NetTv: Cross-Platform Video Retrieval and QA System with Speech Interface

Katsuyuki Tanaka.[1], *Tetsuya Takiguchi* [1], *Tasuo Ariki* [1]

[1] Graduate School of Engineering, Kobe University, 1-1 Rokkodai, Nada, Kobe, 657-8501, JAPAN
katsutanaka@me.cs.scitec.kobe-u.ac.jp, {takigu,ariki}@kobe-u.ac.jp

Abstract The objective of this research is to construct a video searching mechanism and speech interface on the multimedia cross-platform, namely TV and Internet, which requires the capability to deal with dynamic contents. Current NetTv enables users to search both recorded TV contents and news on the Internet by simply speaking keywords as a query; hence the videos related to the keyword spoken are retrieved. Also, the system provides a simple keyword based QA system to answer various questions that may occur to users whilst watching retrieved videos. In this way, NetTv improves the usability of video searching and viewing in a hands free way.

Index Terms: Speech Interface, Video indexing and retrieval, QA System

1. Introduction

Recently, information technology has become increasingly advanced and its dramatic improvements have opened a new world of information to us. Internet technology, such as Web2.0, BB, digital television (DTV) is one such example which enables our access to more information, both in terms of volume and content.

In spite of the advances in information and broadcasting technology, due to the rate of information growth, users are having difficulty in finding the information that they are really interested in, especially in terms of video contents. From such a vast amount of video information with almost no tagged contents descriptions, it takes a huge amount of time and effort to filter this information to find the specific video contents which a user really wants to watch.

When it comes to the interface of interaction between humans and machines, using a mouse, keyboard or remote control has not changed for decades. Even though speech interface maybe the candidate for next generation interface, it has a difficulty of adapting to environments where information changes frequently. Besides, the way of accessing information through different resources on different platforms, which are WWW and TV in this case, is not very natural since the concept of DTV technology has become more and more WWW like. Moreover, whenever users wish to obtain extra information while watching internet videos or/and TV, users are forced to carry out search on the internet by typing a query and clicking on links to find the answer.

In this project, we introduce the concept of combining TV and Internet on the same platform and accessing video or co-relating information under a hands free environment. We built a NetTv system to improve user accessibility and the usability of video based information and its contents information by taking the following aspects into account:

- WWW and TV are considered as one multimedia space since both contain video streaming and data. It merely differs by transmitted sources.
- Manual keyword tagging of video is not rich enough to express video contents. NetTv explores automated video indexing to provide a richer description.
- NetTv generates a speech dictionary automatically and introduces dictionary switching. This method is more suitable to perform speech recognition on dynamic domains.
- Answering various questions users may occur while watching video retrieved.

In this paper, we introduce the architecture of NetTv and we focus on developing speech interface on dynamic world.

2. Previous Work

There are large fields of research on improving the effectiveness of searching on the internet from image indexing using surrounded text on image [1, 2, 3] to user profiling and query expansions to analyse a user's needs [4, 5]. It becomes difficult to have proper information mining when it comes to multimodal data like video contents since it contains a combination of rich uni-modal data, namely the sequences of images and audio. In fact, major search engines like Google [1] and YouTube [6] are relying on manual tag of metadata by users when videos are uploaded. They do not even have automated collection of video. Nevertheless, this method is very time consuming and requires a significant amount of labour as the number of uploaded videos increases.

Speech interface is considered to be one of the popular methods for next generation user interface, for its ease of use as it offers a hands-free environment. A lot of research has been done on speech recognition [7, 8, 9]. However, the current approach lacks the flexibility of adapting to dynamic worlds like WWW and DTV. Typical systems with speech interface are usually conducted under predefined domains [7, 8]. Even though there is research on constructing a large vocabulary dictionary using the web [9], collecting information and maintaining the dictionary is time and resource consuming. Besides, this method has a trade off in terms of its recognition rate and there is a doubt that such a large vocabulary is required on practical speech recognition.

3. NetTv Overview

NetTv is a video and video related information retrieval system with speech interface. It aims to visualise two platforms, namely WWW and TV, as the same multimedia space to aid users for more effective way of searching and viewing of multimedia data in video form. The main purpose of the system is to automate indexing of any form of videos on TV or internet by utilizing internet power as their contents information source and developing speech interface. Henceforth, the term 'video clip' is used in this paper to refer to any media containing sequences of images as moving objects, including TV programs, internet streaming and videos etc.

