(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7216771号
(P7216771)

(45)発行日　令和5年2月1日(2023.2.1)

(24)登録日　令和5年1月24日(2023.1.24)

(51)国際特許分類
Ｇ１０Ｌ　15/00　(2013.01)

ＦＩ
Ｇ１０Ｌ　15/00　２００Ｇ

請求項の数　6　（全19頁）

(21)出願番号　特願2021-96807(P2021-96807)
(22)出願日　令和3年6月9日(2021.6.9)
(65)公開番号　特開2022-188622(P2022-188622A)
(43)公開日　令和4年12月21日(2022.12.21)
審査請求日　令和3年6月9日(2021.6.9)

(73)特許権者　399041158
西日本電信電話株式会社
大阪府大阪市都島区東野田町四丁目１５番８２号
(73)特許権者　593119413
讀賣テレビ放送株式会社
大阪府大阪市中央区城見１丁目３番５０号
(74)代理人　100130513
弁理士　鎌田 直也
(74)代理人　100074206
弁理士　鎌田 文二
(74)代理人　100130177
弁理士　中谷 弥一郎
(74)代理人　100161746

最終頁に続く

(54)【発明の名称】　台本へのメタデータ付与装置、方法、およびプログラム

(57)【特許請求の範囲】

【請求項１】

放送に表示する字幕に用いる台本のテキストにメタデータを付与するメタデータ付与装置であって、

前記放送の少なくとも一部分を音声認識した音声認識テキストと、当該放送の前記一部分を含む発言内容である台本のテキストである台本テキストとをそれぞれ形態素分割する形態素分割手段と、

前記音声認識テキストと、前記台本テキストのそれぞれについて、形態素分割されたテキスト同士を比較し、前記台本テキストの一致度が高い箇所に、前記音声認識テキストに由来するタイムスタンプを含むメタデータを付与するメタデータ付与手段と、

を実行するメタデータ付与装置であって、

上記メタデータ付与手段における、一致度が高い箇所が、

前記音声認識テキストと、前記台本テキストのそれぞれについて、形態素分割されたテキストを連続的に複数個連結させた連結パターン同士を比較し、前記台本テキストの連結パターンができるだけ長い連結数となる照合できた前記連結パターン単位の箇所である

メタデータ付与装置。

【請求項２】

前記メタデータ付与手段により前記タイムスタンプを付与した前記台本テキストの各行について、前記タイムスタンプの整合性を確認する整合性確認手段と、

整合性が満たされなかった行に対して、前後の整合性が満たされた行の前記タイムスタ

ンプに基づいた補正タイムスタンプを付与する補正手段と、
を有する請求項１に記載のメタデータ付与装置。
【請求項３】
前記メタデータ付与手段が、前記照合を行う際に、前記台本内における位置と、前記放送の時間中における位置とを元に探索する範囲を限定する
請求項１又は２に記載のメタデータ付与装置。
【請求項４】
前記メタデータ付与手段が比較する連結パターンに用いる形態素分割されたテキストが、仮名化されたものである、請求項１乃至３のいずれかに記載のメタデータ付与装置。
【請求項５】
台本を有する放送に表示する字幕に用いる台本のテキストにメタデータを付与するメタデータ付与方法であって、
メタデータ付与装置が、
前記放送の少なくとも一部分を音声認識した音声認識テキストと、当該放送の前記一部分を含む台本のテキストである台本テキストとをそれぞれ形態素分割するステップと、
前記音声認識テキストと、前記台本テキストのそれぞれについて、形態素分割されたテキストを連続的に複数個連結させた連結パターン同士を比較し、前記台本テキストの連結パターンができるだけ長い連結数となる照合できた前記連結パターン単位の箇所について、前記音声認識テキストに由来するタイムスタンプを前記台本テキストの当該箇所に付与するステップと、
を実行するメタデータ付与方法。
【請求項６】
コンピュータを、請求項１乃至４のいずれか１項に記載のメタデータ付与装置として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は、放送用の字幕に関する。
【背景技術】
【０００２】
テレビ放送に表示する字幕は、番組内容を人が聞き、トーク部分を正しく認識して、パソコンへテキストで入力する完全手作業で生成する方法が一般的である。ただし、手作業に依存する部分が大きすぎるため、完成までに時間がかかり、ミスを防止するために二重以上の確認作業を行わなければならないといった問題があった。
【０００３】
これに対して、字幕を効率的に生成する手段として、音声認識技術の利用が考えられている。ただし、本出願時点の音声認識技術では音声認識の精度に限界があり、字幕を生成したい番組に例えば話者が複数に及ぶ箇所がある場合や、ＢＧＭ等の効果音が挿入されている箇所などは、正しくトーク部分をテキストへ変換できないという問題があった。正しく変換できなかった部分は手作業で修正を行わなければならず、音声認識技術を利用しても効率の面で十分ではなかった。
【０００４】
それをさらに解決するために、台本のテキストを読み込み、音声認識された音声テキストと比較して相違点を検出する字幕番組データ制作システムが特許文献１に記載されている。台本のテキストを用いることでテキストの正確性が向上する。台本だけではタイミングを得ることができないが、音声認識によって文字列が出現するタイミングは正確に知ることができる。このため、タイミングを音声認識から取得し、テキストのミスを台本を参照して補正することで相互補完的な効果を発揮できる。
【０００５】
また、特許文献２には、場面に登場する人物などを画像認識し、場面の特徴と台本情報

とを対応付けて正確性を向上させる手法が提案されている。
【先行技術文献】
【特許文献】
【０００６】
【文献】特開２００４－３３４１３３号公報
特開２００５－２５４１３号公報
【発明の概要】
【発明が解決しようとする課題】
【０００７】
しかしながら、特許文献１に記載の技術では、相違点を検出することができても、検出された相違データに基づいてオペレータが手作業で修正するので（段落［００４５］等）、手作業をある程度軽減することはできてもその効果は不十分なものであった。
【０００８】
また、特許文献２に記載の技術では、実際の場面の画像と台本内容とを比較して対応させるものの、用いる音声認識結果は認識間違いになり不完全な文字列となってしまう場合がある。音声認識による不完全な文字列では、場面の画像を認識されたテキストと一致しなくなり、機械的な対応づけは難しくなることがあった。
【０００９】
そこでこの発明は、放送用の字幕を作成するにあたって、台本のテキストと音声認識を併用する際の、音声認識の不完全さによる対応付けのために手作業が必要な部分を減らし、機械的に効率よく精度の高い字幕を生成できるようにすることを目的とする。
【課題を解決するための手段】
【００１０】
この発明は、
放送に表示する字幕に用いるテキストにメタデータを付与するメタデータ付与装置であって、
前記放送の少なくとも一部分を音声認識した音声認識テキストと、当該放送の前記一部分を含む発言内容である台本のテキストである台本テキストとをそれぞれ形態素分割する形態素分割手段と、
前記音声認識テキストと、前記台本テキストのそれぞれについて、形態素分割されたテキスト同士を比較し、一致度が高い箇所に、前記音声認識テキスト由来のタイムスタンプを含むメタデータを付与するメタデータ付与手段と、
を有するメタデータ付与装置によって、上記の課題を解決したのである。
【００１１】
音声認識の一部が不正確であっても、音声認識テキストと台本テキストとのそれぞれを形態素分割した上で比較することで、一致度が高い箇所を捜索することが可能となる。形態素分割したテキスト同士の一致度は、オペレータを必要とする手作業ではなくコンピュータにおけるソフトウェア処理によって、所定の一致度の算定方式に従って自動的に行うことができる。タイムスタンプは音声認識から得られる時刻情報を割り当てることができ、話者識別は台本由来でも音声認識由来でもどちらでもよい。
【００１２】
上記の一致度の算定方式としては、前記メタデータ付与手段での一致度が高い箇所を、前記音声認識テキストと、前記台本テキストのそれぞれについて、形態素分割されたテキストを連続的に複数個連結させた連結パターン同士を比較し、前記台本テキストの連結パターンができるだけ長い連結数となる連続して照合できた箇所とする方式を採用することができる。分割された形態素を連続的に複数個連結させた連結パターンを作成すると、形態素が複数個繋がった連結パターン同士でならば一致する部分がある程度は出現する。その一致する部分ができるだけ長く連続して照合できた部分は、音声認識の一部が不正確であっても十分に一致する可能性が高い部分であると言える。文字列同士である連結パターン同士が一致するか否かを照合する作業は、オペレータを必要とする手作業ではなくコン

