# Audiovisual media and automatic speech recognition

## *A methodological reflection*

EMIL STJERNHOLM*

## Abstract

The article maps the pitfalls and affordances of automatic speech recognition in the study of audiovisual media and digital history more broadly. Three key topics are emphasized: first, the collaboration between scholars and audiovisual archives and the potential of automatic speech recognition to transform the way scholars navigate and engage with audiovisual media history; second, what knowledge can be gained from employing automatic speech recognition in the exploration of film and television history; and lastly, thoughts on trajectories for future research. Even though the National Library of Sweden's audiovisual collection is frequently used in humanities and social science scholarship, automatic speech recognition algorithms has not been utilized to analyze material from their collection. As such, this methodological reflection aims to contribute toward methodological development.

*Keywords*: media history, digital history, automatic speech recognition, anslagstavlan, television history.

## Introduction

"Let me put it this way, Mr. Amer. The 9000 series is the most reliable computer ever made. No 9000 computer has ever made a mistake, or distorted information. We are all, by any practical definition of the word, foolproof and incapable of error." In Stanley Kubrick's classic film *2001: A Space Odyssey* (1968) – based on the science fiction novels by Arthur C. Clarke – the highly intelligent computer HAL 9000 is capable of understanding speech and reading lips, recognizing faces and appreciating art.

* Emil Stjernholm, PhD in film studies and associate senior lecturer in media and communication studies at the Department of Communication and Media, Lund University, emil.stjernholm@kom.lu.se




Lychnos 2022, DOI:10.48202/24014

The intelligent machine that understands human voices gained considerable traction in popular culture during the 1960s and 1970s – from the computer in the television series *Star Trek* to the likeable robots R2D2 and C3PO in George Lucas' *Star Wars*. Steve Wozniak, the co-founder of Apple, has credited *Star Trek* as a major inspiration in his work with the world's most famous technology company.[1] Already in 1988, Apple produced a visionary video for the higher education computing conference Educom, which presents a predecessor of Apple's Siri which recognizes gesture and voice controls. The video "The Knowledge Navigator", which is set in 2009, depicts how a professor at Berkley makes use of his voice to instruct the Knowledge Navigator tablet to help him organize his day, discuss new journal articles, and to organize his research data. What once was a futuristic vision has now become commonplace in everyday life. From navigation systems in cars to the smart home, from smartphones to automatic captions––speech recognition converts spoken language into digital text that then can be processed by computers. Automatic speech recognition is thus one type of automation integral to what has been labeled "digital society".[2]

In the past decades, breakthroughs in natural language processing have enabled computers to process and analyze large amounts of natural language data. While there have been significant advances in automatic speech recognition, this is a subfield of natural language processing where many challenges remain.[3] In the wake of mass digitization of cultural artefacts, the development regarding speech recognition has gained increased attention within the historical scholarly community and by cultural heritage institutions such as film and broadcasting archives. For example, in a 2018 report on the establishment of KBLab – a research infrastructure for digital humanities and social sciences – Pelle Snickars takes note of the ambition to make the vast quantities of audiovisual material deposited at the National Library computable and searchable, increasing the availability and usefulness of the audiovisual archive to scholars and students alike.[4]

This article provides a methodological reflection on the use of automatic speech recognition in my ongoing media historical research project Televising Information: Audiovisual Communication of Swedish Government Agencies. Notably, there are major differences between digital history projects carried out by single historians – such as the research project discussed in this methodological reflection – and projects with larger research teams comprising of both historians and computer scientists (or historians with dual competencies in history and computer science). As Mats Fridlund notes, there is a tension between digital history 1.0, which designates traditional historical scholarship increasingly impacted by the

everyday use of digital tools, and digital history 2.0, which posits computational methods and the increasing abundance of digital sources as profoundly transformative to the discipline at large.[5] A key example of the radically new and different digital history 2.0 is the increasingly common practice of "multidisciplinary teamwork" and the shifts such teamwork entails in terms of the scope of research, the scale of data, the process of data curation and the tools utilized. By contrast, my project rather follows what Fridlund labels as "digital history 1.5", that is "a hybrid or mixed methodology in that it is a combination of quantitative and qualitative historical research methodologies, and semi-automatic as it combines a large amount of manual evaluation with the systematic use of automatic analysis in pre-programmed offline and online calculation".[6] In other words, the digital methodology lies at the heart of the project discussed, yet it relies on the work carried out by a single historian using pre-programmed applications.

