Structuring and indexing digital archives of radio broadcasters

Martha Larson
Fraunhofer IMK
Schloss Birlinghoven
53754 Sankt Augustin
Germany
martha.larson
@imk.fraunhofer.de

Thomas Beckers
WDR
Dokumentation & Archive
50600 Köln, Germany
thomas.beckers@wdr.de

Volker Schlögell
Deutsche Welle
Archive-Bibliothek-Dokumentation
53113 Bonn, Germany
volker.schloegell
@dw-world.de

Abstract: This paper describes a pilot project being undertaken by Westdeutscher Rundfunk (WDR) and Deutsche Welle (DW) in cooperation with Fraunhofer Institute for Media Communication (IMK). The project goal is to ascertain the practical usefulness of automatic approaches for the structuring and indexing of digital audio archives of radio broadcasters. Automatic approaches have an enormous potential to complement the conventional annotation methods of radio archivists, who are rapidly becoming overwhelmed by ever-increasing amounts of material that must be archived and growing demands for completely searchable audio collections. Automatic segmentation methods can set cue points in audio broadcasts, which make it possible to skim audio quickly using large intuitive jumps. Classification of segments as speech or non-speech and clustering of speech segments into groups of segments spoken by the same speaker further facilitates browsing. As an additional step, a sort of automatic indexing can be implemented by feeding structured audio through a syllable-based speech recognizer, and performing full-text searches for query words on the resulting syllable transcripts.

1 Introduction

Radio broadcasters maintain extensive audio archives in which radio programs and other recordings of speech or music are reposited, stored either on magnetic tapes or as digital sound files. Audio archives serve as a record of broadcast material; long-term storage preserves an important historical record and short term storage is required by law. Audio archives are also a critical resource for creation of new radio content, since they make it possible to enrich new radio productions with the addition of original material. Of particular value are sound clips containing quotes from public figures of political, historical or cultural importance. As radio broadcasters move to completely digitized workflows encompassing production, broadcast and archival, demand continues to grow for easy access to digital archives for the purpose of research or re-use of audio content.

In order to be accessible, audio material stored in the audio archive of a radio broadcaster must be linked with a formal description (e.g. title, program, date of broadcast, producer) and must be annotated with a description of its content. The size of audio archives is growing at such a rate that it is no longer possible for archivists to

process the huge amounts of incoming new material using exclusively conventional methods. Currently, departments responsible for archiving and preservation are not able to fully meet the requirements of producers/editors, who wish to access archived audio resources for the creation of new productions. In order to reconcile tried-and-true conventional archiving methods and the large-scale archiving methods needed to handle today's volume of digital audio, Westdeutscher Rundfunk (WDR) and Deutsche Welle (DW), two prominent German radio broadcasters, have launched a pilot project in cooperation with Fraunhofer IMK. The project will evaluate the potential of large-scale automatic analysis of digital audio for the generation of metadata suitable for archiving radio broadcasts. This paper describes the scope and goals of this project, which is currently ongoing. The first section depicts the conventional archiving methods and practices currently used. The second discusses automatic structuring of audio. The third describes an approach that makes it possible to find query words directly in the audio essence and provides an outlook for the future of automatic analysis for digital archives.

2 Audio archives of radio broadcasters

A critical division of any large radio broadcaster is the archival department, which bears the responsibility of archiving broadcast audio content in a manner such that it is possible to search for certain specific clips or specific content. On a regular basis the archival department is called upon to provide segments, reports or entire programs from the archives for use in new radio productions. Additionally, the audio archive of a large radio broadcaster reflects the temporal development of broadcast content, and as such fulfills the important function of repository for part of a common cultural memory [Pl04][Sp04].

Audio archives have been accessible for quite some years by means of databases in which metadata describing the audio material is stored. The metadata are cross-referenced with the audio files, so that a search of the metadata base will yield an exact identification of the relevant program as well as time codes of points in the program relevant to the search. The metadata describe the audio material with different levels of granularity and at different levels of detail. The description contains formal information (title, broadcast date, producer) as well as technical information (format, production means) about the audio essence. It also contains information concerning its copyright status. Importantly, the metadata describe the content of the audio material. In a conventional archival department, all metadata are created manually and entered into the database by highly specialized archivists. It is these metadata that make it possible to carry out precise searches in the audio archive [He01].

