# A Computational Look at Oral History Archives

FRANCISCA PESSANHA and ALMILA AKDAG SALAH, Utrecht University

Computational technologies have revolutionized the archival sciences field, prompting new approaches to process the extensive data in these collections. Automatic speech recognition and natural language processing create unique possibilities for analysis of oral history (OH) interviews, where otherwise the transcription and analysis of the full recording would be too time consuming. However, many oral historians note the loss of aural information when converting the speech into text, pointing out the relevance of subjective cues for a full understanding of the interviewee narrative. In this article, we explore various computational technologies for social signal processing and their potential application space in OH archives, as well as neighboring domains where qualitative studies is a frequently used method. We also highlight the latest developments in key technologies for multimedia archiving practices such as natural language processing and automatic speech recognition. We discuss the analysis of both visual (body language and facial expressions), and non-visual cues (paralinguistics, breathing, and heart rate), stating the specific challenges introduced by the characteristics of OH collections. We argue that applying social signal processing to OH archives will have a wider influence than solely OH practices, bringing benefits for various fields from humanities to computer sciences, as well as to archival sciences. Looking at human emotions and somatic reactions on extensive interview collections would give scholars from multiple fields the opportunity to focus on feelings, mood, culture, and subjective experiences expressed in these interviews on a larger scale.

CCS Concepts: • **Computing methodologies** → *Speech recognition*; *Natural language processing;*

Additional Key Words and Phrases: Oral history archives, audio/video collections, social signal processing, computational paralinguistics, automatic breathing detection, heart rate detection

**ACM Reference format:**
Francisca Pessanha and Almila Akdag Salah. 2021. A Computational Look at Oral History Archives. *J. Comput. Cult. Herit.* 15, 1, Article 6 (December 2021), 16 pages.
https://doi.org/10.1145/3477605



## 1 INTRODUCTION

The introduction of computational technologies had an undeniable impact on archival sciences. The launch of the database enforced a new approach in designing the relations of archival items [38]; the interface enabled novel ways of interaction with the archive [10]; the ability to search through the full text of books and print material prompted a completely new approach in thinking about the metadata; and last, the digitization of the archival materials minimized the need to access the original items.

We observe a similar impact of technology on **oral history (OH)** archives, collections of interviews recorded in audio/video format. How these collections were accessed have seen important changes: thanks to **automatic speech recognition (ASR)**, it was possible to generate transcriptions of these archives. Advancement in the

Authors' address: F. Pessanha and A. A. Salah, Information and Computational Sciences, Ultrecht University, Princetonplein 5, Utrecht, Netherlands, 3584 CC; emails: {m.f.pessanhademenesesribeirodosreis, a.a.akdag}@uu.nl.

1556-4673/2021/12-ART6 $15.00
https://doi.org/10.1145/3477605

**natural language processing (NLP)** domain not only gave the possibility to find specific topics in the transcriptions but also offered insights that can only be achieved via machine learning approaches. This especially applies to big collections where human effort to watch all interviews or to read all transcripts would be too time consuming or even impossible. However, the transformation from audio recordings to print is seen by many oral historians as a source of lamentation for the loss of aural (i.e., subjective) cues [6, 17, 36]. In this article, we argue that advancement in automatic analysis of non-verbal cues have the potential to compensate for this loss, albeit in a manner that will require a familiarity with computational tools and the interpretation of their application, rather than offering an intuitive solution.

In 2006, a prominent oral historian, Michael Frisch, celebrated the digitization practices from old formats to CDs and DVDs, and predicted a future of multi-access to the OH archival material, where everyone would be able to generate their own "documentaries" out of the OH video/audio collections [36]. A year later, Thomson [98] questioned whether the impact of digital revolution will amount to a new paradigm in OH archiving, asking appropriately, "Is this technological revolution also a cognitive revolution?"

OH, as a discipline that has memory, narratives, and everyday practices at its focus, already shares common interests and practices with many neighboring disciplines such as ethnography, anthropology, qualitative sociology, literature and cultural studies, and psychology [69]. These neighboring disciplines have different approaches to qualitative research and interview protocols, and generate audio/video collections similar to OH archives in terms of data types, such as audio/video recordings with field notes and transcriptions. However, among all these approaches that collect similar information, it was OR that embraced the subjective side of the interviewing process, using the act of remembering not only as a tool to understand what happened but also as an object that needs to be studied on its own—that is, how what happened is remembered [92, 98]. For example, an observation of the factual errors of the interviewee can be very important, and the oral historian, instead of fixing the errors, annotates them and interprets why they may be happening, adding footnotes to this history [92]. For the oral historians who emphasize the shortcomings of memory and storytelling as the strong points of OH studies, the non-verbal cues of the interviewees, as well as the dialogue between the interviewee and the interviewer, contain important information that needs to be included in the archive and should be analyzed further to complete the research.

The ability to look at human emotions, and somatic reactions while narrating important life events, both on an individual level and on a collective level, gives scholars from many fields the means to focus on the feelings, mood, culture, and subjective experiences on a mass scale. Naturally, such a computational approach also opens new sources and challenges for data/**artificial intelligence (AI)** scientists as well. The goals of this work are multifold: we offer a review of the computational practices in OH archives while discussing the role of computational non-verbal analysis approaches, and their future potential within OH archival practices. We also aim to stimulate social signal processing researchers to tap into the rich data in OH collections and thereby generate the tools that will render OH archives and similar archives open to analyze for emotions and expressions.

The article is structured as follows. In Section 1, we give a brief overview to key technologies that are already applied within OH archives, highlighting the advancements and the challenges in the state of the art of these technologies and how that affects OH archival practices. Section 2 summarizes the non-verbal cue analysis technologies that are not widely used yet. Section 3 covers the main challenges of these computational approaches to implement in OH archives. In our conclusion in Section 4, we also highlight how non-verbal analysis of oral archives will effect the oral historical research and archiving practices.

## 2 THE COMPUTATIONAL TURN IN OH ARCHIVES: A SHORT OVERVIEW

OH archives mainly contain audio- or video-recorded interviews. These interviews can take various shapes including autobiographical narratives by the interviewee, such as lifestyle interviews, or interviews with semi-structured, open-ended questions following a research agenda, such as thematic interviews, or collective

information sessions, where many persons participate in the conversation [99]. Usually, these interviews are accompanied by field notes, interview questions, questionnaires, reports, and other materials. It is desirable to maintain them in a format where in a given point of an audio or video file the correlated notes, materials, and metadata are made available to the researcher [39]. The sheer size of the archive makes modern computer-based search and retrieval applications and interactive interfaces designed for multimedia retrieval very relevant.[1]

The process of archiving and building suitable interfaces to the OH collections is an important challenge in itself. In 2008, de Jong and Oard [49] proposed an excellent research agenda for this purpose while explaining the key technological components already in use. They observed that the state of the art of many technologies were applied for research in domains other than OH archiving but needed to be further developed to be effectively used in archival interfaces. Here, we complement their review by first summarizing the new developments in the key algorithmic advances in automatic speech and language processing. In the following sections, we first focus on the relevant technologies, and we highlight both the advances and the challenges to give a glimpse of what the present holds and what the future may hold in store for archival sciences.

