自然で聞きやすい電子書籍読上げのための文書構造解析技術

Document Structure Analysis for Expressive e-Book Reading

布目　光生　　鈴木　優　　森田　眞弘
■FUME Kosei　■SUZUKI Masaru　■MORITA Masahiro

東芝は，流行の兆しを見せる電子書籍の新しい楽しみ方を提供する目的で，より聞きやすい合成音声による電子書籍読上げの実現を目指している。このために，高音質な音声を合成する技術に加えて，見出しや箇条書きなど文書の構造に応じて適切な間を入れたり，文の内容に応じて音声を切り替えて読み上げたりする技術の開発に取り組んでいる。

今回，この機能開発の一環として，入力文書に応じて音声合成制御用のメタデータ(注1)を自動生成する文書構造解析技術を開発した。この技術は，音声合成システムの前処理として構成され，出力を業界標準のSSML（Speech Synthesis Markup Language）をはじめとしたXML（Extensible Markup Language）形式で得ることができる。そのため，必要に応じて音声合成システムとこの機能を組み合わせることによって，付加価値の高い多様な読上げを実現できる。

To provide new and enjoyable experiences for e-book readers, Toshiba has been developing expressive technologies related to e-book reading. In addition to our current high-quality text-to-speech (TTS) systems, we have focused on the realization of more natural talking e-books by developing new functions such as auto-pause insertion utilizing document logical structures, auto-talker selection, and talking style selection, according to each sentence type in the input document.　As one of these functions, we have developed a technology for document structuring that makes it possible to generate TTS control metadata from the input document features.

This technology is implemented as prefilter for TTS systems, and the resultant data can be obtained in extensible markup language (XML) format as well as in speech synthesis markup language (SSML), which is a standard format used in the TTS domain.　By incorporating this function into TTS systems as needed, we can provide an expressive as well as impressive e-book reading experience to readers.

1　まえがき

電子書籍の普及により，紙媒体では実現できなかった電子書籍ならではの様々な活用方法が広がり始めている。音声合成による朗読機能もその一つである。

一般に，書籍の朗読には，専門のナレーターによる朗読音声を収録した“オーディオブック”が知られているが，提供されている書籍数やジャンルが限られていたり，コストが高いこともあって気軽に利用することができなかった。

音声合成技術[1), 2)]を活用することで，ユーザーは好きな書籍を好きな合成音声で聞くことが期待できる。しかし，従来の音声合成は，一文やフレーズの読上げでは高音質な音声を実現している一方で，書籍データのような長い文書では，平板で淡々とした読上げになってしまい，そのままでは，感情的あるいは情緒的な表現を多く含む小説などの朗読を聞くには不十分だった。

また，例えば，章立て構造のある文書を段落境界やタイトルの後ろなどに一呼吸置きながら読み上げたり，“せりふ”と地の文を含んだ物語文などを，感情や調子を変えて読み分けたりすることが難しかった。

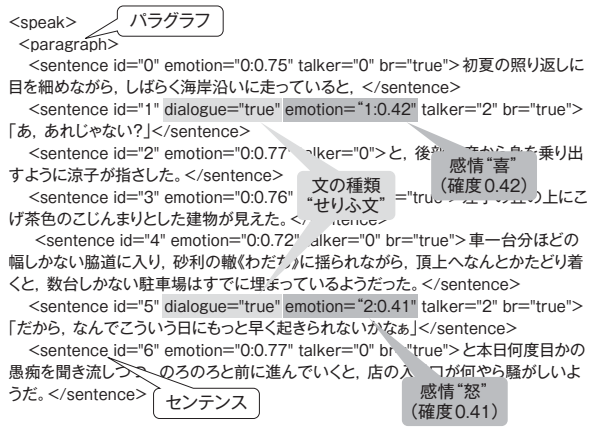

図1．XML構造化によるメタデータの生成例 ― 入力文書から，論理要素や文タイプ，せりふ文に対する感情ラベルを判定して，XMLデータとして出力する。

Example of metadata generation using XML structure

東芝は，こうした電子書籍をはじめとする文書をより自然に読み上げるために，音声合成の制御用メタデータを生成する文書構造解析技術を開発した。この文書構造解析技術[3), 4)]

