C. Baume, M. D. Plumbley, D. Frohlich and J. Ćalić, "PaperClip: A Digital Pen Interface for Semantic Speech Editing in Radio Production,"
J. Audio Eng. Soc., vol. 66, no. 4, pp. 241–252, (2018 April.).
DOI: https://doi.org/10.17743/jaes.2018.0006

PaperClip: A Digital Pen Interface for Semantic Speech Editing in Radio Production

CHRIS BAUME,[1, 2] *AES Associate Member*, **MARK D. PLUMBLEY,**[2] *AES Member*,
(chris.baume@bbc.co.uk) (m.plumbley@surrey.ac.uk)

DAVID FROHLICH[3], **AND JANKO ĆALIĆ**[1]
(d.frohlich@surrey.ac.uk) (janko.calic@bbc.co.uk)

[1]BBC Research and Development, London, UK
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK
[3]Digital World Research Centre, University of Surrey, Guildford, UK

We introduce "PaperClip"—a novel digital pen interface for semantic editing of speech recordings for radio production. We explain how we designed and developed our system, then present the results of a contextual qualitative user study of eight professional radio producers that compared editing using PaperClip to a screen-based interface and normal paper. As in many other paper-versus-screen studies, we found no overall preferences but rather advantages and disadvantages of both in different contexts. We discuss these relative benefits and make recommendations for future development.

0 INTRODUCTION

The radio production workflow typically involves recording material, selecting which parts of that material to use, then editing the desired material down to the final output [1]. Many producers will write transcripts of their recordings, either themselves or using a third-party service, to help them recall what was said and when, identify themes, and make links between different parts of their content. In our previous study [2], we found that some radio producers we tested found this process easier to achieve on paper than directly on the screen, so choose to print the transcript. Reading from paper rather than a screen has been found to improve comprehension [3], recollection [4], sense of structure and cross-referencing [5], and to be faster [6]. Radio producers can use paper to make hand-written annotations to help them structure their program and make editorial decisions. However, after they have decided which parts of the audio they want to use in their program, they must use a digital audio workstation (DAW) to manually execute those editorial decisions, which is a tedious and slow process.

In this paper we describe the design, development, and evaluation of *PaperClip*—a novel system for editing speech recordings directly on a printed transcript using a digital pen. In Sec. 1 we review previous approaches to semantic speech editing and natural annotation of digital content. In Sec. 2 we describe our first study in which we worked with radio producers to design the layout of our system. In Sec. 3 we describe the design of PaperClip, which we developed in collaboration with a digital pen manufacturer. In Sec. 4 we explain the methodology of our second study in which radio producers edited content for their programs using PaperClip, a screen interface and a normal printed transcript. We present the results in Sec. 5 which compares the strengths of the digital pen and screen interfaces, and shows how the accuracy of the transcript and listening affect the editing process. We discuss these results in Sec. 6, present our conclusions in Sec. 7, and propose future work in Sec. 8.

1 BACKGROUND

Our system combines semantic editing of speech with natural annotation of digital content. Previous semantic speech editing systems [7–13, 2] have all used screen-based interfaces. We identified three alternative types of interfaces that could be used to edit digital content, which were based on barcodes [14–17], digital pens [18–21], and digital ink [22–25]. We explore each of these approaches and their applications below.

A number of screen-based systems have previously been developed to explore the benefits of semantic speech editing. *SCANMail* [7] demonstrated the advantages of navigating voicemail recordings using a transcript, but did not include editing capabilities. The *LIDS Editor* [8], and later *TRAED* [9], used automatically-generated transcripts to allow users to navigate and edit lecture recordings by removing and rearranging sentences and words. Rubin [10] created a system for creating audio stories using perfect crowd-sourced transcripts. Similar techniques have been applied to video editing. *SILVER* [11] was a video editor that had an editable transcript window, generated from subtitles, and Berthouzoz et al. [12] developed a system that used crowd-sourced transcripts and image processing to allow text-based editing of multi-camera video interviews. Even though automatically-generated transcripts are imperfect, Whittaker and Amento [13] found they are sufficiently accurate to allow navigation and editing. This was supported by our previous study [2] in which radio producers used our screen interface to edit programs using automated transcripts.

Barcodes printed on paper transcripts have been explored as a method of navigating video recordings by using a device to scan the barcode and play the video from that position. *Video Paper* [14] was a system that embedded video keyframes with barcodes down the side of the page. Each barcode linked to a position in a video, which was downloaded and played on the scanning device. *Books with Voices* [15] was a similar system that tested this approach with oral historians who found it effective for assisting a transcript editing task. Erol et al. [16] went a step further by embedding the video data in the barcode, removing the need for a server. *HotPaper* [17] removed the need for barcodes by using a camera to measure the whitespace between words and matching that to unique patterns in the text.

The “Anoto dot pattern” is a unique non-repeating marking printed onto normal paper, which allows a digital pen to use an on-board camera to track and record its position. This technology can be used as a way to “bridge the gap” between paper and digital documents. *ChronoVis* [18] was a note-taking system that used the Anoto pattern for recording synchronized hand-written notes during playback of a video. An accompanying screen interface allowed users to click on the digital display of the handwritten notes to navigate to that position in the video. *PADD* [19] was a concept for a system of editing documents that allowed users to move from digital to paper and back again. *ProofRite* [20] was an implementation of this, which used the Anoto pattern to link annotations made on paper into a word processor such that they “reflow” with the text they were attached to. *PaperProof* [21] improved on this by interpreting the edit annotations and automatically applying them to the document.

“Digital ink” interfaces capture natural annotations on a screen interface, typically using a tablet PC and stylus. This approach has been explored as a method for annotating and editing video content. *Marquee* [22] synchronized handwritten notes with a live video recording by using a horizontal line gesture to mark a timestamp. *Videotater* [23] was another digital ink interface for segmenting and annotating pre-recorded video clips. A vertical line gesture on a video timeline split the video, and handwritten words could be written over a clip. *WaCTool* [24] extended this functionality by associating user interactions with edit commands. For instance, users could assign a “skip” command by pressing buttons at the start and end of an unwanted region. *Video as Ink* [25] allows users to “paint” video frames onto the tablet interface and then edit the video by using an “eraser mode” to remove unwanted frames. However, these previous approaches to video editing [23–25] have relied on the manipulation of video thumbnails, which cannot be used for radio production.

We have seen that barcodes, digital pens, and digital ink can be used to link natural annotations to digital content. Digital ink interfaces have been successfully applied to editing video content, but because they use screens, they do not benefit from the advantages of reading from paper. Barcodes and digital pens allow users to navigate media while reading and annotating a paper transcript. However, with barcodes the paper annotations are not captured, so would have to be typed into a device. Digital pens can capture handwritten annotations in a digital format, but they have yet to be applied to editing media.

