

明 細 書

発明の名称： 動画編集装置及びプログラム

技術分野

[0001] 本発明は、動画ファイルを生成する技術に関し、特に、外国語の音声を含む動画ファイルから、字幕や吹替音声を含む動画ファイルを生成する技術に関する。

背景技術

[0002] この種の技術を開示した文献として、特許文献1、2、3がある。特許文献1に記載された受信装置は、マルチメディアプレゼンテーションをバッファリングし、このマルチメディアプレゼンテーションのオーディオコンポーネント及びビデオコンポーネントを取り出し、オーディオコンポーネントに音声認識分析を施して字幕テキストデータを生成し、この字幕テキストデータを元のビデオコンポーネントに統合して出力する。特許文献2に記載された文字起こしサーバは、ユーザのプロファイルに従って、音声認識ユニットによって生成されたテキストをユーザのターゲット言語に翻訳し、翻訳結果である翻訳テキストをユーザの携帯端末に送信する。特許文献3に記載された機械翻訳機は、オンラインビデオサービスにおける時限テキストトラックを処理対象とし、時限テキストトラックの時限テキストを、所定の外国語に翻訳し、時限テキストを書き換えたオンラインビデオサービスのコンテンツをユーザのビューアクライアントに送信する。

先行技術文献

特許文献

[0003] 特許文献1：米国特許第8781824号明細書

特許文献2：欧州特許第2725816号明細書

特許文献3：米国特許出願公開第2010/0138209号明細書

発明の概要

発明が解決しようとする課題

[0004] しかしながら、これまでに提案された動画コンテンツ吹替技術は、吹替音声におけるセリフや言説の言葉遣いが画一的であり、発言者の感情や言葉のニュアンスが視聴者に伝わり難いという問題があった。

[0005] 本発明は、このような課題に鑑みてなされたものであり、動画コンテンツ内におけるセリフや言説を発した人物の属性に応じて、言葉遣いを調整できる技術的手段を提供することを目的とする。

課題を解決するための手段

[0006] 上記課題を解決するため、本発明の好適な態様である動画編集装置は、処理対象の動画ファイル内の各音声を、前記音声を示す第1テキストデータ及び前記音声を発した人物の属性データとして認識する認識手段と、前記第1テキストデータを、別の言語の第2テキストデータに翻訳する翻訳手段と、前記属性データに基づいて、前記第2テキストデータが示す翻訳結果の一部の言葉遣いを改変した第3テキストデータを生成し、生成した第3テキストデータを、前記動画ファイル内の画像を示すデータとともに出力する改変手段とを具備することを特徴とする。

[0007] この態様において、前記認識手段は、前記動画ファイル内の音声の音声データに、FFT処理を施し、FFT処理により得られたスペクトラムデータに基づいて、当該音声を発した人物の属性を特定する属性認識処理手段を有してもよい。

[0008] また、前記認識手段は、前記動画ファイルにおける動画コンテンツの開始時から終了時までの間の音声データから、人の発声音の記録された部分の音声データのみを抽出する音声抽出手段と、前記音声抽出手段により抽出された音声データを、各々がひとまとまりの発声音を示す複数の音声データに分割する分割手段と、を具備し、前記属性認識処理手段は、前記分割手段により分割された複数の音声データの各々について、当該音声を発した人物の属性を特定してもよい。

[0009] 本発明の別の好適な態様であるプログラムは、コンピュータに、処理対象の動画ファイル内の各音声を、前記音声を示す第1テキストデータ及び前記

音声を発した人物の属性データとして認識する認識機能と、前記第1テキストデータを、別の言語の第2テキストデータに翻訳する翻訳機能と、前記属性データに基づいて、前記第2テキストデータが示す翻訳結果の一部の言葉遣いを改変した第3テキストデータを生成し、生成した第3テキストデータを、前記動画ファイル内の画像を示すデータとともに出力する改変機能とを実現させる。

発明の効果

- [0010] 本発明によると、外国語の音声を含む動画コンテンツから、発言者の感情や言葉のニュアンスが視聴者に伝わり易い吹替音声情報を含む動画コンテンツを生成することができる。

図面の簡単な説明

- [0011] [図1]本発明の一実施形態である動画編集システムの全体構成を示す図である。
- [図2]図1の動画編集システムの動作を示すフローチャートである。
- [図3]図2におけるステップS201～S203の処理内容を示す図である。
- [図4]図2におけるステップS204～S210の処理内容を示す図である。
- [図5]図2におけるS211～S214の処理内容を示す図である。