### 2.1 Automatic Speech Recognition

The curation of large-scale interview collections benefits from a contextualization of each interview, namely by extracting biographical metadata and events named [16]. This structure facilitates the search for related items in the collection, contributing to meaningful research. For this purpose, transcription of the interviews is necessary. Manually transcribing oral archival interviews is very demanding, both due to their duration (with manual transcription time being around 10 times real time [49]) and their emotionally taxing content, frequently containing descriptions of traumatic events. Thus, there is a clear advantage in automating the transcription process.

ASR consists of converting a speech signal into a textual representation. There are several types of natural speech recognition tasks with different challenges associated, among them spontaneous speech—for example, human-to-human dialogue is the most important for OH collections. Generally, ASR tools are based on a standard version of the language. As expected, the training of ASR systems requires a lot of data, which is not available for a lot of languages due to the cost of manual transcriptions and/or the lack of standardized writing systems. In particular, when applied to large collections with multilingual content, where the recordings cover a wide time span, such as found in OH collections, ASR proves to be a difficult task. A paradigm called *transfer learning* is a potential solution to training complex ASR models for under-resourced languages [25]. In this approach, a model trained on a well-resourced language serves as a basis. This layered model that processes the input in a coarse-to-fine way is re-purposed for the new language, and some of the internal representations learned for the basis language are retained. Many neural network models are suitable for transfer learning. Gref et al. [42] proposed the application of neural networks to train robust acoustic models for speech recognition in German OH interviews. The training set consisted of 128 hours from the GerTV1000h Corpus [94], and multi-condition experiences were conducted to analyze the influence of different mixtures of noise-based data augmentation strategies in the performance of the model on the OH archives interviews. These techniques proved to be efficient to better represent the audio-recording conditions observed in the interviews. Later on, a second stage was added to the pipeline, applying transfer learning to the resulting acoustic model to tackle particular challenges of the OH interviews, focused on speech, especially speed, dialects, and pronunciation [41]. For this step, a "leave-one-speaker-out" evaluation was used, utilizing part of the OH interviews to further train the initial acoustic model. The introduction of this second training stage, with the target data, led to improvements in results, proving to be an interesting approach to tackle the lack of large amounts of annotated speech in the OH archives field.

[1]The preparation of all the materials as an archive is different from analyzing them for research. For such qualitative data analysis, software such as Atlas.ti and Nvivo offer a way to organize and connect all the notes to the audio/video recordings while at the same time offering some basic analytical capabilities to researchers with no computational expertise [104].

Like under-resourced languages, dialects pose a challenge to ASR systems. Particularly when analyzing interviews of older interviewees, the language will likely be non-standard, with a representation of different dialects. Additionally, conversational speech has a greater variation in lexical choice than written text, which makes it challenging to learn from one speaker how another would express a similar idea [49]. In general, spontaneous speech analysis will have a higher error rate than text readings, even when applied to widely studied languages [49]. Ideally, a diverse set of audio recordings, with the dialects in the dataset, should be used for a more accurate audio-to-text model [39]. An adequate vocabulary will be essential for a correct transcription and should include domain-specific words, such as common names and entities expected for the interview topic as well as unusual terms typical from the geographic region and/or socio-economic background of the interviewees. A good vocabulary will reduce the number of out-of-vocabulary words, not recognizable by the system. Furthermore, speaker diarization (i.e., automatic recognition and tracking of speakers in a given recording or in a collection) is desirable for OH archive applications to separate the interviewer from the interviewee intervention.

## 2.2 Natural Language Processing

After manual or automatic transcription of speech, further operations can be applied to the resulting text for archival processing purposes. A quick glance at how NLP technologies can assist the enhancement of audiovisual collections brings a number of topics to the fore. One topic is the application of regular expressions (i.e., a formal language for specifying a string) to the transcribed text for pattern matching—for instance, to detect national identification numbers consisting of a set of characters with a specific pattern. Another research area is named entity recognition to identify defined categories such as names, organizations, or geographic locations from context. A third one is "topic modeling," which aims to extract a finite set of topics from a collection of documents, and represents each document in the collection as a mixture drawn from a small set of topics [45]. Finally, a research area that closely relates to archives is to automatically generate metadata and structured summaries.

A recent trend in NLP is the use of Artificial Neural Networks (ANNs) [68]. Currently, the most commonly used model is **Bidirectional Encoder Representations from Transformers (BERT)** [55] with the recently proposed GPT-3 [18] showing strong performance on many NLP datasets and tasks. BERT is an example of a pre-trained model, where a large corpus is used in the initial training, and the trained model is made available as a system itself, or as a feature extraction module that can be used in another system. Pre-trained language models for multiple languages based on the BERT model are available, which is a great development for under-represented languages [24, 27, 78]. Further work in this area offers new solutions for information retrieval [59], knowledge extraction [2, 54, 58], classification [1], and summarization [106].

Several software tools have been developed in the field of NLP with applications for archival processing, such as ePADD [84] and the BitCurator NLP project [53]. These tools allow to extract, analyze, and produce reports on features of interest in the text, and provide essential NLP functionalities. Although commercial software packages may not include state-of-the-art models, their intuitive user interface makes them appealing to use in the archival processing field.

Similar to what was described for ASR, suitable training data is needed for NLP, ideally produced with manually annotated or corrected metadata. Due to the complexity of the interactions in recorded interviews, with frequent interchanges in topic, location, and times, the search tools should allow complex queries to find implied concepts in the speech [49]. Complex queries would also be desirable for context assessment, since keyword searching tools will have added noise to single words with multiple meanings [45].

## 2.3 Bias in Artificial Intelligence Approaches

Today's computational approaches, especially AI algorithms, thrive with big data and computing power. The rise of AI not only happened due to the mathematical advancements in the field that made it possible to create more powerful neural network based models but also benefited from the publication of domain-specific datasets that

contained millions of images, words, and audiovisual materials [28, 57]. To demonstrate the state of the art in ASR and NLP, we referred to these developments, which rendered improved performance and new capabilities across the board. However, the advancement in AI also brings new problems and challenges.

For example, in the work of Koenecke et al. [51], five state-of-the-art ASR systems (Google, Amazon, Apple, IBM, and Microsoft) for English language are tested on different user groups, and a racial parity is reported. All the systems show nearly twice the word error rate for African American speakers compared to white speakers. This is a common metric developed to measure speech recognition systems, where recognized words by the system are compared to spoken words, and all the insertions, deletions, and substitutions of the system are summed up and divided by the number of spoken words. We describe this study in some detail, as it illustrates some of the techniques in this area.

In the work, a collection of socio-linguistic interviews with African Americans (**Corpus of Regional African American Language (CORAAL)** ) and a collection of interviews done in urban and rural California (Voices of California) are used. To analyze the reasons of the difference in word error rate, the work further checks three measures: the **dialect density measure (DDM)**, the proportion of the words used in the CORAAL and Voices of California datasets to the vocabularies of the machine vocabularies, and the perplexity of sentences in each corpus. The first measure focuses on African American vernacular speech features and checks how much is present in a sub-group of CORAAL. The second feature looks at the number of words used in each corpus and compares that to the underlying vocabulary in each ASR system. Both corpus vocabularies are well represented within the machine vocabularies. If a high percentage of the words used only in CORAAL were not covered by the machine vocabularies, this lack would have been explained the racial bias in the results. The last check (i.e., the perplexity measure) looks if the sentence structure in both corpora follows the statistical models used in the underlying language model. ASR systems are based on language models that are trained with millions of words and sentences, based on which the model predicts the next word that would be used in any given sentence. If the sentences are too complex, and statistically not represented in the language model, that could trigger high word error rates as well. The comparison of perplexity of both corpora shows lower rates for African American speakers, which should result in better performance in ASR. Among the three measures checked, the DDM is the most likely reason; the study concludes that the lack of African American speaker data for training the models is the most likely reason for the bias in the results.

