平 成 29 年 度　修 士 研 究 論 文

題目　アニメ動画における性別判定を用いた声優認識に関する研究

指導教員　　服部 峻

提出者　　室蘭工業大学院 情報電子工学系専攻

氏　名　　榮田　基希

学籍番号　　16043009

提出年月日　　平成 30 年 1 月 31 日

目次

図目次

表目次

第 1 章

研究背景

近年，日本には様々な娯楽メディアがあり，普段の生活の中で目や耳にする機会が多くなっている．情報通信機器の普及により多くの人にとって，パソコンやモバイル端末などの機器で番組や動画の視聴，ゲームなどが今では手軽に行うことが出来る．このような娯楽メディアに触れる機会が多くなって来ると，どこかで聞いたことがある誰かの音声が流れて来ることがある．ミュージックビデオ中の歌手の歌声，テレビドラマや実写映画中の俳優が演じる役柄のセリフ音声，アニメ動画中の声優が演じるキャラクターのセリフ音声などが挙げられる．

音声の発生源がアニメ動画の場合，誰の音声であるかを知る為には，エンディングのスタッフロールまで飛ばしたり，Web で作品のタイトル名やキャラクター名で検索したりするなどの余計な労力を掛ける必要が出て来る．例えば，あるユーザが適当なアニメを視聴していた際，そのアニメの中に出て来たキャラクター A の音声がユーザの聞いたことのある音声であったとする．そこで，そのユーザがキャラクター A の声優について調べようとするならば，エンディングまで飛ばしたり，アニメタイトルやキャラクター名で Web 検索して，そのアニメの公式サイトやウィキペディアなどを探そうとしたりするであろう．しかし，知りたいキャラクター A が作中の目立たない配役だった場合，Web で検索を掛けても中々出て来ないことも考えられる．また，主要なキャラクターではない場合，キャラクター名を記憶していない可能性もあり，エンディングのスタッフロールが流れても分からないであろう．その上，脇役であった場合，スタッフロールには男の子 B，男の子 C というようにキャラクター名を不明瞭に表記していることもあり，どの場面に出て来たキャラクターか分からないことも考えられる．

ここで，ユーザが余計な労力を掛けずに，音声の主の声優名を知ることが出来るようにする為には，アニメ視聴中に音声が流れたらリアルタイムに声優名を認識して自動的に画面に表示するシステムが必要になる．本研究で提案するシステムが実現した場合，図 1.1 に示すように，視聴しているアニメの事前知識が無いユーザに知りたい答えを素早く表示することが出来たり，ユーザに知りたい声優に関連する出演アニメや出演ゲーム，イベント事などの付随情報を提供出来たり，また，声優に関する様々な商品の推薦が可能になるなどのメリットが生まれることが考えられる．

これまでの著者の研究 [1] では，アニメ動画から流れる音声波形データと声優データベースに予め登録してある各声優の音声波形データを使って類似度の計算を行い声優を判定する手法

を提案した．しかし，音声を用いた声優認識の精度として良好な結果を得ることが出来なかった．精度が悪かった理由として主に考えられるのは，音声の振幅の波形データを単純に使用していた点と，声優データベースに登録する各声優の音声波形データを無作為に選出していた点が挙げられる．

そこで，本研究では，声優が個々に持つ特有の周波数スペクトルのパターンから「特有スペクトル」を特定出来れば声優認識が可能になると考えた．予め声優データベースに声優の数だけ特有スペクトルを登録しておくことで，視聴中のアニメ動画から流れるある音声の持ち主が誰であるか認識する手法 [2] を提案する．声優データベースに登録しておく特有スペクトルとは，著者が収集した各声優の音声データから時系列毎に周波数スペクトルパターン抽出を行って，その時系列毎に抽出したスペクトル自身の間で良く出現した波形のことである．それを抽出する為に本研究では，自己における周波数スペクトル同士の類似度計算を行って特定するアルゴリズムを提案する．

また，声優認識の精度向上を狙う為，声優認識を行う前に，実際にアニメ動画から流れる音声を用いて，アニメキャラクターの性別判定と音声の持ち主の性別判定を行うシステムを提案する．Web テキストから抽出されたキャスト情報で声優データベースに絞り込みを掛けた後，アニメ視聴中のリアルタイムに取得される音声の持ち主の性別判定や，アニメキャラクターの性別判定を行うことで，更に声優データベースの中から声優の候補を絞り込む為である．音声の持ち主の性別判定を行う為に，音声の主要な要素の 1 つである基本周波数（F0）を用いる．声優データベースには男性声優，女性声優の基本周波数をそれぞれ予め登録しておく．また，著者の研究 [3] の性別判定手法の発展として，その後の研究 [4] は，声質をより細かく解析することで性別判定をより精度良く行うことが出来ると考えて，音声の声域の高低を判定して分類する新たな手法を提案している．

本研究における最終目標としては，アニメ動画が流れている最中に知りたい音声の持ち主である声優名を判定することであるが，その際に複数の問題が生じる．例えば，BGM と声優のセリフの区別や，オープニングやエンディングの楽曲中における音楽と歌声の区別，及び，1 個の場面で 2 人以上の声優のセリフが重なる場合などがある．そこで本研究では，アニメ動画をフルに視聴している状況を想定するのではなく，アニメに出て来るキャラクター（または，ナレーション）が 1 人で話しているシーンの部分を切り取り，それらの動画の音声を声優認識対象の音声として評価実験を行う方法を採用する．

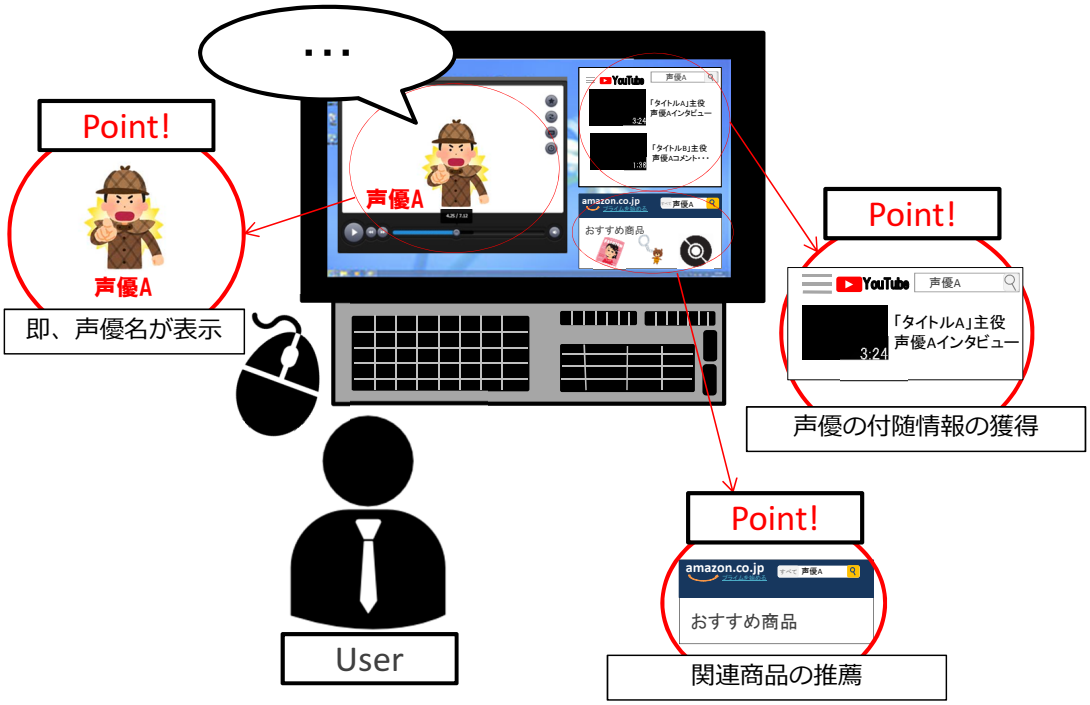


図 1.1: 提案システムのイメージ

第 2 章

関連研究

本章では，関連研究との比較を行うことで，本研究の立ち位置を明確にする．まずは，人物認識の個人の特定方法の中でも，音声に特化した研究についての説明と，本研究との比較を行う．最後に，音声から性別判定を行う為の特徴量について紹介する．

