Supporting radio archive workflows with vocabulary independent spoken keyword search

Martha Larson
ISLA, University of Amsterdam
Kruislaan 403
1098 SJ Amsterdam,
The Netherlands
larson@science.uva.nl

Stefan Eickeler
Fraunhofer IAIS
Schloss Birlinghoven
53754 Sankt Augustin,
Germany
stefan.eickeler@
iais.fraunhofer.de

Joachim Köhler
Fraunhofer IAIS
Schloss Birlinghoven
53754 Sankt Augustin,
Germany
joachim.koehler@
iais.fraunhofer.de

ABSTRACT

Archive departments of large radio archives stand to benefit greatly from speech recognition technology and other audio processing techniques. One of the reasons why automatic digital audio processing has not yet realized its full potential is that it remains unclear how to integrate automatic techniques into existing archive workflows. In order to move towards a practical understanding of how automatic techniques can be used to support archive staff, two large German radio broadcasters, Deutsche Welle and Westdeutscher Rundfunk, commissioned Fraunhofer IAIS to build a German-language radio archive prototype. This paper discusses the development and assessment of the spoken keyword search module of this prototype. The difference between the radio archive prototype discussed in this paper and existing systems for spoken document retrieval is that the prototype was designed and tested in a project group consisting of both multimedia researchers and archive professionals. As a result, the prototype and its evaluation is tuned to the explicit needs of archivists working at large radio archives. First, the paper discusses the special needs of radio archive staff and how they were accommodated in the design of the keyword search capacity. In particular, the archive staff required a vocabulary-independent search facility. This facility was implemented by a fuzzy-matching algorithm that performs a similarity search on syllable transcripts generated by the speech recognizer. Then, the paper presents the results of an evaluation designed to assess whether or not the radio archive prototype fulfilled the needs of archivists.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.4 [**Information Systems Applications**]: Miscellaneous; H.5 [**Information Interfaces and Presentation**]: Multimedia Information Systems

General Terms

Algorithms, Design, Performance

Keywords

spoken document retrieval, spontaneous speech, audio archive, speech recognition, radio broadcaster

1. INTRODUCTION

In 2000, spoken document retrieval was declared a solved problem [6]. Yet we arrive in 2007 and audio search systems are still not in widespread use to either retrieve or archive materials in large audio archives, such as those maintained by radio broadcasters. This paper reports on work in the area of spoken document retrieval carried out within a project commissioned from Fraunhofer IAIS by two large German radio broadcasters, Westdeutscher Rundfunk and Deutsche Welle. The goal of the project was to produce a German-language prototype radio archive that would be used to investigate the practical aspects of integrating automatically generated metadata into archive systems. The project wished to investigate the source of the lag between the availability of audio search technology and its integration into archiving workflows and to clarify which performance factors or design factors might be responsible.

Radio archive departments are eager to explore the possibilities for content-based spoken document retrieval offered by speech recognition technologies because they are faced with the task of annotating more material than would be possible by conventional (exclusively human-based) methods alone. Previously, archive departments stored radio content in the short term primarily for legal reasons. For long term storage, the best segments were removed and painstakingly annotated and archived as historical record or as cultural documents. Currently, the cost of storage medium has fallen far enough that it is becoming feasible to simply store everything broadcast. In other words, it not necessary to choose content for long term storage. Archives are increasingly being called upon to supply journalists and editors with content for reuse. Today's radio programs are enriched by an increasing number of sound bites drawn from past programs. This development has been characterized in the literature as archives moving "out of the basement" [8] to play a key role in production. In short, the trend is towards archiving as much material as possible.

In order to set a clear focus on the practical aspects of integrating automatic digital audio processing technology into archive work flows, the radio archive prototype was designed and tested by a project group consisting of both multimedia researchers, who were experts in the field of audio processing and speech recognition, and archive professionals, who were experts in the field of archiving and retrieval of radio content. Although the project group built on familiarity with previous research in the area of spoken document retrieval and on knowledge of existing systems and prototypes such as [9, 14, 7], an effort was made to eschew preconceptions and to build a prototype explicitly tailored to the needs of the archive departments of the two broadcasters. The prototype would thus offer a clear demonstration of the concrete potential of automatic digital audio processing to support the existing workflows of archive staff. The project group realized that in two areas the prototype radio archive needed to go above and beyond current practices in broadcast news retrieval. First, only a subset of the content of the radio archive is broadcast news or planned speech. Spontaneous speech, such as occurs in interviews, makes up a large portion of the archive content. Second, the information needs of archivists are specialized. Research in information retrieval places a focus on "aboutness", in other words in determining the topic of a document. User queries are considered to be requests for documents dealing with a certain topic. Although staff at radio archives needs to search for documents on certain topics, their information needs tend to transcend "aboutness." Archives receive requests for quotations spoken by certain prominent figures, for excerpts in which politicians express particularly strong negative opinions on popular issues and for segments in which interviewees use particular buzz-words or phrases. The project group also realized that in order to integrate well into the existing workflow, the radio archive prototype must be designed building on established archiving practices. The prototype needed to allow archivists to continue to use tried-and-true archiving conventions and familiar search strategies that allow them to combine detailed world knowledge and knowledge of the archive with a search interface.

