映像検索システムの開発

音声認識技術を応用したビデオのキーワード検索

Development of a Video Retrieval System

Video Keyword Search using Speech Recognition Technology

（電力技術研究所　お客さまネットワークG　情報通信T）

ナレーション付き映像の内容をキーワードで検索できるシステムを開発した。本システムでは、映像中の音声区間に対し音声認識を行うことによって主要な単語に出現時刻を自動付与し、キーワードによるシーンの頭出しを実現している。

(Information and Communication Team, Customer Supply Network Group, Electric Power Research and Development Center)

We have developed a keyword search system that can retrieve video contents with narration. The system generates a time code index of important keywords in narration, using speech recognition, and realizes scene detection via keywords.

1 研究の背景と目的

広報など社内のビデオコンテンツを保有する部署では、デジタル化された映像アーカイブの構築が検討されている。大規模な映像アーカイブシステムの実現にはユーザ支援技術としてビデオ内容の検索技術が求められる。そこで本研究では、音声認識技術を応用した利便性の高い映像検索システムの開発を行った。

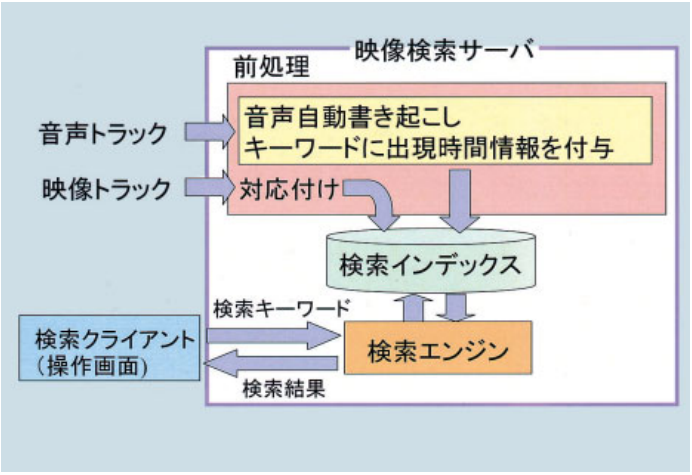
2 音声認識を用いた映像検索

映像のインデックス付与（検索のための索引付け）の手法としては画像処理に基づく方式が多数提案されている。画像処理による手法は、類似シーンの検出など人間の視覚的イメージに基づく検索には適しているが、これらの手法ではキーワードなど意味的な概念によるインデックス付与が難しい。もし対象となる映像が音声トラックを有していれば、これを手掛りに音響処理によって意味的な単語によるインデックス（単語の出現時刻情報）を生成することが可能となる。本システムでは、画像処理は用いずに、音響処理（音声自動書き起こし）のみで映像検索を実現する方式について検討を行った。

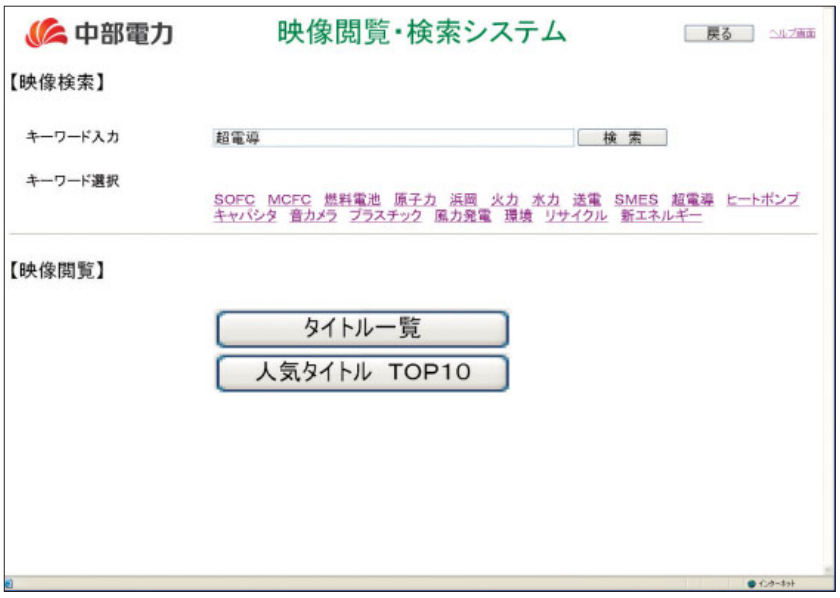
3 システム概要

本システムの概要を第1図に示す。システムは映像検索サーバ（Linux PC）と検索クライアント（Windows PC）より構成され、検索インタフェースにはWebブラウザを用いている。システムの検索画面を第2図に、検索結果の画面例を第3図に示す。検索結果の画面では検索キーワードが発話されたシーンのサムネイル画像を表示している。

