(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2004-233541
(P2004-233541A)
(43)公開日　平成16年8月19日(2004.8.19)

(51)Int.Cl.[7]	F I			テーマコード（参考）
G10L 15/08	G10L	3/00	531W	5C053
G10L 15/00	G10L	3/00	521R	5D015
G10L 15/06	G10L	3/00	537G	
G10L 15/10	G10L	3/00	551G	
G10L 15/18	H04N	5/91	Z	

審査請求 有　請求項の数 5　OL　（全 16 頁）　最終頁に続く

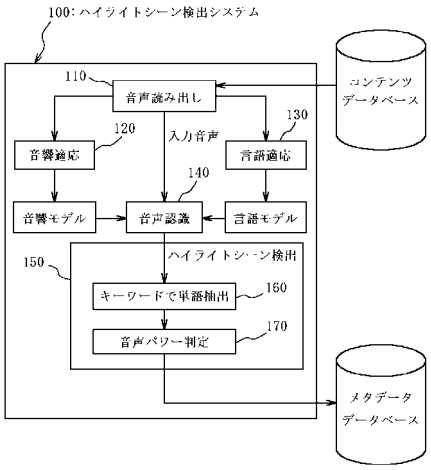
(21)出願番号　特願2003-20643(P2003-20643)
(22)出願日　平成15年1月29日(2003.1.29)

(71)出願人　597065329
学校法人　龍谷大学
京都府京都市伏見区深草塚本町６７番地
(71)出願人　391004104
株式会社毎日放送
大阪府大阪市北区茶屋町１７番１号
(74)代理人　100072051
弁理士　杉村　興作
(72)発明者　有木　康雄
滋賀県大津市瀬田大江町横谷１－５　龍谷大学内
(72)発明者　塚田　清志
大阪府大阪市北区茶屋町１７番１号　株式会社毎日放送内

Fターム(参考)　5C053 FA30 GB11 JA01
5D015 AA04 GG01 HH00 KK02 LL07

(54)【発明の名称】ハイライトシーン検出システム

(57)【要約】
【課題】コンテンツ中の音声情報を利用してコンテンツを構造化する、特に、スポーツ中継のハイライトシーンを自動的により正確かつ効率的に自動検出する手法がなかった。
【解決手段】予め作成した音響モデルおよび言語モデルを参照して、入力音声の音声認識を行って単語を抽出する音声認識手段と、前記抽出した単語のうちハイライトシーンを特徴付けるキーワードと合致するものを検索し、このキーワードと合致した単語のうち予め規定された閾値を超える音声パワーを持つような単語が存在するシーンをハイライトシーンとして検出するハイライトシーン検出手段とを含むハイライトシーン検出システムを提供する。
【選択図】　図１



【特許請求の範囲】
【請求項１】
ハイライトシーン検出システムであって、
予め作成した音響モデルおよび言語モデルを参照して、入力音声の音声認識を行って単語を抽出する音声認識手段と、
前記抽出した単語のうちハイライトシーンを特徴付けるキーワードと合致するものを検索し、このキーワードと合致した単語のうち予め規定された閾値を超える音声パワーを持つような単語が存在するシーンをハイライトシーンとして検出するハイライトシーン検出手段と、
を含むハイライトシーン検出システム。
【請求項２】
請求項１に記載のハイライトシーン検出システムであって、
予備音声データを用いて、ベースライン音響モデルに対して、ＭＬＬＲ法およびＭＡＰ法で教師あり適応を施して基本音響モデルを作成し、さらに前記入力音声から前記基本音響モデルを用いて音声認識を行い、その結果からラベルを自動作成し、このラベルを用いて前記ＭＬＬＲ法およびＭＡＰ法で教師なし適応を施して前記音響モデルを作成する音響モデル作成手段、
を含むことを特徴とするハイライトシーン検出システム。
【請求項３】
請求項１または２に記載のハイライトシーン検出システムであって、
ウェブコーパスと、書き下しコーパスとを結合した結合コーパスを用いて、ベースライン言語モデルに対して適応を施し第１の予備言語モデルを作成し、前記ベースライン言語モデルに対して前記書き下しコーパスを用いて適応を施し第２の予備言語モデルを作成し、これら第１および第２の予備言語モデルを融合して、前記言語モデルを作成する言語モデル作成手段、
を含むことを特徴とするハイライトシーン検出システム。
【請求項４】
請求項１～３のいずれか１項に記載のハイライトシーン検出システムであって、
前記ハイライトシーン検出手段で検出された前記ハイライトシーンの単語と、前記ハイライトシーンの時間情報とを関連付けたメタデータを作成し、このメタデータを記憶装置に格納する手段と、
所望のハイライトシーンを特徴付ける所望のキーワードに基づき記憶装置に格納された前記メタデータを検索して、前記所望のハイライトシーンを探し出すハイライトシーン検索手段と、
を含むことを特徴とするハイライトシーン検出システム。
【請求項５】
請求項１～３のいずれか１項に記載のハイライトシーン検出システムであって、
前記ハイライトシーン検出手段で検出された前記ハイライトシーンの単語と、前記ハイライトシーンの時間情報と、を関連付け、さらに、これらに前記入力音声を取得した映像コンテンツの時間情報、或いは、前記入力音声と同一の対象を撮影した映像コンテンツの時間情報と、を関連付けたメタデータを作成し、このメタデータを記憶装置に格納する手段と、
所望のハイライトシーンを特徴付ける所望のキーワードに基づき、記憶装置に格納された前記メタデータを検索して、前記映像コンテンツのなかから前記所望のハイライトシーンを探し出すハイライトシーン検索手段と、
を含むことを特徴とするハイライトシーン検出システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、ハイライトシーン検出システムに関するものであり、特にスポーツ中継のハイ

