NetTv：NetNewsとテレビ放送のクロスプラットホームにおける音声検索*

○田中克幸,滝口哲也,有木康雄(神戸大学工学部)

１．はじめに

情報網・Web２．０の発展や放送のデジタル化により、情報整理が困難なメディア、映像、画像、音響などの普及が、情報の氾濫を招いている。情報量の爆発とプラットフォームの多様化により、ユーザーが欲しい情報が入手できない状況にあり、効率的にユーザーが欲しい情報だけを入手できる環境が必要とされてきている。本研究では、NetNews 動画と TV 映像上において、クロスプラットフォームの動画インデキシングと音声インタフェイスを用いた検索システムを構築した。すなわち、ユーザーが快適に動画を観覧でき、疑問を解決できる NetTv システムを構築し、情報の統合によるユーザーの検索負担軽減を目指した。

２．ＮｅｔＴｖの概要

本研究では、まず、プラットフォームの統合と（これを、"マルチセマンティックメディア空間"と呼ぶ）、日々変化するダイナミックな環境において音声認識を実現することを目的としている。そのため、インターネット、TV 放送などの映像メディアに対して、以下のような機能を円滑に行える環境の基盤作りに重点を置き、システムを構築した。

- NetNews と TV 映像のクロスプラットフォームにおける動画インデキシング
- 映像のメタ情報の自動付与
- 音声認識のための辞書を自動生成し、情報ドメインに流動性をもたせ、音声インタフェイスによる検索と質問応答機能を実現する

このシステムは、２つのモジュール、NetNews と HddTv からなる。

２.１. NetNews

NetNews はネット上に流れるニュースを、周りのテキスト情報をメタ情報として付与しニュースをインデックス化することにより、音声検索可能にする。

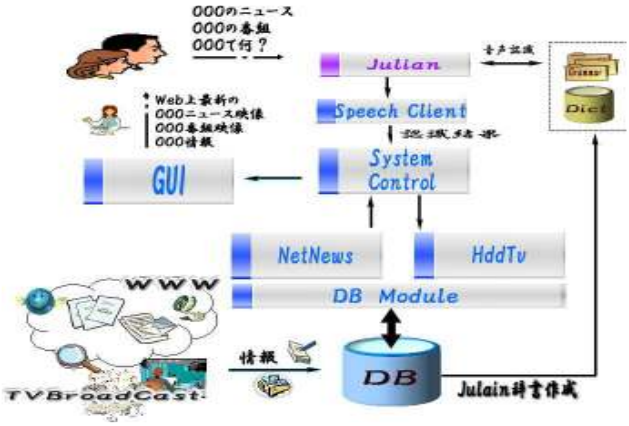

Fig 1 NetTv 概要

ニュースのダイナミックなコンテンツ変化に対応できように、リンクと詳細記事を自動的にニュースサイトから集めてくる。集めた各詳細記事ページに html パーザをかけ、記事部分と動画部分を切り出し、詳細記事とヘッダーを文単位に分解して茶筅[2]に掛け形態素解析を行う。この結果 url、動画、ヘッダー、詳細記事の単語に関するインデックステーブルが作成され、キーワードによる検索が可能となる。“Video Grammar”を使って映像をシーン別に摘出し、メタデータを付与するシーンインデックス化[3]という方法を行うことも可能であると考えられるが、オンラインのコンテンツを解析するのは難しいため、的確に動画コンテンツの内容を示しているニュース記事は、動画のインデックスに最適であると考える。

２.２. HddTv

HddTv はユーザーが見たい TV 番組を HDD 型の映像録画環境で視聴できる環境を想定し、ネット上にある EPG をメタ情報として、映像に付与し、映像をキーワードによりインデックス化し、音声インタフェイス用の辞書を自動的に作成して、音声検索可能にする。

録画した TV 番組に対する関連情報検索のために、EPG を自動的に入手し、人名、番組名、日時等のテーブルを作り、それらをリン

*NetTv： Multimedia Cross-platform video retrieval with speech interface by K.TANAKA T.TAKIGUCHI and Y.ARIKI(Kobe University)

