UniS

The Semi-Automatic Generation of Audio Description from Screenplays

James Lakritz and Andrew Salway

School of Electronics and Physical Sciences

Department of Computing

CS-06-05

THE QUEEN'S ANNIVERSARY PRIZES
FOR HIGHER AND FURTHER EDUCATION
2002

UniS
University of Surrey

The Semi-Automatic Generation of Audio Description from Screenplays

Abstract

Audio Description is an important accessibility medium for visually impaired people, allowing them to enjoy visually-based experiences such as television and cinema. Its use is increasingly widespread, due in part to increased Government legislation. However,the production of audio description is time-consuming – it may take 60 person-hours to describe a 2-hour feature film.

This paper investigates the novel idea of generating a first draft audio description script from a film screenplay. This can be viewed as a text summarisation problem in which relevant sentences are identified in the screenplay and are then adapted to suit the style guidelines of audio description. Through a systematic comparision of screenplays and audio description scripts we discovered that on average a screenplay contains about 60% of the information required for an audio description, though not necessarily expressed in a suitable form. We present algorithms that can recall 80% of the available sentences from a screenplay at a precision rate of 50%. Of the resulting sentences that are not in an appropriate form for audio description, our set of heuristics can then map 66% of them to a suitable form. These results, along with feedback from evaluation sessions with BBC audio describers, suggest that the semi-automatic generation of audio description is possible and applicable in this important real-world scenario.

1. Introduction

Audio Description (AD) is an accessibility medium for visually-impaired audiences. Previously-scripted spoken descriptions of on-screen actions and appearances are played along with television programmes and films. AD scripts are currently produced manually. The process is both time-consuming and repetitive. There are currently some 60 describers employed full-time in the UK, and it may take 60 person-hours to produce a description for a 2-hour film. With over 175 cinemas in the UK offering thousands of showings of films with optional ADs every week, and with legislation requiring digital television broadcasters to provide AD for 10% of their output by the 10th year of their license, the need for a more efficient and reliable process for the production of AD is a problem with both social and commercial consequences.

The widely acknowledged 'semantic gap' means that the automatic generation of AD from video data cannot be foreseen in the near to medium-term future. Instead, we have turned to extant textual material, i.e. screenplays, as a source for the semi-automatic generation of audio description. Screenplays already contain a significant amount of the information concerning the events occurring on screen: we estimate the amount to be 60%. This paper proposes a two-stage method to generate a first-draft audio description automatically from a screenplay: (i) candidate sentences

for the audio description are extracted from a screenplay based on a measure of importance; (ii) the style of each sentence is checked and where necessary it is adapted to meet the style guidelines of audio description. Our algorithms can recall 80% of the available sentences from a screenplay at a precision rate of 50%. Of the resulting sentences that are not in an appropriate form for audio description, our set of heuristics can then map 66% of them to a suitable form.

Section 2 details the results from a comparative analysis of a corpus of audio description scripts and a corpus of screenplays. The analysis focussed on how the two text types describe the same event, and how the language used varies. Section 3 proposes a solution for the generation of AD from screenplays. Sentences from screenplays are extracted based on importance, signified by 3 lists of unusually frequent words. The language in these sentences is then converted to a more suitable language for Audio Descriptions, through defined heuristic mappings. Section 4 presents encouraging results from the evaluation of each of the two stages, and from feedback from professional audio describers at the BBC.

1.2. Background

It is difficult for someone with a visual impairment to enjoy films and television programmes without having someone describe what happens on screen. Audio Description provides this – an additional narration track that is broadcast alongside the original soundtrack for the film on which a narrator describes the events occurring on screen in-between natural pauses (i.e. dialogue gaps). Snyder [10] and Whitehead [12] list the vast number of venues where AD is currently offered including museums, galleries, and theatres, though the focus of this paper will be on film and television.

ADs for film and television have to be very tightly scripted, describing events succinctly and accurately, whilst ensuring that the descriptions do not clash with any existing dialogue. This results in a time-coded script that dictates the exact time sequence when the description should be read. Figure 1.1 shows a sample of an AD script.

01:10:59 (fast) Hana passes Jan some banknotes
01:11:12 Laughing, Jan falls back into her seat as the jeep overtakes the line of lorries
01:11:18 An explosion on the road ahead
01:11:24 The jeep has hit a mine
01:11:27 Hana jumps from the lorry
01:11:35 Desperately she runs towards the mangled jeep
01:11:42 Soldiers try to stop her
01:11:46 She struggles with a soldier who grabs hold of her firmly

Figure 1.1. Sample Audio Description from The English Patient, described by Di Langford.