ライトシーンを検出するシステムに関するものである。
【０００２】
【従来の技術】
近年の放送、通信、インターネットの分野の発展に伴い、アナログ或いはデジタルの膨大な映像や音声のコンテンツが蓄積されてきた。しかしながら、このようなコンテンツは増加の一途を辿り、これら膨大なコンテンツの中から所望の情報を手に入れることが一層困難となってきている。特に、近年のスポーツブームで野球をはじめ、サッカー、ラグビー、アメリカンフットボール、格闘技などのスポーツ中継が盛んとなり、これらスポーツ中継の映像・音声コンテンツが増加しつつあるが、これらのスポーツ映像コンテンツは、重要なシーン（例えば、得点がからむようなハイライトシーン）の占める割合は、コンテンツ全体からみて少ない。
そこで、従来は、膨大なコンテンツの中から必要な情報（例えば、スポーツ中継のハイライトシーン）を必要な時に利用できるように映像に検索時のキーとなるインデックス情報を手作業で登録する手法があった。このようなインデックス情報の手作業登録は、多大な労力と時間がかかり、今後、ますます増大するコンテンツに、手作業でインデックス情報を付加し続けるのは困難である。従って、一旦放送された映像・音声コンテンツは再利用されることなく死蔵されるケースも多かった。
そこで、現在、映像の分野では、映像をコンピュータで解析し、手作業ではなく自動的にインデックス情報を付加し、映像コンテンツの構造化を目指す研究が行われている。例えば、カメラワークから構造化を図る手法（非特許文献１を参照されたい。）、映像中のテロップを解析する手法（非特許文献２を参照されたい。）、また、クローズドキャプションを用いる手法（非特許文献３を参照されたい。）などが提案されている。
【０００３】
【非特許文献１】
野球中継映像における各種プレイシーンの自動検索／編集システム（電子情報通信学会総合大会講演論文集情報・システム２，Ｄ１２－７７，ｐｐ２４７；山本拓他）
【非特許文献２】
野球中継におけるシーン検索（１９９７年１２月、第３回知能情報メディアシンポジウム論文集、ｐｐ１９５～２０２；館山公一他）
【非特許文献３】
メディア理解による映像メディアの構造化（１９９９年７月、信学技報、ＰＲＭＵ９９－４２、ｐｐ３９～４６；館山公一他）
【０００４】
【発明が解決しようとする課題】
上述した３つの手法は、基本的に映像或いは映像に付加された情報（テロップやキャプション）のみに基づく構造化手法である。非特許文献１による手法では、映像コンテンツのカメラワークのみを利用するため、カメラの配置や撮影対象の違いによるカメラワークの変化、カメラの切替えなどの変動要因によって、構造化が不正確になる場合があるといった欠点や、音声コンテンツに利用できないという欠点があった。また、非特許文献２および３の手法は、映像コンテンツにテロップやキャプションが付加されたものしか構造化できないという欠点がある。
そこで、本願発明は、コンテンツ中の「音声情報」を利用したコンテンツの構造化、特に、スポーツ中継のハイライトシーンを自動的により正確かつ効率的に自動検出する手法を提供することを目的とする。
【０００５】
【課題を解決するための手段】
本発明によるハイライトシーン検出システムは、
予め作成した音響モデルおよび言語モデルを参照して、入力音声の音声認識を行って単語を抽出する音声認識手段と、
前記抽出した単語のうちハイライトシーンを特徴付けるキーワードと合致するものを検索