Similar problems are reported in the NLP literature, where especially a bias regarding gender is observed. Various methodologies are developed to test the inherent bias in the pre-trained datasets [97]. In the case of NLP systems, word embeddings that are used for such pre-training reflect the societal and cultural bias that comes from the corpus from which they are generated. A well-known example is when a translation system renders a gender-neutral expression like *he/she is a doctor* as a gendered text, such as *he is a doctor*. If the corpus has higher statistics for jobs being carried by females, then the translation shows that as in the example *she is a nurse*. A recent survey work on algorithmic bias in NLP reports other negative biases against "working-class socioeconomic status, male children, senior citizens, Islamic religious faith, non-religiosity and conservative political orientation," among others [79]. These are the results from the most commonly used word embeddings, and the bias will affect the results in both translation and correction tasks. These bias will be aggravated with the rushed development of larger language models without careful consideration of diverse demographics [13]. As described previously, language can incode biases, such as gendered occupations (doctor vs female doctor), contested framings (undocumented immigrants vs illegal immigrants), or derogatory terms that when included in the dataset can lead to problematic associations. When doing large-scale data collection, for instance, from the public domain, this subtly biased language or extreme ideologies may be incorporated into the system amplifying abusive voices. Despite these risks, the perceived accuracy of machine translation (MT) by the consumers is high, and therefore incorrect translations with a coherent structure can be easily overlooked by the user, proliferating misinformation. As OH collections contain minority views, it is an open question how/if such biases would be amplified in these interviews and how this problem can be solved.

### 2.4 Diffusion of Technologies to OH Archives

The developments in ASR and NLP fields we have summarized in this section apply to many real-life scenarios. From personal artificial assistants that process conversations on the fly to NLP models that are capable of generating new text that matches a given style, many commercial applications are available. However, as de Jong and Oard [49] observed a decade ago, it takes time until these developments are transferred to OH archival practices. There are several reasons for that. A recent work [77] points out the difficulty in applying the latest advancements in AI to record-keeping processes. One of the main reasons for this difficulty is seen (ironically) as the lack of applied examples, which might be a sign of resistance on the side of the back-end users. A more telling reason given is the amount of data annotation and computational resources that are needed to prepare AI algorithms for the task. A last relevant reason is the time needed not only to configure machine learning solutions but also the time and manpower needed for the evaluation of the results. Similarly, a recent survey on the use of technologies within the domain of OH archives found that ASR and NLP are seen as problematic in terms of accuracy and user-friendliness [103].

Today, technologies developed for non-verbal communication are at a stage where development and deployment of applications in various domains are thriving. However, the adaptation of these technologies for OH archival research is not forthcoming yet. Once that stage is reached, it will still take a good while until a transformation in the structure and interface of archives takes place. Hence, in the next section, we will only focus on research practices of non-verbal communication and give examples of how they can be utilized in OH archival (or related) research agendas.

## 3 HOW TO AUTOMATICALLY ANALYZE THE SUBJECTIVE SIDE OF THE ARCHIVES?

The interview collections in OH archives present us with a subjective perspective on historical events [52]. The information is not solely in the narrative, but it also in the breaks and gaps of it; gestures, facial expressions, periods of silences, and other non-verbal interactions carry as much information as the spoken word [52, 69]. It is no wonder that Portelli [74], a leading oral historian, is widely cited in his observation that "transcripts not only fail to convey the essence of the interview space, but also service to flatten the emotional content of speech" (p. 35). We argue in this article that the advancements in affective computing, social signal processing, and automatic analysis of non-verbal communication is providing us with new tools that can capture the subjective and the emotional content of the OH archives as well.

In this section, we will give an overview on the state of the art in social signal processing [70, 101], as well as related advances in affective computing, which produces tools that automatically analyze emotional signals from multiple modalities. A particularly relevant application of emotion detection of OH is finding correlations between non-verbal signals and mental health problems, such as depression or **post-traumatic stress disorder (PTSD)**. Many OH collections contain narratives of traumatic events, told by survivors.[2] Environmental triggers such as natural disasters, war, mass migration, or other violent occurrences due to the unstable politic regimes generate mass trauma, and oral historians take on the task of recording the survivors stories. In OH research, trauma so often takes the focal point that *trauma oral history* has become a commonly referred term in the discipline [76]. Hence, here we will use trauma as a topic or, better still, as a case study to demonstrate the potential of non-verbal communication technologies for oral historians and data scientists, as well as psychiatrists.

### 3.1 Paralinguistics

Speech is a rich source of verbal and non-verbal information that offers glimpses on not only the mind-set of the speakers but also how they feel. Furthermore, through an analysis of the voice quality, the mood and even

[2]Oral historians warn about the dangers of framing an OH archive collection as a source of "traumatic stories," pointing out how refugees or similar minority groups are expected to tell about their traumatic experiences. Especially refugees who search for a safe haven feel the pressure to tell what the authorities want to hear [75].

the personality of the speaker can be categorized [85]. The research area of "paralinguistics" is the analysis of voice quality, and although linguistic analysis explores *what is said*, paralinguistics researches *how it is said*. Beside a large set of acoustic measures (called *prosodic features*) such as intonation, rhythm, tone, and stress, paralinguistics also analyzes filler sounds such as coughs, laughter, moans, gasps, grunts, or utterances, and the silences between speech instances. All these non-verbal signals have a great significance when we try to understand the emotional state of a speaker, even if they are not always explicitly attended to.

Paralinguistic features are useful for many application domains from speech recognition, speakers' intention interpretation [15], and conversation analysis to health-related analysis, to name a few [4]. But they are most commonly needed in emotion recognition. The analysis of linguistic cues can be used to evaluate emotional stability and infer the personality of the speaker [64]. For example, voices from subjects with PTSD or depression were found to have significantly tenser voice quality [83]. Moreover, depressed subjects are prone to possess a low dynamic range of the fundamental frequency, a slow speaking rate, a slightly shorter speaking duration, and a relatively monotone delivery.

Computational paralinguistics offers different feature sets and classification frameworks [86]. There are several toolboxes, such as the openSMILE toolkit [34], or the COVAREP repository [26], which provide feature extraction capabilities for the analysis of paralinguistic features in speech. These toolkits enable the extraction of a diverse set of **low-level descriptors (LLDs)**, such as pitch, Mel-frequency cepstrum coefficient features based on human speech production, signal energy. and spectral features. The processing typically adds various filters, functionals, and transformations to these low-level features. The LLDs are extracted by applying a windowing function to the acoustic signal, either in the time, frequency, or time-frequency domain and by studying the characteristics of the wavelength. LLDs can then be used, for instance, for classification or regression purposes, or model learning.

