間のその時の感情、感じ方に非常に影響を与える場合がある。メロディも同様にコードの構成音を元にすれば大きく外れることはないが、単調になってしまう傾向があるので、多少のランダムさが必要な場合がある。コードに対して大きく外さない範囲で変動させれば単調になるのを防ぐことができる場合がある。

【００４４】

音を作成する方法であって、装置が、動画内のオブジェクトの動作の特徴である動作特徴を特定する方法について開示する（本開示において、「方法」という記載は、明示的に矛盾する記載がない限り「ステップ」という記載に置き換えて解釈することもできる）。少なくとも１の実施形態において、動画とは、1枚1枚の静止画を連続的に表示することで、動きを作り出したもの又はデータである。動画は、音声を含み、映像と音声が同期することで、より豊かな表現をする場合がある。オブジェクトとは、動画の中で操作されるデータ又はそれに関連する処理である。オブジェクトの例として、動画中における動体（生物及び非生物を含む。前者の非包括的な例として、人、ダンスをしている人、動物、植物が風に揺れる様子、など。後者の非包括的な例として、移動体（車など）、波の満ち引き、光の点滅、光の波長又は強弱の変化、コンピューターグラフィックで描かれたデジタルな表現）である。本開示において、人が動画を見た際に、動いていると認知可能な対象は、この「オブジェクト」又は「オブジェクト」に対応するデータであるとみなす。本開示において、「動作特徴」とは、「動作特徴量」である場合がある（本開示全体おいて、明示的に矛盾する記載がない限り同様である）。特徴量は、あるデータセットの中に含まれる、そのデータの特徴を表す数値である。例えば、人の顔を動画で表す場合、特徴量としては「目の大きさ」「鼻の高さ」「肌の色」などが挙げられる。他にも、人がダンスをしている場合において、器官（手、足、腰などを含む）が動く様子も、特徴量とすることができる。これらの数値をコンピュータに与えることで、コンピュータは人々の顔を区別したり、特定の人物を認識したり、ダンスの振り付けを認識できる。他にも、波の満ち引きにおいて、波が満ちる様子と、波が低く様子も、当然に特徴量とすることができる。他にも、コンピューターグラフィックの場合、ある線や面が移動する様子、色が変化する様子も、当然に特徴量になる。特徴量の種類について、非包括的に、以下が挙げられる。数値データは、身長、体重、年齢など、数値で表せるデータ。カテゴリカルデータは、性別、国籍、職業など、カテゴリーで表せるデータ。テキストデータは、文章、単語、キーワードなど、テキストで表せるデータ。画像データ（動画における静止画を含む）は色、形、テクスチャなど、画像から抽出される特徴。動画データは、は色、形、テクスチャ、再生時間におけるこれらの存在又は変化など、動画から抽出される特徴。音声データは、音の高さ、周波数、音色など、音声から抽出される特徴。

【００４５】
少なくとも１の実施形態において、装置は、先述したいずれかの方法、機械又はAIで作成できる。本開示において、「装置」の記載は、明示的に矛盾する記載がない限り、「AI」に置き換えることができる。装置は、対象となる動画のデータを解析する。装置は、動画の中のオブジェクトを特定する。対象となるオブジェクトについて、ユーザーが特定することができる。他の実施形態として、装置は、動画に存在するオブジェクトの中から、出現頻度が相対的に多いオブジェクトを特定する。装置は、そのように特定することを、AIに要求することもできる（本開示全体において、明示的に矛盾する記載がない限り、「装置がある処理をする」という記載は、「装置が、AIにその処理をするよう要求する」と置き換えることができる。この場合において、AIは、先述した通りの学習方法（教師あり学習を含む）によって、その処理や判断を学習したAIを用いることができる）。このような場合、対象となるオブジェクトは、ユーザーが指定するのではなく、装置が自動的に特定する。装置は、オブジェクトの動作特徴を特定する。一例として、ダンサーがダンスをする動画について、ダンサーが手を振る様子が、オブジェクトの動作特徴である。この実施形態において、ダンサーの首を振る様子、腰を振る様子、足の動き、指の動きを動作特徴とすることも、当然にできる。すでに説明したように、動作特徴は、ユーザーが特定することができる。さらに、装置は、オブジェクトの動作の中から、出現頻度が相対的に多い動作を動作特徴として特定することもできる。

【００４６】
類似する動作特徴間の時間を単位時間として、単位時間を拍（拍子）とする音楽を生成することをAIに要求する方法について開示する。本開示において、明示的に矛盾する記載がない限り、「音楽を生成する」という記載は「音を生成する」に置き換えることができる。少なくとも１の実施形態において、動画には、類似する動作特徴が存在する。非包括的な例として、ダンスに関して、ダンサーが同じダンスの振り付けを繰り返し動作する場合がある。この場合、装置は、それぞれの動作の動作特徴同士を比較し、これらを類似する動作特徴と認識する。類似と判断する基準について、ユーザーが定める方法と、機械が自動で定める方法がある。前者においては、装置は、二以上の動作特徴の動画あるいは静止画を、ユーザーの操作する端末の画面に表示する。ユーザーは、その中から、二以上の動作特徴の動画あるいは静止画について、類似する動作特徴とするか、類似と認定しないかについて、装置に指定する。装置は、その指定をもとに、動作特徴の類否を判断する。後者においては、装置は、二以上の動作特徴の動画あるいは静止画について、明らかに異なる動作特徴を含んでいるかどうかを判定する。明らかな特徴又は特徴量の差異を検出できない場合、これらの動画を類似すると判定することができる。当該判定においては、AIが、これらの類否についての教師あり学習を通して学習することができる。装置は、当該AIの機能を活用し、又は当該AIに類否を判定することを要求し、その判定結果を得ることもできる。

【００４７】
装置は、動作特徴間の時間を特定する。非包括的な例として、ダンスに関して、ダンサーが同じダンスの振り付け（類似する動作特徴）を繰り返した場合、動画におけるそれぞれの動作特徴の出現時刻（又は出願時間）の差（又は距離）を特定する。非包括的な例として、あるダンスのパフォーマンスの動画が３分の動画であったとする。そして、ダンサーが動作特徴を1分10秒の時点と1分15秒の時点で繰り返した場合、動作特徴間の時間は5秒であって、単位時間は5秒である。装置は、動作特徴間の時間を単位時間と認識する。装置は、単位時間の情報を、他の装置やAIに送信することができる。他の実施形態において、装置は、先述した通り、教師あり学習によって単位時間の認識を学習したAIを用い、又はそのAIに単位時間を認識するよう要求することもできる。

