(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
02 January 2020 (02.01.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/002941 A1

(51) International Patent Classification:
G06F 17/27 (2006.01) *G10L 13/08* (2013.01)

(21) International Application Number:
PCT/GB2019/051841

(22) International Filing Date:
28 June 2019 (28.06.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1810621.1 28 June 2018 (28.06.2018) GB

(71) Applicant: **QUEEN MARY UNIVERSITY OF LONDON** [GB/GB]; Mile End Road, London Greater London E1 4NS (GB).

(72) Inventors: **REISS, Joshua**; Electronic Engineering & Computer Science, Queen Mary University of London, Mile End Road, London Greater London E1 4NS (GB). **CHOURDAKIS, Emmanouil Theofanis**; Electronic Engineering & Computer Science, Queen Mary University of London, Mile End Road, London Greater London E1 4NS (GB).

(74) Agent: **J A KEMP LLP**; 14 South Square, Gray's Inn, London Greater London WC1R 5JJ (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*

(54) Title: GENERATION OF AUDIO DATA

(57) Abstract: Disclosed herein is a computer-implemented method for generating audio data from unstructured text data, the method comprising: receiving a first data file comprising unstructured text data; performing a semantic analysis process on the first data file to thereby determine semantic data of the unstructured text data in the first data file; generating a second data file in dependence of the first data file and the semantic data, wherein the second data file is a user readable and user editable script; inputting the second data file into a sound production process; and performing the sound production process to thereby generate audio data that is dependent on the script. Embodiments advantageously improve the generation of soundtracks in applications such as radio plays, television programs, films, advertisements, computer games, theatre productions and general prototyping (such as creating an initial pitch or advertisement).

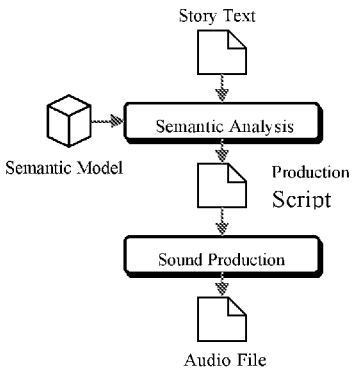

Figure 1

WO 2020/002941 A1

Generation of Audio Data

Field

The field of the invention is the generation of audio data from unstructured text data. Particularly advantageous applications of embodiments include the automatic generation of audio data for applications such as radio dramas, television programs and film soundtracks.

Background

Recent advances provide computational methods for automatically generating works. A human may also be involved the in the processes of generating the work and this results in joint human/machine works.

Such methods have been devised individually for music [1, 2], poetry [3], literature [4], 3d scene generation [5], and film [6]. Techniques for information extraction from stories mostly focus around identifying characters in stories and their social networks. For example, in [7] 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.

Coreference 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 (i.e. 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. [8] builds such clusters incrementally, starting with each mention as its own cluster.

Spatial Role Labelling [9][10] pertains to extracting information from sentences that describe some kind of spatial relation (e.g. "A bull was feeding in a meadow."). [11] 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, [12] deployed Conditional Random Fields to capture the relation constituents and use them as heuristics for relation extraction. When combined with character identification, spatial role labelling 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 [13]. The work therein 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. The mappings therein prove useful in controlling the audio effect from text input.

There is a general need to improve the computational methods and processes for all forms of automatically generated works, in particular joint human/machine works.

Summary

According to a first aspect of the invention, there is provided a computer-implemented method for generating audio data from unstructured text data, the method comprising: receiving a first data file comprising unstructured text data; performing a semantic analysis process on the first data file to thereby determine semantic data of the unstructured text data in the first data file; generating a second data file in dependence of the first data file and the semantic data, wherein the second data file is a user readable and user editable script; inputting the second data file into a sound production process; and performing the sound production process to thereby generate audio data that is dependent on the script.

