Vol.2011-IFAT-102 No.6
Vol.2011-DD-80 No.6
2011/3/28

電子書籍の論理構造に基づくポーズ情報の推定とSSML構造化

布目光生[†1]　鈴木優[†1]　森田眞弘[†1]

電子書籍を音声合成で読み上げる場合に，文書の書式特徴を活用してより聞きやすい朗読を実現するテキスト前処理手段を提案する．文を処理の基本単位とする従来の発話手法では困難な，タイトルや箇条書きと本文を区別したような読み方や，文書全体の構成や流れを考慮したような，自然な読み上げの実現を目指す．今回，具体的なアプローチとして，入力文書テキストの特徴量として論理構造をはじめとする抽出手段と，特にポーズ情報に関連したメタデータの推定手段，そして，音声合成エンジンへ提供するためのXML化，という一連のテキスト処理機能の試作と，ポーズ情報の付与精度評価を行った．本報告では，これらの手法と評価実験結果について述べる．

Pause estimation based on e-book logical structures and SSML transformation

Kosei Fume,[†1] Masaru Suzuki[†1]
and Masahiro Morita[†1]

We conduct feasibility studies for the development of a text preprocessing technique that uses document formatting features for improved natural speech synthesis with the aim of targeting e-book readers. In traditional text-to-speech (TTS) systems, it is difficult to implement a feature by which the different document elements such as the document body, title, and itemized forms are read in a suitable tone. We implement certain functionalities, namely a sentence characteristics extractor that determines the logical nature of a document element, a metadata estimator that generates pause information, and a transformer that converts these results to speech synthesis markup language, which a TTS system can process. Details about these processes and experimental results of a simple implementation of pause estimation are described in this report.

1. はじめに

1.1 電子書籍と音声合成

従来，実用性や使い勝手の面で難があった電子書籍が，近年，Kindle(amazon) やiPad(apple) をはじめとする革新的な閲覧デバイスの発売により徐々に普及し始めてきた．

このように，現在注目されている電子書籍デバイスであるが，その差異化機能の一つに，音声合成による朗読機能がある．英文に限られるが，既にKindleにはText to Speech機能が搭載されており，米国等では従来からオーディオブックが普及している背景から，“聴く” 読書形態が不自然ではないことが窺える．

一方，国内における音声合成読上げのニーズは未知数であり，今のところは稀な利用形態であることは否めないが，今後，様々なコンテンツの普及や，ハードの一機能として音声合成読み上げが簡単に利用できるようになれば，新たな読書形態として，音声合成による朗読が受け入れられる可能性もある．

この電子書籍読み上げを実現するにあたり，デジタルドキュメント処理の観点からは大きく2つの課題がある．

- 閲覧目的のデータ形式から，読上げるべき情報を適切に抽出すること：
  現在，電子書籍フォーマットの一つとして有望視されているePub形式では，本文が事実上XHTML形式で記述されている．閲覧目的のためのレイアウト重視で自由度の高いタグ構成は，webデザイン上大きく貢献してきたことは確かだが，機械処理のためのデータ仕様として考えた場合には，その自由度が足かせとなり，従来web wrapperやモバイル用の画面変換などで取り組まれてきたものと同様の課題が存在する．
  レンダリング用のタグを駆使したり，デザイン・レイアウトを重視した複雑な文書の場合には，パージングした上でメニューリンクやバナー等のノイズ情報を除去し，必要なテキスト内容を図のまわりこみや表形式なども考慮して，読み上げ対象となるテキストエリアを正しい順序で取得する必要がある．また，人が見れば，レイアウト上でその役割が明確なタイトルやパラグラフ・箇条書き，脚注・参考などの論理構造情報も，その役割とタグ名はデザイナーの嗜好やサイトのデザインポリシーの差異などから必ずしも一意に定まっておらず，これらの差異を適切に吸収する必要がある．

†1 東芝研究開発センター
TOSHIBA CORPORATION Corporate Research & Development Center

Vol.2011-IFAT-102 No.6
Vol.2011-DD-80 No.6
2011/3/28

- 読み上げ内容に応じた聞きやすい朗読の定義と実現手段：
  近年の音声合成エンジンは，カーナビや構内アナウンス，エンターテイメントなどに見られるように人と違和感がないほどの高品質な発話を実現している例もあるが，書籍などのテキストデータをそのまま読み上げた場合には，まだまだ人の自然な発話と比べて違和感を感じることが多い．タイトルや本文との区別をはじめ，論理構造や，話者，感情表現，文書内の起承転結や語・フレーズの意味や役割とは無関係に，いつも同じ調子で読み上げられてしまうところに，そのギャップがある．これらの文脈情報のすべてを特徴量として考慮することは現実的には難しいため，効果的な特徴量にフォーカスして段階的に取り組む必要がある．