2.1 個人認識

人物の認識に関する関連研究について述べる．人物認識の個人を特定する研究によく用いられる人物の特徴として声紋，指紋，掌紋，虹彩，DNA，顔認識などが挙げられる．これらの各人の固有の生体情報から個人を認識するのは重要な要素技術の一つであると考えられる．また，その中でも認識する対象が実際の人物であったり，動画像であったり音声だけであったり，それらからどのような情報が得られるかによって認識の方法が変わってくる．音声については，話者認識の手法として混合正規分布（GMM）により個人毎の音声の分布を表現する方法 [5] や音声パラメータ系列のモデル化手法として隠れマルコフモデル（HMM）を用いる方法 [6] がある．

話者認識を行う手法は多々ある中，本研究のようにアニメの音声から声優を認識する研究はほとんど見つからない．これまでの著者の研究 [1] では音声波形をそのまま用いて声優認識を試みてみたが，本研究では音声の主要な要素の一つである基本周波数に注視する．文献 [7] では，音声の有声音に限定して，基本周波数推定の方法として，音声の時間波形に対する周期性に着目した分析法やパワースペクトルの調波構造に着目した方法，及び，特徴量を用いた方法などが用いられている．パワースペクトルに着目した方法では，パワースペクトルの最も低いピークの周波数（f_0）を抽出することや f_0 の整数倍にピークを有する調波構造のピーク間隔を推定することでも基本周波数が推定できると記述されている．この従来研究においては，周波数スペクトルのピーク間隔を推定することで基本周波数を推定する方法が記述されているが，本研究では基本周波数を特定することを目的とするのではなく，周期毎（時間毎）にパワースペクトルを取得し，それらのパターンを観測することで声優が個々に持つ特有のスペクトルを特定する手法を確立することが目的である．関連研究で着目している基本周波数と本研究で着目する特有スペクトルの違いを図 2.1，2.2 に示す．但し，本研究においても，性別判定

では F0 を活用する．本研究では，多数の人の音声特徴を登録しておき誰がいるかを判定する不特定話者認識 [8] を採用する．

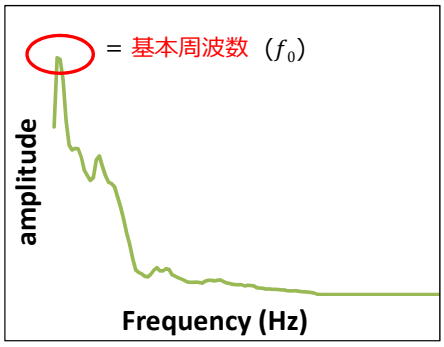


図 2.1: 関連研究の基本周波数（F0）

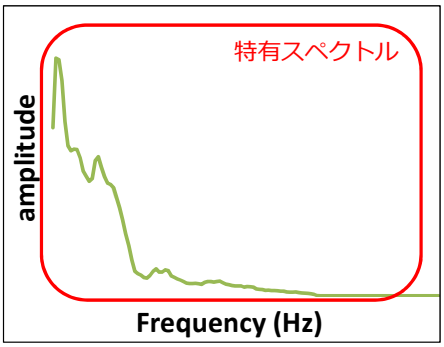


図 2.2: 本研究の特有スペクトル

2.2　性別判定における基本周波数の有用性

朗読した音声から女性の声として判定する方法として，声の高さ，イントネーション，語尾，強弱，速度などの声質が起因するように考えられている [9]．性同一性症者（MtF）や生物学的女性を対象に，女性の声の高さに注目をして，分析を行っている文献 [9,10] から，女性の声の声域の高さの目安は，極端に高くもなく低くもない F0 の値が好ましいと言えることが分かった．聴取者に朗読した音声を聞かせて，女性の音声であると判定した時の F0 値データから，180Hz～230Hz，もしくは 155Hz～254Hz 辺りの声が女性らしく聴こえる高さである．その朗読した音声が女性であると判定された F0 値の平均が 217Hz の値である．結果から，女性判定率と F0 値の関係は，ある高さまでには相関があると示している．また，女性声に関して，メイドが出すような独特な声は地声よりも仕事中の演技声の方が F0 値の平均が高く，上昇する時の振り幅が大きくなることが示されている [11]．

これらの文献を参考に，性別毎に声優の音声から抽出した F0 値を利用して，分析対象である音声の持ち主の性別判定が出来ると仮説を立てた．本研究では，収集した F0 値から，確率密度関数の一つである正規分布に従って，音声の持ち主の性別判定を行う．

また，声優は普段からボイストレーニングを行っており，MtF や一般人とは違い，男性声優，女性声優と共に地声で発生する声域が広域であると仮説を立てた．このボイストレーニングを行うことは，声帯手術より，女声の獲得に有効である [10]．前述した仮説から，本研究では声優の声域における高低分割を行い，本研究で提案した性別判定手法の改良を試みる．

第 3 章

提案システム

本章では，音声の持ち主を特定するシステムについて提案する．これまで著者は，音声の振幅に基づく類似度計算を用いた声優認識システム [1] を研究開発して来たが，図 3.1 のようにキャラクターの性別判定や声優の性別判定を新たに導入することで，声優データベースに登録されているキャスト情報から認識される声優の候補を絞り込み，声優認識の精度向上を試みる．

提案システムの流れについて説明する．まず初めに，「声優認識」の処理を行う前に図 3.1 に示した「キャスト情報の獲得」の処理で，視聴しているアニメタイトルから Web 検索を行い，まうまう [12] や Wikipedia [13] といった，アニメ作品に出演した声優をまとめた Web サイトのテキストからキャスト情報を取得して，声優データベースに対して声優名の候補に絞り込みを掛ける．この処理を行うことにより，声優データの膨大な人数の候補から，声優を特定するまでの処理時間の増加や声優認識精度の悪化といった問題が解消されることが期待出来る．

次の処理で，分析対象であるアニメ動画中の音声が提案システムに入力されると，「キャラクターの性別判定」が行われる．歌舞伎の女形や宝塚歌劇の男役を除き，ほとんどの場合，俳優の性別と，その俳優が演じる役柄の性別は同一であるが，一方，声優の性別と，その声優が演じる（アニメや吹き替えの）キャラクターの性別とは異なる場合も少なくない．しかし，アニメ界の現状において，アニメキャラクターの男性役を担当する声優は，男性，女性と共にキャスティングされている傾向が見られるが，アニメキャラクターの女性役を担当する声優においては，女性が多い傾向があることが分かった．そこで，図 3.1 のキャラ性別判定において，アニメキャラクターの性別を判定することが出来れば，そのキャラクターの音声の主の性別判定を精度良く行うことに繋がると考えた．例えば，あるアニメのある話数に登場するアニメキャラクターが男性一人に女性多数といった場合，アニメ動画に流れる音声からキャラクターが男性の性別であると判定出来れば自動的に声優を絞り込むことが出来たり，アニメ界の現状から声優認識の対象であるアニメキャラクターの性別が女性であると判定出来た場合，その音声の主が男性声優である確率が下がるといった手法を施すことが出来る．