At the moment, there are very few existing scientific works on processing OH archives with the tools of paralinguistics. Working on an OH dataset documenting war and violence in Croatia, de Jong et al. [23] modeled the verbal and prosodic features of emotional expression in narrative data. The correlation between non-verbal emotional expression in the voice (pitch, vocal effort, and pauses) and changes in the intensity of emotional expressions during the interviews were analyzed under the hypothesis that the interviewee would be more emotional after open-ended questions. Such open-ended questions are usually introduced in the latter part of the interview. The emotional expression value was calculated with the percentage of emotion words. A correlation in the intensity of the verbal expression and the pitch and pause duration was observed, but there was not a correlation for vocal effort.

Studies like that of de Jong et al. [23] offer a new set of tools for OH research: during the early history of OH research, to establish an objective approach to the object of study, the focus was in generating interview protocols [90]. Even though with time the majority in the field of OH agreed that every interview is unique and it is not always possible to exactly follow an interview protocol, the common practice still relies on them. Today it is possible to analyze different OH collections focusing on the emotional state of the interviewees in relation to interview protocols in general, and gain new insights to the design of interviews, where, for example, different routes might prove better depending, for example, on the age, or gender of the interviewee.

### 3.2 Facial Expressions

Face expression recognition is a mature field with many applications for human-computer interaction, human behavior understanding, and mental state recognition. Although the link between emotions and facial expression displays is not contested, it is far from trivial. Even a simple smile expression, arguably the easiest expression to detect, can reveal nuances when analyzed in depth, such as posed and spontaneous smiles [29] or smiles signaling embarrassment [5].

Following Darwin, Ekman [32] proposed that there are basic facial emotional expressions that are recognizable across cultures. For two decades, the basic emotions (i.e., happiness, fear, anger, surprise, sadness, disgust)

dominated the affect analysis from faces, and many algorithms and toolboxes have been developed to automatically classify facial images into emotional categories. More recently, these so-called categorical approaches were supplemented by dimensional approaches, where the apparent emotion shown on the face is mapped to a continuous, 2D (i.e., valence and arousal) or 3D (adding dominance) feature space [65]. There were attempts to detect combinations of categorical expressions (e.g., happily surprised) instead of basic emotions [30], and arguments that cross-cultural recognition of basic emotions is far from perfect [47]. An objective annotation system for facial muscle movements, the Facial Action Coding System, was proposed by Ekman [33], and it remains a standard tool in face analysis. However, manual coding of facial movements with the Facial Action Coding System is extremely time consuming and difficult, requiring highly trained coders. Subsequently, automatic analysis tools that can do such a coding from arbitrary videos have been developed [12]. These tools can also provide head pose analysis and gaze direction estimation, which are important social signals. However, the idiosyncratic variations in facial expressions, the highly contextual and semantic interpretation requirements, and the difficulty of processing faces under non-ideal pose and illumination conditions (i.e., the so-called "in the wild" recording conditions) make this problem far from solved. At the time of this writing, there are no algorithms that can detect the traces of sarcasm in a smile.

The potential of face analysis to OH archives is clear, as many archives contain video recordings of speakers' faces. Using visual information such as facial expressions, and how the interviewees gaze travel, it might be possible not only to get a more complete understanding of the interviewee's emotional state but also to reconstruct the relationship between the interviewee and the interviewer. Normally, the position and the role of the interviewer is hidden in these videos, and unless special attention is given, it is not possible to follow the communication between the interviewee and the interviewer. Through such automatic processing tools, it becomes possible to derive analytics of gaze and expressions, highlight distributions of affective moments in the archive, search for patterns, and complement verbal analysis.

### 3.3 Body Language

An important part of non-verbal communication is hidden in the way we use our bodies. Body language conveys mood and affective state, provides non-verbal communication signals, and regulates turn-taking and interactions. Hand gestures can denote specific content, replacing or enhancing certain concepts, and provide indexing and symbolic cues [72]. To decipher an interview and the communicative acts of an interviewee in their fullness, the interpretation of body cues and hand gestures would be essential. Certain hand gestures, called *self-adaptors*, such as hand tapping, stroking, or grooming, have been shown to be correlated to anxiety/depression disorders [35]. A study in trauma oral archives for such hand self-adaptors, for instance, could generate insights about the cultural codes and their relation to self-adaptors.

Body postures, gaze direction, and gestures contain rich signals into the attitude of the talking person, as well as social cues such as agreement and disagreement. For example, in political speech videos, automatic analysis approaches have been used successfully to gain insight into attitudes and the use of persuasive discourse, using gesture and gaze cues [73]. Similar research in an OH collection could bring a new perspective into the discourse of "shared authority" in OH interview protocols [91]. The concept was defined first by Frisch et al. [43], to stress the importance of a dialogue between interviewee and interviewer, in which they negotiate their differences to reach a common ground, to build trust. It was also used to acknowledge the fact that both interviewer and the interviewee have their own agendas [48]. Shared authority is defined, discussed, and expanded to new definitions such as "sharing authority"; however, how it should be applied in research is never explained [91]. A computational study to document the turn-taking, persuasion, and trust in OH interviews would generate a needed example space of how sharing authority in practice is done.

The automatic approach to body language interpretation treats signs and gestures as a pattern classification problem. Here, the analyst specifies a number of gestures, poses, or body postures of interest, and uses a computer

vision based method to automatically detect them in video. For example, touching the face is a prominent signal for certain states (e.g., confusion or concentration), and it is possible to implement a system that can automatically detect when the interviewee is touching their face [14]. Such behavior is also monitored via wearable sensors, but in interview settings, this may be impractical.

Human body analysis and pose estimation received a lot of attention in the past few years because of their vast commercial applications (e.g., activity recognition in smart environments and pedestrian detection for autonomous driving). Deep neural network based methods are developed for this purpose [40, 56, 107], and tools are prepared that are relatively easy to incorporate into software systems. A frequently used tool is OpenPose, which automatically fits a 3D skeleton model to the persons in an image, allowing the quantification of pose, limb angles, and body movement speed [20]. Pose prediction is challenging due to the variety of conditions that affect the appearance (e.g., different lighting conditions, clothing, camera angles, occlusions, and self-occlusions) and requires additional assumptions for robust evaluation [7, 46, 100].

Model-based approaches to body pose estimation expect the full body to be visible. This causes a problem for videos where only the upper body is visible. Unfortunately, many OH interview videos use a camera angle that focuses on the face, and include the upper body only. Moreover, hand gestures may also be missed, especially if the hands remain on the lower body, resting on the knees. We should also note that missing limbs will create problems with many existing automatic approaches, precisely for the same reason.

Affect is also linked to the body pose in a general way. Although it is known that speech and facial analysis are richer sources of affect estimation, the body is an additional source that can be used. Happiness and excitement are revealed with more active body movements, and sadness is perceived by a shrunken body, bowed shoulders, and a bent head [66]. Emotions can be recognized by modeling each body part independently for analysis, or by implementing a structural body model, analyzing the entire pose for emotion recognition [82].

