(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2015-158582
(P2015-158582A)
(43)公開日　平成27年9月3日(2015.9.3)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 15/10	(2006.01)	G10L	15/10	500Z
G10L 15/00	(2013.01)	G10L	15/00	200G
G10L 25/57	(2013.01)	G10L	25/57	

審査請求 未請求　請求項の数 5　OL　（全 20 頁）

(21)出願番号　特願2014-33024(P2014-33024)
(22)出願日　平成26年2月24日(2014.2.24)

(71)出願人　000004352
日本放送協会
東京都渋谷区神南２丁目２番１号
(74)代理人　100064908
弁理士　志賀　正武
(74)代理人　100108578
弁理士　高橋　詔男
(72)発明者　小林　彰夫
東京都世田谷区砧一丁目１０番１１号　日本放送協会放送技術研究所内

(54)【発明の名称】音声認識装置、及びプログラム

(57)【要約】
【課題】音響イベントの情報を付加した字幕を制作する。
【解決手段】音声認識装置１の音声認識部１３は、音声データを音声認識し、発話内容を示す文字列のデータを出力する。音響イベント認識部１５は、音声認識されたものと同じ音声データから得られた音響特徴量に基づいて音響イベントの事後確率を計算し、計算された事後確率に基づいて検出した音響イベントを表す文字列のデータを出力する。認識結果修正部１６は、発話内容の文字列のデータと音響イベントを表す文字列のデータとを修正端末５に表示させ、表示させた中から指定された発話内容の文字列における注釈挿入位置と、表示させた中から選択された音響イベントを表す文字列とを示す注釈挿入指示を受信し、受信した注釈挿入指示に従って発話内容を示す文字列のデータに音響イベントを表す文字列のデータを挿入して注釈付き字幕データを生成する。
【選択図】図２

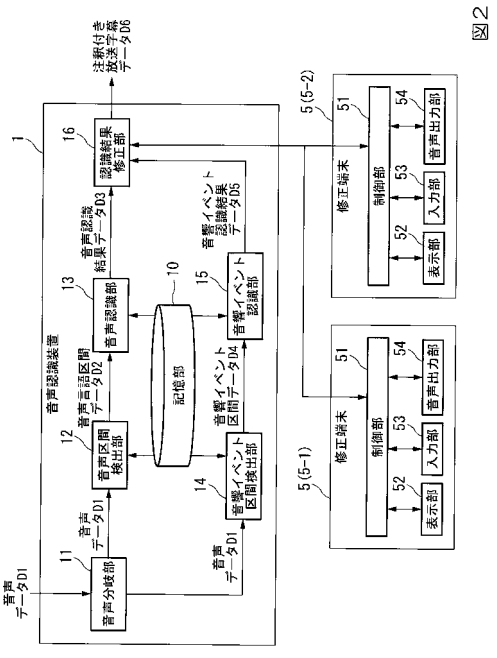

図２

【特許請求の範囲】
【請求項１】
音声データを音声認識し、音声認識結果の発話内容を示す文字列のデータを出力する音声認識部と、
前記音声データから得られた音響特徴量に基づいて音響イベントの事後確率を計算し、計算された前記事後確率に基づいて検出した音響イベントを表す文字列のデータを出力する音響イベント認識部と、
前記音声認識部が出力した前記発話内容の文字列のデータと、前記音響イベント認識部が出力した前記音響イベントを表す文字列のデータとを修正端末に表示させ、表示させた中から指定された前記発話内容の文字列における注釈挿入位置と、表示させた中から選択された前記音響イベントを表す文字列とを示す注釈挿入指示を前記修正端末から受信し、受信した前記注釈挿入指示に従って前記発話内容を示す文字列のデータに前記音響イベントを表す文字列のデータを挿入した注釈付き字幕データを生成する認識結果修正部と、
を備えることを特徴とする音声認識装置。
【請求項２】
前記音声データをフレームに分割し、各フレームの音響特徴量と、無音、音響イベント、及び音声言語それぞれの音響特徴量とを照合して音響イベントを含んだ区間を検出する音響イベント区間検出部を備え、
前記音響イベント認識部は、前記音響イベント区間検出部が検出した前記区間の前記音声データから得られた音響特徴量に基づいて音響イベントの事後確率を計算し、計算された前記事後確率に基づいて検出した音響イベントを表す文字列のデータを出力する、
ことを特徴とする請求項１に記載の音声認識装置。
【請求項３】
前記音響イベント認識部は、前記音声データを分割した時刻順のフレームそれぞれの音響特徴量を並べて畳み込みニューラルネットワークに入力して音響イベントの事後確率を算出し、
前記畳み込みニューラルネットワークは、入力層、隠れ層、プーリング層、及び出力層を有し、
前記入力層は、時刻順に並べた前記フレームそれぞれの音響特徴量を入力とし、
前記隠れ層の各ユニットは、所定フレーム数分のシフトを保ちながら前記入力層の所定数のフレームと結合しており、結合している前記入力層のフレームの音響特徴量を畳み込み演算した結果を示し、
前記プーリング層の各ユニットは、当該プーリング層のユニット数に応じた数の前記隠れ層のユニットと結合しており、結合している前記隠れ層のユニットのうち最大値が伝搬され、
前記出力層の各ユニットは、異なる種類の音響イベントに対応しており、前記プーリング層の全てのユニットと、対応する前記音響イベントの事後確率を算出するためのそれぞれの重みにより結合している、
ことを特徴とする請求項１または請求項２のいずれか１項に記載の音声認識装置。
【請求項４】
前記音響特徴量は、時間周波数領域の特徴量である、
ことを特徴とする請求項１から請求項３のいずれか１項に記載の音声認識装置。
【請求項５】
コンピュータを、
音声データを音声認識し、音声認識結果の発話内容を示す文字列のデータを出力する音声認識手段と、
前記音声データから得られた音響特徴量に基づいて音響イベントの事後確率を計算し、計算された前記事後確率に基づいて検出した音響イベントを表す文字列のデータを出力する音響イベント認識手段と、
前記音声認識手段が出力した前記発話内容の文字列のデータと、前記音響イベント認識

手段が出力した前記音響イベントを表す文字列のデータとを修正端末に表示させ、表示させた中から指定された前記発話内容の文字列における注釈挿入位置と、表示させた中から選択された前記音響イベントを表す文字列とを示す注釈挿入指示を前記修正端末から受信し、受信した前記注釈挿入指示に従って前記発話内容を示す文字列のデータに前記音響イベントを表す文字列のデータを挿入した注釈付き字幕データを生成する認識結果修正手段と、

を具備する音声認識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】

本発明は、音声認識装置、及びプログラムに関する。

【背景技術】

【０００２】

生放送番組の字幕制作に音声認識を利用する技術が実用化されている。放送字幕は、放送番組の音声を音声認識した結果を人手により修正して作成される（例えば、特許文献１参照）。

【先行技術文献】

【特許文献】

【０００３】

【特許文献１】特開２００４－２２６９１０号公報

【発明の概要】

【発明が解決しようとする課題】

【０００４】

放送番組の音声認識は、主に聴覚障碍者や高齢者への情報補償を目的としている。このときの音声認識の対象は、放送番組における音声言語の音声のみである。しかし、多くの放送番組の音声は、音声言語だけから構成されている訳ではない。例えば、番組の演出上の要請から、非言語的な音声（例えば、笑い声）や、拍手、背景音楽などの音響イベントが付加されている。音響イベントは、放送番組のシーンを補足的に説明したり、場面の転換を知らせたりするなど、音声言語同様、情報伝達において重要な役割を担っていると考えられる。このとこから、音響イベントは、視聴者が番組を理解する際に欠かせない要素の一つといえる。

【０００５】

ところが、現在の音声認識による字幕制作では、音響イベントは考慮されておらず、番組理解のための情報が視聴者に十分伝えられていないことがある。音響イベントの持つ情報が字幕に反映されれば、伝達する字幕に彩りやアクセント、あるいはニュアンスといった補足的な情報を付加することとなり、視聴者の番組理解に大いに貢献するものと考えられる。そのためには、音響イベントの情報を付加した字幕制作することが求められる。