The responsibilities of the archivists are broad in scope. An archivist is in charge of selecting which programs or contributions are to be archived and for deciding at what level of details these materials should be annotated and how long they should be retained in the archive. Not all broadcast material receives the same level of attention. Archival material judged to be of historical importance or potentially valuable for reuse is annotated at a high level of resolution and stored in a long-term archive. Annotation of content takes the form of indexing audio material with keywords from a controlled

vocabulary and of creating abstracts summarizing the content of the audio material. Annotation is required to be specific and detailed, since it is not possible to anticipate completely which aspects of the archived content will be important for later search and re-use. Depending on the context of the search and on the point-of-view of the searcher, future searches may be made for the same audio material using radically different query formulations. Manual annotation is a costly process and takes between 3 and 5 times as long as the play-length of the audio. For example, the annotation of a 5 minute report can take up to 25 minutes.

The requirements placed on the archival departments are changing, presenting new challenges. As previously mentioned, audio archives are growing, meaning that an increasingly large volume of audio content must be annotated. Also, radio broadcasters are moving from conventional production systems to completely digital production systems. Producers are demanding fast, exact and intuitive access to archived material. These considerations prompted the initiation of a pilot project to determine the usefulness of automatic approaches for structuring and processing archives that can complement current annotation and indexing procedures and make it possible to store even the lower priority material, which is currently discarded due to archiving costs.

3 Automatic structuring of digital audio archives

Audio collections are notoriously difficult to browse, since unlike text or images, sound inherently extends through time. A useful type of audio browsing could be implemented if cue points were set in the audio at reasonable intervals. With such markers, it would be possible to fast forward or skip over irrelevant sections. For example, if a researcher is interested in the current career of an actor, it would be possible to skip quickly through parts of an extended interview in which the actor discusses his childhood in detail. The setting of such cue points by hand as time markers when the audio file is reposited in the archive is unthinkably expensive. The pilot project has confirmed, however, that automatic segmentation approaches for digital audio recordings make it possible, without human intervention, to set cue points at intuitive time points that facilitate quick jumping through spoken audio. Automatic segmentation is performed using the Bayesian Information Criterion [TG99]. This approach sets cue points at the boundaries of segments containing homogenous audio, for example when the moderator stops talking and the interviewed guest starts or where the news broadcast finishes and music starts.

Further structure can be added to audio materials by classifying audio segments between cue points as either speech or non-speech. The project has determined that current technology performs speech/non-speech classification at a level of robustness that makes such classification useful for archiving and for search in archives. The methods for speech/non-speech classification being developed at Fraunhofer IMK have been reported on in, e.g. [BK03]. A speaker can be presented with a visualization of a radio broadcast in which speech and non-speech parts are marked in contrasting colors. Not only does such a visualization provide an impression of what percent of the broadcast is dedicated to spoken content, it also makes it possible to browse the audio by playing only the speech sections and completely skipping the music.

Finally, audio materials can be given even more structure by identifying all audio segments in an audio broadcast that have been spoken by the same speaker. A speaker can be tracked throughout an entire broadcast and the broadcast can be visualized by marking all segments spoken by the same speaker in the same color. Speaker tracking is performed with speaker clustering algorithms, also under development at Fraunhofer IMK. The benefits of speaker tracking are clear: in an interview program it becomes possible to listen only to the replies of the interviewee and skip segments spoken by the moderator as well as any intervening news programs or traffic announcements.

The project places major focus on setting cue points, speech/non-speech differentiation and speaker tracking, three ways in which structure can be added to radio broadcasts on which the project places its major focus. Other types of structure are also being investigated, or considered for investigation. These include classification of speakers as male or female (intended to provide further information for skimming/browsing) and classification of speakers as speaking in the studio, out of the studio, in an auditorium or over a telephone line (intended to provide information as to whether the audio segment is of the quality necessary to be re-used as a clip.) Separating music from other non-speech will also be explored and particular emphasis will be placed on identifying occurrences of particular jingles. Information concerning jingle occurrence helps to identify programs and provides necessary documentation to billing departments that pay for rights to use jingles by amount of airtime. Speaker tracking represents a clear first step in the direction of speaker identification and the experience being accumulated in the project supports the vision that radio archives will include speaker profiles that will allow known speakers to be detected in new audio material.