（注1）　あるデータが付随して持つ，そのデータ自身についての付加的データ。

では，自然言語処理と機械学習の応用によって，文書中の適切と思われる位置にポーズ（無音区間）情報を入れたり，せりふ文中の感情表現を推定して，それらをXML属性として埋め込むことができる（**図1**）。

この結果に基づき，音声合成システムの抑揚やリズムである韻律の辞書や，声の辞書，韻律に関わるパラメータなどを文書に応じて切り替えることで，自然で聞きやすい合成音声を出力できる。

2　より聞きやすい音声合成による朗読へのアプローチ

より聞きやすい合成音声による朗読を実現するため，次の基本方針に基づき開発を行った。

(1)　ポーズ情報の推定　　文書の論理要素であるタイトルや，章立て構造，パラグラフ，箇条書きの切替わりや文節境界などに応じて適切なポーズ長を推定し，ポーズを挿入して読み上げる。

(2)　感情表現の推定　　せりふ文に対し，文中やその隣接文に出現する表記や単語を手がかりにして，事前定義された複数の感情表現から，もっとも近い感情表現を割り当てる。推定結果に応じて，文ごとに韻律辞書や音声制御用パラメータを切り替えて読み上げることで，擬似的な感情付き読上げを実現する（**図2**）。

2.1　ポーズ情報の傾向分析

これらの基本方針に先立ち，オーディオブックと各種合成音声によるポーズ情報の分析を行った。

ある文書を専門のナレーターが読み上げた場合の，ポーズの長さの傾向を各種の合成音声で読み上げた場合と比較した。対象は市販のオーディオブックで，社会及び経済のニュース解説に関するものである。

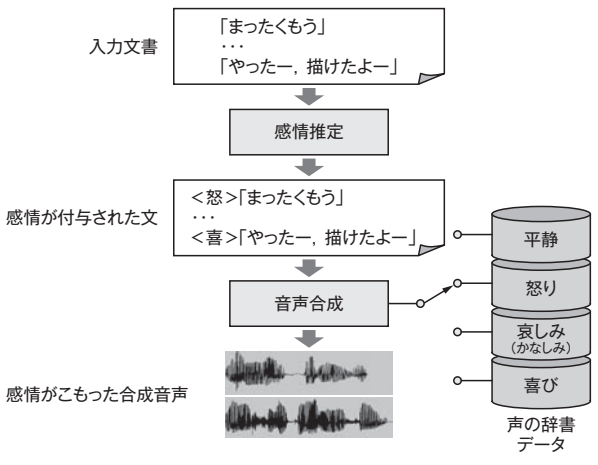

図2．感情付き読上げのアプローチ ― 文ごとの感情推定結果に基づいて，声の辞書データを切り替えながら読み上げる。
Basic approach to expressive e-book reading

ナレーターによる朗読音声では，ポーズの長さが0.25 s付近をピークに3 s程度まで広く分布する傾向があり，更に，タイトルと本文の間や，文と文の間，箇条書きの各項目間，パラグラフの切替わり，文書の結論部などの前後で，ポーズ長が異なっていることが確認できた。一方合成音声では，ポーズ長が句点と読点向けの2種類などに限定されていることが確認できた。内容や文書の論理構造によらず，合成音声で一定の調子で読み進められることは，人による朗読音声と比べた場合に，聞きやすさに影響を及ぼすと考えられる。

2.2　感情付き読上げのユーザー評価

一方，文ごとに感情を切り替えて読み上げる感情推定機能の開発に先立ち，ユーザー満足度に対するこの機能の有効性を検証した。

まず，せりふを比較的多く含む3種類の小説それぞれ約1ページ分を，次の音声で読み上げた。

(1)　感情なしの合成音

(2)　全てのせりふに正しい感情を付与した合成音。ただし，地の文は感情なしで合成

(3)　感情の付与をせりふの60 %に限定し，そのうち10 %程度は誤った感情を付与した合成音

これを30名の被験者にランダムな順番で聞いてもらい，各合成音声サンプルに対して，1（使いたくない）から5（とても使いたい）までの5段階の評価値から選ばせる主観評価試験を行った。

