Mobile Information Access with Spoken Query Answering

Tom Brøndsted[1], Henrik Legind Larsen[2], Lars Bo Larsen[1], Børge Lindberg[1], Daniel Ortiz-Arroyo[2], Zheng-Hua Tan[1], Haitian Xu[1]

[1] Speech and Multimedia Communications, Department of Communication Technology
[2] Software Intelligence and Security Research Center (SIS-RC), Esbjerg
Aalborg University, Denmark
{tb,lbl,bli,zt,hx}@kom.aau.dk, {hll,do}@sis-rc.org

Abstract

This paper addresses the problem of information and service accessibility in mobile devices with limited resources. A solution is developed and tested through a prototype that applies state-of-the-art *Distributed Speech Recognition* (DSR) and *knowledge-based Information Retrieval (IR)* processing for spoken query answering. For the DSR part, a configurable DSR system is implemented on the basis of the ETSI-DSR advanced front-end and the SPHINX IV recognizer. For the knowledge-based IR part, a distributed system solution is developed for fast retrieval of the most relevant documents, with a text window focused over the part which most likely contains an answer to the query. The two systems are integrated into a full spoken query answering system. The prototype can answer queries and questions within the chosen football (soccer) test domain, but the system has the flexibility for being ported to other domains.

1. Introduction

Search engines like Google™ unquestionably belong to the most popular information services accessed from conventional desktop computers with web access. However, on mobile devices such as Personal Digital Assistants (PDAs) and mobile phones, the accessibility to these kinds of services is not optimal as they require keyboard input and normal-sized computer screens for browsing and displaying large numbers of retrieved documents. Consequently, recent research has focused on studying efficient techniques to provide spoken query answering capabilities on mobile devices.

A spoken query system that retrieves textual information from a database over a PDA is presented in [1]. Speech retrieval of broadcast news via mobile devices is presented in [2] where the speech information is recognized and retrieved using spoken natural language queries. A recent project, SmartWeb [3][4], targets mobile use with multi-modal access to the semantic web. However, due to the very limited availability of information in the semantic web, the SmartWeb project is currently investigating the use of language technology and information extraction (IE) methods for the automatic semantic annotation of web pages in the standard XML/HTML format.

The system presented in this paper has some similarities with the approaches previously described, but addresses the *input* and *output* problems of information access with mobile devices in a different way. By *the input problem* we understand the limited input modalities available for text entry on mobile devices, in essence the 12-key keypad or a pen. The use of both devices requires quite substantial motor coordination. We address *the input problem* by substituting textual input with state-of-the-art DSR technology where the computationally demanding part of the speech recognition takes place on a remote server [5]. To provide flexibility to the user, spoken *questions* are entered in natural language. The acoustic models employed in the DSR system are trained for the Danish language.

The output problem refers to the very limited screen size and resolution on mobile devices as compared to desktop computers. This will make a standard web search, likely to produce an abundance of hits, unfeasible to handle on such devices. We address the output problem by using a knowledge-based IR system that retrieves and processes the current information available on the web in the chosen application domain, which in this case is the Danish football league. The system applies a number of advanced techniques, all in the framework of fuzzy logic, to retrieve a small number of the most relevant documents (news articles), that are most likely to contain the answer to a query. These techniques comprise fuzzy term nets [17], fuzzy query answering [18], and fuzzy probabilistic text mining [19], as well as concept clustering [14], and mining of user preferences in the user interaction log. Further, the system supports question answering with the direct answer generated from a database with data extracted from the documents using *Information Extraction* (IE) techniques and knowledge about the domain (soccer games, players, clubs, etc.).

The system interface as experienced by the end-user consists of a *Graphical User Interface* (GUI) consisting of a window that, depending on the application mode, shows either the focused part of the text or the direct answer to the question. Fig. 1 shows the main components of the system.

