volume 7 issue 14/2018
VIEW
JOURNAL OF EUROPEAN TELEVISION HISTORY & CULTURE

SPEECH ANALYTICS IN RESEARCH BASED ON QUALITATIVE INTERVIEWS

EXPERIENCES FROM KA³

Almut Leh
Institut für Geschichte und Biographie
FernUniversität in Hagen
Feithstr. 152
D-58097 Hagen
Germany
almut.leh@fernuni-hagen.de

Joachim Köhler
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
Schloss Birlinghoven
D-53757 Sankt Augustin
Germany
joachim.koehler@iais.fraunhofer.de

Michael Gref
Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS
Schloss Birlinghoven
D-53757 Sankt Augustin
Germany
michael.gref@iais.fraunhofer.de

Nikolaus P. Himmelmann
Institut für Linguistik
Universität zu Köln
D-50923 Köln
Germany
sprachwissenschaft@uni-koeln.de

Abstract: The paper presents aims and results of the project KA³ (***K**ölner Zentrum **A**nalyse und **A**rchivierung von **audio-visual**-Daten*), in which advanced speech technologies are developed and provided to enhance the process of indexing and analysing speech recordings from the oral history domain and the language sciences. Close cooperation between speech technology scientists and digital humanities researchers is an important aspect of the project making sure that the development of the technologies answers the needs of research based on qualitative audio-visual interviews. For practical research reasons, the project focuses on the audio aspect, although visual aspects are of course equally important for the analysis of audio-visual data. The Cologne Centre for Analysis and Archiving of audio-visual data will provide the technologies as a central service.

Keywords: automatic speech recognition, speech retrieval, transcription, oral history, conversation analysis

1 Introduction: What Speech Technologies Can Do for Research in the Humanities

Audio-visual (multimodal) data play an increasingly important role in the humanities and social sciences. Compared to textual data, the possibilities of evaluation, documentation, publication (making it generally available) and archiving of this data using modern information technologies are still little researched and developed. At present, the scientific dissemination and use of this data continues to be primarily based on transcripts, i.e. conversions to the written language medium. Essential characteristics of this data type (voice melody, vocal qualities, gestures, facial expressions, etc.) are lost.

A variety of humanities and social sciences use audio-visual data, which includes interviews with contemporary witnesses or experts, recordings of oral traditions (including recordings of plays and the like), soundscapes, radio and television broadcasts. This often involves interaction data in which several speakers interact verbally (everyday conversations, mediation, debates, classroom interactions, therapy sessions, etc.). The disciplines that regularly work with such data include educational sciences, sociology, ethnology, history (oral history/biography research), all regional sciences (African studies, Oriental studies, Chinese studies, etc.), ethnomusicology, psychology, medicine, media and linguistics. In fact, all humanities and social sciences occasionally encounter such data. The enormous progress made in the development of digital recording devices, which today allow audio-visual recordings to be made with comparatively little financial and personnel expenditure, suggests that this data will become increasingly important in the short and medium term.

These disciplines typically focus on content-related questions, which are approached qualitatively and/or quantitatively. This means that when dealing with this data, it is important to know which topics are dealt with in which order, whether event A follows event B or vice versa, in which segment one can find statements on a search term, which speakers occur when, where conflicts arise and are resolved and so on. Up to now, the further analysis has mainly been done on the basis of transcripts, but this has two crucial disadvantages: Firstly, the production of transcripts is time-consuming and labour-intensive. Secondly, transcripts lose the specific characteristics of speech melody, vocal qualities, gestures, facial expressions, etc. that are specific to audio-visual data. However, the latter are often decisive for an appropriate interpretation of what has been said.

The aim of the KA³-project is therefore to make the handling of audio-visual data more efficient by using speech technologies, in two ways: by reducing the use of resources for the preparatory analysis steps and by carrying out analyses with direct access to the signal, i.e. without the reductions inherent in transcription.

The project focuses on two use cases – oral history and linguistic research on communication – and explores the potential of audio mining technology in these fields of research. Currently, the preparatory steps for analysing qualitative interviews or communicative interactions is mostly performed manually by labelling and annotating speech recordings. The huge effort in terms of time and human resources required thoroughly limits the possibilities of properly including multimedia data in the research process. This is where audio mining steps in. We expect audio mining to facilitate access to audio-visual data through automatic transcription, time alignment with pre-existing transcripts and indexing, and thus their integration into research. Furthermore, we suggest that audio mining allows new approaches in analysing audio-visual data. In analogy to distant and close reading in literature science, audio mining supports the quantitative analysis of large interview collections and a detailed qualitative analysis of individual interviews or sequences including verbal and non-verbal aspects of communication. Sections 2 and 3 provide further detail.

In the digital humanities, the field of audio-visual data has so far played a subordinate role. While the everyday presence of Siri, Alexa, etc., gives the impression that powerful speech recognition has long since become readily available, experience in research practice tells a different story. Interviews featuring spontaneous speech and non-professional recording techniques still present language technologies with a great challenge. In the KA³ project, this application case is being systematically analysed for the first time with the aim of achieving improvements.