The current process for producing these descriptions involves manually identifying all the dialogue gaps where a description can be spoken, and then writing a description that fills that dialogue gap and concisely explains the actions occurring on screen. This is often repetitive as

Hana leans under the tarpaulin, holding some DOLLARS. The two hands - hers and Jan's - reach for each other as the vehicles bump along side by side. They laugh at the effort. Jan's GOLD BRACELET catches the sun and glints. HANA I'm not sewing anything else for you! JAN (getting the money) I love you.	Hana passes Jan some banknotes Laughing, Jan falls back into her seat as the jeep overtakes the line of lorries
The Jeep accelerates away. Hana sighs to the patient. Suddenly AN EXPLOSION shatters the calm as the jeep runs over a MINE. The jeep is THROWN into the air. The convoy halts and there's chaos as soldiers run back pulling people out of the vehicles. Hana runs the other way, towards the accident, until she is prevented from passing by a soldier.	An explosion on the road ahead The jeep has hit a mine Hana jumps from the lorry Desperately she runs towards the mangled jeep Soldiers try to stop her She struggles with a soldier who grabs hold of her firmly

Figure 1.2. Comparison of SP to AD

the initial description written may not fit the gap available, so will need to be rewritten and refined until it does fit the gap. The recent EPRSC-sponsored Television in Words (TIWO) project investigated and prototyped a variety of technologies to assist in the production of audio description [9].

Observations made of Describers from the BBC Audio Description Department (now Red Bee Media) highlighted three key assumptions about the ways in which Audio Describers work.

i. The Audio Describer does not want to draw attention to themselves.
ii. Audio Describers want to be as succinct as possible.
iii. Audio Describers do not want to call attention to elements of film direction, i.e. camera angles etc.

Figure 1.2. shows an example of a screenplay (SP), for the same section of film as the Audio Description sample in Figure 1.1. Screenplays are a pre-written narrative of the contents of a film including dialogue, 'stage directions' and scene information. The 'stage directions' provide information about the events happening on screen. This is often descriptive language and is comparable to the information found in Audio Descriptions. Every film has an accompanying screenplay and describers often refer to the Screenplay when describing scenes, occasionally using sentences directly from screenplays. This report suggests we can exploit the screenplay further.

There is a large difference in the amount of information presented in the two media – a screenplay is approximately 3 times larger than an AD, yet it contains a significant amount of the information needed in Audio Descriptions. This text appears to be a good candidate for a summarisation task, extracting, and then adapting, only the relevant sentences from the screenplay to produce a candidate AD.

2. Corpus Analysis

The aim of this analysis is to inform the design of a system capable of identifying relevant sentences from the screenplay, and adapting them into a form suitable for audio description. This investigation assumes that the most important utterances in the SP will be describing events that should also be described in the corresponding AD.

Luhn [5] and Edmundson [1] suggest that importance in a text is denoted by frequency of words. This analysis looks at the ways the SP describes an AD event based on unusually frequent words (UFW) in the SP utterances.

The AD corpus used in this analysis comprised 70 ADs, constituting over 454,600 words. The AD scripts came from a number of film genres, and were all written by professional describers.

SPs and ADs both make reference to the actions or events occurring on screen. In an SP this information is included in addition to the dialogue, and often takes the form of describing the actions of a character. The same information is also present in the AD, if a sufficient dialogue gap is available. Whilst the tone of the two texts is different, SP has a narrative tone and AD has a descriptive tone, both descriptions are referring to the same event. This analysis compares the quantity of AD events featured in the SP, and compares and contrasts the language used to describe these events. Three lists of words are identified that are unusually frequent amongst the events, and three key differences in the language and grammar used to describe events are discussed.

Utterances in SPs that relate to events in ADs can be paired together; this is termed an SP-AD pair. The creation of these pairs is performed on a sentence level, working from events in the AD to utterances in the SP. During this process, it became apparent that there were two types of pairings. The first type, direct pairings, are explicit matches between an utterance in the SP and an event in the AD. The second type is indirect pairings that contributed to the description of an AD event. Figure 2.1. provides examples.