キーワードの入力手段は通常のタイピングによる方式に加え、あらかじめシステム側で選定したキーワードを画面上に提示する方式も実装した（第2図）。

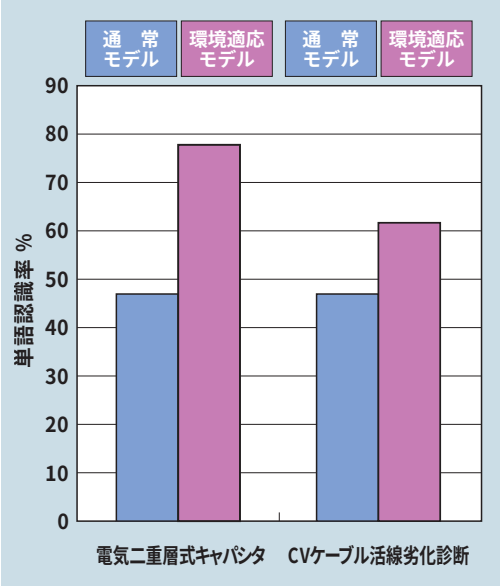


第1図　システム概要



第2図　検索画面

第3図　検索結果画面の例（検索キーワード：超電導）



第4図　認識性能の評価（単語認識率）

4 背景音楽付き音声の認識性能向上

一般にビデオコンテンツでは音声トラックに背景音楽（BGM）が重畳していることが多く、これが認識性能を劣化させる大きな要因となっている。本研究ではこの問題に対するアプローチの一つとして、音響モデル（母音や子音などの音響的照合パターン）の背景音楽に対する環境適応を行った。環境適応の概略手順を以下に示す。

（1）クリーンな新聞読み上げ音声にクラシックやジャズの音楽を重畳し、ベースとなる環境適応モデルを学習する。

（2）上記で学習されたモデルに対し、登録コンテンツの少量の背景音楽付き音声を用いて環境適応を行う。環境適応にはMLLR法を用いた。

5 システムの有効性評価

音声認識性能の評価として、約3分の研究紹介ビデオ2タイトル（学習や適応に用いていないオープンデータ、背景音楽付き）を用いた実験を行った。言語モデル（語彙や文法知識）は当該研究の「技術開発ニュース」の原稿やプレスリリース等のテキストより学習したものを用いた。単語認識率を第4図に示す。なお単語認識率の定義は、認識結果の単語数をN、置換誤り数をS、脱落誤り数をDとした時、次式で定義される。

$$\text{単語認識率}\ (\%) = \frac{N - S - D}{N} * 100$$

第4図より「電気二重層式キャパシタ」のデータでは環境適応手法によって、通常のクリーンな音声から学習した音響モデルを用いた場合から30.7%の性能向上が得られ、「CVケーブル活線劣化診断」のデータでは14.0%の性能向上が得られた。通常の音響モデルと比較して環境適応モデルの効果が高いことがわかる。

第4図の結果は全ての品詞を含む文としての書き起こし精度であるが、実際に映像検索で用いられるのは主に名詞情報である。文としての書き起こしが完全でなくとも、検索という目的に対し実用上は十分な性能が得られていると考えられる。

また、検索システムとしての有効性であるが、システムを実際にユーザに試用させた結果、以下のような評価が得られた。

・映像シーンのキーワードによる頭出しは利便性が高く、ビデオ閲覧の稼働率も高まる。
・専門分野に詳しくないユーザにとって、キーワード提示型の検索インタフェースは有用である。

6 今後の展開

本システムはH19年3月より電力技術研究所本館の展示コーナーに導入され、技術開発本部の研究成果のPRに活用されている。今後は社内のPR施設等への技術展開も検討していきたい。

執筆者／瀬川　修
Segawa.Osamu@chuden.co.jp