ピュータにおけるソフトウェア処理によって実行できる。
【００１３】
この発明にかかるメタデータ付与装置は、上記の手段に加えてさらに、前記メタデータ付与手段により前記タイムスタンプを付与した前記台本テキストの各行について、前記タイムスタンプの整合性を確認する整合性確認手段と、整合性が満たされなかった行に対して、前後の整合性が満たされた行の前記タイムスタンプに基づいた補正タイムスタンプを付与する補正手段と、を実行する実施形態を採用することができる。特に音声認識による正確性の高いタイムスタンプを自動的に台本と照合したテキストに付与し、そのタイムスタンプの整合性を確認して整合性を満たすように補正するという作業を自動的に行うことで、字幕に用いるために必要なメタデータ付テキストを自動化して生成することができる。連結パターン同士で照合したものに自動的に付与したタイムスタンプは、タイミングが同時になってしまったりして、タイムスタンプの時刻が単調増加にならなくなってしまうことがある。また、順番が前後してしまうこともある。さらに、話者識別の整合性がとれない場合もある。そのような前記連結パターンについてはタイムスタンプや話者識別のメタデータを自動的に補正する工程を設けることで、字幕に用いるメタデータに高い正確性を確保することができる。
【００１４】
この発明にかかるメタデータ付与装置は、
前記メタデータ付与手段が、前記照合を行う際に、前記台本内における位置と、前記放送の時間中における位置とを元に探索する範囲を限定する構成を採用することができる。番組が長くテキストが長大になる場合に、番組のテキスト全てを検索して照合すると処理負荷が大きく、本来の箇所とは違う箇所で照合できてしまう可能性も高くなる。探索範囲を予め絞り込んでおくことで、照合の負荷が軽減され、正確性も向上する。
【００１５】
また、この発明にかかるメタデータ付与装置は、
前記メタデータ付与手段が比較する連結パターンに用いる形態素分割されたテキストが、仮名化されたものである構成を採用することができる。音声認識の際に漢字変換が間違っている場合があり、そのままでは正しく分割されていても照合できなくなる場合がある。テキストを仮名化しておくことで、照合できる可能性を向上することができる。
【００１６】
この発明にかかるメタデータ付与方法は、台本を有する放送に表示する字幕に用いるテキストにメタデータを付与する字幕付与方法であって、
前記放送の少なくとも一部分を音声認識した音声認識テキストと、当該放送の前記一部分を含む台本のテキストである台本テキストとを形態素分割するステップと、
前記音声認識テキストと、前記台本テキストのそれぞれについて、形態素分割されたテキストを連結させた連結パターン同士を比較し、前記台本テキストの連結パターンができるだけ長い連結数となる連続して照合できた箇所に、前記音声認識テキストに由来するタイムスタンプを付与するステップと、
前記タイムスタンプを付与した前記台本テキストの各行について、前記タイムスタンプの整合性を確認するステップと、
整合性が満たされなかった行に対して、前後の整合性が満たされた行のタイムスタンプに基づいた補正タイムスタンプを付与するステップと、
を実行する。
【００１７】
この発明にかかるメタデータ付与プログラムは、コンピュータをメタデータ付与装置として機能させるためのプログラムである。
【発明の効果】
【００１８】
この発明にかかるメタデータ付与装置により、タイムスタンプや話者識別などのメタデータを付与した字幕用テキストが、オペレータの手作業を必要とすることなく高い精度で

作成できる。
【図面の簡単な説明】
【００１９】
【図１】この発明の第一の実施形態にかかるメタデータ付与装置が処理するフローの例
【図２】台本の例を示す図
【図３】メタデータ付与テキストの例を示す図
【図４】トークデータの例を示すテーブル
【図５】台本テキストの例を示すテーブル
【図６】（ａ）ＦＳ＝１のときのテキスト照合部における出力フォーマットの例を示すテーブル、（ｂ）ＦＳ＝２のときのテキスト照合部における出力フォーマットの例を示すテーブル、（ｃ）ＦＳ＝３のときのテキスト照合部における出力フォーマットの例を示すテーブル
【図７】図１のメタデータ付与装置のテキスト照合部における処理フローの例図
【図８】音声認識テキストを形態素分割した出力フォーマットの例を示すテーブル
【図９】台本テキストを形態素分割した出力フォーマットの例を示すテーブル
【図１０】図７のテキスト照合部のメタデータ付与ステップにおける処理フローの例図
【図１１】音声認識テキストを形態素分割した結果のフォーマットの例を示すテーブル
【図１２】音声認識テキストの形態素分割結果にメタデータを付与したフォーマットの例を示すテーブル
【図１３】音声認識テキストの形態素分割結果から生成させた連結パターンの例を示すテーブル
【図１４】台本テキストを形態素分割した結果のフォーマットの例を示すテーブル
【図１５】台本テキストの形態素分割結果から生成させた連結パターンの例を示すテーブル
【図１６】台本テキストの形態素分割した連結パターンに照合させたタイムスタンプを付与させた結果の例を示すテーブル
【図１７】図１６の例における各々の形態素の最大連結数の例を示すテーブル
【図１８】図１７の各々の形態素にタイムスタンプ及び話者識別を付与した例を示すテーブル
【図１９】図１８に示す各形態素のタイムスタンプを台本テキストの各行に付与し、台本テキストに代表するタイムスタンプを付与したフォーマットの例を示すテーブル
【図２０】整合性確認ステップを行う台本テキストのフォーマットの例を示すテーブル
【図２１】整合性フラグを付与した台本テキストのフォーマットの例を示すテーブル
【発明を実施するための形態】
【００２０】
以下、この発明について具体的な実施形態とともに詳細に説明する。この発明は、台本を有する放送に表示する字幕に用いるテキストにメタデータを付与するメタデータ付与装置、メタデータ付与方法、およびそのプログラムである。
【００２１】
図１に、この発明の第一の実施形態にかかるメタデータ付与装置１が処理するフローの例を示す。音声ファイル２と、台本３とが入力され、これらから得たデータをもとに、メタデータが付与された字幕用のテキストを生成する。台本３の中身の例を、テキストファイルとしたものを図２に示す。この発明において台本とは、放送の少なくとも一部分を含む発言内容をいう。この台本は具体的には、いわゆる脚本と呼ばれる撮影開始前に予め作られた複数の発言者とセリフとの組み合わせに限らず、一人の人間が読み上げ続けるニュースなどの原稿を含む。また、撮影開始前に作られたものに限られず、即興劇や街頭インタビューなどを含む放送内容を撮影してから速記して作成したテキストも含まれる。図２に示す台本３の例では一人の人間が読み上げる原稿を示している。また、この発明にかかるメタデータ付与装置によって得られるメタデータ付与テキストの例を図３に示す。
【００２２】
メタデータ付与装置１は、一台のコンピュータであってもよく、複数台のコンピュータ

