Spoken News Queries over the World Wide Web

Sebastian Stüker
Research Group 3-01
'Multilingual Speech Recognition'
Karlsruhe Institute of Technology
Postfach 6980
76049 Karlsruhe, Germany
sebastian.stueker@kit.edu

Michael Heck
Interactive Systems Laboratories
Karlsruhe Institute of Technology
Postfach 6980
76049 Karlsruhe, Germany
michael.heck@student.kit.edu

Katja Renner
Interactive Systems Laboratories
Karlsruhe Institute of Technology
Postfach 6980
76049 Karlsruhe, Germany
katja.renner@student.kit.edu

Alex Waibel
Interactive Systems Laboratories
Karlsruhe Institute of Technology
Postfach 6980
76049 Karlsruhe, Germany
waibel@kit.edu

ABSTRACT

In this paper we present our work in expanding the View4You system developed at the Interactive Systems Laboratories (ISL). The View4You system allows the user to retrieve automatically found news clips from recorded German broadcast news by natural spoken queries. While modular in design, so far, the architecture has required the components to at least run in a common file space. By utilizing Flash technology we turned this single machine setup into a distributed set-up that gives us access to our news database over the World Wide Web. The client side of our architecture only requires a web browser with Flash extension in order to record and send the speech of the queries to the servers and in order to display the retrieved news clips. Our future work will focus on turning the monolingual German system into a multilingual system that provides cross-lingual access and retrieval in multiple languages.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Miscellaneous

General Terms

Design

Keywords

information retrieval, automatic speech recognition, View4You, WWW

1. INTRODUCTION

Over the recent years the World Wide Web (WWW) has developed itself from a platform for distributing interlinked hypertext to a provider of multimedia content. This development was made possible by the increase of computing power , the availability of broad band networks in many places at an affordable price via technologies, such as ADSL, Wifi or UMTS, and the availability of efficient audio and video compression algorithms.

Besides the user generated multimedia content, many professional providers, such as broadcasting companies, have started to distribute their content over the WWW. Already when the majority of the content consisted of text only, the amounts of data offered was so large that it required the use of Internet search engines in order to navigate through the web. Searching through hypertext is made easy by the fact that text can be directly processed by machines and current text-based information retrieval technologies. But searching through multimedia content needs more processing, in order to make the unstructured video and audio streams suitable for current search technologies.

Usually, search engines require the user to type a query, usually in the form of relevant keywords and phrases, in order to start the search process. Just like other researchers do [3, 2], we feel that in many situations it would be more suitable and friendlier for the user, to speak a query in a natural way, instead of typing. Especially in situations in which the user is not in front of a classical computer, e.g., when using a home-entertainment system or a mobile device, we see speech as a very suitable input modality.

The View4You system developed at our laboratories [5] enables the user to perform this kind of queries to a database of German TV news shows that were recorded via satellite.

Though modular in design, the system was bound to run at least on a common file space and was not designed to run in a distributed set-up over the Internet. In this paper we therefore present our efforts in expanding the View4You architecture in such a way as to allow the access to the retrieval system via thin client over the World Wide Web.

This extension was done within the scope of the Quaero program[1], a recently started French research and development program with German participation. It targets to develop multimedia and multilingual indexing and management tools for professional and general public applications such as the automatic analysis, classification, extraction, and exploitation of information. The projects within Quaero address five main application areas:

- Multimedia internet search
- Enhanced access services to audiovisual content on portals
- Personalized video selection and distribution
- Professional audiovisual asset management
- Digitalization and enrichment of library content, audiovisual cultural heritage, and scientific information.

Also included in Quaero is basic research in the technologies underlying these application areas, including automatic speech recognition, machine translation, and information retrieval. The vision of Quaero is to give the general public as well as professional user the technical means to access various information types and sources in digital form, that are available to everyone via personal computers, television, and handheld terminals, across languages.

The rest of this article is structured as follows. Section 2 introduces the View4You system developed and maintained at our laboratory. In Section 3 we then describe our new, web-server based architecture for accessing the View4You system through a browser over the World Wide Web. Finally, in Section 4, we give an outlook into our ongoing and future research in turning the View4You system into a multilingual system that provides cross-lingual access to news from different countries by means of spoken queries in multiple languages.