### 3.4 Breathing Analysis

Breathing is a physiological function that is regulated automatically, unless a person consciously focuses on it. Although we do not use breathing to communicate explicitly, it still signals affective cues, as it is an anatomical action that changes with different emotions. Studies show a correlation between respiratory feedback and emotions such as joy, anger, fear, and sadness [71]. Subsequently, looking at breathing patterns of subjects in an archive can provide insights about the affective state of the subjects. For example, narrating painful events can cause a conceivable change in breathing patterns. In a sense, remembering a traumatic event equals to "re-experiencing" the old associations [31]. Salah et al. [3] investigated how breathing patterns change in survivor testimonies and illustrated how these can be used to mark emotional moments during recordings. A broader question they asked is whether traumatic experiences leave somatic traces in subjects that transcend languages, cultures, and geographies. This requires a large-scale analysis of multi-cultural OH collections.

Breathing signals vary in intensity, length, and strength. For example, sighs are audible signals that denote relief, desire, or boredom. On a physiological level, these deep and fast breaths help to gain respiratory control, playing a homeostatic role to restore calmness after an emotional state [102]. Deeper analysis of breathing patterns can be revealing. Expressions such as laughter affect breathing patterns, resulting in large exhalations, which help with "de-escalating" stressful reactions to negative emotions [88]. This may explain the use of humor in some survivor stories. Humor can reduce the intensity of the traumatic stress reactions, proving to be a viable coping mechanism [37]. Personality difference and the level of personal anxiety also influence breathing patterns, especially during mental stress [44].

Analyzing breathing patterns via machine learning approaches in non-verbal communication is a relatively new research line. Most approaches focus on automatic breathing detection, which is especially challenging for uncontrolled recording conditions. Because of the temporal nature of the problem, approaches that model the sound dynamics are preferred, such as long short-term memory models [60, 61, 63]. Long short-term memory

networks contain memory units to preserve and propagate information over time, rendering them especially useful for audiovisual datasets.

There are recent studies that relate breathing patterns to emotional states and mental health. For example, using breath signal information, Cho et al. [21] proposed a neural network based classifier to discriminate between levels of stress while performing a task. In the work of Kaya et al. [50], signals from non-verbal parts of the recordings, such as breathing and silences, are combined with linguistic information for automatic depression detection.

The interest on breathing signals is reflected in the recently organized INTERSPEECH 2020 Breathing Challenge, which was based on a new dataset of spontaneous speech, recorded in a studio setting [87]. The breathing signals were collected with a piezoelectric respiratory belt that returns a continuous ground truth value. In the absence of such a sensor (as is the case with most OH collections), annotation of breathing signals is a difficult and time-consuming task. However, such annotated datasets can be used to develop supervised learning methods that will work on archival data to approximate the functioning of the respiratory belt [60].

### 3.5 Heart Rate Detection

The **heart rate (HR)** is another physiological signal that changes due to physical triggers, as well as due to emotional states. The **heart rate variability (HRV)** is similarly connected to emotional responses. A classical example is the "fight-or-flight" reaction, with an increase in HR and HRV to activate mobilization responses.

The relationship between HR and non-verbal communication is not fully researched, but some direct associations are confirmed. HRV has been proposed as an important marker of emotion regulatory ability, particularly as it relates to social processes and mental health [9]. For example war veterans, when presented with trauma-related cues, such as imaginal material or sounds, experience an increased HR [22, 89]. Subsequently, automatic extraction of HR and HRV can provide valuable insights into the mental state of a subject. Recent progress in computer vision enables such analyses from archival material.

The beating of a heart is a familiar sound for all of us; however, we do not perceive the HR and related biosignals consciously. In medical settings, the HR is typically measured using an electrocardiogram (ECG) with electrodes placed on the body. Recent advances in computer analysis of human behavior enabled non-contact procedures, which can be used with pre-recorded archival videos. A technique called *Eulerian video magnification* was introduced to amplify subtle color and movement changes on the face and the body to "amplify" the HR signal to make it visible [105]. Other visual cues used in automatic approaches include subtle head movements [11], facial features [80], and a combination of speech and facial features [8].

### 3.6 Challenges of Social Signal Processing in OH Archives

OH collections cover a wide range of languages and cultures. When working with data from diverse cultural backgrounds, it is important to keep in mind that social signals vary across culture, gender, age, and identity. There are even community-specific social signals, with variations according to the age and gender of their members. Studies suggest some commonality that surfaces among such rich variations: for example, facial expressions of basic emotions, as well as postures, are similar across cultures to some degree. However, how much emotions are expressed is clearly affected by the cultural codes, and individuals learn how to express or control their emotions [62].

The way culture influences individual behavior is not easy to assess. In an example presented in the work of Summerfield [96], Australian and New Zealand veterans of the First War were observed to construct an idealized soldier, as loyal and patriotic, following the notions of Australian masculinity. However, this construction necessitated them to suppress their painful memories that did not fit the cultural codes of the time. In contrast, the testimonies of women involved in the Home Guard during the Second World War were often fragmented

and deflected within the interview context, possibly due to the lack of a cultural frame of reference for women working in warfare.

Gender has a significant influence on how the individual communicates. Either due to cultural expectations or body composition, men and women tend to move and behave differently [66]. Dire emotional situations like traumatic experiences are also processed differently by men and women. Hence, using gender-dependent models for non-verbal communication cues to detect depression and PTSD raises the performance of automatic approaches [95].

Age also has a considerable effect in verbal and non-verbal communication. For example, the state of the art for both speech recognition and paralinguistic analysis are underdeveloped research lines when it comes to processing the speech of elderly or children. Both age groups have significantly different acoustic features than the norm that is represented in the available datasets, which are used to train the state-of-the-art models [93]. To re-train these models, large datasets for the elderly and children need to be generated. Within the digital humanities studies, a similar (and recurring) problem comes to the fore when working with historical data: a face and gender recognition algorithm trained with millions of images from the 21st century will fail to perform as good in a collection of portrait paintings from the 18th century [81]. The OH collections host rich datasets of interviews done with elderly people. However, to prepare these collections as a training resource needs annotation by field experts—that is, scholars who are familiar with the language, culture, community, and time period in question. This calls for collaboration between many disciplines, and might be the biggest obstacle for the diffusion of AI algorithms to OH archiving research, and similarly, to find new datasets to be used in training automatic systems.

## 4 CONCLUSION

Computing social signals between the interviewee and the interviewer, as well as analyzing the subjectivity in an audiovisual narration, will enrich OH archives and contribute to many audiovisual collections. One obvious benefit lies in understanding the changes in emotions during remembering, storytelling, and conversing. Combined with the transcriptions of a record, the scholars and the other end users of the archives will have a better understanding of what is said, as well as how it is said. Better still, an augmented map of each record and maps of a collection can be generated, with an ability to zoom-in to a specific time in a record or zoom-out to different levels.

In this article, we highlighted the latest developments in key technologies for multimedia archiving practices such as NLP and ASR. We furthermore introduced non-verbal signal processing as a potential source that would enrich OH as well as neighboring domains. OH archives constitute only a portion of all audio/video collections. There are various scholarly practices and communities that generate and maintain similar collections [19]. Even though the research and archiving practices for these show differences, the benefits of developments at the intersection of social signal processing and archival practices will be felt within all these domains. Hence, throughout the article, we argued that once utilized, this potential will have wider influences than OH practices, and by way of concluding, we briefly summarize the expected impact for different domains.

