From my pen to your ears: automatic production of radio plays from unstructured story text

Emmanouil Theofanis Chourdakis
Queen Mary University of London
e.t.chourdakis@qmul.ac.uk

Joshua D. Reiss
Queen Mary University of London
joshua.reiss@qmul.ac.uk

ABSTRACT

A radio play is a form of drama which exists in the acoustic domain and is usually consumed over broadcast radio. In this paper a method is proposed that, given a story in the form of unstructured text, produces a radio play that tells this story. First, information about characters, acting lines, and environments is retrieved from the text. The information extracted serves to generate a production script which can be used either by producers of radio-drama, or subsequently used to automatically generate the radio play as an audio file. The system is evaluated in two parts: precision, recall, and f_1 scores are computed for the information retrieval part while multistimulus listening tests are used for subjective evaluation of the generated audio.

1. INTRODUCTION

Recent advances provide computational methods for generating artistic works that allow a human author to also participate, resulting in joint human/machine creative works. Such methods have been devised individually for music [1, 2], poetry [3], literature [4], 3d scene generation [5], and film [6].

A radio play is a form of drama which exists in the acoustic domain [7]. To the authors' knowledge, there have been no such efforts targeting this form of art. In this paper, we discuss how recent research from the fields mentioned above can be used to devise an autonomous producer of radio drama. The system we propose is divided into two distinct stages: semantic analysis of stories, and sound production of the radio play. Semantic analysis takes stories in their original unstructured text format and produces a human readable semi-structured production script. The production stage generates a radio play from this script. This division into stages allows us to evaluate and report findings about each stage independently. A human user can also intervene between the two stages of the generation process, thus allowing the tool to be used as an assistant in composing radio drama, for example fixing mistakes in the production scripts, changing acting lines or attributes of characters or scenes. The main contributions of this paper therefore are: a methodology for automatically inferring the elements of the story relevant to radio play production, and demonstrating how these elements can be used to produce a finalised play.

The rest of the paper is divided as follows. Section 2 presents relevant literature. Sections 3 and 4 present the methodology used for the implementation of each stage in the generation process. Section 5 evaluates a proof of concept implementation on a corpus of Aesop Fables. Finally, section 6 provides some introspection about the work achieved so far as well as future research directions.

2. BACKGROUND

Here, we present research that formed the basis for our work. We make use of ideas introduced in works related to information extraction from natural language text, as well as novel approaches in audio-based storytelling.

Information extraction from stories is a task which has been tackled in many previous works, mostly focused around identifying characters in stories and their social networks. One of the latest works appears in [8] where the authors use natural language processing techniques such as co-reference resolution, a hand-crafted ontology, and pattern matching to extract such characters as well as their relations. *Co-reference resolution* is the problem of identifying and clustering parts of a text, called mentions, that refer to the same entity. For example, in the following sentence:

> "A bee from Mount Hymettus, the queen of the hive, ascended to Olympus to present Jupiter some honey fresh from her combs."

The mentions *A bee from Mount Hymettus*, *the queen of the hive*, and *her* map to the same entity, or cluster (the bee). The task of co-reference resolution, is to extract such clusters. Co-reference resolution has been a particularly hard task for Natural Language Processing. An interesting recent approach can be found in [9], which builds such clusters incrementally, starting with each mention as its own cluster. In this work, we use co-reference resolution to derive mappings from characters to gendered pronouns and thus extract information about the characters' genders when not explicitly stated.

Spatial Role Labelling was introduced in [10] and pertains to extracting information from sentences that describe some kind of spatial relation (e.g. "*A bull* was feeding *in a meadow*."). Since then, it has led to a large amount of studies and data on the problem [11]. A notable such study [12]

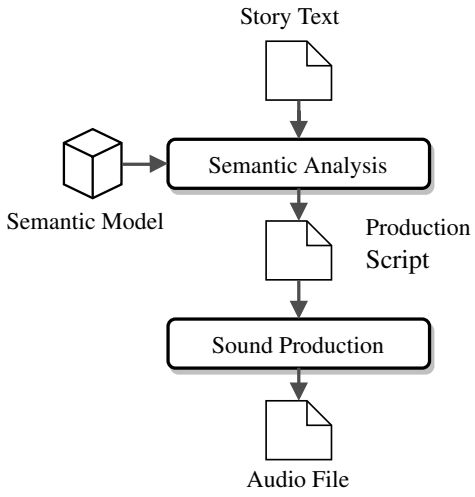


Figure 1: A block diagram of the system.

