Words and Silences is the Digital Edition Journal of the International Oral History Association. It includes articles from a wide range of disciplines and is a means for the professional community to share projects and current trends of oral history from around the world.

http://ioha.org
Online ISSN 2222-4181

Words and Silences
September 2019
"Memory and Narration"

Audio Mining.
Advanced Speech Analytics for Oral History[1]

Almut Leh, Michael Gref and Joachim Köhler
Fernuniversität in Hagen
almut.leh@fernuni-hagen.de

1. Introduction: What speech technologies can do for research in the humanities

Audiovisual (multimodal) data play an increasingly important role in the humanities and social sciences. In fact, all humanities and social sciences occasionally come into contact with such data. The enormous progress made in the development of digital recording devices, which today allow AV recordings to be made with comparatively little financial and personnel expenditure, suggests that these data will become increasingly important in the short and medium term. Compared to textual data, the possibilities of evaluation, documentation, publication (generally accessible provision) and archiving of this audiovisual multimodal data using modern information technologies are still little researched and developed.[2] At present, the scientific dissemination and use of this data continues to be primarily based on transcripts, i. e. converted to the written language medium. This has two crucial disadvantages: Firstly, the production of transcripts is time-consuming and labor-intensive. Secondly, transcripts lose the specific characteristics of speech melody, vocal qualities, gestures, facial expressions, etc. that are specific to AV data. However, the latter are often decisive for an appropriate interpretation of what has been said. This is where audio mining steps in.

---

[1] The paper presents the results of an ongoing research project that aims to further develop language technologies for interview-based research methods. The exemplary application scenarios are the linguistic field interview and the oral history interview. The project "KA³" is funded by the Federal Ministry of Education and Research. The project group consists of Nikolaus P. Himmelmann (Institute for Linguistics, University of Cologne) for coordination, Joachim Köhler and Michael Gref (Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS) and Almut Leh (FernUniversität in Hagen). For in-depth information, see Leh, Almut, Joachim Köhler and Nikolaus P. Himmelmann: Speech analytics in research based on qualitative interviews. Experiences from KA3. *VIEW* Special Issue *Audiovisual Data in Digital Humanities* (2019).

[2] Compare, for example, the current textbook: Fotis Jannidis, Hubertus Kohle, Malte Rehbein (eds.): Digital Humanities. Eine Einführung, Stuttgart 2017.

By using audio mining speech technologies, the handling of AV data should become more efficient, especially in two ways: (a) reducing resources typically employed for the preparatory analysis of interviews and (b) minimizing the reductions inherent in transcription.

Furthermore, we expect that audio mining does not only facilitate the research process but allows new approaches in quantity and quality. Covering more data comparative studies and quantitative analysis become reasonable. The speech technologies also represent verbal and non-verbal aspects of communication more thoroughly and systematically thus opening new dimensions for qualitative research.

2. The need for speech technology in the oral history domain

In Germany, as in many other Western European countries, since the 1970s the use of oral history as a methodology has resulted in a substantial contribution of oral history interviews and projects added to the historical record. The material legacy of all these projects results in thousands of audio recordings, meanwhile also an immense amount of video recordings on different types of data carriers, which tell the technical history of past decades. It is characteristic for oral history interviews that they can be used for different questions and interpretations. Due to the narrative form and the biographical dimension, they contain such an abundance of information that is not exhausted by a single evaluation. For this reason, oral history interviews are often the subject of reanalyses. All the more so when the interviewed contemporary witnesses have already died, so that the interviews are the only access to their experiences. However, archiving oral history interviews as well as analyzing interviews conducted by others and in former times is very challenging. In particular, the retrieval of suitable interviews on certain research questions is a problem. The biographical data can be provided by metadata. However, as soon as the interest is directed to certain contents, a retrieval is very difficult, in the best case possible by means of a full text search.

Traditionally, the transcript plays a central role in the reanalysis of interviews. It is true that the analysis should not be limited to the transcript, but should also include the way of speaking, facial expressions and gestures. But it is also true that the transcript is an indispensable tool both for the analysis of oral history interviews and for archiving them. In other words: the lack of transcriptions will severely limit the use of interviews. In this context, it is a problem that not only a considerable part of the interviews in the archives is not transcribed,[3] but that today many more interviews are produced for which no transcription is planned. All these interviews are of limited use and will be forgotten in many cases. Against this background, it is obvious that oral history research now and in the future has an immense need for automatic speech recognition and other audio mining technologies.

3. The Fraunhofer IAIS Audio Mining system

While the everyday presence of Siri, Alexa and other similar, "virtual assistant" automated technologies gives the impression that powerful speech recognition has long since become state of the art, experience in research practice is clearly different. Interviews in spontaneous speech combined with non-professional recording techniques still present language technologies with a great challenge. While speech recognition works very well in many areas of everyday life, the results with oral history interviews are very weak. For three years we, the archives "Deutsches Gedächtnis" at the Fernuniversität in Hagen, the most comprehensive oral history archive in Germany, have been working with the Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS to improve the performance of speech recognition in oral history interviews. The Fraunhofer Institute (IAIS) has extensive experience in the field of speech recognition and has developed it into an application for

[3] In the archive "Deutsches Gedächtnis" of the FernUniversität in Hagen, the most extensive interview archive in Germany, about 30 percent are not transcribed. See: Almut Leh: Das Archiv „Deutsches Gedächtnis" und seine Bestände: Herkunft – Erschließung – Nutzung, in: Denis Huschka, Hubert Knoblauch, Claudia Oellers und Heike Solga (Hg.): Forschungsinfrastrukturen für die qualitative Sozialforschung. Berlin: SciVero, 2013, pp. 109-116.