•

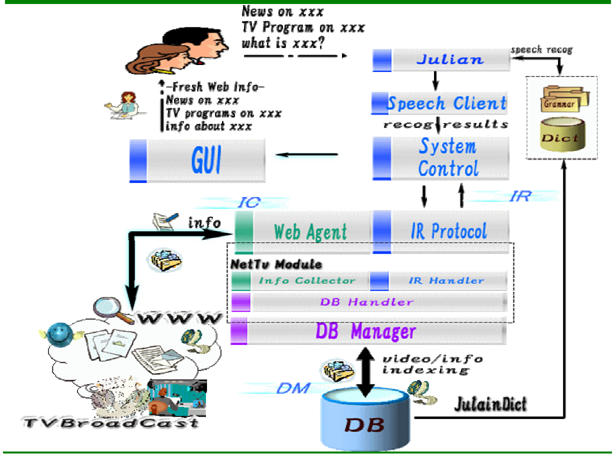

Figure 1 NetTv Overview

NetTv system can be broken into four major components according to the nature of information control, Information Collection (IC), Data Management (DM), Information Retrieval (IR) and Module (MO) which shares three components. The architecture of NetTv is shown in Figure1. The Information Collection component is responsible for all information gathering from WWW and/or TV broadcast. The term "collected data" refers to any form of data collected in IC henceforth. Subsequently, the Database Management component, manages collected data in IC. Information the Retrieval component deals with bridging mechanism between users request and DB, so that users could retrieve various types of information. Module is a pluggable component, which defines behaviours of the system in domain dependent matters. It defines rules of system to behave in particular way by manipulating IC, DM and IR components.

The detailed implementation of the IC, DB, IR and MO are explained in next sections.

4. NetTv Architecture

4.1. Information Collection

Collecting information is one of the major roles of NetTv. The Information Collection component accounts for collecting video clips and information from web pages, especially associated text information for acquired video clips.

The Information Collection component is responsible for all information gathering from WWW and/or TV broadcast, such as video clips, its associated text and any form of text information to support user's search on multimedia data throw NetTv.

Web Agent is literally an agent to collect web pages from WWW by following links automatically using Html Parser. Information extractions are done by a domain based mechanism using NetTv Module's Info Collector since information extraction may differ on different domains. Each module implements Wrapper for its domain in Info Collector to cooperate with Html Parser and Web Agent, thus it could acquire or extract information, such as video clip links and its associated text information from collected data by Web Agent for the domain. Output of IC is a location of video clips, related information for acquired video clips and/or text information that may aid user's search on video clips. These are called “extracted data”.

4.2. Database Management

Subsequently, the Database Management component manages extracted data such as collected video clips and its information in IC. It performs automatic indexing of video clips and text information into storage (database or DB in short). Module’s DB Handler defines a domain based way of handling DB since different domains have different data structures. Hence after, DM Manager stores the extracted data in DB. It also handles domain dependent way of data fetching on requests from IR Handler.

DM also has the responsibility for building a language dictionary for speech recognition system from constructed DB. Thus, it enables an automatic constriction of language dictionary to adapt speech recognition to dynamic worlds.

4.3. Information Retrieval

The Information Retrieval component is responsible for handling request from users and obtaining appropriate answer for the request. NetTv uses graphical user interface (GUI) and speech user interface (SUI) as a retrieval method. It uses DB constructed in DM from collected information in IC. When users make a request to NetTv by speech, speech recognition is performed in the speech by Julian. NetTv employs Julius-3.5.2 [10] complied in Julian mode for speech recognitions. Julius is continuous speech recognition decoder software for large vocabulary. After speech recognition is performed, the requests in xml format are forwarded to Speech Client in where information is extracted by xml parser. From extracted results, System Control forwards the requests to appropriate module's IR Handler through IR Protocol and waits answers for the requests. System Control is a part of the system controlling all interaction between users and the system and decides appropriate actions. Each module implements IR Handler following IR Protocol pre-defined to control requests. Grammar of speech should be defined in IR Handler in each domain so that it could recognise the nature of requests. SUI also follows this grammar for speech recognitions. IR Handler makes appropriate actions of extracted requests and fetches information from DB by communicating with DB Handler. Finally, answers are fed back to Control System hence displaying the answer to the users.