実際に流れる動画から，キャラクターの性別判定を行う時に，声優認識のターゲットとなる音声の前後のセリフ内容の単語をコンテキスト情報として活用することも考えられる．例えば，「兄さん」「姉さん」といった単語が出現した時，その呼称に当たるキャラクターがその場

面に登場する確率が高いと考えた．また，性別を表す単語だけでなく，「先生」「隊長」といったキャラクターの属性情報を獲得出来れば，キャラクターの名称を判定して芋づる式に声優名が判定出来ると考えた．

セリフ内容からアニメキャラクターの性別や名称を判定する為には，予めキャラクターデータベースに声優が演じているキャラクターの名称や性別，属性といった情報を登録しておく必要がある．それらを人の手で実行すると，膨大な数のアニメ作品からキャラクターの性別や属性を確認する手間やコストが掛かる．一人のキャラクターに対して複数の声優が演じている場合も少なからずあるが，一人の声優が複数のキャラクターの音声を演じている方が圧倒的に多いのが現状であり，キャラクターデータベースにおいて膨大な規模になる．そこで本研究では，前準備として，キャラクターデータベースに登録する正しい性別を自動的に獲得する為に，Web テキストを用いたキャラクターの性別判定手法を提案する．

次に，図 3.1 に示す「声優性別判定」で，アニメ動画中の音声の基本周波数（F0）を用いた声優の性別判定を行う．男女の声の違いは，ピッチや抑揚，フォルマント周波数などの情報もあるが，安定した測定値を得ることが出来る基本周波数を用いる．音声の持ち主の性別判定を行うことで，キャスト情報により絞り込まれた声優データベースから更に声優の候補を絞り込むことにより，次に行う声優認識の精度向上が期待される．

また，その性別判定手法の発展として，声質をより細かく解析することで性別判定をより精度良く行うことが出来ると考えて，図 3.1 の「声優性別判定」の部分にターゲットとなる音声の声域の高低を判定して分類する新たな手法も本提案システムに組み込む．図 3.2 に示すように，本研究のシステムは声優認識を行う際，声優データベースに対してキャスト情報の絞り込みを行った上で更に声優の性別で絞り込みを掛けるパターンと，性別判定を行わないでキャスト情報の絞り込みだけを行ったパターンの 2 種類に大別出来る．性別判定の手法にも，声域分割を行わない性別判定手法を組み込んだ場合，そして，声域を分割してから性別判定する手法の 2 つのパターンがある．性別判定を組み込むことで，声優認識の精度にどのような影響を与えるのか，詳細に評価実験していく．

最後に，図 3.1 の「声優認識」において，特有スペクトルを用いた声優名の認識を行う．アニメ動画に流れる音声から声優名を認識するため，声優に限定しない一般の話者認識に関する従来研究 [14,15] を参考にして，人それぞれには周波数毎に個人差があると考えた．そこで，人の声質にはバンド幅毎にそれぞれ違う強さを持つと仮説を立て，声優が個々に持つ特有の周波数スペクトルのパターンである「特有スペクトル」を特定できれば声優認識が可能になると考えた．この仮説から，声優の人数分の特有スペクトルをアニメ動画やラジオ，事務所で公開されているサンプルボイスなどの音声データから分析し，声優毎に一つずつ声優データベースに予め登録しておく．本研究では，声優データベースに登録されている各声優の特有スペクトルと，アニメ動画から流れる音声データの周波数スペクトルとの間の類似度計算を，声優データベースに登録されている声優の数だけ行うことによって，その音声の持ち主が誰であるかを判定するシステムを提案する．

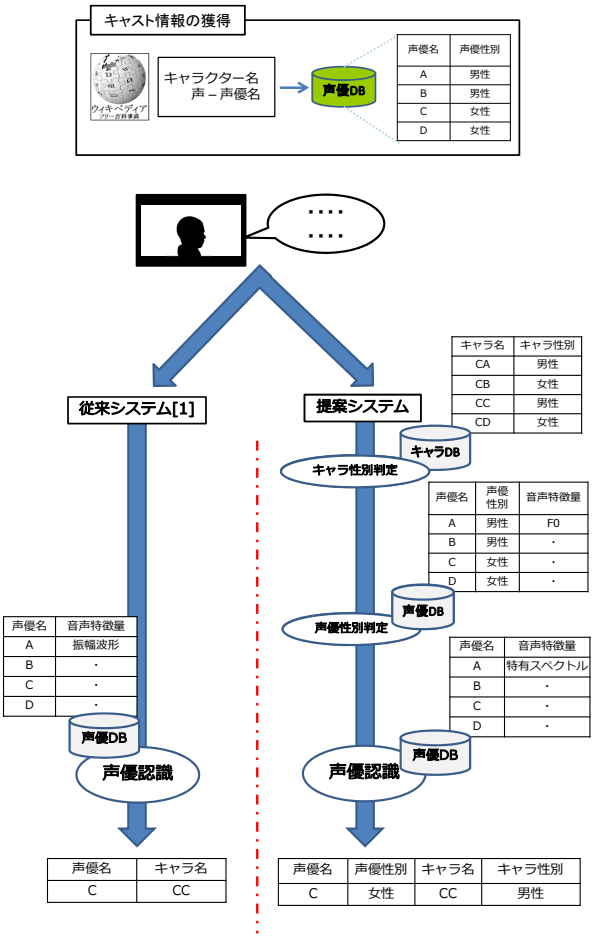


図 3.1: システム全体像

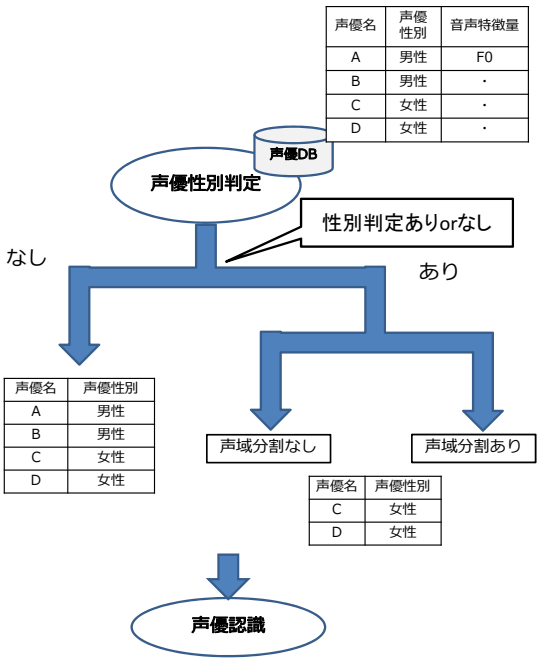


図 3.2: 性別判定手法の種類

第 4 章

声優データベース

声優認識，及び，声優性別判定を行う為には図 4.3 に示すように声優データベースを整えておく必要がある．そこで，まず初めに声優性別判定の為の事前準備として，男女声優それぞれのサンプル音声データから，再生時間の時系列分だけ取得される基本周波数（F0）値のベクトルデータを，音声の持ち主の属性情報として声優データベースに登録しておく．一人の声優に対するサンプル音声のデータ数やその音声データの再生時間は，それぞればらけている．F0 の取得には，図 4.1，4.2 に示すようにアムステルダム大学の Paul Boersma 氏と David Weenink 氏によって開発されたオープンソース・ソフトウェア Praat [16] を用いた．F0 を抽出する為の Praat の設定として，標本化周波数を 16000 Hz に，Pitch の範囲については Pitch floor を 75 Hz，Pitch ceiling を 500 Hz に設定する．この声優データベースに登録されている声優は予め性別毎に分類されており，音声の持ち主の様々なセリフの音声特徴量が，音声の持ち主と紐付けられて登録されている．

