平成 28 年度　卒業研究

機能語に注目した
音声合成朗読システムのための
感情推定手法の提案

東京理科大学　理工学部　経営工学科
西山研究室　7413069　恒川　泰輝

指導教員　西山　裕之

学士論文概要

目 次

図 目 次

表 目 次

第1章　序論

1.1　背景

近年，従来の書籍の他に電子書籍など様々な書籍の楽しみ方が広がっている．その中で，書籍の朗読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．オーディオブックはアメリカを中心に市場規模が拡大している．もともと車社会のアメリカなどの国では早期から市場が確立していたが，近年インターネットを介して気軽にダウンロードして楽しめる環境が整ったことなどによりアメリカとカナダの市場規模が 2015 年には前年比で約 21%拡大している [?]．日本においても定額配信サービスが開始されており，今後さらに普及する可能性がある．

しかし，このようなオーディオブックは書籍から音声化する際には手間やコストが電子書籍にくらべて 10 倍ほどかかっており 2〜3ヶ月ほどかかると言われている．[?]

そこで，電子書籍から人間の音声を人工的に作り出す音声合成技術を用いて機械で自動的に朗読するシステムの研究が行われている．近年の音声合成技術を用いれば喜怒哀楽といった感情を指定することで感情豊かな音声を合成できる．しかし，これらのパラメタは文もしくは単語ごとに人手で設定する必要がある．短い文章など限られた場合は容易であるが，小説といった膨大な文章に対して都度人手でパラメタ調整を行うのは手間がかかる．

そこで本研究では未知の文に対しその文を読み上げるときの感情として最適なものを推測することを目的とする．これにより自然な朗読システムが可能が実現可能になり，人手での手間やコストをかけずにオーディオブックを作成できるようになることが期待される．

本研究では，文にどのような単語が含まれているかという出現情報をもとに

機械学習技術を用いて感情を推定する．名詞，動詞，形容，形容動詞は内容語とよばれるが物語に依存する可能性が高いため，内容を除いた機能語を用いることで内容に依存しない分類器を生成できる．このため内容語を無視して機能語のみを用いて学習し感情の推定を行う．セリフのみに限らずすべての文を対象に，出現情報から機能語に絞った感情推定を行っている研究は筆者の知る限り存在しない．

本研究では，先行研究と同様に Normal，Happy，Sad，Angry の 4 つに感情をクラス分けする．本手法の正しさを確認するための実験として，まず一つの文にそれぞれ 4 つの感情で音声を合成する．そして Web のアンケートシステムを用いて，被験者にこれらの音声を実際に聞いてもらい，その文を読み上げる際にどの感情が最も適しているか判定してもらう．このように生成された学習データを用い交差検証を行うことで本手法の分類性能を評価する．

第2章　関連研究

本研究に関連する過去の研究について述べる．

感情を考慮した音声合成の研究として大谷 [?] らは共有感情モデルを構築してる．このモデルを用いることで音声を合成する際に感情を表出させることが可能である．しかし，感情パラメタ自体は人手で与える必要がある．本研究と組み合わせることで自動的に感情を推定しパラメタを与えることで，文そのもののみで感情豊かな音声合成が可能となる．

吉田ら [3] は自然な朗読システムのために文内・文末表層情報に着目している．文内情報 (命令，否定，意志等) と文末情報 (「〜ある」,「〜いる」,「〜んだ」等) をカテゴリー分けし，実際の朗読音声を元に文間ポーズと基本周波数，話速のモデル化を行うことで推定を実現している．しかし，この研究はあくまで自然な朗読システムの構築を目的としてるため，感情が考慮されてるとは言えない．本研究では物語から喜怒哀楽といった感情を推定することで，感情豊かな朗読システムの実現を目指している．