によって形成されてもよい。ネットワーク上に存在するサーバであってもよく、仮想的なサーバであってもよい。以下に説明する各部、各手段は、コンピュータやサーバ、又はそれらの一部として実装される専用のハードウェアであってもよく、コンピュータ上や仮想サーバ上でソフトウェアとして実行可能な機能群であってもよい。

【００２３】

メタデータ付与装置１は、音声認識部１１を有すると好ましい。音声認識部１１は、字幕を付そうとする放送の一部又は全部を録音した音声ファイル２を取り込んで、音声認識により時刻データ付の音声認識テキストであるトークデータ４に変換する。ここで用いる音声ファイル２は、前記放送の内容を録音した音声ファイル２である。放送を録音して音声ファイル２を生成するにあたっては、マイクとオーディオインターフェースを有する別途の装置（図示せず）で予め行っておくとよい。音声ファイル２の形式はＷＡＶ形式、ＡＩＦＦ形式、ｍｐ３形式など、特に種類は限定されない。

【００２４】

音声認識部１１のために用いるソフトウェアとしては、メタデータ付与装置１全体における話者識別フラグＦＳが、音声認識により話者識別を取得する設定（以下「ＦＳ＝１」）となっている場合は、話者識別結果の出力が可能な音声認識エンジンを採用する。例えば、ＩＢＭ社が提供する音声認識エンジンがこれにあたる。一方、メタデータ付与装置１全体における話者識別フラグＦＳが、文字認識により話者識別を取得する設定（以下「ＦＳ＝２」）か又は話者識別を取得しない設定（以下「ＦＳ＝３」）である場合には、特に種類を限定されず、Ｇｏｏｇｌｅ社、Ｍｉｃｒｏｓｏｆｔ社、ＩＢＭ社などが提供する音声認識エンジンを適宜選択して用いることができる。ただし、単にテキストを生成するだけでなく、音声ファイル２における時刻データ付のテキストを生じるものである必要がある。

【００２５】

なお、メタデータ付与装置１が音声認識部１１を有さない場合は、音声認識部１１と同様の機能を有する別の装置（図示せず）が音声ファイル２からトークデータ４を生成する（図１中Ｏ１）。その別の装置から出力されたトークデータ４を、記憶媒体やネットワークを介してメタデータ付与装置１に入力する。処理としては、図１中Ｏ１の代わりに後述するテキスト照合部１３への入力とする。

【００２６】

トークデータ４は、音声認識テキストとそのテキストに該当する音声が話された時刻についての時刻データとを有する。この時刻データは標準時基準での時分秒まで含めたものでもよいし、音声ファイル２の開始の時点、または音声ファイル２の開始の時点に所定の値を足した時点からの経過時間であってもよい。これは例えば番組開始からそのセリフの出現時刻までの経過時間にあたる。例として図３に示すメタデータ付与テキストに付与されているのは、音声ファイル２の開始の時点からの経過時間である。

【００２７】

また、トークデータ４は、ＦＳ＝１である場合には、音声認識部１１が判別した話者の識別フラグを有する。音声ファイル２に複数の人間の声が含まれている場合、どの人間が喋った内容であるかを識別するものである。ただし、一人の話者の声のみが録音されている場合でも、当該話者の声である識別フラグが付されている形式としてよい。

【００２８】

このようなトークデータ４のフォーマットの例を図４に示す。行番号Ｎｖｏｉｃｅごとに区切られた音声認識テキストＴｅｘｔｖｏｉｃｅが羅列される。音声認識テキストＴｅｘｔｖｏｉｃｅは文節ごとではなく、ある程度の長さを持った文章の塊である。区切られる箇所は音声認識エンジンの設定により、特に限定されない。例えばセリフなどが所定の時間途切れた無声部分で区切られることが挙げられる。また、その文章の塊の開始時間Ｔｖｏｉｃｅ＿ｓｔａｒｔと終了時間Ｔｖｏｉｃｅ＿ｓｔｏｐとが各行に記録されている。時刻のフォーマットは、その音声ファイル２の開始時からの経過時間でもよいし、標準時基準でもよい。さらにＦＳ＝１であるトークデータ４では、各行の音声認識テキストＴｅ

ｘｔｖｏｉｃｅの話者を識別する話者識別Ｓｖｏｉｃｅを有する。話者識別Ｓｖｏｉｃｅのフォーマットは自動的に付される番号などの識別符号であってもよいし、音声認識の際に各話者について入力した名前のテキスト情報であってもよい。なお、ＦＳ＝２，３である場合は、話者識別Ｓｖｏｉｃｅが無いフォーマットとなる。

【００２９】

この実施形態にかかるメタデータ付与装置１は、文字認識部１２を有する。台本３が画像ファイルである場合に、画像ファイルを読み込んで文字認識（ＯＣＲ）により台本のテキストである台本テキスト５を出力する。文字認識を行う文字認識エンジンとしては、Ｇｏｏｇｌｅ社、Ｍｉｃｒｏｓｏｆｔ社、ＩＢＭ社など一般的に提供されているエンジンを適宜用いることができる。また、ＦＳ＝２のとき、画像ファイルにかかれている各セリフの話者を識別した話者識別を含めて出力される。一方、ＦＳ＝１，３のときは話者識別を含まないで出力される。

【００３０】

台本３が紙の状態である場合には、カメラやスキャナなどの光学機器を用いて画像ファイルにしてから上記の文字認識部１２に用いる。

【００３１】

このような台本テキスト５のフォーマットの例を図５に示す。ここではＦＳ＝２の場合を示す。例えば台本３を文字認識する場合は、元の台本３における各行の台本テキストが、それぞれの行番号Ｎｏｃｒを付されて台本テキストＴｅｘｔｏｃｒの各行となる。台本に書かれてある話者の欄も同様に文字認識して読み取り、各行のセリフの話者を識別できるように話者識別Ｓｏｃｒとして出力する。ここで話者識別Ｓｏｃｒはテキスト情報のままであってもよいし、その台本テキストに登場する話者をまとめて区別した識別情報であってもよい。

【００３２】

メタデータ付与装置１は、台本３がテキストデータである場合には、文字認識部１２を有していなくてもよい。セリフが識別できるテキストであれば、そのまま後述するテキスト照合部１３に台本テキスト５として入力してもよい。それぞれのセリフの話者が記録されたテーブル形式やＸＭＬ形式その他の形式のテキストであれば、ＦＳ＝２の条件の台本テキストとしてそのまま用いることができる。そうでない場合には、例えば上記図５に示すようなフォーマットに整形した上でテキスト照合部１３に入力する。

【００３３】

メタデータ付与装置１は、上記の音声認識テキストを含むトークデータ４と上記の台本テキスト５とを入力（Ｏ１，Ｏ２）として、台本テキストにタイムスタンプを付与したメタデータを出力（Ｏ３）するテキスト照合部１３を有する。テキスト照合部１３における出力フォーマットの例を図６に示す。図６（ａ）はＦＳ＝１のときの出力フォーマット例であり、図６（ｂ）はＦＳ＝２のときの出力フォーマット例であり、図６（ｃ）はＦＳ＝３のときの出力フォーマット例である。各行の台本テキスト由来の台本テキストＴｅｘｔｏｃｒに、音声認識テキスト由来のタイムスタンプＴｏｕｔが付されるものとなる。ＦＳ＝１とＦＳ＝２では話者識別Ｓｖｏｉｃｅ又は話者識別Ｓｏｃｒを有するが、その情報の参照元がＦＳ＝１では音声認識テキストであり、ＦＳ＝２では台本テキストとなる。また、ＦＳ＝３では話者識別を有さない。