前者の課題については，文書構造解析の課題として捉え，ここでは同処理を前処理として利用するに留める．本報告では，特に後者の課題にフォーカスし，特に聞きやすさの要因として，ポーズ情報に着目した取り組みについて述べる．

1.2 従来研究

古くから，朗読におけるポーズ情報に関する分析や検討が行われてきた．なお，日本では通常 “ポーズ” といった場合に，無音区間で表現される “間” を指すことが多いが，広義にポーズは 2 種類あり，無音区間に相当する “silent pause” のほか，音声を伴う “filled pause”（フィラー）がある．本報告中での “ポーズ情報” とは，対象が電子書籍の読み上げに関係していることから主に “silent pause” を示すものとする．

ナレータの発話分析として，比企[7] がある．3 通りのテンポで発話されたアナウンサの音声を分析し，ポーズ情報の傾向は，フレーズの読み上げに直接関係するレートなどの他属性と比べてテンポの違いによる影響が大きく，またテンポとポーズ長が比例関係にあることを示した．

また，音声合成時に適切なポーズ位置を推定する手法については，海木[2] や藤尾[9]，海老原[6] らの研究がある．

海木[2] は，規則ベースによるポーズ挿入手法を提案している．話者によらないルール構築のため，10 人の朗読音声を分析して，ポーズ挿入に関わる主要因を特定し，特に文の句構造情報として，1) 当該句境界の係り受けの深さ，2) 先行句境界の右・左枝別れの相違，3) 該当句境界の読点の有無，に着目している．

また，藤尾[9] は，品詞列と係り受け構造のコーパスを用いて確率文脈自由文法を学習し，その文法から導出される単語毎のパラメータを，ニューラルネットの入力として，韻律句境界か否か，またはポーズ挿入位置であるか否かを判定している．独立した単語の属性に加えて，係り受け構造を考慮していることを特徴としている．但し，誤り分析において，係り受け構造が悪影響を及ぼしている事例を示しており，意味的なまとまりとしての係り受け構造（係り受け関係の弱い箇所）は，ポーズの挿入位置とはギャップがあることを示唆している．ポーズ位置の推定精度に関しては，未知文書に対し 85.2%の正解率を得たとしている．

海老原[6] は，文節をノードとするネットワークモデルを仮定し，各ノードの状態遷移先にポーズ状態の遷移を可能として，遷移確率を学習する方式を提案している．入力の属性（意味的特徴）として，体言，用言，その他に，助詞の表層を組み合わせた 7 種のカテゴリーを定義している．正解率はクローズで 75.2%，オープンで 74.6%としている．

こうした 90 年代の規則ベース手法とは別に，尾関[8] は統計的手法に基づき，形態素の系列から，ポーズが入る最適な場所をスコア計算で求めている．ポーズパターンを，ポーズ長を示す連続値とポーズなしを示す離散シンボルの混合系列として扱い，多空間確率分布に基づいた学習・生成を行っている．入力となる属性（コンテクスト）は，前後の品詞，文節境界，文節間の接続強度，直前のポーズの長さ，直前のポーズまでの距離，発話速度などを定義しており，ポーズの予測精度で 99.6%を得たとしている．

また，ポーズと類似した節境界の推定（チャンキング）として，西光[4] らの研究がある．これは，局所的な係り受け関係を考慮することにより，例えば係り受け関係にある二文節間を隣接させる手がかりとして用いる．具体的には，係り受け判定結果である主題や述語・格要素を “構成要素” という単位として境界候補にすることを特徴としている．構成要素による境界判定では，F 値で 0.932-0.963 を得ており，境界の種類によっては，単純に形態素列を推定元として場合よりも効果的であるとしている．