発明を実施するための形態

- [0012] 以下、図面を参照しつつ本発明の一実施形態を説明する。図1は、本発明の一実施形態である動画編集システム1の全体構成を示す図である。この動画編集システム1は、第1言語（例えば、英語）の音声を含む第1動画ファイルから、当該動画コンテンツ内の音声を第2言語（例えば、日本語）の字幕や吹替音声にした第2動画ファイルを生成するものである。
- [0013] 動画編集システム1は、端末10と、動画編集装置20とを有する。端末10と、動画編集装置20は、ネットワーク90を介して接続される。端末10は、ユーザが使用するスマートフォンやパーソナルコンピュータである。動画編集装置20は、サービスの提供者による管理の下に稼働するサーバである。動画編集装置20は、CPU21、RAM22、ROM23、ハー

ドディスク 24、通信インターフェス 25、ディスプレイ 26、キーボード 27、ポインティングデバイス 28を有する。CPU 21は、RAM 22をワークエリアとして利用しつつ、ROM 23やハードディスク 24に記憶されたプログラムを実行する。ハードディスク 24には、動画編集プログラム 29が記憶されている。

[0014] 動画編集プログラム 29は、以下の機能を有する。

a 1. 認識機能

これは、端末 10から、ネットワーク 90を介して、処理対象の第1動画ファイルを受け取り、第1動画ファイル内の第1言語の各音声を、音声を示す第1テキストデータ、及び、音声を発した人物の属性データとして認識する機能である。

b 1. 翻訳機能

これは、第1テキストデータを、第2言語の第2テキストデータに翻訳する機能である。

c 1. 改変機能

これは、属性データに基づいて、第2テキストデータが示す翻訳結果の一部の言葉遣いを改変した第3テキストデータを生成し、生成した第3テキストデータを、動画ファイル内の画像を示すデータとともに出力する機能である。

[0015] 次に、本実施形態の動作を説明する。図2は、本実施形態の動作を示すフローチャートである。図3は、図2におけるステップS201～S203の処理内容を示す。図4は、図2におけるステップS204～S210の処理内容を示す。図5は、図2におけるS211～S214の処理内容を示す。これらの図におけるステップS201～S214、及びS231は、動画編集プログラム 29により実行されるものである。

[0016] 図2のステップS100において、端末10は、アップロード処理を行う。アップロード処理では、ユーザは、端末10のメモリ内から、所望の動画ファイルを選択する操作を行う。この操作が行われると、端末10は、選択

された動画ファイルを、処理対象の第1動画ファイルとして、動画編集装置20に送信する。動画編集装置20のCPU21は、端末10から、第1動画ファイルを受信すると、第1動画ファイルをRAM22に記憶する。

[0017] 図2及び図3のステップS201において、動画編集装置20のCPU21は、ノイズ除去処理を行う。ノイズ除去処理では、CPU21は、RAM22内の第1動画ファイルを復号し、復号結果である動画コンテンツの開始時から終了時までの間の一連のフレーム画像データ、及び当該動画コンテンツの開始時から終了時までの間の音の音波形を示す音声データをRAM22に記憶し、RAM22内の音声データからノイズを除去する。このノイズ除去は、公知のノイズ除去プログラムにより行うとよい。

[0018] 図2及び図3のステップS202において、動画編集装置20のCPU21は、音声抽出処理を行う。音声抽出処理では、CPU21は、RAM22内におけるノイズ除去が施された音声データから、動画コンテンツ内における人の発声音（セリフや言説など）の記録された部分の音声データのみを抽出し、この音声データをRAM22に記憶する。より具体的には、ノイズが除去された音声データにおける複数の音トラックから、人間の発声音が記録されているトラックを探索し、このトラックに記録されている音波形の音声データのみを抽出し、後続処理の対象の音声データとしてRAM22に記憶する。

[0019] 図2及び図3のステップS203において、動画編集装置20のCPU21は、音声分割処理を行う。音声分割処理では、CPU21は、RAM22内における人の発声音の音声データを、各々がひとまとまりの発声音を示す複数の音声データに分割する。より具体的に説明すると、音声分割処理では、CPU21は、元の音声データが示す動画コンテンツの開始時から終了時までの間の全音波形を先頭から走査し、無音のポイントが出現する度に、そのポイントよりも前の音波形を切り取って1つの音声データにする、という処理を繰り返す。この音声分割処理により得られた複数の音声データは、1人の人物が発した1つのセリフや言説を示すものとなる。