2. VIEW4YOU

The original View4You system performed daily recordings of the main German news show—Tagesschau[2] at 8p.m. broadcasted by the ARD TV channel[3]—via satellite. The shows were automatically segmented in coherent news clips and transcribed via an automatic speech recognition system (ASR). The result of segmentation and transcription was stored in a database that contained the 60 latest shows. Via an information retrieval (IR) component the user could then retrieve clips from the database via natural queries of the kind “Was gibt es neues vom Sport?” (English: “What are the latest news from sports?”). These queries could either be typed or spoken, which in the later case were then first recognized by another ASR component.

So, in order for the View4You system to function we need 4 components:

- Data Collection
- Segmentation
- Automatic Speech Recognition
- Information Retrieval

We will briefly review the components of the View4You system here and refer to the corresponding publications which give further details.

2.1 Data Collection

Originally the Tagesschau data was collected via analog satellite TV and digitized into an MPEG-1 video stream with MPEG Audio Layer 2 with a data rate of 192kbit/s and a sampling frequency of 44.1kHz [5]. With the development of the broadcast technology toward the broadcast of digital content, we had temporarily switched to recording the terrestrial digital TV signal broadcasted via DVB-T. As the TV station that broadcasts Tagesschau has started to stream its news program over the Internet, we are currently directly recording the digital stream from the World Wide Web, thus being able to record without additional equipment for receiving either satellite or DVB-T signals.

In addition to recording the main show at 8p.m. we also started to record the daily television news magazine ‘Tagesthemen’[4] and the late night news magazine ‘Nachtmagazin’[5], both broadcasted by the ARD TV station over the Internet as well.

2.2 Segmentation

Once the news show has been recorded the next processing step is to segment it into clips of appropriate length. Ideally the clips cover a single topic, so that later the user does not need to watch the complete show that contains information that interests him, but rather only the shorter segment of the show that covers the topic of interest.

The View4You system uses a model based segmentation approach [10, 6] that was found to performs best at a medium level of recall [5]. The model based segmenter defines four acoustic classes for each of which prior to segmentation one Gaussian Mixture Model (GMM) is trained. The number of Gaussian components was defined by hand individually for the each class. The four different classes are ‘Anchor Speaker’ with 128 Gaussians, ‘Field Speech’ with 128 Gaussians, ‘Music’ with 32 Gaussians, and ‘Silence’ with 2 Gaussians. For the acoustic pre-processing of the data 16 mel-spectral parameters were computed every 50 msec, using a 16 msec window.

Using these models the audio of each recorded show was classified into these four classes using our standard HMM acoustic model based decoder by using each class as one word in the dictionary. By duplicating HMM states for each ‘word’, individual minimum length constraints for each word were enforced: 5 seconds for ‘Anchor Speaker’ and ‘Field Speech’, 2.5 seconds for music, and 0.2 seconds for ‘Silence’ The ‘word’ boundaries in the hypothesis were taken as the segment boundaries.

[1] http://www.quaero.org
[2] http://www.tagesschau.de/
[3] http://www.ard.de
[4] http://www.tagesthemen.de
[5] http://www.nachtmagzin.de

2.3 Automatic Speech Recognition

The ASR component used for transcribing the recorded news programs and for recognizing the spoken user queries [5] was trained and implemented using the Janus Recognition Toolkit (JRTk) featuring the IBIS single pass decoder [7]. The acoustic model consists of context dependent left-to-right Hidden Markov Models with three sub-states per phonemes and without state skipping. For the context-dependent model generalized triphones are used, that were tied by a decision tree and that utilize Gaussian Mixture Models for computing the emission probabilities with 30 Gaussians per model. The preprocessing computes 13 mel-scaled cepstral coefficients from a 16ms Hamming window with a 10ms frame shift. The resulting feature vector is enhanced with its delta and delta-delta coefficients and reduced to 16 dimensions using linear discriminant analysis. The language model is a tri-gram model trained on a selection of newspaper texts and transcriptions of news shows. The recognition dictionary contains the 60,788 most frequent words from the language model training corpus.

