# Two Novel Applications of Speech Recognition Methods for Robust Spoken Document Retrieval

Stefan Eickeler, Konstantin Biatov, Martha Larson, Joachim Köhler
Fraunhofer Institute for Media Communication
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
e-mail: stefan.eickeler@imk.fraunhofer.de

## Abstract

This paper describes two novel applications of speech recognition technology to the indexing of large spoken word collections, an important challenge for Digital Libraries. First, the speech recognizer is used to generate a string of syllables, which can be searched for index terms. This technique circumvents the disadvantages of conventional speech recognition, including restriction to a fixed vocabulary and high sensitivity to spoken audio quality. Audio indexing and retrieval using fuzzy syllable search is described for a collection of radio documentaries from the Deutsche Welle. Second, the speech recognizer is used to align previously existing stenographic transcripts with the corresponding audio files, making it possible for the user to jump between the two media while browsing. A by-product of the alignment procedure is a set of automatically generated keywords, which can be used as indexing terms. With this technique, the speech recognizer enhances the value of a spoken word collection by combining two existing sources of information. The alignment functionality is integrated into our distributed multimedia archiving and retrieval system, called iFinder, and is described here as applied to a collection of videos recorded in the German Parliament.

*Keywords: Speech recognition, spoken document retrieval, MPEG-7,*

## 1 Introduction

Conditions are currently converging to create an unparalleled opportunity in the area of Digital Libraries. Today's technologies make it possible to easily accumulate multimedia collections and digital storage is simple and low cost. At the same time, standards for the description of multimedia content, such as the ISO MPEG-7 standard [1], have been established, allowing for the creation of metadata compatible across platforms and across applications. These circumstances have primed user expectations. The demand for software that makes retrieval, browsing and document summary in large multimedia collections possible has increased accordingly. The need for flexible, intelligent archiving solutions that make use of automatic methods to generate metadata for huge multimedia collections is self evident.

This paper discusses two novel applications of speech recognition technology to the task of automatically generating the metadata needed to allow information retrieval on large spoken word collections. The first application uses the speech recognizer to generate a string of syllables, which can be searched for indexing terms using fuzzy comparison techniques. The second uses the speech recognizer to align previously existing human transcripts with the spoken audio file. By using the speech recognizer to generate syllables or to perform alignment, we are applying speech recognition technology, not to the traditional task of generating an orthographically correct transcript, but rather directly to the generation of metadata optimized to provide search and browsing capacity for large spoken word collections.

We demonstrate our approach on two applications developed within projects funded by the German Ministry of Education and Research. First, the fuzzy syllable search is demonstrated with a syllable search software from the Piavida project. Second, the alignment procedure is demonstrated with the iFinder system, a distributed multimedia archiving and retrieval system, which is an outgrowth of the AGMA project. In the second section of the paper we discuss speech recognition technology and the novel ways in which we apply it to spoken document retrieval. In the third section of the paper we discuss fuzzy syllable search software and its application to archiving a collection of radio documentaries from broadcaster Deutsche Welle. In the fourth section of the paper we discuss the alignment of text and human-generated transcripts. The alignment technology has been implemented in the iFinder system and we discuss the application of this system to archiving a collection of videos from the German Parliament. The fifth section offers conclusion and outlook.

## 2 Speech recognition methods for metadata generation

Speech recognition technology is an important source of two types of metadata for spoken word collection. First, a speech recognition system generates a text transcription of spoken audio. Second, it generates a series of time markers encoding at which point in the audio file each word was spoken. Text transcriptions allow spoken word collections to be searched with the same ease and accuracy as collections consisting of text documents. Time synchronization of text and audio makes it possible jumping freely from one medium to the other, enabling users to effectively browse spoken audio as they would plain text.

### 2.1 Generating syllable transcripts of spoken audio

In a standard spoken document archiving and retrieval systems speech recognition technology is applied in its most generic form. The speech recognizer generates a transcript constrained to containing words that it has recognized from its internal, fixed vocabulary. While this vocabulary might include several hundred thousand words, the people, place and concept names important for information retrieval, are frequently absent. The word recognition rate of the recognizer falls off sharply in the presence of background noise, sound effects, and overlaid voices commonly found in real-world spoken audio collections. If one part of the word presents acoustic difficulty for the recognizer, the entire word will be recognized incorrectly.