Preferably, the sound production process generates audio data by a performing speech synthesis process, a sound effects retrieval process, a reverberation process, a panning process and a mastering process, all in dependence on the second data file.

Preferably, the method comprises: determining elements in the unstructured text data; determining how these elements can be used to produce the audio data, such that the generated audio data is dependent on the determined elements in the unstructured text data.

Preferably, the unstructured text data comprises characters.

Preferably, a character is an entity that performs an action.

Preferably, the semantic analysis process comprises performing a character identification process.

Preferably, the semantic analysis process comprises: performing a lines separation process on the unstructured text data; performing a co-reference resolution process on data processed by the lines separation process; performing a character identification process on data output from the co-reference resolution process; performing an environment detection process on data output from the character identification process; and performing a lines assignment process on data received from the lines separation process and the character identification process.

Preferably, the method further comprises using a Conditional Random Field, CRF, model for Named Entity Recognition, NER, to perform a character recognition process.

Preferably, the semantic analysis process comprises determining landmarks that are related to characters.

Preferably, the semantic analysis process comprises applying a co-reference resolution process to determine mappings from one or more characters to gendered pronouns so as to determine the genders of the one or more characters.

Preferably, the second data file comprises a character list, a scenes list and a timeline.

Preferably, the operation of the semantic analysis process is independent from the operation of the sound production process.

Preferably, the method further comprises outputting the second data file and/or the audio data.

Preferably, the second data file is edited by a user prior to being input to the sound production process.

Preferably, the second data file is automatically input to the sound production process without being edited by a user.

Preferably, the unstructured text data is any data to which semantic analysis can be applied in order to determine audio data related to the unstructured text data.

Preferably, the unstructured text data is a story.

Preferably, the generated audio data is audio data for a radio play, television program or film soundtrack.

According to a second aspect of the invention, there is provided a computing system that is configured to perform the method of the first aspect.

According to a third aspect of the invention, there is provided a computer program product that, when executed by a computing system, causes the computing system to perform the method of the first aspect.

List of Figures

Figure 1 is a block diagram of a system according to an embodiment;

Figure 2 is a block diagram showing the construction of a semantic model according to an embodiment;

Figure 3 is part of a script according to an embodiment; and

Figure 4 is a block diagram showing a sound production process according to an embodiment.

Description

Embodiments of the invention provide a computing system that improves the automatic generation of audio data of works. The audio data of works may be entirely automatically generated without user input, or the automatic generation of the audio data may be dependent on user input.

More particularly, embodiments provide a computer-implemented method of converting unstructured text data to a production script and then to audio data (i.e. audio content).

The method comprises of a cascade of different sub-tasks. The computer-implemented method according to embodiments is divided into two distinct processes, namely:

1) Semantic analysis; and
2) Sound production (also referred to herein as the production stage).

In preferred embodiments the semantic analysis is performed on stories. However, embodiments include performing the semantic analysis on any information/data source to which a semantic analysis can be applied in a process for inferring audio data related to the information/data source. For example, embodiments can be used to generate an audio book in dependence on a novel or an audio advertisement in dependence on a descriptive text.

The sound production process is the production of actual audio data, i.e. the audio content.

Embodiments use techniques related to information extraction from natural language text, as well as novel approaches in automatically generated audio-based storytelling.

The semantic analysis according to embodiments receives stories in an original unstructured text format and automatically generates a human readable semi-structured production script, also referred to herein as just a script, for the stories. Embodiments apply co-reference resolution to derive mappings from characters to gendered pronouns to thereby extract information about the characters' genders when not explicitly stated.

After a script has been generated, a computer-implemented production stage automatically uses the script to generate audio data.

The division into the separate stages of semantic analysis and sound production allows each stage to be automatically implemented independently of each other. The system can act autonomously to thereby automatically generate audio data from an information source. However, embodiments also include a human user providing data input between the two stages of the generation process. Embodiments therefore provide a tool that can be used to either automatically produce audio data without user intervention, or to assist a user when composing audio data.