【００３４】

このテキスト照合部１３における処理フローの例を図７に示す。記載のように、形態素分割ステップＳ０１、探索範囲設定ステップＳ０２、メタデータ付与ステップＳ０３、整合性確認ステップＳ０４、補正ステップＳ０５を行う。以降のステップにおいて使用する変数は次の通りである。

＜音声認識側＞

・ｎｖｏｉｃｅ：処理中行の番号。

・Ｎｖｏｉｃｅ：ｎｖｏｉｃｅに与えられたラベル名。

・Ｌｖｏｉｃｅ：全行数。

・Ｍｖｏｉｃｅ：処理中行の形態素。Ｍｖｏｉｃｅ（ｎｖｏｉｃｅ，ｉ）としてアクセスする。
・ＮＭｖｏｉｃｅ：処理中行の形態素数つまりｉの最大値。
＜ＯＣＲ側＞
・ｎｏｃｒ：処理中行の番号。
・Ｎｏｃｒ：ｎｏｃｒに与えられたラベル名。
・Ｌｏｃｒ：全行数。
・Ｍｏｃｒ：処理中行の形態素。Ｍｏｃｒ（ｎｏｃｒ，ｉ）としてアクセスする。
・ＮＭｏｃｒ：処理中行の形態素数つまりｉの最大値。

【００３５】

まず、上記の音声認識テキストの入力（Ｏ１）と上記の台本テキストの入力に対して、それぞれを形態素分割する形態素分割手段を実行する形態素分割ステップ（Ｓ０１）を行う。形態素分割するエンジンとしては、例えば、ｍｅｃａｂ、Ｊｕｍａｎ等のツールが挙げられる。それぞれのテキストを形態素ごとに区切ることができるのであれば、特に種類は問わない。

【００３６】

また、形態素分割ステップ（Ｓ０１）では、形態素に分割したテキストについて、さらに仮名化しておくと好ましい。仮名としてはひらがなでもカタカナでもよい。仮名化ツールとしては例えばひらがな化ツールであるｋａｋａｓｉが挙げられるが、特に限定されない。

【００３７】

上記の音声認識テキストを形態素分割した出力結果（Ｏ１１）のフォーマット例を図８に示す。ここではＦＳ＝１の例を示す。ＦＳ＝２，３の時は、話者識別Ｓｖｏｉｃｅの項が存在しないフォーマットとなる。元の音声認識テキストに含まれていた音声認識テキストＴｅｘｔｖｏｉｃｅのそれぞれの行について、形態素分割結果として出力されるＭｖｏｉｃｅ（ｎｖｏｉｃｅ、ｉ）と、音声認識テキストＴｅｘｔｖｏｉｃｅ（ｎｖｏｉｃｅ）の形態素数ＮＭｖｏｉｃｅ（ｎｖｏｉｃｅ）の項が付与される。Ｍｖｏｉｃｅ（ｎｖｏｉｃｅ、ｉ）のｉは１以上ＮＭｖｏｉｃｅ（ｎｖｏｉｃｅ）以下の整数となる。例えば、元の音声認識テキストＴｅｘｔｖｏｉｃｅ（Ｘ１）が「音をそれぞれに分割して」であった場合、形態素分割してひらがな化したものは「おと、を、それぞれ、に、ぶんかつ、して」となる。このときＮＭｖｏｉｃｅ（Ｘ１）の値は形態素数である「６」であり、Ｍｖｏｉｃｅ（Ｘ１，１）が「おと」、Ｍｖｏｉｃｅ（Ｘ１，２）が「を」、Ｍｖｏｉｃｅ（Ｘ１，３）が「それぞれ」、Ｍｖｏｉｃｅ（Ｘ１，４）が「に」、Ｍｖｏｉｃｅ（Ｘ１，５）が「ぶんかつ」、Ｍｖｏｉｃｅ（Ｘ１，６）が「して」となる。この出力結果を、メタデータ付与ステップＳ０３で用いる。

【００３８】

一方、上記の台本テキストを形態素分割した出力結果（Ｏ２１）のフォーマット例を図９に示す。ここではＦＳ＝２の例を示す。ＦＳ＝１，３の時は、話者識別Ｓｏｃｒの項が存在しないフォーマットとなる。元の台本テキストに含まれていた台本テキストＴｅｘｔｏｃｒのそれぞれの行について、形態素分割結果として出力されるＭｏｃｒ（ｎｏｃｒ，ｉ）と、台本テキスト（ｎｏｃｒ）の形態素数ＮＭｏｃｒ（ｎｏｃｒ）の項が付与される。ＮＭｏｃｒ（ｎｏｃｒ，ｉ）のｉは１以上ＮＭｏｃｒ（ｎｏｃｒ）以下の整数となる。台本テキストの形態素分割結果Ｍｏｃｒの形式は、上記の音声認識テキストの形態素分割結果Ｍｖｏｉｃｅと同様となる。

【００３９】

この台本テキストを形態素分割した出力結果（Ｏ２１）の各行に対して、０行目から最終行まで順次（図７中Ｂ１におけるｎｏｃｒ＜ＬｏｃｒのＹｅｓ／Ｎｏ分岐による）、音声認識テキストを形態素分割した出力結果（Ｏ１１）と照合してメタデータ付与ステップ（Ｓ０３）を行う。ただし、番組が長い場合に、両方のテーブルの全域について照合するのは時間がかかりすぎる場合がある。また、番組が短くてもテーブル全体に対して照合を

行うと処理負荷が無駄に大きくなる。このため、番組が長い場合や、処理時間を短縮したい場合は、台本テキストの各行に対応する可能性が高く照合のために捜索する範囲を音声認識ファイルの一部に絞り込むように設定する探索範囲設定手段を実行する探索範囲設定ステップ（Ｓ０２）を間に挟むと好ましい。

【００４０】

この探索範囲設定ステップとしては、例えば番組を前半と後半とに分けて、台本テキストの前半に該当する台本テキストＴｅｘｔｏｃｒに対応するテキストを捜索する箇所は、音声認識テキストの前半のみに絞る、という方法が挙げられる。前半と後半とは実時間で分割してもよいが、行番号の前半と後半とで分割してもよい。ただし、前半と後半とを分けるタイミングは音声認識テキストと台本テキストとのどちらも共通させておくとよい。または、タイミングを合わせて前半と後半とを一点で分割するのではなく、前半として捜索する箇所と後半として捜索する箇所との一部が重複するようにしてもよい。例えば、番組の前半にセリフが多く後半にセリフが少ない場合に台本テキストでは後半に入っていても時間経過上は前半のままというケースが想定され、またその逆も想定される。このため、半分の１．ｘ倍（１．０１倍～１．５倍程度）の範囲を捜索する箇所として、適宜倍率を選択できるようにしてもよいし、音声認識や文字認識の総テキスト量などから自動的に倍率を設定するようにしてもよい。

【００４１】

処理フローの例を挙げる。番組の時間長をＴｐｒｏｇｒａｍとする。台本テキストＴｅｘｔｏｃｒの行番号ｎｏｃｒが、ｎｏｃｒ＜Ｌｏｃｒ／２のときすなわち行番号上の前半部分のとき、探索範囲Ｒ（ｎｏｃｒ）は、Ｔｖｏｉｃｅ＿ｓｔａｒｔ（ｎｖｏｉｃｅ）＜（１．ｘ／２）×Ｔｐｒｏｇｒａｍとなるｎｖｏｉｃｅの最小と最大を探索範囲の開始と終了としてＲ（ｎｏｃｒ）に設定する。ｎｏｃｒ＞Ｌｏｃｒ／２のときすなわち台本テキストＴｅｘｔｏｃｒの行番号ｎｏｃｒが行番号上の後半部分のとき、探索範囲Ｒ（ｎｏｃｒ）は、Ｔｖｏｉｃｅ＿ｓｔａｒｔ（ｎｖｏｉｃｅ）≧（（１－０．ｘ）／２）×Ｔｐｒｏｇｒａｍとなるｎｖｏｉｃｅの最小と最大を探索範囲の開始と終了としてＲ（ｎｏｃｒ）に設定する。