布目ら [5] はポーズ情報の推定と感情表現の推定を行っている．ポーズ推定では，タイトルや章立て構造といった文章の論理要素に応じてポーズ長を推定し，ポーズを挿入する．感情推定では「喜び」「怒り」「悲しみ」及び「平静」の各感情を付与した学習データを作成する．ナイーブベイズを用いて学習を行い，推定ではスコアを算出し最もスコアの高い感情を文に付与する．その推定をもとに，文ごとに韻律辞書や音声制御用パラメタを切り替えて読み上げる．しかしながら，感情を付与する対象はセリフのみであり，本研究ではセリフ以外の文についても感情を付与する．

第3章　提案手法

本章では提案する手法の詳細について説明する．

3.1　提案手法の概要

提案手法の概要を図 3.1 に示す．本手法では，物語中のすべての文に対し文中に含まれる単語の出現を手がかりに朗読に最も適切もしくは自然と感じる感情を推定する．感情のクラスは Normal，Happy，Sad，Angry の 4 種類とした．まず，学習データとして各文に，それぞれ適切と思われる感情を人手で割り当てたものを用意する．これに対し前処理を行いランダムフォレストを用いて学習を行う．そして未知の入力文が与えられた場合に，感情クラスの 1 つを自動的に推定する他クラス分類を行うのが本手法である．

3.1.1　機能語のみによる推定

本手法は，学習と推定の際に文から内容語 (名詞，動詞，形容詞，形容動詞) を取り除き，機能語のみで推定を行う．なぜならば，未知の物語の感情を推定を目的としているため，学習データが特定の物語に依存していては推定精度が低くとなると考えられるからである．例えば「鬼」がネガティブに描かれている物語を学習データとして，別の「鬼」がポジティブに描かている物語を推定した場合にネガティブな感情に推定されてしまう恐れがある．一方，機能後は「しまう」や「ところが」など，朗読時の抑揚などに関係すると考えられる重要な助詞やや接続詞を含む．したがって，内容語を排除し排除し，機能語のみで推定を行う．

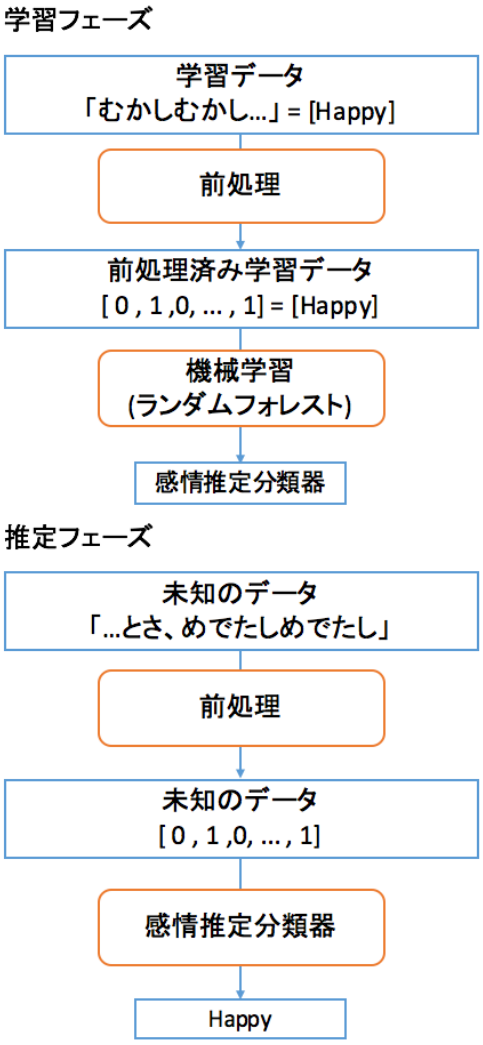

図 3.1: 提案手法の概要

3.1.2　前処理

前処理の概要を図 3.2 に示す．まず，物語の文章を文に分ける際は，基本的に句点で区切る．カギ括弧で囲まれているセリフ部分はカギ括弧で区切り，カギ括弧内のセリフについても句点で区切る．次に，形態素解析を行い単語ごとに分割しそれぞれの単語を基本形に変換する．そして，形態素解析の結果から内容語を削除し機能語のみにする．最後に，文中に単語が含まれるか否かを示すベクトル (Bag-of-words) に変換する．