The audio data may be a soundtrack for produced content, such as a television program, radio drama or film soundtrack, for example fixing mistakes in the production scripts, changing acting lines or attributes of characters or scenes.

Automatic processes that are advantageously performed by embodiments include inferring the elements of the story relevant to produced audio content, and establishing how these elements can be used to produce the finalised audio content.

Embodiments are described in more detail below.

Figure 1 shows a block diagram of a system according to embodiments. Embodiments comprise the two main stages of 'Semantic Analysis' and 'Sound Production'.

Input data to the system according to embodiments is shown in Figure 1 as 'Story Text' and is unstructured text data.

A semantic analysis block uses a semantic model to perform semantic analysis on the received input data to automatically generate a production script.

The production script is user readable and user editable.

A sound production block then automatically generates an audio file, i.e. audio data, in dependence on the production script.

SEMANTIC ANALYSIS

Testing data in the form of text-based stories can be annotated for semantic analysis. Embodiments perform semantic analysis of stories in order to identify key story elements that can later be used to guide sound production. A semantic model is constructed that performs 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 2 is a block diagram showing the construction of a semantic model, according to an embodiment. The semantic model may be constructed by a training process.

The Input Story block is the unstructured text data on which the semantic analysis is performed. The Acting Lines Separation block performs an acting lines separation process on the unstructured text data. The Coref. Resolution block performs a co-reference resolution process on data output from the Acting Lines Separation block. The Character Identification block performs a character identification process on data output

from the Coref. Resolution block. The Environment Detection block performs an environment detection process on data output from the Character Identification block. The Acting Lines Assignment block performs an acting lines assignment process on data received from the Acting Lines Separation block and the Character Identification block. The semantic model is generated in dependence on data received from the Acting Lines Assignment block and the Environment Detection block.

Acting Lines Separation

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

Co-reference resolution

After identifying and replacing speech lines with their tags, the stories are passed through a computer-implemented co-reference resolution algorithm. Apart from not having to deal with unresolved anaphora, it helps in three other ways, namely:

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.

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 computer-implemented co-reference resolution algorithm in embodiments may, for example, be the co-reference resolution algorithm as described in [14] that serves as an adequate baseline for character identification, or another co-reference resolution algorithm.

Character Identification

Characters, as referred to herein, are entities that do some kind of action or say an acting line. For example, consider the following two sentences:

"Jupiter and Venus were arguing..."

"A cat went to Venus"

In the first sentence, both 'Jupiter' and 'Venus' are characters since they are doing something. In the second sentence, only 'A cat' is a character since 'Venus' does not act. The steps described serve as an adequate method for recognizing characters because third person singular pronouns are usually followed by verbs. Co-reference resolution is not designed for character recognition and this can leads to some automatically identifiable mistakes, namely:

1) False positives – These are mainly inanimate non-character elements, identified as characters and leading to lower precision. An example sentence would be

"When the battle was at its height"

2) False negatives – These are 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:..."

Despite these potential errors, co-reference resolution provides an effective way to infer gender information of the characters in the stories.

To reduce co-reference resolution errors, embodiments train a Conditional Random Field (CRF) model for Named Entity Recognition (NER) in order to do character recognition. As a training set, embodiments may use annotated sentences. The features extracted may be the same as used in the CRF model in [13]. Embodiments may use this model to identify the characters in the story, and the co-reference resolution part to assign them a gender attribute. In addition to the above, a dictionary may be 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. Embodiments also include a heuristic that assigns attributes based on the character's text (for example a daughter may be automatically assigned an attribute of female and a grandfather may be automatically assigned an attribute of male).

Environment Detection

Identifying elements relating to the story environment allows embodiments 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 of generated audio data the perception that they are in that specific environment.

For example, 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.