【００４２】

上記はあくまで探索範囲設定ステップＳ０２の一例である。上記例では行数を元に前半後半で２分割しているが、例えば文字数を元に２分割してもよい。また、台本の中の分けられたセクションごとにわけてもよいし、数十秒単位にまで細かく分割してもよい。また、一旦探索して照合することができた台本データの末尾を記憶しておき、その箇所から例えば１００～３００文字程度のｎ文字後までを次の探索範囲とすれば、探索範囲を最小限に絞り込んで処理速度を速めることができる。この場合、その探索範囲で見つからなければ、次のｎ文字後までを次の探索範囲として同様に探索する。また、探索して照合が既にされた台本テキスト部分は、それ以降の探索範囲から除外すると、探索範囲をさらに好適に絞り込んで処理速度を速めることができる。細かく分割するほど照合の負荷は小さくなり、本来の箇所と異なる部分に照合させてしまうエラーは発生しにくくなる。一方で、単純分割でない場合には、各セクションが映像のどの部分であるかを対応させる必要があり、細かく分割するほどその対応させる処理のためにかえって処理負荷が増加する場合がある。単純に行数や時間で分割する場合は、対応関係を一致させる分の処理は容易になる。

【００４３】

次に、形態素分割した音声認識テキストの入力（Ｏ１１）と形態素分割した文字テキストの入力（Ｏ２１）の入力に対して、文字テキストにタイムスタンプを含むメタデータを付与して出力させるメタデータ付与手段を実行するメタデータ付与ステップ（Ｓ０３）を行う。メタデータはタイムスタンプだけでなく、話者識別を含んでいてもよい。また、探索範囲設定ステップ（Ｓ０２）を経ている場合には、文字テキストの入力Ｏ２１が、探索範囲Ｒ（ｎｏｃｒ）の指定とともに入力される。

【００４４】

メタデータ付与ステップ（Ｓ０３）の具体的実施形態を図１０に示すフロー例とともに

説明する。まず音声認識テキスト側の第一の処理Ｓ０３１として、音声認識テキストの形態素分割結果Ｍｖｏｉｃｅについて、それぞれの分割された形態素ごとにタイムスタンプＴＭｖｏｉｃｅ、話者識別Ｓｖｏｉｃｅを付与する。この処理は音声認識テキストＴｅｘｔｖｏｉｃｅの１行ごとに行う。処理対象の音声認識テキストＴｅｘｔｖｏｉｃｅのイメージ各変数は図１１の通り定義する。ここでの内容は入力される図８に示すデータに対応する。すなわち、それぞれの音声認識テキストの形態素分割結果Ｍｖｏｉｃｅ（ｎｖｏｉｃｅ，ｉ）の分割されたそれぞれの形態素についてタイムスタンプを付与する。処理中の行番号がｎｖｏｉｃｅのとき、各形態素Ｍｖｏｉｃｅ（ｎｖｏｉｃｅ，ｉ）へのタイムスタンプＴＭｖｏｉｃｅ（ｎｖｏｉｃｅ，ｉ）、話者識別Ｓｖｏｉｃｅを付与した出力フォーマット（Ｏ１１１）の例を図１２に示す。タイムスタンプＴＭｖｏｉｃｅ（ｎｖｏｉｃｅ，ｉ）は、音声認識結果の行内での話し方のスピードは一定であると仮定し、下式（１）により求める。

TMvoice(nvoice,i) = Tvoice_start(nvoice) + (Tvoice_stop(nvoice) - Tvoice_start(nvoice))*(i-1)/NMvoice(nvoice) ……（１）

【００４５】

また、文節の文字数が１の場合でも対応できるようにした対応式として、下記式（２）を用いてタイムスタンプＴＭＶｏｉｃｅ（ｎｖｏｉｃｅ，ｉ）を求めることもできる。

TMvoice(nvoice,i) = Tvoice_start(nvoice) + (Tvoice_stop(nvoice) - Tvoice_start(nvoice))*num(i)/NUM ……（２）

なお、

・ｎｕｍ（ｉ）：Ｍｖｏｉｃｅ（ｎｖｏｉｃｅ，ｉ）の先頭文字について文頭からの文字数。

・ＮＵＭ：Ｍｖｏｉｃｅ（ｎｖｏｉｃｅ，ｉ）・・（ｉ＜＝ＮＭｖｏｉｃｅ（ｎｖｏｉｃｅ））に含まれる文字総数。

である。

【００４６】

次に、音声認識テキスト側の第二の処理Ｓ０３２として、「探索範囲設定ステップ」で設定した探索範囲Ｒ（ｎｏｃｒ）に出現する処理対象の「音声認識テキストの形態素分割結果Ｍｖｏｉｃｅ」を連続的に複数個連結させた連結パターンを生成する。この生成と併せて、各連結パターンのタイムスタンプＴＭｖｏｉｃｅと話者識別Ｓｖｏｉｃｅをまとめて出力する（Ｏ１１２）。その連結パターンの例を図１３に示す。ここでは、「探索範囲設定ステップ」で設定した探索範囲に出現する「音声認識テキストの形態素分割結果Ｍｖｏｉｃｅ」をＡＢＣＸとして例示している。元の音声認識テキストの該当行が「音をそれぞれに分割して」であった場合、連結パターンとしては「音」「を」「それぞれ」「に」「分割」「して」が連結数１のパターンである。「音を」「をそれぞれ」「それぞれに」「に分割」「分割して」が連結数２のパターンである。「音をそれぞれ」「をそれぞれに」「それぞれに分割」「に分割して」が連結数３のパターンである。「音をそれぞれに」「をそれぞれに分割」「それぞれに分割して」が連結数４のパターンである。

【００４７】

一方、台本テキスト側の第一の処理Ｓ０３３として、処理対象の「台本テキストの形態素分割結果Ｍｏｃｒ」を連続的に複数個連結させた連結パターンを生成する。処理対象の台本テキストＴｅｘｔｏｃｒ（ｎｏｃｒ）のイメージを図１４に示す。ここでの内容は入力される図９に示すデータに対応する。ここでは、「台本テキストの形態素分割結果Ｍｏｃｒ」をＡＢＣＤとして例示している。その生成される「台本テキストの形態素分割結果Ｍｏｃｒ」の連結パターンの例を図１５に示す。これがＯ２１１の出力となる。

【００４８】

なお、上記のＳ０３２とＳ０３３では、照合するテキストとして、それぞれ形態素の文字列を格納しているが、格納する情報はテキストから形態素分割した形態素の文字列に限定されない。例えば、それぞれの形態素を分類した品詞の情報などの、形態素そのものに関する情報を追加したり、文字列の代わりにそれらの情報に置き換えた上で照合してもよ

い。例えば、台本テキストで「富士山へ登山した」という文章を形態素分割すると、「富士山（名詞）」＋「へ（助詞）」＋「登山（名詞）」＋「し（助動詞）」＋「た（助動詞）」となる。この例において品詞の情報で照合するとは、形態素ではなく「名詞＋助詞＋名詞＋助動詞＋助動詞」の組み合わせで、音声テキストの複数行から検索し同一を判断する。また、形態素の文字列だけ見ると同一のパターンが複数ある場合は、形態素だけではなく品詞の情報を比較することで更に同一性を判断することで、照合の正確性を向上させることができる。