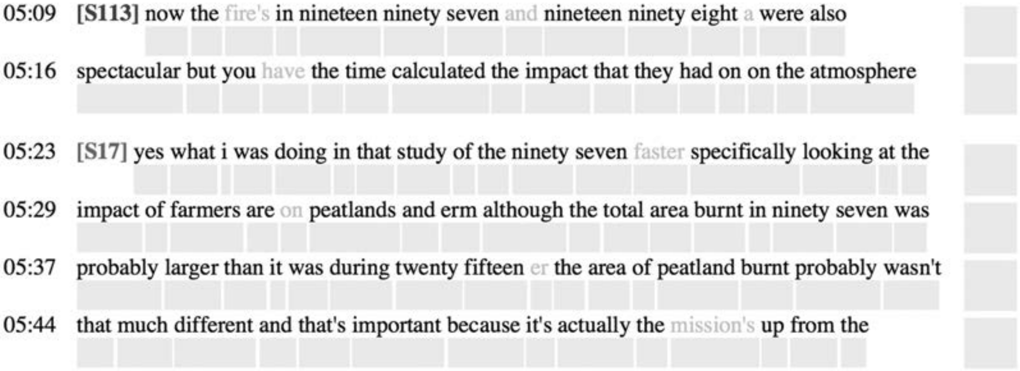
2 SYSTEM REQUIREMENTS

We developed a paper-based semantic speech editor for radio producers, to explore how it affects the production process. We chose to use digital pen technology because it uses paper, which provides better readability, and can capture natural handwritten annotations. Due to the lack of open development platforms, we collaborated with the digital pen manufacturer Anoto to build our system. We used their *Live*^{TM}*Forms* platform, which allowed us to capture digital information from handwritten annotations. The system works by dividing a page into rectangular active zones. When a compatible digital pen draws inside one of these zones, that data is captured digitally and processed.

As there were no previous paper-based media editing systems on which to base the design of our system, we worked with radio producers to evaluate a paper mock-up of the paper interface. The prototype used a normal pen, but otherwise gave an identical experience to that of a digital pen. We were interested in answering three questions: How do producers currently annotate transcripts? Do they prefer to select or remove content? Which additional features (e.g., timestamps, speaker labelling, confidence shading) should be included with the transcript?

2.1 Mock-Up Design

In our previous study [2], we saw that radio producers annotated paper transcripts using underline (for selecting words), strikethrough (for removing words), and drawing a line down the side of the page (for selecting whole lines). We used this information as the basis for the design of our mock-up system, shown in Fig. 1. We used a speech-to-text system to generate the transcript and included the additional information it provided. We wrote a timestamp at the beginning of each line in *minute:second* format, and used confidence shading [26] to “low-light” words with a low confidence score by shading them grey. We put a paragraph break at speaker boundaries and wrote the speaker label at the start of each paragraph. To distinguish speaker gender, we colored the speaker label blue for males and red



05:09 [S113] now the fire's in nineteen ninety seven and nineteen ninety eight a were also

05:16 spectacular but you have the time calculated the impact that they had on on the atmosphere

05:23 [S17] yes what i was doing in that study of the ninety seven faster specifically looking at the

05:29 impact of farmers are on peatlands and erm although the total area burnt in ninety seven was

05:37 probably larger than it was during twenty fifteen er the area of peatland burnt probably wasn't

05:44 that much different and that's important because it's actually the mission's up from the

Fig. 1. Design of the paper prototype

for females. To be able to capture timed edit commands using the *Live*$^{\text{TM}}$*Forms* system, we designed our layout to use rectangular active zones that aligned with the location of each word. We placed an invisible active zone over each word to capture strikethrough, a shaded active zone under each word to capture underline, and a square shaded active zone at the end of each line to capture lines down the side.

2.2 Mock-Up Evaluation Method

To evaluate our proposed layout, we recruited five radio producers (P1–P5) from BBC Radio to use our inactive prototype to annotate real transcripts as if they were editing them. Two of the participants worked in current affairs, two in science, and one in documentaries. The participants had between 7 and 13 years experience in working as a radio producer. Producers are very busy, so to recruit enough participants in the time available, we designed the experiment to take less than one hour. To make the study as realistic as possible, we asked each participant to provide a recent interview recording and used our speech-to-text system to generate their transcript. We directed each participant to employ different strategies when editing each page of the transcript, so they would be forced to try and compare each method.

- *Page 1*: **Undirected** – Edit the speech by annotating the transcript as you would normally.
- *Page 2*: **Underline only** – Edit the speech only by underlining words that you want to keep.
- *Page 3*: **Strikethrough only** – Edit the speech only by putting a line through words you don't want to keep.

To evaluate speaker labelling, we excluded the labels from the first three pages, then included them on *Page 4* and asked the participant to edit the speech how they wished. Timestamps, line selection, and confidence shading were included with all of the prototypes as we expected participants to be able to judge their value in situ.

After the editing task, we conducted a semi-structured interview with each participant. We asked them how they normally edit paper transcripts, whether they prefer to select or remove content, which features they found useful, and whether there were any missing features.

Table 1. Natural edit gestures used by each participant.

	P1	P2	P3	P4	P5	Count
Underline	●	●	●	●		4
Strikethrough	●	●		●	●	4
Line down side	●	●	●		●	4
Comments	●	●			●	3
Corrections	●				●	2
In/out marks	●			●		2
Scribble-out mistake		●	●			2
Lasso					●	1
Line through paragraph					●	1

2.3 Mock-Up Evaluation Results

Table 1 lists the gestures that the participants used when editing undirected on pages 1 and 4. Each participant naturally used a different mixture of gestures for selection, removal, correction, and labelling. The most common gestures for selection were underline and line down side, with strikethrough being the most common removal gesture. Most participants combined line down side for large selections with underline and strikethrough for finer edits.

We asked each participant whether they preferred selecting or removing words when editing the transcript. P1, P3, and P4 reported that they preferred selecting, with P2 and P5 preferring to remove words. P1 commented that selecting *"felt more natural"* to them, but P5 said they prefer to *"get stuff out of the way."* All of the participants were certain about which they preferred, but there was no overall consensus. Additionally, Table 1 shows that most participants used a mixture of select and delete gestures during the undirected stage.