【００４８】
少なくとも１の実施形態において、音における拍は、リズムの単位を含む。非包括的な例として、拍は、一定の時間間隔（音楽の中で、一定の周期で繰り返される基本的な単位）、強弱の対比（強拍と弱拍があり、この対比によってリズムを生成するもの）、又はテ

ンポの基礎（拍の速さがテンポとなり、曲の速さを決めるもの）を含む。拍は、拍子である場合がある。拍子は、曲全体のリズムの基礎となるもので、小節の中にいくつの拍があるかを示す。例えば、4分の4拍子なら、1小節に4つの拍があり、それぞれの拍の長さが同じである。本開示において、「拍」の記載は、明示的に矛盾する記載がない限り、「拍子」に置き換えることができる。装置は、単位時間を拍（拍子）とする音楽を生成することをAIに要求する。非包括的な例として、あるダンスのパフォーマンスの動画が３分の動画であって、かつ単位時間が5秒の場合、装置は、5秒間を1拍とする音楽を生成することをAIに要求する。他の実施形態において、装置は、先述した通り、AIを用いて自らが当該音楽を生成してもよい。

【００４９】

上記の構成によれば、動画に基づき音を自動的に作成することができる。作成された音又は音楽は、動画の特徴量に基づいた拍を有しているので、動画と生成された音との一致度が高い。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音との一致度が高い利用可能性の高い音を自動で生成できる点で、産業上の利用可能性がある。

【００５０】

単位時間を拍とする音楽を生成することをAIに要求する際、動画の時間（動作特徴を特定すべき時間を含む。本開示において、明示的に矛盾する記載がない限り、「動画の時間」とは、「動画の長さ」、「動画の時間の長さ」に置き換えて理解することができる）と同一の長さの音楽を生成することを要求する方法について、開示する。装置は、動画の時間の長さと同一の長さの音楽を生成することを、AIに要求する。非包括的な例として、あるダンスのパフォーマンスの動画が３分の動画であった際は、動画の時間は３分であるから、装置は、３分の音楽を生成することを要求する。他の実施形態において、装置は、動作特徴を特定すべき時間を、動画の長さとする。動作特徴を特定すべき時間とは、少なくとも１の実施形態において、装置又はAIが、動作特徴の特定を許されている、動画内の全部または一部の時間帯のことである。動作特徴を特定すべき時間について、ユーザーが定める方法と、装置が自動的に定める方法がある。前者において、ユーザーは、動画の時間内の限りで、動作特徴を特定すべき時間を定めることができる。非包括的な例として、あるダンスのパフォーマンスの動画が３分の動画であった際、0分0秒から1分30秒の時点までの間を、動作特徴を特定すべき時間と定めることができる。装置は、先述の方法によって、定められた時間の中のみにおいて、動作特徴を特定する。後者の例として、装置が、動画の時間の所定の一部を、動作特徴を特定すべき時間と定めることもできる。当然に、動画の時間全体を、動作特徴を特定すべき時間と定めることもできる。装置は、動作特徴を特定すべき時間と同一の長さの音楽を生成することを要求する。

【００５１】

装置が、動画内の単位時間と、音楽の拍が同一になるよう音楽を生成することを要求する方法について開示する。上記の「同一になる」の記載は、「同期する」に置き換えることができる。少なくとも１の実施形態において、装置は、動画の時間と同一の時間の音楽を生成し、かつ、動画内の単位時間と、音楽の単位時間が同一になるよう音楽を生成することを要求する。非包括的な例として、あるダンスのパフォーマンスの動画が３分の動画であったとする。そして、ダンサーが動作特徴を1分12秒の時点と1分17秒の時点で繰り返した場合、単位時間は5秒である。さらに、装置は、5秒間を1拍とする音楽を生成すること、及び、3分の長さの音楽を生成することをAIに要求する。そして、装置は、動画内の単位時間と、音楽の単位時間が同一になるよう音楽を生成することを要求するから、生成された音は、動画における1分12秒の時点と1分17秒の時点の拍と同一になる（同期する）ように拍を生成する。この場合において、生成された音楽は、0分2秒の時点から5秒の単位時間で拍を有すれば、動画における1分12秒の時点と1分17秒の時点の拍と同一（同期）する。言い換えると、装置は、動画内の単位時間に呼応する音楽の拍を特定拍として認識し、特定拍に基づいて（又は、特定拍に矛盾しないように）、全体の拍を定めた音楽を生成することを要求する。先の例をとれば、1分12秒の時点と1分17秒の時点の拍が

特定拍に該当し、特定拍に基づいて、全体の拍を定めた音楽が生成されている。
【００５２】
上記の構成によれば、動画の拍と同期する音を自動的に作成することができる。作成された音又は音楽は、動画の特徴量に基づいた拍を有するだけでなく、動画の拍と少なくとも同期しているので、動画と生成された音との一致度がより高い。この場合、生成された音は、動画と、生成された音を同時に再生すれば、そのまま産業利用可能である。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音との一致度が高い利用可能性の高い音を自動で生成できる点で、産業上の利用可能性がある。
【００５３】
単位時間を拍とする音楽を生成することをAIに要求する際、類似する動作特徴から生ずる２以上の単位時間の中から、相対的に短いの単位時間を拍とするよう要求する、方法について開示する。少なくとも１の実施形態において、装置は、動画の中から、類似する動作特徴を３以上認識することがありうる。この場合、装置が２以上の単位時間を認識しうる。装置は、類似する動作特徴から生ずる２以上の単位時間の中から、最短の単位時間を、拍とすべき単位時間として認識する。装置は、このような拍とすべき単位時間を、他の装置やAIに送信し、又は、このような拍とすべき単位時間を認識又は特定するようAIに要求することができる。すでに説明した通り、AIは、類似する動作特徴から生ずる２以上の単位時間を有する動画を用いた教師あり学習によって、最短の単位時間を拍とすべき学習をする。非包括的な例として、あるダンスのパフォーマンスの動画が３分の動画であったとする。例えば、Aメロ、Bメロ、サビは、主にポップスやロックなどの歌謡曲でよく使われ、それぞれ頃なるメロディーで構成される。ミュージックビデオにおいては、Aメロ、サビ、Bメロ、サビという構成の組み合わせになり、サビの振り付けは類似していることがある。そして、ダンサーが動作特徴を1分12秒、1分17秒、2分12秒、2分17秒の時点で繰り返したとする。装置は、単位時間として5秒、55秒、1分、1分5秒の４種類を認識することになる。このような場合、装置は、最短の単位時間として5秒を単位時間として認識する。装置は、その最短の単位時間を拍と認識し、又は最短の単位時間を拍と認識するよう装置やAIへ要求する。
【００５４】
上記の構成によれば、類似する動作特徴が動画内において散在するような、複雑な動画であっても、動画の拍と同期する音を自動的に作成することができる。作成された音又は音楽は、動画の特徴量に基づいた拍を有するだけでなく、動画の拍と少なくとも同期しているので、動画と生成された音との一致度がより高い。この場合、生成された音は、動画と、生成された音を同時に再生すれば、そのまま産業利用可能である。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音との一致度が高い利用可能性の高い音を自動で生成できる点で、産業上の利用可能性がある。
【００５５】
単位時間を拍とする音楽を生成することをAIに要求する際、類似する動作特徴から生ずる２以上の単位時間の中から、特定された単位時間の出現頻度が相対的に高い単位時間を拍とするよう要求する、方法について開示する。少なくとも１の実施形態において、装置は、認識した単位時間について、その出現頻度を算出する。出現頻度とは、ある事象や物が、特定の期間や範囲内でどれくらいの回数現れるかを示す数値を含む。一例として、出現頻度は、事象（特定の言葉、数字、行動、現象など）、範囲（特定の文章、データセット、時間、空間など）、文章中での単語の出現頻度（ある文章の中で、「猫」という言葉が何回出てくるか、など）、を含む。装置は、動画の時間又は動作特徴を特定すべき時間の範囲内で、単位時間の出現頻度を計算する。非包括的な例として、あるダンスのパフォーマンスの動画が３分の動画であり、ダンサーの振り付けを動作特徴と認識したとする。そして、装置が、5秒の単位時間が50回、10秒の単位時間が10回、20秒の単位時間が3回、1分の単位時間が1回あると認識したとする。この場合について、装置は、出現頻度が最も高い5秒を、拍とすべき単位時間として認識する。装置は、この単位時間を拍とする音楽を自ら生成し、又はそのように生成することをAIに要求する。