SP Utterance	Type	Pairs With (AD)
Lester's briefcase suddenly springs open and his papers spill all over the driveway.	Direct	His briefcase falls open.
Lester hurries out the front door, carrying a BRIEFCASE.	Indirect	His briefcase falls open.
The Colonel and Barbara are seated on a couch, watching television.	Direct	Now, Frank and his wife are watching a black and white film.
We HEAR a door opening elsewhere in the house, and Ricky enters.	Indirect	At the sound of Ricky coming in, Frank leans back and folds his arms.

Figure 2.1. Screenplay to AD mapping classifications

The results show that almost 60% of the events in the AD form an SP-AD pair, and that almost 20% of all the screenplay utterances are part of an SP-AD pair.

2.1. Unusually Frequent Words

Based on Luhn and Edmundon's theories that frequency indicates importance, three groups of unusually frequent words (UFWs) are present in SP-AD pairs. Here, a word is considered unusually frequent if it is open-classed and is amongst the most frequent words in the text.

2.1.1. UFWs From ADs

The first set of words is based on unusual frequency in ADs. The words with the highest importance in ADs are the most frequently used words. This analysis found the 30 most unusually frequent words from the AD corpus. The words fall into three groups – character descriptions, like body parts, common objects, like door and car, and action verbs like look, and smile (see Appendix A). This is intuitive,

as these words are descriptive. These findings corroborate those of Salway, Vassiliou and Ahmad [8], who show an identical list of words. The lemma groups of these words are also frequent amongst SP utterances, so also can be used as an indicator of importance.

2.1.2. Character Names

SP-AD pairs often contain at least one character name. The linguistic regularities of character names indicate an importance of this subset of words, and the most important character names must be the ones that appear most frequently. Assuming frequency of dialogue is indicative of character importance, a list of UFWs can be compiled of character names. This is achieved by counting the number of separate blocks of dialogue per character in the screenplay. Characters with at least 10% of total dialogue featured in almost all SP-AD pairs where a character name occurs in the SP utterance. Any character with more than 10% of the total dialogue in the screenplay form a second list of UFWs. This wordlist is dynamic, as it changes dependent on the screenplay being analysed.

2.1.3. Domain-Specific UFWs

The words in the above lists have no concept of domain-specificity. For instance, a war film may refer frequently to the concept of trenches, or battlefields. The frequency of these words implies some importance, meaning they likely form part of an SP-AD pair. The above lists have no relevance to events based around these domain-specific concepts. This set of words addresses this weakness. By choosing the most frequent words in SP utterances the final UFW list can be compiled. All closed-class words (recognised by a stoplist) are ignored, and words already included in the above lists are also discarded. Of the remaining words, the 20 most frequently occurring are selected as the most important.

2.1.4. Results

There are a significant number of SP-AD pairs where the SP utterance contains at least 1 of the UFWs defined above. On a sample of 150 SP-AD pairs, the least effective set of UFWs (Domain Specific) still featured in 30% of the SP portion of that pair. When all the sets are combined together, the amount of SP-AD pairs where the SP utterance features a UFW increases to 80%. The total number of utterances in the SP that contain the words is consistently double the number of utterances that form part SP-AD pairs, as shown in Figure 2.2.

	SP-AD pairs where SP contains UFW	Utterances containing UFW in SP
AD UFWs	95	162
Character Names	82	165
Domain Specific UFWs	46	103
All sets of UFWs	121	256

Figure 2.2. UFW Results

2.2. Language Comparison

The language and grammar used in Audio Descriptions and Screenplays differs significantly. The two texts describe the same events, yet still the language is unalike. This section aims to discover the way language and grammar are used to convey the same information.

An Audio Description can be thought of as a subset of its corresponding Screenplay, as the language used in an AD can also be used in SPs.

However, the language differences arise because SPs use language and grammar that are unsuitable for ADs. This is because ADs are restrictive by nature, whereas SPs have no such restrictions.

The three assumptions about the ways describers worked, as discussed in Section 1.2. highlighted three key language differences, which are examined here in terms of frequency. Of interest in this analysis is the frequency of occurrence of these differences in the SP utterances that contain the Unusually Frequent Words already identified.

2.2.1. Point Of View

The grammatical point of view employed in SPs and ADs is often different. Screenplays commonly use the collective first person ('we see', 'looks straight at us'), whereas Audio Descriptions always utilise the third person perspective. This is due to the principle that Describers do not want to draw attention to themselves.