Four of the five participants said that they found the speaker labelling useful, and three of the participants used it to identify where the presenter asked questions. However, P2 said they found it *"distracting"* due to its inaccuracy. All participants said they found the timestamps and confidence shading features useful, but P2 said that the timestamps are *"not needed on every line"* and P5 suggested that one timestamp per page would be sufficient. All of the participants liked being able to select whole lines at a time. P5, who prefers to remove words, asked whether a similar function could be available to delete content.

P3, P4, and P5 remarked that they often highlight important bits of transcripts, usually with asterisks or stars. P1 and P3 also suggested extending the underline gesture so that underlining twice marked words as being more important. Three participants used what little space there was at the side to label the content and make notes for themselves, and P1 and P5 corrected words in the transcript by writing over or above the incorrect word.

2.4 Discussion

The prototype evaluation confirmed our assumptions about underline, strikethrough, and line down side being the most common edit gestures. The alternative gestures were used less than half as often. Most participants valued the additional features we tested—speaker labelling, timestamps, and confidence shading—but reported that timestamps on every line are unnecessarily frequent. There were mixed but strong opinions on whether participants preferred to select or remove content, and most used a mixture of both. We also identified missing functionality for labelling, correction, and highlighting.

Both selection and removal should be made available. Providing an inactive margin would allow users to write labels without inadvertently editing by writing over active zones. The double-spaced text allows enough space for corrections, but the system would need to distinguish between handwriting and edit gestures. Underlining twice would allow users to highlight words, but if all words were exported by default, underlining once could be used for highlighting.

3 SYSTEM DESIGN

Previous semantic speech editing systems have used screen interfaces. To be able to evaluate the effect of paper-based semantic editing on radio production, we implemented both a digital pen interface and a screen interface. This section describes the design and implementation of these systems.

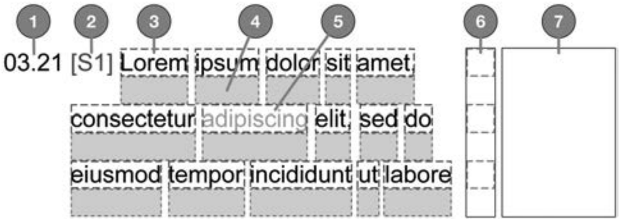
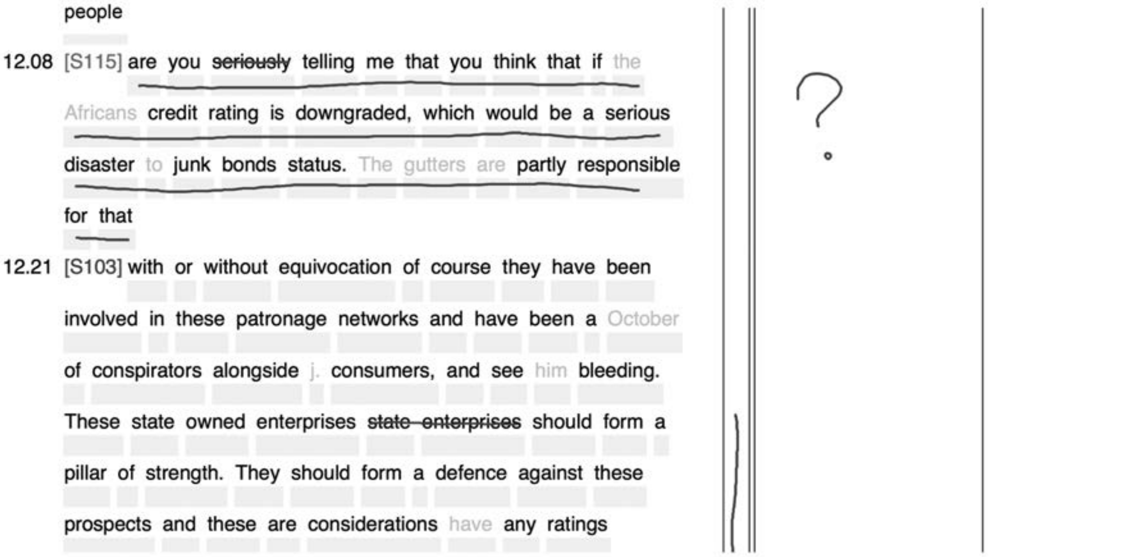

1
2
3
4
5
6
7
03.21 [S1] Lorem ipsum dolor sit amet,
consectetur adipiscing elit, sed do
eiusmod tempor incididunt ut labore

Fig. 2. Layout of the PaperClip interface, which features timestamps at beginning of each paragraph (1), speaker labelling (2), word removal (3), word selection (4), confidence shading (5), line selection (6), and a margin for freehand notes (7). Dotted lines indicate hidden active zones for selection and removal.

3.1 PaperClip

The design for our digital pen interface was informed by the results of our first study in Sec. 2. Based on our findings, we used underline, strikethrough, and line down side as the edit gestures and included speaker labelling and confidence shading. We kept the timestamps, but reduced the frequency to one per paragraph, and included an inactive margin to allow users to make unstructured notes. We collaborated with Anoto to implement PaperClip using their *Live*$^{\text{TM}}$*Forms* platform. As this platform did not allow us to distinguish between lines and handwriting, we could not include any correction functionality. We used two active zones for each word—one on the word to detect a strikethrough and one below the word to detect underline. We drew a long thin rectangle between the transcript and the margin for capturing the line down the side. The final design is shown in Figs. 2 and 3.

Editing was performed using a digital pen, which tracked and digitally recorded the gestures made on the transcript. When the pen was connected to a computer via a USB dock,