The article maps the pitfalls and affordances of automatic speech recognition in the study of audiovisual media and digital history more broadly. Three key topics are emphasized: first, the collaboration between scholars and audiovisual archives and the potential of automatic speech recognition to transform the way scholars navigate and engage with audiovisual media history; second, what knowledge can be gained from employing automatic speech recognition in the exploration of film and television history; and lastly, thoughts on trajectories for future research. Even though the National Library of Sweden's audiovisual collection is frequently used in humanities and social science scholarship, no previous research project has utilized automatic speech recognition algorithms to analyze material from their collection. As such, this reflection aims to contribute toward methodological development.

## Audiovisual media in digital history

With the rise of mass media during the 20th century, new patterns and processes of visual communication were established. In the canonical essay "The Work of Art in the Age of Mechanical Reproduction" (1936), Walter Benjamin theorizes about mass production – in particular the rise of photography and film – and its implications on art and popular culture.[7] Films, manufactured via reproduction, early on reached mass audiences. Soon, television came to transform people's everyday lives, influencing both attitudes and habits.[8] Today, audiovisual media flourish in the social media landscape, with platforms such as YouTube, Instagram and TikTok having great social and cultural impact.[9] Meanwhile, the heterogeneousness of audiovisual sources – ranging from fiction films and television

series to documentaries, from amateur film to news broadcasts – likely contributes to the underutilization of film and broadcasting archives by historical scholars and students alike. In fact, as Andreas Fickers argues, digitized audiovisual media provoke numerous problems regarding selection and contextualization; medium-specific issues which are not necessarily part of historians traditional training in source criticism.[10]

Much like in digital humanities in general, audiovisual media have a relatively peripheral position within the field of digital history. Textual sources, particularly from archives, have traditionally been more valued by historical scholars. As Hannu Salmi notes in his recent book *What is Digital History?,* this has contributed toward the establishment of a "hierarchy of textuality" where archival documents were ranked above private correspondence, where handwritten texts were ranked above popular publications, et cetera.[11] Within this hierarchy, also including maps, statistics, and oral histories, audiovisual sources have consistently ranked low.

Further, as Lev Manovich asserts in the introduction to his book *Cultural Analytics*, digital humanities is a text heavy field: "The larger portion of computational work in the humanities so far has focused on literary texts, historical text records, and spatial data. In contrast, other types of media, such as still and moving images, have received relatively little attention".[12] While digital methods for data-driven text analysis have been developed over the last decades, the use of digital tools in the analysis of audiovisual sources has only recently gained increased attention.[13] In recent years, Manovich has been one of the pioneers in working with digital methods to explore large image and video collections. In their Cultural Analytics Lab, Manovich and his team has been using advances in computer sciences to analyze thousands of images from Museum of Modern Art's (MoMa) collections, the films of the Russian filmmaker Dziga Vertov and millions of images and videos shared on Instagram.[14] Comparably, Taylor Arnold and Lauren Tilton have proposed the concept *distant viewing* – with reference to Moretti's concept of distant reading – to describe the process of analyzing large collections of (audio)visual material computationally. Meanwhile, other computational tools developed specifically for film analysis, such as Lignes de Temps, Cinemetrics and Shotlogger, show how aesthetic variables such as shot length or camera movement can be approached using digital methods.[15]

Today, there is unprecedented access to audiovisual media online – through non-commercial initiatives such as the Internet Archive's moving image database (www.archive.org/details/movies) and UbuWeb (www.ubuweb.tv), large-scale commercial platforms such as YouTube (www.youtube.com) and Vimeo (www.vimeo.com), and the commercial streaming giants. Meanwhile, only a fraction of the vast quantities of audiovisual