4.3.1. Speech Interface

It is fundamental to design a speech interface to achieve high performance in speech recognition in the dynamic world in this project. Stereotypical methods of fixing domain and manual construction of dictionary are not acceptable in NetTv since news and TV programs change frequently. In addition, using a large vocabulary as a dictionary may have lower recognition rate and it is costly to maintain the dictionary. Therefore, this research proposes to produce language dictionary in different domains automatically and provide a capability of switching task domains flexibly by superimposing the required domain. In other words, we do not try to gain better speech recognition performance with neither large vocabulary dictionary nor small vocabulary dictionary from domain restricted method, but we design to create middle size dictionary in different domains and switching them to provide both keyword coverage and recognition rate.

Julian performs a grammar-based recognition by defining patterns or syntaxes of word sequences or sequences of classes containing list of words on given task in grammar. Julian searches defined grammar for the most likely sequence match. IR systems have greater discriminate power using grammar than just a sequence of words.

It keeps separate domain information and switching dictionary to move domains. This enables users to make various queries not merely title keywords such as “what is <xxx>?” and “news/TV program on <xxx>”. Speech Client extracts this information from the results. IR system uses this result for information retrieval. Figure2 shows the Speech Interface description.

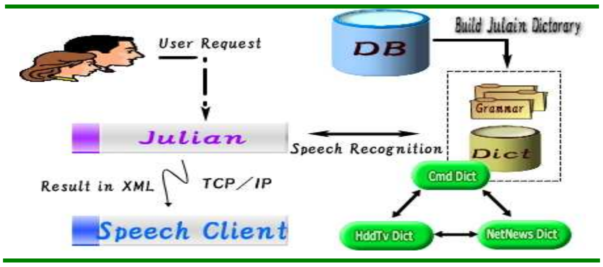

Figure 2 Speech Interface

4.4. Module

Module is a pluggable component of NetTv, defining a domain based rules or policy of data collection or extraction, data representation, and information retrieval in domain dependent matter. Module consists of three parts, Info Collector, DB Handler and IR Handler, which inherit the characteristics of three complements of NetTv, IC, DM and IR correspondingly. Module is the part which manipulates each component of NetTv to define behaviours of NetTv.
Info Collector shares IC of NetTv and uses Web Agent as its interface to collect information from WWW. A method of extracting video clips and associated text from collected information should be defined here to have appropriate IC mechanisms.

DB Handler is a DM part of NetTv. It defines rules for managing collected information and representation of data. Video clips, associated text and any other information to support user's search are stored in DB throw DB Manager.
IR characteristics of NetTv were inherited in IR Handler in Module. IR Handler is in charge of decision making and fetching necessary data from DB by carrying out request analysis on user requests. Thus appropriate rules, such as grammar rules for speech and which information to be retrieved, should be defined in IR Handler to retrieve information from DB. IR Handler communicates with IR parts of NetTv and performs information exchanges, which include receiving request and sending result back, throw IR Protocol.

We implemented the system to build pluggable domain based modules so that it could restrict the domain of the user's usage of multimedia data. A module should be implemented to follow the pre-defined protocol or adapt interface of IC, DM and IR to communicate with them. In other words, Module contains characteristics of three components of NetTv. This architecture also provides extensible features of the system by building a plug-in for new domains. It also provides a more effective speech interface since it could superimpose the target speech vocabulary for speech recognitions. We believe that this architecture is better for speech recognition too.

5. NETNEWS AND HDDTV IMPLEMENTATIONS

For experimental purpose, we build two modules, NetNews and HddTv by applying above architecture.

5.1. NetNews

NetNews is a module enabling users to watch news clips and associated articles on the internet by a speech driven query search.

