E-025

視覚障がい者のための物語テキストの自動朗読システム
Radio Drama Generation System for Visually Impaired People

金子 つばさ† 吉田 有里‡ 田村 直良*
Tsubasa Kaneko Yuri Yoshida Naoyoshi Tamura

1. はじめに

本研究は、視覚障がい者への娯楽提供を想定し、物語テキストから朗読音声を自動生成するシステムの構築を目的とする。音楽や娯楽などの生活の基盤に直接関わらない領域は、障がい者へのボランティア活動の対象としてはとかく軽視されがちであったが、発達したインターネットや情報技術により、障がい者はボランティアを介さずに気兼ねなくそれらを扱える可能性が開けてきた。

自動朗読システムとしては、既に様々なアプリケーションが開発されており、合成音声の品質も年々向上し聴き取りやすい音声になってきている。しかし、DAISY コンソーシアム[1]と日本マイクロソフト株式会社が提供しているコンテンツ自動作成システム DAISY Translator [2]などでは、単一話者による単純なテキストの読み上げに留まっており、物語テキストのようなエンタテイメント性の高い文章を読み上げるには、このような読み上げでは物足りない。

本研究では、物語の音声表現としてラジオドラマに着目し、ラジオドラマのような音声を生成する自動朗読システムを目指す。本システム（以下、ラジオドラマ生成システム）では、物語テキストの原文を読み上げることを目的としており、文生成や要約等によってラジオドラマの脚本を作るようなことはしない。

本稿ではまず、ラジオドラマの構成をモデル化し、実験・評価することによりその有効性を示す。さらに、視覚障がい者の利用を考慮したラジオドラマ生成システムの構想と概要について述べる。

2. ラジオドラマのモデル化

2.1 音声要素

一般に、ラジオドラマの音声は、内容文の読み上げ音声や背景音楽、効果音など、複数の音声要素によって構成される。本モデルでは、ラジオドラマは表 1 の音声要素から構成される。

2.2 意味要素

本モデルでは、実音声である音声要素と別に、意味的なまとまりに着目した意味要素を定義する。意味要素は音声要素を含み、物語の各場面や文間のポーズなど、朗読音声の生成をする際に必要な情報を持つ。本モデルでは、ラジオドラマは表 2 の意味要素から構成される。

† 横浜国立大学 Yokohama National University
‡ 日本ビジネスシステムズ株式会社 Japan Business Systems, Inc.
* 横浜国立大学大学院環境情報研究院 Graduate School of Environment and Information Science, Yokohama National University

Title	物語のタイトルや著者など、書誌情報を読み上げる音声
Text	物語の地の文や発話文など、内容文を読み上げる音声
BGM	場面の雰囲気を形成する背景音楽
SE	オオカミの遠吠えやガラスの割れる音など、単発的に発せられる効果音
BGS	雨音や足音など、持続的に発せられる効果音

表 1：音声要素

Radio Drama		
Biblio	書誌情報に関する意味まとまり	
Act	物語の場面に関する意味まとまり	
	Utterance	文を単位とした意味まとまり
	Pause	Utterance 間のポーズ
	Beginning	Act の発話開始までのポーズ
	Ending	Act の発話終了後のポーズ
Scene	Act 内の BGS に関する意味まとまり	

表 2：意味要素

2.3 ラジオドラマのモデル

ラジオドラマのモデルは以下のように構成される。

［**Radio Drama**］ラジオドラマ全体を表し、単一の Biblio と 1 つ以上の Act を必ず持つ。また、複数の Scene を持つことがある。（図 1 参照）

［**Biblio**］書誌情報に関する意味要素であり、音声要素 Title を持つ。

［**Act**］文脈上もっとも大きな単位として分けられる意味要素であり、物語の場面に相当する。Act は自身にふさわしい音声要素 BGM を持つことができる。また、複数の Utterance と Pause を持ち、単一の Beginning と Ending を持つことができる。