し、このキーワードと合致した単語のうち予め規定された閾値を超える音声パワーを持つような単語が存在するシーンをハイライトシーンとして検出するハイライトシーン検出手段と、
を含むハイライトシーン検出システムである。
本構成によれば、音声情報のみによって高い精度でハイライトシーンを自動的に検出することが可能となる。本構成では、入力音声として映像コンテンツに含まれる音声情報を利用することも可能であるが、ラジオ中継ではアラウンサーがハイライトシーンで興奮し声が高くなるという現象（即ち音声パワーが高くなる）が顕著になる、換言すれば、声を高くすることでハイライトシーンの臨場感を高めるという演出手法をとることが多いため、入力音声としてラジオ番組の音声（そのなかでも特にスポーツ中継）を使用することがより好適である。即ち、ラジオ中継の音声情報を利用することによって、より正確かつ効率良くハイライトシーンを検出することが可能となる。また、ラジオ中継のような番組（音声コンテンツ）は、テレビ放送用の番組（映像コンテンツ）が同時に存在する場合が多いため、ラジオ中継の音声情報から得たハイライトシーンの情報を、同じ対象を撮影した映像コンテンツのハイライトシーンの情報として扱うことが可能である。
【０００６】
また、本発明によるハイライトシーン検出システムは、
予備音声データ（および書き起こしたラベル）を用いて、ベースライン音響モデルに対して、ＭＬＬＲ法およびＭＡＰ法で教師あり適応を施して基本音響モデルを作成し、さらに前記入力音声から前記基本音響モデルを用いて音声認識を行い、その結果からラベルを自動作成し、このラベルを用いて前記ＭＬＬＲ法およびＭＡＰ法で教師なし適応を施して前記音響モデルを作成する音響モデル作成手段、
を含むことを特徴とする。
本構成によれば、ＭＬＬＲ法およびＭＡＰ法を用いて、教師あり適応および教師なし適応の２段階で音響モデル適応を行うことで、より、正確な音声認識が可能となり、これによってハイライトシーンの検出がより正確になる。
【０００７】
また、本発明によるハイライトシーン検出システムは、
ウェブ上のテキスト集合（ウェブコーパス）と、発話から書き下したテキスト集合（書き下しコーパス）とを結合した結合コーパスを用いて、ベースライン言語モデルに対して適応を施し第１の予備言語モデルを作成し、前記ベースライン言語モデルに対して発話から書き下したテキスト集合（書き下しコーパス）を用いて適応を施し第２の予備言語モデルを作成し、これら第１および第２の予備言語モデルを融合して、前記言語モデルを作成する言語モデル作成手段、
を含むことを特徴とする。
或いは、本システムは、ベースライン言語モデルに対してウェブ上のテキスト集合（ウェブコーパス）を用いて適応を施し第１の予備言語モデルを作成し、前記ベースライン言語モデルに対して発話から書き下したテキスト集合（書き下しコーパス）を用いて適応を施し第２の予備言語モデルを作成し、これら第１および第２の予備言語モデルを融合して、前記言語モデルを作成する言語モデル作成手段を含むことを特徴とする。
或いは、本システムは、ウェブ上のテキスト集合（ウェブコーパス）と、発話から書き下したテキスト集合（書き下しコーパス）とを結合した結合コーパスを用いて、ベースライン言語モデルに対して適応を施し言語モデルを作成する言語モデル作成手段を含むことを特徴とする。
本構成によれば、音声認識が正確、即ち、ハイライトシーンを特徴付ける単語の認識がより正確になり、よってハイライトシーン検出の精度がより高まる。
【０００８】
また、本発明によるハイライトシーン検出システムは、
前記ハイライトシーン検出手段で検出された前記ハイライトシーンの単語と、前記ハイライトシーンの時間情報とを関連付けたメタデータを作成し、このメタデータを記憶装置に

格納する手段と、
所望のハイライトシーンを特徴付ける所望のキーワードに基づき前記記憶装置に格納された前記メタデータを検索して、前記所望のハイライトシーンを探し出すハイライトシーン検索手段と、
を含むことを特徴とする。
本構成によれば、迅速、簡易かつ効率良くハイライトシーンを検索することが可能となる。
【０００９】
また、本発明によるハイライトシーン検出システムは、
前記ハイライトシーン検出手段で検出された前記ハイライトシーンの単語と、前記ハイライトシーンの時間情報と、を関連付け、さらに、これらに前記入力音声を取得した映像コンテンツの時間情報、或いは、前記入力音声と同一の対象を撮影した映像コンテンツの時間情報と、を関連付けたメタデータを作成し、このメタデータを記憶装置に格納する手段と、
所望のハイライトシーンを特徴付ける所望のキーワードに基づき、記憶装置に格納された前記メタデータを検索して、前記映像コンテンツのなかから前記所望のハイライトシーンを探し出すハイライトシーン検索手段と、
を含むことを特徴とする。
本構成によれば、迅速、簡易かつ効率良く、映像コンテンツのハイライトシーンを検索することが可能となる。
【００１０】
また、本発明は方法の形態でも実現でき、例えば、本発明によるハイライトシーン検出方法は、
記憶手段に格納された予め作成した音響モデルおよび言語モデルを参照して、入力音声の音声認識を行って単語を抽出する音声認識ステップと、
前記抽出した単語のうちハイライトシーンを特徴付けるキーワードと合致するものを検索し、このキーワードと合致した単語のうち予め規定された閾値を超える音声パワーを持つような単語が存在するシーンをハイライトシーンとして検出するハイライトシーン検出ステップと、
を含む。
また、本発明は、上記方法を実現するプログラムの形態でも実施可能である。その場合は、上記方法の各ステップを実行するプログラムを記憶手段から読み出し、ＣＰＵやＤＳＰなどの演算手段上で各ステップに含まれるインストラクションを実行する。
【００１１】
【発明の実施の形態】
以下、添付する諸図面を参照しつつ、本発明の具体的な実施例を詳細に説明する。
図１は、本発明によるハイライトシーン検出システムを説明する概念図である、本システムの概略を説明するが、まず入力された中継音声（入力音声）を、適応された音響モデルと言語モデルとを用いて音声認識を行う。次に音声認識結果からハイライトシーンに関連するキーワードと一致する単語を取り出す。この単語のうち、所定の閾値よりも音声パワーの大きい区間をハイライトシーンとして検出する。以下、各部の詳細を説明する。図に示すように、本発明によるハイライトシーン検出システム１００は、記憶装置に格納されたコンテンツデータベースから音声情報を読み出す音声読み出し手段１１０、読み出した音声情報に基づき音響モデルを作成する音響モデル作成（適応）手段１２０、読み出した音声情報に基づき言語モデルを作成する言語モデル作成（適応）手段１３０、読み出した音声情報を入力音声として音声認識を実施する音声認識手段１４０、音声認識結果に基づきハイライトシーンを検出するハイライトシーン検出手段１５０から構成される。ハイライトシーン検出手段１５０は、ハイライトシーンを特徴付けるような所望のキーワードに合致する単語を抽出する単語抽出手段１６０と、抽出された単語の音声パワーが所定の閾値を超えるか否かを判定し、超える単語がある時間区間をハイライトシーンとみなし、メ