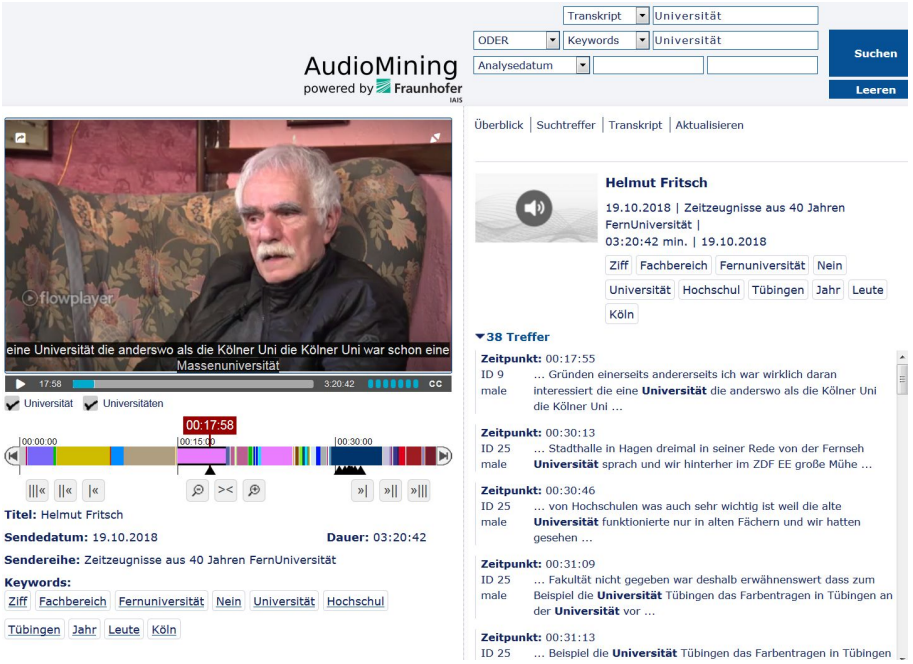
archives in the broadcasting domain. The work with Oral History Interviews is a new challenge for IAIS, which contributes to the further improvement of speech recognition as a whole. The aim of the KA³-project is to improve speech recognition to such an extent that it can be used in practice and represent an added value. In the following I would like to present the Audio Mining tools developed by the Fraunhofer Institute and their current performance.

In general, the Fraunhofer Audio Mining System is designed to analyze the audio-signal of audiovisual media files automatically. The aim is to create a time-aligned transcription of the spoken word, as it is normally made by professional human transcribers. This does not only include automatic speech recognition but an entire workflow by which the files are made text-searchable and structured. On a higher system based level this system generates for each audio recording an additional XML metadata file which contains several kinds of time-code information.

1. Audio Segmentation
2. Concept Detection
3. Speaker Clustering
4. Automatic Speech Recognition
5. Keyword Extraction

In this workflow, there are several processing stages. First, the speech is segmented automatically into homogeneous speech segments. After segmentation, each segment is classified using a speech/non-speech detection. Segments that are assumed to contain speech are additionally classified using a gender detection based on the assignment of female and male voices. In the next processing step a speaker clustering is applied. Here all speech segments of the same speaker get the same ID number. The whole process of speech segmentation is summarized as speaker diarization process. The time-code of each segment and label is stored in the MPEG-7 metadata file. In the next processing step automatic speech recognition technology is applied. Significant progress in speech recognition is based on the

use of deep neural networks, which has decisively advanced the development of artificial intelligence in recent years.[4] Buzzwords are “deep learning” or “machine learning”. The acoustic modelling is trained on 1,000 hours of German broadcast recordings. The speech recognition system contains a *pronunciation lexicon* for about 500,000 words. The phonetic transcriptions for each word are generated automatically using a statistically trained grapheme-to-phoneme converter. The final, automatically generated transcripts provide a search and navigation functionality. For each recognized word the exact time code is generated and stored in the MPEG-7 metadata file. The slide shows the search and retrieval interface of the Fraunhofer Audio Mining system. Providing a Graphical User Interface (GUI) the system enables end-users to quickly navigate within interviews. The interface is shown in this slide:



Graphical Web User Interface of the Fraunhofer IAIS Audio Mining system

[4] Special information on the technical backgrounds can be found at Michael Gref, Joachim Köhler, Almut Leh: Improved Transcription and Indexing of Oral History Interviews for Digital Humanities Research, in: Nicoletta Calzolari, et al. ed.: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, May 2018, 3124-3131. http://www.lrec-conf.org/proceedings/lrec2018/summaries/137.html

An embedded media-player allows users to directly play segments of specific speakers. Different speakers are represented by a unique color for each speaker in the time-bar below the video. The user can search for relevant query items and directly acquire the snippets of the recognised phrases and the entry points (black triangles) to jump to the position where this item was spoken. Additionally, the application provides a list of relevant key words. The result of the speaker diarization analytics is displayed in the coloured segmentation sequence. The tool is capable to process extreme long audio recordings with the duration of about 4 hours.