【００５６】
上記の構成によれば、すでに説明した通り、類似する動作特徴が動画内において散在するような、複雑な動画であっても、動画の拍と同期する音を自動的に作成することができる。さらに、最も出現頻度の高い単位時間を拍としているので、生成された音楽の拍は、動画の拍と不一致になる可能性が少ない。そのため、作成された音又は音楽は、動画の特徴量に基づいた拍を有するだけでなく、動画の拍と少なくとも同期しているので、動画と生成された音との一致度がより高い。この場合、生成された音は、動画と、生成された音を同時に再生すれば、そのまま産業利用可能である。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音との一致度が高い利用可能性の高い音を自動で生成できる点で、産業上の利用可能性がある。

【００５７】
単位時間を拍とする音楽を生成することをAIに要求する際、動画の時間又は動作特徴を特定すべき時間と同一の時間の音楽を生成し、類似しない動作特徴による２以上の単位時間が特定された際には、２以上の音楽を生成し、当該音楽を連結するよう要求する、方法について開示する。少なくとも１の実施形態において、装置は、すでに説明した通り、類似する動作特徴に基づいて、単位時間を認識する。この場合、装置は、Aという動作特徴に基づく単位時間と、Bという動作特徴の基づく単位時間とを認識することになる。これらは、類似しない動作特徴による２以上の単位時間である。装置は、類似しない動作特徴による２以上の単位時間が特定された際には、２以上の音楽を生成する。他の実施形態では、そのように２以上の音楽を生成することをAIや他の装置に要求する。装置は、この生成された音楽を、連結する。連結とは、複数のメロディーが、繋がっている状態を含む。連結は、滑らかで自然な流れでつながる場合がある。非包括的な例として、装置は、動画（ミュージックビデオや映画を含む）からAという動作特徴に基づく単位時間で生成されたAという音楽と、Bという動作特徴の基づく単位時間で生成されたBという音楽を生成する。装置は、Aという音楽と、Bという音楽を、連結する。連結の方法は、ある音楽の終了と同時に別の音楽を再生できるように、音楽のデータを連結する方法がある。他にも、ある音楽を途中で終了し別の音楽を再生するように、音楽のデータの部分と、他の音楽のデータの部分とを連結する方法もある。少なくとも１の実施形態において、２以上の音楽を生成する場合、１の音楽を変調した音楽を、２の音楽とすることができる。

【００５８】
上記の構成によれば、類似しない動作特徴による２以上の単位時間が特定されるような、複雑な動画であっても、動画の拍と同期する音を自動的に作成することができる。さらに、２以上の単位時間に基づいた音であるから、動画の拍と不一致になる可能性が、さらに少なくなる。そのため、作成された音又は音楽は、動画の拍と大部分が同期しているので、動画と生成された音との一致度がより高い。この場合、生成された音は、動画と、生成された音を同時に再生すれば、そのまま産業利用可能である。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音との一致度が高い利用可能性の高い音を自動で生成できる点で、産業上の利用可能性がある。

【００５９】
前記方法において、２以上の動作特徴の出現頻度に基づき特定された時点で、当該音楽を連結するよう要求する、方法について開示する。少なくとも１の実施形態において、装置は、動画の時間又は動作特徴を特定すべき時間と同一の時間の音楽を生成し、２以上の単位時間について単位時間が変わる時点を特定する。非包括的な例として、装置は、各動作特徴に基づく単位時間について、動画の時間帯における出現頻度又は出現確率を算出する。例えば、装置は、動画の時間又は動作特徴を特定すべき時間の範囲内で、Aという動作特徴に基づく単位時間の出現が相対的に多い時間帯と、Bという動作特徴の基づく単位時間の出現が相対的に多い時間帯とを特定する。装置は、これらの出現頻度に基づき、出現頻度が最も高い単位時間に基づく音楽を生成する。装置は、出現頻度が最も高い単位時間が変わった場合には、その新たな単位時間に基づく音楽を生成する。装置は、これらの２つの音楽を連結する。別の実施形態において、装置は、動画の時間（動作特徴を特定す

べき時間）と同一の時間の音楽を生成し、類似しない２以上の動作特徴が特定された際には、２以上の音楽を生成する。装置は、類似しない２以上の動作特徴の出現頻度に基づき特定された時点（本段落において、明示的に矛盾する記載がない限り「時点」や「時刻」の記載は、「時刻帯」に置き換えることができる）を、変更時刻として特定する。例えば、装置は、動画の時間又は動作特徴を特定すべき時間の範囲内で、Aという動作特徴に基づく単位時間の出現が相対的に多い時間帯と、Bという動作特徴の基づく単位時間の出現が相対的に多い時間帯とを特定する。装置は、出現頻度の最も多い動作特徴が変わる時点、又は異なる動作特徴に基づく出現頻度の高い時間帯の間又は略中間時点を、変更時刻として特定できる。非包括的な例として、あるダンスのパフォーマンスの動画が３分であったとする。0分0秒から1分30秒の時刻帯においては、Aという動作特徴に基づく単位時間の出現が相対的に多く、1分30秒から3分の時刻帯においては、Bという動作特徴に基づく単位時間の出現が相対的に多かった場合、装置は、２以上の動作特徴の出現頻度に基づき、その略中間時点である1分30秒を、特定する。装置は、特定された時点で、当該音楽を連結するよう要求する。すなわち、生成された音楽は、0分0秒から1分30秒までが、Aという動作特徴に基づくAという音楽、1分30秒から3分までが、Bという動作特徴に基づくBという音楽である。これらは、1分30秒の時点で連結される。