For practical research reasons, the project focuses on the audio aspect, although visual aspects are of course equally important for the analysis of audio-visual data.

The following two sections (2 and 3) present the two application scenarios with their specific requirements for speech technologies. Subsequently (4), we will demonstrate the Fraunhofer IAIS audio mining system with its functions and current performance capabilities. The paper concludes with assessments of the benefits of the presented technologies and further perspectives for interview-based research.

2 The Need for Speech Technology in the Oral History Domain

Since the early 1980s, biographical interview-based research emerged in almost all areas of the humanities. For all the differences among the various academic disciplines, in terms of their research questions, terminology, methods, and research strategies, all these fields emphasize subjectivity or/and the relationship of the individual to society. This attitude has developed against a background of increasing doubt concerning the explanatory claims of grand historical narratives or large-scale theoretical frameworks. As a result, biographical research has increasingly claimed its own independent research approach and has asserted the efficacy of individual action in relation to the determinative power of structural conditions.

In the historical sciences, research based on interviews with contemporary witnesses has become known as Oral History.[1] To this day, oral history makes an important contribution to contemporary history by providing the perspective of groups that, without oral questioning, would be overlooked by historiography because they do not produce written sources. These can be marginalized social groups, but also victims of political violence, witnesses of historical events or people sharing personal experiences. The autobiographical, narrated memories emphasize the significance of subjectivity and experience for history.

Particularly in Germany, this research has been focused above all on the period of National Socialism and the Second World War. In the meantime, it has also come to include many other topics and historical periods. The past 40 years have seen a multitude of witnesses to a wide range of historical events interviewed by researchers as well as by non-professionals.[2] Essential prerequisites for conducting eyewitness interviews were of a technical nature: the development of easy-to-handle and readily affordable recording technology in the form of a cassette recorder. The material legacy of all these projects results in thousands of audio cassettes, as well as an immense amount of video recordings on different data carriers, which tell the technological history of past decades.

Most of the interviews conducted are characterized by the fact that rather than structuring the interview around questions, the interviewer encourages the interviewee to freely narrate his or her life story. The outcome very often lasts three or four hours or more and is known in the field as a narrative life-story interview.[3] Such an interview represents a highly individual testimony in which the interviewee has presented large parts of his life story and his worldview in a way that is often unguarded and sometimes contradictory. The final result is one in which the interviewer has played a part not only as an initiator but also as an interested and sympathetic listener. Every interview is unique and irreproducible

1 Robert Perks and Alistair Thomson, eds, *The Oral History Reader*, Routledge, 2nd edition, 2006. Donald A. Ritchie, *The Oxford Handbook of Oral History*, Oxford University Press, 2011. Donald A. Ritchie, *Doing Oral History*, Oxford University Press, 3rd edition, 2015. Paul Thompson and Joanna Bornat, eds, *The voice of the past. Oral History*, Oxford University Press, 4th edition, 2017.

2 Almut Leh and Doris Tausendfreund, 'Archiving Audio and Video Interviews', Carlos Nunes Silva, ed, *Online Research Methods in Urban and Planning Studies: Design and Outcomes*, Hershey, 2011, 353–367.

3 Fritz Schütze, 'Zur Hervorlockung und Analyse von Erzählungen thematisch relevanter Geschichten im Rahmen soziologischer Feldforschung – dargestellt an einem Projekt zur Erforschung von kommunalen Machtstrukturen', Arbeitsgruppe Bielefelder Soziologen, ed, *Kommunikative Sozialforschung*, München, 1976, 159–260.

since it is tied to the life of the interviewee and the historical moment of the interview itself. At the same time, it is obvious that due to the open character of the narration and the biographical content, such an interview is open to more than one interpretation and a valuable source for later re-use, especially as more and more witnesses of the time die.

Against this background, the archiving and publication (securing general accessibility) of such interviews for future research is very important and challenging at the same time. Many thousands of such interviews are preserved and are available in archives, museums, historical sites and documentation centres where they are archived alongside other historical sources. Besides, there are also special oral history archives such as the archive *Deutsches Gedächtnis* (*German Memory*), founded in 1994 at the FernUniversität in Hagen, where around 3,000 interviews from more than one hundred research projects conducted over the past four decades are available.[4] A similar archive with a regional focus is the *Werkstatt der Erinnerung* at the Forschungsstelle für Zeitgeschichte in Hamburg with about 2,500 interviews.[5] A central contemporary witness portal with more than 10,000 interviews run by the Haus der Geschichte der Bundesrepublik in Bonn went online in 2017[6].

The fact that interviews with contemporary witnesses are important historical sources, especially but not only on the topics of National Socialism and Second World War, is evidenced by the steadily growing request from contemporary historians. Furthermore, it is hard to imagine the presentation of historical information in exhibitions, documentations and films without the use of witness accounts to the relevant events.