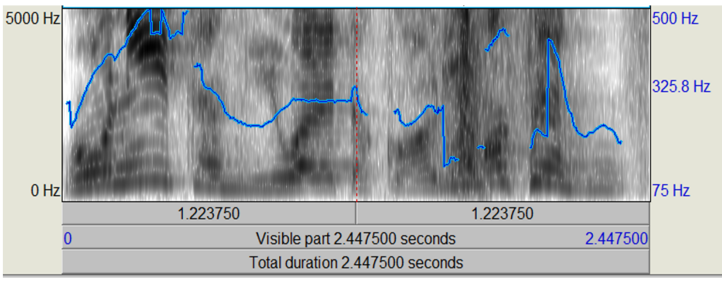


図 4.1: Praat による F0 抽出

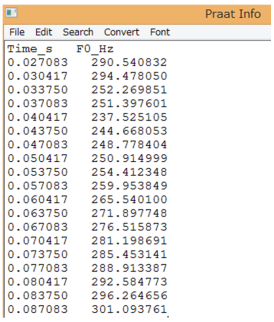


図 4.2: F0 の抽出例

また，声優認識を行う為の音声特徴量の一つとして，個々の声優の声質の特徴が表れている特有スペクトルを登録する必要がある．この特徴量は，個々の声優の声質の特徴である周波数スペクトルであり，声優毎に一つに特定されている．この特有スペクトルを取得する為の音声データは，F0 値を取得する際に使用した音声データと同等のものである．Python の scipy のライブラリを利用して FFT を行い，図 4.4 に示すように周波数スペクトルを出力した．

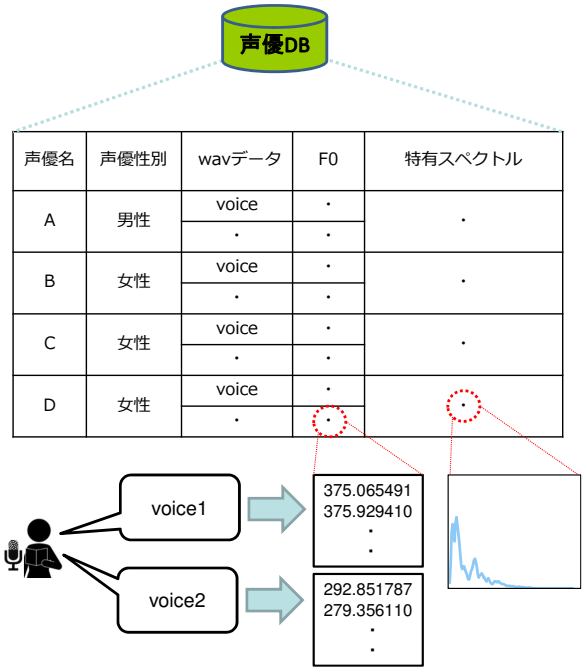


図 4.3: 声優データベースの詳細

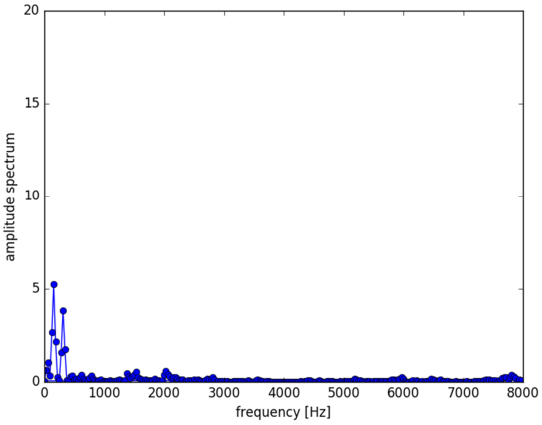


図 4.4: スペクトル波形

声優の特徴が表れているスペクトルを見付ける方法として，ある声優の音声が流れている動画から取得できる周波数スペクトルの数値から，頻出している類似パターンを探し出すことである．良く頻出している周波数スペクトルの類似パターンを見付け出す為に，類似性の計算式として，コサイン類似度を用いる．

声優の特有スペクトルを見付ける方法として，まず初めに声優データベースに登録する声優の音声が流れている数多のアニメ動画やラジオ動画，サンプルボイスの音声データから周波数スペクトルの各バンド幅分の数値を取得する．

目的の声優の音声が流れている部分だけの動画を切り取る際には以下のことに気を付ける．

- BGM が無い
- 効果音が無い
- 時間が短い音声は入れない（例：相槌など）
- こもっている音声を入れない（例：心の声など）
- 他の人の音声とかぶっていない

これらの注意点は，出来る限りノイズを含まない生の音声データの周波数スペクトルの数値を取得する為である．次に，切り取った動画の音声データから周波数スペクトルの数値を取得した後，各声優毎にそれらを一つにまとめる．最後に，声優毎に一つにまとめたスペクトルの集合に対して，繰り返し頻出しているパターンを解析して特有スペクトルを探し出す．

第 5 章

Web テキストを用いたキャラクター性別判定

アニメ視聴中にキャラクターの性別判定を行う為にはキャラクターの性別を予め登録しておく必要がある．まうまう [12] や Wikipedia [13] といった，アニメ作品に出演した声優をまとめた Web サイトのテキストから必要な情報を抽出できれば，キャスト情報を知ることが可能である．しかし，Web テキストから声優とその声優が演じているキャラクターの情報を紐付けられたとしても，キャラクターの性別まで必ず一緒に明記していることは少なく，直接的に知ることは困難である場合が多い．データベースにキャラクターの正しい性別を登録する為に，キャラクターの情報が載っている Web テキストを自然言語処理し解析を行うが，本章では，どのような語彙を含む辞書を作成すれば良いか，また，その辞書を用いてどのような処理を行えばキャラクターの性別を精度良く判定出来るか，検討を行う．

5.1　キャラクター性別判定辞書の選定

本研究では，アニメキャラクターの情報が載っている Web テキストとして，Wikipedia [13]，ピクシブ百科事典 [17]，ニコニコ大百科 [18] を参考にした．解析するアニメキャラクターの情報が載っている Web テキストから，対象とするキャラクターの正しい性別を自動獲得する為に，男性，女性それぞれの性別を表す単語群から成る辞書を用意する．本研究で用意した 5 種類の辞書のそれぞれの違いについて下記に示す．

