

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・金盛 克俊

学士論文概要

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 [1]。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。[3]

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．

目 次

学士論文概要	i
第 1 章 序論	1
1.1 背景	1
1.1.1 オーディオブックの市場拡大と課題	1
1.1.2 音声合成技術の発展	2
1.1.3 朗読システムの現状と課題	3
1.2 本研究の目的	4
1.3 本論の構成	4
第 2 章 関連研究	6
2.1 音声合成の研究	6
2.1.1 様々な感情表現が可能な音声合成手法	6
2.1.2 直接波形接続型音声合成における感情表現	7
2.1.3 HMM 型音声合成における感情表現	7
2.1.4 HMM 型音声合成における共有感情付与モデルを利用し た感情表現	7
2.2 朗読システムの研究	8
2.2.1 表層情報を用いた朗読システム	8
2.2.2 感情推定を用いた朗読システム	8
2.3 本研究の位置づけ	9
2.4 本章のまとめ	9
第 3 章 提案手法	10
3.1 提案手法の概要	10

3.2	前処理	10
3.2.1	文の分割	10
3.2.2	形態素解析	11
3.2.3	分かち書き	11
3.2.4	内容語の削除	11
3.2.5	単語列のベクトル化	11
3.3	機械学習	12
3.3.1	ランダムフォレスト	12
3.3.2	グリッドサーチ	12
3.4	本章のまとめ	13
第4章	データセット	16
4.1	物語文章	16
4.2	音声データの作成	16
4.2.1	前処理	16
4.2.2	音声合成ソフト	17
4.2.3	音声データファイル	17
4.3	アンケートシステム	18
4.3.1	システムの概要	18
4.3.2	システムの利用技術	18
4.3.3	評価アルゴリズム	19
4.3.4	学習データの収集	19
4.4	学習データの概要	20
4.5	本章のまとめ	21
第5章	実験	26
5.1	実験環境	26
5.2	比較実験	26
5.2.1	セリフ文	26
5.2.2	分類器	26

5.3	評価実験	26
5.3.1	評価指標	26
5.3.2	交差検証	26
5.3.3	評価	26
5.3.4	比較実験	27
第 6 章	結果	28
6.0.1	グリッドサーチ	28
6.0.2	評価結果	28
6.0.3	考察	29
第 7 章	結論	31
付 録 A	アンケートの指示書	34

目 次

1.1	感情表現可能な音声合成ソフトの例 (AITalk)	2
3.1	提案手法の概要	14
3.2	前処理の手順	15
4.1	学習データの作成手順	22
4.2	青空文庫データの文分割とタグ削除処理	23
4.3	WEB アンケートシステム選択画面	24
4.4	WEB アンケートシステム評価画面	25

表 目 次

3.1	グリッドサーチで探索したパラメタ	12
4.1	音声合成のソフトウェア仕様	17
4.2	アンケートシステムのソフトウェア仕様	19
4.3	アンケートシステムのサーバースペック	19
4.4	学習データの収集	20
4.5	学習データ (物語別)	20
4.6	学習データ (感情別)	20
6.1	ランダムフォレストのグリットサーチの結果	28
6.2	SVM のグリットサーチの結果	28
6.3	ランダムフォレストでの結果	29
6.4	SVM での結果	29

第1章 序論

序論では、オーディオブックの市場や音声合成について解説し、朗読システムの課題を指摘し、本研究の目的や構成を述べる

1.1 背景

1.1.1 オーディオブックの市場拡大と課題

近年、従来の紙の書籍以外にも電子書籍といった様々な書籍の楽しみ方が増えている。その中でもオーディオブックは近急速な成長をしており、今後も成長が見込まれている。オーディオブックとは書籍の朗読は専門のナレーターによる朗読音声収録したものである。もともと車社会のアメリカなどの国では早期から大きな市場を確立している。近年インターネットを介したダウンロード販売が可能になったことなどによりさらに市場は拡大している。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[1]。

日本においては、欧米に比べて小規模なものにとどまっている。1980年代後半に後半にカセットブックが流行したが、車社会のアメリカに比べ電車などの公共機関による移動が多い日本では、市場規模はなかなか拡大しなかった。しかし、近年、スマートフォンなどが普及しカセットを持ち歩かなくても気軽に楽しめる環境が整ったことにより、インターネット上のダウンロード販売が急速に拡大している。日本の市場規模は2016年現在50億円程度と言われており、約10年後には800億円～900億円ぐらいに見込まれるとされてる。[2]

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電子書籍化に比べて非常に手間やコストがかかる。声優の選定からはじまり録音や読み間違い確認、BGM挿入、再度録音など人での作業が多い。そのため2～3ヶ

月ほどかかるを要し、コストは電子書籍の10倍ほどかかっていると言われている。[3]



図 1.1: 感情表現可能な音声合成ソフトの例 (AITalk)