We compensate these difficulties by training a syllable language model [2]. We chose syllables since they are longer and thus more acoustically distinct than phonemes, and are to be preferred as a base unit for speech recognition. From previous work we have been able to conclude that syllables are long enough to carry distinguishing semantic information [3]. With a closed vocabulary of 5,000 syllables, seldom seen or previously unseen words can be built compositionally by the recognizer during the recognition process. A recognition mistake on the syllable level does not automatically mean that the entire word is wrong. Semantic content can still be derived from adjoining correctly recognized syllables. Using approximate matches between keywords and syllable sequences generated by the speech recognizer, the fuzzy syllable match technique is able to locate indexing terms in spoken audio.

### 2.2 Aligning spoken audio and imperfect text transcripts

Alignment of audio and text transcriptions is fairly standard application of speech recognition technology. The recognizer is given the sequence of words that it must recognize and is required to generate the time markers at which the given words appear. With standard alignment, however, the text transcripts must be perfect. If there are additions or omissions, the alignment will go awry and the time markers will be incorrect. Stenographers do not transcribe speeches word for word, but rather record the meaning of what is said. A stenographic transcript and a literal transcript might only share 70% of the words in common. A surprising number of spoken document collections with parallel transcripts exist, most notably parliamentary proceedings. We modify the speech recognizer so that it is able to perform an alignment despite the great difference between the text transcript and the words spoken in the spoken audio.

## 3 Indexing Radio Documentaries with Syllable Search

We implemented the syllable search technique in a fuzzy syllable search software. The software was developed in the framework of the Piavida project, which concerns the indexing and classification of multimedia data for personalized interactive portal applications (see www.igd.fraunhofer.de/igd-a7/piavida/start.html). Here we describe the application of the software to the identification of index terms for the archival of the Deutsche Welle *Kalenderblatt* database, a collection of radio broadcasts.

### 3.1 The Deutsche Welle *Kalenderblatt* Database

The Deutsche Welle *Kalenderblatt* spoken audio collection contains 850 short radio documentaries from the radio series *Kalenderblatt*. These programs are available on the Internet at www.kalenderblatt.de. The collection is quite heterogeneous since the programs deal with a broad palette of historical, cultural and current interest. Each program contains approximately 650 running words and is about 5 minutes long. This collection represents a truly challenge for the automatic generation of metadata, since the documentaries are each enhanced with background music and sound effects and contain the voices of many different speakers, some overlapping, some speaking foreign languages. These factors contribute to the difficulty of speech recognition and challenge automatic metadata generation as well.

### 3.2 Implementation of fuzzy syllable search

The speech recognition component was build with the ISIP speech recognition toolkit [4], a standard HMM-based speech recognizer. We trained a syllable trigram language model on a text corpus from the German dpa newswire consisting of 64 million words. The transcription module of the Bonn Open Source Synthesis System (BOSSII) [5] was used to decompose the text corpus into syllables. The result of the recognition stage is the sequence of recognized syllables with time stamps.

The syllable transcriptions are stored as metadata accompanying the spoken audio documents. These transcriptions can be used in one of two ways. They can be scanned for a fixed list of indexing terms at file time, or they can be dynamically scanned for query terms at query time. The syllable string is scanned for a keyword in the following manner. First, the keyword is broken down into a syllable string containing its component syllables. The distance between the keyword syllable string and syllable sequences in the transcriptions is

calculating using the minimal edit distance weighted by a syllable similarity score. The syllable similarity score is calculated by breaking both syllables into their constituent phonemes and comparing the phoneme strings using the minimal edit distance weighted with phonetic confusability. The fuzzy syllable search returns an identified keyword whenever the distance between the keyword syllable string and a syllable string in the recognizer output is less than an empirically determined threshold. The time stamp makes it possible for the user to browse the collection by clicking on a keyword and jumping directly to the position where it is spoken.

## 4 Archiving recorded parliamentary proceedings with their stenographic transcripts

We propose a technique which exploits the existence of imperfect transcripts, which are in a surprising number of cases available parallel to spoken audio collections. We automatically extract keywords from the spoken audio and use them to help perform an alignment between imperfect transcripts and spoken audio. The alignment technique is implemented in the iFinder system [6]. We describe the application of this software to the task of archiving videos recorded in the German Parliament using the accompanying stenographic transcripts.