一方，応用として音声合成による朗読，という観点では吉田[10] がある．付与手法としては，まず，事前の分析から，文中に出現する表層表現で文を 13 のカテゴリに分類すると共に，文末表現についても表層から 20 のカテゴリに分類しておく．そして，入力文がこれらの特定カテゴリにマッチするか否かによって，事前に定義しておいた韻律パターンや話速パターンを付与する，というシンプルなものである．文間ポーズに関しては，3 名の被験者により「適切か/不適切か」の定性評価を実施しており，その結果，提案手法ではオリジナル音声よりも，80%以上で適切であるとの評価を得たとしている．

また，文脈を考慮したショートポーズ挿入という観点からは，太田[5] がある．形態素列に，ポーズを挿入すべきであるか (P)，そうでないか (O) というラベルを定義し，形態素に対するラベリング問題として CRF を用いることで，ショートポーズ挿入モデルを構築している．ランダム付与のモデルなどと比較して，提案手法で構築したモデルはパープレキシ

Vol.2011-IFAT-102 No.6
Vol.2011-DD-80 No.6
2011/3/28

ティが改善されたとしている．最終的な結果として，国会会議録に対する認識精度を対象とした場合，提案手法でショートポーズを挿入したモデルが，最も良い精度を得られたとしている．(単語認識精度で 59.1%)

このように，ポーズ情報は古くからその分析が行われている．また，合成においては規則ベースや統計ベースで適切な位置にポーズを挿入する試みがある．これらのポーズ挿入の手がかりとなる特徴量には，音響モデルで一般的に用いられる属性のほか，単語や品詞の Ngram や，係り受け関係を利用して挿入位置の条件とするものもある．

こうした従来手法を，今回のターゲットである書籍の朗読に適用する場合には，以下の課題がある．

(1) センテンスよりも大きな単位の文書特徴 (文書の論理構造) が考慮されない：
一般に，音声合成では，センテンスを入出力を基本単位とした処理が行われている．その結果，一文単位の発話では，ひとつの文として完結した自然な抑揚が実現されている．しかしこの前後の文脈を必要とする場合に対応できない．例えば対話インターフェースでは，これ以上のコンテクスト（文の意味役割，感情表現，文書内容による読み分け，ユーザプロファイル）を考慮しなければならない状況は多々ある．また，本報告が対象とする電子書籍では，予め文書に論理構造（タイトル，本文，箇条書きや見出し語，章節構造）がメタデータとして付けられていても，読み上げに時はこれらを区別することができずに，一定の調子で読み上げることになる．

(2) コンテンツに応じたいろいろな読み上げバリエーションが拡充できない：
同じ文書であっても，ポーズ挿入位置やポーズ長は，読み手の個性により異なる．また，同一人物であっても，文書の内容によっては，異なる間合いで読まれる．人とコンテンツのバリエーションを考慮して，既存の音響モデルや言語モデルにより，すべての可能性を標準モデルとして構築しておくことは現実的ではない．

(3) 個別のシステムに依存した制御手段のため，連携や拡張性が困難：
システム設計として，音声合成の共通コア部分と，入力のドメインや文書に応じて選択的に利用される部分を切り分け，開発者やユーザのニーズに追従できることが現実的には望ましい．しかしながら，実際のフレームワークでは，アプリ開発者から見た場合，例えば HMM の学習データとしてコンテクストもすべて素性に組み込んで，内部処理に深く依存したモデルを構築するか[1)3)]，逆に後処理として限定的パラメータで出力を制御するなどの利用形態になっている．

要素名	付与手段の例	用途例
Speak	固定要素 (root 要素) の単純変換	発話範囲指定の必須要素
Xml:lang	XML 固定要素の流用	文書全体の主要言語 (日本語/英語)
		読みの切り替えなど
Xml:base 等	アプリ側で固定的に追記	メタ情報の提供
Metadata	構造化 + 固定要素抽出	書誌情報のまとめを提供
p,s	テキスト抽出 + 構造化	行間の区切りや連結に依存する
		不自然な読み上げの解消
say-as	XSLT 変換で可能な範囲でルビを属性化	読み指定
phoneme	ルビ抽出, 外部辞書参照	正しい発音の強制指定
sub	外部辞書参照による固定表記の属性化	略称に対する正式名称
voice	セリフ文中や近傍の表層情報	発話者の変更 (性別・年齢)
		読み上げ言語の変更 (ja,en)
emphasis	原文中の強調表記の単純変換	強調
	キーワード抽出と初出表現	
break	実音声からの傾向抽出	自然な間や呼吸区間の模倣
prosody	文書ごとのパラメータセット準備と選択適用	ピッチ，大きさ，Range，Rate の指定