1.1.2 音声合成技術の発展

音声合成とは、人間の音声を人工的に作り出すことである。この技術は文字を読むことが困難な障害者、外国人や幼児などに画面読み上げソフトとして長く利用されてきており、言葉を発することが困難な人が代替手段として利用することも多い。さらに、21世紀に入ってから家電製品の音声ガイドや公共交通機関のアナウンス、ロボットの発話用途などとして広く使用されるようになってきている。近年では声の切り替えや声の高さの調整などが可能になり、さらには指定した感情で音声合成ができるシステムも実用化されている。例えば、

株式会社エーアイのAITalkは図1.1に示すように怒り・悲しみ・喜びの感情をそれぞれ10段階で指定して合成することが可能である。

1.1.3 朗読システムの現状と課題

音声合成技術を用いることで、物語の自動朗読システムを実現することは可能であり、実際に実用化されている。しかし、品質は人間が読み上げて録音したものには及ばないのが現状である。単に音声合成を使うだけではなく、物語にあわせて読み上げる朗読に特化したシステムが実用化されている例は筆者が知る限り存在しない。

このようなシステムの実現が難しい理由としていくつか考えられる。まず、そもそもの音声合成の質の問題である。音声合成を情報提供を目的としたアナウンスとして用いる場合は問題がなかったとしても、物語といった長文の場合には平坦で淡々とした読み上げになってしまう。次に、朗読におけるポーズ長の重要性である。章立てやパラグラフといった文章構造の他に意味内容によるポーズが重要であることが杉藤ら [4] によって示されている。意味内容まで考慮したポーズ長の推定は現状では難しいと考えられる。最後に感情表現の問題である。1.1.2で述べたように、音声合成自体は感情表現が可能になった。しかし、感情は人の手によって調整する必要があり、物語といった長文の場合はコストがかかる。文章から自動的に感情を推定して、その推定された感情にしたがって読み上げる朗読システムは実用化されていない。一方、テキストマイニングの分野で盛んに感情推定が行われているが、あくまで商品や映画などのレビューやクレームといった事実に対する感情を推定するものが多い。そのため物語の文章に対してそのまま適応できるかは疑問である。さらに、朗読の場合の感情推定は文中の人物の感情を推定するのではなく、あくまでどの感情表現で読むべきかを推定する必要がある。

1.2 本研究の目的

本研究では 1.1.3 で述べた朗読システムの課題の中で，感情推定の問題を取り扱う．すなわち，朗読システムのために，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定することを目的とする．これにより音声合成システムの感情パラメータを自動的に調整することが可能になり，より自然な読み上げが自動的に行えるようになる．さらに，自然な朗読システムが可能が実現可能になることで，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，先行研究に従い 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．

1.3 本論の構成

本章では，本論文の背景となるオーディオブック市場の拡大と音声合成技術を説明した上で朗読システムの現状と課題について説明し，それを踏まえ本研究の目的を述べた．第 2 章では既存の関連する研究について述べる．また第 3 章では本論文が提案する手法の詳細を述べ，第 4 章でその評価実験について説明する．そのうえで第 5 章で提案手法の効果を測定するために行なった実験の

第1章 序論

結果と考察を述べ，第6章でその結論と今後の展望について述べる．

第2章 関連研究

本章では、音声合成や朗読システムにの既存研究について述べるとともに本研究の明確な位置づけを行う。

2.1 音声合成の研究

感情表現が可能な音声合成技術の研究を紹介する。

2.1.1 様々な感情表現が可能な音声合成手法

近年、日本語の音声合成手法として広く利用されている2002年のYoshimura[5]が提案した隠れマルコフモデル(HMM)に基づく合成手法がある。この手法以前もカーナビや音声翻訳システムなどに音声合成が利用されていた。しかし、様々な話者の声質で話したり、嬉しそうに、怒ったようになど様々な発話スタイルで話すことができるものは少なかった。なぜなら、それ以前の波形接続型のシステムで様々な声質、発話スタイルを実現するためには、様々な声質、発話スタイルで収録された膨大な量の音声データを処理しなければならず、また合成するには膨大な量の素片を格納する記憶媒体が必要となるため、実現は非常に困難であったためである。このHMMに基づく合成手法を用いることで、抽象化された関数を利用し音声波形に揺らぎあってもあってもその背後に潜んでいる特徴を見出すことができる。これによって少ない音声データで学習することが可能になり、パラメータ調整することで別の人物の声を真似たり、様々な感情表現を行うことが可能になった。

2.1.2 直接波形接続型音声合成における感情表現

飯田ら [6] は自然音声直接波形接続型音声合成システム CHATER を用いて、表現したい感情ごと (喜び・怒り・悲しみ) にその感情に対応する音声コーパスから最適な音声波形素片を選択し接続することで、音声合成した。聴取実験を行い、有意水準 1% で検定を行ったところ、各感情は有意に判別され、各感情は有意に判別された。しかし、波形接続型の音声合成では合成したい音素の基本周波数予測エラーによりイントネーションが不自然に聞こえる場合があり、そもそもの音声合成の質が高いとは言えない。また、前小節で述べた通り波形接続型の音声合成では学習に多くの音声データが必要であり、様々な感情のデータを取得するには多くのコストがかかる