【０００６】

本発明は、このような事情を考慮してなされたもので、音響イベントの情報を付加した字幕を制作することができる音声認識装置、及びプログラムを提供する。

【課題を解決するための手段】

【０００７】

本発明の一態様は、音声データを音声認識し、音声認識結果の発話内容を示す文字列のデータを出力する音声認識部と、前記音声データから得られた音響特徴量に基づいて音響イベントの事後確率を計算し、計算された前記事後確率に基づいて検出した音響イベントを表す文字列のデータを出力する音響イベント認識部と、前記音声認識部が出力した前記発話内容の文字列のデータと、前記音響イベント認識部が出力した前記音響イベントを表す文字列のデータとを修正端末に表示させ、表示させた中から指定された前記発話内容の文字列における注釈挿入位置と、表示させた中から選択された前記音響イベントを表す文字列とを示す注釈挿入指示を前記修正端末から受信し、受信した前記注釈挿入指示に従っ

て前記発話内容を示す文字列のデータに前記音響イベントを表す文字列のデータを挿入した注釈付き字幕データを生成する認識結果修正部と、を備えることを特徴とする音声認識装置である。

この発明によれば、音声認識装置は、音声データを音声認識して得た発話内容を示す文字列と、当該音声データについて検出された音響イベントを表す文字列とを修正端末に表示させる。音声認識装置は、修正者が修正端末において指定した発話内容の文字列における注釈挿入位置と、挿入する注釈として選択した音響イベントを表わす文字列とに従って、発話内容に音響イベントを表す文字列を挿入して注釈付き字幕を生成する。

これにより、音声認識装置は、修正者が修正端末の表示を見ながら、注釈を挿入したい発話内容の位置と、注釈として挿入したい音響イベントを表す文字列を選択する簡易な操作によって、音響イベントの情報を付加した字幕を生成することができる。

【０００８】

本発明の一態様は、上述する音声認識装置であって、前記音声データをフレームに分割し、各フレームの音響特徴量と、無音、音響イベント、及び音声言語それぞれの音響特徴量とを照合して音響イベントを含んだ区間を検出する音響イベント区間検出部を備え、前記音響イベント認識部は、前記音響イベント区間検出部が検出した前記区間の前記音声データから得られた音響特徴量に基づいて音響イベントの事後確率を計算し、計算された前記事後確率に基づいて検出した音響イベントを表す文字列のデータを出力する、ことを特徴とする。

この発明によれば、音声認識装置は、音声データから音響イベントを含んだ区間を検出し、検出した区間の音声データを対象に音響イベント認識を行う。

これにより、音声認識装置は、音響イベントが含まれている区間のみを音響イベント認識の対象とするため、音響イベント認識の精度を良くすることができる。

【０００９】

本発明の一態様は、上述する音声認識装置であって、前記音響イベント認識部は、前記音声データを分割した時刻順のフレームそれぞれの音響特徴量を並べて畳み込みニューラルネットワークに入力して音響イベントの事後確率を算出し、前記畳み込みニューラルネットワークは、入力層、隠れ層、プーリング層、及び出力層を有し、前記入力層は、時刻順に並べた前記フレームそれぞれの音響特徴量を入力とし、前記隠れ層の各ユニットは、所定フレーム数分のシフトを保ちながら前記入力層の所定数のフレームと結合しており、結合している前記入力層のフレームの音響特徴量を畳み込み演算した結果を示し、前記プーリング層の各ユニットは、当該プーリング層のユニット数に応じた数の前記隠れ層のユニットと結合しており、結合している前記隠れ層のユニットのうち最大値が伝搬され、前記出力層の各ユニットは、異なる種類の音響イベントに対応しており、前記プーリング層の全てのユニットと、対応する前記音響イベントの事後確率を算出するためのそれぞれの重みにより結合している、ことを特徴とする。

この発明によれば、音声認識装置は、音声データを音響イベント認識における音響特徴量の処理単位であるフレームに分割し、分割した各フレームの音響特徴量を、対応するフレームの時刻順に並べて畳み込みニューラルネットワークに入力することにより、各音響イベントの事後確率を算出する。

これにより、音声認識装置は、音声データから得られた各フレームの音響特徴量を用いて、各音響イベントの事後確率を得ることができる。

【００１０】

本発明の一態様は、上述する音声認識装置であって、前記音響特徴量は、時間周波数領域の特徴量である、ことを特徴とする。

この発明によれば、音声認識装置は、音声データの時間周波数領域の特徴量を用いて音響イベントを認識する。

これにより、音声認識装置は、周波数領域の特徴量を所定時間分以上連結して音響イベントを認識することができるため、音響イベントの認識の精度を良くすることができる。

【００１１】

本発明の一態様は、コンピュータを、音声データを音声認識し、音声認識結果の発話内容を示す文字列のデータを出力する音声認識手段と、前記音声データから得られた音響特徴量に基づいて音響イベントの事後確率を計算し、計算された前記事後確率に基づいて検出した音響イベントを表す文字列のデータを出力する音響イベント認識手段と、前記音声認識手段が出力した前記発話内容の文字列のデータと、前記音響イベント認識手段が出力した前記音響イベントを表す文字列のデータとを修正端末に表示させ、表示させた中から指定された前記発話内容の文字列における注釈挿入位置と、表示させた中から選択された前記音響イベントを表す文字列とを示す注釈挿入指示を前記修正端末から受信し、受信した前記注釈挿入指示に従って前記発話内容を示す文字列のデータに前記音響イベントを表す文字列のデータを挿入した注釈付き字幕データを生成する認識結果修正手段と、を具備する音声認識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】

【００１２】

本発明によれば、音響イベントの情報を付加した字幕を制作することができる。

【図面の簡単な説明】

【００１３】

【図１】本発明の一実施形態による字幕制作手法と、従来の字幕制作手法との比較を示す図である。

【図２】同実施形態による字幕制作システムの構成を示す機能ブロック図である。

【図３】同実施形態による音声認識装置の全体処理フローを示す図である。

【図４】同実施形態による音響イベント区間検出用のＨＭＭを示す図である。

【図５】同実施形態による音響イベント区間検出部の音響イベント区間検出処理フローを示す図である。

【図６】同実施形態による音響イベント認識用のニューラルネットワークを示す図である。

【図７】同実施形態による音響イベント認識部の音響イベント認識処理フローを示す図である。

【図８】同実施形態による修正端末の表示部に表示される修正作業画面を示す図である。

【発明を実施するための形態】

【００１４】

以下、図面を参照しながら本発明の実施形態を詳細に説明する。

字幕制作を目的とした音声認識では、遅延のない認識結果文字列の出力が重要視されている。従来は、視聴者への情報伝達に重要な音声言語のみが音声から文字列へと変換する字幕化の対象であり、音響イベントのような非言語音は字幕化の対象外であった。これは、特に生放送の番組では、音声認識誤りの修正のための時間が十分に取れず、音声言語以外の情報を字幕化することが困難であったためである。

【００１５】

ニュースなどの番組では、音声言語が極めて重要なウェイトを占めており、効果音などの音響イベントはほとんど含まれていない。よって、音声言語のみを字幕化するだけで、必要な情報を視聴者に伝達することが可能である。一方、スポーツ番組や情報番組では、非言語音である笑い声や拍手、歓声などの音響的なイベントがより大きな役割を果たしている。ニュースが事実を伝えることに主眼を置いている一方で、その他の番組は、臨場感を伝えるなどの演出上の要請から、非言語音の重要性が増すことが一因である。演出上重要な存在である音響イベントは、従来の生放送を対象とした字幕制作では、どちらかといえば重要視されてこなかったという背景がある。しかし、聴覚障碍者や高齢者が放送番組をより楽しむ、あるいは、理解するという観点から見た場合、非言語音である音響イベントを字幕として充実させることが求められるのは当然といえる。

【００１６】

図１（ａ）は、従来の字幕制作手法を示す図である。従来の字幕制作手法では、入力音声に含まれるテキスト化可能な音声言語のみを字幕制作の対象としているため、入力音声

から音声言語を含む音声区間を検出し、該当区間を切り出している。次に、切り出した音声区間を音声認識し、認識結果である単語列のテキストデータを出力する。この認識結果には通常認識誤りが含まれているため、人手により認識結果中の誤りを修正し、修正結果を放送字幕として送出する。

この一連の手続きは、音声区間が切り出されるたびに逐次的に行われ、低遅延で字幕制作を行うことができる。

【００１７】

音声認識に基づく従来の字幕制作手法において音響イベントを挿入する場合、非言語音が表す内容を修正者が適宜解釈した上で、キーボード等の入力方法を用いて、音響イベントを表す文字列を注釈として音声認識結果に挿入することが考えられる。しかし、キーボード入力には時間を要するため、修正者が、音声認識結果を修正しながら、さらに追加のキーボード入力作業を行うことは現実的には非常に困難である。