The growing interest in oral history interviews is also reflected in another trend. Parallel to the use of established oral history archives, a large number of interview projects continue to be carried out. Whereas in earlier oral history projects interview survey and analysis were closely related, these projects focus on the producing documentation, emphasizing the aspect of preserving the documented narrations from oblivion in the face of the death of contemporary witnesses. Any interpretation is postponed for later research.

Typical for these documentation projects is the presentation of the interviews in an internet portal. The Shoah Foundation Institute's *Visual History Archive*, initiated by Steven Spielberg and the world's largest archive of videographed oral history interviews with 52,000 interviews with survivors and Holocaust witnesses, was perhaps the most important source of inspiration.[7] The successors include the *Mauthausen Survivors Documentation Project* with 850 interviews[8] and the archive *Forced Labor 1939–1945: Memories and History* which offers 600 interviews with people who were forced to work for Nazi Germany during World War II in 25 languages with translations into German and English and excellently curated[9].[10] On behalf of many other ongoing documentation projects, the project *Menschen im Bergbau* (*People in Mining*) is named, conducted by the Stiftung Geschichte des Ruhrgebiets in cooperation with the German Bergbaumuseum Bochum, whose one hundred interviews are to be made available in an online portal.[11]

4 **https://www.fernuni-hagen.de/geschichteundbiographie/deutschesgedaechtnis/**

5 **http://www.werkstatt-der-erinnerung.de/index.php**

6 **https://www.zeitzeugen-portal.de/**

7 **https://sfi.usc.edu/vha**. See also: Leh & Tausendfreund 2011.

8 **https://msrp.univie.ac.at/project-information/msdp/**

9 **http://www.zwangsarbeit-archiv.de/en/index.html**

10 Alexander von Plato, Almut Leh and Christoph Thonfeld, eds, *Hitler's Slaves. Life Stories of Forced Labourers in Nazi-occupied Europe*, Berghahn Books, 2010.

11 *Digitaler Gedächtnisspeicher: Menschen im Bergbau (Digital memory storage. People in mining)*, Bochum, **http://isb.rub.de/sbr/drittmittelprojekte/gedaechtnisspeicher.html.de.** Other current projects are: *Sprechen trotz allem. Das Videoarchiv der Stiftung Denkmal für die ermordeten Juden Europas (Speaking in spite of it all. The video archive of the Memorial to the Murdered Jews of Europe Foundation's video archive)*, Berlin, **www.sprechentrotzallem.de**; *Archiv der anderen Erinnerungen. Zeitzeug_innen-Interviewprojekt der Bundesstiftung Magnus Hirschfeld (Archive of other memories. Contemporary witness interview project of the Magnus Hirschfeld Federal Foundation)*, Berlin, **http://mh-stiftung.de/en/zeitzeug_innen-interview-projekt-der-bundestiftung-magnus-hirschfeld/**; *Individuelle Erinnerung und gewerkschaftliche Identität (Individual memory and trade union identity)*, Bonn/Düsseldorf, **http://www.zeitzeugen.fes.de/**; *Museum für Hamburgische Geschichtchen (Museum of Hamburg Stories)*, Hamburg, **http://toepfer-stiftung.de/museum-fuer-hamburgische-geschichtchen/**. See: Linde Apel, 'Oral History reloaded. Zur Zweitauswertung von mündlichen Quellen', *Westfälische Forschungen. Zeitschrift des LWL-Instituts für westfälische Regionalgeschichte*, Bernd Walter and Thomas Küster, eds, 65/2015, p. 245.

The problem is that these documentary projects not only dispense with research, but often also with transcribing the interviews. Transcribing is undoubtedly very time-consuming and therefore a considerable cost factor in project calculation. However, it is clear that the lack of transcriptions will severely limit the use of interviews.

As already explained, the transcript is in many respects an inappropriate reduction of the interview to the text alone. Nevertheless, it is an indispensable tool both for the archiving of oral history interviews and, in some ways, for analysing them. When analysing, it is an instrument in measuring critical distance to the source material efficiently handling the amount of information. For publications in conventional print media, the transcript is indispensable for quoting references. However, the fact that transcripts are indispensable for the purposes mentioned and easier to handle than the audio recordings, should not lead to an interpretation based on the text version alone. In many cases, the interviewee's manner of speech is an important key to an appropriate interpretation. Especially when analysing interviews conducted by others, in the case of secondary analyses, the audio or video recording must be the basis of interpretation. Ideally, the archive must provide the researcher with video or at least audio and text files simultaneously, i.e. with subtitles. This is exactly what advanced speech technologies can do.

In archival practice, the transcript of the interview is currently the most important instrument for retrieving relevant interviews for secondary analyses. Second in importance is the indexing by means of keywords. In fact, retrieving relevant interviews with specific content is the greatest challenge for oral history archives. Given the fact that not only has a considerable part of the interviews in the archives not yet been transcribed (up to 50 percent of the collections), but that many more interviews continue to be produced with no plan for transcription, an immense need for automatic speech recognition and other audio mining technologies is obvious. Interviews without transcripts, whether manually or automatically generated, are only of limited use if not completely unavailable for secondary analysis.