2.1.3 HMM 型音声合成における感情表現

都築ら [7] は HMM 音声合成システムを用いて少ない学習データを用いて感情表現のモデル化を行っている。学習に用いる音声データは英語音声で目的の感情を表現しやすい文章を読み上げたものである。感情は平静・怒り・喜び・悲しみの 4 種類の感情について合成を行い、聴取による主観評価実験を行っている。結果としては、判別誤りが多く見られ芳しい結果は得られなかった。原因として文章と感情の関係や音声品質、学習データの不足などが挙げられている。

2.1.4 HMM 型音声合成における共有感情付与モデルを利用した感情表現

大谷 [10] らは HMM 型音声合成において加算構造に基づく感情表現を提案している。この手法では複数の話者の感情音声データを用いて、学習者共有の感情成分を持つ共有感情付与モデルを学習し、このモデルを任意の平静音声モデルへ適応する。これにより HMM 型音声合成において従来法に比べ、音質が高く、安定した感情表現が可能となった。

2.2 朗読システムの研究

物語の文章内容に応じて、音声合成を調整する音声合成システムの研究について説明する。

2.2.1 表層情報を用いた朗読システム

吉田ら [9] は朗読文に朗読者の音声の間 (ポーズ) 及び韻律的特徴 (基本周波数, 話速) を解析し特徴のモデル化を行うことで, 文章に応じてポーズや韻律を付与するシステムを提案している. 物語の分は文内表層情報 (命令, 否定, 意志等) と文末表層情報 (「～ある」, 「～いる」, 「～んだ」等) よりカテゴリー分けし, それを特徴としている. 聴取実験の結果では調整を行っていない音声に比べ調整を行った音声の方が自然と回答した割合が80%前後であった. しかし, この研究では予め「情景描写」や「緊迫」といった場面抽出を人手で行っているため, それ以外の場面において有効であるか疑問である. また, 基本周波数と話速の調節しか行っておらず, 前節で述べた感情付与モデルを利用して音声合成を調節することで, さらに自然になる可能性がある.

2.2.2 感情推定を用いた朗読システム

布目ら [11] はセリフ文に対し, 文中やその隣接分に出現する表記や単語を手がかりにして, 事前に定義された複数の感情から最も近い感情表現を割り当てるシステムを提案している. 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. 精度評価の結果, 喜, 怒, 哀の3種の感情ラベルに関しては90%前後の精度を得た. しかしながら, 感情を付与する対象はセリフのみであり, セリフ文以外にも付与することでより自然な朗読が可能になると考えられる. また, 感情ラベルの付け方が明示されておらず, 推定が容易な文のみを対象としていたり客観的なラベル付けが行われていない可能性がある.

2.3 本研究の位置づけ

本研究では自然な朗読システムのために物語の全文を対象に一文ずつ予め用意された感情の中からで音声合成すべきか推定を行う。

大谷 [10] らの研究ではすでに感情表現が可能な質の高い音声合成は可能であることがわかった。しかし、この技術だけでは人による感情のパラメタ指定が必要であり、コストがかかる。本研究で期待される感情推定技術と組み合わせることで、自然で質の高い自動朗読システムを実現することが可能になる。

本研究と同じ目的の研究は他にもあり、それらとの違いや類似点を説明する。

吉田ら [9] の研究では基本周波数と話速の調整のみを行っていたが、本研究ではあくまで感情ラベルの推定を行う。これによって感情を考慮した朗読システムが構築できる可能性がある。ただし、この研究によって文脈や内容そのものを理解せずとも理解せずとも、文の表層情報からある程度、どのように朗読すべきか推定可能性が示された。これを受けて、本研究では機能語のみによる推定も行った。

布目ら [11] はセリフ文のみに着目している。本研究ではセリフ文以外の文章に対して感情推定を行う重要性の検討を行い、それに対する推定も行う。また、ナイーブベイズのみの推定になっているが、本研究ではその他にランダムフォレストやSVMでの比較実験を行い評価する。

さらに既存の研究では、正解ラベルの付け方、使用する文の選定に不明瞭な点が多く、実際の運用の際に未知の文に対応できるのか疑問であった。本研究では、選択した物語からすべての文を対象に無作為に抽出を行い、複数の第三者によって評価を行わせ正解データを作成した。

2.4 本章のまとめ

本章では感情表現が可能な音声合成の研究と朗読システムの研究をいくつか挙げた。また、音声合成技術と本研究を組み合わせることで得られるメリットを説明した。さらに既存研究の現状とそれらが抱える問題点を指摘した。それを踏まえ本研究が目指す領域について説明した。