【００４９】

上記のＳ０３２とＳ０３３とを受けた次の処理Ｓ０３４として、Ｓ０３２の出力（Ｏ１１２）と、Ｓ０３３の出力（Ｏ２１１）とを照合する。すなわち、これらは音声認識テキストを形態素分割した結果を連続的に複数個連結させた連結パターンと、台本テキストを形態素分割した結果を連続的に複数個連結させた連結パターンとを、探索範囲で一致する範囲で照合する。照合できた箇所には、台本テキスト由来の連結パターン（例：図１５）のそれぞれについて、それと照合できた音声認識テキストの連結パターン（例：図１３）が有するタイムスタンプＴＭｖｏｉｃｅを、タイムスタンプＴＭｏｃｒとして付与する。照合できなかった部分については空欄のままとする。またＦＳ＝１の場合、台本テキスト由来のそれぞれの連結パターンに、それと照合できた連結パターンの音声認識テキスト由来の話者識別Ｓｖｏｉｃｅも併せて付与する。このＦＳ＝１の場合の照合させた出力結果（Ｏ２１２）の例を図１６に示す。「ＡＢＣ」までは一致する連結パターンが互いに存在するが、「Ｄ」は台本テキスト由来の形態素分割に現れるものの、音声認識テキスト由来の形態素分割には現れない。このため、「Ｄ」が含まれる連結パターンは照合することができず、タイムスタンプＴＭｏｃｒと話者識別Ｓｖｏｉｃｅが空欄となっている。一方、照合できた連結パターンについては、その連結パターンの冒頭部の開始時刻に対応するタイムスタンプが付される。

【００５０】

上記のＳ０３４を受けた次の処理Ｓ０３５として、処理対象の台本テキストの形態素への最大連結数を付与する。最大連結数とは、その形態素が含まれる連結パターンのうち、照合ができたものの中から連結した形態素の数が最も多くなった数である。上記の図１６の例であると、形態素分割結果「Ａ」「Ｂ」「Ｃ」は、様々に組み合わせた連結パターンのうち、「ＡＢＣ」とした連結パターンが、照合できた中では最も多い個数の形態素が連結されたものである。したがって、これらの形態素分割結果Ｍｏｃｒ（ｎｏｃｒ，ｉ）の最大連結数Ｎｃｏｎｎｅｃｔ（ｉ）としては３を付与する。一方、「Ｄ」を含む連結パターンはいずれも照合できなかった。このため、「Ｄ」の最大連結数Ｎｃｏｎｎｅｃｔ（ｉ）としては０を付与する。このように出力されるフォーマットの例を図１７に示す。このように最大連結数が付されたものが、最大連結数付与結果Ｏ２１３として出力される。

【００５１】

上記のＳ０３５を受けた次の処理Ｓ０３６として、台本テキストを形態素分割した形態素のうち、最大連結数が２以上の形態素に、タイムスタンプを付与する。また、ＦＳ＝１，２の場合は話者識別も付与する。そのフォーマットの例を図１８に示す。さらに、台本テキストの行番号単位（図９参照）で、各行を代表するタイムスタンプＴｏｕｔ（ｎｏｃｒ）と、話者識別Ｓｏｕｔ（ｎｏｃｒ）を設定して出力する。このように出力されるフォーマットの例を図１９に示す。ここで、各行を代表するタイムスタンプＴｏｕｔ（ｎｏｃｒ）は、ＴＭｏｃｒ（ｎｏｃｒ，ｉ）の最小値を設定することや、ｉ＝１の値を設定することが挙げられる。代表として有用な選択手法であれば特にこれらに限定されない。この代表するタイムスタンプは後述する整合性確認手段と補正手段で補正されるため厳密なものではないが、補正が少なくなるほど負荷も小さくなる。また、話者識別Ｓｏｕｔ（ｎｏｃｒ）は、Ｓｖｏｉｃｅ（ｉ）の中で最頻の話者識別を採用することが考えられる。これは、自動的な話者識別が低い確率で誤っていたとしても、最頻の話者識別を採用するようにすることで、一部が誤っていても訂正しやすい。こうして暫定的なタイムスタンプＴｏｕｔと、ＦＳ次第では話者識別Ｓｏｕｔとがメタデータとして付与された台本テキスト（

Ｏ２１４）が出力される。
【００５２】
ここまでがメタデータ付与ステップＳ０３で行われるメタデータ付与手段の実施形態例である。探索範囲を設定している場合（Ｓ０２）、一つの探索範囲についてメタデータの付与を行ったら（Ｏ２１４）、最後の探索範囲に到達するまで（Ｂ１→Ｙｅｓ）、順次次の探索範囲について同様の処理を行う（Ｓ０２，Ｓ０３）。最後の探索範囲に到達したら、又は最初から探索範囲が全体であった場合には、次の整合性確認ステップＳ０４へ移る。
【００５３】
上記のメタデータが付与された台本テキスト（Ｏ２１４→Ｏ２１５）に対して、次の処理により整合性を確認して整合性フラグＦｃを追加する整合性確認手段を実行する整合性確認ステップＳ０４を行う。整合性確認手段を適用する前のフォーマットの例を図２０に示す。各行の内容は図１９と同様の構成であり、それが台本テキストにおける行番号の全てについて揃ったものである。
【００５４】
整合性確認ステップＳ０４としてはまず前段として、この各行に対して、行内のタイムスタンプＴＭｏｃｒが単調増加になっているか否かを判定する。この判定に従い、各行の暫定的なタイムスタンプのうち、問題があるものに対して、第一補正を行う。一つの行を構成する複数の形態素のタイムスタンプが、前の形態素のタイムスタンプに対して次の形態素のタイムスタンプが単調増加になっていない、すなわちタイムスタンプが同一又は減少になっているタイムスタンプとなった行に対して、最大連結数が最大となる形態素のタイムスタンプＴＭｏｃｒ（ｎｏｃｒ，ｉ）のみを残し、それ以外を除外する。さらに、タイムスタンプＴｏｕｔ（ｎｏｃｒ）は、最大連結数が最大となる形態素に付与されたタイムスタンプＴＭｏｃｒ（ｎｏｃｒ，ｉ）のうちの最小値に変更する。これはすなわち、その行のタイムスタンプとして最も信頼性の高いことが見込まれる数値に修正している。このような前段の処理により、予備的な補正がされ、Ｓ０４内の後段の処理の精度を上げる効果がある。Ｓ０４の前段としてこの各行への処理を全行に亘って行った後、次の処理へ移る。
【００５５】
なお、上記の整合性確認ステップＳ０４の前段の処理をこの段階で行うのではなく、メタデータ付与ステップＳ０３の中で行ってもよい。その場合、整合性確認ステップＳ０４としては前段の処理を省略し、次の後段の処理のみを行うようにする。
【００５６】
整合性確認ステップＳ０４の後段としては次に、整合性の確認結果を付与する。具体的には、上記の判定と第一補正を行った後、各行について前後の行のタイムスタンプＴｏｕｔを比較し、行間のタイムスタンプＴｏｕｔが単調増加になっているか否かを判定する。前の行に対して単調増加になっている場合には、整合性が満たされたものとして、その行の整合性フラグＦｃ＝０とする。前の行に対して単調増加になっていない場合には、整合性が満たされなかったものとして、その行の整合性フラグＦｃ＝１とする。この整合性フラグを付したフォーマットの例を図２１に示す。このような整合性確認結果を付与したデータを出力する（Ｏ２１６）。
【００５７】
整合性を確認し、整合性が満たされなかったフラグを付された台本テキスト（Ｏ２１６）に対して、補正手段を実行する補正ステップＳ０５を行う。整合性が満たされなかった行であるＦｃ＝１の行に対して、Ｆｃ＝０である前後の行のタイムスタンプＴｏｕｔから補正タイムスタンプＴｏｕｔを求める。なお、Ｆｃ＝１の行が複数行連続している場合はそれらの複数行をまとめて、Ｆｃ＝０である前後の行から補正タイムスタンプＴｏｕｔを求める。Ｆｃ＝１である行には、求められた補正タイムスタンプＴｏｕｔを付与する。すなわち、Ｆｃ（ｎｏｃｒ－１）＝０、Ｆｃ（ｎｏｃｒ＋ｐ－１）＝１，Ｆｃ（ｎｏｃｒ＋ｐ）＝０（ｐ＞０）の場合には、それらのＦｃ＝１である行の補正タイムスタンプＴｏｕｔを次式（３）により求める。