3 Experiences from the Point of Linguistic Research on Communication

Everyday conversation is at the heart of human sociality.[12] In conversational exchanges, interlocutors seamlessly switch speaker and hearer roles. First speaker A speaks and speaker B listens, then vice versa. There are no explicit rules for taking turns, but in general, there are no major interruptions when speaker B starts her turn after having listened to speaker A. No major pauses arise at turn changes. It is also rare that two speakers overlap for more than a few milliseconds at such transitions. In fact, looking at everyday conversations across different cultures, Stivers et al.[13] find that the most common gap at turn transitions is zero, i.e. no gap, no overlap, speaker B starts exactly at the point where speaker A stops. How is this very fine-grained coordination possible, how can speaker B predict exactly when speaker A stops? What steps are necessary to launch one's own opening utterances at exactly the right point in time? Moreover, of course, conversations are not limited to two parties, but may involve three, four, five or even more interlocutors. How does coordination work in such larger groups?

Turn taking is one of a number of riddles that everyday conversation posits for the cognitive sciences in general and linguistics in particular. However, everyday conversation actually involves a host of other tasks where speaker and hearer (groups) have to cooperate to communicate efficiently. Thus, for example, they have to agree on how to refer to a person or object they are both familiar with ("Mr Jones", "my neighbour", "the man at the corner", etc.). They have to agree on how to keep track of persons and objects already introduced into the discourse. They have to make decisions on when it is appropriate to ask a question and what the appropriate form for a question is at a given point in the discourse.

12 Stephen C. Levinson, 'On the human 'interaction machine'', Nicholas J. Enfield and Stephen C. Levinson, eds, *Roots of Human Sociality. Culture, Cognition, Interaction*, Berg, 2006, 39–69. Michael Tomasello, *Origins of human communication*, MIT Press, 2008.

13 Tanja Stivers, N. J. Enfield; Penelope Brown, Christina Engler, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoonf and Stephen C. Levinson, 'Universals and cultural variation in turn-taking in conversation', *PNAS,* 106, 2009, 10587–10592.

An additional line of inquiry compares communicative routines across different cultures. Which aspects of everyday communication are universal, possibly reflecting an interaction engine common to all humans?[14] In addition, where does conversational behaviour vary in culturally specific ways? Answers to the latter question are at the core of studies on intercultural communication. Given that basic aspects such as the system of seamless turn taking appear to be universal, interlocutors from different cultural backgrounds tend to overlook smallish differences, which may easily lead to misunderstandings and communication failures.

To date, investigations of these issues are largely based on smallish excerpts of longer conversations, often not more than 2 to 5 seconds in length. These excerpts are usually transcribed on a very detailed level and the analysis is based on the transcript rather than the actual recording. Transcription is enormously time-consuming. Informed guesses on transcription time for conversational exchanges range from 1 to 10 hours of transcription time per minute of recording, depending in part on the number of interlocutors and the kinds of events noted in the transcripts (e.g. laughs, lengthening of sounds, in the case of videotaped conversations also gaze, body posture, gesture and so on).

The potential of audio mining technology to support and also to transform the research process in this domain should be obvious. If (partially) automatic transcription of conversations were to become possible, the venue for serious quantitative studies of phenomena of specific interest would be wide open. The ability to search for specific words and other types of relevant events such as laughs or feedback items such as *uhm* would allow users/researchers to quickly assemble databases for specific research questions linked to these items across a larger corpus of recorded conversational exchanges. Automatic keyword generation would allow for providing fast and precise access to large collections of audio-visual recordings.

However, as it stands, there is still a long way to go to make audio-visual data as accessible as is standard in the domain of written texts. In fact, it is very likely that if (semi-)automatic transcription will become available for audio recordings at all, it will be limited to well-resourced major national languages where speech recognizers can be trained with hundreds and thousands of hours of training data. As the linguistic component of the KA³ project has a special focus on smaller languages spoken in remote locations, the main focus of exploring possible applications of speech technologies in this domain is on aspects of conversational speech which can be safely assumed to be universal. Two topics in particular are of major concern: speaker diarization and overlap detection.

The goal of speaker diarization is to determine segments of a recording produced by the same speaker, thus allowing estimates on how much talk was produced by speaker A, how much by speaker B, etc. Successful diarization also helps transcription, as it would allow pre-segmenting a recording along the contributions of individual speakers. The challenge for diarization of conversational recordings lies in the fact that speaker turns are often short and fast, each speaker only contributing a short sentence or even only a word or two before the next speaker begins his or her turn. At current standards, speaker diarization of dialogues works reasonably well when individual turns are 5 seconds or longer. Problems arise when more than two speakers are involved and when turn changes are faster. The former problem can be partially addressed by providing the audio mining tool with the information of how many speakers to expect in a given recording, i.e. fixing in advance the number of speakers to be recognized. This information is easily available for many audio-visual recordings from the accompanying metadata.