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．まず，提案手法全体の概要を示し，前処理の具体的な手法や分類に使用する機械学習の手法について解説する．

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．

3.2 前処理

機械学習で前処理の手法を順に追って説明する．その概要を図 3.2 に示す．

3.2.1 文の分割

本手法では文単位で感情の推定を行う．それゆえ，物語の文章を文に分ける必要がある．意味内容を解説しそれにしたがって分割した方が自然な読み上げが可能になる可能性はあるが，朗読システムの構築を考えると単純な分割が好ましいと判断した．基本的に，句点で文章を分ける．カギ括弧で囲まれた箇所はそれを一文とし，その前後の文もそれぞれ一文とする．

3.2.2 形態素解析

次に文ごとに形態素解析を行う。形態素解析とは文法的な情報の注記の無い自然言語の文から、文法や辞書と呼ばれる単語の品詞等の情報にもとづき、言語で意味を持つ最小単位である形態素の列に分割し、それぞれの形態素の品詞等を判別する作業である。本手法では次節以降で述べる分かち書きと機能語の削除に形態素解析で得られた情報を用いる。

3.2.3 分かち書き

次に分かち書きを行う。分かち書きとは文を語ごとにわけ作業である。英文の場合は語と語の間に空白(スペース)が入っているが、日本語の場合は入っていないため単純には行えない。そこで前節の形態素解析の結果を用いて分かち書きを行う。

3.2.4 内容語の削除

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。このとき、形態素解析を行った結果を用いて、内容語と機能語の判別を行い内容語はデータから削除する。

3.2.5 単語列のベクトル化

分かち書きされた文を機械学習で扱える形式に変換がある。本手法では、文のベクトル表現の1つである bag-of-words を用いる。解析で用いるすべての単

語文の次元をもつベクトルを用意し文中に単語があれば1とし、なければ0とする。なお今回は頻度は考慮せず、出現の否かのみを考慮する。

3.3 機械学習

3.3.1 ランダムフォレスト

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機械学習の教師あり学習が適応できる。本手法では、その中の手法の1つであるランダムフォレストを用いて機械学習を行う。波部ら [8] によると、ランダムフォレストは複数の決定木を用いて森を構成し識別などを行う機械学習アルゴリズムである。個々の決定木は高い識別性能をもつわけではないが、それらを複数用いてそれぞれの結果を補うことによって高い予測性能を得ることが1つの特徴である。これは機械学習の分野ではアンサンブル学習と呼ばれており、個々の決定木がアンサンブル学習における弱識別器に相当する。

3.3.2 グリッドサーチ

本手法ではより精度を向上させる手法として学習を行う前にグリッドサーチを行う。グリッドサーチとは学習の際に与えるパラメタそれぞれに対していくつかの値を与え、それらの組み合わせについて学習と交差検証を行いつつ全探索し、最も良いスコアのパラメタを探索する手法である。本手法で探索するパラメタは表 3.1 の通りである。

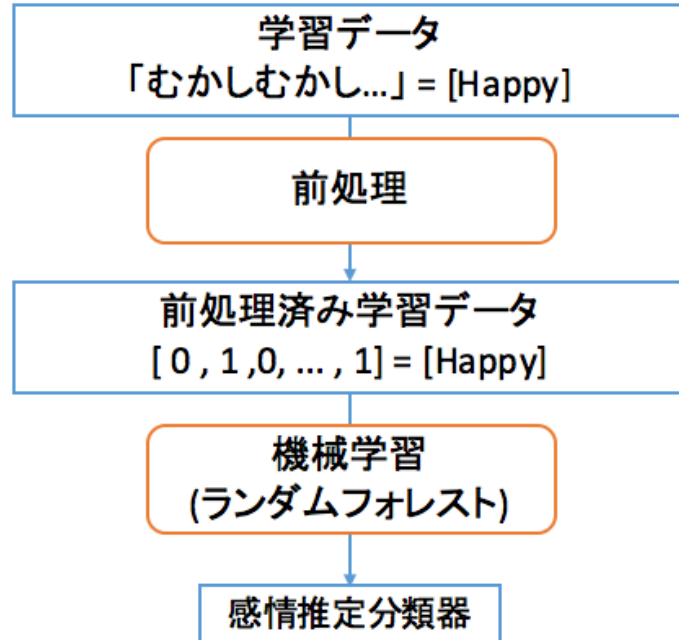
表 3.1: グリッドサーチで探索したパラメタ

パラメタ名	意味
max_depth	木の最大の深さ
n_estimators	決定木の数
max_features	特徴量の最大の数
criterion	重要度計算の尺度
min_sample_split	葉ノードの最少分割数
min_samples_leaf	葉ノードに用いる特徴量の最小数