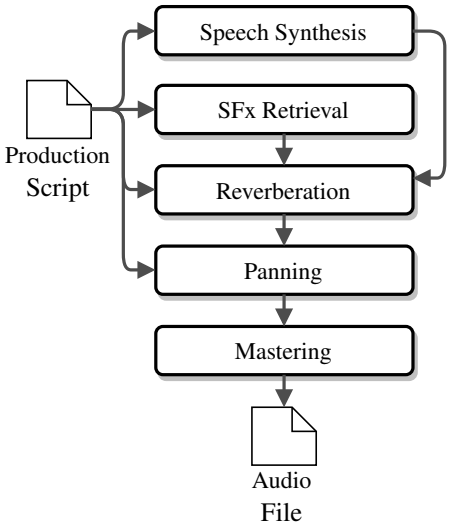


Figure 2: The sound production process

uses high recall heuristics to mine candidate relation constituents (such as *the bull*, *a meadow*, and *in* in the previous sentence) and train a binary SVM classifier to identify such relations. In a similar fashion, [13] deployed Conditional Random Fields to capture the relation constituents and use them as heuristics for relation extraction. When combined with character identification, spatial role labeling can be useful in establishing the character's spatial position in a specific sentence in a story.

A mapping from crowd-sourced reverberation effect labels (such as *dry*, *wet*, or *underwater*) to reverberation effect settings was presented in [14]. Their work is aimed at newcomers in music production, allowing them to apply the effect of reverberation to a piece of audio without getting lost in complicated audio effect control settings often found in mainstream reverberation effect plug-ins. Their mappings also prove useful in controlling the audio effect from text input, such as stories, as we found in our case.

An approach to evaluating a format closely related to radioplays can be found in [15]. In that work, various participants were asked to identify characters and spaces in a novel audio film format without narration designed for sight-impaired users. The experiment aimed to discover factors that aided identification or characters and spaces. In our work, we use some of these factors to convey information about characters and spaces in the radio plays generated by our method.

3. SEMANTIC ANALYSIS

Semantic analysis of stories is performed in order to identify key story elements that will later guide sound production. For this reason, a semantic model is constructed to do co-reference resolution, character identification, dialog lines separation between character acting lines and narrator lines, and detection of the environments the stories take place in. Figure 3 shows how the model was constructed.

3.1 Annotated Corpora

For the purpose of this work, a corpus of 360 Aesop Fables was curated from various websites. From this corpus, a subset of 20 Aesop Fables was annotated using BRAT [16] and kept as a testing dataset for evaluating the semantic analysis. The 20 fables were chosen at random but on the condition that the resulting dataset contains a balanced representation of the semantic elements examined. From the remaining 340 fables, we sampled and annotated 67 sentences to be used as training where required. Furthermore, 3 story segments were extracted from the initial corpus in order to generate audio files for the listening tests in Section 5.

3.2 Acting Lines Separation

Acting lines are parts of the text that will be spoken by an actor or by a speech synthesis engine when the play is generated. At this stage, the text is split between speech lines for our characters and speech lines for the narrator. Observing the corpus, character lines can be easily distinguished by the surrounding quotation marks (‘"’). Everything outside those quotations is considered narrator speech. Character and narration lines are stored in a separate file and character lines are replaced in the original text with a special tag, in order to not interfere with the analysis in subsequent steps.

3.3 Co-reference resolution

After identifying and replacing speech lines with their tags, the stories are passed through a co-reference resolution algorithm. This step is not uncommon for information retrieval in folktales [8, 17]. Apart from not having to deal with unresolved anaphora, it helps in three other ways:

1. The algorithm serves as a heuristic for identifying characters. Characters are usually referred to in many places in a story and such an algorithm captures those references. The algorithm might miss some cases, such as when the character is mentioned only once, or capture false positives, such as when the reference is on objects.