その結果，各条件に対する評価値の平均値であるMOS（Mean Opinion Score）値は，感情なしの(1)より，感情が正しく付与された(2)が0.6ポイント高い結果になった。現実的なシステムを想定した(3)についても，感情なしの(1)よりも0.4ポイント高い結果であった。

これらの結果から，ある程度の精度で感情が推定できれば，その推定結果に従って感情を付与することによって，ユーザーの評価が高くなる可能性があることを確認した。

次章では，これらの基本方針を実現するための具体的な方法について述べる。

3　聞きやすい合成音声を実現する方法

3.1　文書の論理構造に基づくポーズ推定手段

ポーズ情報の推定処理は，音声データとその書起こしテキストから推定モデルを獲得する学習フェーズと，与えられた入力文書にこの推定モデルを適用して入力文中に入れるべきポーズの位置とその長さを出力する推定フェーズで構成される（**図3**）。

まず学習フェーズでは，音声ファイルから発声区間と無音区間を測定し，それぞれの区間と書き起こされたテキストとの

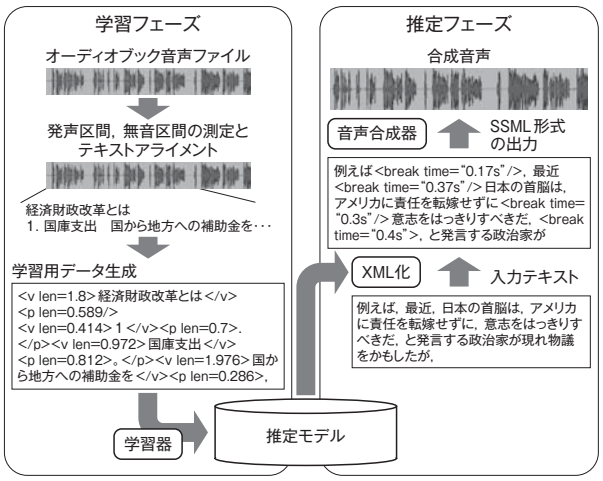

図3．ポーズ推定処理の概要 — 推定モデルを構築する学習フェーズと，入力テキスト中のポーズ情報を推定する推定フェーズで構成される。

Outline of estimation process for pause information

対応付けを行う。

次に，発声区間と無音区間の情報をテキスト中の部分文字列と対応付けながら，学習用データであるトレーニングベクトルを構成する。ここで学習に用いる特徴量には，各文節に出現する品詞種別や，文書の論理要素種別，文字種，隣接する文節情報などがある。これらの特徴量と，後続するポーズの有無，更にポーズが存在する場合にはそのポーズ長を対応付けて学習する。

学習方式として，特にベクトル表現されたトレーニングデータを学習するには，様々な手法が提案されているが，ここでは，ニューラルネットワークを採用した。一定数の学習を繰り返してエラーレートが十分低減した後に，その時点のニューラルネットワークの状態である各ノードの重みを，推定モデルとして保持しておく。

推定フェーズでは，入力文から取り出した特徴量を推定モデルの入力として与えることで，ニューラルネットワークの出力としてポーズ位置とその長さを得ることができる。推定結果は，音声合成システムが受け取ることのできる汎用のXMLデータとして出力される。

3.2　感情表現の推定手段

感情表現の推定では，まず学習データとして，各せりふ文に，それぞれ適切と思われる感情表現をラベルとして人手で割り当てたものを用意する。ここでは，感情ラベルとして，喜，怒，哀，怖，恥，好，厭（えん），昂（こう），安，驚，激怒，及び平の12種類を定義した。未知の入力文が与えられた場合に，この中の一つを自動的に割り当てることが感情推定手段の目的である。

一般に，入力データに対して，あらかじめ定義された複数のカテゴリーから一つの候補を割り当てるには，様々な方式が提案されている。ここでは，取り扱う文書が増えた場合に精度を保持するためのメンテナンスが容易であることと，内部処理が簡潔に実現できることから，ナイーブベイズに基づく分類方式を採用した。

ナイーブベイズを用いた分類手段では，ある文sが与えられた場合に，それが特定のカテゴリーcに属する確率を求めるため，ベイズの定理と呼ばれる式(1)の性質を用いる。

$$P(\mathrm{c}|\mathrm{s}) = P(\mathrm{c})\,P(\mathrm{s}|\mathrm{c})/P(\mathrm{s}) \quad (1)$$