2.2.2. Camera Instructions

Screenplays feature an abundance of camera instructions, a result of the text being written to convey how a film should be shot. However, Audio Descriptions do not include these instructions for two reasons. Firstly, they do not need to convey the manner in which a scene has been shot, and secondly, in doing so would draw attention away from the actual description of the action on screen.

2.2.3. References to Sound Effects

ADs give no indication of sound effects being made, or the dialogue being said, as these will already be heard along with the AD narration itself. However, SPs often reference sound effects or the manner in which dialogue is delivered, as instructions for how these elements should be included.

2.3 Discussion

The results of our analysis suggest four main findings.

i. 60% of events in ADs are also described in SPs

The types of film analysed are very different – a war film, an old classic, a modern drama – yet this value of 60% remained within a 10% margin for each film.

ii. In 80% of SP-AD pairs, the SP utterance contains at least 1 word that is unusually frequent in SPs.

This indicates that both SPs and ADs describe the events in similar ways, using similar language. The high percentage of utterances containing at least 1 UFW demonstrates that these are the words used to describe important events in the SP. This implies that SPs describe events that also appear in ADs using specific language – often the event descriptions from the SP will feature a character name, or a UFW from ADs, illustrating a consistency in event description in SPs.

iii. 50% of SP utterances containing at least 1 UFW form part of an SP-AD pair.

Of all the utterances containing at least 1 UFW, half of these form part

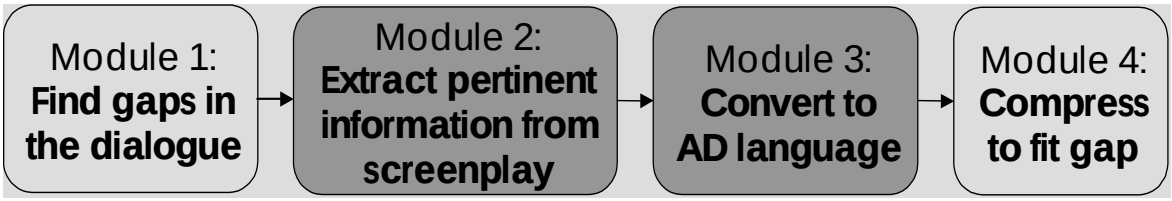

Figure 3.1. Modular System Design

of an SP-AD pair. This complies with the theory of frequency denoting importance, as the most unusually frequent words occur in 50% of the SP utterances that form SP-AD pairs.

iv. 20% of SP utterances containing at least 1 UFW use different language to that of ADs.

One in five SP-AD pairs uses different language and grammar to describe the same events.

These results support the feasibility of automatic SP to AD generation. Crucial is the high percentage of events described in the screenplay that can map to the AD. The linguistic regularities described above could aid the automatic extraction of the relevant information. The perfect system would be able to automatically generate 60% of the AD using the SP. A system built on the knowledge obtained in this paper would be able to identify 80% of those events using the UFWs as cue-words, whilst returning an extra 50% of 'noisy' SP utterances.

In the automatic generation of an Audio Description, the language differences should also be considered, and used as a starting point for the creation of a set of heuristics to 'map' a SP sentence to a corresponding sentence in an AD using appropriate language for ADs. Using the above analysis, it is possible to design a mapping such that the major language differences are minimised, thereby reducing the workload for a Describer when refining the first draft output of the system. This can be achieved using simple mappings based on heuristic evidence.

3. System Design

Figure 3.1 shows our design of a system capable of automatically generating a first draft Audio Description.

This architecture firstly identifies gaps in the soundtrack of the programme. These gaps form placeholders for candidate descriptions. The second module identifies important information contained within the screenplay. These sentences form the basis of the candidate descriptions that are used in the gaps found in module 1. Module 3 then attempts to convert the local grammar and language used in the screenplay to a local grammar and language that more closely resembles that of an AD. The converted sentences are passed to the final module, which compresses the sentences to fit the gap identified by module 1.

A modular approach was chosen as this allows separation of processing, and the creation of well-defined interfaces between the modules. This paper will focus on modules 2 and 3 of the architecture, as these are the pivotal modules. By ensuring that well-defined interfaces exist between the modules, future work

can easily extend on the work detailed here.

3.1 Extraction Module

The design for this module can be seen in Figure 3.2. The first stage in the design is cue-word identification, generating the wordlists that are used to extract sentences. These sentences are then identified, anaphora are resolved and the output is formatted correctly to resemble an Audio Description.