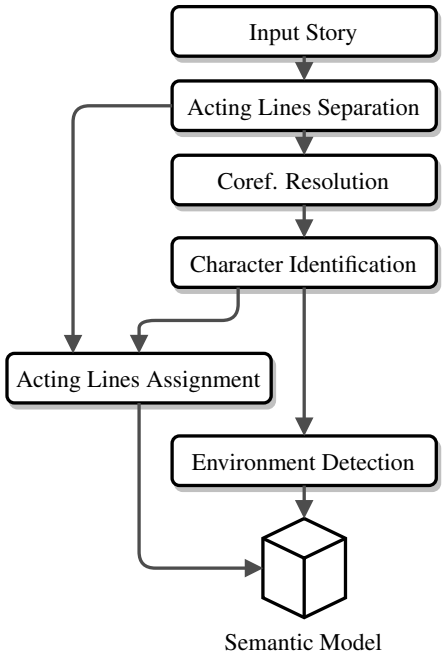


Figure 3: The training process of the semantic model

2. Clusters of mentions provide candidates for characters as well as information about their perceived gender (in the scope of the corpus). Character mentions that are grouped with 'he' or 'him' pronouns are assigned as 'male' and ones that are grouped with 'she' or 'her' as 'female'. Characters with a neutral pronoun are not assigned a gender.

3. Sentences that include pronouns become sentences that include the referenced character in their text. This helps subsequent semantic analysis tasks by providing them with more examples that include the original characters.

The algorithm used is described in [18], and was chosen because of its easy-to-use implementation in NEURALCOREF [1]. This algorithm already serves as an adequate baseline for character identification, as seen in Section 5.

3.4 Character Identification

We consider as Characters entities that do some kind of action or say an acting line. Consider the following two sentences:

"*Jupiter and Venus were arguing...*"

"*A cat went to Venus*"

In the first one, both *Jupiter* and *Venus* are characters since they are doing something. In the second, only *A cat* is a character since *Venus* does not act.

The steps described in the previous section already serve as an adequate method for recognizing characters since third person singular pronouns are usually followed by verbs. Since co-reference resolution is not designed for character recognition, it leads to some easily identifiable mistakes:

[1] https://github.com/huggingface/neuralcoref

1. False positives – Those are mainly inanimate non-character elements, identified as characters and leading to lower precision. An example sentence would be (the references are given in *italics*):

   "When *the battle* was at *its* height"

2. False negatives – Cases where a character is mentioned only once in the text and cannot be assigned with a pronoun. These types of errors lead to lower recall. An example would be the following sentence at the very end of the story:

   "...until *an old mouse* got up and said:..."

Ignoring those errors however, co-reference resolution presents an elegant way to infer gender information of the characters in the stories.

To reduce co-reference resolution errors, we train a Conditional Random Field (CRF) model for Named Entity Recognition (NER) in order to do character recognition. As a training set, we use the 67 annotated sentences introduced in Section 3.1. The features extracted are the same as used in the CRF model in [13] and the library used is CRFSUITE [19] with the SKLEARN [20] interface. We use this model to identify Characters in the story, and the co-reference resolution part to assign them a gender attribute.

In addition to the above, a dictionary is introduced that annotates text as characters (for example the Olympian gods). This heuristic lowers precision by a small amount (some of the Characters found this way do not participate in actions) but increases recall. We also include a heuristic that assigns attributes based on the character's text (for example a daughter is automatically assigned an attribute of female and a grandfather is always male).

3.5 Environment Detection

Identifying elements relating to the story environment allows us to consider relevant sound effects for the composition of the audio scene, and also to choose appropriate reverberation settings in order to give the listener the perception that they are in that specific environment. Consider the following sentence:

"A bull was feeding in a *meadow*"

Here *meadow* refers to a specific environment that can be conveyed given sound effects that relate to a meadow, like wind or ruffling grass. It also has a specific impulse response, maybe an open space.

We approached the problem of identifying such environments as a Spatial Role Labeling (SpRL) task. SpRL tackles the problem of identifying a spatial relation, its *spatial indicator* (or indicator for short), its *trajector* and its *landmark*. In the sentence above, *in* is considered a spatial relation, *bull* is considered its *trajector* and *meadow* its landmark. For the purpose of recognizing these aspects, we trained the model with our training dataset which included 35 spatial relations. We first identify tokens as landmarks,

using the same CRF model we used for character recognition. Then, together with the characters identified in Section 3.4 (as trajectors) and trigger words (such as: in, to) as spatial indicators, we construct candidate spatial relations which we then classify as valid or not using the method described in the second part of [13].