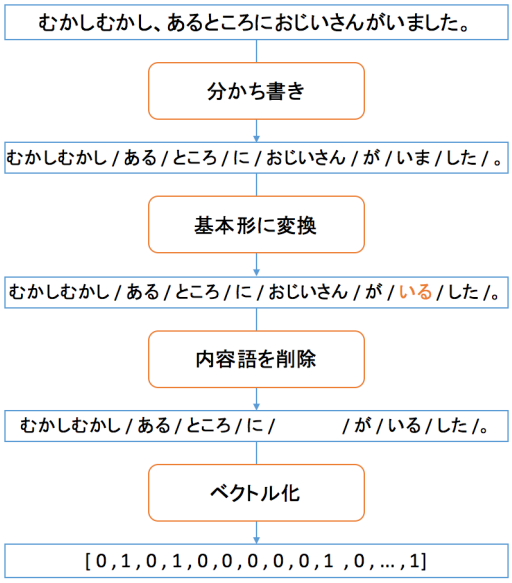

図 3.2: 前処理の手順

3.1.3　分類器の生成

一般に，入力データに対して，予め定義された複数のクラスから一つを推定する手法としては機械学習の教師あり学習が適応できる．本研究では，その 1 つであるランダムフォレストを用いて推定を行う．ランダムフォレストは複数の決定木を用いて識別などを行うアンサンブル学習アルゴリズムである．また，精度を向上させるために，正確度を指標としてグリットサーチを用いて最適なパラメタを探索する．

第4章　実験

本実験の概要を図 4.1 に示す．本実験では，まず物語文を元に各感情を指定した音声データを生成し，Web のアンケートシステムを用いて複数の被験者に評価してもらい，学習データを作成する．その後，学習データを用い交差検証を行う．

図 4.1: 学習データの作成手順

4.0.1　データの作成

提案手法で述べた通り物語データを文に区切り，一文につきそれぞれ 4 種類の感情を指定して音声データを生成する．

本実験では物語データとして，青空文庫 [2] にある 5 つの物語を用いる．今

回は文体を近づけるために同じ訳者の童話を中心に「白雪姫」,「赤ずきんちゃん」,「浦島太郎」,「ジャックと豆の木」,「ヘンゼルとグレーテル」を用いた．なお，ルビのデータが含まれているため予めタグを除いておく．

本研究の音声合成にはオープンソフトの Open JTalk[4] を用いる．Open JTalk は形態素解析部に MeCab[1]，発音辞書に NAIST Japanese Dictienary[?] を用いてる．波形生成部には MMDAgent[?] にある Mei のサンプルを用いてる．NOMAL，HAPPY，ANGRY, BASHFUL,SAD の 5 種類のサンプルがあるが，先行研究に従いこの内の BASHFUL を除いた 4 種類を用いた．なお，音声データは WAVE 形式として保存される．

4.0.2　学習データの収集

図 4.2: WEB アンケートシステム画面

Web 上で学習データを収集するためのシステムを構築した．そのシステム画

面を図 4.2 に示す．被験者には文ごとに各感情のパラメタで合成した音声をそれぞれ聞いてもらい内容にもっとも適切 (自然) であると思われる感情を 1 つだけ選択してもらう．しかし，感情は主観的な尺度であるため一人だけの評価では信頼性が低い．そこで，一文に対して同じ感情の評価が二票集まった時点で，その文の感情を決定することとした．したがって，同じ感情の評価が二票あるまで他の被験者に評価し続けてもらう．被験者等に関する詳細を表 4.1 に示す．