4 Outlook: Keyword search in spoken audio and beyond

The goal that research in the area of spoken document indexing and retrieval strives to attain is to make spoken audio as easily searchable as text, without the aid of manual pre-processing. Currently an exact quote spoken by, for example, a high-profile politician cannot be found in an audio archive unless an archivist has previously located the quote, transcribed it by hand and entered the transcription along with an accurate time marker into the metadata base. An important goal of the project is to evaluate in how far speech recognition technology can be used for the automatic generation of transcripts of radio broadcasts that would make possible query- word search in spoken audio documents. Speech recognition systems process audio files and produce a transcript of what was spoken in those files. Ideally, a full text search would be performed on this transcript, constituting a reduction of spoken audio search to full text search. However, such an ideal system cannot be realized due to the errors produced by the speech recognition process. Top recognition performance can only be achieved on audio recorded in studio conditions spoken by trained speakers. Although these conditions sometimes hold in the case of broadcast news, typical radio content is backgrounded by music or atmospheric effects to make it more interesting. Interview guests are not trained speakers and interviews are characterized by stops, starts incomplete sentences, interruption and often chuckles if not outright laughter. Radio listeners are not bothered by the difference between trained and untrained speakers, in

fact naturally speaking interview partners provide highly interesting and entertaining radio content. Speech recognition systems, however, are seriously challenged by such conditions. Separation of speech from non-speech audio is an important pre-processing step which insures that only speech is sent to the speech recognition module for processing. In fact, all the methods of structuring mentioned in the previous section may prove to be beneficial. The project plans to investigate whether optimization of the recognizer to deal with certain classes of speech yields additional tangible benefit. Separate processing of male/female speech and studio/non-studio/telephone speech may make it possible to boost recognition performance. Not only audio quality, but audio content influences the performance of speech recognition systems. When a system encounters a word that it has not explicitly been trained to identify, it will never produce anything except a recognition error. State of the art recognizers have vocabularies of several hundred thousand words, but new words, especially proper names, occur in radio broadcasts frequently, for example as current events launch new personalities and new localities into the international spot lights. In the project we will address the vocabulary limitations of conventional speech recognizers explicitly and use a speech recognition system that generates a transcript that consists not of words, but of syllables.

Automatic systems will not replace human archivists in the foreseeable future. However, the possibilities for computer-supported structuring, processing and search represent an important aid to the work of the archival department of large radio broadcasters. When automatic annotation methods have made spoken audio as accessible as speech audio, it becomes possible to envision huge multimedia archives, encompassing text, audio and even video.

References

[Pl04] Pleitgen, Fritz. 2004. Mediensammlungen als Kulturgut, In: *Mediensammlungen in Deutschland im internationalen Vergleich. Bestände und Zugänge, Bonn*: 20-30.

[Sp04] Spree, Ulrike. 2004. Trends in der Mediendokumentation. Herausforderungen für Medienunternehmen, In: *Info* 7(3):130-148.

[He01] Herla, Siegbert. 2001l Online-Archive - mit Metadaten zum Erfolg. In: *Rundfunktechnische Mitteilungen* 45(H. 1):8-15,

[AZ02] ARD/ZDF-AG *Archive, Content- und Produktionsmanagement: Archivierung im digitalen Produktionsprozess*, Ms. München 2002;

[Hä95] Häfner, Albrecht. 1995. Digitalisierung in den Schallarchiven der ARD am Beispiel des Südwestfunks Baden-Baden. In: *Info* 7(10)(H. 1):25-29.

[TG99] Tritschler, Alain and Gopinath, Ramesh. 1999. Improved speaker segmentation and segments clustering using the Bayesian Information Criterion. In *Proceedings Europseech*, vol. 2: 261-264.

[BK03] Biatov, Konstantin & Köhler, Joachim. 2003. An audio stream classification and optimal segmentation for multimedia applications. Proc. 11th ACM Multimedia: 211-214.