This paper recounts the development, implementation and test of the spoken keyword search module of the radio archive prototype designed by the project group and implemented at Fraunhofer IAIS. It begins by a discussion of the needs of the archive staff and how these were incorporated into the functionality and interface design of the radio archive prototype. Then it discusses the testing of the prototype in order to determine if it met the needs of the archivists. Finally, it concludes with comments about lessons learned.

2. NEEDS OF ARCHIVE STAFF

In order to understand the radio archive domain and the broadcast news domain, the project group made a thorough investigation of the workflow and needs of archive staff. Both the content of the archive and the needs of the users (i.e. of the archive staff) turned out to constitute important differences between the radio archive domain and the broadcast news domain.

2.1 Radio archive content

Radio content is archived for several reasons. Broadcasters are typically legally required to keep a record of what they broadcast for a specified period of time. Also, a broadcaster, especially a public broadcaster, may have a mandate to preserve culturally relevant modern recordings and to curate a collection of historical audio recordings. Finally, a broadcaster archives material as resources to be rebroadcast or to be used in future productions.

The broadcasters who commissioned the radio archive prototype maintain broadcast news in their archives. But they also store a wealth of other content including documentaries, interviews and talk shows. Archive professionals pointed out that the most pressing need for content-based retrieval of radio content was for those programs for which there was little or no formal metadata. A news show typically has a minimal description of each of the report segments that was used to produce the show and which, if all goes well, follows it from production to the archive. Interview talk shows, however, are largely unplanned. In fact, their appeal to the listening public lies exactly in the spontaneous and free form discussion between show host and guest. An interview talk show arrives at the desk of the archivist with long sections which are unsegmented and not described in any way. The greatest potential for content-based retrieval for archive staff is being able to access these sorts of segments. Often, there are insufficient resources available to allow human annotation of interview talk shows and they are stored in the archive without annotation and are effectively lost. In sum, although radio archives contain planned speech such as broadcast news, archive departments need speech retrieval systems in order to retrieve programs containing unplanned speech, since these have no production data.

2.2 Archiving workflow

The first step in the archive workflow is for the archive staff to screen radio material for selection. They select which recordings, or which subsections of recordings, will be archived and for how long. They also select the level of detail at which each radio recording will be annotated. Naturally, material selected for long-term archival will be earmarked for a high level of annotation detail. The next step is to annotate each recording that is selected for archiving. In this step, the archivist produces, by hand, a description of the recording that will make it possible to find the recording in the archive. This description takes the form of a summary or a list of keywords. Depending on the level of granularity required, this description can include a division of the recording into topical sections, each marked with a time stamp. Each topical selection is then annotated separately. Sometimes a list of program segments is available from the production metadata. If such a list is available, the archivist is able to use it as a skeleton on which to build up the description. A final responsibility of the archive staff is to maintain the archive, protecting it against inconsistencies.

The radio archive prototype was conceived to support the archiving workflow of archive staff. Automatically generated metadata such as segmentations and speech recognition transcripts are to be used to aid conventional archiving practices. Annotation becomes a collaboration between archivist and machine. In contrast to broadcast news systems that fully automate the generation of metadata, the radio archive prototype, aims to support the workflow of archive staff.

2.3 Retrieval from the archive

The archive staff is responsible for responding to requests for recordings from the archives. Emphasis in retrieval from the archive is on precision and not recall, since archivists' task is to find something that is suitable and not necessarily to find everything that is suitable. The project group surveyed the types of requests generally received and assessed the response potential of the content-based search functionality of the prototype radio archive. The types of requests received by archive departments can be grouped into different groups with respect to the support that content-based keyword search potentially provides for responding to them.

In the first group are the requests that can be handled by using only the formal metadata of a recording, namely information such as the title of the series, the title of the particular program and the date of first broadcast. This group includes requests for a particular program by title, title words, producer, or by broadcast date.

In the second group are requests that are abstract and require knowledge of the archivist about the archive, about the programming of the broadcaster and about the world in general. Examples of highly abstract queries would include requests as, "Find statements of politicians who are pessimistic about the economy," "Find segments discussing prejudices of and against Germans," and "Find excerpts on the negative side of being famous." Archive staff see very limited potential for keyword search for these types of requests. They rely largely on knowledge accumulated during annotation work about which shows might include such fragments and which politicians or public figures might be inclined to make such statements.