- [0020] 図2及び図4のステップS204において、動画編集装置20のCPU21は、属性認識処理を行う。属性認識処理では、CPU21は、RAM22内の複数の音声データの各々について、当該音声データが示す音声を発した人物の属性を認識し、認識した属性を示す属性データを当該音声データと対応づけてRAM22に記憶する。この属性には、音声を発した人物の大まかな年齢（例えば、幼児、10代未満、20代、30代、40代、50代、60代以上）、性別（例えば、女性または男性）、声のトーン（例えば、高い、普通、低い）がある。より具体的には、属性認識処理では、各音声データにFFT（Fast Fourier transform）処理を施し、FFTにより得られたスペクトラムデータに基づいて、当該音声を発した人物の属性を特定する。スペクトラムデータに基づく属性の特定は、音声データを属性認識用の学習済みモデルに入力することにより行うとよい。
- [0021] 図2及び図4のステップS205において、動画編集装置20のCPU21は、トランスクリプション処理を行う。トランスクリプション処理では、CPU21は、RAM22内における音声分割処理により分割された音声データの各々について、当該音声データに文字越し処理を施し、文字越し処理の処理結果を、当該音声データの音声を示す第1テキストデータとし、この第1テキストデータを、音声コンテンツ内における当該音声の出現タイミングを示すタイムスタンプと対応付けてRAM22に記憶する。
- [0022] 図2及び図4のステップS206において、動画編集装置20のCPU21は、辞書調整処理を行う。辞書調整処理では、ステップS205のトランスクリプション処理により得られた単語を辞書ライブラリに登録する。
- [0023] 図2及び図4のステップS207において、動画編集装置20のCPU21は、再構成処理を行う。再構成処理では、CPU21は、RAM22内における第1テキストデータを、より長い纏まりの第1テキストデータに再構成する。例えば、RAM22内における連続する6つの音声の第1テキストデータTX(30)、TX(31)、TX(32)、TX(33)、TX(34)、TX(35)（カッコ内は、動画コンテンツ内の先頭から数えた当

該音声の出現順を示す数)が、意味のある1つのパラグラフをなしている場合、それら6つの音声の第1テキストデータTX(30)、TX(31)、TX(32)、TX(33)、TX(34)、TX(35)を纏めて1つの第1テキストデータTX'(30)にする。

[0024] 図2及び図4のステップS208において、動画編集装置20のCPU21は、翻訳処理を行う。翻訳処理では、CPU21は、RAM22における再構成済みの第1テキストデータを、第2言語に翻訳し、この翻訳結果を、元の音声データが示す動画コンテンツの開始時から終了時までの間の各パラグラフの翻訳字幕を示す第2テキストデータとしてRAM22に記憶する。

[0025] 図2及び図4のステップS209において、動画編集装置20のCPU21は、パラグラフ分割処理を行う。パラグラフ分割処理では、CPU21は、RAM22における翻訳を経た第2テキストデータを、動画コンテンツ内における同じ画像に表示させる字幕と対応する第2テキストデータに分割する。例えば、RAM22内における1つのパラグラフの第2テキストデータTX'(30)が、2つの画面に分かれて表示される字幕と対応するものである場合、この第2テキストデータTX'(30)を、先の画面の字幕と対応する第2テキストデータTX''(30-33)と、後の画面の字幕と対応する第2テキストデータTX''(34-35)に分割する。

[0026] 図2及び図4のステップS210において、動画編集装置20のCPU21は、言語品質保証(LQA:Linguistic Quality Assurance)処理を行う。言語品質保証処理では、CPU21は、RAM22内の属性データを基に、RAM22内の第2テキストデータが示す翻訳結果の一部の言葉遣いを改変した第3テキストデータを生成し、生成した第3テキストデータをRAM22に記憶する。例えば、RAM22内における1つの画面の字幕と対応する第2テキストデータTX''(30-33)が、4つの音声からなり、これら4つの音声と対応する属性データが、大まかな年齢=20代、性別=女性、声のトーン=普通、のものであった場合、言語品質保証処理では、第2テキストデータTX''(30-33)を、20代女性の普通のトーンの言葉

遣いにしたものを、第3テキストデータTX' ' ' (30-33)とする。

また、次の画面の字幕と対応する第2テキストデータTX' ' (34-35)が、2つの音声からなり、これら2つの音声と対応する属性データが、大まかな年齢=20代、性別=女性、声のトーン=高い、のものであった場合、言語品質保証処理では、第2テキストデータTX' ' (34-35)を、20代女性の高いトーンの言葉遣いにしたものを、第3テキストデータTX' ' ' (34-35)とする。この第3テキストデータの生成は、第2テキストデータを、言語品質保証処理用の学習済みモデルに入力することにより行うとよい。