On average the ASR component achieves a WER of 22.7% on the Tagesschau news data [8]. Depending on the background an recording condition the WER varies between 11.9% for clean anchor speech and about 30.0% for very noisy conditions [5]

2.4 Information Retrieval

The View4You information retrieval component [8] uses the Okapi distance measure [1] for finding the documents d that are closest to the user query q, since past evaluations by NIST have found this distance to be quite powerful [6]. The Okapi distance can be parameterized to meet special requirements. The View4You system uses a parameterization that has been found to work well for short queries [9].

$$d(q,d) = \sum_{t \in q \wedge t \in d} \left(\frac{f_{d,t}}{f_{d,t} + \frac{\sqrt{f_d}}{E(\sqrt{f_d})}} \right) log \left(\frac{N - f_t}{f_t} \right) \qquad (1)$$

$$= Okapi(k_1 = 1, k_2 = 0, k_3 = 0, b = 1, r = 0, R = 0) \qquad (2)$$

Here $E(.)$ denotes the expected value, N is the number of documents in the collection, f_t is the number of documents containing term t, $f_{d,t}$ is the frequency of term t in document d, and f_d is the number of terms in document d, as an approximation of the document length. A *term* in this context refers to a word, except that the 500 most frequent words are excluded—a simple stop word list. Morphological stemming is applied to the query and the database. The IR component computes the distance between the query and every document in the database and returns the documents sorted by decreasing order of similarity to the query.

3. RESTRUCTURED ARCHITECTURE

The part of the original View4You system that accepts user queries, runs the information retrieval, and displays the results of the query runs on a single client machine. The architecture allows, in theory, to distribute these components over several servers, but only under the condition that they share a common file space. In reality this is of little use, if one wants to make such a database available to several distributed clients, e.g., via the Internet.

We therefore extended the system in such a way that it is possible to submit spoken queries and display the new clips retrieved using a thin client that only runs a web browser but none of the components of the View4You system. Instead, the ASR, IR and the database is hosted on a remote server architecture orchestrated by the web server.

Current web browsers implementing the HTML standard up to version 4 do not offer by themselves the capability to record audio and send it over the web, nor to receive video streams and display them. Instead extension technologies exist—such as Java, Java Script, and Flash—that provide these capabilities. When it comes to recording audio on the client by a web browser and sending the recordings over the WWW, security concerns become important. Therefore, the different technologies offer different security mechanisms to prevent malicious software from eavesdropping on the user. While Java uses public private key certificates to identify trustworthy applets, Flash uses a simple dialog box to obtain the necessary security privileges for recording from the microphone of the client.

Due to the easier handling of the security aspects when recording audio, and due to the ease of programming when streaming and displaying multimedia content, we decided to use Flash for developing our distributed architecture. The client side thus only requires a web browser with flash extension.

Figure 1 shows this distributed client server architecture.

3.1 Client Side

The Flash application on the client side needs to be able to record and send spoken queries over the WWW to the server and to receive the list of the resulting clips and the ability to play them. We built the Flash application using the Adobe Flex Builder 3.5 Educational[7]. As a consequence of this, the client needs an Adobe Flash Player 9 or higher in order to run the application. For displaying the results, playing the videos, and for recording the spoken queries we used several controls, such as the mx.control.VideoDisplay, that are part of the Flex 3.5 packages. For submitting a spoken query we use a push and hold interface that requires the user hold down a button while speaking.

3.2 Server Side

The server side receives the spoken query, converts and passes it on to the ASR component. The resulting transcription is then relayed by the server to the IR component which returns the list of results to the server. The list of results is processed and then sent to the client. When the client requests a clip to be played the corresponding video is streamed to the client.

For the implementation of the server we use the Red5 server version 0.9.1 final[8]. Red5 is an open source software Realt Time Messaging Protocol (RTMP) media server written in Java. A Java routine in the web server is called when the user starts to stream a spoken query and which orchestrates the subsequent calls to ASR and IR, as well as the return of the list of results.