The third group of requests can be approached by using full-text search of speech recognition transcripts. The most straight-forward cases are when the archivist is searching for a known quotation, such as the famous 1997 quote of Roman Herzog, President of Germany, "Durch Deutschland muß ein Ruck gehen." In the case of the original speech, the archive staff have probably annotated it already with a transcription of this quote. However, full text search on speech recognition transcripts makes it possible to find which politicians have quoted Roman Herzog since the original speech. Speech recognition transcripts are also invaluable to find buzzwords or currently important phrases spoken in different contexts by different people, such as "Harz IV", the labor reform. Also, many topics have indicative keywords which can be assumed to appear in the transcripts when these topics are discussed.

3. RADIO ARCHIVE PROTOTYPE

During the course of the project, the project group met on a regular basis to define the functional specifications of the prototype and to design the prototype interface. This section gives information about the prototype definition process and about the resulting system.

3.1 Data

The radio archive prototype needed to contain the full range of types of radio programs that the broadcaster archives must handle. Four programs were chosen to cover these types, two from each radio broadcaster. Deutsche Welle contributed approximately 80 hours of material from *Funkjournal* and *Wiso* two programs containing news reports and interviews. If the interview is not conducted in German, a clip of the interviewee responding in the interview language is played before the German translation is blended in. If the interviewee is speaking in English, the whole answer is played and then repeated translated into German. WDR contributed 40 hours from *Der Tag*, which contains reports, interviews, opinions and music. WDR also contributed 40 hours from *Montalk*, a two hour interview talk show featuring prominent figures from media, sports, politics and culture. Each show includes surprise guests, some physically present in the studio and some participating in the show via telephone. It also contains collages or short recordings made of people answering interview questions on the street as well as music. *Montalk* is the most challenging of the 4 programs because it contains nearly exclusively conversational speech characterized by laughter, interruptions and speakers with regional and/or colloquial speech. The radio archive prototype contained 160 hours of total material broadcast in 2005. As is discussed later in the paper, 12 hours was chosen to be annotated as a test set for the evaluation of the performance of the system.

3.2 Prototype functionality

The archive professionals in the project group defined the functionality that was most critical for the radio archive prototype. In this section, three of these functionalities that are related to spoken keyword search are discussed. In the following section implementation of these functionalities in the user interface is discussed.

First, it was deemed essential that the audio archive prototype not be dependent on a speech recognition vocabulary. The requests for information that archivists must respond to deal with a disproportionately large number of proper names. Archive staff need to be able to respond to requests concerning a rare name or a previously rarely mentioned place. The speed with which new proper names can break into the media was dramatically illustrated by the data set which included material broadcast in the weeks after the 2004 tsunami. This material contained many names of smaller places in Thailand and of people who were feared to be missing. Even the largest vocabulary cannot guarantee complete coverage of the large variety of human names. Archives are interested in exploiting marginal topics in radio collections, such as people or places that were mentioned only fleetingly. Such sound bites are of great value if the mentioned people or places subsequently achieve a high public profile.

Second, it was important that the prototype allow intelligent skimming of audio. Currently, archivists screening a radio program need to make best guess jumps when they fast forward. Intelligent skimming means offering the archivists signposts so that jumps can be informed. These signposts can take the form of segment boundaries or of segment labels. Archivists are then able to skip over music within a radio program and listen in only to speech. Signposts can also take the form of keywords. Archivists wanted to be able to type in a keyword and see at which places it is spoken within a radio program. In this way, an archivist can use