表 1 SSML の主要要素

2. 提案手法

2.1 ターゲットの検討

入力文書に対して，自動で読み上げ表現に貢献できるタグや属性を付与するために，ターゲットを定める必要がある．音声合成入出力の仕様を定める SSML を対象として，書籍の読み上げにおいて効果の期待できる要素を検討した．

表 1 に，要素の一覧を示す．アプリ側の作りこみや固定的なフォーマット変換で対応可能なものは除き，入力文書の表記や内容から推定が期待できてかつ出力音声の制御に効果的な要素として，break や prosody の付与や属性値の推定がターゲット候補として考えられる．

また，実際の朗読音声を分析し，発話者の違いによる傾向を概観した．

図 1 に話者の違いによる無音区間長の傾向を示す．対象は，市販のオーディオブックで社会・経済のニュース解説に関するものである．プロのナレーションによる音声と，書き起こしたテキストを，また 3 種の音声合成で出力させた音声から，無音区間長を検出し，その出現頻度の分布をグラフ化した．なお，音声合成エンジンの入力はプレーンテキストとして与えており，話速等のパラメータはデフォルト値を利用した．

人による朗読音声は，無音区間の分布がバリエーションに富んでいることがわかる．実際の音声を聞いた場合にも，タイトルと本文，本文中の文と文，箇条書きの各項目や近傍のパ

Vol.2011-IFAT-102 No.6
Vol.2011-DD-80 No.6
2011/3/28

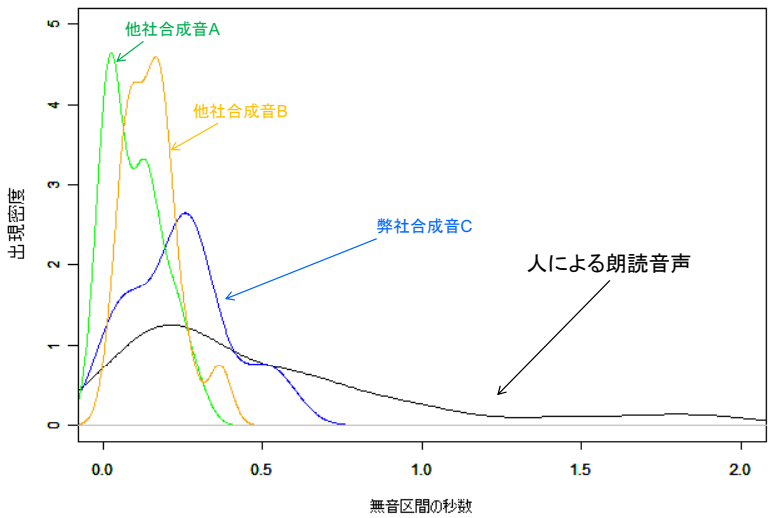

図 1 朗読音声の実例による無音区間長の傾向

ラグラフ，文書の結論部，などの前後の無音区間長が異なっていることがわかる．

一方，合成音声では，秒数の分布区間が限定されており，かつ特定のピークを有することから，句点向けと読点向けなど無音長のバリエーションも比較的限定されていることがわかる．内容や論理構造によらず，一定の調子で読み進められてしまうことは，人による朗読音声と比較した場合に聞きやすさに影響を及ぼすと考えられる．

2.2 基本方針

こうした背景のもと，本報告では以下のアプローチに基づいたフィージビリティスタディを行う．

(1) 文書の局所的な論理構造を考慮した読み分けの実現：
電子書籍に予め付与されている電子書籍のタグ情報を利用する．また，階層構造や箇条書きなどが X(HT)ML タグとして明示されていない場合にも，文書構造化処理を適用することで，これらの構造情報を抽出し，メタデータとして後段の処理へ提供する．

(2) サンプル音声の特徴を学習してモデル化するモデルベース手法：
聞きやすい読み上げに影響を与える入出力の特徴量は，古くからの分析でいくつかの知見が得られており，コーパスベースの音声合成に大きく貢献しているが，例えば音響モデルで用いられる 38 次元の素性に解釈を与えることは難しい．逆に言うならば，ユーザが説明的に条件を指定することで，自由に所望の合成音声を得る，という手段はまだ確立していない．そこで，このような複雑な素性に解釈を与えることはせず，目標とする音声から直接特徴量を入手して学習することで，その特徴量を模倣できるようなモデルを構築する，というアプローチを取る．また，この手法によれば，本報告で対象としたポーズ情報だけでなく，例えばレートやピッチ，小説の会話文に対する話者特定などの課題にも，この学習・モデル構築の枠組みを拡張することで対応することが期待できる．