タデータとして記録する音声パワー判定（閾値処理）手段１７０とを含む。

【００１２】

次に、本発明で使用する音響モデルについて詳細に説明する。本システムで作成した音響モデルのベースラインとなる音響モデルは、比較的話し言葉に近い特徴を持った学会講演音声を用いて学習している。このベースライン音響モデルに、対象の音声情報の話し手であるアナウンサーを教師とした教師有り適応などを施して音響モデルを作成した。話者適応におけるモデルパラメータの推定手法においては、ＭＬＬＲ（Ｍａｘｉｍｕｍ Ｌｉｋｅｌｉｈｏｏｄ Ｌｉｎｅａｒ Ｒｅｇｒｅｓｓｉｏｎ）法などのようにモデルパラメータ間での情報の共有化を利用し、パラメータ空間へ一括移動させる方法や、ＭＡＰ推定法などのように適応学習における事前知識を効率的に利用した方法を用いることができる。本発明によるハイライトシーン検出システムでは、これら２つの手法を組み合わせた適応手法であるＭＬＬＲ＋ＭＡＰを用いている。即ち、まずＭＬＬＲによってモデルパラメータの変換を行い、それを事前知識としてＭＡＰ推定を行う。また、本発明によるハイライトシーン検出システムでは、ＭＬＬＲ＋ＭＡＰによる適応処理を行う際に、教師あり適応と、教師なし適応とを併用している。

【００１３】

まず、「教師あり適応」と「教師なし適応」との差異について述べる。一般的に、音響モデルの適応処理には、適応データの書き起こし文（ラベル）が必要である。「教師あり適応」とは人手で書き起こした正確なラベルを用いて適応を行う処理である。一方、「教師なし適応」とは、適応前の音響モデル即ちベースライン音響モデル（ＨＭＭ（隠れマルコフモデル）で表現されているためベースラインＨＭＭとも呼ぶ場合がある。）を用いて、一旦、音声認識を行った結果（基本音響モデル或いは）からラベルを作成し、そのラベルを用いて適応を行う処理である。各適応の決定的な違いは、適応処理の際に使用するラベルが正確であるか、否か（即ち誤りを含むか）の違いである。教師なし適応では音声認識結果をラベルとして用いるため、ラベルに誤りがあり、適応の精度は教師あり適応に劣るが、人手でラベルを作成する必要がなく、自動で適応処理を行えるという利点がある。

【００１４】

ここで、スポーツなどの実況中継を行う放送局などのアナウンサーの人数は限られているため、事前に該当するアナウンサー全員の音声データと、その音声データに対応する人手による正確な書き起こし文（ラベル）を用意することが可能である。このようにアナウンサー別の音声データおよび書き起こし文（ラバル）を用いて教師あり適応により、個々のアナウンサー毎に適応された音響モデル（基本音響モデル）を事前に用意しておくことが可能となる。このようにアナウンサー毎に適応された音響モデルを使用して、この音響モデルに対応済みのある特定のアナウンサーが発話した音声データの音声認識を行うと音声認識精度が向上する。

【００１５】

しかしながら、適応された音響モデルと実際に評価する音声では、時間差の問題により、微小ではあるが、話者性のミスマッチが生じているものと考えられる。また、収録時期や収録場所など収録環境によって観客の歓声等の周囲の雑音が変動するため、収録環境に関するミスマッチもまた生じるものと考えられる。本発明によるシステムでは、上述した話者性や収録環境のミスマッチを吸収即ち除去するために、このように事前に教師あり適応により適応された音響モデル（基本音響モデル）に対して、ハイライトシーン検出対象の入力音声を適応データとして、再度適応処理を行う。ここで、入力音声の書き起こし文を人手で事前に入手することは、実況中継の性質から不可能であるため、前述の基本音響モデルを用いて一旦当該入力音声を音声認識し、その結果（自動作成されたラベル）を用いて教師なし適応を行う。

【００１６】

図２は、本発明によるハイライトシーン検出システムの音響モデル作成手段で行われる音響適応の手順を示すフローチャートである。図に示すように、２段階の適応（ベースライン音響モデル→基本音響モデル→最終の音響モデル）を施すことによって、より高精度に

音声認識を行うことができる音響モデルを提供することができる。

【００１７】

中継音声の特徴としては、通常の読み上げの発話などに比べて発話速度が速い。特に、ラジオ中継では映像がないため、試合状況や臨場感を音声のみで伝達しなければならず、発話速度がより速くなる。これらの各種音声の特徴を比較するための表を示す。

【００１８】

【表１】

各種音声の特徴の比較

	発話速度 (mora/sec)	雑音レベル (SNR)	感情	発話スタイル
読み上げ音声	7.26	静か	平坦	読み上げ
講演音声	7.31	比較的静か	やや平坦	話し言葉
実況中継	8.51	騒々しい	起伏が激しい	話し言葉

表に示すような中継音声の特徴から、従来研究されている新聞読み上げ音声の音響モデルで認識することは困難であると考え、本発明では、講演音声からベースラインの音響モデルを作成し、それを各種データで適応させることにより精度の向上を図ることとした。

次に、本実施態様で使用する言語モデルについて詳細に説明するが、本実施例では、最も多いスポーツ中継である「野球」に適応させてある。しかしながら、本発明は、野球以外のスポーツ中継などにも対応できることは言うまでもない。さて、第１の適応の際は、言語モデルの作成のベースとなるテキストの集合にウェブ上から収集したテキストから不要な記号を取り除いたものを使用して第１の予備言語モデル（ウェブコーパスによる言語モデル）を作成する。この収集のとき、例えば野球中継のコンテンツを対象にしてハイライトシーンの検出をしたい場合は、野球に関するページを集めることで、より音声認識の精度を高めることができる。