### 4.1 Stenographic transcripts and video recordings from the German Parliament

The sheer volume of German Parliament video that exists makes this domain a classic case in which automatic methods are needed to generate metadata. German Parliament recordings represent a challenging domain, since the audio is not of extremely high quality and contains much background noise. The spoken audio differs in speech quality from almost read speech to very spontaneous dialogues. A typical day in the Parliament has a duration of 6 hours. Some speeches are only a few minutes, other speeches are more than 40 minutes. The AV-data was recorded as MEPG-2 stream from the Phoenix TV-channel. The stenographic transcripts used for automatic alignment were downloaded from the German Parliament website, www.bundestag.de where they are available in .pdf format. Because we work with a version of the video recordings broadcast on TV, the audio track is interspersed with the commentary of the different TV moderators, which often overlaps with the voice of the speaker on the floor of the Parliament. The presence of the voice of the TV moderator means that the correlation between the stenographic transcripts and the recorded spoken audio is quite loose. If a perfect word-for-word transcription of the spoken audio is used as ground truth, the stenographic transcripts have about a 70% word accuracy.

### 4.2 Automatic selection of salient words

We implement an automatic method which selects informative and reliable words from the speech recognizer output. First, the audio of the parliamentary proceedings for a single day is broken down into usable segments by using a silence detector to determine the position of pauses. No segment is allowed to exceed 30 seconds in length. A speech/non-speech classifier is used to discard segments containing only applause or other extraneous noise. Next, the segments are transcribed by the speech recognizer, again we use the ISIP speech recognition toolkit [4], using a language model trained on the stenographic transcripts for that day. We remove all common words from the output of the speech recognizer. We also remove all words that cannot be verified as having been reliably recognized. For the rest of the words in the speech recognizer output we calculate the conditional entropy of speech fragment $f$ under condition the keyword $k$.

$$H(f) = p(f/k)\log P(f/k)$$

For each speech fragment we select $N$ keywords with the largest conditional entropy. This list of keywords includes the most salient words with respect to the speech fragments. These keywords are used directly as indexing terms for the spoken audio from the German Parliament. The keywords are also used as landmarks to help align the spoken audio with the stenographic transcripts.

### 4.3 Implementation of the alignment of imperfect transcripts

A full day (about 6 hours) of spoken material is too long to align to the transcripts in a single step, and for this reason we use a two-stage alignment process, performing first a global alignment and afterwards a local alignment. During the global alignment stage we use the keywords extracted with the procedure described in the previous subsection to align the recognizer transcription of each audio segment with a neighborhood of sentences in the stenographic transcripts. The neighborhoods of sentences assigned to adjoining audio segments overlap. In this way we prevent global alignment error from feeding error in local alignment.

During the local alignment stage, each sentence in the text neighborhood corresponding to an audio segment is considered in turn. Each sentence is used to build a finite state automaton, in the manner described in [7]. The speech recognizer recognizes the audio segment using each finite state automaton in turn. The automaton corresponding to the correct sentence for the audio segment is assumed to be the automaton that yields the most reliable recognition result. If no automaton yields a recognition result that can be verified as reliable, the speech

segment can be assumed to be moderator speech. A second iteration of local alignment can be applied in order to insure that each text sentence is assigned to only one audio segment. The output of the alignment procedure is a time stamp for each sentence of the stenographic transcript. The time stamps allow the implementation of the text-video browser functionality of the iFinder system, which enables the user to jump between browsing text and browsing video, and is described in the next section.