表 4.1: 学習データの収集

被験者	東京理科大学の学部生及び大学院生
人数	学部生 15 名，大学院生 2 名
取得期間	2017 年 1 月 9 日〜25 日
評価取得数	2641

4.0.3　評価

本実験では，leave-one-out 交差検証を行い，判定結果に対応する入力データの集合を TP，FP，TN，FN を次のように定義する．

True Positive(TP) 実際の感情のものを実際の感情であると予測したものの件数

True Negative(TN) 実際の感情でないものをその感情でないと予測したものの件数

False Positive(FP) 実際の感情でないものを実際の感情であると予測したものの件数

False Negative(FN) 実際の感情のものを実際の感情でないと予測したものの件数

以上をふまえ，分類器の性能評価を式 (1)，(2)，(3)，(4) で行う．本研究は分類推定を目的としているため特に F 値 (4) に注目する．

$$\text{正解率}\,(Ac) = \frac{TP + TN}{TP + TF + NP + NF} \tag{4.1}$$

$$\text{適合率（}Pr\text{）} = \frac{TP}{TP + FP} \tag{4.2}$$

$$\text{再現率（}Re\text{）} = \frac{TP}{TP + FN} \tag{4.3}$$

$$F\,\text{値} = \frac{2 * Pr * Re}{Pr + Re} \tag{4.4}$$

4.0.4　比較実験

提案手法の有効性を検証するために，同様な実験をセリフ文のみに絞った場合とさらに機能語に絞らなかった場合とそれぞれ行った．さらに，ランダムフォレストの比較として SVM を用いた実験も行った．このとき，ランダムフォレストと同様に SVM でもグリットサーチで最適なパラメタを導出した．

第5章　結果

5.0.1　学習データ

表 5.1: 学習データ (物語別)

タイトル	文数 (セリフ)	評価確定数 (セリフ)
白雪姫	287 (90)	258 (86)
赤ずきんちゃん	109 (54)	108 (54)
浦島太郎	206 (48)	78 (26)
ジャックと豆の木	206 (49)	78 (26)
ヘンゼルとグレーテル	319 (114)	260 (90)
合計	1096 (364)	765 (283)

表 5.2: 学習データ (感情別)

感情	全文	セリフのみ
Normal	459	63
Happy	134	110
Sad	99	60
Angry	73	50
合計	765	283

学習データの概要を表5.1 と表5.2 に示す．全体で評価が確定したものは全体で 69.8%であった．全文とセリフのみに絞った場合の比較を行う．得られた学習データは表5.2 の通り，Normal 以外の感情はセリフに多く含まれることがわかる．したがってセリフの感情推定の精度を上げることで全体の精度をあげることができることがわかる．全文にくらべセリフのみを対象とした場合はより均等に感情が別れているため分類がより難しい．

表 5.3: ランダムフォレストのグリットサーチの結果

パラメタ名	全文	全文 (機能語のみ)	セリフ	セリフ (機能語のみ)
ceriterion	entropy	entropy	entropy	entropy
min_samples_leaf	12	8	3	8
n_estimators	80	250	30	30
max_features	None	None	None	None
min_samples_split	12	10	3	10
max_depth	17	20	20	15

表 5.4: SVM のグリットサーチの結果

パラメタ名	全文	全文 (機能語のみ)	セリフ	セリフ (機能語のみ)
kernel	sigmoid	sigmoid	poly	rbf
gamma	0.001	0.001	3	0.001
C	100	100	1000	1

5.0.2　グリッドサーチ

グリッドサーチの結果を表 5.3 表と表 5.4 に示す．それぞれの場合で値が大きく異なるパラーメタが得られる場合があった．

5.0.3　評価結果

表 5.5: ランダムフォレストでの結果

対象	正確度	適合率	再現率	F 値
全文	0.82	0.57	0.64	0.57
全文 (機能語)	0.82	0.54	0.64	0.57
セリフのみ	0.70	0.37	0.40	0.37
セリフのみ (機能語)	0.70	0.35	0.40	0.34