第２の適応の際は、実際のスポーツ中継音声のアナウンサー発話を書き下したものを使用して第２の予備言語モデル（書き下しコーパスによる言語モデル）を作成する。それぞれの予備言語モデルに対して、形態素解析を行い（本実施例では、奈良先端科学技術大学院大学の「茶筅」という形態素解析ツールで形態素解析を実施した。）、ＣＭＵ－ＣａｍｂｒｉｄｇｅＴｏｏｌｋｉｔにより言語モデル・発音辞書を作成する。また、幾つかの野球用語に関しては、形態素解析時の辞書に追加し、１形態素となるようにした。

【００１９】

言語モデルの適応としては、ＭＡＰ推定によるものや、Ｎ－ｇｒａｍ出現回数の重み付き混合によるものなどが報告されている。本実施例では、長友他による「相補的バックオフを用いた言語モデル融合ツールの構築（情報処理学会研究報告、２００１－ＳＬＰ－３５－９）」に開示された融合ツールを用いてＮ－ｇｒａｍ言語モデルの重み付き融合を行う。

本実施例では、以下の３つの手法で３つの言語モデルを作成した。

（１）ウェブコーパスによる言語モデルと書き下しコーパスによる言語モデルを融合する手法

（２）ウェブコーパスと書き下しコーパスを結合したコーパス（結合コーパス）により言語モデルを作成する手法

（３）結合コーパスによる言語モデルに書き下しコーパスによる言語モデルを融合する手法（請求項３に相当）

なお、（２）および（３）において融合時の比率は、それぞれで最も低い単語正解率を出したものを与えた。上述した３つの手法で作成した各言語モデルを用いて、正解精度を確

かめるために予備実験を行った結果を表に示す。

【００２０】

【表２】

言語モデルに関する予備実験結果

	手法（１）の言語モデル	手法（２）の言語モデル	手法（３）の言語モデル
テストセット１	４８．６５	４９．４３	５１．８８
テストセット２	３３．８６	３６．０８	３８．４２

【００２１】

表のとおり、手法（３）で作成した言語モデルが最も高い正解精度を出した。上述した長友他の文献によると、手法（１）は手法（２）と同等もしくは手法（２）よりも劣る結果が出るとされている。しかし、上記文献では異なるタスクの言語モデルの融合に関しても実験が行われており、本実施例のような１つのタスク内での適応は試されていない。今回のタスクでは、野球に絞ってコーパスを作成したため、元々どちらのテキストにも含まれていた単語の割合が高かったため、相補的バックオフが有効に働いたことよりも、出現確率の重み付き混合が「話し言葉」である中継音声の発話スタイルをＮ－ｇｒａｍの出現確率でうまく表現したものと思われる。

【００２２】

本発明で対象とする中継音声には、プロスポーツ選手などの個人名が随所で発話される。言語モデルを作成する際のテキスト中にも多数の個人名が出現し、データスパースの問題によりその出現確率を言語モデル内で表現するのは困難である。そこで、本実施例では、音声に多く含まれる人名として、選手名と解説者名との二種類をクラス化することとした。これにより、過去の知識を用いる統計的言語モデルにおいて出現しない人名に関しても表現することができるようになった。

【００２３】

また、このようにして作成した言語モデルを用いた音声認識結果の中には、話者の発音変形によって起こったと思われる認識誤りが幾つか存在した。これを改善するために発音辞書内の発音表記に、幾つかのパターンを持たせた。また、ウェブから集めたテキストには存在したが、音声には含まれないであろう単語の発音を無音に置き換えた。以下の表にその例を示す。

【００２４】

【表３】

発音辞書内の表記例

言語モデル内の表記	出力記号	発音表記
ﾎﾞｰﾙｶｳﾝﾄ+ﾎﾞｰﾙｶｳﾝﾄ+2	[ﾎﾞｰﾙｶｳﾝﾄ]	Bo: ru ka u N
ﾎﾞｰﾙｶｳﾝﾄ+ﾎﾞｰﾙｶｳﾝﾄ+2	[ﾎﾞｰﾙｶｳﾝﾄ]	Bo: ru ka u N to
@+ｱｯﾄ+77	[@]	sp

【００２５】

例えば、野球中継のアナウンサーの発音では、「ボールカウント」は最後の音が極端に小さく発音される傾向が見られたため、そのような発音が精度良く認識されるような発音表記を加えた。一例ではあるが、表の「ボールカウント」の例では「ト」の発音が無いパターンの発音記号を加えてある。アットマーク（@）は、ウェブテキスト中多く見られる記

号であるが、話し言葉では使用される機会が非常に少ないと思われるため、無音の発音記号を割り当ててある。このように、プログラミングによる自動削除によって大まかに不要な部分は削除したが、不要な記号には無音を意味する発音記号「ｓｐ」を割り当てた。

【００２６】

次に、音声認識で抽出された単語のうちハイライトシーンを特徴付けるキーワードと合致するものを検索する。本実施例では野球中継の音声を対象としているため、下記の表に記載したキーワードを用いた。なお、野球以外の中継音声を対象とする場合は、別途それにふさわしいキーワードを用意することが好適である。