【００６０】

上記の構成によれば、すでに説明した通り、類似しない動作特徴による２以上の単位時間が特定されるような、複雑な動画であっても、動画の拍と同期する音を自動的に作成することができる。さらに、２以上の単位時間に基づいた音であるから、動画の拍と不一致になる可能性が、さらに少なくなる。そのため、作成された音又は音楽は、動画の拍と大部分が同期しているので、動画と生成された音との一致度がより高い。この場合、生成された音は、動画と、生成された音を同時に再生すれば、そのまま産業利用可能である。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音との一致度が高い利用可能性の高い音を自動で生成できる点で、産業上の利用可能性がある。

【００６１】

少なくとも１の実施形態において、動画は、映画、ダンス、振り付けを含む。一例として、Youtube（登録商標）で公開される動画を含む。この場合、公開される動画に合わせた音楽を自動生成することができる。動画は、ライブ配信される動画を含む。この場合、装置は、ライブ配信と同時に、本開示のいずれかに記載の条件又は構成に従って、音楽を生成する。装置は、この生成された音楽を、ライブ配信の動画と同時に配信することができる。上記の構成によれば、ライブ配信と同時に、動画の拍と一致する音楽を提供できる。このような柔軟性の高い生成方法は、人間の作曲家にはできないことであるから、従来の技術に比べると優れた効果があって、産業上の利用可能性がある。少なくとも１の実施形態において、装置は、配信中の環境音や観客の声をリアルタイムで解析し、その１以上を特定音、特徴、特徴量と認識し、それらに応じた音や効果音を生成することを、AIに要求することができる。

【００６２】

少なくとも１の実施形態において、動画は、VR又はARで用いられる動画を含む。VR（仮想現実）は、完全に仮想の世界に没入する体験を提供する技術である。VRゴーグルなどのデバイスを装着することで、ユーザーはあたかも別の世界にいるかのような感覚を味わえる。AR（拡張現実）は、現実世界にデジタル情報を重ね合わせる技術である。スマートフォンのカメラを通して、現実の風景に仮想のオブジェクトを表示したり、情報を付加したりすることができる。装置は、VR又はARを通じてユーザーに提示する動画について、本開示のいずれかに記載の条件又は構成に従って、音楽を生成することができる。装置は、VR又はARを通じてユーザーに提示するのと同時刻に合わせて、この生成された音楽を音声出力することができる。音声出力は、ユーザー端末から音声出力する実施形態を含む。上記の構成によれば、VR又はARの動画の拍と一致する音楽を提供できる。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音との一致度が高い利用可能性の高い音を自動で生成できる点で、産業上の利用可能性がある。

【００６３】
少なくとも１の実施形態において、AIは、教師なし学習で、次の音を予測する学習を繰り返す。すでに説明した通り、作曲の流れは、その構成によって階層化されている。音楽を構成する要素は、少なくともスケール（調）とコード（和音）とメロディ（旋律）とされている。スケールが決まればその構成音に合わせてコードも決まる。コード進行はセオリーがあり、これまでの曲の中で良い進行とされるパターンも数多くあり蓄積されている。コードは曲のムードを大きく左右し、人間のその時の感情、感じ方に非常に影響を与えると言われている。AIは、音楽のデータによる教師なし学習で、次の音を予測する学習を繰り返す。さらに、AIは、音楽のデータについて、その音楽がどのような感情を示すかについて教師あり学習を行う。
【００６４】
少なくとも１の実施形態において、装置又はAIは、著作者が登録されているか否かにかかわらず、すでに完成したものとしてインターネット上にアップロード又はデータベースに記録されている音楽又はそれに関連する情報をもとに、音楽を生成する。非包括的な例として、Aという動作特徴に基づく単位時間が5秒であった際に、装置は、すでに完成したものとしてインターネット上にアップロードされている5秒を拍とする音楽を特定する。複数の音楽が特定された場合は、本開示に記載されているいずれかの条件により合致する音楽を特定する。装置又はAIは、もしくは特定された音楽を改変し、音楽を生成する。
【００６５】
少なくとも１の実施形態において、装置又はAIは、著作者が登録されているか否かにかかわらず、すでに完成したものとしてインターネット上にアップロード又はデータベースに記録されている音楽又はそれに関連する情報をもとに、１の音楽を特定する。本実施形態においては、装置又はAIは、音楽を生成するのではなく、条件に合致する既成の音楽を特定する構成を持つ。装置は、特定された音楽又は音楽に関する情報を、ユーザーに提示する。
【００６６】
少なくとも１の実施形態において、装置又はAIは、著作者が登録されているか否かにかかわらず、すでに完成したものとしてインターネット上にアップロード又はデータベースに記録されている音楽又はそれに関連する情報を利用して、音楽を生成する。
【００６７】
少なくとも１の実施形態において、装置は、AIによって生成した音楽のデータに、AIによって生成されたことを示すデータを付加する。ユーザーに動画を生成した音楽と共に提供する際に、装置は、音楽が当該データがあることを認識した場合には、ユーザーに、音楽がAIによって生成されたことを提示し、又はユーザー端末にその旨を表示する。別の実施形態において、装置は、すでに完成したものとしてインターネット上にアップロード又はデータベースに記録されている音楽を利用して音楽を生成した際に、生成した音楽のデータに対して、利用元の音楽の名称又は著作権者に関する情報を付加する。装置は、音楽に当該情報が含まれていることを認識した場合には、ユーザーに、音楽の名称、その著作権者が創作したこと、又はその音楽に依拠してAIが生成したことを提示し、又はユーザー端末にその旨を表示する。少なくとも１の実施形態において、装置は、配信プラットフォームに適合したフォーマットの著作権表示を行うデータ又はプロンプトを生成する。一例として、Youtube(登録商標)において、そのプラットフォームの定める著作権表示をするデータ又はプロンプトを生成し、もしくはそれらのデータを生成した動画に付随して生成する。これらを、著作権管理ツールと呼ぶ場合がある。上記の構成によれば、AIによって付加又は生成した音楽のデータにおいて、著作権侵害が発生するリスクを減らすことができる点で、産業上の利用可能性がある。
【００６８】
少なくとも１の実施形態において、ユーザーは、装置に対して、音楽の生成に際して依拠すべき音を指定することができる。一例として、装置は、依拠すべき音データを受け付ける。装置は、ユーザーが指定できる、２以上の音源が収録されており、ユーザーが任意