Recognition of overlapping speech is of interest for a number of issues. Minimal overlaps of a few hundred milliseconds, for example, are typical for feedback tokens such as *uhm, okay*, etc. which in turn can be used to identify conversational hotspots. As found by Kawahara et al.[15], for example, the use of such tokens increases at points where particularly relevant and possibly contested information is exchanged. Longer overlaps of a second or more typically also point to interactional problems where several speakers compete for the floor. Extended overlap of several seconds, on the other hand, often occurs at points where speakers co-construct their utterances, with one speaker

14 Levinson, 2006.

15 Tatsuya Kawahara, Kouhei Sumi, Zhi-Qiang Chang and Katsuya Takanashi, 'Detection of Hot Spots in Poster Conversations based on Reactive Tokens of Audience', *Proceeding of INTERSPEECH 2010,* 3042–3045.

speaking along with another one, and then finishing the ongoing turn on their own. Finally, overlap is one of the major stumbling stones for proper speaker diarization. Hence, marking points of overlap as additional input for speaker diarization has the potential for leading to major improvements in this regard. Within the KA³ project, we are exploring the potential of *Deep Convolutional Neural Networks* for overlap detection. Here, the availability of good and sufficient training materials has proven the major obstacle, as this requires very precise and detailed transcripts where overlaps are clearly marked as such.

4 The Fraunhofer IAIS Audio Mining System

The Fraunhofer IAIS audio mining system[16] bundles a number of tools for analysing audio-visual data using advanced speech technologies. These tools enhance the process of transcribing and indexing audio-visual recordings semi-automatically and provide additional speech-related analysis features. On a higher system-based level, this system generates an additional XML metadata file, which contains several kinds of time-code information, for each audio recording. The metadata scheme used is the MPEG-7 standard. The audio recording represents the input source of the processing containing a sequence of speech samples. The processing itself contains several processing stages. The first step is the automatic segmentation into homogeneous speech segments. For this, the Bayesian Information Criterion (BIC) is applied on full covariance Gaussian models of the mel frequency cepstral coefficients (MFCC).[17] After segmentation, each segment is classified using a speech/non-speech detection. Segments that are assumed to contain speech are additionally classified using provisional gender detection and then passed on to the following processing steps. The two detection algorithms are Gaussian Mixture Model-Universal Background Model (GMM-UBM) approaches trained for the respective classification task. In the next processing step, a speaker clustering is applied. Here, all speech segments of the same speaker receive the same ID number. Recently, we adapted iVectors[18] for this task and achieved increased performance compared to our previous BIC-based algorithm. 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. The speech segments are transcribed with a speech recognition system based on Kaldi technology.[19] This open-source framework contains the most advanced speech recognition approaches. For the latest acoustic modelling approaches, the LF-MMI trained models[20] with both LSTM and TDNN layers in a single network are used. The training is performed with 1,000 hours of a German broadcast corpus, namely the GerTV1000h corpus.[21] The speech recognition system contains a pronunciation lexicon for about 500,000 words. The phonetic transcriptions for each word are automatically generated using a statistically trained grapheme-to-phoneme converter. For speech recordings from the broadcast domain, the speech recognition systems achieve word error rates of 8 percent. This yields high quality transcriptions, which can be directly used for reading and understanding purposes.

16 Christoph Andreas Schmidt, Michael Stadtschnitzer and Joachim Köhler, 'The Fraunhofer IAIS audio mining system: Current state and future directions', *Speech Communication 12,* ITG Symposium, Paderborn, Germany, 2016, 115–119.

17 Alain Tritschler and Ramesh Gopinath, 'Improved speaker segmentation and segments clustering using the Bayesian information criterion'. *EUROSPEECH'99*, Budapest, Hungary, September 1999.

18 Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel and Pierre Ouellet, 'Front-end factor analysis for speaker verification', *IEEE Transactions on Audio, Speech, and Language Processing*, 19, 4, May 2011, 788–798.

19 Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yammin Qian, Peter Schwarz, Jan Silovsky, Georg Stemmer and Karel Vesely, 'The kaldi speech recognition toolkit'. *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB, December 2011.

20 Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang and Sanjeev Khudanpur, 'Purely sequence-trained neural networks for asr based on lattice-free mmi', *Interspeech 2016*, 2751–2755.

21 Michael Stadtschnitzer, Jochen Schwenninger, Daniel Stein and Joachim Köhler, 'Exploiting the large-scale German broadcast corpus to boost the Fraunhofer IAIS speech recognition system', in Nicoletta Calzolari et al., eds, *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA), May 2014.

Oral history recordings present a much greater challenge for automatic speech recognition systems than recordings from the broadcast domain due to different speaking styles, out-of-vocabulary words and much more difficult recording conditions. Dialects are also a major challenge for speech recognition systems as studied in Stadtschnitzer[22]. Usually, off-the-shelf speech recognition systems perform poorly when transcribing oral history interviews. To measure and evaluate the performance of automatic speech recognition systems on German oral history interviews, an in-house data set was recently proposed.[23] This evaluation set is a subset of the archive "Deutsches Gedächtnis" representing the wide spectrum of interviews in terms of recording technique, interview methodology and pronunciation. The test sets include early interviews from the 1980s as well as recent interviews and present interview methods from various academic disciplines. In terms of gender and age, the test data represent the archive collections. The selection should reflect the key aspects of the entire archives. The set consists of overall 3.5 hours of audio from 35 different speakers and contains 27,708 transcribed spoken words (with 4,582 different words).