本実施形態の音声認識装置は、このような問題を解決し、音響イベントに関する情報伝達を視聴者に行うための字幕制作を行う。

【００１８】

そこで、本実施形態の音声認識装置は、従来の字幕制作手法と同様の音声認識結果とともに、音響イベントの認識結果を注釈として出力する。ここで「注釈」とは、音声言語に対する付加情報である音響イベントを言語表現としてテキスト（文字列）で表したものである。また、音声言語の音声認識結果に基づく従来の字幕に対して注釈が挿入されたものを「注釈付き字幕」と記載する。

【００１９】

図１（ｂ）は、本実施形態の音声認識装置による字幕制作手法を示す図である。

同図に示すように、本実施形態の音声認識装置による字幕制作手法においては、従来の音声区間検出処理及び音声認識処理に併せて、音響イベント区間検出処理及び音響イベント認識処理を並列で実行する。音響イベント区間検出処理では、入力音声から音響イベントを含む音声区間を検出し、該当区間を切り出す。音響イベント認識処理では、切り出された音響イベント区間の音響イベントを認識し、認識した音響イベントを表す単語列のテキストデータを出力する。音声認識処理と音響イベント認識処理の並列動作により、本実施形態の音声認識装置は、個々の認識処理に対して独立に最適なアルゴリズムを実装することが可能となる。また、音響イベントの認識が不要であれば、音響イベント認識処理の実行プログラムを動作させないように本実施形態の音声認識装置に設定すればよい。これにより、字幕制作者のニーズに合わせた字幕制作手法を選択することも可能である。

【００２０】

そして、本実施形態の音声認識装置による字幕制作手法においては、人手による音声認識結果の修正作業時に音声認識結果と音響イベント認識結果とを統合し、放送する注釈付き字幕である注釈付き放送字幕を制作する。上述のように、本実施形態の音声認識装置が、音声認識処理と音響イベント認識処理を並列に実行した場合、最終的な音声認識結果と、注釈として与えられる音響イベント認識結果とを統合する必要がある。通常は、音声認識結果に対して修正端末において人手による修正が行われる。本実施形態の音声認識装置は、修正端末に表示させた音声認識結果に対して修正者が修正指示を入力する際に、音響イベント認識結果である注釈についても修正端末に表示させ、音声認識結果に挿入するための効率的なインタフェースを有する。このインタフェースにより、キーボード入力による音響イベント文字列作成の省力化を図る。

【００２１】

上記のような音声認識処理と音響イベント認識処理の並列実行、及び、修正作業時の音声認識結果と音響イベント認識結果の統合により、本実施形態の音声認識装置は、従来困難であった、音響イベントに関する注釈を付加した効率的な字幕制作を可能とする。

【００２２】

図２は、本発明の一実施形態による字幕制作システムの構成を示すブロック図であり、本実施形態と関係する機能ブロックのみ抽出して示してある。同図に示すように、字幕制

作システムは、音声認識装置１と修正端末５とを備えて構成される。音声認識装置１と修正端末５とはネットワークを介して接続される。同図においては、字幕制作システムが、２台の修正端末５を備える場合を示しているが、修正端末５を１台のみ備えてもよく、３台以上備えてもよい。２台の修正端末５をそれぞれ、修正端末５－１、５－２とする。

【００２３】

音声認識装置１は、コンピュータ装置により実現される。同図に示すように、音声認識装置１は、記憶部１０、音声分岐部１１、音声区間検出部１２、音声認識部１３、音響イベント区間検出部１４、音響イベント認識部１５、及び認識結果修正部１６を備えて構成される。

【００２４】

記憶部１０は、音声区間検出用の統計的音響モデルと、音声認識用の統計的音響モデル及び統計的言語モデルを格納する。さらに、記憶部１０は、音響イベント区間検出用の統計的音響モデルと、音響イベント認識用のニューラルネットワークを格納する。音声分岐部１１は、音声認識装置１に入力された音声データＤ１を２つに分岐し、音声区間検出部１２と音響イベント区間検出部１４に出力する。

【００２５】

音声区間検出部１２は、記憶部１０に記憶されている音声区間検出用の統計的音響モデルを用いて、音声分岐部１１から入力された音声データＤ１において、テキスト化の対象となる音声言語の音声区間である音声言語区間を検出する。音声区間検出部１２は、検出した音声データＤ１の音声言語区間である音声言語区間データＤ２を音声認識部１３に出力する。音声認識部１３は、記憶部１０に記憶されている音声認識用の統計的音響モデル及び統計的言語モデルを用いて音声言語区間データＤ２を音声認識する。音声認識部１３は、発話内容の音声認識結果を設定した音声認識結果データＤ３を認識結果修正部１６に出力する。

【００２６】

音響イベント区間検出部１４は、記憶部１０に記憶されている音響イベント区間検出用の統計的音響モデルを用いて、音声分岐部１１から入力された音声データＤ１において、音響イベントが含まれる音声区間である音響イベント区間を検出する。音響イベント区間検出部１４は、検出した音声データＤ１の音響イベント区間である音響イベント区間データＤ４を音響イベント認識部１５に出力する。音響イベント認識部１５は、記憶部１０に記憶されている音響イベント認識用のニューラルネットワークを用いて音響イベント区間データＤ４の音響イベントを認識する。音響イベント認識部１５は、音響イベント認識結果を設定した音響イベント認識結果データＤ５を認識結果修正部１６に出力する。音響イベント認識結果は、検出した音響イベントを表すテキスト表現（文字列）である。

【００２７】

認識結果修正部１６は、音声認識部１３から出力された音声認識結果データＤ３と、音響イベント認識部１５から出力された音響イベント認識結果データＤ５を修正端末５へ出力し、表示させる。認識結果修正部１６は、修正端末５から受信した修正指示に基づいて音声認識結果を修正するとともに、修正端末５から受信した注釈挿入指示に基づいて注釈文字列を音声認識結果に挿入し、注釈付き放送字幕データＤ６を生成する。修正指示は、音声認識結果における修正箇所と、その修正箇所における文字の削除、挿入、置換などの修正内容を示す。注釈挿入指示は、音声認識結果における注釈挿入箇所と、その注釈挿入箇所に挿入する注釈文字列を示す。注釈文字列は、修正端末５に表示させた音響イベント認識結果データＤ５の音響イベントのテキスト表現の中から、修正者が選択したものである。認識結果修正部１６は、生成した注釈付き放送字幕データＤ６を出力する。

【００２８】

修正端末５は、例えば、パーソナルコンピュータなどのコンピュータ装置により実現される。修正端末５は、制御部５１、表示部５２、入力部５３、及び音声出力部５４を備えて構成される。表示部５２は、ディスプレイであり、画面を表示する。入力部５３は、キーボードやマウスなどであり、修正者による操作を受ける。本実施形態では、修正端末５

がタッチパネルと、キーボードを備える場合を例に説明する。タッチパネルは、表示部５２と入力部５３を兼ねる。音声出力部５４は、ヘッドホンやスピーカーであり、音声データＤ１の再生音声を出力する。制御部５１は、音声認識装置１から受信した音声認識結果データＤ３と音響イベント認識結果データＤ５を表示部５２に表示させる。また、制御部５１は、入力部５３により修正者が入力した音声認識結果の修正指示や、音声認識結果への注釈挿入指示を音声認識装置１に出力する。さらに、制御部５１は、音声データＤ１の再生音声を音声出力部５４から出力させる。

【００２９】

次に、音声認識装置１の動作について説明する。

まず、音声認識装置１は、音声区間検出用、音響イベント区間検出用それぞれの統計的音響モデルと、音声認識用の統計的音響モデル及び統計的言語モデルと、音響イベント認識用のニューラルネットワークを記憶部１０に格納する。音声区間検出用の統計的音響モデルや、音声認識用の統計的音響モデル及び統計的言語モデルは、従来と同様のものを用いることができる。本実施形態では、音響イベント区間検出用の統計的音響モデルとして、ＨＭＭ（Hidden Markov Model、隠れマルコフモデル）及びＧＭＭ（Gaussian Mixture Model、ガウス混合分布）を用いる。この音響イベント区間検出用のＨＭＭ及びＧＭＭは、音声、音響イベント、及び無音の３つのクラスそれぞれのラベルがつけられた音声データを学習データとして用い、従来技術と同様の学習方法により学習される。なお、音声のラベルは、音声言語の音声データにつけられる。例えば、音響イベントのＧＭＭの場合、混合されるガウス分布のそれぞれが、異なる種類の音響イベントの特徴を表すようにする。また、音響イベント認識用のニューラルネットワークの学習には、各音響イベントのラベルが付けられた音声データを学習データとして用い、従来技術と同様の学習方法により学習される。音響イベント区間検出用のＨＭＭについては図４を用いて、音響イベント認識用のニューラルネットワークについては図６を用いて後述する。