*For humanities scholars.* As we have stressed more than once, the oral qualities of the OH collections are too valuable to lose, too cumbersome to access in a user-friendly way, and too difficult to have an overview without the help of computational technologies. As noted by Frisch [36]: "Everyone knows that there are worlds of meaning that lie beyond words; nobody pretends for a moment that the transcript is in any real sense a better representation of reality than the voice itself. Meaning inheres in context and setting, in gesture, in tone, in body language, in expression, in pauses, in performed skills and movements. To the extent we are restricted to text and transcription, we will never locate such moments and meaning, much less have the chance to study, reflect on, learn from, and share them" (p. 2). The successful application of non-verbal computational tools in the analysis of OH archives will generate new insights to oral historians. Similarly, humanities and social science

scholars working with qualitative interview data, as well as multimedia resources such as movies, will find a valuable toolkit. However, until these technologies can be integrated into the interfaces to audiovisual collections, collaboration with data scientists, and a familiarity or, better still, a computational literacy might prove necessary.

*For (OH) archivists.* Collections that are transcribed (i.e., turned into a printed format) fit the traditional approach of archival studies where the document (i.e., written word) was the norm. The oldest archival principles were formulated to accommodate the physical characteristics of the text. Hence, to generate easy access to the oral and visual qualities of audiovisual collections was a challenging task. With the newest multimedia interfaces, it is possible to locate specific instances within collections, watch the corresponding instance in a video file, or listen to it in an audio file. The introduction of non-verbal computational methodologies to OH collections and interfaces will bring new opportunities, as well as new challenges, as these will generate new datasets of the collections that need to be added as data, metadata, and new layers to interfaces. In archival sciences, arguments for the need of a modular, flexible, and agile infrastructure are already put to implementation and discussion [67]. To adapt to new technologies such as non-verbal communication analysis asks not only for computational and economical resources but also for new guidelines and maybe a new understanding in archival studies. Adding emotional states as metadata to an audiovisual collection will not only enhance the information retrieval experience of OH archive users but also will become a milestone in the archival science.

*For AI/data scientists.* AI technologies work with computational and data resources. In theory, thanks to social media, many audiovisual datasets can be collected from the generously growing user-generated content. However, in practice, to collect, prepare, and generate metadata for such datasets is a very time consuming activity. It is difficult to find a topically focused but subjectively diverse set of collections like OH archives can offer. Moreover, OH interviews are usually long and follow basic standards (the same point of view in camera angles, the same recording equipment for all participants). These qualities make OH collections interesting and potentially very informative datasets.

Each tool becomes an extension of its user and changes the physical or cognitive capabilities of the user. Social signal processing approaches, as they mature, will be packaged into easy-to-use software tools that will give scholars of audiovisual collections new ways of interacting with the archives. New technologies such as these will offer not only novel ways of perusing the archives but also will generate new spaces within the archive, where information retrieval can amount to doing research on various levels, bringing new patterns to the light of day.

## REFERENCES

[1] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. DocBERT: BERT for document classification. *arXiv preprint arXiv:1904.08398* (2019).

[2] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1638–1649.

[3] Almila Akdag Salah, Albert Ali Salah, Heysem Kaya, Metehan Doyran, and Evrim Kavcar. 2020. The sound of silence: Breathing analysis for finding traces of trauma and depression in oral history archives. *Scholarship in the Digital Humanities* 36, Suppl. 2 (2020), ii2–ii8.

[4] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. 2018. Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing* 9, 4 (2018), 478–490. https://doi.org/10.1109/TAFFC.2016.2634527

[5] Zara Ambadar, Jeffrey F. Cohn, and Lawrence Ian Reed. 2009. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *Journal of Nonverbal Behavior* 33, 1 (2009), 17–34.

[6] Kathryn Anderson and Dana C. Jack. 2002. Learning to listen: Interview techniques and analyses. In *The Oral History Reader*. Routledge, 171–185.

[7] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. 2018. PoseTrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5167–5176.

[8] Haydar Ankışhan. 2019. Estimation of heartbeat rate from speech recording with hybrid feature vector (HFV). *Biomedical Signal Processing and Control* 49 (2019), 483–492.

[9] Bradley M. Appelhans and Linda J. Luecken. 2006. Heart rate variability as an index of regulated emotional responding. *Review of General Psychology* 10, 3 (2006), 229–240.
[10] Linda H. Armitage and Peter G. B. Enser. 1997. Analysis of user need in image archives. *Journal of Information Science* 23, 4 (1997), 287–299.
[11] Guha Balakrishnan, Fredo Durand, and John Guttag. 2013. Detecting pulse from head motions in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3430–3437.
[12] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings of the 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG'18)*. IEEE, Los Alamitos, CA, 59–66.
[13] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.
[14] Cigdem Beyan, Matteo Bustreo, Muhammad Shahid, Gian Luca Bailo, Nicolo Carissimi, and Alessio Del Bue. 2020. Analysis of face-touching behavior in large scale social interaction dataset. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 24–32.
[15] Daniel Bone, Chi-Chun Lee, and Shrikanth Narayanan. 2014. Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features. *IEEE Transactions on Affective Computing* 5, 2 (2014), 201–213.
[16] Jessie Both, Didi de Hooge, Ramses IJff, Oana Inel, Victor de Boer, and Lora Aroyo. 2017. Linking Dutch World War II cultural heritage collections with events extracted by machines and crowds. In *Proceedings of the SEMANTICS Workshops*.
[17] Tim Bowden. 2005. Let's not throw the baby out with the bathwater. *Oral History Association of Australia Journal* 27 (2005), 63.
[18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
[19] Silvia Calamai and Francesca Frontini. 2018. FAIR data principles and their application to speech and oral archives. *Journal of New Music Research* 47, 4 (2018), 339–354.
[20] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008* (2018).
[21] Youngjun Cho, Nadia Bianchi-Berthouze, and Simon J. Julier. 2017. DeepBreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In *Proceedings of the 2017 7th International Conference on Affective Computing and Intelligent Interaction (ACII'17)*. IEEE, Los Alamitos, CA, 456–463.
[22] J. Mark Davis, Henry E. Adams, Madeline Uddo, Jennifer J. Vasterling, and Patricia B. Sutker. 1996. Physiological arousal and attention in veterans with posttraumatic stress disorder. *Journal of Psychopathology and Behavioral Assessment* 18, 1 (1996), 1–20.
[23] Franciska de Jong, Khiet Truong, Gerben Westerhof, Sanne Lamers, and Anneke Sools. 2013. Emotional expression in oral history narratives: Comparing results of automated verbal and nonverbal analyses. In *Proceedings of the Workshop on Computational Models of Narrative (CMN'13)*. 310–314.
[24] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT model. *arXiv preprint arXiv:1912.09582* (2019). http://arxiv.org/abs/1912.09582.
[25] G. Deekshitha and Leena Mary. 2020. Multilingual spoken term detection: A review. *International Journal of Speech Technology* 23, 3 (2020), 653–667.
[26] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP–A collaborative voice analysis repository for speech technologies. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, Los Alamitos, CA, 960–964.
[27] Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: A Dutch RoBERTa-based language model. *arXiv preprint arXiv:2001.06286* (2020). http://arxiv.org/abs/2001.06286.
[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 248–255.
[29] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. 2012. Are you really smiling at me? Spontaneous versus posed enjoyment smiles. In *Proceedings of the European Conference on Computer Vision*. 525–538.
[30] Shichuan Du, Yong Tao, and Aleix M. Martinez. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* 111, 15 (2014), E1454–E1462.
[31] Kirsten Ekerholt and Astrid Bergland. 2008. Breathing: A sign of life and a unique area for reflection and action. *Physical Therapy* 88, 7 (2008), 832–840.
[32] Paul Ekman. 1999. Basic emotions. In *Handbook of Cognition and Emotion*, Tim Dalgleish and Mick J. Power (Eds.). John Wiley & Sons, 45–60.
[33] Rosenberg Ekman. 1997. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press.
[34] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*. 1459–1462.