- 辞書 1（単語数：男 2 件，女 2 件）
  - 男性の辞書には「男」「男性」の単語のみ，
    女性の辞書には「女」「女性」の単語のみ含まれている．
- 辞書 2（単語数：男 20 件，女 20 件）
  - 男性の辞書，女性の辞書共に，著者らが主観で決めた複数の単語が含まれている．
    表 5.1 に全てを示す．
- 辞書 3（単語数：男 3385 件，女 2833 件）
  - 男性の辞書，女性の辞書共に，日本語シソーラス連想類語辞典 [19] の
    Web サイトを利用して検索フォームや索引から，辞書 2 に含まれている単語を
    入力して出力された類語，同義語，連想語の集まりである．
- 辞書 4（単語数：男 206 件，女 379 件）
  - 男性の辞書，女性の辞書共に，日本語 WordNet [20] を利用して，辞書 2 に含まれている単語を入力して出力された下位語の集まりである．
- 辞書 5（単語数：男 3472 件，女 3059 件）
  - 男性の辞書，女性の辞書共に，辞書 3 と辞書 4 に含まれている単語を混合させたものである．

辞書 1 は，男性辞書には「男」「男性」の単語だけを，女性辞書には「女」「女性」の単語だけを登録した最も単純な辞書であると言える．辞書 2 は，キャラクター情報が載っている Web テキストを著者が分析して，上手く性別判定出来るであろう単語を推測して手動で登録した．辞書 2 に登録した単語の種類は，性別を指す名詞や代名詞，家族構成を表す単語，気質（性質）を表す単語である．辞書 3 は，日本語シソーラス連想類語辞典から，辞書 2 に含まれている単語を指定して検索フォームや索引から出力された類語，同義語，連想語を登録した．辞書 4 は，日本語 WordNet から，辞書 2 に含まれている単語を指定して出力された下位語を登録した．日本語 WordNet とは，プリンストン大学で開発された Princeton WordNet の synset に対応して日本語が付与された，日本語の概念辞書のことであり，自由にダウンロードが出来る．その概念辞書は，語概念を上位下位関係などの多様な関係でまとめているデータベースである．以上の辞書の関係を図 5.1 に示す．辞書 5 は，辞書 3 と辞書 4 に含まれている単語を組み合わせた辞書である．

また，男性辞書と女性辞書との間で単語の重複が出て来る可能性や自身の辞書に含まれる単語同士で重複する可能性を考慮して，いずれの辞書にもノイズ除去処理を施している．

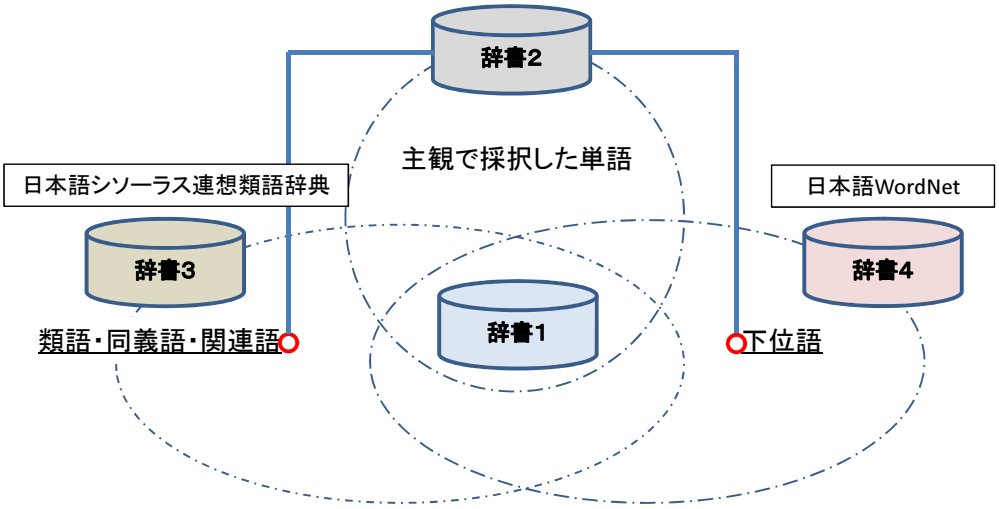


図 5.1: 辞書間の関係

表 5.1: 辞書 2 の単語一覧

男性辞書	男 / 男性 / 男子 / 男の子 少年 / ボーイ / 婿 / 父 夫 / 兄 / 弟 / 翁 祖父 / 息子 / 彼 / 王子 ナルシスト / イケメン / ハンサム / 青年
女性辞書	女 / 女性 / 女子 / 女の子 少女 / ガール / 嫁 / 母 妻 / 姉 / 妹 / 婆 祖母 / 娘 / 彼女 / 王女 ブリッ娘 / レディ / ギャル / ヒロイン

5.2　キャラクター性別判定アルゴリズム

本節では，アニメキャラクターの情報が載っている Web テキストから，対象キャラクターの性別を判定する手法について述べる．前節で用意した辞書を用いて Web テキストを解析して，各キャラクターの性別を判定する．キャラクターの紹介が記載されている Web テキスト内には，キャラクターの性別に応じた関連語や連想語などが頻出していると仮説を立てた．例えば，男性キャラクターの紹介欄の場合，Web テキストには「お父さん」「兄」などの男性を連想する語が多く含まれ，女性キャラクターの紹介欄の場合，Web テキストには「お母さん」「姉」などの女性を連想する語が多く含まれていると予測した．さらに，キャラクター紹介の文章の大部分は一人称視点や二人称視点で記載されているのではなく，三人称視点で記載されており，男性の場合「彼」，女性の場合「彼女」といった単語で紹介されている場合が多い．また，紹介されている対象キャラクターについての詳細が，図 5.2 に示すように，主に前半部分に集中して記載されている傾向があると考えて，男女それぞれの性別判定辞書に含まれる単語の出現順に重みを変化させて付与する必要があると考えた．

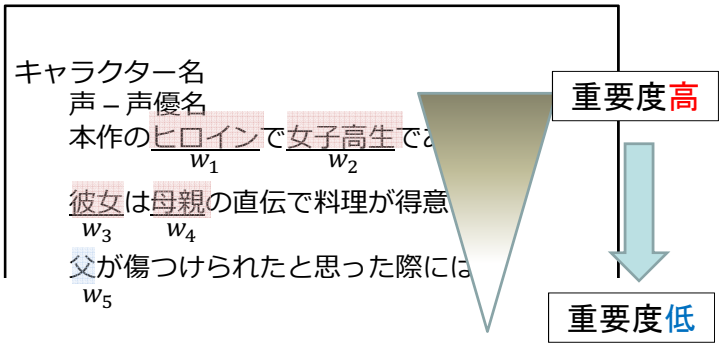


図 5.2: Web テキストに出現する注目単語の位置

前述したキャラクター紹介の文章に関する考察から，それぞれのキャラクターに対して，その Web テキストから性別に関連する単語の出現回数を数えることでアニメキャラクターの性別判定が可能であると考えた．そこでまず，日本語係り受け解析器である CaboCha [21] を使用して，Web テキストの文章を対象に形態素解析を行い，単語を抽出する．Web テキストを対象に，初めの文章から順番に解析して抽出した単語が，男女の性別判定辞書に含まれている単語であった場合，マッチした単語順に k 番目の単語を w_k（$k \in \{1, 2, 3, \cdots, N\}$）と表し，男女それぞれの辞書において出現回数をカウントする．単語 w_k の男性の辞書 D_M における出現回数を関数 $DC_M(w_k)$ で，女性の辞書 D_F における出現回数を関数 $DC_F(w_k)$ で定義する．さらに，男女の辞書に含まれる単語の出現順毎に重みを付ける為，減衰率 δ（$\delta \in \{0.00, 0.01, 0.02, \cdots, 1.00\}$）を設けて，単語 w_k に対してカウントする値を δ^{k-1} とした関数を以下に示す（前述の $DC_{M/F}(w_k)$ では，加算されるのは常に 1.00 である）．Web テキストの解析が終了するまでに行われる，場合分けによる男性の辞書における最終的な重み付き出現回数 S_M，女性の辞書における最終的な重み付き出現回数 S_F のカウント方法を以下の数式と図 5.3 に示す．