【００３０】

図３は、音声認識装置１の全体処理フローを示す図である。音声認識装置１は、音声データＤ１が入力される度に、同図に示す処理を行う。

音声認識装置１に放送番組の音声データＤ１が入力されると、音声分岐部１１は、入力された音声データＤ１を、音声認識及び音響イベント認識それぞれの入力とするために２つに分岐する。これは、音声言語と音響イベントに重なりがあるためである。音声認識処理と音響イベント認識処理を分割することにより、それぞれ独立に最適な認識アルゴリズムを適用できるようにする。音声分岐部１１は、２つに分岐した音声データＤ１のうち一方を、音声認識の前処理を行う音声区間検出部１２に出力し、もう一方を、音響イベント認識の前処理を行う音響イベント区間検出部１４に出力する（ステップＳ１）。

【００３１】

音声区間検出部１２は、従来技術によって、音声データＤ１においてテキスト化が必要となる音声言語区間を検出して切り出す（ステップＳ２）。この音声言語区間には、背景音などの音響イベントとの重なりが含まれ得る。本実施形態では、特開２００７－２３３１４８号公報や、特開２００７－２３３１４９号公報に記載の技術により、音声区間を検出する。音声区間検出部１２は、検出した音声データＤ１の音声言語区間である音声言語区間データＤ２を音声認識部１３に出力する。

【００３２】

具体的には、音声区間検出部１２は、音声データＤ１が入力される度に、音声データＤ１が示す音声を、所定の時間間隔の１処理単位のフレームである入力フレームに分割する。音声区間検出部１２は、時刻が早い順に選択した所定数の入力フレームそれぞれの音響特徴量を計算する。発話区間検出用の状態遷移ネットワークは、発話開始から発話終了までに、非音声言語、音声言語、無音の３状態を飛越しなく遷移するｌｅｆｔ－ｔｏ－ｒｉｇｈｔ型のＨＭＭである。なお、無音の状態に代えて、非音声言語の状態を用いてもよい。音声区間検出部１２は、記憶部１０から非音声言語、音声言語それぞれの音響モデルを読み出し、読み出したこれらの音響モデルを用いて各入力フレームの音響スコア（対数尤

度）計算を行う。非音声言語の音響モデルは、無音や音響イベントなどのＨＭＭを表す。また、音声言語の音響モデルは、各音素の音素ＨＭＭからなる。音声区間検出部１２は、各入力フレームの状態遷移の記録を記憶しておき、現在の状態から開始状態に向かって状態遷移の記録を遡り、状態遷移ネットワークを用いて処理開始（始端）の入力フレームからの各状態系列の累積の音響スコアを算出する。音声区間検出部１２は、各状態系列の累積の音響スコアのうち最大のものと、始端の音響スコアとの差が閾値より大きい場合、最大の累積の音響スコアが得られた系列において最後に非音声言語の状態であった時刻から所定時間遡った時刻を発話開始時刻とする。

音声区間検出部１２は、さらに発話開始時刻検出後の入力フレームについて、上記と同様に処理開始の入力フレームからの現在の入力フレームまでの各状態系列の累積の音響スコアを算出する。音声区間検出部１２は、各状態系列の中で最大の累積の音響スコアと、各状態系列のうち音声言語から非音声言語の終端に至る状態系列の中で最大の累積の音響スコアとの差が閾値を超えたかを判断する。音声区間検出部１２は、閾値を超えた状態が所定時間経過した場合、その経過した時刻から所定時間遡った時刻を発話終了時刻とする。

音声区間検出部１２は、発話開始時刻から発話終了時刻までの区間の入力フレームをまとめた音声言語区間データＤ２を出力する。

【００３３】

音声認識部１３は、従来技術により、記憶部１０に記憶されている音声認識用の統計的音響モデル及び統計的言語モデルを用いて音声言語区間データＤ２を音声認識する（ステップＳ３）。本実施形態では、音声認識部１３は、統計的音響モデルに、ＨＭＭ、及びＧＭＭを用いる。また、本実施形態では、音声認識部１３は、統計的言語モデルに単語ｎ－ｇｒａｍ言語モデルを用いたマルチパス音声認識により認識結果を得る。この認識結果は、単語を単位とした分かち書きであり、音声認識部１３は、各単語に、当該単語が発話された時刻情報を付与する。音声認識部１３は、音声認識結果を設定した音声認識結果データＤ３を認識結果修正部１６に出力する（ステップＳ４）。

【００３４】

一方、音響イベント区間検出部１４は、音声データＤ１において背景音等の音響イベントを含む音響イベント区間を検出して切り出す（ステップＳ５）。この音響イベント区間には、音声認識によりテキスト化が必要となる部分との重複が含まれ得る。音響イベント区間検出部１４は、音声区間検出部１２と同様のアルゴリズムにより、記憶部１０に記憶されている音響イベント区間検出用のＧＭＭとＨＭＭを用いて音響イベント区間の検出を行う。ただし、音声区間検出部１２が、音声言語の音声区間（音声言語区間）を検出対象としているのに対し、音響イベント区間検出部１４は、非言語音の音声区間を検出対象とする点が異なる。また、発話区間検出用の状態遷移ネットワークに代えて、音響イベント区間検出用のＨＭＭを用いる。

【００３５】

図４は、記憶部１０に記憶されている音響イベント区間検出用のＨＭＭを示す図である。本実施形態では、ＨＭＭの構成を、いわゆるエルゴディックＨＭＭとする。同図に示すように、このエルゴディックＨＭＭは、音声、音響イベント、無音の３クラスの遷移を表現したＨＭＭである。各遷移には、学習により得られた遷移確率が付与されている。

【００３６】

図５は、音響イベント区間検出部１４の音響イベント区間検出処理フローを示す図であり、図３のステップＳ５における詳細な処理を示す。まず、音響イベント区間検出部１４は、音声データＤ１が入力される度に、音声データＤ１を、所定の時間間隔の１処理単位のフレームである入力フレームＤ１１に分割する。

【００３７】

音響イベント区間検出部１４は、まだ処理対象としていない入力フレームＤ１１のうち、時刻が早い順に所定数の入力フレームＤ１１を取得する（ステップＳ５１）。音響イベント区間検出部１４は、取得した各入力フレームＤ１１の音響特徴量を計算する。音響イ

ベント区間検出部１４は、記憶部１０からＨＭＭの各状態である音声、音響イベント、及び無音それぞれのＧＭＭを読み出す。音響イベント区間検出部１４は、読み出したこれらのＧＭＭと各入力フレームＤ１１の音響特徴量とを照合して各入力フレームＤ１１の音響スコア計算を行い、必要があればＨＭＭの状態間の遷移を行う（ステップＳ５２）。音響イベント区間検出部１４は、トレースバックに必要な定められた数の入力フレームを処理していない場合（ステップＳ５３：ＮＯ）、ステップＳ５１に戻って新たな入力フレームＤ１１を取得し、音響スコアの計算を行う。

【００３８】

音響イベント区間検出部１４は、トレースバックに必要な定められた数の入力フレームを処理した場合（ステップＳ５３：ＹＥＳ）、現在の状態に至るまでの状態系列のリストをトレースバックにより求める（ステップＳ５４）。つまり、音響イベント区間検出部１４は、現在の状態から開始状態に向かって状態遷移の記録を遡り、図４に示すエルゴディックＨＭＭを用いて、処理開始の入力フレームＤ１１の状態（開始状態）から現在の状態までの各状態系列の累積の音響スコアを算出する。この際、音響イベント区間検出部１４は、累積の音響スコアが大きい順に系列をソートしておく。

【００３９】

音響イベント区間検出部１４は、トレースバックにより得られたＨＭＭの状態系列から、第１位の系列と第２位の系列を比較する（ステップＳ５５）。音響イベント区間検出部１４は、累積の音響スコアの差が予め定めた閾値以下である場合、区間が確定しないと判断し（ステップＳ５６：ＮＯ）、ステップＳ５１に戻って新たな入力フレームＤ１１に対して音響スコアの計算を行う。音響イベント区間検出部１４は、累積の音響スコアの差が予め定めた閾値を超えたと判断した場合（ステップＳ５６：ＹＥＳ）、第１位の系列を確定区間とする。音響イベント区間検出部１４は、最後に音響イベントの確定区間のフレームをまとめあげたフレーム列を、音響イベント区間データＤ４として出力する（ステップＳ５７）。