To expand on this, consider the example sentence above: *the bull* is annotated as a character and can serve as a trajector (labelled as tr below) by our character identification process and *meadow* as a place (lm) by the model from [13]. We also have the word *in* which acts as an indicator (ind, implies there is a possibility of a spatial relation). So the sentence can be seen as:

"[A bull]$_{tr}$ was feeding [in]$_{ind}$ a [meadow]$_{lm}$"

The candidate relation we extract can be expressed as a triplet $< tr, ind, lm >$. In our case:

$$< \text{A bull}, \text{in}, \text{meadow} > \quad (1)$$

We then extract features for this triplet and predict a relation label for it by using the SVM classifier described in [13]. The classifier will label it as either being SPATIAL or None.

One could ask whether simple NER instead of spatial relation extraction, or even use of simple dictionaries to recognize tokens that relate to environments, could serve the same purpose. We are interested however in landmarks that are related to our characters, and only them. Consider the following quote:

> "A [Woodman]$_{tr}$ was felling a tree [on]$_{ind}$ the [bank of a river]$_{lm}$, when his [axe]$_{tr}$, glancing off the trunk, flew out of his hands and fell [into]$_{ind}$ the [water]$_{lm}$."

We identify one character, the *Woodman*, and two possible environments, the *bank of a river*, and the *water*. If we were to associate an environment for subsequent use, we could also end up with *water* since it also appears in the text. We can solve this issue by figuring out spatial relations with characters. By doing this, we notice that only the *bank of a river* is eligible as an environment. After each environment in the story is identified, it is assigned to a separate scene number in our play.

3.6 Acting Line assignment

After tagging the acting lines (Section 3.2) and identifying the Characters in the story (Section 3.4), we need to identify who speaks when. This is done in a similar manner as Section 3.5 but instead of spatial indicator trigger words, we detect words that relate to speaking and instead of landmarks, we use the *acting lines* as extracted in Section 3.2. Trigger words related to speaking can either be identified with our CRF model, or extracted with high recall by matching the lemma of a word with a known word related to speaking. As an example we annotate the elements of the sentence below as such:

"[*<CLINE1>*]$_{al}$ [*said*]$_{sw}$ [*the mouse*]$_{ch}$ "

```
Cast List:
Narrator - male or female  -- panned center
Young Mouse - male - panned left
Old Mouse - male or female - panned right

Scenes:
1 - room - room.wav - clearer

Script:
-- Scene 1 --
(...)
[Young Mouse] By this means (...)
[Narrator] This proposal  (...)
(...)
```

Figure 4: Excerpt from a generated production script

Here al denotes tagged acting lines, sw (stands for *sayword*) a synonym to *saying* and ch a character. To determine whether a character says something, we create candidate relations of all characters, saywords, and acting lines, and classify them as valid using the same SVM classifier as in section 3.5.

4. SOUND PRODUCTION

After semantic analysis is completed, a production script is created which contains a character list, a scenes list and a timeline of acting lines. The script is targeted at an amateur human radio-drama producer. The radio producer can change any of the elements presented in the script and proceed to further produce their own play, or feed it back into the system in order for a play to be automatically generated. An excerpt of such a script can be seen in Fig. 4.

4.1 Media Content Retrieval

After the environments are detected and assigned to scenes, a sound effect from a local sound library is assigned to each based on the text of the detected environment. While the play is on that particular scene, that sound is looped at a lower volume level. For character voices we allow one of three different methods:

1. Assign voices based on a line-to-audio dictionary.
2. Populate the dictionary using a speech recognition system.
3. Synthesize the voices.

For (1), we use a dictionary that maps acting lines to sound files containing character speech. Those files can be recorded in advance by the user. For (2), we use the DEEPSPEECH [2] speech recognition system to convert speech recorded by the user to text and match it against the acting lines based on a string similarity measure. And for (3), we use information about gender to select an appropriate voice for the FESTIVAL speech synthesis engine and synthesize the acting lines with this voice.

[2] https://github.com/mozilla/DeepSpeech

Label	p	r	f_1
character (NC)	0.944	0.648	0.768
character-gender (NC)	0.548	0.381	0.449
character (NC+H)	0.950	0.724	0.822
character-gender (NC+H)	0.800	0.648	0.716
character (CRF+H)	0.955	0.800	0.870
character-gender (CRF+H)	0.303	0.419	0.352

Table 1: Results for the character recognition task on the test set. (NC) are the results obtained by using NEURALCOREF, (CR+H) are the results obtained by also using a dictionary of known names and heuristics, and (CRF+H) are the results obtained by the CRF model, again augmented by heuristics.