3.4 本章のまとめ

本章では本研究で提案する文に対する感情推定の手法について，文章からベクトルへの変換手法や内容語削除などについて述べた．次章では，本手法の有効性を示すための実験について説明する．

学習フェーズ



推定フェーズ

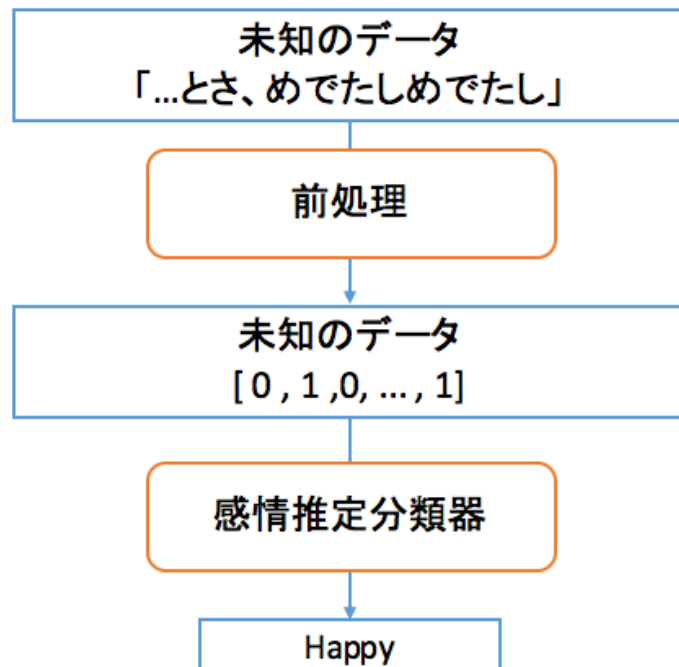


図 3.1: 提案手法の概要

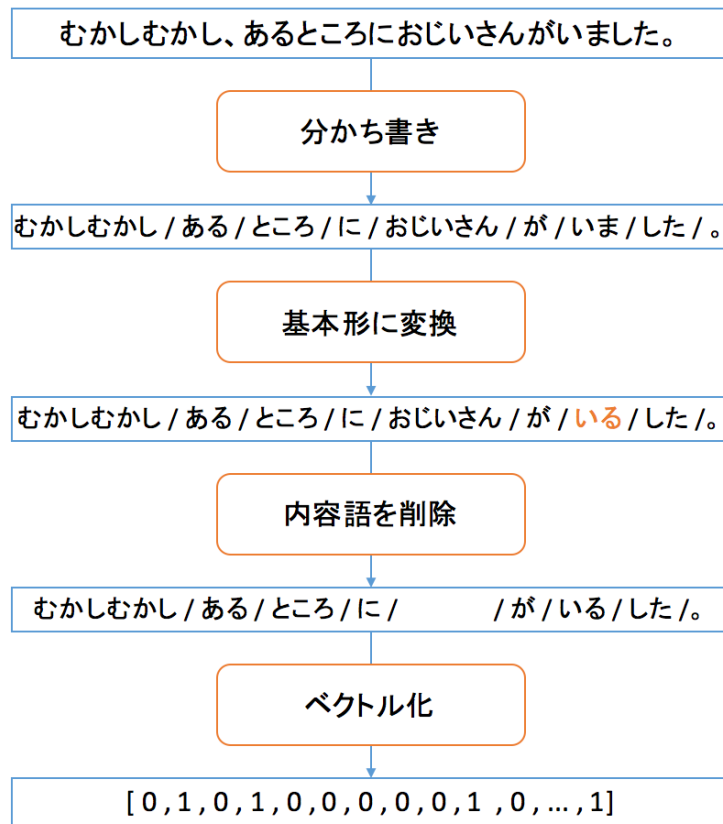


図 3.2: 前処理の手順

第4章 データセット

本章では実験に用いる教師データを収集するためのシステムや集計手法について説明する。教師データの作成手順を図 4.1 に示す。本研究には、文に対してどの感情表現で音声合成を行うべきかという、文と感情表現ラベルが対になった学習データが必要である。このために、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評価者に評価してもらい、学習データを作成する。

4.1 物語文章

本実験では物語データとして、青空文庫 [12] にある 5 つの物語を用いる。青空文庫とは、著作権が消滅した作品や著者が許諾した作品のテキストを公開しているインターネット上の電子図書館であ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。

4.2 音声データの作成

4.2.1 前処理

3.2.1 で述べたように物語の文章を文に分ける。朗読システム実現するためにはこの処理も自動化する必要がある。また、青空文庫の書籍データにはルビや改行といった HTML タグが挿入されているためこれを除く必要がある。プログラミング言語 Ruby [18] を用いて、これらの処理を実装した。これらの処理の例を 4.2 に示す。