【００４０】

図３において、音響イベント認識部１５は、記憶部１０に記憶されている音響イベント認識用のニューラルネットワークを用いて、音響イベント区間検出部１４において得られた音響イベント区間データＤ４から音響イベントを認識する（ステップＳ６）。そこでまず、音響イベント認識部１５は、音響イベント区間データＤ４を構成する音響イベントのフレーム列を、フレーム列連結により予め定めた長さＮフレーム以上に至るまで連結する。これは、短すぎるフレーム列からは音響イベントの周波数特性の時間変化をとらえることが困難となり、精度よく音響イベントを推定することは困難なためである。音響イベント認識部１５は、フレーム連結によりＮフレーム以上のフレーム列からなる入力フレーム列を得ると、記憶部１０に記憶されているニューラルネットワークを用いて、音響イベント認識を行う。

【００４１】

図６は、記憶部１０に記憶されている音響イベント認識用のニューラルネットワークを示す図である。同図に示すように、本実施形態では、音響イベント認識部１５は、音響イベント認識に、ニューラルネットワークの一種である畳み込みニューラルネットワークを用いる。畳み込みニューラルネットワークの例は、例えば、文献「Andrew L. Maas et al., "Word-level Acoustic Modeling with Convolutional Vector Regression", ICML Representation Learning Workshop, 2012」に記載されている。

【００４２】

同図に示す畳み込みニューラルネットワークは、入力層、隠れ層、プーリング層、出力層の４層から構成される。入力層は、音響イベント区間検出部１４で出力された時刻順の複数のフレームに対応し、入力層の値は、対応するフレームから得られたメル周波数ケプストラムなどの時間周波数領域の音響特徴量である。この音響特徴量は、例えば、ベクトルで表される。本実施形態において、入力層の音響特徴量の総フレーム数N_s（≧Ｎ）は可変である。

【００４３】
隠れ層の各ユニット（素子）は、入力層の総フレーム数N_sのフレーム（素子）のうち、連続するn_s個のフレームのみと結合している。隠れ層の各ユニットが結合している入力層のn_s個のフレームは、１つ前の隣接するユニットが結合しているn_s個のフレームよりも後の時刻に対応するが、一部が重複するようにｋフレームずつシフトしている（$k < n_s$）。例えば、入力層のｉ～（ｉ＋２）番目のフレームが隠れ層のｉ番目のユニットに結合しているとする。隠れ層のｉ番目のユニットの値は、入力層のｉ～（ｉ＋２）番目のフレームの値の加算（畳み込み演算）となる。ただし、入力層のｉ番目のフレーム、（ｉ＋１）番目のフレーム、（ｉ＋２）番目のフレームそれぞれと隠れ層のｉ番目のユニットとの結合重み（加算の際の重み）は均等でなくてもよい。例えば、入力層の１～３番目のフレームが隠れ層の第１番目のユニットに結合し、入力層の２～４番目のフレームが隠れ層の第２番目のユニットに結合し、入力層の３～５番目のフレームが隠れ層の第３番目のユニットに結合する。このとき、（入力層の１番目のフレームから隠れ層の１番目のユニットの結合重み）＝（入力層の２番目のフレームから隠れ層の２番目のユニットの結合重み）＝（入力層の３番目のフレームから隠れ層の３番目のユニットの結合重み）＝…である。同様に、（入力層の２番目のフレームから隠れ層の１番目のユニットの結合重み）＝（入力層の３番目のフレームから隠れ層の２番目のユニットの結合重み）＝（入力層の４番目のフレームから隠れ層の３番目のユニットの結合重み）＝…である。つまり、隠れ層のユニットと入力層のフレームとの結合は、ｋフレーム分のシフトを保ちながら、入力層と隠れ層の各素子の間を同じ結合重みで結んでいる。隠れ層のユニット数N_hは、入力層のユニット数に応じた数になる。

【００４４】
隠れ層の上位のプーリング層は、予め定められた固定のユニット数N_pのユニットにより構成される。プーリング層の各ユニットは、隠れ層のユニットのうち可変のユニット数$n_h = N_p / N_h$のユニットと結合している。プーリング層のユニットと隠れ層のユニットとの結合は、同じプーリング層のユニットに結合されている隠れ層のユニットの値のうち、最大値のみプーリング層に伝搬するという特質をもつ。

【００４５】
プーリング層と出力層は、互いに各ユニットが全て結合している。出力層の値は、プーリング層の値に、プーリング層の各ユニットと出力層の各ユニットとの間それぞれの重みを表す重み係数行列を作用させた後、Softmax関数を用いて出力層の各ユニットの出力を正規化して計算される。出力層のユニットは、音響イベントに対応したテキスト表現（文字列）を表しており、音響特徴量が与えられたときのテキスト表現の事後確率を与える。
なお、本実施形態では、プーリング層と出力層を連結しているが、この間には任意の数の隠れ層及びプーリング層を挿入可能である。

【００４６】
図７は、音響イベント認識部１５の音響イベント認識処理フローを示す図であり、図３のステップＳ６における詳細な処理を示す。
音響イベント認識部１５は、畳み込みニューラルネットワークの入力特徴量が十分な長さとなるよう、音響イベント区間検出部１４からの出力である音響イベント区間データＤ４のフレーム列を時刻順にフレーム連結し、入力フレーム列を生成する（ステップＳ６１）。入力フレーム列の長さがＮに達していない場合（ステップＳ６２：ＮＯ）、音響イベント認識部１５は、ステップＳ６１に戻り、Ｎフレーム以上の入力フレーム列が得られるまで新たな音響イベント区間データＤ４のフレーム列をフレーム連結する。入力フレーム列の長さが音響イベント認識に必要なＮ以上となった場合（ステップＳ６２：ＹＥＳ）、音響イベント認識部１５は、記憶部１０に記憶されている畳み込みニューラルネットワークにより音響イベント認識を行う（ステップＳ６３）。音響イベント認識部１５は、入力フレーム列を構成する各フレームの音響特徴量を計算する。音響イベント認識部１５は、入力フレーム列の各フレームについて計算した音響特徴量を、図６に示す畳み込みニューラルネットワークの入力層の入力とし、隠れ層、プーリング層、出力層の各ユニットの値

を計算する。

【００４７】

最後に音響イベント認識部１５は、畳み込みニューラルネットワークの出力層のユニットを、各ユニットの出力が示す事後確率に基づいて選択する。例えば、音響イベント認識部１５は、事後確率が最大のものから順に所定数のユニットを選択してもよく、事後確率が所定以上のユニットを選択してもよく、事後確率が所定以上の中から事後確率が大きい順に所定数までのユニットを選択してもよい。記憶部１０には、予め、出力層のユニットの番号と、その番号のユニットが表す音響イベントについてユーザが選んだテキスト表現とを対応付けて記憶しておく。音響イベント認識部１５は、選択したユニットに対応する音響イベントのテキスト表現を記憶部１０から読み出す。

【００４８】

本実施形態では、以下の表１から表５に示すような分類に従った音響イベントのテキスト表現を用いる。

【００４９】

【表１】

(1) 人間の発声に基づくもの

音響イベントの種類	テキスト表現の例
笑い声	(笑い)
泣き声	(泣き声)
せき	(せき)
いびき	(いびき)
くしゃみ	(くしゃみ)
あくび	(あくび)
舌打ち	(舌打ち)
鼻歌	(鼻歌)
歓声	(歓声)

【００５０】

【表２】

(2) 体の部位によるもの

音響イベントの種類	テキスト表現の例
手をならす音	(拍手)

【００５１】

【表３】

(3) 自然現象に基づくもの

音響イベントの種類	テキスト表現の例
雨に関する音	(雨音)
雷に関する音	(雷鳴)
動物の鳴き声	(犬の鳴き声)

【００５２】
【表４】

(4) 音楽・楽器等によるもの

音響イベントの種類	テキスト表現の例
音楽	♪～
楽器による音	(ピアノ)

【００５３】
【表５】

(5) 人工的な音

音響イベントの種類	テキスト表現の例
道具・機械による音	(金づちでたたく音)
呼び鈴	(チャイム)
銃撃音	(銃声)
爆発音	(爆発音)