ランダムフォレストと SVM の結果を表 5.5 と表 5.6 に示す．なお (機能語) とは学習，推定時に機能語のみを用いた場合を示す．全体として F 値は高くない結果となった．特に SVM の場合は SVM では，全文に機能語を絞らずに分類を行った場合を除く他のすべての場合で，推定が一つの感情に偏ってしまった．

表 5.6: SVM での結果

対象	正確度	適合率	再現率	F 値
全文	0.82	0.57	064	0.59
全文 (機能語)	0.80	0.36	0.60	0.45
セリフ	0.70	0.38	0.22	0.22
セリフ (機能語)	0.70	0.15	0.39	0.22

第6章　考察と今後の課題

全体としてのF値は高くない結果となった．原因として学習データが少ないことや出現を示すベクトルの形式に問題がある可能性がある．また，グリットサーチを正確度を基準に行ってしまったためF値を基準にやり直す必要がある．

ランダムフォレストとSVMを比較する．SVMでは，全文に機能語を絞らずに分類を行った場合を除く他のすべての場合で，推定が一つの感情に偏ってしまった．したがって，本研究の目的のためにはSVMよりランダムフォレストの方が有用であると言える．

機能語に絞った場合ととそうでない場合を比較する．ランダムフォレストの値ではほぼ同じもしくは機能語に絞らない方がわずかに良い結果が得られている．これは，学習データに用いた物語が５つと少ないことやleave-one-outを用いたことで推定する文と同じ物語の文を用いて学習を行っているからであると考えられる．したがって，機能語だけでも感情の推定を行える可能性はまだある．実際の運用では未知の物語の文に対して推定を行うので，機能語だけの学習・推定の方が精度が高い推定が行えるかもしれない．この検証を行うためには物語数を増やし学習データを増やした上で，leave-one-outではなく一つの物語をテストデータして他の物語を学習データとして検証を行う必要がある．また，決定木を用いて各単語の重要度を算出することで，機能語が感情推定にどれほど寄与するのか確認することができる．

第7章　結論

本研究では未知の文に対しその文を読み上げるときの感情として最適なものを推測することを目的とした．このための手法として物語に依存しがちな内容語を除いて機能語のみを用いてランダムフォレストで学習・推定する手法を提案した．

実験はネット上の5つの物語を使用して音声データを作成しWebのアンケートシステムを用いてNormal，Happy，Sad，Angryの４つのに分類し学習データを作成した．また，比較実験として機能語のみで学習・推定するか否かやランダムフォレストの他にSVMでの実験やセリフ文のみに絞った場合を行った．

結果とした全体的に高い精度を得ることはできなかった．しかし，本研究にはSVMよりランダムフォレストが有用であることや内容語を取り去って機能語のみで学習・分類を行っても，精度に大差はないことがわかった．したがって，ランダムフォレストを用いて，物語を増やし学習データを増やして学習を行い未知の物語に対して推定する検証を行うことで本手法の有用性が証明される可能性がある．

参考文献

[1] Mecab: Yet another part-of-speech and morphological analyzer. http://taku910.github.io/mecab/.

[2] 青空文庫. http://www.aozora.gr.jp/.

[3] 吉田有里, 奥平康弘, and 田村直良. 音声合成による朗読システムに関する研究. **情報科学技術フォーラム講演論文集**, 8(2):337–380, aug 2009.

[4] 大浦 圭一郎, 酒向 慎司, and 徳田 恵一. 日本語テキスト音声合成システム open jtalk. **日本音響学会春季講論集**, 1(2-7-6):343–344, 2010.

[5] 布目光生, 鈴木優, and 森田眞弘. 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. **東芝レビュー**, 66(9):32–35, 2011.

付 録A　WebページのURLリスト