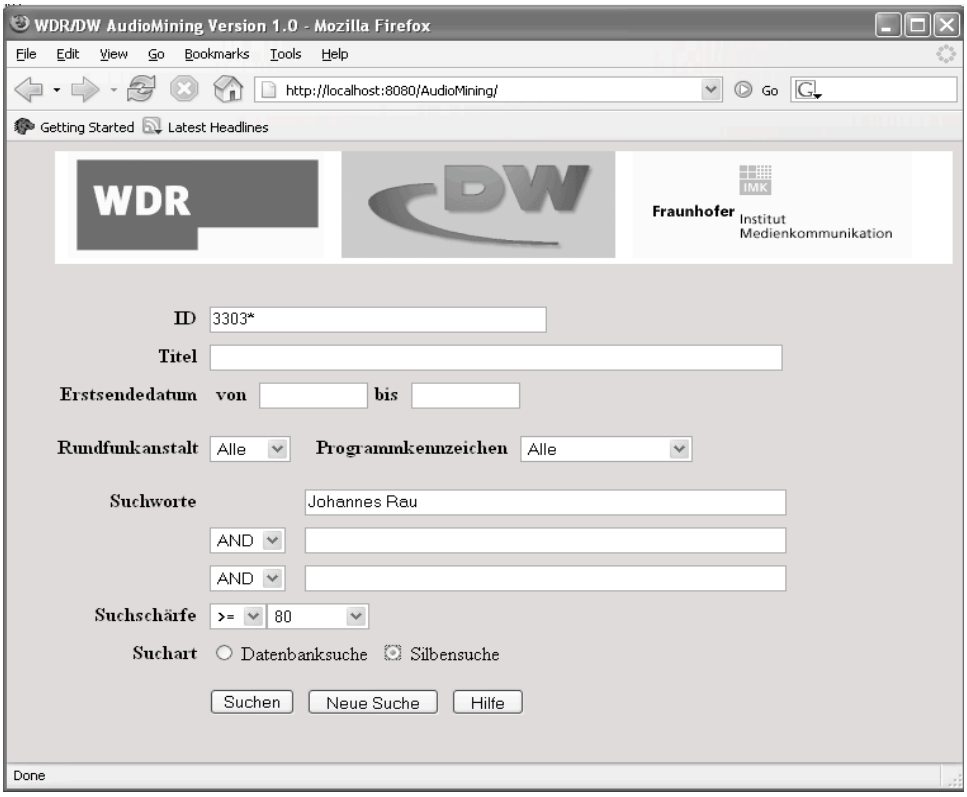

Figure 1: The search interface for the radio archive prototype

previously existing notes or titles to chose keywords and perform a search using these keywords to localize certain topics. In order to skim through an interview, archivists can jump from question to question by tracking audio segments containing the voice of the host through the audio file.

Third, it was important that the archive allow combined search on formal data and on spoken audio content. As mentioned above, many topics have indicative keywords such as "Theo van Gogh" or "Arctic Monkeys." Using the formal metadata to restrict search to shows broadcast around the time of van Gogh's death is a form of integration of world knowledge that archivists often exploit. Archivists would also tend to expand a query for information on Arctic Monkeys with the names of the band members. Archivists use outside sources or their own knowledge to implement these refinements. The project group realized how important it is to maintain possibilities of combining formal metadata with content for search and of maintaining the possibility to use familiar search strategies involving the integration of outside information. Exploring the possibilities of automatic query expansion fell outside the scope of the project.

3.3 Prototype interface

The prototype interface was a joint design created by the archive professionals and the multimedia researchers in the project group. This section discusses how the required functionality was integrated into the user interface.

Figure 1 depicts the search mask of the radio archive prototype. This mask adopts the search fields for formal metadata currently used for search in the archive metadata database. The fields are ID, title, broadcast date (specified as a range), broadcaster and station. These fields are augmented with fields that make possible search in the spoken content. In the area labeled "Suchwort," three terms can be input and joined by and- or or-operators. Archivists input orthographic words, but phonetic transcriptions of words are also accepted as input for system test and development purposes. In the field labeled "Suchschärfe," the user can input the degree of acoustic match required between the query and the transcript. This degree is a match-score which represents the distance between the input string and the syllable transcript, as will be explained in detail later. For the purpose of the interface, the match-score is chosen on a scale of 1-100, although it is not technically a percent. The match-score does not have a direct relation to the underlying algorithm, but was chosen to resemble a percent because users had best intuition of its purpose this way. Finally, the user can specify the kind of search. "Silbensuche" is syllable search in the syllable transcripts and "Datenbanksuche" is search in the database in which popular syllable searches have previously been stored. At the bottom are buttons labeled "search," "new search," and "help."

In Figure 1, the query that is being entered is for "Johannes Rau" and the system is being constrained to operate on particular archive numbers beginning with "3303." The possibility to constrain the system in this way is necessary so that archivists can continue to integrate their knowledge about the world and about the contents of the archive. Some of the functionality requested by the archive staff initially did not