【００２７】

【表４】

キーワード一覧

ホームラン、２ラン、３ラン、満塁、グランドスラム、タイムリー、ホームスチール、先取点、先制点、追加点、駄目押し、押し出し

ここで表のようなキーワードを用いて、仮に１００％キーワードを検出したとしても、キーワードはハイライトシーン以外の箇所でも多数出現するものであり、例えば、ホームランを例に説明すると以下のような場合が想定される。

ある、打者がホームランを打ったとすると、当然アナウンサーはそのことを伝えこれが真の「ハイライトシーン」となる。しかし、実際にホームランを打ったシーンが終わってからも、そのシーンを振り返り「ホームラン」と発話する場合がある。打者が打席に入ったときに、「打率３割、ホームラン２０本」などのように打者に関する情報を伝える場合が多いため、このように実際のハイライトシーンではない時間区間でも「ホームラン」というキーワードが出現する。即ち、様々な場面において「ホームラン」というキーワードが多数出現するため、実際のハイライトシーンではない区間で多数検出されるという問題が発生する。

【００２８】

そこで、本発明によるハイライト検出システムでは、ハイライトシーンでは、アナウンサーが興奮し感情を込めて発話することが多く、そのような区間では音声の持つパワーが他の部分に比べて非常に大きいという特徴に着目し、それを利用して実際に生じたハイライトシーン以外のシーンを除去することとした。図３に、本発明のハイライトシーン検出手段における音声パワーの処理手順を示す。このような特徴を利用して音声認識結果と認識結果の各単語に割り当てられた時間情報を用いて、図に示すようにキーワード区間のみの音声を切り出し、単位時間（１秒）あたりの音声パワーを算出した。その後、音声認識により検出された単語のうちキーワードに一致するものであり、かつ、そのキーワード区間の音声のパワーが所定の閾値よりも大きい区間をハイライトシーンと判定する。例えば、「ホームラン」の閾値は、８０（デシベル）、「満塁」の閾値は６０（デシベル）というように予め設定しておく。本実施例では、図に示すように、「ホームラン」と発話している時間区間の音声パワーを計算し、それが、キーワード別に規定された閾値を超えるものを残し、それ以外のものを除去することで「実際のハイライトシーン」を高精度で検出するようにした。

【００２９】

本実施例で用いた音声データは、ＭＤから連続的に取り込んだ音声データを人手で切り出したものを利用している。下記の表にテストセットの概要を示す。

【００３０】

【表５】

テストセットの概要

	適応データ（予備音声データ）	評価データ（入力音声）
テストセット１	9/8 (A 球団 vs B 球団)	9/24 (A 球団 vs C 球団)
テストセット２	8/29 (A 球団 vs D 球団)	9/8 (A 球団 vs E 球団)

中継音声の発話スタイルは、「読み上げ音声」と比較すると、「話し言葉」のスタイルに近い。従って、ベースラインとなる音響モデルは比較的話し言葉の特徴に近い学会講演音声から作成した。また、ベースラインとなる言語モデルはウェブ上に存在する野球に関するテキスト（約５７万形態素）から作成した。

【００３１】

下記の表に音響分析条件とＨＭＭを示す。

【００３２】

【表６】

音響分析条件とHMM

音響分析	サンプリング周波数		16KHz
	特徴パラメータ		MFCC(39 次元)
	フレーム長		20ns
	フレーム周期		10ms
	窓タイプ		ハミング窓
HMM	タイプ		音節
	混合数		32 混合
	状態	母音(V)	5 状態 3 ループ
		子音(CV)	7 状態 5 ループ

表に示すように、音響モデルには長母音化を考慮した音節ＨＭＭ（詳細には、有木康雄他による「日本語話し言葉音声認識における話者間のための音節に基づく高精度な音響モデルの検討」（電子情報通信学会、ＳＰ２００２－１２９，ｐｐ４９－５４（２００２年１２月）を参照されたい。）を使用し、１状態あたりの混合分布数を３２とした。また、母音は５状態３ループ、子音は７状態５ループとした。サンプリング周波数は１６ｋＨｚ、音響特徴量には、１２次元のＭＦＣＣと対数パワーの１３次元、およびそれに１次微分、２次微分を加えた計３９次元である。

【００３３】

次に、本発明によるハイライトシーン検出システムにおける音声認識手段で使用する言語モデルの性能を評価するために音声認識実験を行った。ベースライン言語モデルにウェブコーパスで適応させたウェブコーパス言語モデルと、上述の予備事件で最も正解精度の高かった結合コーパスを用いて適応させた言語モデルと書き下しコーパスを用いて適応させた言語モデルとを融合させた融合言語モデル（請求項３の言語モデルに相当する）とを比較したものを下記の表に示す。

【００３４】

【表７】

言語モデルによる音声認識結果の比較

		単語正解率(%)	単語正解率(%)	単語正解精度(%)	Keyword(%)
ﾃｽﾄｾｯﾄ１	ｳｪﾌﾞｺｰﾊﾟｽ言語モデル	258.43	43.84	35.11	80.00
	融合言語モデル	75.82	58.32	51.88	80.00
ﾃｽﾄｾｯﾄ２	ｳｪﾌﾞｺｰﾊﾟｽ言語モデル	248.72	38.31	27.28	77.78
	融合言語モデル	69.16	49.23	38.42	70.38