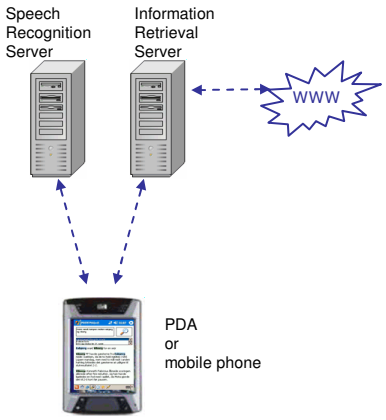

Fig. 1. The integrated mobile information access system

This paper is organized as follows. In Section 2, the system's architecture is presented. Sections 3 and 4 describe the speech recognizer and the IR system, respectively. Finally, in Section 5, we present some conclusions and future work.

2. Architecture

The system employs a fully distributed architecture that includes both the speech recognizer and the IR system. The overall architecture of the system is depicted in Figure 2.

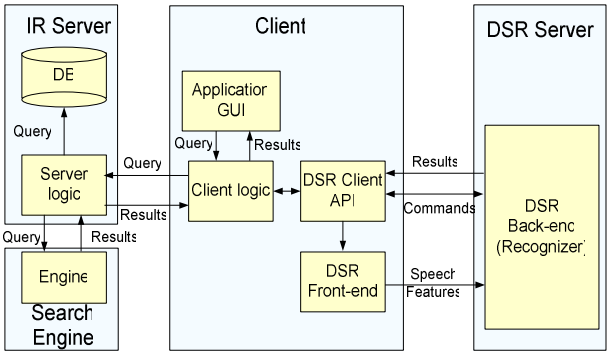

Fig. 2 Architecture of the speech-enabled information access system

The speech recognizer is implemented using the DSR-scheme where the speech recognition processing is split into the client-based *DSR front-end* feature extraction and the server-based *DSR back-end* recognition. Speech features are transmitted from the *DSR client* to the *DSR server*. Subsequently, the output in the form of either the best result or an N-best list is sent from the remote server back to the client and from there to the IR server. In order to enable the configuration of the DSR back-end recognizer commands are transmitted back and forth between the *DSR client* and the *DSR server*. The distributed speech recognizer is described in Section 3.

The IR server consists of a server logic module that receives the user's query in text form and determines whether the query should be sent to the search engine or can be answered directly. A *Really Simple Syndication* (RSS) feed client (not shown on the figure) constantly retrieves sports news from diverse news providers and stores these documents in the data base for further processing. The data base contains also the user's preference profile that is utilized to improve the precision of the IR system.

3. Distributed speech recognizer

As illustrated in Fig.3, the DSR system [6] is developed on the basis of the ETSI-DSR advanced front-end (AFE) [7], [8] and the SPHINX IV recognizer [9].

3.1. The DSR system

The advanced front-end client-side module extracts noise-robust *Mel-Frequency Cepstral Coefficient* (MFCC) features which together with *Voice Activity Detection* (VAD) information are encoded sequentially and packed into speech packages for network transmission.

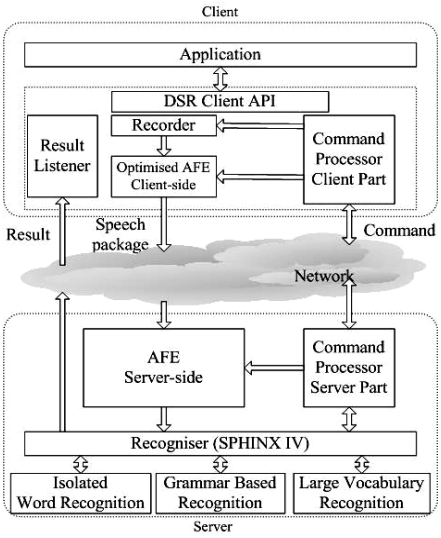

Fig.3 The DSR architecture

At the server side the received speech packages are processed by the advanced front-end server-side module. First, on the detection of transmission errors, error concealment is conducted for feature reconstruction. Then, the error-corrected speech packages are decoded into a set of cepstral features and VAD information. Subsequently, the cepstral features are processed by the SPHINX speech recogniser. The recogniser presents its result (either the best or N-best results) at the utterance end, detected by the VAD information, and transmits back to the client. To increase system usability and flexibility, three typical recognition modes are represented, namely: "Isolated word recognition", "Grammar based recognition" and "Large vocabulary recognition". Each is defined by a set of prototype files at the server side. The choice is done at system initialisation, and specific settings can be changed at any time. The setting may be different across a group of end-users.