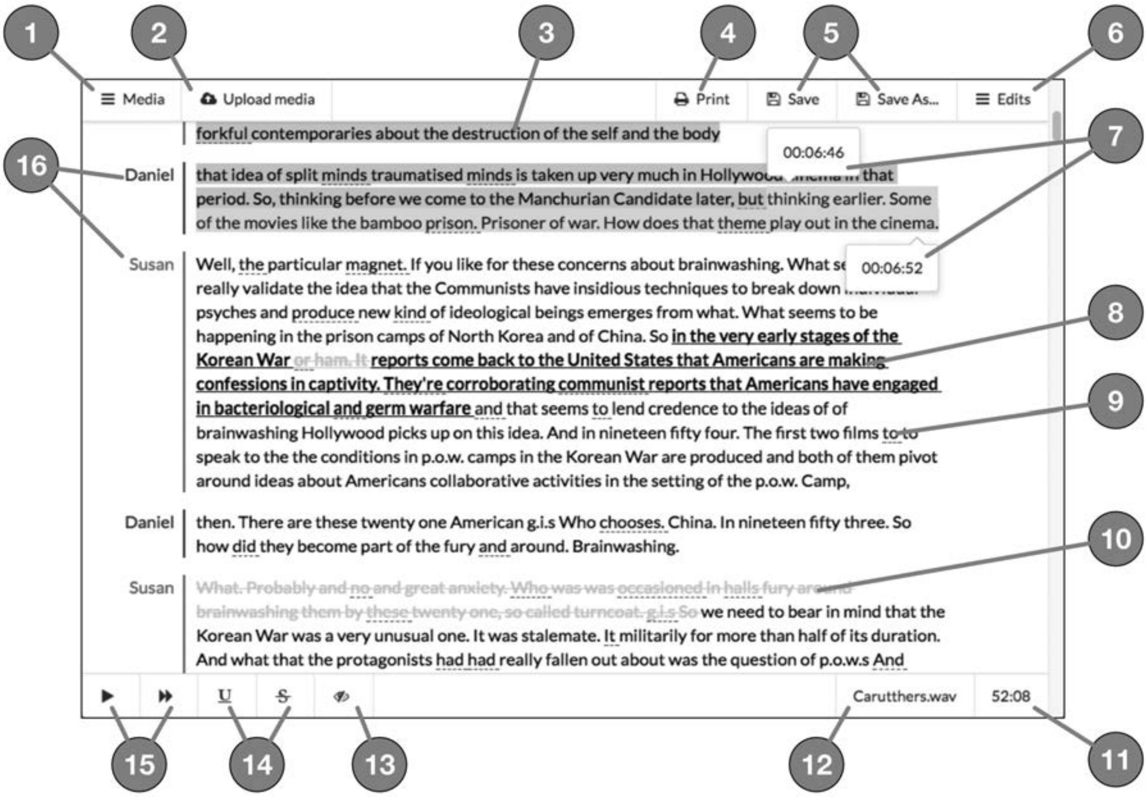
people

12.08 [S115] are you ~~seriously~~ telling me that you think that if the
Africans credit rating is downgraded, which would be a serious
disaster to junk bonds status. The gutters are partly responsible
for that

12.21 [S103] with or without equivocation of course they have been
involved in these patronage networks and have been a October
of conspirators alongside j. consumers, and see him bleeding.
These state owned enterprises ~~state enterprises~~ should form a
pillar of strength. They should form a defence against these
prospects and these are considerations have any ratings

?

Fig. 3. Example of the PaperClip interface, with gestures that demonstrate its use.



forkful contemporaries about the destruction of the self and the body

Daniel that idea of split minds traumatised minds is taken up very much in Hollywoo[illegible] that period. So, thinking before we come to the Manchurian Candidate later, but thinking earlier. Some of the movies like the bamboo prison. Prisoner of war. How does that theme play out in the cinema.

Susan Well, the particular magnet. If you like for these concerns about brainwashing. What se[illegible] really validate the idea that the Communists have insidious techniques to break down [illegible] psyches and produce new kind of ideological beings emerges from what. What seems to be happening in the prison camps of North Korea and of China. So in the very early stages of the Korean War ~~or ham. It~~ reports come back to the United States that Americans are making confessions in captivity. They're corroborating communist reports that Americans have engaged in bacteriological and germ warfare and that seems to lend credence to the ideas of of brainwashing Hollywood picks up on this idea. And in nineteen fifty four. The first two films to ~~to~~ speak to the the conditions in p.o.w. camps in the Korean War are produced and both of them pivot around ideas about Americans collaborative activities in the setting of the p.o.w. Camp,

Daniel then. There are these twenty one American g.i.s Who chooses. China. In nineteen fifty three. So how did they become part of the fury and around. Brainwashing.

Susan ~~What. Probably and no and great anxiety. Who was was occasioned in halls fury around brainwashing them by these twenty one, so called turncoat. g.i.s So~~ we need to bear in mind that the Korean War was a very unusual one. It was stalemate. It militarily for more than half of its duration. And what that the protagonists had had really fallen out about was the question of p.o.w.s And

Fig. 4. Layout of the screen interface, which features media storage (1), media upload (2), highlight of the current playback position (3), printing the transcript (4), saving edits and corrections to transcript (5), edit storage and export (6), displaying timestamps of the current selection (7), underlining words (8), confidence shading (9), strikethrough of words (10), display of edited audio duration (11), name of current asset (12), show/hide words with strikethrough (13), underline/strike buttons (14), playback buttons (15), and speaker labelling (16).

the gestures were processed and translated into edit commands. We integrated PaperClip with our screen interface (see Sec. 3.2) to handle audio import, printing transcripts, viewing/changing edits, viewing the margin notes, and exporting the edits. We supported two export formats— audio as a .wav, or an edit decision list (EDL) for the DAWs used for radio production the BBC.

3.2 Screen Interface

For the screen interface, we used the system from our previous study [2], which we updated to reflect user feedback. The original design used a drag-and-drop system for creating clips from selected text. We replaced this with underline and strikethrough gestures to provide better support for large selections, and to align with the design of PaperClip. We also added a double-speed playback feature to allow faster than real-time listening, and a “save-as” feature to allow multiple edits of the same material. Fig. 4 shows the layout of the screen interface and lists its features.

The screen interface included integrated playback, which allowed the user to listen to and navigate the audio while they edit. The current playback position was shown in the text and the user could jump to a word by double-clicking it on the transcript. Any edits made to the transcript were reflected in the audio. The user could also correct any mistakes in the transcript by editing the text as they would in a word processor.

4 EVALUATION METHODOLOGY

The objective of our second study was to discover whether professional radio producers could use PaperClip as part of their workflow and to compare how the workflow was affected by PaperClip and our screen interface. To find out, we ran a within-subjects qualitative user study in which we tested radio producers editing speech recordings under three different conditions:

C1. PaperClip digital pen interface;
C2. Screen interface;
C3. Normal printed transcript.

The normal printed transcript included speaker labels and timestamps but did not use the PaperClip layout or Anoto dot pattern. The transcripts for all three conditions were generated by a speech-to-text system developed by the BBC, which used the Kaldi toolkit[1] and was trained on television recordings.

[1] http://kaldi-asr.org/

Table 2. Evaluation study participants.

ID	Experience	Department	Computer literacy
P1	13 years	Current affairs	Medium
P2	16 years	Documentaries	Low
P3	8 years	Current affairs	High
P4	10 years	Science	High
P5	18 years	Current affairs	Low
P6	16 years	Current affairs	Medium
P7	28 years	Documentaries	Medium
P8	20 years	Science	Low

We recruited eight radio producers from the current affairs, science, and documentaries teams in BBC Radio. Table 2 lists the participants and their self-reported professional experience, department in which they work, and computer literacy, as rated by the investigator based on their observations. Only one of the participants overlapped with our first study in Sec. 2. As producers are very busy, we designed our study to take less than a day to complete. Despite this, it took us 12 months to recruit the participants and collect the data as producers often cancelled or re-arranged due to their demanding role.