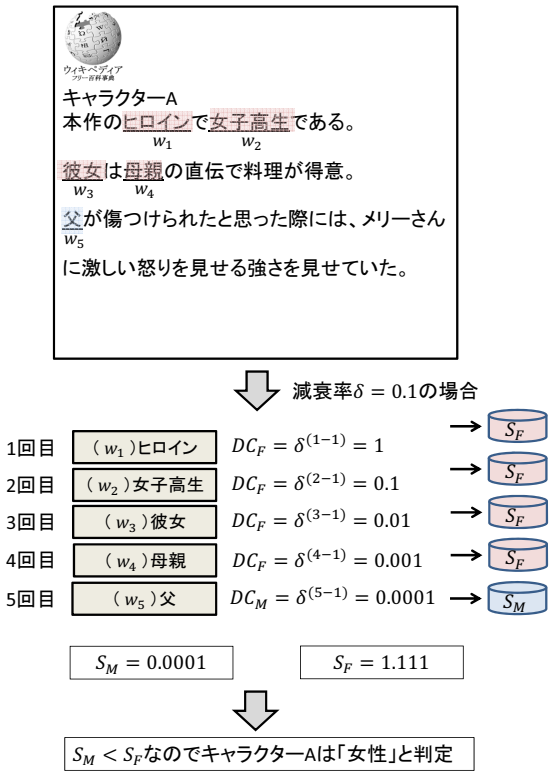


図 5.3: キャラクター性別判定手法

1. 男性辞書 D_M における出現回数（減衰率 δ 付き）

$$DC^*_M(w_k) = \begin{cases} \delta^{k-1} & (w_k \in D_M \text{のとき}) \\ 0 & (otherwise) \end{cases}$$

$$S_M = \sum_{k=1}^{N} DC^*_M(w_k)$$

2. 女性辞書 D_F における出現回数（減衰率 δ 付き）

$$DC^*_F(w_k) = \begin{cases} \delta^{k-1} & (w_k \in D_F \text{のとき}) \\ 0 & (otherwise) \end{cases}$$

$$S_F = \sum_{k=1}^{N} DC^*_F(w_k)$$

Web テキストの文章を全て解析した後，男女に関する辞書の単語の出現回数（減衰率付き）が大きかった性別を採択する．

$$S_M > S_F \rightarrow \text{男性キャラクター}$$
$$S_M < S_F \rightarrow \text{女性キャラクター}$$
$$else \rightarrow \text{判定出来ず}$$

第 6 章

声優性別判定

本章ではアニメ動画から流れる音声から，その音声の持ち主の性別判定を行う手法について説明する．まず，音声の F0 に基づいた性別判定手法の説明を行い，次に，音声の高低に基づいた声域分割の性別判定の概要と手法の説明を行う．

6.1　基本周波数に基づく声優性別判定

実際にアニメ動画から流れる音声から，その音声の持ち主である声優の性別を判定する為，声優に限定しない性同一性障害者の話声位に関する研究 [9] を参考にして，声優の特徴である「アニメ声」であっても男女によって F0 の値に差が生じると仮説を立てた．そこで，声優データベースにおける男女それぞれの F0 に対して特徴を突き止めることにより，分析対象であるアニメ動画から流れる音声データの性別判定が可能になると考えた．この仮説から，まず初めに行う工程として，男女声優それぞれのサンプル音声の収集を行って，それらの音声データから取得される F0 値を声優データベースに登録しておく．この男女毎に登録されている声優データベースの F0 値の標本データから，男性の声質の特徴を表しているであろう F0 の値と，同様に，女性の声質の特徴を表しているであろう F0 の値の両方を探し出す．その声優データベースにおける男女の声質の特徴を見つけ出して，それを活用する手法として，本研究では，確率密度関数の一つである正規分布の計算式を使用することにより，分析対象であるアニメ動画から流れる音声の持ち主の性別の判定を行う．

まず初めに，図 6.1 に示すように，正規分布の計算の為に，声優データベースに登録されている F0 の標本データを用いて，男女毎に平均 μ，分散 σ^2 を求めておく．分析対象である声優の音声データから得られた F0 の羅列であるテストデータを x_i（$i \in \{1, 2, 3, \cdots, T\}$）と表す．$x_i$ を対象に，男女各々の正規分布において算出された確率を別々に加算していく（以下，PS）．男性の正規分布に基づいて加算された PS を PS_M，女性の方を PS_F と表す．性別判定が行われるまでの計算式を以下に示す．f は正規分布の確率密度関数である．

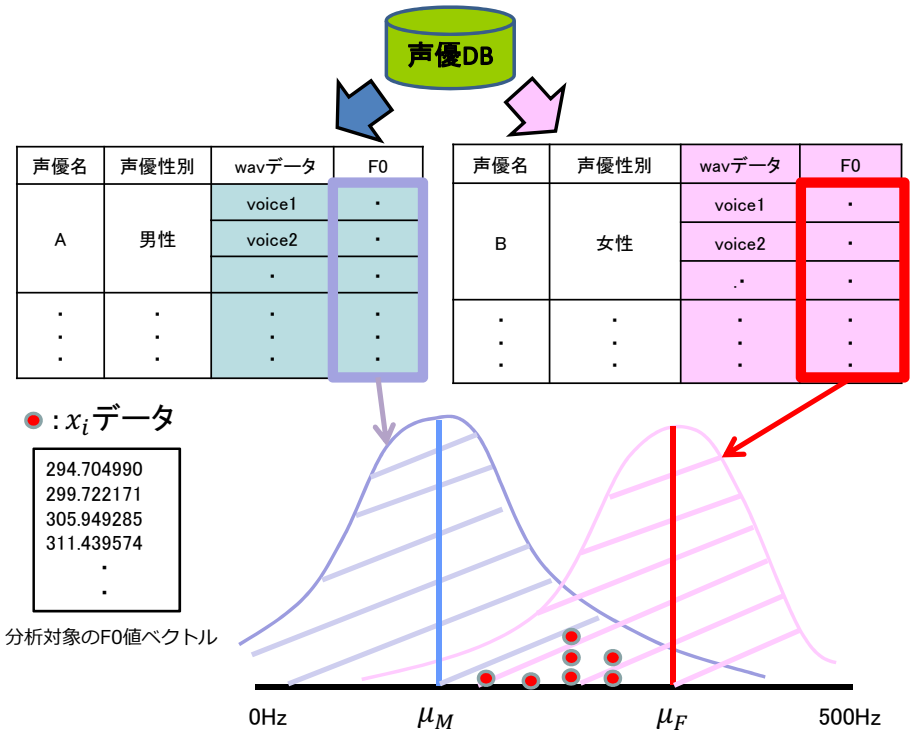


図 6.1: 正規分布による性別判定

$$f(x_i, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

$$PS_M = \sum_{i=1}^{T} f(x_i, \mu_M, \sigma_M)$$

$$PS_F = \sum_{i=1}^{T} f(x_i, \mu_F, \sigma_F)$$

テストデータ x_i が T まで計算された後，最終的な PS_M と PS_F の大小比較を行い，大きい値であった性別を採択する．

$$PS_M > PS_F \rightarrow \text{男性声優}$$
$$PS_M < PS_F \rightarrow \text{女性声優}$$
$$else \rightarrow \text{判定出来ず}$$