Utterance は物語文中の 1 文（“。”や“「”、“」”で区切られる）に対応し、音声要素 Text を持つ。また必要な場合はその文に付随する音声要素 SE を 1 つ以上持つ。文にはセリフや地の文によって、ふさわしい話者が設定される。話者には、性別やタイプ（人間・ロボット・動物等）、年齢（幼少・大人・老人等）の特徴プロファイルが設定される。また、SE は文に対して「同時に鳴らす」や「先に鳴らす」、「一部重ねる」等の付与位置を設定する。Beginning と Ending は、一定の長さのポーズを持ち、効果的な始まりや余韻をもたらす効果を持つ。

［**Scene**］Act 内の Utterance から Utterance 間を単位とする意味要素であり、音声要素 BGS を持つ。Scene 同士は一部や全体が重なり合うことがある。

2.4 レベルの設定

実現されるラジオドラマの品質、構成要素の種類、構造の複雑さにより、モデルに 0〜4 の 5 段階のレベルを設定する。

Level 0. 単一話者による物語文すべての読み上げ。従来の読み上げに相当する。読み上げの際は文間ポーズ長を考慮する。

Level 1. Level 0 に加え、単一の話者ではなく各登場人物のセリフに適切な話者を割り当てた読み上げ。

Level 2. Level 1 に加え、場面毎にふさわしい BGM が付与される。また、読み上げ文に適した単一の SE が付与可能である。

Level 3. Level 2 に加え、BGS が付与される。BGS 間に重複（2 種の BGS が同時に発せられること）はない。SE は文に対して複数付与することが可能とする。

Level 4. Level 3 での BGS 間において重複が可能とする。

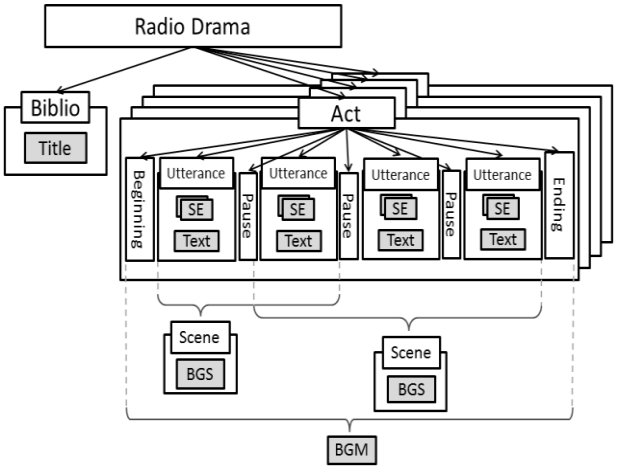

図 1：ラジオドラマモデルのイメージ（Level 4）

2.5 本モデルで表現できない状況

以下のような状況が考えられるが、本モデルでは表現できない。今後の課題とする。

- 1 文を複数話者によって同時に発話するような状況
- 右の方から聞こえるなど音の発生源の位置を考慮した状況
- 遠くからの呼びかけなど音量を考慮した状況
- 文の読み上げ速度を考慮した状況
- 登場人物の感情を考慮した状況
- 文未満の単位での音声表現

3. ラジオドラマモデルの実験・評価

3.1 実験設定

先に示したラジオドラマモデルの有効性を評価するために聴取実験を行う。

被験者は日本語話者 3 名で、物語テキストから Level 0〜4 のモデルの仕様に基づいた XML 構造データ（以下、朗読 XML）を人手で作成し、ラジオドラマ生成システムの音声部によって生成された朗読音声を聴かせる。生成した音声をレベルの低い順から聴かせ、以下の 4 つの項目について 1〜5 の 5 段階で評価させる。

1. 内容の聞き取りやすさ
2. 内容の理解しやすさ
3. 物語の読み上げとしての適切さ
4. ラジオドラマとしての自然さ

評価は値が大きいほど高評価とする。

実験に使用する物語テキストは、青空文庫[1]で公開されている『ブレーメンの町楽隊』（グリム兄弟作、楠山正雄訳）を用いた。

Level 0（単一話者による物語文の読み上げ）による出力音声をベースライン音声とし、それと比較することにより上位レベルのラジオドラマモデルの有効性を検証する。

3.2 実験準備

3.2.1 テキストデータの修正

『ブレーメンの町楽隊』の原文には、児童用のルビが「胸算用《むなざんよう》」のように書かれていることにより、音声合成時に正しい構文解析ができないことが考えられる。そのため、音声合成前にルビを消す作業を行う。