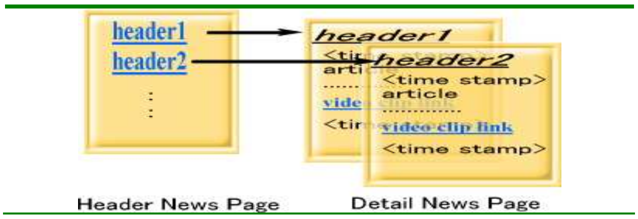

Figure 3 News Sites Layout

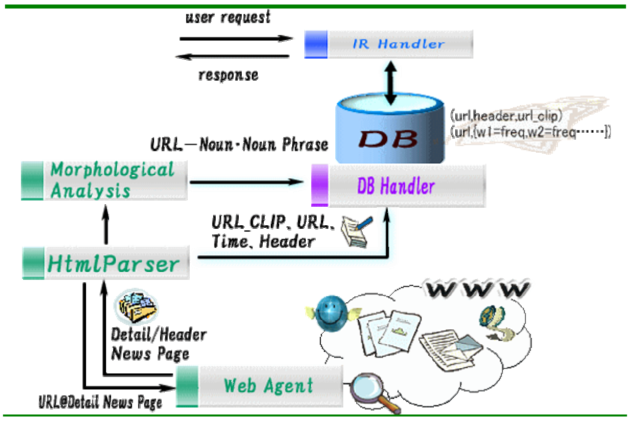

Figure 4 NetNews Module

NetNews gathers information from news sites offering video clips and articles. A typical news website has "header news page" containing a list of news headers linked to its detailed news page. In the "detailed news page", it contains its article and video clip with time stamp. A typical News sites layout shows in Figure3. The procedure of NetNews module proceeds as follows:

1) Web Agent fetches "header news page" from a news site containing list of news headers from WWW.
2) Html Parser extracts URL links from the header news page for next crawl to fetch the detailed news page.
3) For all URLs found in (2), the web spider collects all the detailed news pages.
4) A collected page is parsed by Html Parser to extract actual article of the news, news header and URL for video clip link.
5) A Morphological Analysis is performed on the article and the header extracted. They are broken down into a sequence of morphemes. This process mines nouns and calls these keywords. Also,

sequences of consecutive nouns construct noun phrases and these are also considered keywords.

6) Finally, a DB is constructed by using these keywords extracted from detailed news page as video indexer and augmentation.

Repeat steps 4-6 for all detailed news pages collected from a header news page to obtain keywords. By following the above procedure for all header news pages could construct fully automated video indexing on news.

All requests are made by speech as mentioned. Possible services that NetNews offers are:

- On requesting "news on <xxx>", responds with video news on keyword <xxx> and plays the clip.
- A "detail article" request displays the detailed articles on clip played at the moment.
- Requesting "next/previous clip" forwards/backwards to next/previous clip if there are any.

5.2. HddTv

HddTv also has a similar mechanism to NetNews. This module offers searching and viewing of TV programs that a user has recorded on HDD based environments by speech. Instead of collecting video clip, HddTv assumes that the video clip is already collected by users by recording their favourite TV programs on HDD. Thus, instead of collecting clips from internet, HddTv automatically collects information of the video clip from internet, that is EPG and builds a video indexing database.

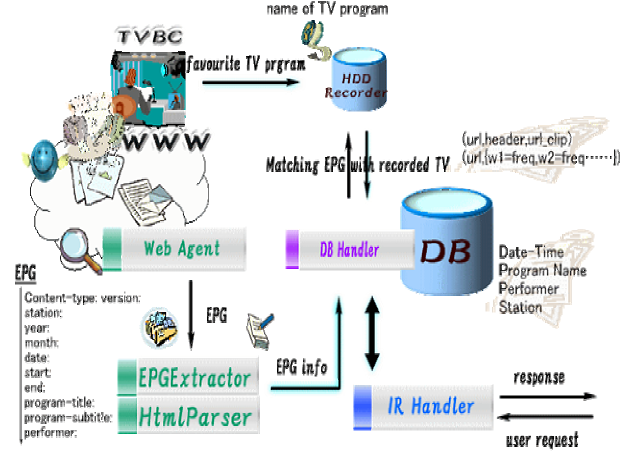

Figure 5 HddTv Module

HddTv focuses on constructing video indexing from recorded video clip and its EPG. This phase makes good use of internet resources to collect EPG information. The flow of the collection and DB construction is explained below.

1) Web spider collects a web page on "date X" containing a list of TV programs from internet (EPG list page).
2) The web page is parsed by Html Parser and extracts URL links to actual EPG.
3) For all links found in 2), spider collects EPGs and EPGExtrator mines detailed information, for example date, time, program name, performer on the program etc, about a program regarding to the EPG.
4) Build EPG DB to facilitate user search on program name, time/date, performer and various another ways.
5) By synchronising recorded video clip and EPG, construct index table for the clip.

The Web Spider needs to crawl through the internet occasionally to collect EPG, since the information about programs is only released weekly in advance, by repeating the above procedure on a specified date. At the moment only the title of the TV program is used as an index keyword for video clips. A morphological analysis is preceded like NetNews but keywords are choices as sequence of all morphemes that appear on title. This should be expanded to the "program_subtitle" field on EPG to accommodate flexibility in a user query and the user's needs in the future, thus enables richer augmentations of video.