source material stored in film, radio and television archives is available for easy consumption. Whereas digital texts are easy to access, and there are established methodologies on how to make use of these sources in research (for example searching in digitized newspaper databases or text analysis using topic modelling), access to the material stored in film and broadcasting archives tends to be restricted due to intellectual property issues. Historians working with audiovisual sources tend either to use collections that are open and freely available (such as the Swedish Film Institute's initiative www.filmarkivet.se and the European broadcasters joint effort EUscreen http://www.EUscreen.eu) or alternatively spend time on-site at archives browsing large audiovisual data collections. Notably, the ways in which to navigate these archival collections, relying heavily on metadata supplied by archivists over the years, yields many challenges to untrained scholars. Moreover, as Roeland Ordelman and co-authors point out, manually created metadata "is typically sparse, quite diverse, and often incomplete",[16] a problem which often also applies to open collections such as Filmarkivet and EUscreen.[17] In this context, automatic speech recognition appears as a promising technology to facilitate further engagement with audiovisual media in historical scholarship, possibly serving to "bridge the gap between metadata sparsity and distant reading requirements of scholars".[18] In other words, automatic speech recognition of large audiovisual archives could allow scholars to identify patterns through "distant reading" and to enable new ways to search for and discover specific videos for close reading.

## Infrastructures

Today, non-commercial actors like Språkbanken Speech at KTH Royal Institute of Technology as well as commercial companies like Google and All Ears provide automatic speech recognition of the minor language Swedish, and make possible the extraction of speech-to-text data from different types of media. Whereas some automatic speech recognition models primarily have been trained on spoken language data (such as the one stored at Språkbanken), others have been trained on contemporary media such as radio, podcasts and social media videos (such as All Ears). Media historical material, meanwhile, poses particular challenges to automatic speech recognition. For example, sub-optimal sound quality, diffuse background noise and archaic language all impact the performance of the automatic speech recognition tool.[19] While many challenges remain in turning sound and vision into data, this is a prioritized area for many heritage institutions, among them the National Library of Sweden (Kungliga Biblioteket, KB).[20]

Recently, the National Library has been developing new automatic speech recognition models whose performance is optimized for media historical material. The Wav2Vec model, labelled "Wav2vec 2.0 large VoxRex Swedish", been trained on Swedish language radio material and common voice data.[21] For this model, the word error rate for common voice test set is 8.49%.[22] Much work likely remains in optimizing the model to deal with a variety of audio and audiovisual media, and currently the speech-to-text transcriptions generated from the model are not made available online for scholarly exploration (e.g. through The Swedish Media Database, SMDB).

Internationally, high ambitions have been voiced in this area. There are several multidisciplinary teamwork initiatives that aim to make broadcasting archives serve researchers needs with regard to the exploration of audiovisual collections. In this context, The Netherlands Institute for Sound and Vision's Media Suite, which serves as an online interface for an underlying media infrastructure that offers scholars the opportunity to explore, browse and compare content and metadata, is a particularly interesting example.[23] As Ordelman and co-authors note, the Media Suite allows scholars to create "personal virtual collections" and it offers a range of tools for working with these media collections including automatic annotation, visualization and analysis.[24] As a postdoc researcher at Utrecht University's Institute for Cultural Inquiry, I was able to take part in several workshops on the Media Suite, and explore the data and tools within this closed environment, which is only available to authenticated scholars. Notably, already in 2018, The Netherlands Institute for Sound and Vision's speech recognition service could operate faster than real time, and could process about 1000 hours per day.[25] Including a range of material such as radio, television, film, photography and newspapers, the Media Suite allows scholars to search across material categories. Meanwhile, time-coded transcripts are available for the audiovisual media that has been processed by the speech recognition service so far. Within the service, you can also browse available metadata and make annotations.

Another state-of-the-art project within this domain is the Connected Histories of the BBC project, which is a collaboration between the British public service broadcaster BBC and the University of Sussex. The BBC is a frontrunner in curating and making its collections available to a broader public, for example through the BBC Genome initiative which makes all program listing published in the *Radio Times* between 1923–2009 searchable (a publication comparable to the Swedish tv and radio-journal *Röster i Radio–TV*, 1930–1994).[26] Connected Histories of the BBC aims to further engagement with the BBC's vast oral history archive, making it searchable and interconnected.[27] For this purpose, the open source automatic speech

recognition tool Kaldi was developed in-house by the BBC.[28] Besides this, Named-entity recognition (NER) helped locate and classify named entities in this collection. Similarly, this task relied on a software developed by the BBC, Starfruit, which helped advance the accessibility of the speech-to-text transcriptions.[29] Similar to the Netherlands Sound and Vision's Media Suite, Connected Histories developed an interface, Macroscope, facilitating both close and distant reading. While no such advanced interface exists in the Swedish context, this international infrastructural development is noteworthy and provides an example of what the future might have in store.