3.2.2 音声合成

音声部では、商用の音声合成器[2]を用いて発話音声を合成し、文間ポーズ長等を考慮しつつ、各音声要素を時間軸上に配置する（図 2）。SE は指定されたファイルから読み込まれ、発話音声との時間関係を考慮して時間軸上に配置される。BGM や BGS は指定されたファイルから必要な長さ分のデータを取得し、フェードイン、フェードアウトや音量調節を行った後で時間軸上に配置される。全ての音声要素は、場面ごとに飽和加算され WAV ファイルに合成される。

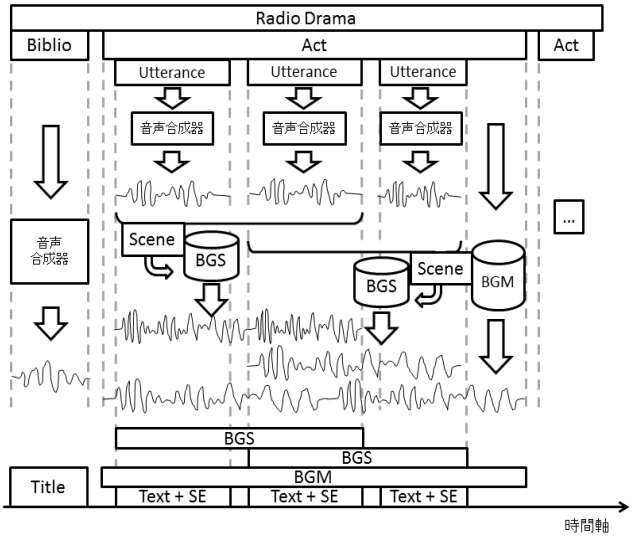

図 2：Level 4 における音声合成のイメージ

発話音声合成時の話者の声質設定に関しては、あらかじめ人物の特徴と声のパラメータを組にした役プロファイルを複数用意しておく。その中から朗読 XML より取得した話者の特徴パラメータに合った特徴を持つ役プロファイルを選択し、声パラメータを割り当てる。一度使用した役プ

[1] http://www.aozora.gr.jp/

[2] FineVoice （NTT アイティ株式会社）
FineSpeech SDK V2.2 （株式会社アニモ）

ロファイルは使用済みとし、二度以上使用されないようにする。

物語では、複数人の声質の違いを表現する必要があるが、使用した 2 種類の音声合成器では、合計で高々男性二名、女声三名しか表現できない。これに対して、各声の音声パラメータを調節することにより、実際の声数以上の話者の声を擬似的に表現する。音声パラメータの基本周波数（高・中・低）と抑揚（有・無）とを調整することにより、5 人の話者から 30 人分の話者を表現する。

3.3 評価・考察

各項目に対する評価を表 3〜6 に示す。

レベル	0	1	2	3	4
評価平均	1.6	2.3	2.7	3.7	4.0

表 3：内容の聞き取りやすさ

レベル	0	1	2	3	4
評価平均	1.6	3.7	3.7	4.0	4.3

表 4：内容の理解しやすさ

レベル	0	1	2	3	4
評価平均	2.3	3.3	3.7	3.7	4.0

表 5：物語の読み上げとしての適切さ

レベル	0	1	2	3	4
評価平均	2.0	2.7	4.0	4.3	4.7

表 6：ラジオドラマとしての自然さ

全ての項目について、レベルの上昇に伴い評価が同等か上昇している。特に、単一話者による文の読み上げに対応するモデルであるベースライン(Level 0)音声に対して、すべての上位モデル音声において評価が上回った。したがって、本稿で提案したラジオドラマモデルの有効性が示されたと言える。