4.2 Mixing & Mastering

The effects used during mixing are *panning* and *reverberation* and are only applied on the acting lines. Narrator is panned to the center and no reverberation is applied to their lines. The characters are hard panned to the left and right based on their order of appearance, to clearly position them in space relative to the listener and give the impression of a dialog happening between them.

For reverberation we used the method given in [14]. They provide a mapping from text descriptors (such as dry, clean, underwater) to reverberation effect parameters. We match those descriptors to our environments by using a simple dictionary and we apply the effect on the acting lines, similar to the way we applied panning. For mastering, a $80Hz$ highpass filter was used and the final mixdown normalized at $-9dB$.

5. EVALUATION & RESULTS

Our method consists of a cascade of different sub-tasks, which allows us to evaluate each task independently. We therefore give the evaluation methodology and results separately for each subtask evaluated.

5.1 Semantic Analysis

Semantic analysis tasks are evaluated using precision, recall and f1-score metrics for the tasks of character identification, identifying gender, and assigning characters to acting lines and to environments. Precision, recall, and f_1 scores are calculated as:

$$p = \frac{d_r}{d_r + d_n}, \quad r = \frac{d_r}{d_r + d_R}, \quad f_1 = 2\frac{p \times r}{p + r}, \qquad (2)$$

where d_r is the number of *relevant* and *retrieved* documents, d_n the number of *retrieved* but not relevant, d_R the number of relevant documents that were not retrieved, and d_N the number of non-relevant documents that were not retrieved. In our case, a document refers to the elements we want to identify (e.g. Characters, Landmarks and their spatial relations). Table 1 shows the result for the task of identifying story characters and their attributes. We note that while the CRF model designed for character identification gives a higher f_1 score than NEURALCOREF aided by heuristics, it struggles to correctly extract gender attributes (much lower precision, recall, and f_1 scores). To compensate for that we can do character identification with our CRF model, and gender identification with NEURALCOREF.

Label	p	r	f_1
SAYS (CV)	0.957	0.917	0.936
SPATIAL (CV)	0.909	0.714	0.800
SAYS	0.698	0.857	0.769
SPATIAL	0.633	0.383	0.477

Table 2: Results for the relation extraction task. Top (CV) are results on our validation set, and bottom the results on the test set. SAYS is the relation assigning acting lines to characters, and SPATIAL is the relation assigning characters to landmarks.

ID	CHARA	PAN	REVB	SFX
0000				
0011				✓
1111	✓	✓	✓	✓
1011	✓		✓	✓
1101	✓	✓		✓
1110	✓	✓	✓	

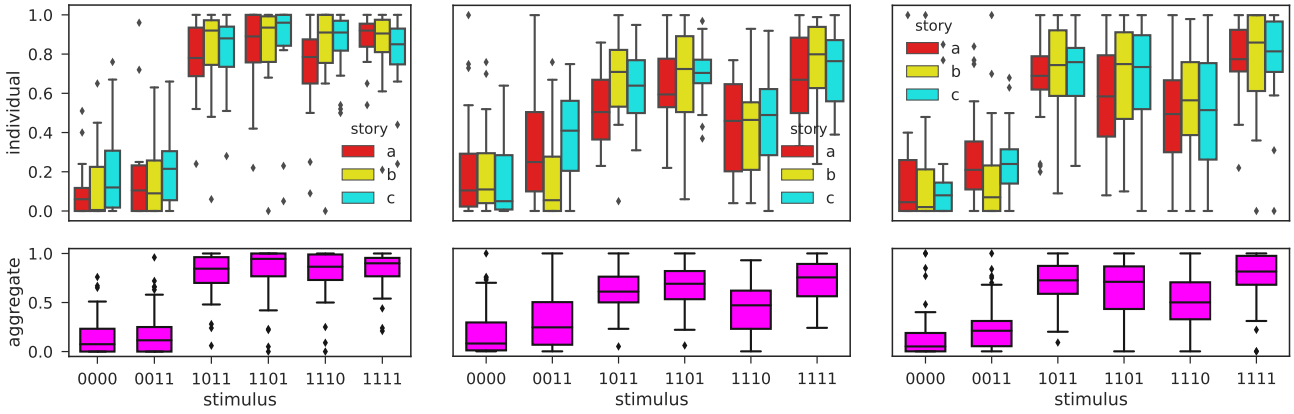
Table 3: Evaluation segments and audio story elements they represent in Fig. 5. CHARA pertains to whether the story has different character voices or not, PAN whether it contains spatial panning, REVB whether it contains reverberation and SFX whether it contains spatial sound effects.