The problem of identifying such environments is posed 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, embodiments train the model with a training dataset which includes spatial relations. Embodiments first identify tokens as landmarks, using the same CRF model used for character recognition. Then, together with the characters identified (as trajectors) and trigger words (such as: in, to) as spatial indicators, embodiments construct candidate spatial relations which are then classified 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 the character identification process of embodiments and 'meadow' as a place (lm) by the model from [13]. There is also 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:

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

The candidate relation can be extracted and expressed as a triplet < tr,ind,lm >. In the present example:

< A bull,in,meadow >

Embodiments 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.

Preferred embodiments only determine landmarks that are related to characters, and the above-techniques provide this data. Alternatively, in less preferred embodiments, simple NER instead of spatial relation extraction, or even use of simple dictionaries to recognize tokens that relate to environments, could be used to also determine environments that are not related to characters.

The advantage of only determining landmarks that are related to characters is demonstrated by 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]$_{lm}$."

It is possible to identify one character, the 'Woodman', and two possible environments, the 'bank of a river', and the 'water'. If the applied technique were to associate an environment for subsequent use, the technique could also end up with 'water' since it also appears in the text. Preferred embodiments solve this problem by determining spatial relations with characters. By doing this, the technique determines 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 the work, e.g. play.

Acting Line assignment

After tagging the acting lines and identifying the characters in the story, the techniques according to embodiments identify who speaks when. This is done in a similar manner as with spatial indicator trigger words, but instead detecting words that relate to speaking. Instead of landmarks, embodiments use the extracted acting lines. Trigger words related to speaking can either be identified with the CRF model, or extracted with high recall by

matching the lemma of a word with a known word related to speaking. As an example the sentence below has its elements annotated as:

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

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, embodiments create candidate relations of all characters, saywords, and acting lines, and classify them as valid using the same SVM classifier described above.

After the semantic analysis process on the unstructured text data has been completed, a production script is automatically created which contains a character list, a scenes list and a timeline of acting lines. The automatically generated script is in a form that can be easily read, understood and edited by a user. The script may be output to a user interface for inspection and editing by a user. This is particularly helpful for user such as a radio-drama producer. The user can change any of the elements presented in the script and proceed to further produce their own audio data for a work, or feed the script back into the system in order for audio data for the work to be automatically generated.

Embodiments also include a sound production process automatically being performed on the script, either without the script being output to the user interface or in addition to the script being output to the user interface.

An excerpt of an example of an automatically generated script according to the techniques of embodiments is shown in Figure 3.

SOUND PRODUCTION

Figure 4 is a block diagram showing the sound production process according to an embodiment. The content of the production script is used for speech synthesis, sound effects retrieval, applying reverberation and panning effects and mastering the data in order to generate audio data. Embodiments also include other techniques, such as

equalisation and dynamic range compression, being applied in the sound generation process.

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 data of the detected environment. While the play is on that particular scene, that sound is looped at a lower volume level. For character voices embodiments allow one of three different methods, namely:

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), embodiments may 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), embodiments may use the DEEPSPEECH (https://github.com/mozilla/DeepSpeech, as viewed on 27/7/2018) speech recognition system, or any other 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.

For (3), embodiments use information about gender to select an appropriate voice for the FESTIVAL speech synthesis engine, or other speech synthesis engine, and synthesize the acting lines with this voice.

Mixing and 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, embodiments may use the method given in [14]. The technique therein provides a mapping from text descriptors (such as dry, clean, underwater) to reverberation effect parameters. Embodiments match those descriptors to the determined environments by using a dictionary and embodiments apply the effect on the acting lines, similar to the way applied for panning. For mastering, a 80Hz highpass filter was used and the final mixdown normalized at –9dB.

Accordingly, embodiments provide the automatic generation of audio data from unstructured text data.

An example of an embodiment is provided below.