4. ラジオドラマ生成システムの構想と概要

上記で示したモデルに基づき、物語テキストを入力とし、DAISY 規格のデジタル録音図書（以下、DAISY 図書）を出力とするラジオドラマ生成システムの構想を述べる。

DAISY は、視覚障がい者や印刷物を読むことに困難を伴う人々のためのデジタル録音図書の国際標準規格である[3]。DAISY 図書にはいくつかの種類があるが、本システムではその中から、音声とフルテキストを含み、各テキストとそれに対応する音声とを同期させた「マルチメディア DAISY 図書」と呼ばれる DAISY 図書を出力する。マルチメディア DAISY 図書は、テキストを含む html ファイルと音声ファイル（mp3, wav など）、テキストと音声の同期関係を記述する smil ファイル、セクション間の移動や再生の制御を記述する NCC(Navigation Control Center)ファイル（HTML 構造）から構成される。DAISY 図書は、再生用の専用機器や専用ソフトウェアを使って再生できる。

ラジオドラマ生成システムは、処理面から言語処理部と音声合成部に分けられ、後者のみ実装されている（図 3）。

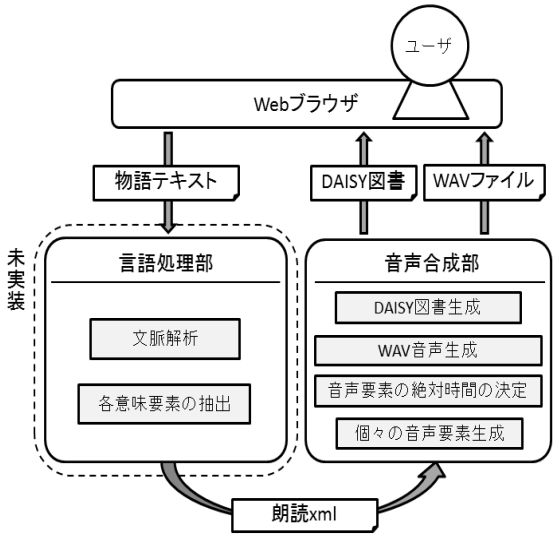

図 3：ラジオドラマ生成システムの構成

4.1 言語処理部

言語処理部では、自然言語処理技術により音声合成に必要となる情報を物語テキスト中から抽出する。抽出に当たっては、セリフの話者や場面境界の推定、文間のポーズ長の決定など、単文の集合としての物語の文脈を踏まえた文脈解析が必要になる。言語処理部は抽出結果を朗読 XML として出力し、音声合成部に送る。

4.2 音声合成部

音声合成部ではまず、朗読 XML を受け取り、3.2.2 節で示した方法で各音声要素の生成を行う。また物語テキストを含む html ファイルの生成を行う。次に、1 文とそれに対応する発話音声を同期させるための smil ファイルを生成する。最後に、先に生成した smil ファイルを基に NCC ファイルを生成する。これらをまとめて DAISY 図書として出力する。

5. おわりに

本稿では、ラジオドラマのモデル化を行い、聴取実験を行うことでその表現力の有効性を示した。また、物語テキストを入力としてラジオドラマのような音声を生成するラジオドラマ生成システムの構想について述べた。

今後の課題としては、ラジオドラマモデルの改良とラジオドラマ生成システム言語処理部の実装が挙げられる。前者については、2.4 節で触れたような、より複雑な状況を表現できるようにモデルを拡張することによって、より豊かなラジオドラマを生成することが可能になると考えられる。後者については、言語処理部において、モデルに基づいた各項目の抽出処理の実装を行う必要がある。

参考文献

[1] DAISY Consortium, http://www.daisy.org/
[2] http://www.dinf.ne.jp/doc/daisy/software/save_as_daisy.html
[3] （公財）日本障がい者リハビリテーション協会 (JSRPD), http://www.dinf.ne.jp/doc/daisy/index.html
[4] 吉田有里, 奥平康弘, 田村直良, ”音声合成による朗読システムに関する研究”, 第 8 回情報科学技術フォーラム論文集, E-051, 2009