[0027] 図2及び図5のステップS211において、動画編集装置20のCPU21は、タイムスタンプアジャスト処理を行う。タイムスタンプアジャスト処理では、CPU21は、動画編集装置20のディスプレイ26に、タイムスタンプアジャスト画面を表示させる。タイムスタンプアジャスト画面は、RAM22内における一連の第3テキストデータが示す字幕と、動画コンテンツ内における当該第3テキストデータが示す字幕の出現タイミングとを対応付けて示すものである。オペレータは、タイムスタンプアジャスト画面において、字幕の出現タイミングを修正する操作を行うことができる。字幕の出現タイミングを修正する操作が行われると、CPU21は、この操作に応じて、RAM22内における該当の第3テキストデータと対応するタイムスタンプを変更する。

[0028] 図2及び図5のステップS212において、動画編集装置20のCPU21は、後編集処理を行う。後編集処理では、CPU21は、動画編集装置20のディスプレイ26に、後編集画面を表示させる。後編集画面は、RAM22内における一連の第3テキストデータが示す字幕とその修正欄とを示すものである。オペレータは、後編集画面において、字幕を修正する操作を行うことができる。字幕を修正する操作が行われると、CPU21は、この操作に応じて、RAM22内における該当の第3テキストデータを変更する。

[0029] 図2及び図5のステップS213において、動画編集装置20のCPU2

1 は、テキスト音声変換処理を行う。テキスト音声変換処理では、CPU 21 は、RAM 22 内の一連の第3テキストデータを、そのデータと対応する吹替音声データに変換し、この吹替音声データをRAM 22 に記憶する。テキストから吹替音声への変換は、第3テキストデータを、変換用の学習済みモデルに入力することにより行うとよい。

[0030] 図2及び図5のステップS214において、動画編集装置20のCPU 21 は、音声調整処理を行う。音声調整処理では、CPU 21 は、動画編集装置20のディスプレイ26に、音声調整画面を表示させる。音声調整画面は、RAM 22 内における吹替音声データの音波形とその調整操作子とを示すものである。オペレータは、音声調整画面において、吹替音声を調整する操作を行うことができる。吹替音声を調整する操作が行われると、CPU 21 は、この操作に応じて、RAM 22 内における吹替音声データを変更する。

[0031] 図2のステップS130において、端末10は、レビュー&アジャスト処理を行う。レビュー&アジャスト処理では、端末10は、動画編集装置20のRAM 22 内における1つの動画コンテンツをなす画像データ、第3テキストデータ、及び吹替音声データを受信し、これらのデータを時系列に従って配置したレビュー&アジャスト画面を表示させる。ユーザは、レビュー&アジャスト画面において、第3テキストデータが示す字幕、または、吹替音声データが示す吹替音声を修正する操作を行うことができる。字幕、または、吹替音声を修正する操作が行われると、端末10は、修正内容を示すデータを動画編集装置20に送信する。動画編集装置20のCPU 21 は、修正内容を示すデータを受信すると、このデータに従って、RAM 22 内における第3テキストデータ、及び吹替音声データを変更する。

[0032] 図2及び図5のステップS231において、動画編集装置20のCPU 21 は、動画公開処理を行う。動画公開処理では、CPU 21 は、RAM 22 内における1つの動画コンテンツをなす画像データ、第3テキストデータ、及び吹替音声データを、指定された動画投稿サイトにアップロードする。このアップロードは、画像データと第3テキストデータとを含む字幕版の動画

コンテンツの動画ファイルとして行ってもよいし、画像データと吹替音声データとを含む吹替版の動画コンテンツの動画ファイルとして行ってもよい。また、動画コンテンツをなす画像データ、第3テキストデータ、及び吹替音声データは、非圧縮のファイルとしてアップロードしてもよいし、圧縮符号化したファイルとしてアップロードしてもよい。字幕版の動画コンテンツのアップロードには、2つの態様がある。第1の態様は、動画コンテンツをなす画像データを動画ファイルとし、第3テキストデータを字幕ファイルとし、画像と字幕を別のファイルとしてアップロードする態様である。第2の態様は、動画コンテンツの画像データの動画ファイルに、第3テキストデータを字幕として書き込み、画像と字幕を1つのファイルとしてアップロードする態様である。第1の態様では、動画ファイルを再生するアプリケーションが字幕ファイルに対応していない場合、字幕を表示させることができないが、第2の態様であれば、動画ファイルを再生するアプリケーションが字幕ファイルに対応していない場合でも、字幕を表示させることができる。