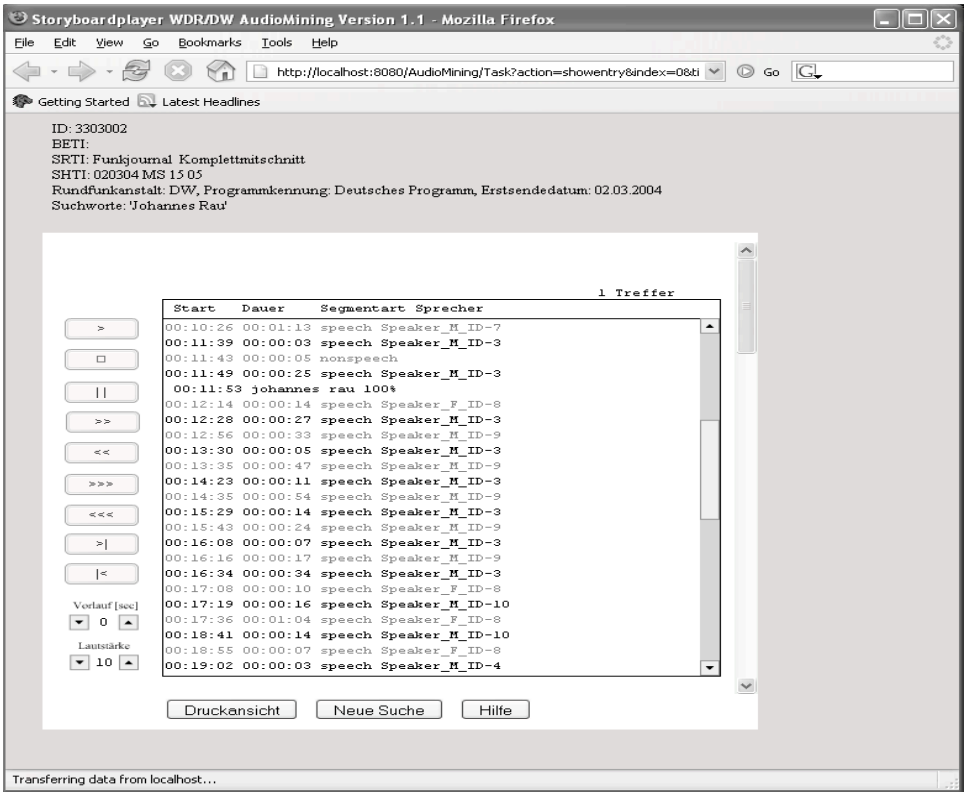

Figure 2: Display of a radio program with a keyword hit

seem particularly user-friendly to the multimedia researchers in the project group. As the researchers learned more about archivist workflows, however, it became clearer why some of the features were recommended. In this case the possibility of inputting a truncated archive number relates to the semantic structure of the archive IDs within the archive, which archive staff are thoroughly familiar with and use daily.

Figure 2 displays the structured audio player in which the retrieved radio programs are played. The programs are depicted as a list of segments. Keywords are depicted below the segments containing them together with their match scores. Note that the interface represents hits as orthographic words. The audio file can be played back using the buttons at the left, which allow for normal playback as well as accelerated playback at two different speeds. Clicking on a keyword jumps into the audio file and starts playback at the moment when the keyword is spoken. If the user wishes to add lead time, the button "Vorlauf" makes it possible to set the number of seconds before the keyword that the playback starts. The user can also adjust volume with "Lautstärke." This list can be exported for print via the button "Druckansicht". Archive staff indicated that a possible integration into the currently existing workflow would be to transfer this list per cut-and-paste from the print view to another application. This list can then form the basis for the hand generated annotations, eliminating the tedious work of setting time codes leaving archivists free to concentrate on creating high level summaries. The structured audio player is described in detail in [12].

In general, archive staff recommended a minimalist interface. Archive staff prefer to see as much information as possible on a single screen and avoid clicking "next" or scrolling to see additional hits or cuts. When the system displays results to the user query, the individual fields are marked with standard abbreviations, as can be seen at the top of Figure 2. These abbreviations are part of the daily vocabulary of (German-speaking) archivists, but appear cryptic to lay persons. In sum, the radio archive system was designed to be used on a regular basis by highly trained experts with a very different view on what is intuitive and user-friendly than a non-specialist or an occasional user.

The balance of this paper focuses on the keyword search functionality of the radio archive prototype, explaining the algorithm used to perform the fuzzy match between query and transcript and presenting the results of the evaluation of the system.

4. VOCABULARY INDEPENDENT SEARCH

The vocabulary independent search calculates a distance score between the user query and transcripts generated by automatic speech recognition (ASR). This score is calculated in a way that is intended to capture acoustic similarity. The system is required to find places in the speech recognition transcripts that "sound" the same as the user query. Search by acoustic similarity rather than by word match has the benefit of freeing the system from dependence on the vocabulary of the speech recognition. The approach promises another advantage as well. Retrieval systems that perform exact match in word transcripts are sensitive to speech rec-

ognizer errors. If a "sounds like" match is performed rather than an exact match, it opens the possibility that a spoken word or phrase is identified correctly despite the presence of speech recognition errors. The corrective power of the fuzzy-match technique applied by the radio archive prototype relies on the insight that speech recognizer errors are often caused by acoustic confusion. This section first describes the generation of the ASR transcripts and then details the distance calculation.