3.2. Language Conversion Module

Heuristic mappings can be used to convert the language. This involves deletion, reordering and editing of the existing sentences to remove any unsuitable information. The approach is based on pattern-matching within sentences, identifying specific word arrangements and then performing a transformation of those words.

4. Evaluation

The two modules discussed above were implemented and evaluated in turn and user feedback was obtained from the BBC Audio Description Department .

4.1. Extraction Module Evaluation

The extraction algorithm was run on a sample set of 3 screenplays that had been fully mapped into 719 SP-AD pairs. The three films each covered a different genre. The extraction module returns results with approximately 80% recall and 50% precision. The high recall rates demonstrate that the analyser is capable of extracting the SP-AD pairs from the screenplay. As mentioned above, 20% of the utterances in a SP feature in SP-AD pairs. Therefore, returning the whole SP would result in a precision of 20%. The precision of this analyser is 3 times greater than that achieved by returning the whole SP. Salton [7] and Veit *et al* [11] suggest that automatic retrieval methods generally have a better recall than precision, but conversely manual methods have a stronger precision and weaker recall. Veit et al further suggest that the best extraction systems combine automatic methods, to benefit from the high precision, and manual methods, to optimise on precision. This system follows this approach, as the AD generated is a first draft and intended to be refined manually by a Describer.

Anaphora resolution focussed on pronoun resolution only, based on a 'recency strategy' as suggested by Jurafsky and Martin [3].The pronoun resolution algorithm has an 80% success rate. Jurafsky and Martin consider a method of pronoun resolution that uses a variety of factors to identify referents, and report that these have an optimal performance in the mid-80% range. Using just a recency strategy, this module manages to achieve success rates close to this. This was achieved by ensuring pronoun gender agreement. It is believed that

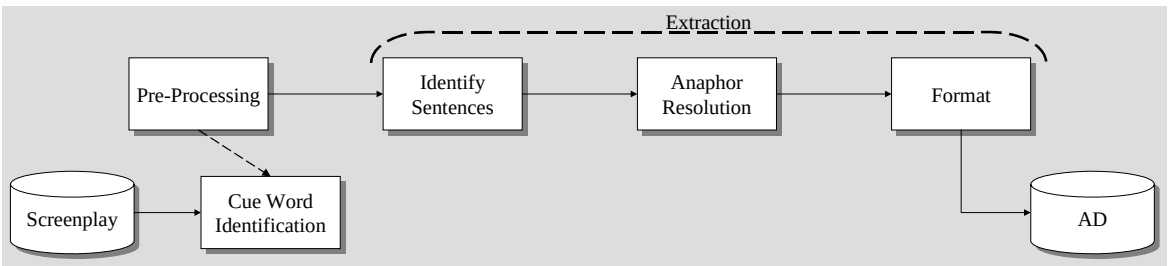

Figure 3.2. Sentence Identification Module Design

this success rate will remain consistent if the resolution algorithm is extended to include all anaphora in addition to pronouns.

4.2. Language Conversion Module Evaluation

The language conversion module was built around heuristics based on the results of the analysis in Section 2 (see Appendix A). Of the approximate 20% of sentences not suitable for inclusion in ADs, 66% of these were correctly converted to a suitable description using these heuristics. Where the module fails to correctly convert a sentence, it is usually due to a stylistic problem, rather than a grammatical one. For instance, the sentence, "*Jane turns and walks quickly towards her house, flipping off as she goes*", was rejected by Describers as 'flipping off' is an Americanism, so unsuitable for inclusion. As the approach detailed here is shallow and syntactical and not semantic in nature, the meaning of the sentences is not considered, so these types of incompatibilities are not addressed.

4.3. User Feedback Evaluation

The BBC Audio Description Department evaluated the concept by using the system to describe a short sample of a film with the aid of some sample system output from module 3. The time taken was compared to the time taken to describe a similar length of film using their standard methods. Both clips were taken from the same film, and it was hoped that this would mean they were of similar complexity to describe.

The output from the two modules was not able to increase the efficiency of the describing process. It took almost double the time to describe the film clip using the system output. It should be noted that the describer suggested that the two samples of film were not of similar complexity – the film clip described with the aid of the system output was deemed to be more complex as it contained longer pauses, and less dialogue. This, and the learning curve required for the new system, may have led to the longer description time, but nonetheless the describer felt the system required the additional modules to be more beneficial. They indicated that as the system output was not time-coded it made it difficult to manually identify the gaps, and then find the right description from the output for that gap. However, they suggested that the output sentences were useful as a reference.