(3) テキスト前処理モジュールとしての音声合成エンジンとの連携：
本開発内容は，音声合成エンジンをはじめとする他のアプリケーションから容易に連携して利用できることを目指す．そのため，入力および出力は XML を仮定し，音声合成エンジンに向けた今回の課題設定では SSML 形式のデータを出力するものとする．SSML であれば，一般の音声合成エンジンへ入力として受け渡すことが容易に可能であり，また本モジュールの出力結果に対しても，XSLT を適用することで，簡単に要素や属性のカスタマイズが可能である．

次節では，本提案手法の概要を述べる．

2.3 全体構成

全体構成を図 2 に示す．本処理は，大きく分けて 3 点で構成される．

まず，入力文書の書式情報を利用するため，入力文書からテキストの書式に情報に関連した情報をメタデータとして抽出する．入力文書が XHTML の場合には，タグの入れ子情報や，原文中のインデント，空行（複数，単数の違いで吸収）などを保持してテキスト情報を抜き出す．抽出されたテキストは，文書構造化処理によって階層構造などを抽出し，docbook ライクの論理構造タグが付与される．

次に，読み上げる表現のバリエーションを与えるために，前段のタグ情報と，テキスト内容から，テキストに関連した詳細な属性や属性値の付与を行う．これらは，テキストの解析結果（出現する形態素列や含有する論理構造情報）と，事前に用意されたモデルとをマッチングすることにより，属性の有無や値を判定する．

本報告では，目的とする属性としてポーズタグと，その属性値（ポーズ区間長）の推定・付与を行うこととする．

最後に，目的とするタグ仕様に沿った要素に変換する．ここでの出力は，SSML を想定し

Vol.2011-IFAT-102 No.6
Vol.2011-DD-80 No.6
2011/3/28

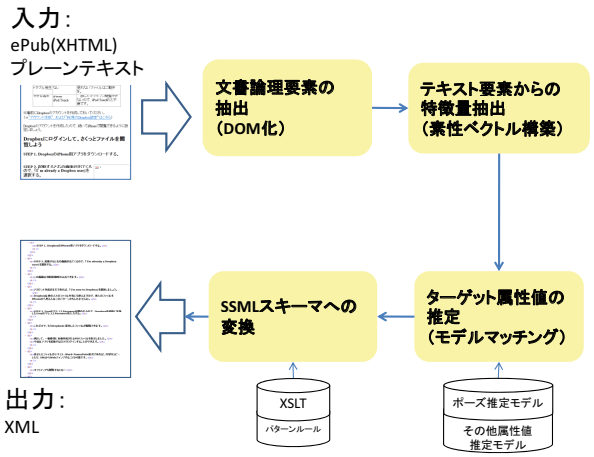

図 2 全体構成概要

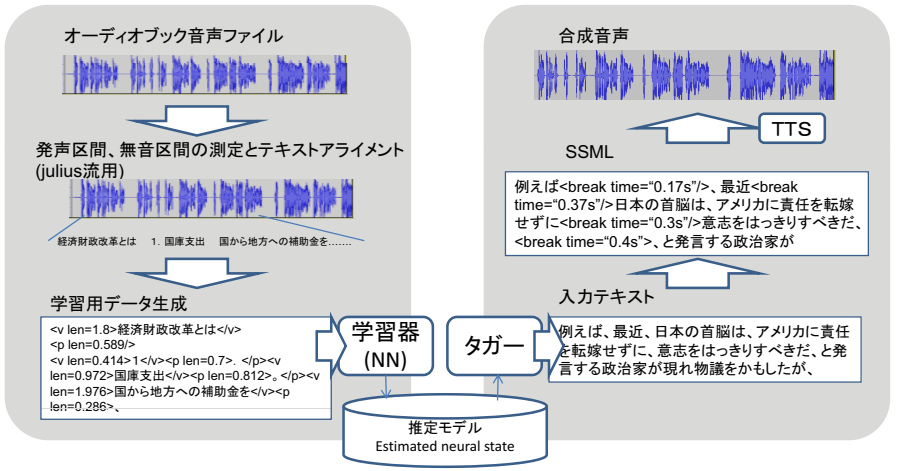

図 3 モデルベースのアプローチ

ているので，SSML の要素に準拠していないタグを変換したり削除することで，整形式の文書に整形する．

2.4 各部の詳細

こうした分析を元に，読み上げに貢献できる要素として，主に break 要素の位置と区間長の推定を実現するための試作を行った．

人のナレーション特徴として，ポーズ情報（発声長情報）の抽出と自動付与手段を検討し，未知テキストに対して，ナレーション特徴を模倣した読みを実現することが目的である．