The input data to the system according to an embodiment is the following version of Aesop's fable, 'The Lion and the Mouse':

> 'Once when a Lion was asleep a little Mouse began running up and down upon him; this soon wakened the Lion, who placed his huge paw upon him, and opened his big jaws to swallow him. "Pardon, O King," cried the little Mouse: "forgive me this time, I shall never forget it: who knows but what I may be able to do you a turn some of these days?" The Lion was so tickled at the idea of the Mouse being able to help him, that he lifted up his paw and let him go. Some time after the Lion was caught in a trap, and the hunters who desired to carry him alive to the King, tied him to a tree while they went in search of a waggon to carry him on. Just then the little Mouse happened to pass by, and seeing the sad plight in which the Lion was, went up to him and soon gnawed away the ropes that bound the King of the Beasts. "Was I not right?" said the little Mouse.'

The input data is an example of unstructured text data. In the present embodiment, the unstructured text data is a short story with sufficient information to indicate the main information and plot elements to be conveyed. Other examples of unstructured text data include condensed or summarised stories, novels, chapters of a novels and other forms of descriptive text.

Unstructured text data is fundamentally different from structured text data. An example of structured text data is a screenplay. A screenplay presents information according to a known structure that clearly identifies who is talking, what background sounds are occurring and other details, and this information from the structure of the input data can be used in the construction of an audio environment. When the input data to a system is unstructured text data, for example a short story, the input data is not presented in such a structure. Information for generating an audio environment is therefore not provided by a structure of the input data. Accordingly, when determining information for generating an audio environment, unstructured text data is processed in a fundamentally different way from structured text data.

The input data of systems according to embodiments is unstructured text data and there are therefore substantially no restrictions on the structure of the input data. There are also substantially no restrictions on the processing required, such as the sounds or actions that should be rendered and the total length of the scene when rendered.

The present embodiment includes performing a semantic analysis on the above example of input data.

An acting lines separation process identifies and tags the spoken parts: “Pardon, O King,” “forgive me this time, I shall never forget it: who knows but what I may be able to do you a turn some of these days?” and “Was I not right?”. The remainder of the text is assigned to the narrator.

A co-reference resolution process establishes characteristics for the characters, e.g., the mouse 'cried'. Almost all references to the mouse note that he is 'little', and so that aspect is re-enforced.

A character identification process is performed for the mouse. Since the mouse is 'little', it may be assigned a child's voice. The Lion is referred to as 'King' and 'King of the beasts', so may be assigned an adult, male voice.

An environment detection process determines that the scenes involve animals, hunters and a tree, indicating that it is outside and in a natural environment.

An acting line assignment process determines that all spoken text is assigned to the mouse.

A sound production process is then performed. A sound effect retrieval process comprises first identifying text with associated sounds; such as a roar when the Lion is first introduced or when it is trapped, a thump for placing the paw on the mouse, and the sound of gnawing at ropes. Such sounds are readily available from sound effect libraries, and may be retrieved if they are labelled with metadata.

An audio effects process may be performed that positions the characters and narrator at different locations, that may be in a stereo field or an immersive environment, and applies reverberation appropriate to the environment.

Embodiments therefore provide a tool that can be used in the development radio dramas, audio books, audio advertisements and other such works. Due to the computer-implemented techniques of embodiments, the audio data may be generated based on more appropriate determinations that a human would make as well as the audio data being generated faster and more energy efficiently than human generated audio data.

In addition, the time required for generating audio data is reduced due to the splitting the processes into distinct tasks. The generation of a user readable and editable production script

advantageously provides an easy and efficient way for a user to edit the audio data, with it being possible for the user to only edit specific parts of the audio data.

Embodiments are able to generate audio data from a broad corpora, such as the one introduced in [15], which has characters and dialogs in an easily exploitable format.