【００５４】
表１から表５では、該当する音響イベントのテキスト表現の例を示しているが、ある音響イベントに対応するテキスト表現を一意に定めることは難しい。そこで、過去に行われた字幕放送のテキストを解析し、頻度の高い代表的な表現をテキスト表現として選んでおく。例えば、これらの表現は、字幕放送のト書き（場面の説明を行う脚注）として表現されるものである。そして、出力層のユニットの番号と、その番号のユニットが表す音響イベントとして選んだテキスト表現とを対応付けて記憶部１０に記憶しておく。
【００５５】
図３において、音響イベント認識部１５は、読み出した音響イベントのテキスト表現に、事後確率が大きい順に順位を付与する。音響イベント認識部１５は、順位が付与された音響イベントのテキスト表現である注釈文字列を音響イベント認識結果データＤ５に設定し、認識結果修正部１６に出力する（ステップＳ７）。
【００５６】
認識結果修正部１６は、音声認識結果データＤ３が示す音声認識結果と、音響イベント認識結果データＤ５が示す注釈文字列とを統合して、最終的な放送字幕を作成する（ステップＳ８）。本実施形態の音声認識装置１は、両者を効率的に実施可能な効率的なインタフェースを提供する。このインタフェースの提供方法には、以下の２つがある。

【００５７】
第１のインタフェースの提供方法は、修正者が認識結果を修正する際に、注釈を挿入する方法である。認識結果の修正は、タッチパネルを具備したコンピュータ装置によって実現される修正端末５を用い、操作者の入力に基づいて行われる。
【００５８】
図８は、修正端末５の表示部５２に表示されるコンピュータディスプレイ画面である修正作業画面８を示す。修正作業画面８は、音声認識結果表示ウィンドウ８０、音響イベント認識結果表示ウィンドウ８３、音響イベント認識結果候補ウィンドウ８６、履歴表示ウィンドウ８７を含む。
音声認識結果表示ウィンドウ８０は、音声認識結果と、音声認識結果に修正や注釈文字列の挿入を行った文字列とを表示する。音響イベント認識結果表示ウィンドウ８３は、注釈文字列を表示する。音響イベント認識結果表示ウィンドウ８３に表示される注釈文字列は、音響イベント認識結果データＤ５に設定されている順位が最も高い注釈文字列である。音響イベント認識結果候補ウィンドウ８６は、注釈文字列の候補を表示する。注釈文字列の候補は、音響イベント認識結果データＤ５に設定されている順位が２番目以下の注釈文字列である。履歴表示ウィンドウ８７は、音声認識結果に対する修正文字列を表示する。
【００５９】
音声認識装置１の認識結果修正部１６は、音声認識部１３から出力された音声認識結果データＤ３と、音響イベント認識部１５から出力された音響イベント認識結果データＤ５を、修正端末５に随時出力する。このとき、認識結果修正部１６は、音声認識結果データＤ３に対応した音声データＤ１も修正端末５に出力する。認識結果修正部１６は、修正端末５に出力した音声認識結果データＤ３が示す音声認識結果を作業中字幕とする。
【００６０】
各修正端末５の制御部５１は、受信した音声データＤ１の再生音声を音声出力部５４から出力する。制御部５１は、音声認識結果表示ウィンドウ８０に、受信した音声認識結果データＤ３から読み出した音声認識結果を、修正対象の文字列として音声認識結果表示ウィンドウ８０の最下行に表示させる。このとき、制御部５１は、音声認識結果を、単語間に縦棒を挟んだ文字列により表示させる。なお、音声認識結果表示ウィンドウ８０にすでに最下行まで修正済みの音声認識結果が表示されていた場合、制御部５１は、表示していた修正済みの音声認識結果の中で最も先の時刻の修正済みの音声認識結果を消去する。消去後、制御部５１は、残りの修正済みの音声認識結果を現在よりも上の行に移動し、受信した音声認識結果データＤ３から読み出した音声認識結果を、音声認識結果表示ウィンドウ８０の最下行に表示させる。
【００６１】
また、各修正端末５の制御部５１は、音響イベント認識結果表示ウィンドウ８３の右端から順に最新の注釈文字列を表示させる。つまり、制御部５１は、音声認識装置１から新たな音響イベント認識結果データＤ５を受信する度に、音響イベント認識結果表示ウィンドウ８３に表示していた注釈文字列を左にシフトして表示させる。制御部５１は、新たに受信した音響イベント認識結果データＤ５から読み出した、最も順位の高い注釈文字列を、音響イベント認識結果表示ウィンドウ８３の右端に表示させる。また、制御部５１は、音響イベント認識結果候補ウィンドウ８６に、受信した音響イベント認識結果データＤ５に設定されている２位以下の順位の注釈文字列をメニュー表示させる。
【００６２】
音声認識結果の修正作業は、以下のように行う。修正者は、番組音声を聞きながら、音声認識結果表示ウィンドウ８０により表示部５２が表示している文字列の中から、修正対象の文字列を含む文字の表示部分を指などにより触れる。修正者は、指を移動させて、複数の文字に触れてもよい。入力部５３は、接触を検知した画面位置の情報を制御部５１に出力する。制御部５１は、接触を検知した画面位置に表示させていた文字が含まれる単語を選択し、選択された単語を特定する指摘情報を音声認識装置１に送信する。例えば、指

摘情報には、単語が発音された時刻を用いることができる。音声認識装置１の認識結果修正部１６は、修正端末５－１からの指摘情報を最も早く受信したとする。認識結果修正部１６は、修正端末５－１から受信した指摘情報により示される文字列の表示を赤色等の選択色に変更するよう各修正端末５に指示する。各修正端末５の制御部５１は、音声認識装置１からの指示に基づき選択された文字列の表示を選択色に変更する。さらに、認識結果修正部１６は、修正端末５－２には、選択色に変更に併せて修正ガードを指示する。修正ガードが指示された修正端末５－２においては、修正作業や注釈の挿入作業はできない。

【００６３】

修正端末５－１を使用している修正者は、入力部５３を用いて、選択色で表示されている文字列に対する置換、挿入、消去などの修正作業を行う。例えば、修正者は、単語が選択された状態で、キーボードにより文字を入力する。修正者は、修正作業が終了すると、修正作業終了操作として、キーボード上でＥｎｔｅｒ等のキーを押下する。制御部５１は、修正作業終了操作の入力を受けると、修正作業の内容を音声認識装置１に送信する。音声認識装置１の認識結果修正部１６は、作業中字幕における選択文字列を、修正端末５－１から受信した修正作業内容に従って修正し、新たな作業中字幕を生成する。認識結果修正部１６は、新たな作業中字幕と、修正作業において修正者がキーボードから入力した文字列を各修正端末５に送信する。各修正端末５の制御部５１は、音声認識装置１から受信した作業中字幕により、音声認識結果表示ウィンドウ８０に表示されている音声認識結果の表示を置き代える。また、各修正端末５の制御部５１は、一覧の作業の履歴として、修正者がキーボードから入力した文字列を履歴表示ウィンドウ８７に表示させる。修正端末５－２は、修正ガードを解除する。

【００６４】

注釈の挿入作業は、以下のように行う。修正者は、番組音声を聞きながら、音響イベント認識結果表示ウィンドウ８３に表示されている任意の注釈文字列を、音声認識結果表示ウィンドウ８０に表示されている文字列の任意の箇所に挿入していく。

例えば、文字列８１が示す音声認識結果（あるいは修正済み音声認識結果）「お料理が上手ですね。」の直後に、音響イベント認識結果表示ウィンドウ８３に表示されている注釈文字列８４「（笑い）」を挿入する場合、修正者は次の操作を行う。修正者は、注釈文字列を挿入したい文字列８１の最後の文字「。」に触れる。入力部５３は、接触を検知した画面位置の情報を制御部５１に出力する。制御部５１は、接触を検知した画面位置に表示させていた文字が含まれる単語「。」を選択し、選択された単語を特定する指摘情報を音声認識装置１に送信する。つまり、このときの指摘情報は、注釈挿入位置を示す。音声認識装置１の認識結果修正部１６は、修正端末５－１からの指摘情報を最も早く受信したとする。認識結果修正部１６は、修正端末５－１から受信した指摘情報により示される文字列の表示を赤色等の選択色に変更するよう各修正端末５に指示する。各修正端末５の制御部５１は、音声認識装置１からの指示に基づき、選択された文字列の表示を選択色に変更する。さらに、認識結果修正部１６は、修正端末５－２に、選択色への変更に併せて修正ガードを指示する。