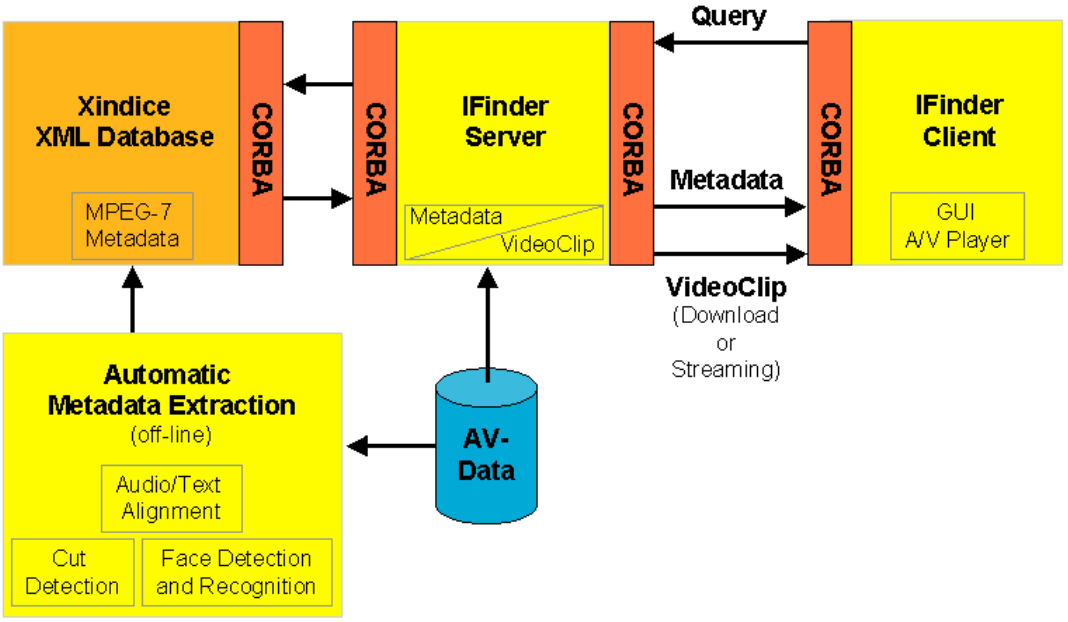


**Figure 1:** Architecture of iFinder system

### 4.4 Automatically generated metadata in the iFinder system

The iFinder system is a multimedia system for archiving large audio visual databases. The iFinder system has been conceptualized and designed for media editors, journalists, historical researches or users from the general public searching for information. The system is being developed at Fraunhofer IMK and had its inception in the AGMA project on automatic generation of MPEG-7 compliant metadata, (see www.imk.fraunhofer.de).