Accordingly, embodiments are directed towards the production of audio data from unstructured text data. A particularly advantageous application of embodiments is the automatic generation of appropriate audio data for stories. Embodiments automatically generate audio data, such as sound effects, music and/or speech, for stories that may be the basis of radio dramas, television programs, film soundtracks, audio advertisements and any other application that requires audio data. Embodiments provide a modular and multistage approach to computer-implemented audio data generation that allows sounds to be sourced and edited from modular components. The computer-implemented method of embodiments receives as an input a story in the form of unstructured text data. A production script for the story is automatically generated. Appropriate audio data for the story can be generated from the production script.

Embodiments include a number of modifications and variations of the techniques as described above.

Embodiments include the spatial positioning of sound sources being in the stereo field. Embodiments also include sounds being rendered in an immersive environment so that sound sources may be positioned behind, above and/or below the listener.

Embodiments include the sound generation processes retrieving sound effects from a local sound library. Embodiments also include the sound generation processes alternatively, or additionally, retrieving sound effects from online libraries. For example, sound effects for text that has been tagged may be obtained from 'freesound.org'. Embodiments also include the sound generation processes using procedural audio techniques to generate sound effects. When procedural audio techniques are used, parameters for the sound effect generation may be obtained based on semantic information extracted from the text.

Embodiments include the individual sound components (i.e. each sound effect, each line of dialogue, every musical excerpt) being stored as object-based audio that may be re-rendered for different conditions.

The input data to the system according to embodiments is unstructured text data. The input data may be a story or any other type of descriptive text.

Embodiments generally include the application of semantic and sound production process for generating semantic dependent audio data. Embodiments include semantic process being performed on other data sources that stories, such as records of communications, or intercepted communications, in order for audio data for re-creating the environment of the communication to be generated.

Embodiments also include applying the process to more elements. For example, the generated audio data could contain sound effects related to actions or states of characters, and/or music.

Embodiments can 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.

Embodiments can be combined with previous work on automatic story generation [16] that proposed a method for a system to learn to tell coherent stories given short story segments and arbitrary user defined criteria. Embodiments can be combined with that method, 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).

Embodiments can be applied in any application in which a soundtrack is generated in dependence on an actual description, or inferable description, of the scenario of the soundtrack, such a story, an account of events, script, etc. The soundtrack may comprise speech, music and/or sound effects. Applications include the generation of radio plays,

audio books, television programs, films, audio advertisements, computer games, theatre productions and general prototyping (such as creating an initial pitch or advertisement).

The flow charts and descriptions thereof herein should not be understood to prescribe a fixed order of performing the method steps described therein. Rather, the method steps may be performed in any order that is practicable. Although the present invention has been described in connection with specific exemplary embodiments, it should be understood that various changes, substitutions, and alterations apparent to those skilled in the art can be made to the disclosed embodiments without departing from the spirit and scope of the invention as set forth in the appended claims.

Methods and processes described herein can be embodied as code (e.g., software code) and/or data. Such code and data can be stored on one or more computer-readable media, which may include any device or medium that can store code and/or data for use by a computer system. When a computer system reads and executes the code and/or data stored on a computer-readable medium, the computer system performs the methods and processes embodied as data structures and code stored within the computer-readable storage medium. In certain embodiments, one or more of the steps of the methods and processes described herein can be performed by a processor (e.g., a processor of a computer system or data storage system). It should be appreciated by those skilled in the art that computer-readable media include removable and non-removable structures/devices that can be used for storage of information, such as computer-readable instructions, data structures, program modules, and other data used by a computing system/environment. A computer-readable medium includes, but is not limited to, volatile memory such as random access memories (RAM, DRAM, SRAM); and non-volatile memory such as flash memory, various read-only-memories (ROM, PROM, EPROM, EEPROM), magnetic and ferromagnetic/ferroelectric memories (MRAM, FeRAM), phase-change memory and magnetic and optical storage devices (hard drives, magnetic tape, CDs, DVDs); network devices; or other media now known or later developed that is capable of storing computer-readable information/data. Computer-readable media should not be construed or interpreted to include any propagating signals.

Incorporation by reference

The entire disclosures of all of the following documents are incorporated herein by reference:

[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 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] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

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

[14] 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.