Tout(nocr+q) = Tout(nocr -1) + (Tout(nocr +p) - Tout(nocr -1))/(p+1) * q (0 = 〈 q 〈 p) ・・（３）

【００５８】

また、上記の補正ステップでは合わせて、話者識別を補正した補正話者識別を付与すると好ましい。補正話者識別Ｓｏｃｒ（ｎｏｃｒ＋ｑ）は、Ｆｃ＝０となる連続するｐ行において最頻の話者識別Ｓｏｕｔに置換する。

【００５９】

この発明にかかるメタデータ付与装置、メタデータ付与方法を用い、以上の補正ステップにより補正された補正タイムスタンプを付与された台本テキストは、音声認識テキストとの照合を連結パターン同士の比較によって行うことで照合の精度を高めて暫定的なタイムスタンプを付与された上で、さらに前後関係を踏まえて補正された補正タイムスタンプに修正されているため、人の判断が入らない機械的な処理ながら、正確性の高いタイムスタンプを有する台本テキストが得られる。これにより、台本のある放送において正確性の高い字幕の表示が自動的に行える。

【００６０】

特に、日本語を音声認識した場合、誤変換ではないが人名や同音異義語など当該番組で適切な漢字に変換できない場合が多い。台本テキストを元にした字幕では人名や同音異義語の変換の誤りが極めて少ないことから、単純な音声認識テキストを用いるよりも、固有名詞の正確性が高くなる。また、完成した字幕について、音声認識テキストと台本テキストとの変換の規則性を学習することで、音声認識テキストを得るための音声認識エンジンの精度を向上させることができる。

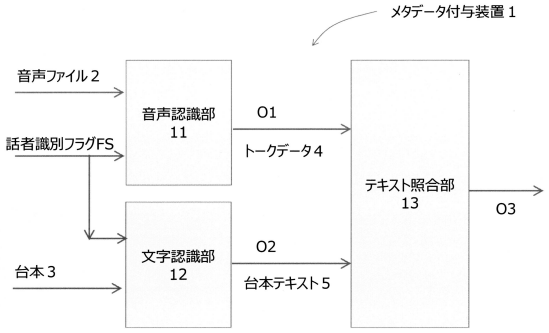
【符号の説明】

【００６１】

１　メタデータ付与装置
２　音声ファイル
３　台本
４　トークデータ
５　台本テキスト
１１　音声認識部
１２　文字認識部
１３　テキスト照合部

【図面】

【図１】



【図２】

0	こんにちは。
1	○月○日、○時になりました。
2	今日は、肌寒いですね。
3	最近衣替えしたところだったのですが
4	朝あまりに寒かったので
5	冬服を出してきました。
6	ただ、この後、
7	お昼頃になると
8	暖かくなるようです。
9	体調管理に気を付けて
10	くださいね。
11	それでは、本日は、
12	こちらの話題から開始です。

【図３】

	開始時刻	
0	3.5	こんにちは。
1	4.7	○月○日、○時になりました。
2	6.8	今日は、肌寒いですね。
3	7.3	最近衣替えしたところだったのですが
4	10.8	朝あまりに寒かったので
5	12.1	冬服を出してきました。
6	14.2	ただ、この後、
7	15.4	お昼頃になると
8	16.6	暖かくなるようです。
9	19.5	体調管理に気を付けて
10	20.7	くださいね。
11	22.2	それでは、本日は、
12	23.9	こちらの話題から開始です。

【図４】

行番号 Nvoice	開始時間 Tvoice_start	終了時間 Tvoice_stop	音声認識テキスト Textvoice	話者識別 Svoice
0	Tvoice_start(0)	Tvoice_stop(0)	Textvoice(0)	Svoice(0)
1				
2				
…				
nvoice	Tvoice_start(nvoice)	Tvoice_stop(nvoice)	Textvoice(nvoice)	Svoice(nvoice)
…				
Lvoice				

【図５】

行番号 Nocr	台本テキスト Textocr	話者識別 Socr
0	Textocr(0)	Socr(0)
1		
2		
・・・		
nocr	Textocr(nocr)	Socr(nocr)
・・・		
Locr		

【図６】

(a)

行番号 Nocr	台本テキスト Textocr	**タイムスタンプ Tout**	**話者識別 Svoice**
0	Textocr(0)	**Tout(0)**	**Svoice (0)**
1			
2			
・・・			
nocr	Textocr(nocr)	**Tout(nocr)**	**Svoice (nocr)**
・・・			
Locr			

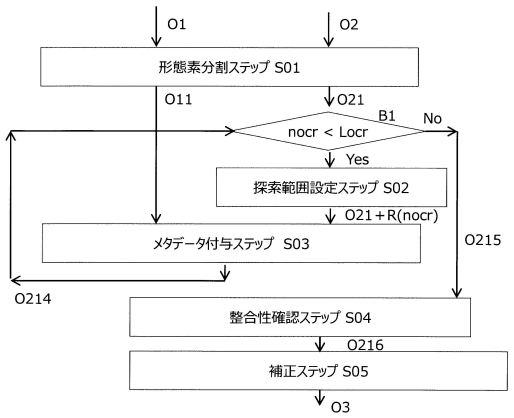
(b)

行番号 Nocr	台本テキスト Textocr	**タイムスタンプ Tout**	話者識別 Socr
0	Textocr(0)	**Tout(0)**	Socr(0)
1			
2			
・・・			
nocr	Textocr(nocr)	**Tout(nocr)**	Socr(nocr)
・・・			
Locr			

(c)

行番号 Nocr	台本テキスト Textocr	**タイムスタンプ Tout**
0	Textocr(0)	**Tout(0)**
1		
2		
・・・		
nocr	Textocr(nocr)	**Tout(nocr)**
・・・		
Locr		

【図７】



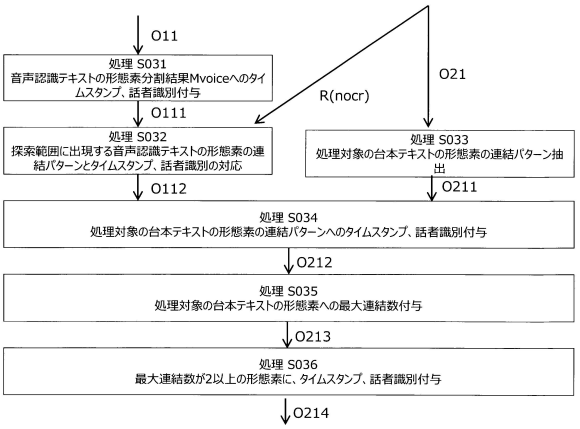
【図８】

行番号 Nvoice	開始時間 Tvoice_start	終了時間 Tvoice_stop	音声認識テキスト Textvoice	話者識別 Svoice	音声認識テキストの形態素分割結果 Mvoice	音声認識テキストの形態素数 NMvoice
0	Tvoice_start(0)	Tvoice_stop(0)	Textvoice(0)	Svoice(0)	Mvoice(0,i) (i<=NMvoice(0))	NMvoice(0)
1						
2						
・・・						
nvoice	Tvoice_start(nvoice)	Tvoice_stop(nvoice)	Textvoice(nvoice)	Svoice(nvoice)	Mvoice(nvoice,i) (i<=NMvoice(nvoice))	NMvoice(nvoice)
・・・						
Lvoice						