以下に各処理を示す．

2.4.1 書式情報を利用した前処理

本報告では，電子書籍のフォーマットとして ePub を仮定している．ePub は目次やナビゲーションのためのファイルや，本文ファイル，本文から参照される図などのファイルで構成されている．この ePub データから取り出された本文ファイルに対して，テキスト抽出と構造化処理を適用し，書式付テキスト，および構造化の適用結果である XML ファイルを生成する．

なお，ePub 以外のデータフォーマットを取り扱う場合には，本文テキストを書式付きで抽出できるアダプタを作成すれば，以降の処理を同じ枠組みで実行できる．

2.4.2 読み上げ表現の多様化のための要素詳細化

前段の処理結果で取得された XML ファイルを対象にして，事前に用意されている読み上げ表現のための推定モデルを適用し，推定結果をもとに各テキスト要素に属性や属性値を付与する．

本処理のフローを図 3 に示す．

図の左側はモデル構築のフェーズ，右側は，モデルを適用して入力文書の属性や属性値を判定するフェーズである．モデルを構築する場合には，学習用のデータが必要である．ここでの学習用データは，SSML に準拠した XML 形式とした．

この学習用データを作成するには，実際の朗読音声から自動でテキスト書き起こしと無音区間長を計測して XML 形式で出力する手段を用いたり，手作業でテキストの書き起こしや無音長を埋め込んだ XML 形式で記述してもよい．ここでは，julius を流用して音声データからテキスト文字列と区間長を抽出し，学習用データ形式へ変換した．

この学習用データ (XML データ) を学習器に与えて，モデルを構築する．学習器は任意の手段を用いることができるが，ターゲットとする属性値を考慮すると，実際には離散値だけでなく，連続値を扱えることが必要である．また，昨今の学習器はその入出力を素性ベクトルとして記述しておくのが一般的であり，今回の試作内容も学習データから，入力テキス

Vol.2011-IFAT-102 No.6
Vol.2011-DD-80 No.6
2011/3/28

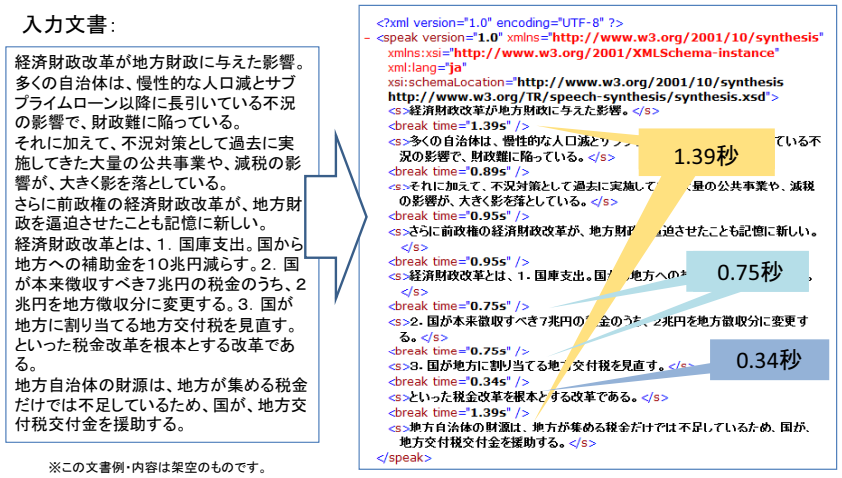

図 4 無音区間の推定結果例

トから一旦各種の特徴量を抽出した素性ベクトルを構成することとした．ここでは，学習器にニューラルネットを流用して推定モデルを構築したが，素性記述を変更すれば SVM や CRF などの学習手法をそのまま流用できる．

こうして構築した推定モデルを使って，未知の入力文書に対し，ポーズ位置やポーズ長の推定結果を得ることができる．ここではタガーを用意し，推定結果をタグとして埋め込んだ SSML データを生成することとした．