の音源を選択可能である。「２以上の音源が収録される」とは、装置自体がこれらの音源を記憶されている実施形態だけではなく、装置が外部の記憶媒体からこれらの音源を得る実施形態を含む。例えば、これらの音源は、既成の音楽である。ユーザーは、依拠すべき１以上の音データを選択し、装置は、選択された音楽に基づいて、音楽を生成する。非包括的な例として、動画が単位時間が５秒であり、選択された音データが７秒の拍の既成の音楽であった際には、装置は、依拠すべき既成の音楽の再生速度を調整して５秒の拍の音楽に変更する。この例においては、変更された音楽が、上記の実施形態における「生成した音楽」に相当する。上記の構成によれば、動画の拍と同期する既成の音楽を自動的に特定することができる。さらに、特定された既成の音楽は、単位時間に基づいた音であるから、動画の拍と一致する。この場合、提示された音は、動画と、生成された音を同時に再生すれば、そのまま産業利用可能である。このことによって、音を作成する者の作業量負担の軽減を図り、拍が合致する音楽を探索する作業負担を軽減し、又は動画と生成された音との一致度が高い利用可能性の高い音を自動で用意できる点で、産業上の利用可能性がある。これらの実施形態において、装置は、使用した既成の音に関するライセンス表示をすべきデータ又はプロンプトを、生成した音楽のデータ内に生成またはデータに付随して保存することができる。このような実施形態の利点は、少なくとも、著作権管理や使用許諾を円滑に行うシステムが提供できる点である。

【００６９】

少なくとも１の実施形態において、装置は、動画内の動作特徴（一例として、ダンスの動作）をより詳細に解析し、動作特徴（一例として、ダンスのステップ）ごとに異なる音楽要素（ビートやメロディ）を生成する。これらの音楽要素は、本開示におけるいずれかの方法に基づいて、採用され、又は連結される。これらの音楽要素は、生成された音楽の拍に矛盾しない限り、生成された音楽に付加されてもよい。

【００７０】

少なくとも１の実施形態において、装置は、動画内のオブジェクトの所定の動作特徴を特定動作特徴として認識し、特定動作特徴が発生した時刻に、所定の音である特定音を生成する。装置は、生成される音楽に付加して、特定音を生成することができる。別の実施形態において、装置は、生成される音楽とは独立して、特定音を生成することができる。ユーザーは、特定動作特徴又は特定音に関する情報を、装置に登録する。非包括的な例として、ユーザーは、人が驚く様子又は所定のリアクションを特定動作特徴、驚くという様子またはプロンプトを特定音に関する情報として、装置に登録する。特定音は、ユーザーが指定する既成の音楽や効果音を含む。装置は、このような既成の音楽や効果音の２以上を、自らが記憶し、又は外部の記憶装置から転送を受けることができる。装置は、そのような利用可能な特定音の２以上を、ユーザーに提示することができる。ユーザーは、その中から、利用すべき特定音を指定することができる。二以上の特定音を指定した場合、装置は、交互に、ランダムに、又は本開示の実施形態により合致する特定音を、生成すべき特定音として利用する。動画が２時間の映画であり、1時間20分43秒の時点で、映画の中の人物が驚いた様子を示した際、装置は、その時点において、びっくりした音を表す特定音を生成する。上記の構成によれば、特定動作特徴に基づく特定音を、自動的に生成することができる。さらに、生成された特定音は、特定動作特徴の発生した時刻に基づいた音であるから、動画の特定動作特徴の発生するタイミングと一致する。この場合、提示された音は、動画と、生成された音を同時に再生すれば、そのまま産業利用可能である。このことによって、特定音を作成する者の作業量負担の軽減を図り、又は動画と生成された音との一致度が高い利用可能性の高い音を自動で用意できる点で、産業上の利用可能性がある。なお、すでに説明した通り、本構成は、矛盾のない限り、本開示される別の実施形態における態様の１以上と組み合わせ、又は置き換えることができる。すなわち、配信中の環境音や観客の声をリアルタイムで解析し、それに応じたBGMや効果音（特定音）を生成することもできる。他の実施形態において、びっくりした動作に対してびっくりした効果音を自動で入れるアプリケーションとして、実施することができる。

【００７１】

少なくとも１の実施形態において、動画は、映画を含む。本実施形態によれば、映画のシークエンスごとに、最適な音楽又は特定音を別々に生成することができる。

【００７２】

上記とは異なる実施形態について開示する。

【００７３】

装置が、動画内のオブジェクトの特徴を特定し、特徴に基づいて音楽を生成することをAIに要求する、方法について開示する。この方法は、動作の特徴である動作特徴を特定し、動作特徴に基づいて音楽を生成することをAIに要求する、方法を含む。少なくとも１つの実施形態によれば、装置又はAIは、ビデオや画像をLLMのテキストトークンに似た小さなデータ単位であるビジュアルパッチの集合として表現する。装置は、映像認識AIとLLMを組み合わせて、動画の各シーンを理解する。具体的には、映像認識AIを活用し、シーンを構成する人物、車、建物、動物、樹木などの自然物、天気などの様々な物体や環境と、それらの変化を個別に認識する。AIは、先述した通りの学習方法（教師あり学習を含む）によって、ビジュアルパッチに対応する感情を学習する。さらに、AIは、対象分野のサンプル映像を使ってLLMを事前にファインチューニングされる場合がある。例えば、特定のビジュアルパッチについて、ムード又は雰囲気、感情（本開示において、これらを総称して「コンセプト」という）に対応することを、AIは学習する。装置は、動作特徴を基に、特徴に基づいて音楽を生成することをAIに要求する。一例として、動画が３分のミュージックビデオであり、ダンサーの表情が笑顔であった場合に、装置は、明るい、楽しいなどといった動作特徴を特定し、その特徴に基づいて音楽を生成することをAIに要求する。他の実施形態において、装置は、動画内のオブジェクトの動作の特徴である動作特徴を特定することも、AIに要求する。他の実施例として、動画が３分のミュージックビデオであり、ダンサーの背後が演出的に廃墟であった場合、装置は、付属的な特徴として廃墟、暗いなどといった付属特徴を特定し、その付属特徴に基づいて音楽を生成することをAIに要求する。他の実施例において、動画は映画であり、 動画のシーンや感情の流れを解析し、それに応じた音楽の構成（イントロ、クライマックス、エンディング）を生成することができる。また、コンセプトに対応した音楽は、音楽ジャンル（クラシック、EDM、ヒップホップなど）に対応する実施形態を含む。この実施形態によれば、動画の趣向又はコンセプトに沿った音楽を自動生成することができる。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音とのコンセプトの一致度が高い利用可能性の高い音を自動で用意できる点で、産業上の利用可能性がある。