Table 2 presents results for spatial role labeling. The bottom two rows on that table shows the result on our test set while the top two rows on our validation set. In both of them, relation extraction tasks first identify their arguments (characters, indicators, landmarks in the case of spatial relations and characters, saywords and acting lines in the case of the say relation) and then classify the relation as being SAYS, SPATIAL or neither. We suspect that the large differences are due to poor annotation of the testing set, which causes entities to be extracted a little differently (e.g. $[$the bank$]_{lm}$ instead of $[$ the bank of a river $]_{lm}$) and as a consequence hurts performance.

5.2 Sound Production

Evaluation of the produced play takes the form of a listening test. This subjective evaluation has the goal of identifying the extent to which the various parts of the sound production system contribute to story character recognition (task 1) and listener immersion (task 2), and how well they rank on the listeners' preference (task 3).

Each test was presented on 9 pages (3 stories for each of 3 tasks) implementing the MUSHRA [21] listening test environment using the Web Audio Evaluation Tool (WAET)



(a) Character Identification (b) Immersion (c) Preference

Figure 5: Box plots from the listening tests

[22]. This environment consists of samples that can be player in the browser, and each sample can be given a rating from 0 to 100, by using a vertical scroll bar. Scales at 0, 25, 50, 75, and 100 of the scroll bar are annotated according to each test. The samples were randomized in each page but the pages retained the same order across all participants. MUSHRA tests generally have a hidden anchor and reference as stimuli. Our tests have a hidden anchor but not a reference, since it is difficult to generate an objective reference for our tasks.

We gathered 21 subjects, mostly non-english speakers. Subject age was between 23-31. The subjects were also asked whether they had experience in radio/tv production, with theater and whether they were regular consumers of radio-plays/audiobooks. From the subjects we omitted one person who reported not understanding the test.

5.3 Listening Segments

Three story segments were selected, with each having a narrator and two additional story characters. The first segment had all of the characters in the story as *male*, the second as *female*, and the third had a male narrator, a male story character, and a female story character. The choice of genders were made in order to consider character recognition based on the difference in genders between different characters since we expect both gender and voices to be important factors in recognizing different characters [15]. In addition to the character voices, two environmental background sound effects were used (*a meeting room*, and *a forest*), the reverberation descriptors *clearer* and *dry* from the Reverbalize social reverberation map [14], and stereo panning. While Reverberation and Spatialization have not been found to be very high in importance while identifying spaces [15], with elements such as actions and context being higher, our segments lack action sounds and are too short to provide a context. For the rest of the evaluation section, we will refer to Character voices, Sound Effects, Reverberation, and Panning as *audio story elements*. The segments were created by combining those elements but leaving one out each time. This approach allows us to avoid introducing a "the more, the better" bias to our listeners, something that could happen if we built each segment from the previous one with one more additional element. In addition, a hidden anchor with all elements disabled and an extra segment with all the elements enabled were used. The table of listening segments and their elements can be seen in Table 3.

5.4 Character Recognition

This task pertains to how each audio story element contributes to the improvement of listener's ability to distinguish between 3 different characters. We expect however that the listener has already some cues from the story text. The question asked on the related pages was:

> "How easy does each segment make it to distinguish between the 3 characters (based on both sound and text)?"

The answers were on a continuous scale from 0 (*very hard*) to 100 (*very easy*). Full use of the scale was not required. If our system performs well, we expect the ratings for segments including character voices, to be rated much higher than the segments with just the narrator reading the text, and panning and reverberation to contribute to the ability of our system to convey character differences.

Results are shown in figure Fig 5a. Though the bars mostly overlap, stimuli with different character/gender voices are rated much higher than the ones without. This appears to agree with the observation in [15]. From between the elements with sound, there is a hint of preference towards the ones with spatial panning (1111, 1101, and 1110) compared to the one without (1011). This again seems to agree with [15].