Users are allowed to have request according to the protocols defined and the module handles them to make the right response. The chart below shows the available services on HddTv for speech information retrieval. It is also possible to ask question such as "who is <xxx>" about the performers in the program in this module.

- “program <xxx>” plays TV program with name <xxx> and list of the video clip sorted by the date.
- “display EPG” shows the EPG on viewing video clip.
- User could ask question such as "who is <xxx>?", "what is <xxx>?".
- User could search for appearance of favourite performer on the TV program currently watching by saying "List program <xxx> appears". The name of TV programs listed to users that the requested performer scheduled to appear on the show for next one week.

5.3. QA System

There are often situations when a user would like to know more details about some terms or events mentioned in news. In such circumstances, NetTv also provide QA like mechanism to answer various questions. This is achieved by queries like "what is <xxx>?" where <xxx> is the keyword that the user wishes to have explained. This system works like a mini encyclopaedia trying to solve various questions that users might face during news viewing. The keyword must be in the index table before a user could make any request, otherwise speech recognition would not be able to detect the keyword. At the moment the answer to the questions are forward to online encyclopaedia, Wikipedia [12]. In the future, this function may be refined to make better question and answering mechanism. Proper speech interfaced QA System is one of the main our major goal and it is left as our feature works.

6. Experiments

It is impractical to consider the whole domain of internet or TV as a source of gathering video clips for experiments since it is time and resource consuming. Therefore, we implemented two modules NetNews and HddTv for news on internet and user's recorded TV broadcast for experiment.

NetNews is a module enabling users to watch news clips and associated articles on the internet by a speech driven query search. NetNews gathers information from news sites offering video clips and articles.

HddTv also has a similar mechanism to NetNews. This module offers searching and viewing of TV programs that a user has recorded on HDD based environments by speech. Instead of collecting video clip, HddTv assumes that the video clip is already collected by users by recording their favourite TV programs on HDD. Thus, instead of collecting clips from internet, HddTv automatically collects information of the video clip from internet, that is EPG and builds a video indexing database.

One of the purposes of our research is to build a QA system to answer various questions which may occur while they are watching video clip. Current QA system only deals with factoid-based question, that is “what is <keyword>?”, and query is only forward to Wikipedia [12] at the moment. Construction of proper QA system with non-factoid type question is left as future works.

Table. 1: #Speech on Task & Retrieval Success rate

#Speech Tried	Task Success	Retrieval Success
1	63%	48%
2	69%	57%
3	70%	57%

Table. 2: User's System Evaluation

Satisfaction measurements	1	2	3	4	5
#users	1	3	4	1	1

Table. 3: Trained User's Task and Retrival Sccess Rate

#Speech	Task Success	Retrieval Success
1	85%	67%

6.1. Evaluation Methods

Evaluation of the system is measured by result of success of the system response on user's requests and feedback in terms of their sufficiency of the system usage and performance. Experiments are performed with 10 users (9 male, 1 female) who used NetTv under the condition of 117 news (6918 keywords) and 18 TV programs (140 keywords). Each user made 10 queries to the system. All Morphological Analysis are done using ChaSen [11].

The term “user speech” is defined as all requests made by user to the system, and “retrieval speech” as the user requests regard to information retrieval, which is requesting for video clips by keyword. Using these terms we evaluated NetTv’s success rate in tasks that the user requested as follows:

- Task Success: the rate of succeeded response made by system against “user speech”.
- Retrieval Success: the rate of succeeded retrieval response made by system against "retrieval speech".
- Satisfaction Measurement: users feedback of how much users satisfied with NetTv described in 5 levels, 1. unsatisfied, 2. mildly unsatisfied 3.normal, 4.mildly satisfied, 5. satisfied.

Table1-3 shows the results of the experiments.

6.2. Evaluation

Table1 describes different numbers of speech tried by a user before the task or retrieval succeeded. Here, systems may not make the right response to the user's request due to the nature of speech recognition technology. Hence, “number of speech tried” is the number of times a user made the same request to the system. Task success and Retrieval success is also measured for speech trained user and its results stated in Table3.

Results show that the systems achieved about 70% of task success rate on user's request to the system if speech repetition once is tolerable for users. Due to the limitation of the speech recognition, there are words that are not very easy to decode.