A "Command Processor" is implemented at both the client and server side to support the interchange of configuration commands. Potential commands include control commands to start or stop recognition, choice of recognition mode, commands providing feedback information from the server to a client (e.g. success or failure of any user request), etc.

The DSR client has been evaluated on a H5550 IPAQ with a 400MHz XScale CPU and 128 MB memory. With speech data sampled at 8 KHz the client is able to conduct the speech processing 0.82 times real-time.

3.2. Acoustic models

The acoustic models [10] are trained on 91 hours of speech from the SpeechDat(II) [11] database, which contains telephone quality speech sampled at 8kHz stored in A-law format. Training of the models is performed using tools from SphinxTrain and using the Advanced DSR Front-end [8].

The models trained are between-word triphones and word-internal triphones based on 61 phones (diphthongs are treated as phones). Furthermore three context independent models are trained for non-speech events. Both the triphones and the non-

speech event models are modeled by three state left-to-right continuous density *Hidden Markov Models* (HMMs) with 16 Gaussians per state.

In order to handle unseen and rarely seen triphones reliably decision tree based clustering and training is performed using 4000 senones.

3.3. Vocabulary, grammar, language model

The current target domain for the recognition task is Danish football news. For the modelling of the spoken query language, the system uses a rule-grammar that allows only a fixed sentence structure. The sentences allowed are questions about players, teams and matches (17 different question types). The vocabulary size is 709, where 361 are player names, 47 are team names, and 101 is the base vocabulary for numbers.

Applying this grammar for recognition a sentence accuracy of 90% is achieved when testing on 260 sentences (approximately 14 minutes) recorded by one person using a Plantronics M2500 bluetooth headset. It should however be noted that it is also possible to use the embedded microphone in the PDA as input device. In addition, another test was carried out with the built-in microphone on a Qtek 9090 Pocket PC. This test revealed a doubling of the *Word Error Rate* (WER) for users who partly obscured the microphone with the hand holding the PDA (about 50% of the test users did so). This result indicates that other problems than those strictly related to the acoustic decoding must be addressed for the PDA platform before a real-life service can be successfully deployed.

4. Knowledge-based IR system

The knowledge-based IR system applies a number of models for mining term association knowledge, adaptive personalization, conceptual characterization of documents and queries, and matching of the presentations of documents and queries. These models are applied in the framework of fuzzy sets and fuzzy logic allowing us to utilize vague and uncertain knowledge as is characteristic in intelligent IR. The IR system supports document retrieval where the query answer is a ranked list of the most relevant documents, and question answering where the direct answer is constructed from a database containing structured information (facts) extracted from the document base. The document retrieval system and the question answering system are presented in the following subsections.

4.1. The document retrieval system

To reduce the set of representative keywords, all texts employed in the system, comprising queries and documents, are first processed by stemming and stop-words removal.

The IR engine employs a model consisting of the user's context representation, a fuzzy term net [17] and concept clustering processing using fuzzy aggregation operators [18, 13]. The user's context representation consists of the historical list of queries, the user's preferences, and the list of documents that the user has selected for viewing as a result of previous queries. This context is used as supplementary knowledge for reducing the uncertainty in the semantic interpretation of the query.

A fuzzy term net is created (before the user enters any query) from the document corpora obtained from the diverse RSS feed sport news providers. The term net essentially contains term associations mined from the corpora through fuzzy data mining [19]. The model employs this term net to index the main concepts contained in the documents. To create the term net, the internal structure of the documents (e.g. title, abstract, source) is parsed and used to weight the significance of the index terms.