5.5 Listener Immersion

This task pertains to how well our system can immerse the listeners in the story environment by usage of panning, reverberation, and environmental sound effects. The same segments as in Section 5.4 were used. The question asked was:

> "How easy does each segment make it to imagine yourself in the environment of the story?"

The question seeks to identify what elements act as cues to communicate the environment of the story to the listener. We expect environmental sound effects to contribute to listener immersion, and reverberation and panning to add to that contribution. The answers were again on a continuous scale from 0 (*very hard*) to 100 (*very easy*). The resulting boxplots can be seen in Fig. 5b.

As in [15], sound effects seem to be an important factor since there is a large difference between stimuli with (1011, 1101, 1111) and without (1110, 0000) sound effects. They are not the single factor for immersion though, as observed by the relative low rating of the stimulus with just sound effects (0011). Panning seemed to play a bigger role for immersion. We can attribute this to two reasons; selected reverberation was not adequate for the environment, and positioning characters using panning does indeed bring the listener in the story.

5.6 Listener Preference

The last test was a generic listener-preference test. The same segments used in the two previous subsections were used, but this time they were ranked based on how well each user preferred each one. The goal of this test was to check how much the listener liked each element. The question asked was:

> "Please listen to each segment again, how would you rank them in regards of preference?"

The answer was on a continuous scale from 0 (*very low preference*) to 100 (*very high preference*). This was the only part of the test which encouraged full use of the scale, since it was checking for relative preference and not absolute user liking. The results can be seen in Fig.5c. In this case, we can only conclude that users prefer stories with character voices and sound effects (1011, 1101, 1111) but the boxplots overlap too much to make an observation about whether they prefer panning over reverberation or vice versa.

6. DISCUSSION

We presented a method from converting unstructured story text to a production script and then to an (amateur) radio-play. We also presented evaluation results for a proof-of-concept implementation. We believe that such research would lead to development of tools for assisting radio drama or relevant format production and thus make the format more accessible for non-professionals. For example our prototype [3] reduces the time for producing a play by splitting the production process into distinct tasks and presenting the user only the ones where they must intervene. During the progress of our work many obstacles surfaced both for semantic analysis and for production of the play. A major issue, which is also an issue for many NLP tasks, is that the system we described relies heavily on the quality of annotations. Due to lack of resources, our annotation corpora remained quite small and of poor quality. This hindered both training of our models and their evaluation. A more rigid corpus construction attempt must be performed. A related issue is that we implemented our proof-of-concept using input only from Aesop Fables, which restricts it to this domain. Future work could be made to tackle more broad corpora such as the one introduced in [23], which has characters and dialogs in an easily exploitable format. Another limitation of our work is due to the small number of valid test samples which forbid us to make rigid observations. While this was a preliminary study effort, we plan for tests of larger sample sizes in the future.

In our effort to produce a radio-play, we deliberately left out questions relating to mixing and mastering for radio. Instead we focused on utilizing audio effects to convey parts of the story while leaving out other important effects such as equalization. We addressed whether panning and reverberation are needed to convey information and not how much of it is needed. Automatically applying audio effects in multi-track mixes has been tried for Panning [24], Equalization [25], Reverberation [26], and Dynamic Range Compression [27], and while the the context of such works are usually multitrack music mixes, one would expect them to be easily applicable to radio plays.

A very obvious limitation to our work remains the fact that radio plays contain more elements than what we tried to control. They could contain sound effects pertaining to actions or states of characters, and even music. We will examine those elements in a future work.

Finally, the method we presented could be combined with an automatic story generation system to generate an automatic storyteller. An example would be the work done in [4] where a human sci-fi author worked alongside a generative model to compose joint human/machine sci-fi stories. In a similar fashion, the work could be combined with our previous work on automatic story generation [28]. In that work, we proposed a method for a system to learn to tell coherent stories given short story segments and arbitrary user defined criteria. We could combine that method with our system, with the addition of including sound-related elements to the story segments, and user criteria that related to the radio play as a whole (such as time constraints).

Acknowledgments

This research has been supported by RPPTv Ltd. We would like to thank Beici Liang for her help in gathering participants for the listening tests, the participants, and everyone participating in the reviewing process.

7. REFERENCES

[1] A. Papadopoulos *et al.*, "Assisted lead sheet composition using FLOWCOMPOSER," in *Proceedings of the International Conference on Principles and Practice of Constraint Programming*, 2016.