## Working with automatic speech recognition

My current research project Televising Information sets out to study the narrative, aesthetic and rhetorical development of the influential public information television program *Anslagstavlan* over time. From the early 1970s until today, the public service broadcaster has circulated information on topics such as rights and responsibilities, public health and traffic security. I characterize this as a mixed methods project centering on two levels of analysis: textual analysis using digital methods and media production analysis drawing on archival research. In total, some five thousand *Anslagstavlan* programs have been transmitted, equivalent to approximately 160 hours of television, most of which has been preserved and deposited at the National Library. From this, a smaller sample of 600 programs has been used for this project. This methodological reflection builds on my experiences of collaborating with KBLab in turning this large sample of *Anslagstavlan*'s spoken messages into text.

The workflow comprised of four steps. First, the audiovisual dataset had to be manually curated on-site in Stockholm. In 1978, the legislation about the deposit of audiovisual material to the National Library changed, and from then onward the public service broadcaster and later also commercial radio and television companies have deposited material to the National Library. The files are often long and the metadata describing their structure is often missing or incorrect. To curate the sample, I had to identify and manually cut the relevant *Anslagstavlan*-segment (5 minutes) from a larger video block, containing 3 to 4 hours of video. In the second step, the National Library's computer scientists ran the sample through the Wav2Vec model, resulting in speech-to-text transcripts. In the third step, these transcripts were cleaned manually. As noted previously, the word error rate with this model is relatively low, however, a number of recurring problems were noted, such as the misinterpretation of noise as words, misprocessing of children's speech and people with thick

dialects, insufficient processing of background dialogue heard on top of the main narration, as well as the mislabeling of non-Swedish words or phrases. Lastly, having performed these three steps, the speech-to-text transcripts were available to explore using the corpus analysis tool AntConc.

All-in-all, the transformation of *Anslagstavlan* into text resulted in a mid-sized corpus comprised of approximately 180,000 tokens. For larger amounts of data, more advanced statistical models such as topic modelling would have been preferable. However, given the size of the data, the user-friendly application AntConc was chosen to explore factors such as keyness, concordances and keywords in context.[30] One of the most hands-on ways to use AntConc is to make comparisons between corpora. In this case, I compared different decades of public information to get a sense of how *Anslagstavlan* has transformed thematically over time. Keyness can be described as the frequency of a word in a smaller corpus when measured against the frequency of the word in a reference corpus. For my purposes, I used a smaller corpus (such as all text lists from the 1980s) against a reference corpus (comprising of all *Anslagstavlan*-transcripts). In doing so, I could arrive at a list of overrepresented words from each decade, that is, words that are statistically unexpected or unusual in relation to the whole corpus. While words such as "swimming" and "helmet", related to specific health and safety measures, were overrepresented in the 1970s, a word such as "discrimination" becomes overrepresented in the 2000s.

AntConc also makes it possible to explore keywords in context. By typing a manual search for a term in the search box, the concordance view shows every time the word appears in the corpus of television spots. Zooming in on keywords in context is a useful way to search for patters. For example, "must" (*måste*) is a modal verb used to indicate that it is important or necessary for something to happen. Which words appear near the word "must"? Having looked at keywords in context, you can then generate a list of words appearing most frequently in the company of the keywords, so-called collocates. Using a concordance plot, you can also trace how the usage of the word "must" has transformed over time. In this case, it should be noted, "must" was used more frequently during the 1970s and 1980s than during later decades, mirroring a transformation of the tone in government agencies communication.

Additionally, the speech-to-text transcripts helped provide an overview of potential topics for closer analysis. For example, during the 1970s, "energy" emerges as an overrepresented word in relation to the whole corpus. It is not surprising that the oil crisis of the 1970s and the early 1980s makes an imprint on televised public information. In fact, The Swedish Energy Conservation Committee commissioned numerous spots

for *Anslagstavlan* during this time, alongside a wide range of other information activities. However, in this context, the speech-to-text transcripts presents the scholar with an opportunity to search and discover time-coded segments on a particular topic (such as energy or the environment) from what originally was an unstructured and vast collection of video material with lacking metadata. These can then be explored and contextualized further through close reading.