クさせDBを作成する。録画したTV番組のタイトルを形態素解析にかけ、記号以外の品詞の連続を動画検索のキーワードとし、録画映像に付与する。これより、EPG 上の出演者、日時、番組名などをもとに情報検索することが可能になる。付与されたEPG内のキーワードと先１週間迄のEPGをマッチングすることで、その出演者が出演する他の番組などの情報を、録画番組に自動的に付与することができるため、PCやTVの前に座って、ユーザーが検索をかけて探すという負担を軽減できる。

2.3. 音声インタフェイス

NetTvでは、各モジュールで作成したDBから音声認識用の辞書をドメインごとに自動生成し、ユーザーがドメインを変えるたびに、また、ドメインの内容が日々更新されるたびに、音声認識用の辞書をダイナミックに変化さることができ、多様なドメインの最新情報に関する柔軟なドメイン遷移を実現している。情報が日々更新される環境では、本手法がより効果的であると考える。音声認識にはJulian[1]を使用しグラマーを設定することで、“ＯＯＯのニュース”、“ＯＯＯの番組”、“ＯＯＯの出演番組”といった発話による検索を可能としている。

３．実験

10人（男９女１）にシステムを使ってもらい、ユーザーのシステムへの全ての発話をユーザー発話、ユーザー発話のうち検索を行うものを検索要求発話と定義し、NetTv の評価を行った。評価方法と定義を以下に述べる。

- **タスク成功率：**ユーザー発話を認識して意図したタスクが成功した割合を表す。
- **検索成功率：**検索要求発話に対して検索が成功した割合を表す
- **満足度：**５段階でシステムの使いやすさや満足度を表示。１．不満、２．やや不満、３．普通、４．やや満足、５．満足

3.1. 実験結果

１１７ニュース（６９１８語彙）と１８番組（１４０語彙）の環境下で実験を行った。結果をTable1-3に示す。

１回の発話に対する音声の正確な認識は語彙の多いNetNewsでは難しかった。タスク、の２回までの言い直しにより成功率は約７０％の精度で行うことができた。

Table. 1発話回数とタスク成功率・検索成功率

発話回数	タスク成功率	検索成功率
１回	６３％	４８％
２回	６９％	５７％
３回	７０％	５７％

Table. 2ユーザー満足度

満足度	1	2	3	4	5
人数	1	3	4	1	1

Table. 3発話訓練者による成功率

発話回数	タスク成功率	検索成功率
１回	８５％	６７％

検索の成功率は、データベース上にないキーワードの発話を除くと、約６５％となる。音声検索した結果、検索トピックにそぐわないものは出てこなかった。

満足度については、ユーザーの発話の繰り返し回数に満足度が左右される傾向にあった。

認識精度の問題点としては、マイクと口元の距離が離れ過ぎ、文法の規則通りに決まった言い方をしないと認識できない、辞書にない未知語を発話すると誤った単語に認識されてしまい誤作動を起こす、などが考えられる。発話訓練されたユーザーがシステムを使用するとTable3のような結果になる。

４．おわりに

本研究では、スピーチインタフェイスにより、ユーザーが動画や動画に関する情報を検索することができるシステムを構築した。今後、多様な面でユーザーのユーズにそった検索や、質問可能なシステムが構築できるよう、検討を行っていく予定である。

参考文献

[1] 河原達也, 李晃伸, "「連続音声認識ソフトウエア Julius」",人工知能学会誌, Vol.20, No.1, pp.41- 49, 2005

[2] 松本裕治,"形態素解析システム「茶筅」", 情報処理, Vol.41, No.11, pp1208-1214, 2000

[3] 天野美紀, 上原邦昭, 熊野雅仁, 有木康雄, et al,"映像文法に基づく映像編集支援システム"情報処理学会論文誌, Vol.44, No.03, pp.915-924, 2003