To index documents, a technique based on concept clusters similar to the one proposed in [14] was used. However, we modified that technique to make use of the user's context information during cluster creation. Clusters are created considering a document as "a conglomeration of concepts". Indexes are established through the concepts extracted, and weights are used to signify the degree of semantic importance of the terms in the text. The *Andness – directed Importance Weighted Averaging* (AIWA) and *Andness-directed Averaging* (AA) operators [13] are applied inside and between these clusters to obtain the level of satisfaction to which a document satisfies the user's query. Additionally, each term belonging to a query is expanded using the user's context information and the fuzzy term net. The method used for this purpose is the same as the one used during document indexing, namely the creation of concept clusters. However, in this case concepts are created amongst the top-ranked documents. The terms coming from these representative concepts are used to expand the query. Using this technique the model is able to find documents containing the same concepts employed by the user in a query. A detailed description of the model is presented in [15].

The retrieved documents are ordered according to the relevance to the query. Moreover, to improve the system's response to the user, the most relevant document's text is sent from the IR server to the client, jointly with the title of other relevant documents. Once a user selects a new document on the GUI, the document's text is retrieved from the server.

We performed a preliminary evaluation of the IR engine comparing its performance with Lucene, an open source IR engine. Lucene is a high performance, full-featured text search engine library based on the vector space model. Lucene provides several searching types: lexical proximity search, word occurrence proximity search, wildcard etc. In our experiments, we performed 18 queries over a corpus of one hundred documents. Our preliminary results indicate that our search engine outperforms Lucene by approximately 5% in terms of precision and recall.

4.2. The question answering system

In addition to retrieve documents from RSS feeds, the IR system monitors sport-related web sites to retrieve information about players, matches, clubs, tournaments, etc. An IE system is applied to extract relevant information about football through identifying valid tables contained in the retrieved web pages. Subsequently, the tables are filtered and processed to store the relevant information in a data base. The data base is structured to represent the knowledge about the football soccer domain using an *Entity-Relationship* modeling (ER). Finally, a list of question formats is stored in the system. These formats contain common questions that users interested in soccer may ask. When one of these questions is recognized by the system, the data base is

queried to obtain the direct answer. A detailed description of the information extraction processing is presented in [16].

5. Conclusions and Future Work

In this paper we presented a mobile information access system with spoken query answering. The system adopts a distributed architecture in which the speech recognizer and the knowledge-based IR system are located in different servers. The spoken queries are processed using the DSR technology. The IR system is characterized by application of fuzzy text mining, concept clustering and personalization, all in the framework of advanced fuzzy logic based representation and operations. The integrated system has the capability of answering spoken queries over a PDA.

Future work will consider using *n*-gram language models instead of the current rule grammar to enable natural spoken language queries. One obvious consequence of this will be an increase of the current vocabulary with a factor of 10-50, which will truly reveal the power of the client-server architecture, but also increase the demands to the acoustic decoder. Future work will also include extensive user test scenarios to identify and solve potential usage problems such as those mentioned in Section 3.3. Furthermore, experiments with the presentation of the query answers must also be carried out to optimize the user satisfaction. One important factor will be to include learned contextual information, such as user profile, dialogue history, result sets from previous queries, etc. Indexing and searching video contents using voice recognition of the sound track also provide challenges for future work.

To carry out user tests we have built a test facility, where different heterogeneous networking and environmental contexts can be simulated in end user tests.

6. Acknowledgements

The project was supported by Center for TeleInFrastruktur (CTIF) in the project POSH under CTIF's C3 programme. Dan Albæk Majgaard managed the software engineering of the client server solution of the IR system, and contributed creatively in the GUI design. Morten Højfeldt Rasmussen and Morten Thunberg Svendsen contributed to the training of acoustic models. The following master students participated enthusiastically and constructively in the development of the IR system: Marthe Buffière, Henrik Mathiassen, Niels Nygaard Nielsen, Allan Pedersen, and Frédéric Pichon.