4.1 Protocol

The protocol for our study had three stages:

- *Stage 1*: **Training** – The participant performed a scripted series of tasks that used all of the features of each interface and was given an opportunity to use the interfaces until they were comfortable.
- *Stage 2*: **Task observation** – The participant provided three recent speech recordings that they needed to edit. Our previous study [2] found that there was no benefit in using transcripts for short recordings, so each recording was at least 20 minutes in length. The participant edited each recording under one of the three conditions (C1, C2 or C3) at their desk. The order of conditions was counterbalanced. The investigator observed the task, made written notes about their behavior, and logged the duration of each audio file and the time taken to edit it, excluding any interruptions. After each task, the participant filled out a questionnaire to measure the usefulness and usability of the interface, using Perceived Usefulness [27] and the Software Usability Scale (SUS) [28], respectively. After completing all three tasks, the participant was asked to select which system they would prefer to continue using.
- *Stage 3*: **Interview** – The investigator conducted a semi-structured interview that asked:
  1) Can you please describe your existing process for editing audio?
  2) What did you like and dislike about using the digital pen system?
  3) What did you like and dislike about using the screen system?
  4) What did you like and dislike about using normal paper?
  5) Overall, which of these systems would you most prefer to continue using, and why?

The order of questions 2–4 was adjusted to match the order in which the conditions were presented to the participant. An audio recording was made of the interview for later analysis.

4.2 Analysis

We transcribed the interview recordings and corrected the words manually using the screen interface described in Sec. 3.2. Using grounded theory [29], the investigator then openly coded the transcripts and observation notes using *RQDA* [30], which produced 229 initial codes. The investigator then used *FreeMind* mind-mapping software to group the codes into categories, and the categories into themes.

As the time taken to edit an audio file depends upon its length, we divided the edit speed of each task by the audio file duration to calculate the "relative edit time." We used the procedures in [27, 28] to convert the perceived usefulness and SUS ratings into percentage scores. Within-subjects one-way ANOVA [31] was used to test for differences between the systems in the relative edit time, perceived usefulness, and usability (SUS) metrics.

5 EVALUATION RESULTS

5.1 Metrics

When asked which system they would prefer to continue using, four of the eight participants chose PaperClip, two (P3 and P6) chose the screen interface and two (P1 and P4) chose the normal paper transcript. Although it did not include any semantic editing functionality, P1 and P4 said they preferred the normal paper transcript as it allowed them to use their existing workflow and tools, which they found easiest and most comfortable.

For the SUS metric, the mean ratings for PaperClip, screen interface, and normal paper were 73%, 75%, and 82%, and for perceived usefulness they were 75%, 78%, and 85%, respectively. However, there were only eight participants, and a one-way within-subjects ANOVA found that there was no statistically significant difference between the systems for usefulness [$F(2, 14) = 0.788$, $p > 0.05$], nor usability [$F(2, 14) = 1.068$, $p > 0.05$].

For each task we divided the edit time by the audio duration to calculate the relative edit time. The screen and normal paper interfaces had the same mean relative edit time (x0.99 real-time), but PaperClip was 16% faster (x0.83 real-time). However, a one-way within-subjects ANOVA did not find any statistically significant difference [$F(2, 14) = 0.931$, $p > 0.05$].

The metrics results show that although half of participants preferred the PaperClip interface and it had the fastest relative edit time, it was rated least useful and least usable.

Table 3. Summary of comments by participants.

Editing	P1	P2	P3	P4	P5	P6	P7	P8
Decision-making				●	●			●
User-friendliness					●	●		●
Information processing		●						●
Strict boundaries			●		●		●	
Undo			●			●		
Collaboration			●			●	●	
Travel							●	●
Comfort					●		●	
Edit iterations	●					●		●

Transcript	P1	P2	P3	P4	P5	P6	P7	P8
Paper easier to read		●			●			●
Paper easier on the eye	●	●				●	●	
Tangibility of paper		●			●		●	
Orientation	●				●			
Accuracy						●	●	●
Distraction of errors		●		●		●		
Custom training			●				●	
Transcript is largest benefit	●		●					

Listening	P1	P2	P3	P4	P5	P6	P7	P8
Information processing	●			●		●		
Error identification	●					●		
Comprehension		●						●
Navigation		●		●	●		●	
De-umming							●	●
Audio-only editing				●		●	●	●

5.2 Thematic Codings

To better understand the ratings in Sec. 5.1, we now turn to the interview and observational data. Table 3 summarizes the comments made by participants in the interviews and during observation. We have grouped the comments into three themes, which we identified during the analysis. We present the results from each theme below:

5.2.1. Editing

Participants P4, P5, and P8 reported that they could make editorial decisions faster and more easily on paper compared to the screen because of the reduced functionality of the interface, uninterrupted playback of the audio, natural edit gestures, and faster reading speed. P4 said that the lack of correction features in PaperClip allowed them to edit faster than the screen, as it didn't interrupt their flow.

"With the pen, I couldn't [correct] the transcript so there's no point stopping. [. . .] I don't think I've ever done an edit that fast, where it was literally real time." (P4)

P5, P6, and P8 felt that the physicality of the PaperClip interface made it user friendly, intuitive, and simple.

"It feels like you're working analogue, but you're actually working digitally. [. . .] It's nice to hold a pen and go on real paper, which has the feel of every day life." (P7)

P8 said they felt that the digital pen allowed them to be more precise with their edits than with the screen. Although the screen is just as precise, the digital pen can be used to start making a selection without knowing the endpoint, which may give a feeling of better control over precision.

P2 and P5 reported that they could process the information faster when reading on paper compared to the screen. P5 said that when using the screen, they would select more than necessary because their decision-making couldn't keep up with the audio.

"The [screen] just felt too quick and much much harder to make a decision. It was like 'just keep everything,' because you don't want to miss something." (P5)

The design of PaperClip forced users to select or delete content by drawing lines within strictly defined zones that are interpreted literally. P3, P5, P7, and P8 said they did not like that they could not freely draw on the page and were concerned about potential errors that could be introduced by straying outside of the boundaries.

"[PaperClip] doesn't have the convenience of paper, which is that there's no real rules [and] you can write anywhere on the paper." (P3)

P3 and P6 said that they did not like that there wasn't any way to undo the edits using PaperClip. P6 suggested that the lack of undo functionality may force them to be more decisive.