表に示したＫｅｙｗｏｒｄ（％）は、表３で示したキーワードに関して、認識できた文章数をキーワードを含む実際の文章数で割ったものである。この評価から分かるように、ウェブコーパスのみによって適応させた言語モデルに比べて融合言語モデルの方は、単語正解率を下げると同時に、単語正解率、単語正解精度を向上している。キーワードの正解率については有意差を見つけることはできなかたが、挿入誤りについては向上が確認されている。なお、この実験では音響モデルには前記ベースライン音響モデルを使用した。また、本発明による手法で適応させた２段階適応音響モデルを使用した場合も同様の結果が確認された。

【００３５】

次に、本発明によるハイライトシーン検出システムにおける音声認識手段で使用する音響モデルの性能を評価するために音声認識実験を行った。比較のために、ベースライン音響モデルと、このベースライン音響モデルに対して本発明による手法で２段階の適応を施した２段階適応音響モデルとでデータを取った。実験結果を下記の表に示す。

【００３６】

【表８】

音響モデルによる単語認識結果の比較

		単語正解率(%)	単語正解精度(%)	Keyword(%)
ﾃｽﾄｾｯﾄ１	ﾍﾞｰｽﾗｲﾝ音響語モデル	58.32	51.88	80.00
	2段階適応音響モデル	78.63	74.58	90.00
ﾃｽﾄｾｯﾄ２	ﾍﾞｰｽﾗｲﾝ音響語モデル	49.23	38.42	70.38
	2段階適応音響モデル	72.82	63.77	96.30

ベースライン音響モデルに比べて、本発明による２段階適応音響モデルは最大３０％近い大幅な改善が見られた。これは、音響適応を用いることにより、アナウンサーにより近づいた音響モデルを作成することが出来たためと考えられる。

【００３７】

下記の表に本発明によるハイライトシーン検出システムによるハイライトシーン検出結果を示す。

【００３８】

【表９】

ハイライトシーン検出結果

	検出結果（正解検出数／正解区間数）	湧き出し誤り数
ﾃｽﾄｾｯﾄ１	２／２	2
ﾃｽﾄｾｯﾄ２	２／４	0

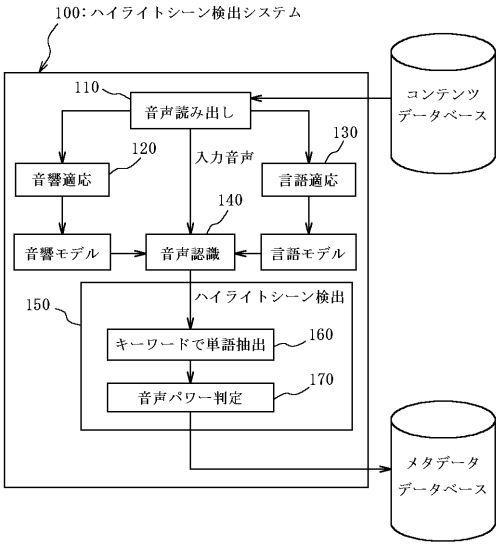
上の表に示すように、テストセット１では、データに含まれていた２つのハイライトシーンを２つとも正確に検出することができた。しかしながら、検出ミスである湧き出し区間数が２箇所存在している。この２つの湧き出しは、それぞれ以下のような要因により発生したものと推測される。１つは、キーワード区間音声パワーの閾値処理による湧き出しである。該当する区間では、音声認識によるキーワードの検出は正しく行われていたが、実際にはハイライトシーンではなく、選手の紹介を行っている区間であった。しかし、この湧き出し区間では、攻撃側チームの得点チャンスが続いていたため、アナウンサーが常に興奮して発話しており、大きなパワーを持った音声が継続していた。このような理由で、当該キーワードは、設定された閾値を超える音声パワーであったため誤ってハイライトシーンであると誤認識されたものである。

もう一方は、音声認識の誤りに起因する湧き出しである。該当する区間の音声認識結果を確認したところ、音声認識の誤りにより実際の音声には存在しないキーワードが湧き出していた。この湧き出し単語の音声パワーが閾値以上であったため、ハイライトシーンと誤