4.2.2 音声合成ソフト

本研究では音声合成にオープンソフトのOpen JTalk[13]を用いる。Open Jtalkは日本語テキスト向けのHMM型の音声合成のオープンソースのソフトウェアである。2.1.3で述べた通りHMM型音声合成手法は少ない学習データで感情表現が可能になっている。また、再現実験を考慮し広く利用が可能なオープンソースソフトウェアであるOpen Jtalkを用いた。

音声波形データにはMMDAgent[16]にあるMeiのサンプルを用いた。この音声波形データは1人の女性の848文の音声サンプルをもとに作成されている。そのうち、503文は音声波形作成に広く用いられている用いられるATR音素バランス503文[17]であり、残りのうち215文は普通の発話スタイルで、残る75文は4つの感情(Angry, Bashful, Happy, Sad)の発話スタイルで収録されている。それゆえ、MMDAgentにはNormal, Angry, Bashful, Happy, Sadの5つの音声波形データが用意されている。本研究では布目ら[11]などの研究に従い、このうちのNormal, Angry, Happy, Sadの4つを利用した。

発音辞書にはNAIST Japanese Dictionary[15]を用いる。

各ソフトウェアの仕様は表??に示す。

表 4.1: 音声合成のソフトウェア仕様

名称	バージョン
Open JTalk	1.10
NAIST Japanese Dictionary	0.4.3
MMDAgent	1.7

4.2.3 音声データファイル

本研究では、前章のOpen JTalkとMeiの音声波形データ等を用いて、一つの文に対して4つの異なる感情表現の音声ファイルを生成する。この音声ファイルはWAVフォーマット形式で出力される。WAVフォーマット形式は、非圧縮形式でありリニアPCMのサンプリングデータ用のフォーマットとして扱わ

れる．Open JTalk の出力ではサンプルレートが48,000Hz, 16bps, モノラルの WAV ファイルが得られる．

4.3 アンケートシステム

本研究では効率よく学習データを採取するために Web 上で文に対して感情のラベル付けが行えるシステムを構築した．

4.3.1 システムの概要

サインアップもしくはログインを行い評価者は図 4.3 のインデックスページの評価ボタンがから図 4.4 に示す評価ページに遷移する．評価ページにおいて後述する評価アルゴリズムに基づいて自動選択された文について評価者は4つの音声それぞれを聞き，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．評価ボタンをクリックすると評価はデータベースに保存され，自動選択された別の文の評価ページへと遷移する．

4.3.2 システムの利用技術

Web アプリケーションフレームである Ruby on Rails[19] を採用し，データベースには SQLite3[20] を用いた構築した．

評価ページの音声再生部分には HTML5 に定められている Web Audio API と audio タグによる2つの実装を行った．これは，評価者のブラウザ環境によってはどちらかが使えない場合があるからである．なお，Play ボタンを押すと Web Audio API が呼び出されるがこの部分の実装には JavaScript を用いた．これにより，Open JTalk で生成された WAV ファイルが評価自身の端末で再生可能となった．

また，得られた学習データは CSV 形式で出力する機能も備えており実用的である．

その他，本システム用いたサーバー機のスペックはソフトウェアの仕様 4.2 に，表 4.3 に示す

表 4.2: アンケートシステムのソフトウェア仕様

名称	バージョン
Ruby	2.3.3p222
Ruby on Rails	4.2.7.1
SQLite3	3.7.17

表 4.3: アンケートシステムのサーバースペック

名称	スペック
マシン	マウスコンピューター MASTERPIECE i1550PA7-CL-W7
OS	CentOS Linux release 7.3.1611 (Core)
CPU	Intel Core i7-3970X 3.50GHz × 2
メモリサイズ	64.0GB

4.3.3 評価アルゴリズム

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1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2票集まった時点で評価を確定することとした。異なる評価が集まった場合はもう1票だけ評価を続け、既存の評価と同じ感情であればその評価で確定することにした。3票ともに異なる評価が集まった場合には、その文は学習データとしては利用しないこととした。なお、セリフ文を優先的に評価するように設定した。

4.3.4 学習データの収集

大学生に自身が所有するPCやスマートフォンで、アンケートシステムにアクセスさせ、評価を行わせた。なお、評価の期限や数は評価者の任意とした。その他詳細は表 4.4 に示す。

表 4.4: 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

4.4 学習データの概要

学習データの概要を表 4.5 と表 4.6 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. セリフは全体の 33.21%であり, セリフの評価完了率は 81.04%であった. 表 4.6 の通り, Normal 以外の感情はセリフに多く含まれており, セリフの方が感情豊かであり, それ以後の文は Normal であることが多いことがわかる.