4.1 Syllable transcripts

The radio archive prototype implements vocabulary independent keyword search on the basis of syllable level speech recognition transcripts. The syllable constitutes a basic building block of speech. Words that are not contained in the training data can be reconstructed from the syllable transcripts by searching for the appropriate sequence of component syllables. Approaches using linguistic units at the phoneme level have long been popular [4, 3]. These units have the advantage that they form a very small and closed set, but have the disadvantage that they are too small to provide the large acoustic contexts needed for optimal speech recognition performance. Larger units, such as morphemes and in our case syllables are also popular [2, 13, 1]. Although such units do not form a closed vocabulary, it is possible to attain good coverage of a language with a relatively restricted inventory. The project also aimed to explore other advantages of syllables such as the potential for smaller, faster language models that require less training data.

The speech recognition transcripts used in the radio archive prototype are generated by the ISIP speech recognition system HMM-based speech recognition toolkit[5]. Instead of a word-level vocabulary, however, a syllable-level vocabulary is used. The language model is trained on a corpus consisting of 64 million running words from German newswire. A tri-gram syllable language model is trained by decomposing the word level text into a syllable level text using the transcription module from a speech synthesis system [15]. The syllable vocabulary contains the top 5000 most frequent syllables. Previous work has shown that at this vocabulary size, the performance of syllable recognition levels off [11]. Previous work has also shown that this tri-gram syllable model attains a syllable accuracy of 75% on studio quality speech, which is the same syllable rate achieved by our 91k word-level bi-gram language model. A 75% syllable rate was estimated to correspond to a 68% word accuracy [10].

4.2 Fuzzy syllable search

The algorithm that matches query words with acoustically similar points in the syllable transcripts is based on a two-stage Levenshtein distance. First, the query word is decomposed into syllables using the same transcription module that decomposed the training data for the syllable language model. Then, the fuzzy match algorithm calculates the Levenshtein distance between the query syllable string and each position in the syllable transcript. This Levenshtein distance is weighted using a acoustic similarity score between syllables. The acoustic similarity score is itself another weighted Levenshtein distance between the strings of phonemes that compose the syllables. The weights are calculated using confusion information derived from analyzing the performance of the speech recognizer. Substitutions between phonemes easily confused by the recognizer receive a lower penalty than substitutions between phonemes rarely confused by the recognizer. Finally, positions in the syllable transcript that receive a similarity score above a certain threshold are hypothesized by the system to be hits for the query word. This threshold is determine empirically and reflects the "fuzziness" or "exactness" of the match between the query and the hit. The interface gives the user the ability to adjust the match-score threshold, providing control over the tradeoff between precision and recall.

5. EVALUATION

The system was evaluated by using 213 queries that were chosen by archive professionals to reflect the kinds of information requests they receive. The queries consisted of both single words and multi-word phrases and the system was required to return the positions in the audio files at which the word or phrase was spoken. A lot of effort was devoted to creating a representative and well-distributed query list, since the performance of the system was to be evaluated on the basis of whether or not it was able to provide archivists with appropriate responses.

First and foremost the system was evaluated on speech recorded in studio conditions. The project group placed primary emphasis on attaining adequate performance under studio conditions since the group was pessimistic about potential retrieval performance on telephone speech, speech recorded on the street, or speech with music or foreign speech background. Notice that studio speech comprises the greater portion of the interview talk show *Montalk*. Recall that *Montalk* was particularly important for the radio archive prototype since it contains long expanses for which no production metadata are available. *Montalk* stands to benefit greatly from being made accessible to archivists through content-based keyword search. *Montalk* is also the most challenging of the 4 programs contained in the archive because it contains nearly exclusively spontaneous speech.

Retrieval performance was tested on 12 hours of material from the radio prototype archive, 4 hours each from *Montalk* and *Der Tag* and 2 hours each from *Funkjournal* and *Wiso*. Roughly estimated, the test material contains 50% spontaneous speech and 10% music and commericals. The data was annotated with segment boundaries between speakers and between speech and non-speech. Each segment was assigned a label relating to the acoustic quality of the audio in that segment. The data were transcribed by a professional transcription service and then automatically aligned with the audio files, so that each individual word spoken was associated with a time code. Then, human annotators listened to all 12 hours and checked the time codes of the words, correcting the alignment when the ASR alignment software had committed an error. The remainder of this section details the tests that were performed on this test set using the archivist-defined queries.