【００７４】

少なくとも１の実施形態において、装置は、ユーザーから指定される特徴を動作特徴又は付属特徴とすることもできる。ユーザーは、生成すべき音楽のコンセプトを、文字で装置に入力する。これらの文字は、プロンプトである場合がある。装置は、ユーザーから入力されたコンセプトに基づいて、音楽を生成することをAIに要求する。さらに、装置は、動画の時間又は動作特徴を特定すべき時間と同一の時間の音楽を生成することを、AIに要求する場合がある。ある実施例において、ユーザーがムードや雰囲気を選択し、選択されたムードに合わせた音楽を生成することができる。この実施形態によれば、ユーザーが好みの趣向を指定することができるので、ユーザーの趣向に沿った音楽を自動生成することができる。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音とのコンセプトの一致度が高い利用可能性の高い音を自動で用意できる点で、産業上の利用可能性がある。

【００７５】

少なくとも１の実施形態において、装置は、ユーザーから指定される特徴として、歌詞を解析する。装置又はAIは、歌詞に含まれるコンセプトを特徴として認識する。例えば、装置又はAIは、歌詞に基づいて、恋愛、失恋、悲しい、といったコンセプトを抽出する。装置またはAIは、抽出されたコンセプトに基づいて、音を生成する。歌詞に基づいて２以上のコンセプトが認識された際は、後述する通り、特徴が認められる出現頻度が最も高い特徴に基づくことをAIに要求する。この方法によれば、複雑な歌詞であったとしても、主

要となる歌詞のコンセプトに基づいて音楽を生成できる。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音とのコンセプトの一致度が高い利用可能性の高い音を自動で用意できる点で、産業上の利用可能性がある。

【００７６】

少なくとも１の実施形態において、装置は、動画のオブジェクトを文字に変換することをAIに要求する。すでに説明した通り、AIは、装置又はAIは、ビデオや画像をLLMのテキストトークンに似た小さなデータ単位であるビジュアルパッチの集合として表現する。装置は、映像認識AIとLLMを組み合わせて、動画の各シーンを理解する。具体的には、映像認識AIを活用し、シーンを構成する人物、車、建物、動物、樹木などの自然物、天気などの様々な物体や環境と、それらの変化を個別に認識する。AIは、先述した通りの学習方法（教師あり学習を含む）によって、ビジュアルパッチに対応する文字を学習する。この方法によって、AIは、与えられた動画を文字で描写（文字に変換）することができる。AIは、変換された文字に基づいて、音楽を生成する。具体的には、装置は、変換された文字に基づいて特徴を特定し、特徴に基づいて音楽を生成することをAIに要求する。特徴に基づいて音楽を生成する方法については、本開示の別の箇所で開示する通りである。

【００７７】

特徴に基づいて音楽を生成することをAIに要求する際、２以上の特徴があった際には、少なくとも、特徴が認められる時間が相対的に長い特徴に基づくことをAIに要求する、方法について開示する。少なくとも１の実施形態において、装置が２以上の特徴を認めることがあり得る。この場合、動画において特徴が認められる時間が最長の特徴に基づくことをAIに要求する。装置は、各特徴が認められる時間を特定する。非包括的な例として、動画が３分の映画であり、そのうちの２分４０秒は失恋の様子を描写しているが、２０秒は過去の回想として、楽しく高揚した恋愛中の様子を描写していた場合、「失恋」の特徴は２分４０秒であり、「楽しい」の特徴は２０秒である。つまり、装置は、最長の特徴である「失恋」の特徴に基づいて音楽を作成することをAIに要求する。なお、この方法は、１の特徴に限定する趣旨ではなく、２以上の特徴に基づいて音楽を生成することも可能である。上記の構成によれば、複数のコンセプトを有する動画であっても、主要なコンセプトに基づいた音楽を生成できる。そのため、たとえ背反するコンセプトを内包する動画であっても、主要なコンセプトに基づく音楽であるから、動画とのコンセプトが一致する可能性が高い。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音とのコンセプトの一致度が高い利用可能性の高い音を自動で用意できる点で、産業上の利用可能性がある。

【００７８】

特徴に基づいて音楽を生成することをAIに要求する際、２以上の特徴があった際には、少なくとも、特徴が認められる出現頻度が相対的に高い特徴に基づくことをAIに要求する、方法について開示する。すでに説明した通り、少なくとも１の実施形態において、装置が２以上の特徴を認めることがあり得る。この場合、動画において特徴が認められる頻度が最も高い特徴に基づくことをAIに要求する。装置は、各特徴が認められる回数を特定する。非包括的な例として、動画が３分の映画であり、そのうち演者が笑っているの回数が１５回であり、泣いている回数が１回であり、怒っている回数が１回である場合、頻度が最も高い特徴は「笑い」である。つまり、装置は、最長の特徴である「笑い」の特徴に基づいて音楽を作成することをAIに要求する。なお、この方法は、１の特徴に限定する趣旨ではなく、２以上の特徴に基づいて音楽を生成することも可能である。上記の構成によれば、複数のコンセプトを有する動画であっても、主要なコンセプトに基づいた音楽を生成できる。そのため、たとえ背反するコンセプトを内包する動画であっても、主要なコンセプトに基づく音楽であるから、動画とのコンセプトが一致する可能性が高い。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音とのコンセプトの一致度が高い利用可能性の高い音を自動で用意できる点で、産業上の利用可能性がある。

【００７９】

特徴に基づいて音楽を生成することをAIに要求する際、動画の時間又は動作特徴を特定すべき時間と同一の時間の音楽を生成し、類似しない２以上の特徴が特定された際には、２以上の音楽を生成し、当該音楽を連結するよう要求する、方法について開示する。すでに説明した通り、装置またはAIは、２以上の特徴が特定された際には、２以上の音楽を生成することができる。また、すでに説明した通り、装置又はAIは、２以上の音楽を連結することができる。上記の構成によれば、類似しない２以上の特徴を有する複雑な動画であっても、動画のコンセプトと類似する音を自動的に作成することができる。さらに、２以上のコンセプトに基づいた音であるから、動画のコンセプトと不一致になる可能性が、さらに少なくなる。そのため、作成された音又は音楽は、動画のコンセプトと大部分が同期しているので、動画と生成された音との一致度がより高い。この場合、生成された音は、動画と、生成された音を同時に再生すれば、そのまま産業利用可能である。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音とのコンセプトの一致度が高い利用可能性の高い音を自動で生成できる点で、産業上の利用可能性がある。