NetTv accomplished a retrieval success of about 57% on two tries. The rate improves to 65% success if we omitted any user's query where keywords did not exist in the DB.

The system could not gain high user's satisfaction rate. It seems that users have a tolerance of repeating their request to a maximum of three times. The number of repeating requests is a factor to gain user's support for the system.

The number of problems could be raised to improve the performance of the system. One of the reasons for the low rate of the task success may be due to the lack of instruction to the participating users for speech. While Julian introduces flexibility in terms of what to say, it restricts how to speak under grammatical constraints. This prevents users from using spontaneous speech which in turn prevents users from making proper requests to the system. By training users in speech grammar and/or how to use the system more effectively, NetTv could successfully perform about 85% of user's requests. The retrieval success rate improves to 80% if we discount any user’s query of keywords that do not exist in DB. This shows that a few instructions in how to speak or how to use the system could achieve higher performance of NetTv. If this is the case, users may not have to make repeated requests, thus, it is possible to expect the user sufficiency of NetTv would increase.

7. Conclusions

In this project, we introduced the architecture of NetTv, which enable to search video clips in and view them by speech interface, with two modules NetNews and HddTv. We argued that a more effective and natural way of accessing multimedia data from TV broadcasting and internet is to combine two platforms in one multimedia space. We believe that this proposal assists users to improve their information retrieval since it prevents unwilling platform crossing when they, for example, like to watch TV and/or videos from the WWW. Additionally, indexing of the multimedia information becomes increasingly important, thus, using web resources definitely serves the purpose of multimedia augmentation. Automatic dictionary generation and domain dictionary switching for speech recognition is also introduced in this paper. Flexibility and automatic dictionary generation is very important in adapting dynamic worlds for speech interface to become a trend of next generation user interface.

Future work on NetTv research will mainly focus on improving usability of the system and video contents enrichment. Our interest in future research lies in building intelligence to understand the contents of video clip by using such technologies as image, speech as well as natural language processing. It may be possible to produce richer description of video clip contents by utilising web power or extract speech from video clip using expanded knowledge. Also, from collected data, it is possible to deduce possible questions that users may have while watching video clips and to find solutions for these in advance from internet resources to build proper QA system.

Multimedia information retrieval will take a very big part of next generation information age, hence, a rich uni-modal and multi-modal analysis are an essential technology to extract information from multimedia data. Understanding multimedia contents and video contents expansion or augmentations is one of the aims of our future research.

8. References

[1] Google. http://www.google.com

[2] Fujimoto, N., Hagihara, K., Idehara, H., Takeno, H. A Sentence Extraction Technique Based on HTML Parsing Three Structures around Images for WWW Image Retrieval. IEICE Technical Report, DE2005-136, pp. 19--24, 2005.

[3] Zettsu, K., Kastumi, T. Extraction and Visualization of Image Contexts from Web, DBSJ Letters, Vol.2, No.1, pp.99-102, May 2003

[4] Doi, T., Honiden, S., Niwa, S. Web Page Recommender System Based on Folksonomy mining. IPSJ, Vol47, No5, pp1382-1391, 2006

[5] Li, J., Zaiane, O. Combining Usage, Content and Structure Data to Improve Web Site Recommendation. Proc. WebKDD-2004 work-shop on Web Mining and Web Usage, 2004

[6] YouTube. http://www.youtube.com

[7] Misu, T., Kawahara, T., “Speech-based Interactive information Guidance System Using Question-Answering Technique”, In Proc. IEEE-ICASSP, pp.145-148, 2007.

[8] Misu, T., Kawahara, T., Shoji, T., Minoh, M., “Speech-based Interactive Information Guidance System Using Question-Answering and Information Recommendation”, IPSJ, Vol.48, No.12, pp.3602-3611, 2007.

[9] Fujii, A., Itou, K. “Building a Test Collection for Speech-Driven Web Retrieval”, Proceedings of the 8th European Conference on Speech Communication and Technology, pp.1153-1156, 2003.

[10] Kawahara, T., Lee, A. “Open-Source Speech Recognition Software Julius”, JSAI, Vol.20, No.1, pp.41-49, 2005

[11] Matsumoto, Y, Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., Asahara M. Japanese Morphological Analysis System ChaSen version 2.2.1. Dec, 2000.

[12] Wikipedia, The Free Encyclopedia, http://www.wikipedia.org/