4. Oral history interviews as a challenge

Applied to broadcast scenarios, the most advanced speech recognition approaches achieved word error rates of 8%, which allows a meaningful understanding of the text. This yields in high quality transcriptions which can be directly used for reading and understanding purposes. Applied to oral history data the error rates are much higher. For recordings of the oral history domain word error rates are much higher due to different speaking styles and difficult recording conditions. First test runs at the beginning of our project resulted in word error rates of up to 60 percent. But four years of research were able to reduce the word error rate to 25 percent (May 2019). In view of the very different speech and recording quality, it is not surprising that the results vary widely. For some interviews, error rates of 15 percent and less are achieved − for other interviews, the error rate is still high.

One would expect the interview partners' pronunciation and dialect to be the biggest problem. In fact, the recording quality causes the bigger problems. Audio signal quality is one of the main challenges when adapting the ASR system to oral history interviews. The broadcast audio signals used for training the acoustic model were recorded and post-processed using highly professional equipment. The recordings usually have no background noise,

barely perceptible reverberation and the levels are well adjusted. Of those oral history interviews tested, a majority were conducted "in the field," for example, an interviewee's living room, or kitchen, or other spaces where external noise levels are less controllable. The use of table microphones contaminates the recordings with reverberation, which is not a problem for the human ear but interferes with speech recognition.

There are two main approaches how to face these challenges: Speech Enhancement and Multi-Condition Training. Speech enhancement aims to modify the signal itself and increase the audio quality by reducing noise and compensating corruptions. Multi-Condition Training on the other hand aims to make the acoustic model robust against corruption by showing the model a wide range of features with corrupted audio signals during training. This approach aims to let the model generalize to unseen noise-types and distortions by relying only on robust acoustic features. Fraunhofer utilize the latter approach and try to train such a robust acoustic model. But of course, other big challenges posed by Oral History interviews are colloquial language used in spontaneous speech, hesitations, age- and health-related changes in the way of speaking and domain specific words used in the interviews that usually not occur in everyday speech (e.g. *Kriegerwitwenso¨hne*, German for *sons of a war widow*). These challenges must be addressed by adapting the language model and will be part of future work.

5. Conclusion and Perspectives

Although the results of speech recognition are currently not sufficient for the automatic production of transcripts even with good source material, speech recognition can already be used at a profit. On the one hand, incorrect transcripts generated by speech recognition can be corrected manually, saving time and resources compared to full transcription. On the other hand, these tools improve the archival practice. Direct access to the

audio signal and pre-structuring of the content enable the retrieval of large amounts of data, so that specific data can be made available for research queries from the large pool. In this way, speech recognition allows to include non-transcribed interviews in the search query, so that they can be used for secondary analysis, while being lost without speech recognition for further research. By displaying the search results in context, the relevance of a hit can be immediately assessed, which increases the efficiency of the search. Automatically generated keywords are also a step forward in archival practice when it comes to finding relevant interviews or interview passages irrespective of the transcription status. In view of the current error rates in speech recognition, the hits for term searches and keywords are incomplete, but they are already a significant added value because they include the large number of non-transcribed interviews in the search.

Also in the analysis, direct access to the audio signal is a considerable gain for oral history and biography research, i. e. the implementation of the previously seldom fulfilled claim to treat the audio recording as a primary source and thus to include the way of speaking (voice melody, voice quality, pauses etc.) in the interpretation. In fact, interview analysis is often based solely on knowledge of the transcript, which can be the cause of misinterpretations, especially in the evaluation of interviews conducted by others. On the other hand, access to the audio signal enables synchronous representation of audio and transcription and thus the reception of the primary source as a prerequisite for an interpretation of the interview that is appropriate to the intention of the statement. In addition, the tools open up new dimensions for research questions. While the case numbers for conventional analysis methods are usually around 30 interviews, much larger case numbers can be processed with technical support and thus evaluated quantitatively under a variety of comparative questions. At the same time, the tools also offer new dimensions for qualitative

analysis, in that linguistic and non-linguistic aspects of communication can be recorded, documented and thus made accessible to research questions in a differentiated way. Future pilot studies will show what concrete possibilities these quantitative and qualitative approaches offer.

The use of speech technologies in interview-based research has become more than just a promise. Automatic speech recognition can already be used profitably in analysis and archiving. Above all, however, it becomes clear what potential the language technologies offer. Ongoing research activities investigates how to model the challenging recording conditions of the oral history interviews. Therefore, multi-condition training on robustness of the acoustic models are applied. The conclusion I would like to draw: Much has been achieved and much more needs to be done.