[15] 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.

[16] 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.

Claims:

1. A computer-implemented method for generating audio data from unstructured text data, the method comprising:

receiving a first data file comprising unstructured text data;

performing a semantic analysis process on the first data file to thereby determine semantic data of the unstructured text data in the first data file;

generating a second data file in dependence of the first data file and the semantic data, wherein the second data file is a user readable and user editable script;

inputting the second data file into a sound production process; and

performing the sound production process to thereby generate audio data that is dependent on the script.

2. The computer-implemented method according to claim 1, wherein the sound production process generates audio data by a performing speech synthesis process, a sound effects retrieval process, a reverberation process, a panning process and a mastering process, all in dependence on the second data file.

3. The computer-implemented method according to claim 1 or 2, wherein the method comprises:

determining elements in the unstructured text data;

determining how these elements can be used to produce the audio data, such that the generated audio data is dependent on the determined elements in the unstructured text data.

4. The computer-implemented method according to any preceding claim, wherein the unstructured text data comprises characters.

5. The computer-implemented method according to claim 4, wherein a character is an entity that performs an action.

6. The computer-implemented method according to claim 4 or 5, wherein the semantic analysis process comprises performing a character identification process.

7. The computer-implemented method according to claim 4, or any claim dependent thereon, wherein the semantic analysis process comprises:

performing a lines separation process on the unstructured text data;

performing a co-reference resolution process on data processed by the lines separation process;

performing a character identification process on data output from the co-reference resolution process;

performing an environment detection process on data output from the character identification process; and

performing a lines assignment process on data received from the lines separation process and the character identification process.

8. The computer-implemented method according to claim 7, the method further comprising using a Conditional Random Field, CRF, model for Named Entity Recognition, NER, to perform a character recognition process.

9. The computer-implemented method according to claim 4, or any claim dependent thereon, wherein the semantic analysis process comprises determining landmarks that are related to characters.

10. The computer-implemented method according to claim 4, or any claim dependent thereon, wherein the semantic analysis process comprises applying a co-reference resolution process to determine mappings from one or more characters to gendered pronouns so as to determine the genders of the one or more characters.

11. The computer-implemented method according to any preceding claim, wherein the second data file comprises a character list, a scenes list and a timeline.

12. The computer-implemented method according to any preceding claim, wherein the operation of the semantic analysis process is independent from the operation of the sound production process.

13. The computer-implemented method according to any preceding claim, the method further comprising outputting the second data file and/or the audio data.

14. The computer-implemented method according to any preceding claim, wherein the second data file is edited by a user prior to being input to the sound production process.

15. The computer-implemented method according to of claims 1 to 13, wherein the second data file is automatically input to the sound production process without being edited by a user.

16. The computer-implemented method according to any preceding claim, wherein the unstructured text data is any data to which semantic analysis can be applied in order to determine audio data related to the unstructured text data.

17. The computer-implemented method according to any preceding claim, wherein the unstructured text data is a story.

18. The computer-implemented method according to any preceding claim, wherein the generated audio data is audio data for a radio play, television program or film soundtrack.

19. A computing system that is configured to perform the method of any preceding claim.

20. A computer program that, when executed by a computing system, causes the computing system to perform the method of any of claims 1 to 18.

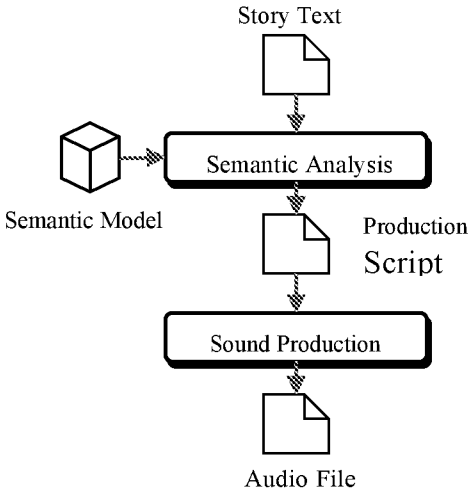
Story Text
Semantic Analysis
Semantic Model
Production
Script
Sound Production
Audio File