The first version of the audio mining speech recognition system evaluated on the oral history data set only achieved a word error rate of 55 percent. The use of the aforementioned, more advanced LF-MMI models drastically improved overall recognition performance of audio mining and enabled the system to achieve a word error rate of 34.2 percent on this task. However, even these advanced models have difficulty transcribing interviews recorded under difficult acoustic conditions such as noise or reverberation. To address these issues, multi-condition training for oral history interviews was studied in Gref, Schmidt, Köhler[24] using data augmentation to artificially degrade the audio signal quality of the 1,000 hour broadcast training data. This approach aims at adapting the training data to the acoustic conditions of oral history interviews. Overall, the proposed approach achieved an average word error rate of 29.5 percent, further improving performance significantly.

The error rates depend heavily on the technical quality and speaking style of the oral history interview. For some interviews, error rates of 15 percent and less are achieved – for other interviews, the error rate is still high. There are several error sources:

- Out-of-vocabulary: If the topic of the interview contains words and terms which are not in the vocabulary of the speech recognizer, the transcription quality becomes quite low. To overcome this problem, the unknown words have to be added to the audio mining system.
- Background noise, recording quality: Especially the poor voice recording quality of older interviews leads to higher error rates. One special problem is that the speaker diarization and speech detection algorithms are quite sensitive to such recording conditions.
- Speaking style: If the speaking style is completely different from the training set, the error rate increases. For example, some female speakers tend to whisper and some speakers have a very strong accent or dialect.

As learned from other domains, like broadcast news, we made the experience that a transcription rate of more than 85 percent leads to a readable and comprehensive document. In addition, the effort and time for manual correction is reduced significantly. Hence, the interviews with less than 15 percent word error rate can be used for reading purposes without any manual post-correction.

Although we have seen great advances in automatic speech recognition for oral history interviews in recent years, the transcription quality has not yet reached the level achieved on broadcast recordings. Thus, speech recognition on such challenging recordings remains the subject of ongoing research.

22 Michael Stadtschnitzer: 'Robust Speech Recognition for German and Dialectal Broadcast Programmes', PhD Thesis, Mathematisch-Naturwissenschaftliche Fakultät, Rheinische Friedrich-Wilhelms-Universität Bonn, 2018, urn:nbn:de:hbz:5n-52369.

23 Michael Gref, Joachim Köhler and Almut Leh, 'Improved Transcription and Indexing of Oral History Interviews for Digital Humanities Research', Nicoletta Calzolari, et al., eds, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan, May 2018, 3124–3131.

24 Michael Gref, Christoph Schmidt and Joachim Köhler, 'Improving Robust Speech Recognition for German Oral History Interviews Using Multi-Condition Training', *Speech Communication*, 13. ITG Symposium, Oldenburg, Germany, October 2018, 256–260.

Nevertheless, the transcripts of oral history interviews generated by the current audio mining system provide a much better search and navigation functionality. The system automatically calculates the exact time code for each recognized word and stores the information in the MPEG-7 metadata file. All recognised words and corresponding time codes are indexed in the Solr search engine.

The following figure shows the search and navigation capabilities of an automatically generated MPEG-7 file, which is an output of the Fraunhofer IAIS Audio Mining System.