【図９】

行番号Nocr	台本テキスト Textocr	話者識別 Socr	台本テキストの形態素分割結果 Mocr	台本テキストの形態素数 NMocr
0	Textocr(0)	Socr(0)	Mocr(0,i) (i<=NMocr(0))	NMocr(0)
1				
2				
…				
nocr	Textocr(nocr)	Socr(nocr)	Mocr(nocr,i) (i<=NMocr(nocr))	NMocr(nocr)
…				
Locr				

【図１０】



【図１１】

行番号 Nvoice	開始時間 Tvoice_start	終了時間 Tvoice_stop	音声認識テキスト Textvoice	話者識別 Svoice	音声認識テキストの形態素分割結果 Mvoice	音声認識テキストの形態素数 NMvoice
nvoice	Tvoice_start(nvoice)	Tvoice_stop(nvoice)	Textvoice(nvoice)	Svoice(nvoice)	Mvoice(nvoice,i) (i<=NMvoice(nvoice))	NMvoice(nvoice)

【図１２】

音声認識テキストの形態素分割結果 Mvoice	タイムスタンプ TMvoice	話者識別 Svoice
Mvoice(nvoice,1)	TMvoice(nvoice,1)	Svoice(nvoice)
…		
Mvoice(nvoice,i)	TMvoice(nvoice,i)	Svoice(nvoice)
…		
Mvoice(nvoice, NMvoice(nvoice))	TMvoice(nvoice, NMvoice(nvoice))	Svoice(nvoice)

【図１３】

「音声認識テキストの形態素分割結果Mvoice」の連結パターン	タイムスタンプ TMvoice	話者識別 Svoice
A TMvoice(nvoice,1)	TA	Svoice(nvoice)
B TMvoice(nvoice,2)	TB	Svoice(nvoice)
C TMvoice(nvoice,3)	TC	Svoice(nvoice)
X TMvoice(nvoice+1,1)	TX	Svoice(nvoice+1)
AB	TA	Svoice(nvoice)
BC	TB	Svoice(nvoice)
CX	TC	Svoice(nvoice)
ABC	TA	Svoice(nvoice)
BCX	TB	Svoice(nvoice)
ABCX	TA	Svoice(nvoice)

【図１４】

行番号Nocr	台本テキスト Textocr	台本テキストの形態素分割結果 Mocr	台本テキストの形態素数 NMocr
nocr	Textocr(nocr)	Mocr(nocr,i) (i<=NMocr(nocr)) =ABCD	NMocr(nocr) =4

【図１５】

連結数	台本テキストの形態素分割結果Mocrの番号i	「台本テキストの形態素分割結果Mocr」の連結パターン
1	1	A
1	2	B
1	3	C
1	4	**D**
2	1	AB
2	2	BC
2	3	CD
3	1	ABC
3	2	BCD
4	1	ABCD

【図１６】

連結数	台本テキストの形態素分割結果Mocrの番号i	「台本テキストの形態素分割結果Mocr」の連結パターン	タイムスタンプ TMocr(nocr,i)	話者識別 Svoice
1	1	A	TA	Svoice(nvoice)
1	2	B	TB	Svoice(nvoice)
1	3	C	TC	Svoice(nvoice)
1	4	**D**		
2	1	AB	TA	Svoice(nvoice)
2	2	BC	TB	Svoice(nvoice)
2	3	CD		
3	1	ABC	TA	Svoice(nvoice)
3	2	BCD		
4	1	ABCD		

【図１７】

台本テキストの形態素分割結果Mocrの番号i	台本テキストの形態素分割結果 Mocr(nocr,i)	最大連結数 Nconnect(i)
1	A	3
2	B	3
3	C	3
4	**D**	0

【図１８】

台本テキストの形態素分割結果Mocrの番号i	台本テキストの形態素分割結果 Mocr(nocr,i)	最大連結数 Nconnect(i)	タイムスタンプ TMocr(nocr,i)	話者識別 Svoice
1	A	3	TA	Svoice(nvoice)
2	B	3	TB	Svoice(nvoice)
3	C	3	TC	Svoice(nvoice)
4	**D**	0		

【図１９】

行番号 Nocr	台本テキスト Textocr	タイムスタンプ Tout	話者識別 Sout	タイムスタンプTMocr(nocr,i)と最大連結数			
				1	2	3	4
nocr	Textocr(nocr)	Tout(nocr) =TA	Sout(nocr) =Svoice(nvoice)	TA、3	TB、3	TC、3	

【図２０】

行番号 Nocr	台本テキスト Textocr	タイムスタンプ Tout	話者識別 Sout	タイムスタンプTMocr(nocr,i)と最大連結数				
				1	2	3	4	…
0	Textocr(0)	Tout(0)	Sout(0)					
…								
nocr	Textocr(nocr)	Tout(nocr)	Sout(nocr)					
…								
Locr								

【図２１】

整合性フラグ Fc	行番号 Nocr	台本テキスト Textocr	タイムスタンプ Tout	話者識別 Sout	タイムスタンプTMocr(nocr,i)と最大連結数				
					1	2	3	4	…
	0	Textocr(0)	Tout(0)	Sout(0)					
	…								
	nocr	Textocr(nocr)	Tout(nocr)	Sout(nocr)					
	…								
	Locr								

フロントページの続き

弁理士　地代 信幸
(72)発明者　駒井 友香
大阪府大阪市中央区馬場町３番１５号 西日本電信電話株式会社内
(72)発明者　川嶋 喜美子
大阪府大阪市中央区馬場町３番１５号 西日本電信電話株式会社内
(72)発明者　安楽 沙希
大阪府大阪市中央区馬場町３番１５号 西日本電信電話株式会社内
(72)発明者　洞井 晋一
大阪府大阪市中央区馬場町３番１５号 西日本電信電話株式会社内
(72)発明者　谷知 紀英
大阪府大阪市中央区城見１丁目３番５０号 讀賣テレビ放送株式会社内
(72)発明者　松田 慎一郎
大阪府大阪市中央区城見１丁目３番５０号 讀賣テレビ放送株式会社内
(72)発明者　浅井 拓登
大阪府大阪市中央区城見１丁目３番５０号 讀賣テレビ放送株式会社内
審査官　菊池 智紀

(56)参考文献
特開２００５－２５８１９８（ＪＰ，Ａ）
特開２０００－２７０２６３（ＪＰ，Ａ）
特開２００３－１８６４９１（ＪＰ，Ａ）
特開２００３－２４４５３９（ＪＰ，Ａ）
特開２００９－１８２８５９（ＪＰ，Ａ）
特開２０１０－２３３０１９（ＪＰ，Ａ）
谷村正剛、外1名，テレビドラマのシナリオと音声トラックの自動対応付け，情報処理学会研究報告，日本，社団法人情報処理学会，1999年05月28日，第99巻、第49号，第23-29ページ
丸山一郎、外3名，ワードスポッティングと動的計画法を用いたテレビ番組に対する字幕提示タイミング検出法，電子情報通信学会論文誌，日本，社団法人電子情報通信学会，2002年02月，第85巻、第2号，第184-192ページ
西沢容子、外1名，字幕表示のための音声とテキストの自動対応付け手法とその評価，電子情報通信学会技術研究報告，日本，社団法人電子情報通信学会，2004年01月30日，第103巻、第633号，第7-12ページ

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　１５／００－１５／３４
Ｈ０４Ｎ　５／２７８
Ｇ０６Ｆ　１６／００－１６／９５８