【００８０】

前記方法において、２以上の特徴の出現頻度に基づき特定された時点で、当該音楽を連結するよう要求する、方法について開示する。すでに説明した通り、装置は、動画の時間帯における各特徴の出現頻度又は出現確率を算出できる。例えば、装置は、動画の時間又は動作特徴を特定すべき時間の範囲内で、Aという特徴に基づく単位時間の出現が相対的に多い時間帯と、Bという特徴の基づく単位時間の出現が相対的に多い時間帯とを特定する。装置は、これらの出現頻度に基づき、出現頻度が最も高い単位時間に基づく音楽を生成する。装置は、出現頻度が最も高い単位時間が変わった場合には、その新たな単位時間に基づく音楽を生成する。装置は、これらの２つの音楽を連結する。別の実施形態において、装置は、動画の時間（動作特徴を特定すべき時間）と同一の時間の音楽を生成し、類似しない２以上の動作特徴が特定された際には、２以上の音楽を生成する。装置は、類似しない２以上の動作特徴の出現頻度に基づき特定された時点を、変更時刻として特定する。例えば、装置は、動画の時間又は動作特徴を特定すべき時間の範囲内で、Aという特徴に基づく単位時間の出現が相対的に多い時間帯と、Bという特徴の基づく単位時間の出現が相対的に多い時間帯とを特定する。装置は、出現頻度の最も多い動作特徴が変わる時点、又は異なる動作特徴に基づく出現頻度の高い時間帯の間又は略中間時点を、変更時刻として特定できる。非包括的な例として、ある映画の動画が１０分であったとする。０分００秒から１分３０秒の時刻帯においては、Aという特徴（例えば、舞台が廃墟の中であるという特徴）に基づく単位時間の出現が相対的に多く、1分30秒から3分の時刻帯においては、Bという特徴（例えば、舞台が華やかな建物（城など）の中であるという特徴）に基づく単位時間の出現が相対的に多かった場合、装置は、２以上の動作特徴の出現頻度に基づき、その略中間時点である1分30秒を、特定する。装置は、特定された時点で、当該音楽を連結するよう要求する。すなわち、生成された音楽は、0分0秒から1分30秒までが、Aという特徴に基づくAという音楽、1分30秒から3分までが、Bという特徴に基づくBという音楽である。これらは、1分30秒の時点で連結される。

【００８１】

上記の構成によれば、すでに説明した通り、類似しない特徴による２以上の単位時間が特定されるような、複雑な動画であっても、動画のコンセプトと類似する音を自動的に作成することができる。さらに、２以上の単位時間に基づいた音であるから、動画のコンセプトと不一致になる可能性が、さらに少なくなる。そのため、作成された音又は音楽は、動画のコンセプトと大部分が同期しているので、動画と生成された音との一致度がより高い。この場合、生成された音は、動画と、生成された音を同時に再生すれば、そのまま産業利用可能である。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音とのコンセプトの一致度が高い利用可能性の高い音を自動で生成できる点で、産業上の利用可能性がある。

【００８２】

少なくとも１の実施形態において、歌詞又は文字に基づいて音楽を生成する際に、装置

は、文中においてリズムを示す文字又はコンセプトを認識し、その文字又はコンセプトに基づいたリズムの音楽を生成することをAIに要求する。上記の構成によれば、指定のリズムによる音楽を生成することができるので、生成された音楽は、動画のコンセプトと不一致になる可能性が、さらに少なくなる。そのため、生成された音は、動画と、生成された音を同時に再生すれば、そのまま産業利用可能である。このことによって、音を作成する者の作業量負担の軽減を図り、又は動画と生成された音とのコンセプトの一致度が高い利用可能性の高い音を自動で生成できる点で、産業上の利用可能性がある。

【００８３】

上記とは異なる実施形態について開示する。装置は、音楽データを基に歌詞を作ることを、AIに要求する。装置又はAIは、音楽データを、スケール（調）、コード（和音）又はメロディ（旋律）などの各単位に分解する。AIは、教師あり学習によって、各単位が示すムード、雰囲気又は感情を学習する。さらに、AIは、すでに公開されている音楽のスケール（調）、コード（和音）又はメロディ（旋律）及び歌詞の組み合わせによって、学習とファインチューニングを行う。このことによって、AIは、スケール（調）、コード（和音）又はメロディ（旋律）を与えられた場合に、どのような文字や歌詞が来るかを予測することができる。さらに来た文字または歌詞について、次にどのような単語が来るかを予測することもできる。このようにして、AIは、音楽データを基に歌詞を作ることができる。装置は、音楽データを基に歌詞を作ることを、AIに要求する。上記の構成によれば、音楽の伝統的なコンセプトに基づいた歌詞を生成することができるので、生成された歌詞は、音楽のコンセプトと不一致になる可能性が、少なくなる。さらに、AIは自然言語処理と学習をするので、生成された歌詞が文法的に誤りを含む可能性も、確かに低い。そのため、生成された歌詞は、音楽と付随してそのまま産業利用可能である。このことによって、歌詞を作成する者の作業量負担の軽減を図り、又は音楽と生成された歌詞とのコンセプトの一致度が高い利用可能性の高い歌詞を自動で生成できる点で、産業上の利用可能性がある。少なくとも１の実施形態において、装置は、さらに、生成した歌詞に基づいて、歌声を生成することをAIに要求する。少なくとも１の実施形態において、装置は、 音楽を生成することをAIに要求し、生成した音楽を基に歌詞を作ることをAIに要求し、生成された歌詞を基に歌声を生成することをAIに要求し、生成した歌声と音楽のデータを合わせる（統合する、組み合わせる）ことをAIに要求する（又は、装置自らがデータを合わせる）。歌声を生成するAIのアプリケーションとして、ボーカロイド（登録商標）を含む。この方法によれば、動画のコンセプトに合致する歌声を自動的に生成できる点で、産業上の利用可能性がある。

【００８４】

以下、上記において説明した実施形態の概要について開示する。

【００８５】

音を作成する方法であって、
装置が、
動画内のオブジェクトの動作の特徴である動作特徴を特定し、
類似する動作特徴間の時間を単位時間として、単位時間を拍とする音楽を生成することをAIに要求する、
方法。

【００８６】

単位時間を拍とする音楽を生成することをAIに要求する際、
動画の時間又は動作特徴を特定すべき時間と同一の時間の音楽を生成し、
動画内の単位時間と、音楽の拍が同一になるよう音楽を生成することを要求する、
方法。