図 4 に推定結果の例を示す．ポーズ位置のコンテクスト（隣接する形態素列や論理構造情報など）に応じて，異なるポーズ区間長が推定されていることがわかる．

2.4.3 後処理としての XSLT 適用

最後に，後処理整形として XSLT 等により要素の付与や削除，変更を行う．ここまでの処理結果や推定結果は，すでに一部が SSML 形式で埋め込まれており，そのまま出力として，音声合成エンジンに与えることができる．ただし，中間処理結果としてタグを付与している場合には，必要に応じて除去したり，適切なタグ名称に変更する必要がある．(一般に，未対応のタグは音声合成エンジン側でスキップされる．)

またその他に，後処理の可能性として，たとえば推定結果を，早い，遅い，変更なしの 3 値など（レートの場合）の抽象的な値にしておき，本処理でユーザのニーズに応じた具体的な数値を埋め込む（通常の “早い” は+15%を割り当てるが，早聞きモードの場合には “早い” に+80%を割り当てるなど) という手段が考えられる．

3. 実　　験

3.1 目　　的

今回の試作内容による，無音区間領域の判定精度を概観する．また，時間長推定結果の参考としてロングポーズ (本実験では 1sec 以上の長さのものとする) の判定精度を概観する．

3.2 準　　備

市販のビジネス書に相当するオーディオブックから，Disc7 枚組のコンテンツ（以下データセット A (A)）と異なる Disc4 枚組のコンテンツ（以下データセット B (B)）の朗読音声ファイルから，ポーズ推定モデルと評価データを作成した．データ規模として，データセット A は 7disc 分の 66 章，データセット B は 4disc 文 103 章から成る．

なお，ポーズ推定モデルのために学習として与えた XML データは，wav ファイルから julius を用いて書き起こされたテキストと無音区間長である．無音区間長は SSML 形式 (break 要素と time 属性) で表記し，それ以外の文字列には便宜的にセンテンスタグ (s) を付与した．なお書き起こしテキスト中の読点（、）は各データセットから削除している．(句点は含む）そのため，句読点を特定することが直接ポーズの特定に一致するというタスク設定になっているわけではない．

この XML データから素性ベクトルを作るために，テキストの品詞列や論理構造情報，そのコンテクストとして近傍の品詞列や論理構造など 37 の属性と，それに対応するポーズの有無，ポーズがあればその長さを教師データとして抽出し，学習器に与えた．

3.3 実　　験

実験では二通りのタスクを設定した．一つは，それぞれの書籍データで，モデル構築と評価を行う 10-fold cross validation による評価であり，もう一つは，片方の書籍データをトレーニングセットとしてモデルを構築し，別の書籍データでテストを行うクロス評価である．各文節がポーズ挿入位置かどうかを判断する 2 クラス判定問題として，ポーズの判別評価結果を適合率・再現率・F 値で算出した．

なお，データセット A では，全データのほか，少量データに対する傾向を見るため，任意に抽出した 1 章分だけの実験結果も掲載した．

3.4 結果および考察

10-fold cross validation による各無音領域の判定結果を表 2 に示す．評価結果のカッコ

Vol.2011-IFAT-102 No.6
Vol.2011-DD-80 No.6
2011/3/28

対象データ	推定ターゲット	適合率	再現率	F-Measure
1 章相当分 (A)(disc2 の 9 章)	無音位置のみ	0.93	0.99	0.96
	ロングポーズ	0.78	0.58	0.67
全編 (A)(Disc1-Disc7)	無音位置のみ	0.97	0.99	0.98
	ロングポーズ	0.64	0.90	0.75
全編 (B)(Disc1-Disc4)	無音位置のみ	0.94	0.98	0.96
	ロングポーズ	0.68	0.97	0.8

表 2 10 交差検定によるポーズ位置の推定精度結果

training データ	test データ	推定ターゲット	適合率	再現率	F-Measure
(A)	(B)	無音位置のみ	0.59	0.96	0.73
(B)	(A)	無音位置のみ	0.97	0.99	0.98

表 3 トレーニングとテストで異なる書籍データを用いた場合のポーズ位置推定精度結果

内の数値は，システムが推定した数と実際の正解データの出現数である．

評価結果表 2 から，無音位置のみの判定（ロングポーズとショートポーズの区別なし）の F 値では，小規模データで 0.96，全編データで 0.98 の値を得た．しかしながら，位置に加えてロングポーズを特定する推定するタスクでは，F 値でそれぞれ 0.67，0.75 と大幅な精度低下がみられた．(なお，ロングポーズの推定に，無音位置の推定結果は入力要因として考慮されない.)