Drawing on my experiences, there are some lessons for scholars who are interested in using automatic speech recognition to explore audio or audiovisual collections. *Anslagstavlan* relies heavily on voice-of-God narration and can thus be described as a "text-driven form of audiovisual culture", to quote Salmi who used this classification in relation to newsreels.[31] This likely had a positive impact on the quality of the automatic speech recognition. Working with other genres of audiovisual media – for example fiction films with fast-pace dialogue, background music and noise, or lots of switching between different languages – will likely result in a higher word error rate, and thus more time-consuming manual cleaning. This is something that is important to take into account when sketching a project's time plan and implementation. Moreover, as mentioned, working with film and broadcasting archives brings forth questions concerning intellectual property rights that can be hard to overcome. In this case, due to copyright reasons, the manual curation of the dataset had to take place on location at KB's department for audiovisual media. Additionally, following a discussion with KB's legal department, it was deemed that the speech-to-text transcripts were also protected by copyright, an interpretation that presents problems to scholars and research groups based outside of Stockholm. Given that this project is based at Lund University, this entailed a lot of travelling to work with both the curation and analysis of the data. Naturally, this also hampered the iterative process so integral to work with digital methods. However, I would argue that the optimal way for scholars to adapt to these circumstances would be set aside ample time for travelling and, if possible, schedule a prolonged visit to KBLab for the analysis stage.

## Trajectories for further research

In recent years, there has been a turn within the digital history toward the analysis of sound,[32] visual content such as photographs, paintings and maps,[33] and audiovisual material.[34] On the one hand, it might seem counterintuitive to call for the transformation of complex multimodal audiovisual media into simple text documents, thereby overlooking the importance of visual aesthetics, storytelling and mediation. On the other hand,

automatic speech recognition is merely one tool of many in the growing domain that Sander Münster and Melissa Terras call "the visual side of digital humanities".[35] Here, it is important to note that the multimodality of audiovisual data—including sound, vision and language—makes automated, computer-supported analysis more complex than working with text. For example, tools that enable object detection, facial recognition and image clustering are data-intensive and thus require much computing power as well as knowledge of Python. In this sense, lacking computational skills constitutes a difficult threshold for many digital historians interested in audio, visual and audiovisual media. In these circumstances, automatic speech recognition opens up for the exploration of audiovisual collections using established text mining tools, something which could speak to a larger research community.

As is evident in this special section of *Lychnos*, methods for large-scale computational text analysis have been utilized to study a wide range of text sources, from parliamentary data to publicly accessible newspaper repositories.[36] By contrast, audiovisual media culture has rarely been approached in this way. In the future, automatic speech recognition of large audiovisual collections could allow scholars to explore discursive changes in some of the most influential media during the 20th century such as film, radio and television. As noted above, automatic speech recognition models are well-disposed to handle text-driven forms of audiovisual culture. On this basis, newsreels, a type of short news documentaries with large cultural impact between the 1910s and the 1970s, stands out as a genre of film well-suited for this. After the breakthrough of sound film during the late 1920s and early 1930s, Salmi observes, newsreels were made with a silent camera that did not capture sound on location.[37] Instead, scripted voice-over narration was added at a later stage. Other film and television genres similarly relied on scripted narration, primarily in the so-called "useful cinema" domain, which describes the wide variety of non-fiction films that were used to instruct, inform or sell products to contemporary audiences.[38] In *The Arclight Guidebook to Media History and the Digital Humanities,* film scholars Charles R. Acland and Eric Hoyt describe these types of documentaries, sponsored films and instructional films as "the great unread" of media studies.[39] Similarly, historical radio and television programs such as *Dagens Eko* ("Echo of the Day", 1937–) and *Aktuellt* (1958–) built on presenters' narration until the widespread adoption of portable sound and camera technologies during the 1960s.[40] What themes have prevailed in these radio and television programmes? When do topics emerge and when do they disappear? With current automatic speech recognition models, text-driven forms are likely a good place to start for historians who want to explore discursive changes in audiovisual culture over time.

As media historians Asa Briggs and Peter Burke remind us in *A Social History of the Media*, "[t]o think in terms of a media system means emphasizing the division of labour between the different means of communication available in a given space and at a given time, without forgetting that old and new media can and do coexist and that different media may compete with or echo one another as well as complement one another."[41] The notion that media always relate to one another is crucial also to digital media history: increasing possibility to search and explore across material categories, including digitized newspapers, literature, photography or audiovisual media, promises to enrich historians investigations of various themes and discourses.