[35] Lynn A. Fairbanks, Michael T. McGuire, and Candace J. Harris. 1982. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of Abnormal Psychology* 91, 2 (1982), 109.
[36] Michael Frisch. 2006. Oral history and the digital revolution: Toward a post-documentary sensibility. In *The Oral History Reader* (2nd ed.), Robert Perks and Alistair Thompson (Eds.). Routledge, London, UK, 102–114.
[37] Jacqueline Garrick. 2006. The humor of trauma survivors: Its application in a therapeutic milieu. *Journal of Aggression, Maltreatment & Trauma* 12, 1–2 (2006), 169–182.
[38] Katharine Gavrel and UNESCO. 1990. *Conceptual Problems Posed by Electronic Records: A RAMP Study*. Unesco.
[39] Anne-Sophie Ghyselen, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen, and Arjan van Hessen. 2020. Clearing the transcription hurdle in dialect corpus building: The corpus of Southern Dutch dialects as case-study. *Frontiers in Artificial Intelligence (Lausanne)* 3 (2020), 1–17.
[40] Ross Girshick. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*. 1440–1448.
[41] Michael Gref, Christoph Schmidt, Sven Behnke, and Joachim Köhler. 2019. Two-staged acoustic modeling adaption for robust speech recognition by the example of German oral history interviews. In *Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME'19)*. IEEE, Los Alamitos, CA, 796–801.
[42] Michael Gref, Christoph Schmidt, and Joachim Köhler. 2018. Improving robust speech recognition for German oral history interviews using multi-condition training. In *Proceedings of the 13th ITG Symposium on Speech Communication*. 1–5.
[43] Steven High. 2009. Sharing authority: An introduction. *Journal of Canadian Studies* 43, 1 (2009), 12–34.
[44] Ikuo Homma and Yuri Masaoka. 2008. Breathing rhythms and emotions. *Experimental Physiology* 93, 9 (2008), 1011–1021.
[45] Tim Hutchinson. 2020. Natural language processing and machine learning as practical toolsets for archival processing. *Records Management Journal* 30, 2 (2020), 155–174.
[46] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. 2017. ArtTrack: Articulated multi-person tracking in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6457–6465.
[47] Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* 109, 19 (2012), 7241–7244.
[48] Erin Jessee. 2011. The limits of oral history: Ethics and methodology amid highly politicized research settings. *Oral History Review* 38, 2 (2011), 287–307.
[49] Franciska de Jong and Douglas W. Oard. 2008. Access to recorded interviews : A research agenda. *Journal on Computing and Cultural Heritage* 1, 1 (2008), 1–27. https://doi.org/10.1145/1367080.1367083
[50] Heysem Kaya, Dmitrii Fedotov, Denis Dresvyanskiy, Metehan Doyran, Danila Mamontov, Maxim Markitantov, Alkim Almila Akdag Salah, Evrim Kavcar, Alexey Karpov, and Albert Ali Salah. 2019. Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. In *Proceedings of the 9th InternationalAudio/Visual Emotion Challenge and Workshop*. 27–35.
[51] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.
[52] Dominick LaCapra. 2016. Trauma, history, memory, identity: What remains? *History and Theory* 55, 3 (2016), 375–400.
[53] Christopher A. Lee, Matthew Kirschenbaum, Alexandra Chassanoff, Porter Olsen, and Kam Woods. 2012. Bitcurator: Tools and techniques for digital forensics in collecting institutions. *D-Lib Magazine* 18, 5–6 (2012), 14–21.
[54] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 65–71.
[55] Hsiao-Yun Lin, Tien-Hong Lo, and Berlin Chen. 2019. Enhanced BERT-based ranking models for spoken document retrieval. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'19)*. IEEE, Los Alamitos, CA, 601–606.
[56] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2980–2988.
[57] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. 740–755.
[58] Xiao Liu, Heyan Huang, and Yue Zhang. 2019. Open domain event extraction using neural latent variable models. *arXiv preprint arXiv:1906.06947* (2019).
[59] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1101–1104.
[60] Maxim Markitantov, Denis Dresvyanskiy, Danila Mamontov, Heysem Kaya, Wolfgang Minker, and Alexey Karpov. 2020. Ensembling end-to-end deep models for computational paralinguistics tasks: ComParE 2020 mask and breathing sub-challenges. In *Proceedings of the 2020 INTERSPEECH Conference*.
[61] John Mendonça, Francisco Teixeira, Isabel Trancoso, and Alberto Abad. 2020. Analyzing breath signals for the INTERSPEECH 2020 ComParE challenge. In *Proceedings of the 2020 INTERSPEECH Conference*. 2077–2081.