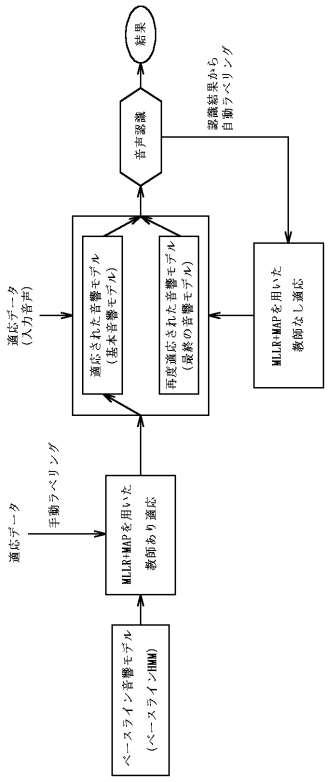
認識されたものである。
【００３９】
テストセット２では、湧き出し区間数はゼロに抑えることができたが、２つのハイライトシーンが未検出となった。１つの未検出区間については、キーワードは検出できていたが、音声パワーが閾値を超えなかったために発生した未検出である。もう１つの未検出区間については、音声認識の段階でキーワードが検出されなかったために発生した未検出である。従って、これらの未検出や湧き出しに対応してハイライトシーン検出の精度を向上させるためには、キーワード検出精度の改善、前記閾値の最適化や適応的な設定などを行う必要がある。
【００４０】
図４は、本発明によるハイライトシーン検出システムを映像コンテンツに適用させたシステムの概念図である。図に示すように、中継映像に含まれている音声データを、本発明によるハイライトシーン検出システムにおいて適応された音響モデルと言語モデルを用いて音声認識を行い、その後、キーワードを見つけハイライトシーンを検出する。配信用映像生成部では、入力された中継映像データに含まれる映像データをモバイル配信用の映像フォーマットに変換する。メタデータ生成部では、ハイライトシーンの映像検索に使用するためのメタデータを生成する。メタデータ生成の流れとして、まず、中継映像データと中継音声データとの時間的な同期を取り、それぞれ別系統のＰＣに取り込む。次に、映像解析ＰＣにより、入力映像を基本シーン（ピッチャーの投球場面でバックスクリーンから撮影されているシーン）ごとに切り出し、メタデータとしてその時の時刻情報をＸＭＬ形式のファイルに出力する。一方、音声解析ＰＣでは、本発明の手法に従い、入力された中継音声データを無音区間を基準に自動切り出しを行い、切り出した順に音声認識し、単語を抽出し、その単語のうち所定のキーワードに該当するものを検出する。検出されたキーワードは、出現した音声区間の始終端の時間情報と共にＸＭＬ形式ファイルに出力する。最終的に、映像、音声から出力されたＸＭＬファイルをデータベースの各層に登録する。データベースの内部は、ＸＭＬによって構造化されており、所望のキーワードを検索キーとして入力すると、対応するハイライトシーンを検索して、当該シーンの音声・映像を再生させることができる。
【００４１】
本明細書では、様々な実施態様で本発明の原理を説明してきたが、当業者であれば、本発明の開示に基づき、本発明の構成に幾多の修正や変形を施し得ることは自明であり、これらも本発明の範囲に含まれるものと理解されたい。例えば、本明細書では、本発明を主としてシステム（装置）として説明してきたが、本発明は、これらに相当する方法、その方法をコンピュータ上で実現するプログラム、当該プログラムを格納した記憶媒体の形態でも実施し得ることに注意されたい。
【図面の簡単な説明】
【図１】本発明によるハイライトシーン検出システムを説明する概念図である。
【図２】本発明によるハイライトシーン検出システムの音響モデル作成手段で行われる音響適応の手順を示すフローチャートである。
【図３】本発明のハイライトシーン検出手段における音声パワーの処理手順を示す図である。
【図４】本発明によるハイライトシーン検出システムを映像コンテンツに適用させたシステムの概念図である。
【符号の説明】
１００　ハイライトシーン検出システム
１１０　音声読み出し手段
１２０　音響モデル作成手段
１３０　言語モデル作成手段
１４０　音認識手段
１５０　ハイライトシーン検出手段

１６０　単語抽出手段
１７０　音声パワー判定手段

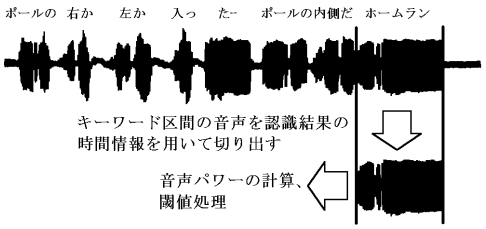
【図１】



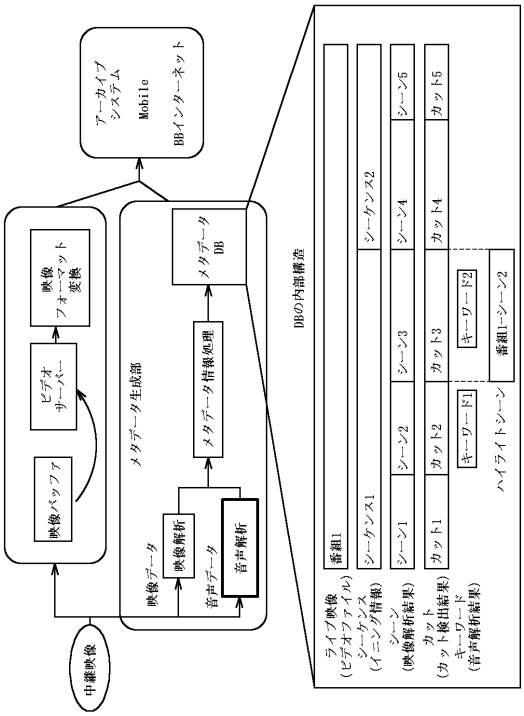
【図２】



【図３】


ボールの 右か 左か 入っ た- ボールの内側だ ホームラン
キーワード区間の音声を認識結果の
時間情報を用いて切り出す
音声パワーの計算、
閾値処理

【図４】


アーカイブ
システム
Mobile
BBインターネット
メタデータ生成部
映像データ
音声データ
映像解析
音声解析
メタデータ情報処理
メタデータ
DB
映像
フォーマット
変換
ビデオ
サーバー
映像バッファ
中継映像
DBの内部構造
ライブ映像
（ビデオファイル）
シーケンス
（イニシング情報）
シーン
（映像解析結果）
カット
（カット検出結果）
キーワード
（音声解析結果）
番組1
シーケンス1
シーケンス2
シーン1
シーン2
シーン3
シーン4
シーン5
カット1
カット2
カット3
カット4
カット5
キーワード1
キーワード2
ハイライトシーン
番組1-シーン2

フロントページの続き

(51)Int.Cl.[7]	FI	テーマコード（参考）
Ｈ０４Ｎ 5/91	Ｇ１０Ｌ 3/00 ５３７Ｊ	