6.2　声優の音声を用いた声域分割による性別判定

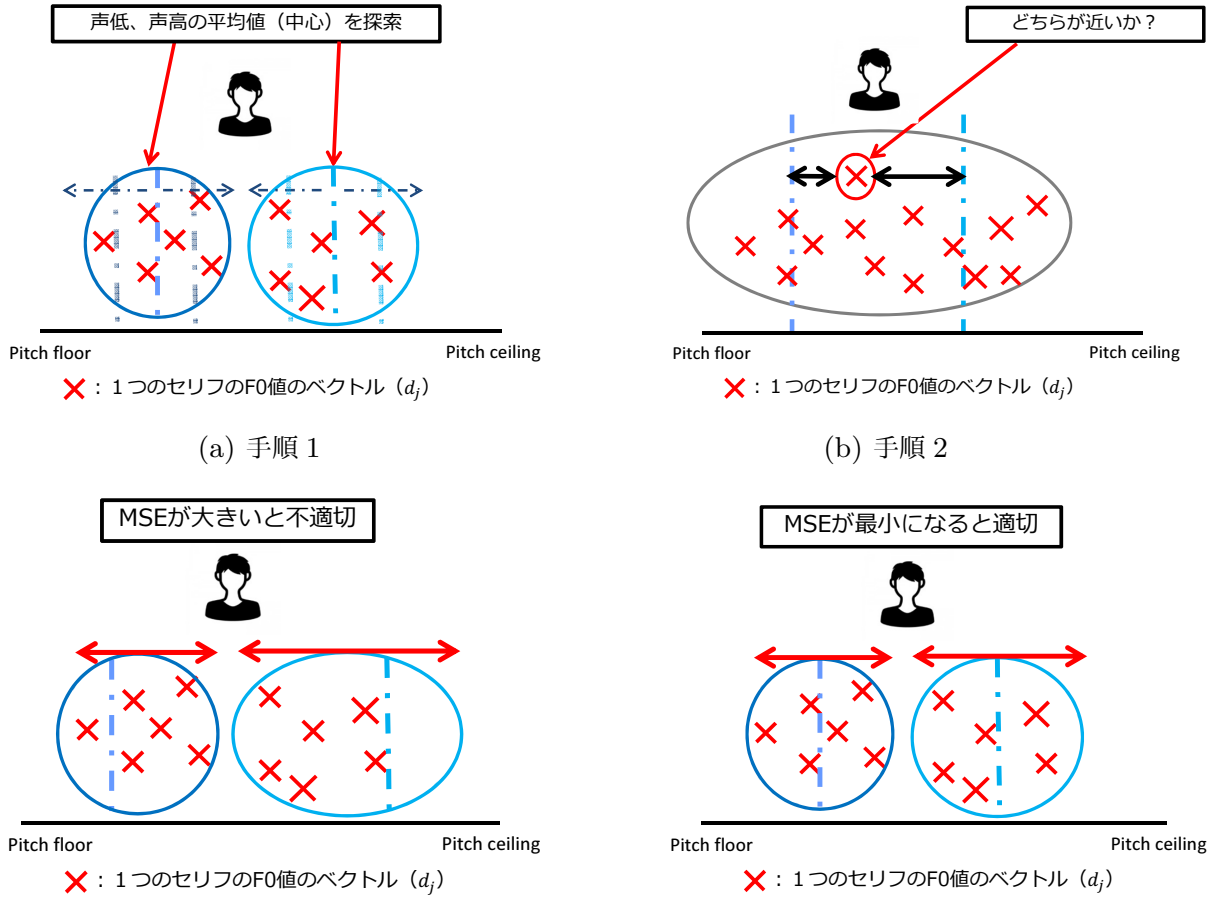
まず，声優の音声における声域分割の概要について述べる．次に，声域分割のアルゴリズムの説明と声域分割を用いた性別判定手法の説明を行う．

6.2.1　声域分割手法の概要

前節の手法による性別判定を行った結果，精度良く性別判定を行えたが，実験対象のある男性声優に対してだけは，全体の精度よりも比較的悪かったと言える．その男性声優の声質は他の男性声優と比べて，主観ではあるが比較的高いと感じた．この考察から，声域が高い男性声優のセリフの F0 の集合が，女性である特徴を表していると従来手法による性別判定システムに判定されたと考えられる．

以上のように，これまでの研究で行った評価実験に対する考察を踏まえて，男女共に声優は職業柄，養成所などの施設でボイストレーニングを行っており，一般人の声域とは違い広域であると仮説を立てた．この仮説から，声優に関して性別毎に声域を高低に分割することが可能であると考えた．予め声優データベースに登録されている各声優の音声データを，性別毎に声域の高低に基づいて分類を行っておくことで，分析対象である音声が男性で声質が高い人，女性で声質が低い人であっても性別判定を精度良く行えると考えた．本研究では，声優データベースに登録されている各声優のセリフの F0 値の標本データ（以下，d_j）が，声域の高低のどちらかに属するのかを判定する為に，F0 に基づいて分割する手法を提案する．

声域の高低判断は実際に人間が音声を聞くことにより，感覚的に判定することは可能であるが，コンピュータに分類を任せると，どのような大きさを持つ F0 値が声域において高いのか，低いのか，境目となる値の判断がつかない．そこで，図 6.2 に示すように，まず初めに，図 6.2(a) で男女それぞれにおいて声域が低いセリフのクラスタ，高いセリフのクラスタの平均値を Pitch floor から Pitch ceiling の範囲で探索を行う．図 6.2(b) で，その 2 つの平均値から d_j がどのくらい数値が離れているか，大小比較を行い，より小さい（距離が近い）クラスタの平均値を持つ方へ振り分ける工程を全ての d_j において行う．声域の高低のクラスタの組み合わせの，それぞれのクラスタの範囲が図 6.2(c) に示すように平均二乗誤差（MSE）が最小でなければ分類は失敗として扱い，図 6.2(d) に示すように最小であったならば分類は成功と見なす．