5.1 The effects of fuzzy match

As previously mentioned, the user interface provides the user with control over how much acoustic mismatch the system should admit between the query and the syllable transcripts. Table 1 reports precision, recall and F1-Value for five differ-

Table 1: System Performance at different match-score levels

Level	Precision	Recall	F1-Value
60	0.28	0.72	0.41
70	0.45	0.62	0.52
80	0.56	0.51	0.54
90	0.5	0.39	0.44
100	0.45	0.31	0.37

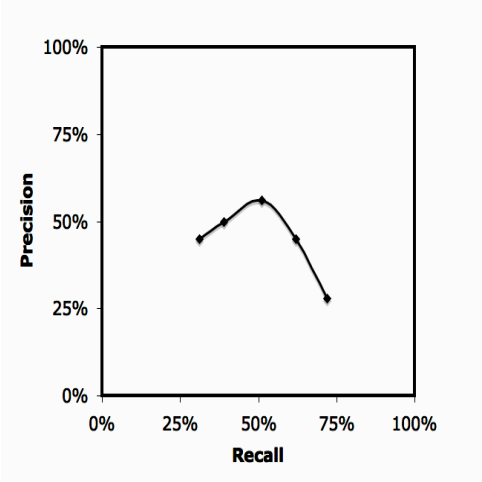

Figure 3: Recall vs. Precision on studio quality audio

ent settings of the match-score parameter. At low match scores, precision is low and recall is high. An interesting effect is that precision hits a peak and then deteriorates as the match is forced to be more and more exact. This effect is due to the fact that at high levels of exactness, many correct matches are not longer contained in the list of the hits returned by the system and false positives returned due to ASR errors dominate the shorter hit list. Figure 3 shows plots precision vs. recall, illustrating clearly the rise and fall of precision. After the radio archive prototype had been implemented, archivists experimented with the system and discovered that the most useful operating point was one at which the precision and the recall were approximately matched, i.e. the break-even point. This operating point occurs at match-score 80. In the remainder of this discussion, the system is evaluated at match-score 80, since this point had the highest utility for supporting archive workflows. Although it was disappointing that the system did not achieve a higher level of performance, the advantages of the fuzzy-match approach are clear, since the system at match-score level 80 clearly outperformed exact match (i.e. system at match-score 100).

5.2 The effects of implicit decomposition

A side-effect of the fuzzy-match approach used by the radio archive prototype is to introduce an implicit decomposition into the keyword search. The speech recognition transcripts are strings of syllables and do not represent word boundaries, which are not hypothesized by the ASR system. For this reason, a position in the speech transcripts that has a high match score with respect to the user query might actually correspond to a partial word in the spoken audio. There is no simple strategy for implementing a way to "turn off" this effect. The implicit decomposition means that a user query *Kinder* ("children") will return points in audio files at which the word *Kindergarten* ("kindergarten") is spoken. For this example, returning a compound containing the query word is probably not going to hinder the archivists' work. Indeed *Kindergarten* does have relevance to children. However, the situation is different if the original query was for *Garten* ("garden"). Here the system also returns points in audio files at which *Kindergarten* is pronounced. Such hits are clearly semantically far afield from the original query and effectively lower the precision of the system.

Table 2: Effects of implicit decompounding at match-score 80

Type	Precision	Recall	F1-Value
Admit partial word matches as correct	0.56	0.51	0.54
Exclude partial word matches as correct	0.52	0.53	0.53

Table 3: Effects of adding out-of-studio, telephone quality and music-background material at match-score 80

Type	Precision	Recall	F1-Value
Studio quality only	0.56	0.51	0.54
All except music	0.51	0.46	0.48
All audio qualities	0.51	0.45	0.48

The project group decided to evaluate the performance of the system under the stringent requirement that only exact matches be counted as correct hits. The purpose of this evaluation was to determine to what degree the return of compound words containing the query word lowered the precision of the system. The results on studio speech at match-score level 80 are reported in Table 2. If partial words (i.e. compound sub-units) are excluded as correct hits the precision declines somewhat. At the same time, recall improves slightly, since the system no longer was required to find every instance of an acoustic match. The over-all effect was only a slight, possibly insignificant, deterioration of system performance. If the query list compiled by the project group is taken to be representative of the queries that an archivist would submit to the system during the normal course of responding to information requests, it can be concluded that the implicit compounding of the system is not an aspect of the prototype that will cause an increased burden on archivists in their work.

5.3 The effects of including non-studio speech

At the end of the project, the performance of the radio archive prototype was evaluated on all speech types in order to ascertain what kind of deterioration of performance could be expected. Table 3 provides a comparison of system performance on studio speech with performance on studio plus non-studio speech and with performance on all audio qualities, including spoken audio with music background. It can be seen that when the system moves beyond its self-imposed restriction to studio speech, performance deteriorates. The

same pattern of deterioration was observed for all levels of match-score, although it is reported here only for match-score 80. The level of deterioration is not such that it would motivate the exclusion of non-studio speech from the system. Indeed the performance of the system on all audio qualities introduced a tolerable drop in system precision.