表 4.5: 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 4.6: 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

4.5 本章のまとめ

本章では，本研究に用いるデータセットの説明や収集方法，アンケートシステムの詳細について説明した．

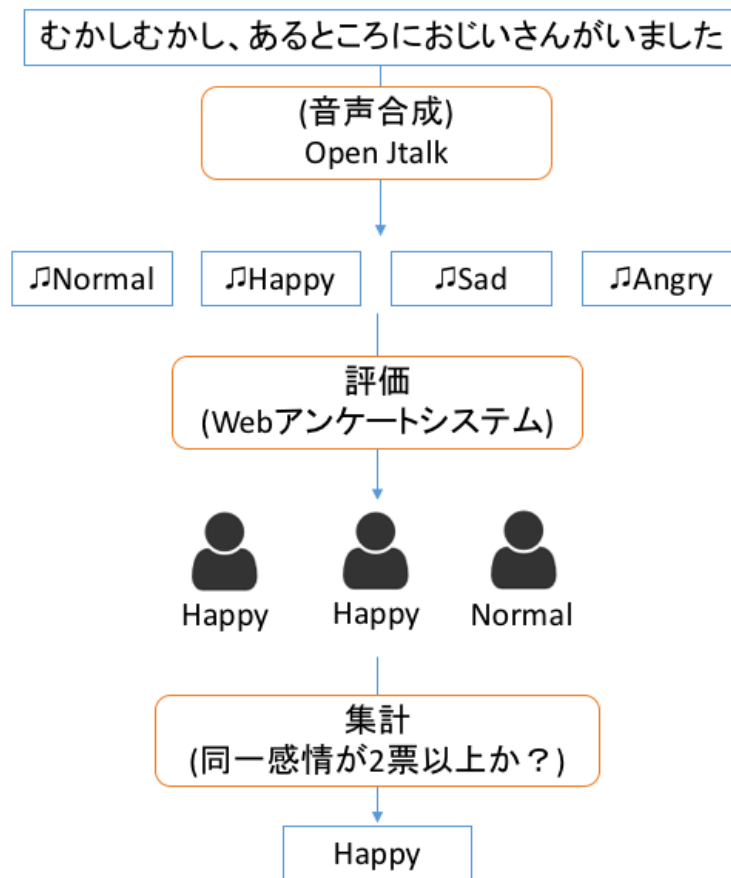


図 4.1: 学習データの作成手順

青空文庫の元データの例

むかし、むかし、あるところに、ちいちゃいかわいい女の子がありました。それで、あるとき、おばあさんは、赤いびろうどで、この子にずきんをこしらえてやりました。すると、それがまたこの子によく似あうので、もうほかのものは、なんにもかぶらないと、きめてしまいました。そこで、この子は、赤ずきんちゃん、赤ずきんちゃん、とばかり、よばれるようになりました。


ある日、おかあさんは、この子をよんでいました。

「さあ、ちよいといらっしゃい、赤ずきんちゃん、ここにお<ruby><rb>菓子</rb></rp></rt><rt>かし</rt></rp></rt></rp></ruby>がひとつと、ぶどう<ruby><rb>酒</rb></rp></rt><rt>しゅ</rt></rp></rt></rp></ruby>がひとつびんあります。これを赤ずきんちゃん、おばあさんのところへもっていらっしゃい。おばあさんは、ご病気でよわっていらっしゃるが、これをあげると、きっと元気になるでしょう。それでは、あつくならないうちにおでかけなさい。それから、そとへでたら気をつけて、おぎょうぎよくしてね、やたらに、しらない横道へかけだしていったりなんかしないのですよ。そんなことをして、ころびでもしたら、せっかくのびんはこわれるし、おばあさんにあげるものがなくなるからね。それから、おばあさんのおへやにはいったら、まず、おはようございます、をいうのをわすれずにね。はいると、いきなり、おへやの中をきょろきょろみまわしたりなんかしないね。」

変換後のデータ

むかし、むかし、あるところに、ちいちゃいかわいい女の子がありました。
それで、あるとき、おばあさんは、赤いびろうどで、この子にずきんをこしらえてやりました。
すると、それがまたこの子によく似あうので、もうほかのものは、なんにもかぶらないと、きめてしまいました。
そこで、この子は、赤ずきんちゃん、赤ずきんちゃん、とばかり、よばれるようになりました。
ある日、おかあさんは、この子をよんでいました。
「さあ、ちよいといらっしゃい、赤ずきんちゃん、ここにお菓子がひとつと、ぶどう酒がひとつびんあります。
これを赤ずきんちゃん、おばあさんのところへもっていらっしゃい。
おばあさんは、ご病気でよわっていらっしゃるが、これをあげると、きっと元気になるでしょう。
それでは、あつくならないうちにおでかけなさい。
それから、そとへでたら気をつけて、おぎょうぎよくしてね、やたらに、しらない横道へかけだしていったりなんかしないのですよ。
そんなことをして、ころびでもしたら、せっかくのびんはこわれるし、おばあさんにあげるものがなくなるからね。
それから、おばあさんのおへやにはいったら、まず、おはようございます、をいうのをわすれずにね。
はいると、いきなり、おへやの中をきょろきょろみまわしたりなんかしないね。」