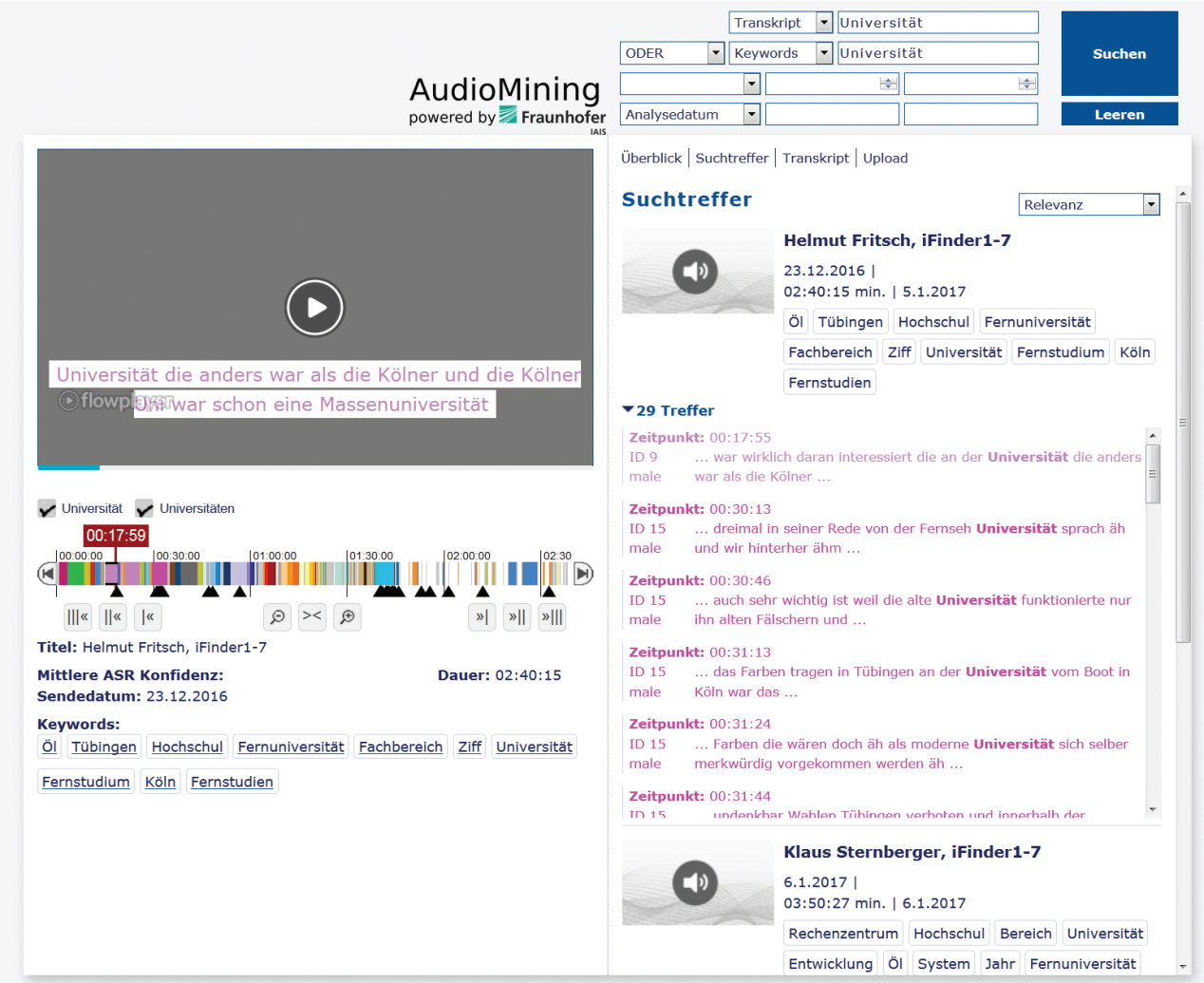

Figure 1. Search and Retrieval Interface of the Fraunhofer IAIS Audio Mining System for Oral History.

The user can search for relevant query items and directly access 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, which are calculated by a modified tf-idf algorithm to roughly describe the interview. The result of the speaker diarization analysis, which has not yet been satisfactory, is displayed in the coloured segmentation sequence. Only two colours for two speakers should appear here. In fact, the analysis “detects” significantly more speakers. The tool is capable of processing extreme long audio recordings with a duration of 4 hours and more. Ongoing research activities investigate the further improvement of the speaker diarization, such as the automatic detection of double-talks (overlapping speech) in audio recordings of natural conversations using deep convolutional neural networks.

The double-talk events are also considered as back-channel elements, which are important for a natural conversation. Currently, linguists have to find and annotate these double-talk segments manually using annotation tools like ELAN. In our research work, we apply and adjust Deep Neural Networks (DNN) to perform this annotation task automatically. Early results show that, for specific situations, the neural network detector can find such double-talk situations. However, the overall performance (recall, precision) is still poor because the overlapping speech segments are often very short (e.g. 50 ms). Due to the high complexity of this segmentation and classification task, we are convinced that only powerful deep learning approaches can solve this task. For machine learning, a huge amount of manually annotated training data is required. Currently, we are investigating how to use existing speech databases (e.g. Fisher speech database) to train the detector for overlapping speech segments.

5 Conclusion and Perspectives

Initially, the project confirmed that current speech recognition software has considerable difficulties with qualitative interviews recorded in the field using simple technology. Word error rates of more than 50 percent led to completely unusable transcripts. The fact that error analyses, new language models and the adaptation of algorithms reduced word error rates to below 30 percent on average and for good quality audio recording to under 10 percent is a great success of the project. What this improvement means for archival practice and for scientific analysis was evaluated in pilot studies parallel to technological development.

Due to the speech and recording quality of a large part of the interview collections, speech recognition often does not yet produce perfect transcripts. Nevertheless, the evaluation in the archive “Deutsches Gedächtnis” was able to show that speech recognition can already be used profitably with current performance. Retrieval with direct access to the audio signal and pre-structuring of the content enable access to 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 users/researchers to include non-transcribed interviews in the search query, so that they can be used for secondary analysis, whereas they would be lost to research without the use of speech recognition technology.

Testing in practice also showed that automatically generated transcripts save resources even at current word recognition rates. These transcripts have to be post-processed manually. Depending on the quality of the transcript, however, this saves up to 50 percent of time or more. Automatically generated keywords also proved to be a step forward in archival practice. It is true that key terms for the retrieval of qualitative data are basically of limited use. For finding relevant interviews or interview sequences, they are nevertheless a useful auxiliary tool – all the more so, if the interviews are not transcribed. By displaying the search results in context, the relevance of a hit can be immediately assessed, which increases the efficiency of the search. In view of 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.