$P(\mathrm{c}|\mathrm{s})$：ある文 s が与えられた時，カテゴリー c である確率

$P(\mathrm{c})$　：あるカテゴリー c が出現する確率

式(1)の右辺が最大になるカテゴリーcを出力することが目的である。ここで分母$P(\mathrm{s})$は，各カテゴリーに依存しないため省くことができるので，求めるものは，$P(\mathrm{c})\,P(\mathrm{s}|\mathrm{c})$を最大にするカテゴリー$\mathrm{c}_{\max}$（式(2)）となる。

$$\mathrm{c}_{\max} = \underset{\mathrm{c}}{\operatorname{argmax}}\, P(\mathrm{c})\,P(\mathrm{s}|\mathrm{c}) \quad (2)$$

$\underset{x}{\operatorname{argmax}}\, F(x)$：関数 $F(x)$ を最大にする x の値

ここで，式(2)の右辺の$P(\mathrm{s}|\mathrm{c})$は，あるカテゴリーに出現する文を示しているが，あらゆる文について取りえる単語の種類と組合せの出現頻度を調べることは現実的ではない。そのため，学習データ文中に含まれる単語を$\mathrm{word}_1, \cdots, \mathrm{word}_n$として，$P(\mathrm{s}|\mathrm{c})$を式(3)のように近似する。

$$P(\mathrm{s}|\mathrm{c}) = P(\mathrm{word}_1|\mathrm{c})P(\mathrm{word}_2|\mathrm{c})\cdots P(\mathrm{word}_n|\mathrm{c}) \quad (3)$$

これは直観的に，$P(\mathrm{s}|\mathrm{c})$があるカテゴリーcで出現する各単語の頻度情報で成り立っている，とみなすことができる。

この頻度情報をカテゴリーごとに調べ，その結果得られた，学習文書中に出現する単語とその頻度情報のリストをここでの推定モデルとする。

新しい入力文が属するカテゴリー，すなわち付与される感情ラベルを推定する場合には，この推定モデルから，文中に含まれる各単語の出現頻度情報を取り出し，ラベルごとのスコアを計算することで，入力文にもっとも近いラベルを推定することができる（**図4**）。

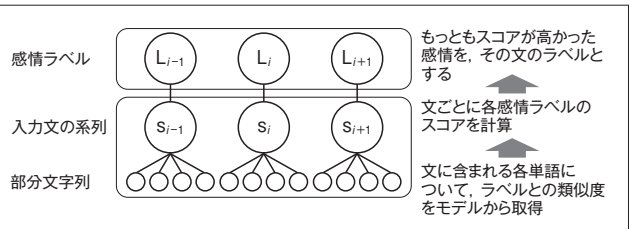

図4．各文に対する感情ラベルの推定処理のイメージ — 入力文に含まれるそれぞれの単語の特徴を総合して，文全体の感情ラベルを推定する。

Conceptual image of estimation process for emotion labels

4　評価実験と応用例

ポーズ推定機能及び感情推定機能の推定精度を実験によって評価した結果と，この機能を組み込んだ電子書籍読上げのインタフェース例について以下に述べる。

4.1　ポーズ推定に関する精度評価

市販のビジネスジャンルのオーディオブック2種類（ディスク合計13枚）から，自動検出されたポーズ情報と書起こしテキストをテストセットとして評価実験を行った。

テストセットの一部を学習データ，残りを評価データとした10交差検定によって，推定された文書中のポーズ位置が実際のポーズ位置と比べて正しいかどうかを評価した。

推定結果のうち，正解が含まれている割合を示す適合率で0.95，文書中のポーズ位置をカバーできている割合を示す再現率で0.98となり，この手法の有効性を確認できた。

4.2　感情推定に関する精度評価

Webサイトから入手できる書籍データのうち，数十種類から抜粋した約3,000文を対象に精度評価実験を行った。その結果，特に読み上げた時に音声として効果がわかりやすい喜，怒，及び哀の3種の感情ラベルに関しては，全体の2/3程度を学習データとして用いた場合に，適合率で約0.9の値を得た。