[0033] 以上が、本実施形態の詳細である。本実施形態である動画編集装置20は、処理対象の動画ファイル内の各音声を、音声を示す第1テキストデータ、及び、音声を発した人物の属性データとして認識する認識手段と、第1テキストデータを、別の言語の第2テキストデータに翻訳する翻訳手段と、属性データに基づいて、第2テキストデータが示す翻訳結果の一部の言葉遣いを改変した第3テキストデータを生成し、生成した第3テキストデータを、動画ファイル内の画像を示すデータとともに出力する改変手段とを具備する。よって、外国語の音声を含む動画コンテンツから、発言者の感情や言葉のニュアンスが視聴者に伝わり易い吹替音声情報を含む動画コンテンツを生成することができる。

[0034] 以上、本発明の実施形態について説明したが、この実施形態に以下の変形を加えてもよい。

[0035] (1) 上記実施形態において、属性認識処理で認識する属性の種類は、音声を発した人物の適当な年齢、性別、及び声のトーンに限らない。これら

以外の属性を特定してもよい。例えば、音声を発した人物の時代背景（中世や近世などの生きた人物であるか、それとも、現代に生きている人物か）をその人物の属性として認識してもよい

[0036] （２）上記実施形態において、タイムスタンプアジャスト処理及び後編集処理の両方、及び一方を実行しなくてもよい。タイムスタンプアジャスト処理及び後編集処理の両方を実行しない態様では、言語品質保証処理の実行後、直ちに、テキスト音声変換処理に進み、言語品質保証処理による言葉遣いの改変を経た第３テキストデータを、そのデータと対応する吹替音声データに変換する。

[0037] （３）上記実施形態において、音声調整処理及びレビュー＆アジャスト処理の両方、及び一方を実行しなくてもよい。音声調整処理及びレビュー＆アジャスト処理の両方を実行しない態様では、テキスト音声変換処理の実行後、直ちに、動画公開処理に進み、動画コンテンツをなす画像データ、第３テキストデータ、及び吹替音声データを、指定された動画投稿サイトにアップロードする。

符号の説明

- [0038] 1 動画編集システム
- 10 端末
 - 20 動画編集装置
 - 21 CPU
 - 22 RAM
 - 23 ROM
 - 24 ハードディスク
 - 25 通信インターフェース
 - 27 ディスプレイ
 - 28 キーボード
 - 29 ポインティングデバイス
 - 29 動画編集プログラム

90 ネットワーク

請求の範囲

- [請求項1] 処理対象の動画ファイル内の各音声を、前記音声を示す第1テキストデータ及び前記音声を発した人物の属性データとして認識する認識手段と、
- 前記第1テキストデータを、別の言語の第2テキストデータに翻訳する翻訳手段と、
- 前記属性データに基づいて、前記第2テキストデータが示す翻訳結果の一部の言葉遣いを改変した第3テキストデータを生成し、生成した第3テキストデータを、前記動画ファイル内の画像を示すデータとともに出力する改変手段と
- を具備することを特徴とする動画編集装置。
- [請求項2] 前記認識手段は、
- 前記動画ファイル内の音声の音声データに、FFT処理を施し、FFT処理により得られたスペクトラムデータに基づいて、当該音声を発した人物の属性を特定する属性認識処理手段を有する
- ことを特徴とする請求項1に記載の動画編集装置。
- [請求項3] 前記認識手段は、
- 前記動画ファイルにおける動画コンテンツの開始時から終了時までの間の音声データから、人の発声音の記録された部分の音声データのみを抽出する音声抽出手段と、
- 前記音声抽出手段により抽出された音声データを、各々がひとまとまりの発声音を示す複数の音声データに分割する分割手段と、
- を具備し、
- 前記属性認識処理手段は、前記分割手段により分割された複数の音声データの各々について、当該音声を発した人物の属性を特定する
- ことを特徴とする請求項2に記載の動画編集装置。
- [請求項4] コンピュータに、
- 処理対象の動画ファイル内の各音声を、前記音声を示す第1テキス

トデータ及び前記音声を発した人物の属性データとして認識する認識機能と、

前記第1テキストデータを、別の言語の第2テキストデータに翻訳する翻訳機能と、

前記属性データに基づいて、前記第2テキストデータが示す翻訳結果の一部の言葉遣いを改変した第3テキストデータを生成し、生成した第3テキストデータを、前記動画ファイル内の画像を示すデータとともに出力する改変機能と

を実現させるプログラム。