精度低下の要因としてはいくつかの原因が考えられるが，ロングポーズの場合は再現率が高い一方で適合率が低下していることから，ショート・ロングを問わず，ポーズの位置自体は正確に把握できている一方で，過剰に割り当てられた事例が多いことがわかる．

また，書籍データを変えてトレーニングとテストを実施した場合の無音領域の判定結果を表 3 に示す．こちらも同様に，(A) でトレーニングを行い (B) でテストした場合には，無音判定において再現率が高いが適合率が低下するという同様の傾向が見られることから，特定の文書特徴（ポーズが付与される傾向にある形態素列特徴）が，過剰に適合していることがわかる．

今後の適合率向上のための方策として，現在制約としては含まれていないような一般的な性質を取り込むことがまずは考えられる．単位文や単位時間当たりポーズ数や頻度，前回出現したポーズからの隔たり長，ストップワードに相当するような回避すべき表層語などがそれにあたるが，こうしたヒューリスティクスは，朗読における読み上げ手段のバリエーションを抑制してしまう場合もあるため，注意深い素性化が必要であると考えている．

4. おわりに

本報告[⋆1]では，電子書籍を音声合成でより聞きやすく読み上げるために，文書の論理構造を手掛かりとする方式検討を行った．聞きやすい音声出力の要因の一つとして，ポーズ情報に着目し，音声データと文書の論理構造から，モデルベースの手法を採用し，必要な機能を試作した．市販のオーディオブックをコーパスとしてポーズ推定モデルの構築を確認し，ポーズ位置に関する簡易精度評価の結果，F 値で 0.98（ポーズ長さを問わない場合）および 0.75（ロングポーズに限定した場合の精度) を得て，提案方式の有効性を確認できた．

今後は，ポーズ情報以外の異なる特徴量でも試行を行い，本手法の有効性を確認する．また，対象文書の種別や朗読パターンのバリエーションを拡充し，提案方式のロバスト性や評価や課題の洗い出しを検討する予定である．

参 考 文 献

1) University ofCambridge. HTK - hidden markov model toolkit - speech recognition toolkit. http://htk.eng.cam.ac.uk/.
2) 海木延佳, 匂坂芳典. 局所的な句構造によるポーズ挿入規則化の検討. 電子情報通信学会論文誌, Vol.79, No.9, pp. 1455–1463, 1996.
3) 京都大学河原研究室ほか. 大語彙連続音声認識システム - Julius. http://julius.sourceforge.jp.
4) 西光雅弘, 河原達也, 高梨克也. 隣接文節間の係り受け情報に着目した話し言葉のチャンキングの評価. 情報処理学会研究報告, Vol. 2006-SLP-61, No. (4), pp. 19–24, 2006.
5) 大田健吾, 土屋雅稔, 中川聖一. 音声認識用言語モデルにおけるポーズの有効利用. 日本音響学会講演論文集, Vol. 2-5-8, pp. 59–62, 2009.
6) 海老原充, 石川泰. 音声合成におけるネットワークモデルによるポーズ位置予測. 電子情報通信学会技術研究報告. SP, 音声, Vol.96, No. 566, pp. 45–50, 1997.
7) 比企静雄, 金森吉成, 大泉充郎. 連続音声中の音韻区分の持続時間. 日本音響学会誌, Vol.23, No.5, pp. 314–317, 1967.
8) 尾関創, 益子貴史, 小林隆夫. 多空間確率分布に基づくポーズのモデル化. 電子情報通信学会技術研究報告. SP, 音声, Vol. 104, No.30, pp. 41–46, 2004.
9) 藤尾茂, 匂坂芳典, 樋口宜男. 確率文脈自由文法を用いた韻律句境界とポーズ位置の予測. 電子情報通信学会論文誌, Vol.80, No.1, pp. 18–25, 1997.
10) 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. FIT2009(第 8 回情報科学技術フォーラム), Vol. E-051, pp. 377–380, 2009.

⋆1 本報告中で使われているシステム・製品名は，一般に各社の商標または登録商標です．