They also commented that the SP sometimes contained extraneous and inaccurate information. They found that the output sometimes mentioned events that did not occur on the screen – this is due to the 50% precision, meaning some noisy sentences are included in the output.

The describer summarised that the combination of these factors made it difficult to describe the film clip, and required additional thought processes in order to complete the description. However, they suggested that the system would be useful if it identified the dialogue gaps, and mapped the extracted text to these gaps. The describer felt that this would make the editing of the sentences easier, and could improve the efficiency of the process. They suggested they would be likely to use a fully implemented version of the system.

This comment suggests that by implementing the remaining modules the system could improve efficiency in describing and be beneficial to the process. Whilst the current conception of the system does not improve the process, it is believed that a full implementation of the system architecture will.

4.4. Evaluation of Approach

The success of the approach taken can be quantified by showing the number of sentences that could feature in an AD through each stage of the analyser. This is based on the relevancy, the grammatical accuracy and the stylistic suitability of the sentences. It has been shown that the extraction analyser has a precision of 50%, so the highest that can be expected is 50%. Figure 4.1 shows the results achieved.

	% sentences suitable for AD.
Original Screenplay	**20%**
After Sentence Extraction	41%
After Language Conversion	**48%**

Figure 4.1. Sentences that Could Feature in AD

The table shows a continual improvement in the number of sentences that can be used in an AD. Only 20% of all the sentences in a SP are relevant to the AD. The extraction analyser addresses the issue of relevance by extracting only relevant sentences. After extraction the number of relevant sentences increases to approximately 50% (the precision of the analyser). However, not all of these sentences can be used for grammatical and stylistic reasons. 41% of the extracted sentences remain appropriate. The next analyser addresses the grammatical and stylistic problems, and manages to increase the number of appropriate sentences further still to 48%.

Screenplays can be incorrectly ordered as the final version of the screenplay is produced pre-editing, and once the film or programme is edited scenes can be re-ordered. This means that the generated AD would also be incorrectly ordered, as it outputs in the inherent chronological order of the screenplay. However, this is only a first draft so it is feasible to suggest that during the manual refinement stage, a Describer can re-order the AD.

This re-ordering could also be attempted automatically, using a method similar to Assisted Subtitling. Evans [2] describes Assisted Subtitling as a process of using the transcript of a soundtrack along with the soundtrack itself to identify when a phrase is said, allowing subtitles to be displayed. The process makes use of voice recognition technology, but combines this with having the actual transcript of what is being said. As the voice recognition software knows what words it is expecting to hear, it can more accurately identify the phrases being said, and in turn find the correct times for the subtitles. Using a similar idea to this, voice recognition software can be used on the soundtrack of a film, identifying when a particular dialogue block is said. This process will identify areas where the soundtrack and screenplay do not match – these discrepancies are likely due to a re-ordering of the programme. The system can then reorder the screenplay to produce a correctly ordered version that can

then be used as an input to this module.

5. Conclusions

To generate an AD, each of the modules in the system architecture (Figure 3.1) must to be implemented. This report has demonstrated that the functionality needed to meet the requirements of modules 2 and 3 has been achieved. It is possible to identify and extract 80% of the important information from screenplays that should feature in ADs. This process, however, also returns a number of noisy sentences that will need to be reduced for maximum benefit. It is also possible to convert 66% of unsuitable sentences from SPs to a suitable form for inclusion in ADs.

Extracting sentences based on the presence of automatically selected cue-words has proven effective. This method was chosen as it does not require semantic understanding of the text, produces output already in natural language (so does not need to be able to generate natural language) and allows a shallow approach to be taken, which Mani [6] suggests is often effective. However, this method does have limitations. Significantly, the usefulness of the output depends entirely on the accuracy of the screenplay – if a screenplay is inaccurate, the generated AD will also be inaccurate. Nonetheless, until visual processing has advanced to such a stage where meaning can be gleaned from the pixels of a scene alone, screenplays provide the only source of information capable of bridging this gap.

5.1. Future Work

Additional work is needed to enable the ideas proposed in this paper to be fully realised. Most notably, the remaining modules from Figure 3.1 need to be designed and implemented. Whilst suggesting and detailing a design for the remaining modules is outside the scope of this paper, a number of starting points have been considered and will be discussed here.