"It's harder to say 'oh no I've changed my mind, I want to go back,' so you almost have to be much more decisive, which maybe is a good discipline." (P6)

Often transcripts can be very long, so printing them requires a large amount of paper. P2 used quite a long recording for the experiment, which required over 50 sheets of paper. The Anoto system also requires access to a color laser printer. This is not usually a problem in an office environment, but can be an issue when travelling, or when working from home.

Radio producers work with a variety of people including presenters, assistant producers, contributors, and organizations. P3, P6, and P7 said that transcripts make it easier to collaborate as they create a common reference point that is easy to share and annotate.

"The way we're doing it is printing out our transcripts and we can all go 'page 15' [. . .] there's a common reference, whereas if you're just doing audio it's harder." (P6)

The physical nature of paper allows people in the same room to hand around transcripts, point at words, and lay pages out. However, the digital nature of the screen means it can be used for remote collaboration. For example, P6 reported that they use Google Docs to simultaneously write and edit the script remotely with the presenter.

P1, P5, P7, and P8 said that they often prefer to work away from the office, such as at home, to help them focus and get more work done. P7 and P8 suggested that PaperClip was well-suited for travel, such as during commuting, which may provide an additional opportunity to be productive in what would otherwise be considered downtime. Although, P7 pointed out that the screen interface could be used on-the-road with a laptop and noise-cancelling headphones.

"With the pen you could do stuff on the train [. . .] or on a bus. You could do it anywhere as long as it's not too bumpy." (P8)

P5 said they did not enjoy spending too long sitting upright at their desk, and P7 cited comfort as a factor in where they prefer to work.

"I would feel more comfortable with a nice digital pen and a sheet of paper sitting on a couch [. . .] You could do it in bed - that would really have your work-life balance sorted, wouldn't it?" (P7)

P1, P2, P6, and P8 reported that editing was an iterative process. P2 said this was because they are not sure what they need in the early stages, so they select too much then reduce it later. P8 said that what they select, or how much they select, depends on what was said in other interviews; and P1 said they often have to go back to re-edit clips in a different way.

P1, P6, and P8 reported that all three systems we tested were only suitable for the first iteration, known as a "rough edit," because they were missing two features—re-ordering and labelling. Re-ordering is used to to see and hear how different clips from separate interviews would work together, and labelling is used to help the producer navigate, organize, and structure their content. P5 used PaperClip's margin to write labels and mark questions, but these were not digitized or made available in the exported edit.

"I was just labelling by summarizing a paragraph in about two or three words – just who is speaking and the substance of it – or maybe just putting a cue to say that was a question." (P5)

P3 suggested that it might be possible to automatically generate labels using the text of a selected clip.

5.2.2 Transcript

Most participants commented that working with paper had a number of benefits to their workflow. P2, P5, and P8 said they found it easier to read from paper than screen. P1, P2, P6, and P7 said that it was easier on the eye and gave them a break from working on screen. P2, P5, and P7 said they enjoyed that paper was a physical, tangible medium that they could touch. P1 and P5 commented that using paper transcripts made it easier for them to orientate themselves. P1 said the paper interface allowed them to think more widely, and P8 reported that they found it easier to remember the content of the transcript when reading on paper rather than a screen.

"I find it easier to read off paper, and easier to remember stuff." (P8)

All of the participants were successfully able to use the automatically-generated transcript to edit their material as part of the production of their radio program, and all reported that the transcripts were sufficiently accurate for the purpose of editing their content. Similarly to our previous study [2], the most common complaints were of reduced accuracy due to heavy accents or background noise, and problems with speaker labelling and confidence shading. For example, the speech-to-text system would occasionally give a high confidence score to an incorrect word, or vice-versa, which caused P3 to mistrust the confidence shading.

"The things it wasn't sure about weren't actually very often the real mistakes." (P3)

P6 normally works with perfect transcripts and found that the errors by the speech-to-text caused them to rely more on the audio than they normally would, although P7 and P8 said they could use their memory to ignore many of the mistakes in the transcript. P8 reported that lower accuracy transcripts caused them to make rougher edits than they would normally.

We observed that all of the participants chose only to correct errors that impacted on their ability to read the transcript. P2, P4, and P6 said that gross inaccuracies in the transcript distracted them, which caused them to read slower and impacted the editing speed.

"It's good to have the option to sharpen it up as you go along because, obviously, reading back it'll slow you down if it's completely the wrong word." (P2)

We observed that the speech-to-text system would often make repeated mistakes on an unknown word by mistranscribing it as a variety of words, which made it difficult to fix. This usually occurred with names of contributors, or words specific to the topic of the program. P3 and P7 asked whether it would be possible to provide custom training to the speech-to-text system to tailor it for their specific program.

"If you're doing a story about AIDS, there's going to be stuff about anti-retrovirals [. . .] The ability to teach it some words would be really good." (P3)

Other than P7, who already uses speech-to-text, the participants reported that they normally write transcripts themselves manually. P1 and P3 stated that the speech-to-text element was the largest benefit of the semantic speech editing systems, as it freed up that time.

"The transcription thing for me is eighty percent of the advantage." (P3)

5.2.3 Listening

All of the participants chose to listen to the audio while editing with the transcripts. They gave four reasons for doing so: processing information, efficient navigation, judging quality, and identifying non-speech sounds.

P1, P4, and P6 reported that listening while editing made it easier for them to process the information that was being communicated in the interviews. P1 and P6 said this helped them to find where corrections needed to be made and find words that were inaudible or not actually present. P2 and P8 suggested that the multi-modal input of listening and reading helped them to understand the content and make edit decisions.

"I think reading and listening at the same time makes it easier to take that amount of information on. It's going into two sensory inputs so it's easier." (P8)

P2, P4, P5, and P7 spoke of how they used listening in combination with the transcript to efficiently navigate and edit the audio by skipping forward when what they were hearing was not usable, jumping backward to review content that had already been listened to, and seeing if the upcoming audio was something of interest. If it was not, then they could avoid listening to it altogether, which would save them time.

"You can glance at the transcript and just see there's a paragraph of stuff that really is not really relevant [. . .] and just discount it, whereas with your ears you've got to listen to the whole thing." (P5)

Although a transcript can tell you what was said, it does not tell you how it was said. This can change the meaning of the words, and make the difference between an edit that works or not. One thing the participants were looking out for were any low quality sounds such as "umm"s and breaths, which are distracting to listeners and can reduce the intelligibility of the speech. The speech-to-text process does not attempt to transcribe "umm"s, breaths or non-speech sounds. This means that producers must listen to identify if and when they occurred. P7 and P8 showed an interest in using the transcript to remove these noises.