6. CONCLUSIONS

The radio archive prototype discussed in this paper was built with the goal of acquiring a concrete, practical understanding of how automatic digital audio processing can be integrated into the existing workflows in archive departments at large radio broadcasters to support the work of archive staff. This focus of this paper was vocabulary-independent keyword search. It was shown that the syllable based fuzzy search algorithm delivers tolerable performance without relying on a pre-defined vocabulary and is not derailed by acoustically challenging audio. Although the prototype was designed to retrieve keywords from German audio only, the fact that the system occasionally returns a proper name from the small fraction of English audio in the archive suggests that the method holds promise for keyword search in a multilingual archive.

The tests performed on the system demonstrate that higher precision rates can only be achieved with significant sacrifices in the area of recall. Archivists do not consider the system at its current level of performance to provide significant support to their workflow. This conclusion must be seen against the backdrop of the fact that many requests for information can be satisfied by searching formal metadata only or are so abstract that they could not be met using keyword search, even if the performance were perfect.

The keyword search functionality implemented in the radio archive prototype demonstrates three aspects which confirm its clear potential in the future for archive staff support. First, vocabulary independent archive access is indeed possible. Second, the implicit de-compounding that is a by-product of the fuzzy-match search approach has a very limited negative impact on precision. Third, inclusion of all types of audio in the archive and not just audio recorded under studio conditions does cause deterioration of system performance, but not to an extreme degree. In sum, the vocabulary independent keyword search method implemented in the radio archive prototype continues to hold promise for the future, even in the face of the challenges offered by a collection containing a large amount of spontaneous speech such as occurs in interview talk shows.

7. ACKNOWLEDGMENTS

We would like to acknowledge Deutsche Welle and WDR, who commissioned the project and provided the data. Thank you to the members of the archives departments who worked in the project group. The final work for this paper was carried out by the first author while being supported by the EU IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104.

8. REFERENCES

[1] G. Choueiter, D. Povey, S. Chen, and G. Zweig. Morpheme-based language modeling for Arabic LVCSR. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 1, 2006.

[2] M. Elbeze and A. Derouault. A morphological model for large vocabulary speech recognition. *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 577–580, 1990.

[3] A. Ferrieux and S. Peillon. Phoneme-level indexing for fast and vocabulary-independent voice/voice retrieval. *ESCA ETRW on Accessing Information in Spoken Audio*, pages 60–63, 1999.

[4] J. Foote, S. Young, G. Jones, and K. Sparck Jones. Unconstrained keyword spotting using phone lattices with application to spoken document retrieval. *Computer Speech & Language*, 11(3):207–224, 1997.

[5] A. Ganapathiraju, N. Deshmukh, J. Hamaker, V. Mantha, Y. Wu, X. Zhang, J. Zhao, and J. Picone. ISIP Public Domain LVCSR System. *Proceedings of the Hub-5 Conversational Speech Recognition (LVCSR) Workshop*, 1999.

[6] J. Garofolo, C. Auzanne, and E. Voorhees. The TREC spoken document retrieval track: A success story. *Text Retrieval Conference (TREC)*, 8:16–19, 1999.

[7] J. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2):89–108, 2002.

[8] N. Hans and J. de Koster. Taking care of tomorrow before it is too late: A pragmatic archiving strategy. *116th convention of the Audio Engineering Society*, May 2004.

[9] A. Hauptmann and M. Witbrock. Informedia: News-on-demand multimedia information acquisition and retrieval. *Intelligent Multimedia Information Retrieval*, pages 215–239, 1997.

[10] M. Larson and S. Eickeler. Using syllable-based indexing features and language models to improve German spoken document retrieval. *Eurospeech'03*, pages 1217–1220, 2003.

[11] M. Larson, S. Eickeler, K. Biatov, and J. Köhler. Mixed-unit language models for German language automatic speech recognition. *Elektronische Sprachsignalverarbeitung, Tagungsband der*, 13:127–134, 2002.

[12] M. Larson and J. Köhler. Structured Audio Player: Supporting radio archive workflows with automatically generated structure metadata. *Proceedings of RIAO 2007*, 2007.

[13] K. Ng. *Subword-based approaches for spoken document retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.

[14] G. Rigoll. The ALERT system: Advanced broadcast speech recognition technology for selective dissemination of multimedia information. *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pages 301–306, 2001.

[15] K. Stöber, P. Wagner, J. Helbig, S. Köster, D. Stall, M. Thomae, J. Blauert, W. Hess, R. Hoffmann, and H. Mangold. Speech synthesis by multilevel selection and concatenation of units from large speech corpora. *W. Wahlster (ed.), Verbmobil: Foundations of speech-to-speech translation*, 2000.