7. References

[1] Chang, E., Seide, F., Meng, H.M, et al.: A system for spoken query information retrieval on mobile devices, *IEEE Trans. Speech and Audio Proc*., 10(8):531–541, 2002.

[2] Chen, B., Chen, Y.-T., Chang, C.-H., et al.: Speech retrieval of Mandarin broadcast news via mobile devices, *Interspeech 2005*, Lisbon, Portugal, Sep. 2005.

[3] Reithinger, N. and Sonntag, D.: An integration framework for a mobile multimodal dialogue system accessing the semantic web, *Interspeech 2005*, Lisbon, Portugal, Sep. 2005.

[4] SmartWeb: Mobile Broadband Access to the Semantic Web. http://www.smartweb-project.org

[5] Tan, Z.-H., Dalsgaard, P. and Lindberg, B.: Automatic speech recognition over error-prone wireless networks, *Speech Communication*, **47**(1–2), 220–242, 2005.

[6] Xu, H., Tan, Z.-H., Dalsgaard, P., Mattethat, R. and Lindberg, B.: A configurable distributed speech recognition system, *Biennial on DSP for in-Vehicle and Mobile Systems*, Sesimbra, Portugal, Sep. 2005.

[7] 3GPP TS 26.243: *ANSI-C code for the Fixed-Point Distributed Speech Recognition Extended. Advanced Front-end*, December, 2004.

[8] ETSI Standard ES 202 212. *Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithm, back-end speech reconstruction algorithm*, November 2003.

[9] W.Walker, P.Lamere, P.Kwok et.al.: *Sphinx-4: A Flexible Open Source Framework for Speech Recognition,* Technical report TR-2004-139, Sun corporation, USA, 2004.

[10] Rasmussen, M.H. and Svendsen, M.T.: *Large Vocabulary Continuous Speech Recognizer for Danish and Language Model Adaptation,* Master Thesis, Aalborg University, 2005.

[11] Lindberg, B.: *Speechdat, Danish FDB 4000 speaker database for the fixed telephone network*, pp. 1–98, March 1999. cdrom://biblo/SpeechDat2_DK.pdf.

[12] Larsen, H., L. Importance weighted similarity based soft interpretation of crisp criteria queries, *Proceedings of EUROFUSE Workshop on Preference Modeling and Applications*, Granada, Spain, pp 175-180. 2001

[13] Larsen, H.L.: Efficient Andness-directed Importance Weighted Averaging Operators. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **12** (Suppl.), pp 67–82. 2003

[14] Kang, B., Y., Kim, D., W., and Lee, S.J.: Exploiting concept clusters for content based information retrieval, *Information sciences 170,* pp. 443–462, 2005

[15] Buffiere, M. and Pichon, F.: *Knowledge based flexible query answering*, MSc. Thesis, Aalborg University Esbjerg, 2005.

[16] Mathiassen, H., Nielsen N. N, and Pedersen, A.: *Mining Tables from Domain Specific HTML Text*, Information Retrieval Project Report, Aalborg University Esbjerg, 2005.

[17] Larsen, H.L., and Yager, R.R.: The use of fuzzy relational thesauri for classificatory problem solving in information retrieval and expert systems. *IEEE J. on System, Man, and Cybernetics* **23**(1):31–41, 1993.

[18] Larsen, H.L., and Yager, R.R.: Query Fuzzification for Internet Information retrieval. In Dubois, D., Prade H., and Yager, R.R., Eds., *Fuzzy Set methods in Information Engineering: A Guided Tour of Applications*, John Wiley & Sons, pp. 291–310, 1997.

[19] Larsen, H.L.: Fuzzy data mining of term associations for flexible query answering. *Proc. International Conference in Fuzzy Logic and Technology, Leicester, England, 5–7 September 2001 (EUSFLAT'2001).*

[20] The CMU Sphinx Group Open Source Speech Recognition Engines. http://cmusphinx.sourceforge.net/html/cmusphinx.php