P4, P6, P7, and P8 all said that they sometimes edit using only the audio itself. When the audio recording is short enough that the producer can remember what was said and where, then there is less need for a transcript. P4 put the cut-off threshold as 15–25 minutes.

6 DISCUSSION

Through our study, we wanted to learn how PaperClip affected professional radio production compared to a screen-based interface. We found that neither interface was best in all situations, but that each was better suited to different uses and circumstances.

Participants reported that the paper transcripts were easier to read and remember, and made it easier for them to think widely and orientate themselves. They reported that PaperClip was simple, intuitive, precise, and allowed edit decisions to be made faster and easier. However, PaperClip didn't include integrated playback, which made it difficult for the participants to navigate the audio.

The screen interface included integrated playback and correction, which made it easier for the participants to find mistakes in the transcript and fix them. This also made it easier to handle content that required more listening, such as old or unfamiliar recordings. As such, PaperClip may be better suited to quick and simple edits where listening is not as critical, such as with high accuracy transcripts, or very recent recordings.

The restrictions of the system we used to implement PaperClip prevented us from including integrated playback, correction or undo features. The system also interpreted the edit gestures literally, so accidentally drawing outside the boundaries could introduce errors. Both systems lack re-ordering and labelling features, which currently prevent them from being useful beyond the "rough edit" stage.

The physical nature of the digital pen and paper made it better suited to travel, working away from the desk, and collaborating with others face-to-face. However it requires access to a color laser printer, uses considerable amounts of paper, and involves carrying a digital pen. Screen interfaces require a display, which are bulky and less suited to travel. However the digital nature of the screen makes it easier to integrate and better suited for remote collaboration.

The accuracy of transcripts is crucial to the success of both systems, and two participants reported that the transcripts themselves provided the largest benefit. The speech-to-text system was accurate enough for the participants to complete their editing tasks, but participants reported that errors in the transcript resulted in more correction, slower reading, more reliance on listening, and selecting more than needed. Many errors that the participants encountered were specific to the program content, such as names and topic-specific words.

Listening is an important part of the editing process and is used to process information, judge quality, and identify non-speech sounds. With short recordings, transcripts are not needed as the audio can be edited by listening alone.

7 CONCLUSION

We introduced a novel digital pen interface for semantic speech editing and presented the results of a user study of professional radio producers that compared editing using our digital pen interface to a screen interface. We found that the digital pen and screen interfaces both worked well in different situations.

The benefits of reading from paper and the simplicity of the digital pen interface made it better for fast, simple editing with familiar audio and accurate transcripts. The integrated playback and correction features of the screen interface made it better for more complex editing with less familiar audio and less accurate transcripts. The screen interface is capable of remote collaboration, but the pen interface may work better when working with others face-to-face. The digital pen provides greater flexibility to work away from the desk, but its dependence on printing makes it difficult to work on the road. The lack of re-ordering and labelling features in both systems prevented them from being used beyond the first edit iteration.

8 FUTURE WORK

PaperClip did not include integrated playback, but this could be added by using a system that supports wireless digital pens. The audio playback could be controlled using real-time information about the pen's position. Unintentional mistakes by users when drawing edit gestures in PaperClip introduced errors. This could be fixed by developing a system that can detect and ignore these mistakes.

Participants listened to the audio in part to identify unwanted noises, such as "umm"s and breaths. By training a speech-to-text system to explicitly transcribe these sounds rather than ignore them, these could be marked in the transcript. This may help producers better judge the quality of the material and make it easier to remove unwanted noises. The speech-to-text system could also provide better transcripts by using prior information provided by the user about contributors and topics. This could be used to expand the system's dictionary and adjust the probability of specific words appearing.

Finally, the screen interface could be extended to use operational transformation techniques [32]. This would allow multiple remote users to edit audio simultaneously using a shared transcript.

9 ACKNOWLEDGMENTS

We would like to thank all of the producers who took time to contribute to the study. Thanks also to Carl Garner, Richard Sargent, and Liam Bowes from Anoto for their support in the development of PaperClip. This work was funded by the BBC and conducted as part of the BBC Audio Research Partnership.

10 REFERENCES

[1] C. Baume, M. D. Plumbley, and J. Ćalić, "Use of Audio Editors in Radio Production," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9237.

[2] C. Baume, M. D. Plumbley, J. Ćalić, and D. Frohlich, "A Contextual Study of Semantic Speech Editing in Radio Production," *Int. J. Human-Computer Studies* (2018), in press.

[3] A. Mangen, B. R. Walgermo, and K. Brønnick, "Reading Linear Texts on Paper versus Computer Screen: Effects on Reading Comprehension," *Int. J. Educational Research*, vol. 58, pp. 61–68 (2013), doi:10.1016/j.ijer.2012.12.002.

[4] L. M. Singer and P. A. Alexander, "Reading Across Mediums: Effects of Reading Digital and Print Texts on Comprehension and Calibration," *J. Experimental Educ.*, vol. 85, no. 1, pp. 155–172 (2017), doi:10.1080/00220973.2016.1143794.

[5] K. O'Hara and A. Sellen, "A Comparison of Reading Paper and On-line Documents," *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI '97, pp. 335–342 (1997), doi:10.1145/258549.258787.

[6] S. Kurniawan, S. H. Kurniawan, and P. Zaphiris, "Reading Online or on Paper: Which Is Faster?" *Proceedings of the 9th International Conference on Human Computer Interaction*, pp. 5–10 (2001).

[7] S. Whittaker, J. Hirschberg, B. Amento, L. Stark, M. Bacchiani, P. Isenhour, L. Stead, G. Zamchick, and A. Rosenberg, "SCANMail: A Voicemail Interface that Makes Speech Browsable, Readable and Searchable," *Proc. SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pp. 275–282 (2002), doi:10.1145/503376.503426.

[8] M. Apperley, O. Edwards, S. Jansen, M. Masoodian, S. McKoy, B. Rogers, T. Voyle, and D. Ware, "Application of Imperfect Speech Recognition to Navigation and Editing of Audio Documents," *Proceedings of the SIGCHI-NZ Symposium on Computer-Human Interaction*, CHINZ '02, pp. 97–102 (2002), doi:10.1145/2181216.2181233.

[9] M. Masoodian, B. Rogers, D. Ware, and S. McKoy, "TRAED: Speech Audio Editing Using Imperfect Transcripts," *Proceedings 12th International Multi-Media Modelling Conference* (2006), doi:10.1109/MMMC.2006.1651371.