図 4.2: 青空文庫データの文分割とタグ削除処理



ID	タイトル	参加人数	完了率(個人)	完了率(全体単純)	完了率(全体完全)
1	白雪姫	3	0.00 %	158.71 %	89.90 %
2	赤ずきんちゃん	3	0.00 %	158.26 %	96.33 %
5	浦島太郎	2	0.00 %	92.24 %	37.93 %
6	ジャックと豆の木	3	0.00 %	93.93 %	39.81 %
9	ヘンゼルとグレーテル	3	0.00 %	115.36 %	82.76 %

図 4.3: WEB アンケートシステム選択画面



図 4.4: WEB アンケートシステム評価画面

第5章 実験

形態素解析部に MeCab[14]

5.1 実験環境

5.2 比較実験

5.2.1 セリフ文

5.2.2 分類器

5.3 評価実験

5.3.1 評価指標

5.3.2 交差検証

5.3.3 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．

$$\text{正解率 (Ac)} = \frac{TP + TN}{TP + TF + NP + NF} \quad (5.1)$$

$$\text{適合率 (Pr)} = \frac{TP}{TP + FP} \quad (5.2)$$

$$\text{再現率 (Re)} = \frac{TP}{TP + FN} \quad (5.3)$$

$$F \text{ 値} = \frac{2 * Pr * Re}{Pr + Re} \quad (5.4)$$

5.3.4 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6章 結果

6.0.1 グリッドサーチ

表 6.1: ランダムフォレストのグリッドサーチの結果

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 6.2: SVM のグリッドサーチの結果

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

グリッドサーチの結果を表 6.1 表と表 6.2 に示す．それぞれの場合で値が大きく異なるパラメタが得られる場合があった．

6.0.2 評価結果

ランダムフォレストと SVM の結果を表 6.3 と表 6.4 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 6.3: 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

表 6.4: SVM 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	0.59
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

6.0.3 考察

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。

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 4.6 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。

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] Jennifer Maloney, "The Fastest-Growing Format in Publishing: Audiobooks", <http://www.wsj.com/articles/the-fastest-growing-format-in-publishing-audiobooks-1469139910>, Wall Steet Journal
- [2] 佐藤和也, "高い継続率は「耳がさみしくなるから」-オトバンクに聞くオーディオブック市場と利用動向", <https://japan.cnet.com/article/35076656/>
- [3] 上田 渉, "「耳で聴く読書文化」を築く", http://www.ajec.or.jp/interview_width_ueda1/, 一般社団法人日本編集制作協会
- [4] 杉藤美代子; 大山玄. 朗読におけるポーズと呼吸一息継ぎのあるポーズと息継ぎのないポーズー. 音声言語. 1990.
- [5] YOSHIMURA, Takayoshi. Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems. 2002. PhD Thesis. Nagoya Institute of Technology.
- [6] 飯田朱美, and 安村通晃. "感情表現が可能な合成音声の作成と評価." 情報処理学会論文誌 40.2 (1999): 479-486.
- [7] 都築亮介, et al. HMM 音声合成における感情表現のモデル化 (合成, 韻律, 生成, 一般). 電子情報通信学会技術研究報告. SP, 音声, 2003, 103.264: 25-30.
- [8] 波部斉, "ランダムフォレスト", 情報処理学会研究報告 2012

- [9] 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”, 情報科学技術フォーラム講演論文集, 2009:p337-380
- [10] 大谷大和, et al. ”HMM に基づく感情音声合成のための共有感情付与モデル (オーガナイズドセッション「文脈や状況に合った発声を実現する音声合成技術及び周辺技術」, 合成, 韻律, 生成, 音声一般).” 電子情報通信学会技術研究報告. SP, 音声 114.303 (2014): 13-18.
- [11] 布目光生, 鈴木優, 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, 2011:p32-35
- [12] 青空文庫, <http://www.aozora.gr.jp/>
- [13] 大浦 圭一郎, 酒向 慎司, 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, 2010:p343-344
- [14] ”MeCab: Yet Another Part-of-Speech and Morphological Analyze”, <http://taku910.github.io/mecab/>
- [15] ”NAIST Japanese Dictionary”, <http://naist-jdic.osdn.jp/>
- [16] LEE, Akinobu; OURA, Keiichiro; TOKUDA, Keiichi. MMDAgent—A fully open-source toolkit for voice interaction systems. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013. p. 8382-8385.
- [17] 磯健一; 渡辺隆夫; 桑原尚夫. 音声データベース用文セットの設計. 昭 63 年春音講論, 1988, 2-2.
- [18] ”オブジェクト指向スクリプト言語 Ruby”, <https://www.ruby-lang.org/ja/>
- [19] ”Ruby on Rails”, <http://rubyonrails.org/>
- [20] ”SQLite”, <https://www.sqlite.org/>

付 録 A アンケートの指示書