【００８７】

単位時間を拍とする音楽を生成することをAIに要求する際、
類似する動作特徴から生ずる２以上の単位時間の中から、相対的に短いの単位時間を拍とするよう要求する、

方法。
【００８８】
単位時間を拍とする音楽を生成することをAIに要求する際、
類似する動作特徴から生ずる２以上の単位時間の中から、特定された単位時間の出現頻度が相対的に高い単位時間を拍とするよう要求する、
方法。
【００８９】
単位時間を拍とする音楽を生成することをAIに要求する際、
動画の時間又は動作特徴を特定すべき時間と同一の時間の音楽を生成し、
類似しない動作特徴による２以上の単位時間が特定された際には、２以上の音楽を生成し、
当該音楽を連結するよう要求する、
方法。
【００９０】
前記方法において、
２以上の動作特徴の出現頻度に基づき特定された時点で、当該音楽を連結するよう要求する、
方法。
【００９１】
音を作成する方法であって、
装置が、
動画内のオブジェクトの特徴（特徴量）を特定し、
特徴に基づいて音楽を生成することをAIに要求する、
方法。
【００９２】
特徴に基づいて音楽を生成することをAIに要求する際、
２以上の特徴があった際には、少なくとも、特徴が認められる時間が相対的に長い特徴に基づくことをAIに要求する、
方法
【００９３】
特徴に基づいて音楽を生成することをAIに要求する際、
２以上の特徴があった際には、少なくとも、特徴が認められる出現頻度が相対的に高い特徴に基づくことをAIに要求する、
方法
【００９４】
特徴に基づいて音楽を生成することをAIに要求する際、
動画の時間又は特徴を特定すべき時間と同一の時間の音楽を生成し、
類似しない２以上の特徴が特定された際には、２以上の音楽を生成し、
当該音楽を連結するよう要求する、
方法。
【００９５】
前記方法において、
２以上の特徴の出現頻度に基づき特定された時点で、当該音楽を連結するよう要求する、
方法。
【００９６】
少なくとも１の実施形態において、「類似する動作特徴から生ずる２以上の単位時間の中から、相対的に短いの単位時間を拍とするよう要求する」の記載は「類似する動作特徴から生ずる２以上の単位時間の中から、最短の単位時間を拍とするよう要求する」と置き換えることができる。「類似する動作特徴から生ずる２以上の単位時間の中から、特定された単位時間の出現頻度が相対的に高い単位時間を拍とするよう要求する」の記載は「類似する動作特徴から生ずる２以上の単位時間の中から、特定された単位時間の出現頻度が最

も高い単位時間を拍とするよう要求する」と置き換えることができる。「２以上の特徴があった際には、少なくとも、特徴が認められる時間が相対的に長い特徴に基づくことをAIに要求する」の記載は「２以上の特徴があった際には、少なくとも、特徴が認められる時間が最長の特徴に基づくことをAIに要求する」と置き換えることができる。「２以上の特徴があった際には、少なくとも、特徴が認められる出現頻度が相対的に高い特徴に基づくことをAIに要求する」の記載は「２以上の特徴があった際には、少なくとも、特徴が認められる出現頻度が最も高い特徴に基づくことをAIに要求する」と置き換えることができる。

【００９７】

音を作成する方法であって、
装置が、
動画の時間又は指定された時間と同一の時間の音楽を生成し、
動画に対応する歌詞の特徴に基づいて音楽を生成することをAIに要求する、
方法。

【００９８】

上記の構成において、指定された時間とは、ユーザーが指定する時間を含む。装置は、ユーザーから指定された時間の範囲内で、音楽を生成することをAIに要求する。歌詞の特徴に基づいて音楽を生成する方法及びその利点は、先述した通りである。

【００９９】

少なくとも１の実施形態において、ユーザーは、AIが生成した音楽又は歌声について、細かい音楽の調整（テンポ、キー、エフェクト）を、装置を通じて行うことができる。装置は、ユーザーの調整に応じて、生成された音楽または歌声のデータについて、テンポ、キー、エフェクトなどを変更する。当該調整は、動画がリアルタイムの動画又は配信動画であっても、行うことができる。

【０１００】

少なくとも１の実施形態において、２以上のユーザーは、AIが生成した音楽又は歌声について、生成すべきリズムやコンセプトの装置への入力、及び、細かい音楽の調整（テンポ、キー、エフェクト）を、装置を通じて共同で行うことができる。装置は、二以上のユーザーによる情報入力を受け付ける。２以上のユーザーは、装置を通じて、１つのデータを共同で編入することができる。当該編集は、動画がリアルタイムの動画又は配信動画であっても、行うことができる。非包括的な例として、この編集機能を「コラボレーションモード」又は「リアルタイムモード」という。

【０１０１】

少なくとも１の実施形態において、装置は、歌詞やコンセプトに応じて、１以上の外国語で歌詞を作成することをAIに要求する。装置は、入力された外国語の歌詞又はコンセプトに応じた音楽を作成することを、AIに要求する。AIは、外国語に応じたコンセプト、音楽についてファインチューニングを行う。非包括的な例として、日本というコンセプトには琴、ヨーロッパやアイルランドというコンセプトにはバグパイプ、といった民族音楽に関するコンセプトのファインチューニングを行う。また、装置は、外国語の歌詞が入力された場合、当該外国のコンセプトで音楽を生成することを、AIに要求する。

【０１０２】

少なくとも１の実施形態において、装置は、生成された音楽や動画が、主要な配信プラットフォーム（非包括的に、YouTube, TikTok, Instagramなど。いずれも登録商標）で動作するよう、互換性のある形式で音楽又は動画を生成することをAIに要求する。他の実施形態において、装置は、AIが生成した音楽や動画を、主要な配信プラットフォームで動作するよう、互換性のある形式に変換する。このような実施形態の利点は、少なくとも、プラットフォーム互換性のあるデータを作成し、利用者の利便性を向上させる点にある。

【０１０３】

本開示に係る発明は、上述した各効果のうち、少なくとも１つを奏することができればよい。

【要約】
１以上の実施形態が、音を作成する方法であって、装置が、動画内のオブジェクトの動作の特徴である動作特徴を特定し、類似する動作特徴間の時間を単位時間として、単位時間を拍とする音楽を生成することをAIに要求する、方法。
【選択図】なし

フロントページの続き

(56)参考文献　特開２００２－２８７７４６（ＪＰ，Ａ）
特表２０１８－５３７７２７（ＪＰ，Ａ）
特許第３５７８４６４（ＪＰ，Ｂ２）
特表２０２３－５１３５８６（ＪＰ，Ａ）
米国特許出願公開第２０２１／００２０１４９（ＵＳ，Ａ１）
韓国登録特許第１０－２７０３７６７（ＫＲ，Ｂ１）

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｈ　１／００－１／４６