[62] Batja Mesquita and Nico H. Frijda. 1992. Cultural variations in emotions: A review. *Psychological Bulletin* 112, 2 (1992), 179.
[63] Venkata Srikanth Nallanthighal and H. Strik. 2019. Deep sensing of breathing signal during conversational speech. In *Proceedings of the 2019 INTERSPEECH Conference*. 4110–4114.
[64] Michael Neff, Nicholas Toothman, Robeson Bowmani, Jean E. Fox Tree, and Marilyn A. Walker. 2011. Don't scratch! Self-adaptors reflect emotional stability. In *Proceedings of the International Workshop on Intelligent Virtual Agents*. 398–411.
[65] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2, 2 (2011), 92–105.
[66] Fatemeh Noroozi, Dorota Kaminska, Ciprian Corneanu, Tomasz Sapinski, Sergio Escalera, and Gholamreza Anbarjafari. 2018. Survey on emotional body gesture recognition. *arXiv preprint arXiv:1801.07481* (2018).
[67] Roeland Ordelman, Carlos Martínez Ortíz, Liliana Melgar Estrada, Marijn Koolen, Jaap Blom, Willem Melder, Jasmijn Van Gorp, et al. 2018. Challenges in enabling mixed media scholarly research with multi-media data in a sustainable infrastructure. In *Proceedings of the Digital Humanities 2018 Conference*.
[68] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* 32, 2 (2020), 604–624.
[69] Cord Pagenstecher and Stefan Pfänder. 2017. Hidden dialogues: Towards an interactional understanding of oral history interviews. In *Oral History Meets Linguistics*, Erich Kasten, Katja Roller, and Joshua Wilbuer (Eds.). Kulturstiftung Sibirien, Furstenberg/Havek, Germany, 185–207.
[70] Alex Pentland. 2007. Social signal processing [exploratory DSP]. *IEEE Signal Processing Magazine* 24, 4 (2007), 108–111.
[71] Pierre Philippot, Gaëtane Chapelle, and Sylvie Blairy. 2002. Respiratory feedback in the generation of emotion. *Cognition & Emotion* 16, 5 (2002), 605–627.
[72] Pramod Kumar Pisharady and Martin Saerbeck. 2015. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding* 141 (2015), 152–165.
[73] Isabella Poggi and Laura Vincze. 2008. Gesture, gaze and persuasive strategies in political discourse. In *Proceedings of the International LREC Workshop on Multimodal Corpora*. 73–92.
[74] Alessandro Portelli. 2009. What makes oral history different. In *Oral History, Oral Culture, and Italian Americans*. Springer, 21–30.
[75] Katherine Randall, Katrina M. Powell, and Brett L. Shadle. 2020. Resisting the trauma story: Ethical concerns in the oral history archive. *Displaced Voices: A Journal of Archives, Migration and Cultural Heritage* 1 (2020), 6–11.
[76] Robert Reynolds. 2012. Trauma and the relational dynamics of life-history interviewing. *Australian Historical Studies* 43, 1 (2012), 78–88.
[77] Gregory Rolan, Glen Humphries, Lisa Jeffrey, Evanthia Samaras, Tatiana Antsoupova, and Katharine Stuart. 2019. More human than human? Artificial intelligence in the archive. *Archives and Manuscripts* 47, 2 (2019), 179–203.
[78] Willem Röpke, Roxana Radulescu, Kyriakos Efthymiadis, and Ann Nowé. 2019. Training a speech-to-text model for Dutch on the Corpus gesproken Nederlands. In *Proceedings of the 31st Benelux Conference on Artificial Intelligence (BNAIC'19)*.
[79] David Rozado. 2020. Wide range screening of algorithmic bias in word embedding models using large sentiment lexicons reveals underreported bias types. *PLoS One* 15, 4 (2020), e0231189.
[80] Aibek Ryskaliyev, Sanzhar Askaruly, and Alex Pappachen James. 2016. Speech signal analysis for the estimation of heart rates under different emotional states. In *Proceedings of the 2016 International Conference on Advances in Computing, Communications, and Informatics (ICACCI'16)*. IEEE, Los Alamitos, CA, 1160–1165.
[81] Cihan Sarı, Albert Ali Salah, and Alkım Almıla Akdag Salah. 2019. Automatic detection and visualization of garment color in Western portrait paintings. *Digital Scholarship in the Humanities* 34, Suppl. 1 (2019), i156–i171.
[82] Nikolaos Savva, Alfonsina Scarinzi, and Nadia Bianchi-Berthouze. 2012. Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience. *IEEE Transactions on Computational Intelligence and AI in Games* 4, 3 (2012), 199–212.
[83] Stefan Scherer, Giota Stratou, Gale Lucas, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Louis-Philippe Morency, et al. 2014. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing* 32, 10 (2014), 648–658.
[84] Josh Schneider. 2016. ePADD: Supporting archival appraisal, processing, and research for e-mail collections. *MAC Newsletter* 43, 3 (2016), 8.
[85] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Muller, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language—State-of-the-art and the challenge. *Computer Speech & Language* 27, 1 (2013), 4–39.
[86] Bjorn W. Schuller. 2012. The computational paralinguistics challenge [social sciences]. *IEEE Signal Processing Magazine* 29, 4 (2012), 97–101.
[87] Björn W. Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, et al. 2020. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, breathing and masks. In *Proceedings of the 2020 INTERSPEECH Conference*.
[88] Sophie K. Scott, Nadine Lavan, Sinead Chen, and Carolyn McGettigan. 2014. The social life of laughter. *Trends in Cognitive Sciences* 18, 12 (2014), 618–620.

[89] Arieh Y. Shalev, Tali Sahar, Sara Freedman, Tuvia Peri, Natali Glick, Dalia Brandes, Scott P. Orr, and Roger K. Pitman. 1998. A prospective study of heart rate response following trauma and the subsequent development of posttraumatic stress disorder. *Archives of General Psychiatry* 55, 6 (1998), 553–559.
[90] Anna Sheftel and Stacey Zembrzycki. 2019. Who's afraid of oral history? Fifty years of debates and anxiety about ethics. *Oral History Review* 43, 2 (2019), 338–366.
[91] Linda Shopes. 2003. Commentary-sharing authority. *Oral History Review* 30, 1 (2003), 103–110.
[92] Linda Shopes. 2015. Community oral history: Where we have been, where we are going. *Oral History* 43, 1 (2015), 97–106.
[93] Gizem Sogancıoglu, Oxana Verkholyak, Heysem Kaya, Dmitrii Fedotov, Tobias Cadée, Albert Ali Salah, and Alexey Karpov. 2020. Is everything fine, Grandma? Acoustic and linguistic modeling for robust elderly speech emotion recognition. In *Proceedings of the 2020 INTERSPEECH Conference*.
[94] Michael Stadtschnitzer, Jochen Schwenninger, Daniel Stein, and Joachim Köhler. 2014. Exploiting the large-scale german broadcast corpus to boost the fraunhofer IAIS speech recognition system. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. 3887–3890.
[95] Giota Stratou, Stefan Scherer, Jonathan Gratch, and Louis-Philippe Morency. 2013. Automatic nonverbal behavior indicators of depression and PTSD: Exploring gender differences. In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, Los Alamitos, CA, 147–152.
[96] Penny Summerfield. 2004. Culture and composure: Creating narratives of the gendered self in oral history interviews. *Cultural and Social History* 1, 1 (2004), 65–93.
[97] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976* (2019).
[98] Alistair Thomson. 2007. Four paradigm transformations in oral history. *Oral History Review* 34, 1 (2007), 49–70.
[99] Alistair Thomson and Alistair Thomson. 1998. Fifty years on: An international perspective on oral history. *Journal of American History* 85, 2 (1998), 581. https://doi.org/10.2307/2567753
[100] Alexander Toshev and Christian Szegedy. 2014. DeepPose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1653–1660.
[101] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27, 12 (2009), 1743–1759.
[102] Elke Vlemincx, James L. Abelson, Paul M. Lehrer, Paul W. Davenport, Ilse Van Diest, and Omer Van den Bergh. 2013. Respiratory variability and sighing: A psychophysiological reset model. *Biological Psychology* 93, 1 (2013), 24–32.
[103] Iain Walker and Martin Halvey. 2017. On designing an oral history search system. *Journal of Documentation* 73, 6 (2017), 1281–1298.
[104] Megan Woods, Trena Paulus, David P. Atkins, and Rob Macklin. 2016. Advancing qualitative research using qualitative data analysis software (QDAS)? Reviewing potential versus practice in published studies using ATLAS.ti and NVivo, 1994–2013. *Social Science Computer Review* 34, 5 (2016), 597–617.
[105] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. 2012. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics* 31, 4 (2012), 1–8.
[106] Haoyu Zhang, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243* (2019).
[107] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2019. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055* (2019). http://arxiv.org/abs/1905.05055.

Received November 2020; revised May 2021; accepted July 2021