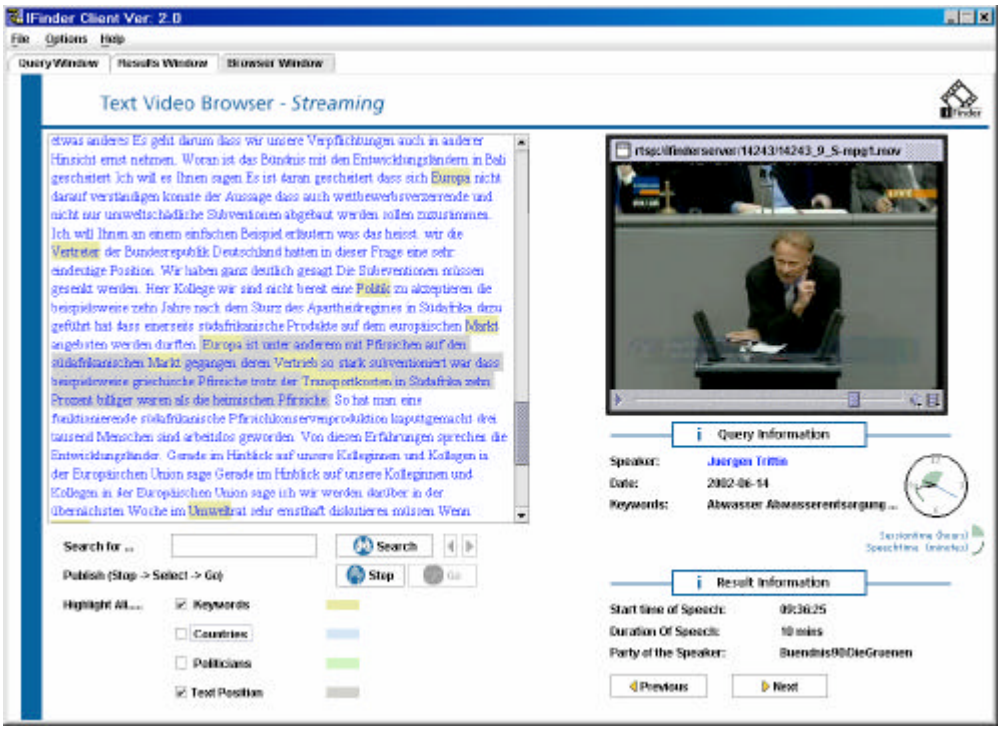


**Figure 2:** Screenshot of the text-video browser

The iFinder system is a distributed environment based on a client/server architecture. The system design enables multiple users in a distributed environment to access several multimedia archives containing audio-visual content and the accompanying metadata. The key characteristics of the iFinder system are its support of MPEG-7-based metadata and its use of an open system architecture, which can be extended and integrated in other archiving environments very easily. The architecture of the iFinder system is shown in figure 1. The main system components are the iFinder server and client applications, an XML database for metadata storage and a metadata extraction module. The recognition and alignment techniques described in the previous subsection are integrated into the iFinder system in the metadata extraction module.

The metadata with which the parliamentary recordings have been annotated makes it possible to search by keyword, date or by politician name. An important feature of the iFinder system is the text-video browser, pictured in figure 2. The text-video browser is used to display the result of a query. The user is able to simultaneously browse the stenographic transcripts and the video of the parliamentary proceedings, and freely switch between the two media. On the right hand side is the video window. On the left hand side is the text window with the current sentence highlighted. Additionally some named entities are highlighted. The user is able to navigate in the text-video stream by clicking on sentences or using the scrollbars of the video window.

## 5 Summary and Outlook

This paper has described two useful applications of speech recognition to the problem of automatically generating the metadata needed to archive large volumes of spoken audio material. By using the speech recognizer to produce a string of syllables, which can be scanned for the presence of keywords, we are able to mitigate common problems of generic speech recognition, including vocabulary-dependence. By using the speech recognizer in a two-stage alignment process we are able to align imperfect transcripts with spoken audio files, effectively using an existing resource to generate high quality metadata. Spoken audio with synchronized transcripts is effectively as searchable and browsable as text since it enables the user to jump freely back and forth between the two media.

Future work will include the integration of the indexing functionality of the fuzzy syllable search into the iFinder extraction module. Additionally, we plan to address the problem of providing a compacter representation of long multimedia documents. We believe that providing document summaries or abstracts of the multimedia documents will enhance the convenience of the iFinder system and increase user comfort. Digital Libraries cannot be developed for international use unless the problem of multilinguality is addressed. In the future we will be porting the techniques described here to languages other than German, the language for which our system is now optimized. Developing multilingual multimedia archiving and retrieval systems would be a fruitful area of cooperation with international research teams from other countries.

## 6 Literature

[1] Martinez, J. M., *Overview of the MPEG-7 Standard,* ISO/IEC JTC1/SC29/WG11 N4031, 2001.

[2] Larson, M., Eickeler, S., Biatov, K., and Koehler, J., "Mixed-unit language models for German language automatic speech recognition," *Proceedings of the 13. Konferenz Elektronische Sprachsignalverarbeitung*, 2002.

[3] Larson, M., Eickeler, S., Paaß, G., Leopold, E., and Kindermann, J., "Exploring sub-word features and linear support vector machines for German spoken document classification," *Proceedings of the International Conference on Spoken Language Processing,* 2002.

[4] Picone, J., et al., *A public domain decoder for large vocabulary conversational speech recognition*, Mississippi State University, 1998.

[5] Stöber, K., et al., "Speech Synthesis by Multilevel Selection and Concatenation of Units from Large Speech Corpora," In: W. Wahlster, ed., *Verb-mobil*, Springer, 2000.

[6] Löffler, J., Biatov, K., Eckes, C., and Köhler, J., "iFinder: An MPEG 7-based Retrieval System for Distributed Multimedia Content, " *Proceedings of the ACM Multimedia Conference*, 2002.

[7] Biatov, K. and Köhler, J., "Methods and Tools for Speech Data Acquisition Exploiting the German Parliamentary Speeches Database and Transcript from Internet," *Proceedings of the Third International Conference on Language Resources and Evaluation*, 2002.