[6] http://trec.nist.gov/

[7] http://www.adobe.com/de/products/flex/

[8] http://code.google.com/p/red5/

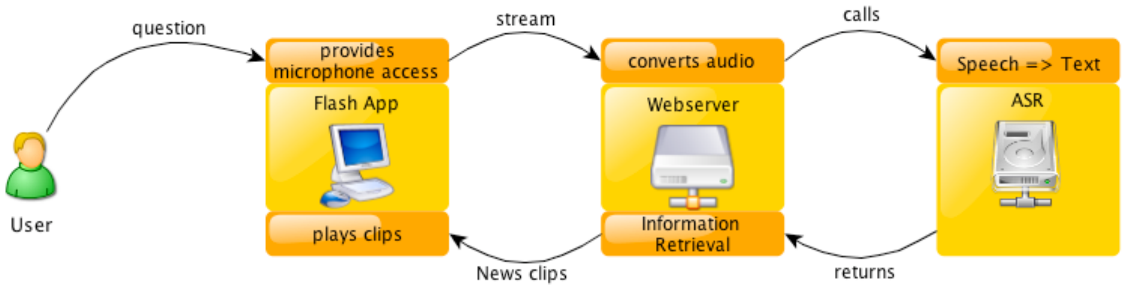

Figure 1: The Client-Server setup

4. WORK IN PROGRESS

Currently the View4You system only contains a database of German news shows and only accepts German queries. Similar to the Informedia system [4] we are planning to expand the View4You system to cover multiple languages and to allow cross-lingual queries. In contrast to the Informedia system we plan to not only allow textual queries across languages, but also spoken queries in multiple languages.

In preparation for this we have started to collect data from CNN via YouTube, and are preparing to collect French news programs from TF1. The main question in dealing with the data in a cross-lingual way will be the trade-off between the two strategies of either translating the query into the two other languages or translating the whole databases into all languages.

We further plan to enhance the interface with an automatic language identification, so that the user does not need to manually select his input language, but can rather just speak his query and view the results in the language corresponding to the query.

5. CONCLUSION

In this paper we presented our recent extension of the View4You system which records German TV news and allows the user to search for automatically segmented clips by the way of spoken, natural queries. While the system so far had to run on a single client, we now have rearranged the architecture in such a way, that the system can be accessed over the World Wide Web. By making use of Flash Technology, the client for accessing the View4You database only needs to run a web browser with Flash capability. In this way the processing can be completely done on a remote server, while only the recording of the spoken query and the result in form of a list of retrieved clips must be transferred between the server and client.

For the future we plan to extend the currently monolingual, German system into a multilingual system that allows user queries in multiple languages and access to information across languages by utilizing machine translation technology.

6. ACKNOWLEDGMENTS

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. 'Research Group 3-01' received financial support by the 'Concept for the Future' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative.

7. REFERENCES

[1] M. Beaulieu, M. Gatford, X. Huang, S. Robertson, S. Walker, and P. Williams. Okapi at trec-5. In *Proceedings the 5th TREC*, Gaithersburg, MD, USA, January 1997.

[2] A. Franz and B. Milch. Searching the web by voice. In *Proceedings of the 19th international conference on Computational linguistics*, Morristown, USA, 2002.

[3] H. Gu, J. Li, B. Walter, and E. Chang. Spoken query for web search and navigation. In *Poster Proc. 10th Int. World-Wide Web Conf*, 2001.

[4] A. Hauptmann, N. Moraveji, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Baron, W.-H. Lin, J. Yang, T. Ng, N. Papernick, C. Snoek, G. Tzanetakis, H. Wactlar, R. Yan, and R. Jin. Informedia at trecvid 2003: Analyzing and searching broadcast news video. In *Proceedings of (VIDEO) TREC 2003*, Gaithersburg, MD,USA, November 2003.

[5] T. Kemp, M. Weber, and A. Waibel. End to end evaluation of the isl view4you broadcast news transcription system. In *In the Proceedings of RIAO*, Paris, France, April 2000.

[6] A. Sankar, F. Weng, Z. Rivlin, A. Stolcke, and R. Gadde. The development of sri's 1997 broadcast news transcription system. In *Proceedings the DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA, 1998.

[7] H. Soltau, F. Metze, C. Fügen, and A. Waibel. A one pass-decoder based on polymorphic linguistic context assignment. In *Proceedings of the ASRU*, Madonna di Campiglio Trento, Italy, December 2001.

[8] M. Weber and T. Kemp. Evaluating different information retrieval algorithms on real-world data. In *Proceedings of the ICSLP*, Beijing, October 2000.

[9] R. Wilkinson, J. Zobel, and R. Sacks-Davis. Similarity measures for short queries. In *Proceedings of the 4th TREC*, Gaithersburg, MD, USA, November 1995.

[10] P. Woodland, T. Hain, S. Johnson, T. Niesler, A. Tuerk, and S. Young. Experiments in broadcast news transcription. In *Proceedings of the ICASSP 1998*, Seattle, WA, USA, May 1998.