[10] S. Rubin, F. Berthouzoz, G. J. Mysore, W. Li, and M. Agrawala, "Content-Based Tools for Editing Audio Stories," *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pp. 113–122 (2013), doi:10.1145/2501988.2501993.

[11] J. Casares, A. C. Long, B. A. Myers, R. Bhatnagar, S. M. Stevens, L. Dabbish, D. Yocum, and A. Corbett, "Simplifying Video Editing Using Metadata," *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, DIS '02, pp. 157–166 (2002), doi:10.1145/778712.778737.

[12] F. Berthouzoz, W. Li, and M. Agrawala, "Tools for Placing Cuts and Transitions in Interview Video," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 67:1–67:8 (2012 Jul.), doi:10.1145/2185520.2185563.

[13] S. Whittaker and B. Amento, "Semantic Speech Editing," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pp. 527–534 (2004), doi:10.1145/985692.985759.

[14] J. J. Hull, B. Erol, J. Graham, and D.-S. Lee, "Visualizing Multimedia Content on Paper Documents: Components of Key Frame Selection for Video Paper," *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pp. 389–392 vol.1 (2003 Aug.), doi:10.1109/ICDAR.2003.1227695.

[15] S. R. Klemmer, J. Graham, G. J. Wolff, and J. A. Landay, "Books with Voices: Paper Transcripts as a

Physical Interface to Oral Histories," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pp. 89–96 (2003), doi:10.1145/642611.642628.

[16] B. Erol, J. Graham, J. J. Hull, and P. E. Hart, "A Modern Day Video Flip-Book: Creating a Printable Representation from Time-based Media," *Proceedings of the 15th ACM International Conference on Multimedia*, MM '07, pp. 819–822 (2007), doi:10.1145/1291233.1291419.

[17] B. Erol, E. Antúnez, and J. J. Hull, "HOTPAPER: Multimedia Interaction with Paper Using Mobile Phones," *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, pp. 399–408 (2008), doi:10.1145/1459359.1459413.

[18] A. Fouse, N. Weibel, E. Hutchins, and J. D. Hollan, "ChronoViz: A System for Supporting Navigation of Time-Coded Data," *Proceedings of the Extended Abstracts on Human Factors in Computing Systems (CHI)*, CHI EA '11, pp. 299–304 (2011), doi:10.1145/1979742.1979706.

[19] F. Guimbretière, "Paper Augmented Digital Documents," *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology*, UIST '03, pp. 51–60 (2003), doi:10.1145/964696.964702.

[20] K. Conroy, D. Levin, and F. Guimbretière, "ProofRite: A Paper-Augmented Word Processor," *Proceedings of the ACM Symposium on User Interface Software and Technology* (2004).

[21] N. Weibel, A. Ispas, B. Signer, and M. C. Norrie, "PaperProof: A Paper-Digital Proof-Editing System," *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, pp. 2349–2354 (2008).

[22] K. Weher and A. Poon, "Marquee: A Tool for Real-time Video Logging," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, pp. 58–64 (1994), doi:10.1145/191666.191697.

[23] N. Diakopoulos and I. Essa, "Videotater: An Approach for Pen-Based Digital Video Segmentation and Tagging," *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*, UIST '06, pp. 221–224 (2006), doi:10.1145/1166253.1166287.

[24] R. G. Cattelan, C. Teixeira, R. Goulart, and M. D. G. C. Pimentel, "Watch-and-Comment as a Paradigm Toward Ubiquitous Interactive Video Editing," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 4, no. 4, pp. 28:1–28:24 (2008 Nov.), doi:10.1145/1412196.1412201.

[25] D. Cabral and N. Correia, "Video Editing with Pen-Based Technology," *Multimedia Tools and Applications*, pp. 1–26 (2016), doi:10.1007/s11042-016-3329-y.

[26] S. Vemuri, P. DeCamp, W. Bender, and C. Schmandt, "Improving Speech Playback Using Time-compression and Speech Recognition," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pp. 295–302 (2004), doi:10.1145/985692.985730.

[27] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340 (1989), doi:10.2307/249008.

[28] J. Brooke, *Usability Evaluation In Industry*, chap. "SUS: A 'Quick and Dirty' Usability Scale," pp. 189–194 (Taylor & Francis, 1996).

[29] D. Silverman, *Qualitative Research* (Sage, 2016).

[30] R. Huang, "RQDA: R-Based Qualitative Data Analysis. R package version 0.2-8." (2016), http://rqda.r-forge.r-project.org/.

[31] H. Rouanet and D. Lépine, "Comparison between Treatments in a Repeated-Measurement Design: ANOVA and Multivariate Methods," *British J. Mathematical and Statistical Psych.*, vol. 23, no. 2, pp. 147–163 (1970), doi:10.1111/j.2044-8317.1970.tb00440.x.

[32] D. Sun, S. Xia, C. Sun, and D. Chen, "Operational Transformation for Collaborative Word Processing," *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, pp. 437–446 (2004), doi:10.1145/1031607.1031681.

THE AUTHORS

Chris Baume

Mark D. Plumbley

David Frohlich

Janko Ćalić

Chris Baume is a Senior Research Engineer at BBC R&D in London, where he leads the BBC's audio production tools research and the BBC's role in the Orpheus EU H2020 project. His research interests include semantic audio analysis, interaction design, object-based audio, and spatial audio. Chris is a Chartered Engineer and a Ph.D. researcher at CVSSP at the University of Surrey.

•

Mark Plumbley is Professor of Signal Processing at CVSSP at the University of Surrey, Guildford, UK. He is known for his work on analysis and processing of sound signals, using techniques such as sparse representations, source separation, and deep learning. He currently leads two major projects on audio source separation and making sense of everyday sounds, and two EU-funded research training networks in sparse representations and compressed sensing.

•

David Frohlich is Director of Digital World Research Centre at the University of Surrey and Professor of Interaction Design. His current work explores digital storytelling, personal media collections, and augmented paper. David previously worked for 14 years at HP Labs, conducting research on the future of mobile, domestic, and photographic technology. He has written several books, numerous studies, and patents in the field of digital photography.

•

Janko Ćalić is a Senior Research Engineer at BBC R&D and a visiting lecturer at CVSSP at the University of Surrey. His research on multimedia communications ranges from video coding and processing to user aspects in multimedia delivery and interaction. Previously he was a Deputy Director of the Digital World Research Centre, University of Surrey and a Research Fellow at the Department of Computer Science, University of Bristol. He is the founding member of the Multimedia & Vision Research Group, Queen Mary University of London, where he was awarded his Ph.D.