Figure 1

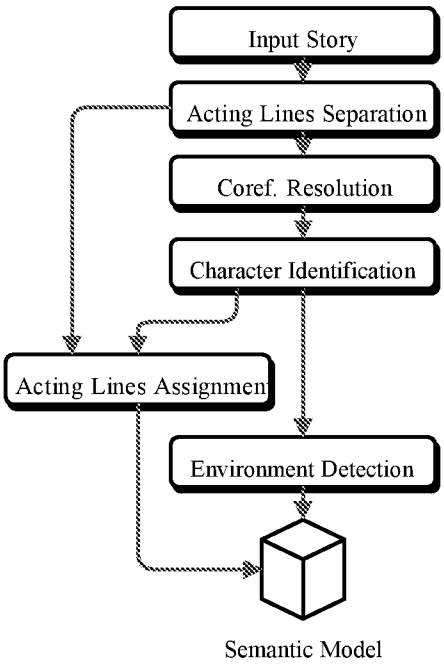
Input Story
Acting Lines Separation
Coref. Resolution
Character Identification
Acting Lines Assignmen
Environment Detection
Semantic Model

Figure 2

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 3

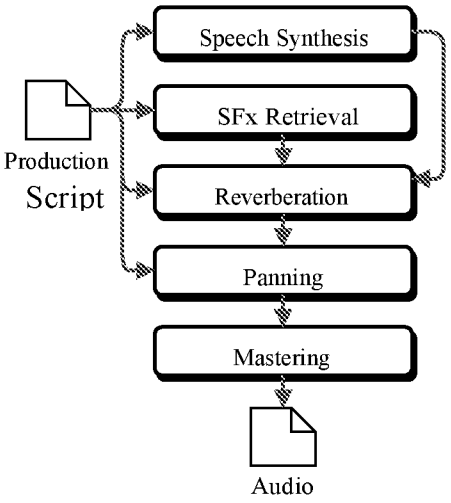
Speech Synthesis
SFx Retrieval
Production
Script
Reverberation
Panning
Mastering
Audio

Figure 4

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2019/051841

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/27 G10L13/08
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 99/66496 A1 (ONLINE ANYWHERE [US]) 23 December 1999 (1999-12-23) page 7 - page 15; figure 1	1-20
Y	US 2007/055527 A1 (JEONG MYEONG-GI [KR] ET AL) 8 March 2007 (2007-03-08) paragraph [0029]	1-20
A	US 2003/163314 A1 (JUNQUA JEAN-CLAUDE [US]) 28 August 2003 (2003-08-28) paragraph [0011] - paragraph [0015]	1-20

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 September 2019	02/10/2019

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Abram, Robert

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2019/051841

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CECILIA OVESDOTTER ALM ET AL: "Emotions from text", HUMAN LANGUAGE TECHNOLOGY AND EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, N. EIGHT STREET, STROUDSBURG, PA, 18360 07960-1961 USA, 6 October 2005 (2005-10-06), pages 579-586, XP058318305, DOI: 10.3115/1220575.1220648 the whole document -----	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2019/051841

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9966496 A1	23-12-1999	AT 336775 T	15-09-2006
		AU 4681699 A	05-01-2000
		BR 9911315 A	15-01-2002
		DE 69932819 T2	16-08-2007
		EP 1086450 A1	28-03-2001
		JP 2002518711 A	25-06-2002
		US 6446040 B1	03-09-2002
		WO 9966496 A1	23-12-1999
US 2007055527 A1	08-03-2007	KR 20070028764 A	13-03-2007
		US 2007055527 A1	08-03-2007
US 2003163314 A1	28-08-2003	NONE	