The most feasible method of finding dialogue gaps, as required by module 1, utilises signal processing. A simplistic approach mimics the workings of a karaoke filter. A karaoke filter is a form of signal processing usually rendered on a music file with the aim of removing the vocal track to obtain just the backing music. This is achieved by reversing one audio channel (say the right channel), and combining it with the other channel. This has the effect of removing any sound that occurs equally on both the right and left channel. As spoken dialogue is usually centred between the two channels, this method should be able to noticeably reduce the vocal track. The new audio soundtrack can then be compared against the original soundtrack, and any differences should correspond to periods of dialogue. This method has inherent problems, for example if a line of dialogue is spoken only in the right channel then it will not be recognised. However, initial feasibility testing shows promising results.

Another possible approach is to utilise the same mechanism used for Assisted Subtitling. This has already been discussed in relation to re-ordering the screenplay, but could be further employed to identify when the ends of dialogue blocks occur through the assisted speech recognition. The lengths of the

dialogue gaps can be measured, allowing an accurate representation of the gaps in which descriptions can be spoken. The most efficient implementation of this approach would also re-order the screenplay at the same time, thus allowing the next stage, the extraction module, to run effectively.

The precision of module 2 also needs to be addressed. It currently stands at 50%, but by increasing this figure, more of the irrelevant text can be removed from the system output. This report suggests two starting points for any future work. Edmundson [1] suggests a set of words termed 'stigma-words' that carry a negative weighting on sentences in which they are included. Stigma-words could be used to remove sentences from the extracted output, hopefully removing sentences that are irrelevant. Stigma words could be found through word-frequency analysis on all the sentences that are incorrectly extracted by the analyser. A second approach is to only extract sentences that contain more than one cue-word. Initial testing of this approach shows that precision is generally increased by 10%, but that recall falls significantly.

The heuristic mappings in module 3 can be improved. They currently correct approximately 66% of unsuitable SP sentences. To increase this coverage requires additional, and more finely tuned, mappings to be designed. For instance, the analysis in section 2 identified 3 key differences between the language used in ADs and SPs, but Lakritz [4] details 7 differences in total. A more comprehensive implementation, exploring each of the remaining differences, would be of significant use in increasing the suitability of sentences extracted from SPs.

Mapping extracted sentences to time-coded gaps, the requirement of module 4, also needs to be possible for an effective solution. The difficulties here arise because there is no correlation between an extracted sentence and when that event occurs on screen. Three initial ideas may merit further consideration in future research. Firstly, the first extracted sentence could be mapped to the first dialogue gap, and so on until every extracted sentence has been mapped to a gap. Whilst this method is overly simple, it would be interesting to see how effective it could be. Secondly, the overall length of the film could be ascertained by measuring the length of the soundtrack or video file. This length can then be divided by the number of lines in the screenplay – this will give a crude guide to the time associated with each line of the screenplay. For instance, if the film was 100 minutes long, and the SP was 10,000 lines long then each line in the SP would be assigned a half second value (0.01 of a minute). This provides a guide for where each extracted sentence would occur in terms of the time-coded dialogue gaps. Lastly, elaborating the Assisted Subtitling concept again, each dialogue block could be time-coded through use of the voice recognition. Extracted sentences will have been positioned between these blocks of dialogue, so an approximate time of its occurrence can be judged. For instance, if two contiguous blocks of dialogue are spoken at 66 minutes and 68 minutes respectively, then any extracted sentences that occur between these dialogue blocks must

refer to events occurring between 66 and 68 minutes. Extracted sentences can then be mapped to the closest time-coded dialogue gap.

There may also be a need to compress the extracted sentences so they fit the dialogue gap. This requires the knowledge of how long a description will take to speak, and also a compression algorithm. Discussion on compression is beyond the scope of this report, but initial research could be conducted into the recent work by CNTS - Language Technology Group of the University of Antwerp in Belgium. They have been involved in work on the MUSA project to enable automated subtitling, including the ability to compress sentences. See *http://sifnos.ilsp.gr/musa/* and an online demo at *http://www.cnts.ua.ac.be/cgi-bin/anja/musa* for more information.