## Conclusions

Traditionally, most scholarly attention within digital history has been devoted to literary and historical texts. However, as noted, computers' ability to comprehend images, video, and audio has progressed significantly in recent years. Meanwhile, working with Swedish sources and historic content that may be of poor sound and image quality presents methodological challenges. This methodological reflection draws on my experiences as a single historian collaborating with KBLab in using automatic speech recognition to turn spoken messages from the influential television program *Anslagstavlan* into text. The reflection begins by discussing audiovisual media's position in what Salmi labels a "hierarchy of textuality" in historical scholarship, underlining that film and broadcasting archives pose particular challenges when it comes to digital source criticism, for example when contextualizing audiovisual media with sparse and incorrect metadata. Three main observations have been teased out: First, the current infrastructure for automatic speech recognition in relation to film and broadcasting archives is portrayed, outlining state-of-the-art examples of multidisciplinary teamwork projects in the Netherlands and the United Kingdom. Secondly, drawing on my own experiences, I discuss the pitfalls and affordances of these methods, providing examples of how speech-to-text transcripts can facilitate distant reading as well as an improved overview and searchability, something which supports close reading. The text ends with a reflection on possible avenues for further research, highlighting text-driven forms of audiovisual culture as a type of exciting data particularly well-suited for automatic speech recognition.

This methodological reflection scratches the surface of the potential role of automatic speech recognition and the text mining of audiovisual media culture in digital history. Notably, the most key challenges discussed here will be explored further within the multidisciplinary teamwork

project Modern Times 1936 (2022–2025), which I am also part of. Within this project, automatic speech recognition will be utilized to study expressions of modernity in a range of sonic and visual datasets from 1936, among them all surviving radio programs from Swedish Radio and all weekly newsreels and short films produced by Svensk Filmindustri. Here, an important ambition is to learn more about the possibilities and limitations of automatic speech recognition and various other machine listening techniques. As noted, the number of tools to explore "the visual side of the digital humanities" is increasing. While advanced infrastructure for the exploration of film and broadcasting archives remains rare – with The Netherland's Media Suite serving as a benchmark – these underutilized historical sources have much untapped potential. In the future, automatic speech recognition might thus not only lower the threshold for scholars who want to engage with audiovisual collections, but also assist historians in gaining new knowledge.

## Notes

1. Zaki Hasan: "Steve Wozniak on sci-fi, comic books, and how *Star Trek* shaped the future" in *Huffington Post*, 19 April 2017: https://www.huffpost.com/entry/interview-steve-wozniak-on-sci-fi-comic-books-and_b_58f7e86de4b081380af51897, accessed 22 March 2022.

2. Simon Lindgren: *Digital media and society* (Thousand Oaks, 2017).

3. Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al.: "Deep neural networks for acoustic modeling in speech recognition. The shared views of four research groups" in *IEEE Signal Processing Magazine,* 29:6 (2012): 82-97.

4. Ibid, 19

5. Mats Fridlund: "Digital history 1.5. A middle-way between normal and paradigmatic digital historical research", in *Digital histories. Emergent approaches within the new digital history*, (eds.) Mats Fridlund, Mila Oiva & Petri Paju (Helsinki, 2020), 69–87.

6. Ibid, 79–80.

7. Walter Benjamin: *The work of art in the age of mechanical reproduction* (London, 2008).

8. Lynn Spigel: *Make room for TV. Television and the family ideal in postwar America* (Chicago, 1992).

9. Pelle Snickars & Patrick Vonderau: *The YouTube reader* (Stockholm, 2009)

10. Andreas Fickers: "Towards a new digital historicism? Doing history in the age of abundance" in *VIEW journal of European television history and culture*, 1:1 (2012).

11. Hannu Salmi: *What is digital history?* (Cambridge, 2021), 43.

12. Lev Manovich: *Cultural analytics* (Cambridge, 2020), 25.

13. Andreas Fickers, Mark J. Williams & Pelle Snickars: "Audiovisual data in digital humanities" in *VIEW. journal of European television history and culture*, 7:14 (2018), 1–4.