【００６５】

修正端末５－１を使用している修正者は、キーボード上の「挿入（Insert）」キーを押下し、さらに、注釈文字列８４「（笑い）」のいずれかの文字に触れる。入力部５３は、「挿入（Insert）」キーの押下と、接触を検知した画面位置の情報を制御部５１に出力する。制御部５１は、接触を検知した画面位置に表示させていた文字が含まれる注釈文字列を判断すると、その注釈文字列を特定する情報、あるいは、注釈文字列を設定した挿入注釈情報を音声認識装置１に送信する。先に送信した指摘情報と挿入注釈情報とを併せたものが注釈挿入指示に相当する。音声認識装置１の認識結果修正部１６は、挿入注釈情報により特定される、あるいは、挿入注釈情報が示す注釈文字列を、作業中字幕における選択された単語「。」の直後に挿入し、新たな作業中字幕「お料理が上手ですね。（笑い）」を生成する。認識結果修正部１６は、新たな作業中字幕を各修正端末５に送信する。各修正端末５の制御部５１は、音声認識装置１から受信した作業中字幕により、音声認識結果

表示ウィンドウ８０に表示されている音声認識結果（あるいは修正済み音声認識結果）の表示を置き代える。修正端末５－２は、修正ガードを解除する。

【００６６】

なお、修正者は、注釈文字列「（笑い）」を挿入したい場合、音響イベント認識結果表示ウィンドウ８３に表示されている注釈文字列８４「（笑い）」に代えて、注釈文字列８５「（笑い）」のいずれかの文字に触れてもよい。

また、例えば、音声認識結果表示ウィンドウ８０に表示されている文字列８２が示す修正済みの認識結果「○○さんの趣味はなんですか。」の直後に、注釈文字列を挿入する場合、文字列８２の最後の文字「。」に触れればよい。

【００６７】

音響イベント認識結果が誤っている場合、音響イベント認識結果表示ウィンドウ８３から正しい注釈文字列を選択することができない。この場合、作業者は、音響イベント認識結果候補ウィンドウ８６にメニュー表示される注釈文字列の候補の一覧の中から、挿入する注釈文字列を選択する。

【００６８】

第２のインタフェースの提供方法は、修正後の文字列の装飾時に注釈文字列を挿入する方法である。情報番組やスポーツ中継の字幕制作では、話者（番組出演者）に応じて、該当する字幕の色を、白、青、黄等に色分けすることが行われる。色分けは、修正後の字幕について別の作業者が行うことが多い。この場合は、図８に示す画面において、文字列を修正する代わりに、表示されている文字列の各行に対して適切な色を指定する同時に、音響イベント認識結果表示ウィンドウ８３から適切な音響イベント認識結果を挿入すればよい。以下では、修正端末５－１により音声認識結果の修正を行い、修正端末５－２により修正後の音声認識結果に装飾を行う場合について、第１のインタフェースの提供方法との差分を中心に説明する。

【００６９】

音声認識装置１の認識結果修正部１６は、音声認識部１３から出力された音声認識結果データＤ３、及び対応する音声データＤ１と、音響イベント認識部１５から出力された音響イベント認識結果データＤ５を、修正端末５に随時出力する。各修正端末５の制御部５１は、受信した音声データＤ１の再生音声を音声出力部５４から出力し、図８に示す修正作業画面８を示す。修正端末５－１の修正者による音声認識結果の修正作業は、第１のインタフェースの提供方法と同様である。ただし、音声認識装置１の認識結果修正部１６は、音声認識結果の修正を行う他の修正端末５がある場合には修正ガードを送信するが、修正後の音声認識結果に装飾を行う修正端末５－２には、修正ガードを送信しなくてもよい。

【００７０】

続いて、音声認識装置１の認識結果修正部１６は、新たに生成された音声認識結果データＤ３と、対応する音声データＤ１を音声認識装置１に出力する。各修正端末５の制御部５１は、新たに受信した音声データＤ１の再生音声を音声出力部５４から出力する。さらに、制御部５１は、第１のインタフェースの提供方法と同様に、受信した音声認識結果データＤ３から読み出した音声認識結果を、修正対象の文字列として音声認識結果表示ウィンドウ８０の最下行に表示させる。

【００７１】

修正端末５－２の修正者は、番組音声を聞きながら、音声認識結果表示ウィンドウ８０により表示部５２が表示している文字列の中から、色を変えたい修正済みの音声認識結果（例えば、文字列８２）を含む文字の表示部分を指などにより触れ、文字色を入力する。文字色は、キーボードなどにより入力してもよく、音声認識結果表示ウィンドウ８０に文字色を選択するボタンを設け、そのボタンに触れることにより入力してもよい。入力部５３は、接触を検知した画面位置の情報を制御部５１に出力する。制御部５１は、接触を検知した画面位置に表示させていた文字が含まれる行を選択し、選択された行を特定する情報と、入力された文字色とを示す装飾情報を音声認識装置１に送信する。音声認識装置１

の認識結果修正部１６は、修正端末５－２から受信した装飾情報により示される作業中字幕における行の文字列を、装飾情報により示される文字色に変更し、新たな作業中字幕を生成する。認識結果修正部１６は、選択された行の文字列を、変更後の文字色により表示するよう各修正端末５に指示する。各修正端末５の制御部５１は、音声認識装置１からの指示に従って、音声認識結果表示ウィンドウ８０の指定された行（修正済みの音声認識結果）の文字列を変更後の文字色により表示する。

【００７２】

さらに、修正端末５－２の修正者は、音響イベント認識結果表示ウィンドウ８３に表示されている任意の注釈文字列を、音声認識結果表示ウィンドウ８０に表示されている修正済みの音声認識結果の任意の箇所に挿入していく。

例えば、文字列８２が示す修正済みの音声認識結果「○○さんの趣味はなんですか。」の直後に、注釈文字列８４「（笑い）」を挿入する場合、修正者は、キーボード上の「挿入（Insert）」キーを押下し、さらに、文字列８２の最後の文字「。」に触れる。入力部５３は、「挿入（Insert）」キーの押下と、接触を検知した画面位置の情報を制御部５１に出力する。制御部５１は、接触を検知した画面位置に表示させていた文字が含まれる単語を選択し、選択された単語を特定する注釈挿入位置情報を生成する。さらに、修正者は、注釈文字列８４「（笑い）」のいずれかの文字に触れる。入力部５３は、接触を検知した画面位置の情報を制御部５１に出力する。制御部５１は、接触を検知した画面位置に表示させていた文字が含まれ注釈文字列を判断すると、その注釈文字列を特定する情報、あるいは、注釈文字列を設定した挿入注釈情報を生成する。制御部５１は、注釈挿入位置情報と挿入注釈情報を設定した注釈挿入指示を音声認識装置１に送信する。音声認識装置１の認識結果修正部１６は、注釈挿入位置情報により、作業中字幕における注釈挿入対象の単語「。」を特定する。認識結果修正部１６は、挿入注釈情報により特定される、あるいは、挿入注釈情報が示す注釈文字列を、作業中字幕における注釈挿入対象の単語「。」の直後に挿入し、新たな作業中字幕を生成する。認識結果修正部１６は、新たな作業中字幕を各修正端末５に送信する。各修正端末５の制御部５１は、音声認識装置１から受信した作業中字幕により、音声認識結果表示ウィンドウ８０に表示されている修正済みの音声認識結果の表示を置き代える。

【００７３】

図２において、音声認識装置１の認識結果修正部１６は、上記の音声認識結果の修正作業と、注釈の挿入作業とが反映された作業中字幕を設定した注釈付き放送字幕データＤ６を出力する（ステップＳ９）。注釈付き放送字幕データＤ６は、放送局内で放送波に重畳されて放送される。

【００７４】

上記のように、修正者は、音響イベントのテキスト表現である注釈を、簡易な操作によって音声認識結果に挿入し、注釈付き字幕を制作することができる。よって、キーボード入力により注釈文字列を挿入する場合と比較し、大幅に作業を効率化することが可能となる。

【００７５】

なお、字幕制作システムが修正端末５を１台のみ備える場合、第１のインタフェースの提供方法において、音声認識装置１の認識結果修正部１６は、上述した処理のうち、最も早く指摘情報を送信した修正端末５以外の修正端末５との間の動作は実行しない。

また、認識結果修正部１６は、音響イベント認識結果が変わったタイミングで、音響イベント認識結果データＤ５を修正端末５に出力して表示させるようにしてもよい。これにより、音響イベント認識結果表示ウィンドウ８３に、同じ注釈文字列が連続して表示されないようにすることができる。