5. References

1. Edmundson, H. P., (1969) New Methods in Automatic Extracting, Journal of the ACM,, Vol 16, Issue 2, Pages 264-285, 1969. Available: http://doi.acm.org/10.1145/321510.321519
2. Evans, M.J. (2003) Speech Recognition in Assisted and Live Subtitling for Television, BBC R&D White Paper, WHP 065. Available: http://www.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP065.pdf
3. Jurafsky, D; and Martin, J. (2000). Speech And Language Processing. New Jersey: Prentice Hall.
4. Lakritz, J. (2006). Towards the Automatic Generation of Audio Descriptions, Undergraduate Dissertation, University of Surrey
5. Luhn, H. (1958) The Automatic Creation of Literature Abstracts, Available: http://www.research.ibm.com/journal/rd/022/luhn.pdf
6. Mani, I., (2001) Automatic Summarization. Philadelphia: John Benjamins Publishing Company.
7. Salton, G. (1986) Another Look At Automatic Text Retrieval Systems, Communications of the ACM, Vol 29, Issue 7, Pages 648-656, 1986.
8. Salway, A; Vassiliou, A; and Ahmad, K. (2005). What Happens in Films? in Procs. IEEE International Conference on Multimedia and Expo, ICME 2005 Available: http://www.computing.surrey.ac.uk/personal/pg/A.Salway/pdfs/What Happens in Films FINAL.pdf
9. Salway, A., (2005) TIWO: Television In Words, Available: http://www.computing.surrey.ac.uk/personal/pg/A.Salway/tiwo/TIWO.htm
10. Snyder, J., (2005) Audio description: The visual made verbal, International Congress Series, Volume 1282, Pages 935-939, 2005
11. Veit, D. Müller, J. P. Weinhardt, Ch. (2002) An Empirical Evaluation of Multidimensional Matchmaking. In Poster Proceedings of the Workshop on Agent Mediated Electronic Commerce IV (AMEC-IV), held at the First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), Bologna, July 16th 2002. Available: http://www.iw.uni-karlsruhe.de/Publications/Veit Mueller Weinhardt - Empirical Eval MM.pdf
12. Whitehead, J., (2005) What is Audio Description, International Congress Series, Volume 1282, Pages 960-963, 2005.

Appendix A

UFWs from ADs							
man	head	face	eyes	hand	hands	men	woman
looks	turns	takes	walks	goes	stands	steps	smiles
stares	puts	watches	opens	looking	door	room	car
window	table	water	bed	house			

Figure 4a. UFWs from ADs

Common Pattern	Suggested Mapping	Example
Looks [up|down|back] at us	**Looks [up|down|back]**	**"Jane** looks up at us**"** **→ "Jane looks up"**
Looks ___ at us	**Looks ahead**	**"Carolyn looks directly at us" → "Carolyn looks ahead**
We see ___ <verb>	**___ is <verb>-ing**	**"**We see Lester dance**"** **→ "Lester is dancing"**
We see	**null**	**'**We see a street being decorated**"** **→ "A street being decorated"**
We look ___ [at|on]	**null**	**"**We're looking right at **Jane dancing" → "Jane dancing"**
We <verb>	**<verb>-ing**	**"**We're flying **above clouds" → "Flying above cloud"**
We	**null**	**"**We're **now in an untidy room" → "Now in an untidy room"**

Figure 5a. Heuristics for Point of View language conversions

Common Pattern	Suggested Mapping	Example
[Medium | Tight | Extreme] [Closeup | Close Up | Close On]	**null**	"**Extreme close on** a framed photograph as Lester picks it up." **→ "A framed photograph as Lester picks it up"**
[A] series of shots [of | as]	**null**	**'**Series of shots as **four or five boys make the slide down the hill and out onto the ice**.**" → "Four or five boys make the slide down the hill and out onto the ice"**
<char> POV:	**null**	**"**Cole's POV: **a BMW speeds toward them, passes, its radio blaring" → "A BMW speeds toward them, passes, its radio blaring"**

Figure 5b. Heuristics for Camera Instruction Language Conversion

Common Pattern	Suggested Mapping	Example
We ___ hear	**Delete offending clause from sentence**	**"**We barely hear **a sound, as she sleeps" → "She sleeps"**

Figure 5bc Heuristics for References to Sound Effects Language Conversion

Department of Computing

University of Surrey
Guildford, Surrey
GU2 7XH UK

Tel: +44 1483 683133
Fax: +44 1483 686051
E-mail: a.salway@surrey.ac.uk
www.surrey.ac.uk

UniS
University of Surrey