14. Manovich 2020.

15. Olivier Fournout, Valérie Beaudouin & Estelle Ferrarese: "De l'utopie numéri-

que à la pratique. le cas de l'annotation collaborative de films" in *Communication & langages*, 180:2 (2014), 95–120; Jeremy Butler: "Statistical analysis of television style. What can numbers tell us about tv editing?" in *Cinema journal*, 54:1 (2014), 25–44; Yuri Tsivian: "Cinemetrics, part of the humanities' cyberinfrastructure" in Michael Ross, Manfred Grauer & Bernd Freisleben (ed.): *Digital tools in media studies: analysis and research* (Bielefeld, 2009), 93–101.

16. Roeland Ordelman, Liliana Melgar, Jasmijn Van Gorp & Julia Noordegraaf: "Media Suite. Unlocking audiovisual archives for mixed media scholarly research" in *Selected papers from the CLARIN annual conference 2018, Pisa, 8–10 October 2018*, no. 159 (2019), 134.

17. See Mats Jönsson & Pelle Snickars: "Filmens arkiv" in Mats Jönsson & Pelle Snickars (eds.): *"Skosmörja eller arkivdokument": Om filmarkivet.se och den digitala filmhistorien* (Stockholm, 2012), 1–33.

18. Ibid.

19. Sanna Dannert & Stefani Platakidou: "Performance of automatic speech recognition. A study investigating the performance of google speech-to-text on modern and historical Swedish audio material", Bachelor Thesis, Uppsala University Department of Statistics, 2020.

20. Pelle Snickars: *Datalabb på KB. En förstudie* (Stockholm, 2018).

21. https://huggingface.co/KBLab/wav2vec2-large-voxrex-swedish, accessed 24 March 2022.

22. On-going pilot experiments within the new research project Modern Times 1936 (http://modernatider1936.se/en/) conducted by Ester Lagerlöf show that the word error rate is higher for historical source material.

23. https://mediasuite.clariah.nl/about, accessed 24 March 2022.

24. Ordelman et al. 2018.

25. Ordelman et al. 2018, 136.

26. https://genome.ch.bbc.co.uk, accessed 29 March 2022.

27. Anna-Maria Sichani & David Hendy: "Connected histories of the BBC. Opening up the BBC oral history archive to the digital domain" in *Journal on computing and cultural heritage*, 15:1 (2021).

28. http: //kaldi-asr.org, accessed 29 March 2022.

29. Sichani & Hendy 2021, 13. See http://starfruit.virt.ch.bbc.co.uk, accessed 29 March 2022.

30. Heather Froehlich: "Corpus analysis with Antconc" in *Programming historian* (2015), https://doi.org/10.46430/phen0043, accessed 30 March 2022.

31. Salmi 2021, 65.

32. Mary Caton Lingold, Darren Mueller & Whitney Trettien (eds.): *Digital sound studies* (Durham, 2018).

33. Kevin Kee & Timothy Compeau (eds.): *Seeing the past with computers. Experiments with augmented reality and computer vision for history* (Ann Arbor, 2019).

34. Taylor Arnold, Lauren Tilton & Annie Berke: "Visual style in two network era sitcoms" in *Journal of cultural analytics*, 20 July 2019.

35. Sander Münster & Melissa Terras: "The visual side of digital humanities. A survey on topics, researchers, and epistemic cultures" in *Digital scholarship in the humanities*, 35:2 (2020).

36. See for example *Welfare state analytics. Text mining and modeling Swedish politics, media & culture, 1945–1989*, https://www.westac.se/en, accessed 30 March 2022; *Mining*

*for meaning. The dynamics of public discourse on migration,* https://www.statistik.uu.se/forskning/projekt/mining-for-meaning-the-dynamics-of-public-discourse-on-migration, accessed 30 March 2022.

37. Salmi 2021, 65–66.

38. Charles R. Acland & Haidee Wasson (eds.): *Useful cinema* (Durham, 2011).

39. Charles R. Acland & Eric Hoyt (eds.): *The arclight guidebook to media history and the digital humanities* (Sussex, 2016), 12.

40. Betsy McLane: *A new history of documentary film* (New York, 2012), 192.

41. Asa Briggs and Peter Burke: *A social history of the media. From Gutenberg to the internet* (Cambridge, 2002), 22–23.

## Acknowledgements

This work is conducted in collaboration with KBLab and funded by the Swedish Research Council (dnr 4.3-2019-06414).