この値は，2.2節で述べた，ユーザーが有用性を認める一定の基準と同程度であることが確認できた。

4.3　電子書籍読上げインタフェース

一般的な電子書籍ビューアを想定し，読上げ機能を検証するためのデモシステムを構築した。

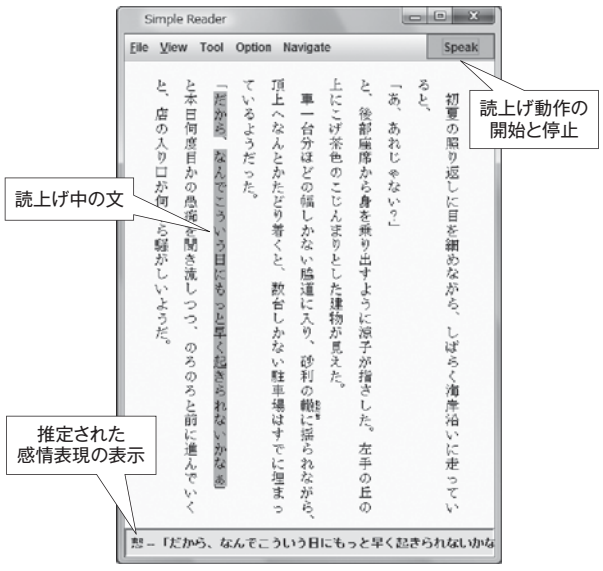

図5．読上げインタフェースの画面構成例 ― 任意のテキスト文書を表示しながら，各文に感情推定結果を適用した擬似的な感情付き読上げができる。

Example of e-book reading interface display

ユーザーが用意した任意のテキストデータに構造化処理を適用し，その結果を縦書き及びルビ付きの書籍形式で表示する。

読上げ時には，せりふと地の文の違いで話者を切り替えたり，せりふ文で推定された感情ラベルに従って，感情別に用意された韻律辞書を自動的に切り替えたりしながら読み進むことができるため，表現豊かな合成音声による朗読を簡単に聞くことができる（**図5**）。

5　あとがき

当社は，電子書籍を合成音声で朗読するために，入力文書の論理要素や内容に応じて，音声合成用の制御用メタデータを生成する文書構造解析技術を開発した。メタデータとして得られた制御情報を，当社の高品質な音声合成技術と組み合わせることで，自然で聞きやすい合成音声で朗読を聞くことができる。

今後は，音声合成システムの前処理や電子書籍リーダの読上げ付加機能として製品化を目指す。また，合成音声の特色を生かした新機能開発によって，ユーザー満足度に直接貢献できる技術開発を進めていく。

文　献

(1) 平林　剛 他. 次世代音声合成システムToSpeak$_{TM}$V2を支える多様性向上技術. 東芝レビュー. **65**, 4, 2010, p.43－47.

(2) 籠嶋岳彦. 高音質で聞きやすい音声合成システム ToSpeak$_{TM}$. 東芝レビュー. **62**, 12, 2007, p.34－37.

(3) 布目光生 他. 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. 情報処理学会 デジタルドキュメント（DD）. **2011-DD-80**, 6, 2011, p.1－7.

(4) Fume, K. et al. "Model-based document categorization employing semantic pattern analysis and local structure clustering". Document Recognition and Retrieval XV. San Jose, CA, USA, 2008-01, SPIE. 2008, p.68150R.1－68150R.8, (Proc. of the SPIE, **6815**).

布目　光生　FUME Kosei
研究開発センター 知識メディアラボラトリー研究主務。
文書構造解析及びナレッジマネジメント技術の研究・開発に従事。ACM，人工知能学会，情報処理学会会員。
Knowledge Media Lab.

鈴木　　優　SUZUKI Masaru
研究開発センター 知識メディアラボラトリー主任研究員。
情報検索及び自然言語処理技術の研究・開発に従事。人工知能学会，情報処理学会会員。
Knowledge Media Lab.

森田　眞弘　MORITA Masahiro
研究開発センター 知識メディアラボラトリー主任研究員。
音声合成技術の研究・開発に従事。日本音響学会会員。
Knowledge Media Lab.