An advantage of the audio mining system, which is already usable, is the simultaneous presentation of audio or video recording and transcript in the form of subtitles, because this brings the very source, the speech act, to the centre of attention. This feature addresses the previously seldom fulfilled claim of treating the audio recording as a primary source and thus to include the way of speaking (voice melody, voice quality, pauses etc.) in the interpretation. Especially when it comes to the interpretation of interviews conducted by others, the reduction to transcript form is a source of misinterpretations.

For a simultaneous display of audio and transcripts by means of subtitling, existing manually created transcripts can also be used. Therefore, the transcripts are time-coded using forced alignment. Until now, this was only possible for short audio sequences. As a result of the project, forced alignment can now also process audio recordings lasting several hours. The audio mining system is currently being used in the archive “Deutsches Gedächtnis” to automatically

add timecodes to all existing transcripts. After completion of this process, all interview collections in the audio mining system will be available for archive research and research purposes.

As far as analysis is concerned, the benefit for oral history and biographical research can currently only be estimated in initial approaches, because the performance of speech recognition does not yet permit comprehensive use for analysis and interpretation, but the evaluation has provided promising first insights. 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 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. In a two-year follow-up project that has just started, these initial findings are to be followed up in more extensive studies.

The challenges in the analysis of conversational data are even higher than in the case of interviews due to the fast changes that occur between speakers and overlap. The KA³ project explores the perhaps most difficult case scenario, i.e. conversational exchanges in small language communities where no big data are available that would allow for training state-of-the-art speech recognition software. At current estimates, such training minimally requires 100 hours of transcribed speech, which, among other things, would allow training a pronunciation lexicon for the variety under investigation.

Therefore, the KA³ project focuses, on the one hand, on aspects of conversation that occur language-independently, specifically speaker diarization and overlap. Speaker diarization becomes easier the longer the turns of one speaker last. The current model produces usable results at turn lengths of five seconds and longer, provided that the number of participating speakers is set in advance. Some sections of conversations have such extended turns, but more typically, turns are considerably shorter, often shorter than one second.

As for overlaps, the challenge pertains to the fact that these are usually very short, rarely extending for more than a few hundred milliseconds. Experiments with deep neural learning algorithms have produced some promising results, but the amount of training items required is still very high. Current experimentation investigates whether it is possible to combine training results from different languages and corpora on the hypothesis that the basic characteristic of overlap (the co-occurrence of two voices) is in principle language-independent.

Biographies

Almut Leh - PhD in history, since 1994 research assistant and director of the archive “Deutsches Gedächtnis” of the Institute for History and Biography at the University of Hagen, editor and publisher of “BIOS - Zeitschrift für Biographierforschung, Oral History und Lebensverlaufsanalysen”, member of the Council of the International Oral History Association. Research on the history of German mentalities in the 20th century, various stakeholder studies, publications on research-ethical and methodological questions of oral history and the archiving of biographical interviews.

Dr. Joachim Köhler received his diploma and Dr.-Ing. degree in Communication Engineering from the RWTH Aachen and Munich University of Technology in 1992 and 2000, respectively. In 1993 he joined the Realization Group of ICSI in Berkeley where he investigated robust speech processing algorithms. From 1994 until 1999 he worked in the speech group of the research and development centre of the SIEMENS AG in Munich. The topic of his Ph.D. thesis is multilingual speech recognition and acoustic phone modelling. Since June 1999 he has been with Fraunhofer IAIS in Sankt Augustin and head of the department NetMedia. His current research interests include pattern recognition, artificial intelligence, deep learning, speech recognition, spoken document, multimedia retrieval and multimedia information systems. He has published over 50 papers in the area of speech and multimedia research and served as reviewer in several speech-related conferences.

Michael Gref received his B.Eng. degree in Industrial Engineering from Düsseldorf University of Applied Sciences in 2014 and his M.Eng. in Electrical Engineering from Niederrhein University of Applied Sciences in 2017. In 2017, he joined the NetMedia department at Fraunhofer IAIS in Sankt Augustin as a research engineer. He is also a research assistant at the iPattern Institute of the Niederrhein University of Applied Sciences. He is currently working towards the PhD degree in the field of robust speech recognition and speech signal processing.

Nikolaus P. Himmelmann - Professor in General Linguistics specializing in discourse and conversation analysis as well as language description and documentation. Extensive fieldwork in the Philippines, Indonesia and East Timor. Major theoretical and practical work on language documentation, centring on lasting audio-visual records of communicative practices.

VIEW Journal of European Television History and Culture Vol. 7, 14, 2018
DOI: 10.18146/2213-0969.2018.jethc158
Publisher: Netherlands Institute for Sound and Vision in collaboration with Utrecht University, University of Luxembourg and Royal Holloway University of London.