[2] P. Pestana and J. Reiss, "Intelligent audio production strategies informed by best practices," in *Proceedings*

[3] `https://code.soundsoftware.ac.uk/projects/chourdakisreiss2018smc`

of the 53rd International Audio Engineering Society Conference: Semantic Audio, 2014.

[3] H. G. Oliveira, "O poeta artificial 2.0: Increasing meaningfulness in a poetry generation twitter bot," in *Proceedings of the INLG 2017 Workshop on Computational Creativity in Natural Language Generation*, 2017, pp. 11–20.

[4] E. Manjavacas *et al.*, "Synthetic literature: Writing science fiction in a co-creative process," in *Proceedings of the INLG 2017 Workshop on Computational Creativity in Natural Language Generation*, 2017.

[5] A. Chang *et al.*, "Text to 3d scene generation with rich lexical grounding," in *Proceedings of the International Joint Conference on Natural Language Processing*, 2015.

[6] D. Grba, "Avoid setup: Insights and implications of generative cinema," *Journal of Science and Technology of the Arts*, vol. 9, no. 1, 2017.

[7] T. Crook, *Radio drama: theory and practice*. Psychology Press, 1999.

[8] A. Groza and L. Corde, "Information retrieval in falktales using natural language processing," in *Proceedings of the 11th International Conference on Intelligent Computer Communication and Processing*, 2015.

[9] K. Clark and C. D. Manning, "Improving coreference resolution by learning entity-level distributed representations," in *Proceedings of the 54th Annual Meeting of the ACL*, Berlin, Germany, 2016.

[10] P. Kordjamshidi, M.-F. Moens, and M. van Otterlo, "Spatial role labeling: Task definition and annotation scheme," in *Proceedings of the 7th conference on International Language Resources and Evaluation*, 2010.

[11] J. Pustejovsky *et al.*, "Semeval-2015 task 8: Spaceeval," in *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, 2015.

[12] K. Roberts and S. M. Harabagiu, "Utd-sprl: A joint approach to spatial role labeling," in *First Joint Conference on Lexical and Computational Semantics*, 2012.

[13] E. Nichols and F. Botros, "Sprl-cww: Spatial relation classification with independent multi-class models," in *9th International Workshop on Semantic Evaluation*, 2015.

[14] P. Seetharaman and B. Pardo, "Crowdsourcing a reverberation descriptor map," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014.

[15] M. Lopez, "Perceptual evaluation of an audio film for visually impaired audiences," in *138th Audio Engineering Society Convention*, 2015.

[16] P. Stenetorp *et al.*, "Brat: a web-based tool for nlp-assisted text annotation," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012.

[17] D. Suciu and A. Groza, "Interleaving ontology-based reasoning and Natural Language Processing for character identification in folktales," in *10th International Conference on Intelligent Computer Communication and Processing*, 2014.

[18] K. Clark and C. D. Manning, "Deep reinforcement learning for mention-ranking coreference models," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, November 2016.

[19] N. Okazaki, "CRFsuite: a fast implementation of conditional random fields (CRFs)," 2007. [Online]. Available: http://www.chokkan.org/software/crfsuite/

[20] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

[21] R. S. ITU, "Recommendation bs. 1534-2: Method for the subjective assessment of intermediate quality level of audio systems," June 2014.

[22] N. Jillings *et al.*, "Web audio evaluation tool: A framework for subjective assessment of audio," in *Proceedings of the 2nd Web Audio Conference*, April 2016.

[23] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 2011.

[24] E. Perez Gonzalez and J. D. Reiss, "A real-time semi-autonomous audio panning system for music mixing," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 1, 2010.

[25] S. Hafezi and J. D. Reiss, "Autonomous multitrack equalization based on masking reduction," *Journal of the Audio Engineering Society*, vol. 63, no. 5, pp. 312–323, 2015.

[26] E. T. Chourdakis and J. D. Reiss, "A machine-learning approach to application of intelligent artificial reverberation," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, 2017.

[27] Z. Ma *et al.*, "Intelligent multitrack dynamic range compression," *Journal of the Audio Engineering Society*, vol. 63, no. 6, 2015.

[28] E. T. Chourdakis and J. D. Reiss, "Constructing narrative using a generative model and continuous action policies," in *Proceedings of the INLG 2017 Workshop on Computational Creativity in Natural Language Generation*, 2017.