【００７６】

本実施形態によれば、音声認識装置１は、従来の音声認識に加え、音響イベントの認識を並行して行って修正端末５にそれらの認識結果を表示させ、修正者は、修正端末５の表示から注釈挿入位置と、挿入する注釈（音響イベントのテキスト表現）を指定する。従っ

て、人手による注釈付き字幕制作の負荷を大幅に軽減することが可能となる。また、音声認識装置１は、様々な種類の音響イベントについてのテキスト表現を認識結果として得ることができるため、得られた音響イベントのテキスト表現を注釈として字幕に挿入することによって、より豊かな字幕表現が可能となる。

【００７７】

なお、上述の音声認識装置１は、内部にコンピュータシステムを有している。そして、音声認識装置１の動作の過程は、プログラムの形式でコンピュータ読み取り可能な記録媒体に記憶されており、このプログラムをコンピュータシステムが読み出して実行することによって、上記処理が行われる。ここでいうコンピュータシステムとは、ＣＰＵ及び各種メモリやＯＳ、周辺機器等のハードウェアを含むものである。

【００７８】

また、「コンピュータシステム」は、ＷＷＷシステムを利用している場合であれば、ホームページ提供環境（あるいは表示環境）も含むものとする。

また、「コンピュータ読み取り可能な記録媒体」とは、フレキシブルディスク、光磁気ディスク、ＲＯＭ、ＣＤ－ＲＯＭ等の可搬媒体、コンピュータシステムに内蔵されるハードディスク等の記憶装置のことをいう。さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回線等の通信回線を介してプログラムを送信する場合の通信線のように、短時間の間、動的にプログラムを保持するもの、その場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリのように、一定時間プログラムを保持しているものも含むものとする。また上記プログラムは、前述した機能の一部を実現するためのものであっても良く、さらに前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるものであっても良い。

【符号の説明】

【００７９】

１…音声認識装置、５…修正端末、１０…記憶部、１１…音声分岐部、１２…音声区間検出部、１３…音声認識部、１４…音響イベント区間検出部、１５…音響イベント認識部、１６…認識結果修正部、５１…制御部、５２…表示部、５３…入力部、５４…音声出力部

【図１】

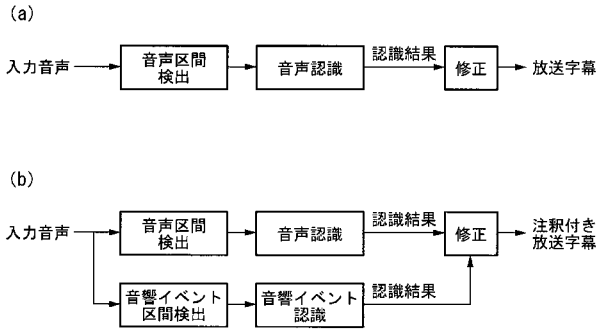

図１

【図２】

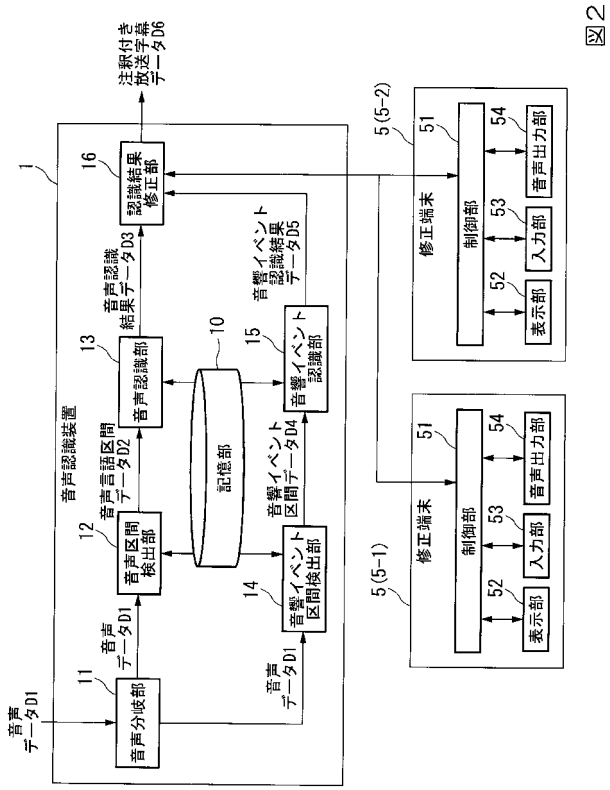

図２

【図３】

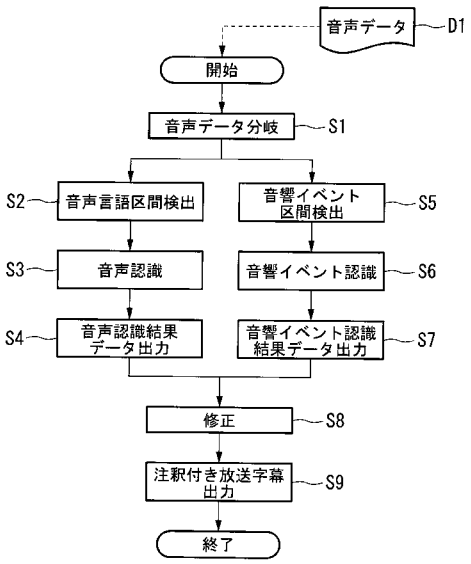

図３

【図４】

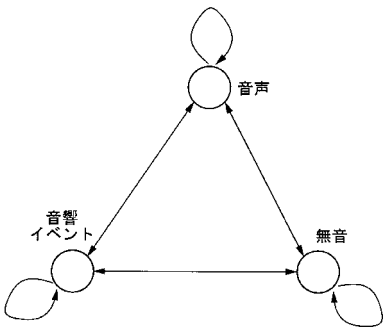

図４

【図５】

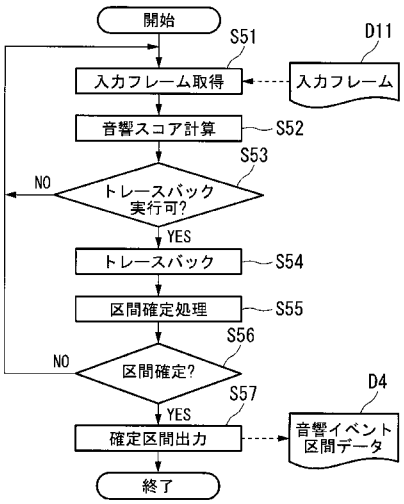

図5

【図６】

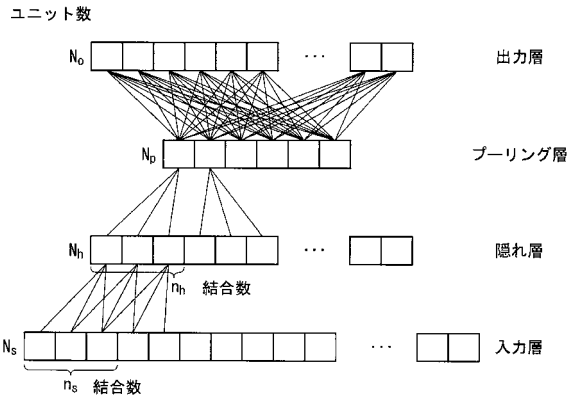

図6

【図７】

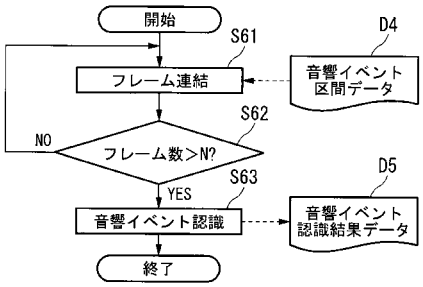

図7

【図８】

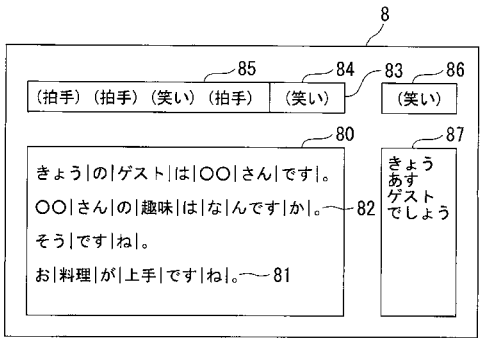

図8