(a) 手順 1　(b) 手順 2

(c) 分類失敗例　(d) 分類成功例

図 6.2: 性別判定のクラスタ分割イメージ

6.2.2　声域分割アルゴリズムを用いた性別判定手法

本節では，前節に記載した手法のアルゴリズムの詳細説明を行う．高低における Pitch range（75〜500Hz）のパラメータの組み合わせ (b_{low}, b_{high}) と，d_j との平均二乗誤差（MSE）が最小になる組み合わせを探索する．分類の最適解を求める為，図 6.3 に示すように，ある時点での b_{low}，b_{high} と d_j との MSE を求めて，b_{low}，b_{high} のうち MSE が小さかった方の値を SUM に加算するという作業を全ての d_j に対して行う．この工程を全パターンの組み合わせ (b_{low}, b_{high}) に対して行い，SUM の値が最小になった組み合わせの時の分類結果を最適解とする．高低それぞれに分類された d_j を用いて男女毎における声高の平均値 μ_{high} と分散 σ^2_{high}，声低の平均値 μ_{low} と分散 σ^2_{low} を求める．しかし，性別各々において分割された後の分散 σ^2_{high}，σ^2_{low} が分割する前の分散 σ^2_M あるいは σ^2_F よりも値が大きかった場合，分割は不適切として行わない．

分析対象の音声の性別判定の手法として，図 6.4 に示すように，分析対象であるアニメのセリフの音声データから得られた F0 値のベクトル x_i が入力された時，男女毎の声域の高低それぞれの正規分布の数式において算出された確率を，それぞれの関数 PS_{M_low}，PS_{M_high}，PS_{F_low}，PS_{F_high} に加算する．x_i が T まで正規分布の数式により計算された後，関数

PS_{M_low}，PS_{M_high}，PS_{F_low}，PS_{F_high} の値で大小比較を行い，最も大きい値であった関数 PS が属している性別を採択する．

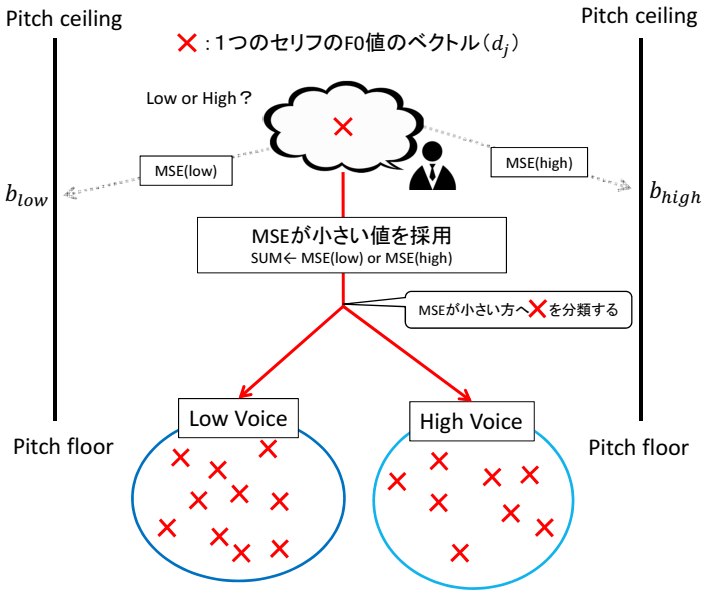
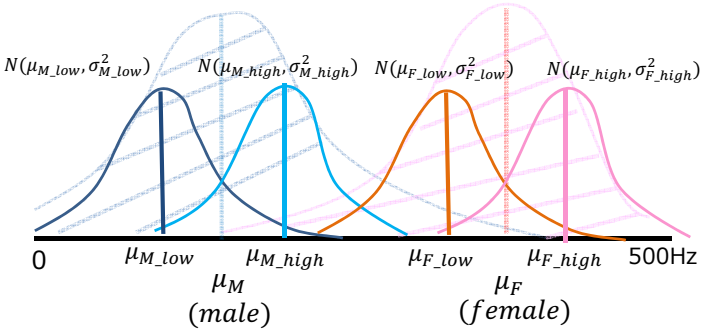


図 6.3: F0 値のベクトルのクラスタの振り分け



$$PS_{M\_low} = \sum_{i=1}^{T} f(x_i, \mu_{M\_low}, \sigma_{M\_low}) \qquad PS_{F\_low} = \sum_{i=1}^{T} f(x_i, \mu_{F\_low}, \sigma_{F\_low})$$

$$PS_{M\_high} = \sum_{i=1}^{T} f(x_i, \mu_{M\_high}, \sigma_{M\_high}) \qquad PS_{F\_high} = \sum_{i=1}^{T} f(x_i, \mu_{F\_high}, \sigma_{F\_high})$$

最終的なPSが最大である性別と判定する

図 6.4: 声域分割による性別判定

第 7 章

声優認識

本章では分析対象である音声の声優認識について述べる．まず，個々の声質を表す特有スペクトルの探索のアルゴリズムの説明を行う．最後に，分析対象の音声の持ち主が誰であるか判定する手法について説明を行う．

7.1　特有スペクトル探索アルゴリズム

図 7.1 のように，ある声優における音声データから取得した周波数スペクトルの時系列パターンの中で，自己と類似するパターンが繰り返し頻出する特有スペクトルを探索する為の類似性の計算式について説明する．各声優毎に一つにまとめた周波数スペクトルの時系列パターンを $\boldsymbol{P_t}$（$t \in \{t_1, t_2, \cdots, T\}$）と表す．ある瞬間の t 秒における周波数スペクトル $\boldsymbol{P_t}$ のバンド幅の数は，FFT を行う際のサンプル数（N）の変動による周波数分解能の性能や，著者によって使用するバンド幅 B を定めることで左右される．$\boldsymbol{P_{t'}}$（$t' \in \{t_1, t_2, \cdots, T\}$）は，$\boldsymbol{P_t}$ のコピーであり，全く同じ周波数スペクトルの時系列パターンを表している．本研究では，自己の時系列パターンとの類似性を計算する関数 $\mathrm{Auto}(\boldsymbol{P_t}, \boldsymbol{P_{t'}})$ として，以下のコサイン類似度 $\mathrm{sim}(\boldsymbol{P_t}, \boldsymbol{P_{t'}})$ を用いる．

$$\boldsymbol{P_t} = (P_{t,1}, \cdots, P_{t,B}), \quad \boldsymbol{P_{t'}} = (P_{t',1}, \cdots, P_{t',B})$$

$$\mathrm{sim}(\boldsymbol{P_t}, \boldsymbol{P_{t'}}) = \frac{\sum_{j=1}^{B} P_{t,j} \cdot P_{t',j}}{\sqrt{\sum_{j=1}^{B} P_{t,j}} \sqrt{\sum_{j=1}^{B} P_{t',j}}}$$

本節では，各声優のセリフ毎に切り取った動画から取得した周波数スペクトルの時系列パターンを一つにまとめたファイルから，声優の特有スペクトルを探索する手法について説明する．図 7.1 のように，時系列毎のあるパターン $\boldsymbol{P_t}$ を一つずつベースとして，もう一つの比較対象の全パターン $\boldsymbol{P_{t'}}$ 各々との類似度を計算する．計算した類似度が予め定めた閾値を上

回った場合，ベースにしているパターン $\boldsymbol{P_t}$ のカウント数を増やす．$\boldsymbol{P_t}$ と $\boldsymbol{P_{t'}}$ との全ての組み合わせの計算を終えた後，閾値を上回ったカウント数を一番多く獲得した $\boldsymbol{P_t}$ を，動画の音声データから得られた音声の持ち主の特有スペクトルと特定する．しかし，閾値を上回ったカウント数が同数になることがある．その場合に，計算から得られた類似度のうち閾値を上回った場合のみを加算した平均で比較して，類似度の平均が最も大きかったパターン $\boldsymbol{P_t}$ を動画の音声データから得られた音声の持ち主の特有スペクトルとして特定する．

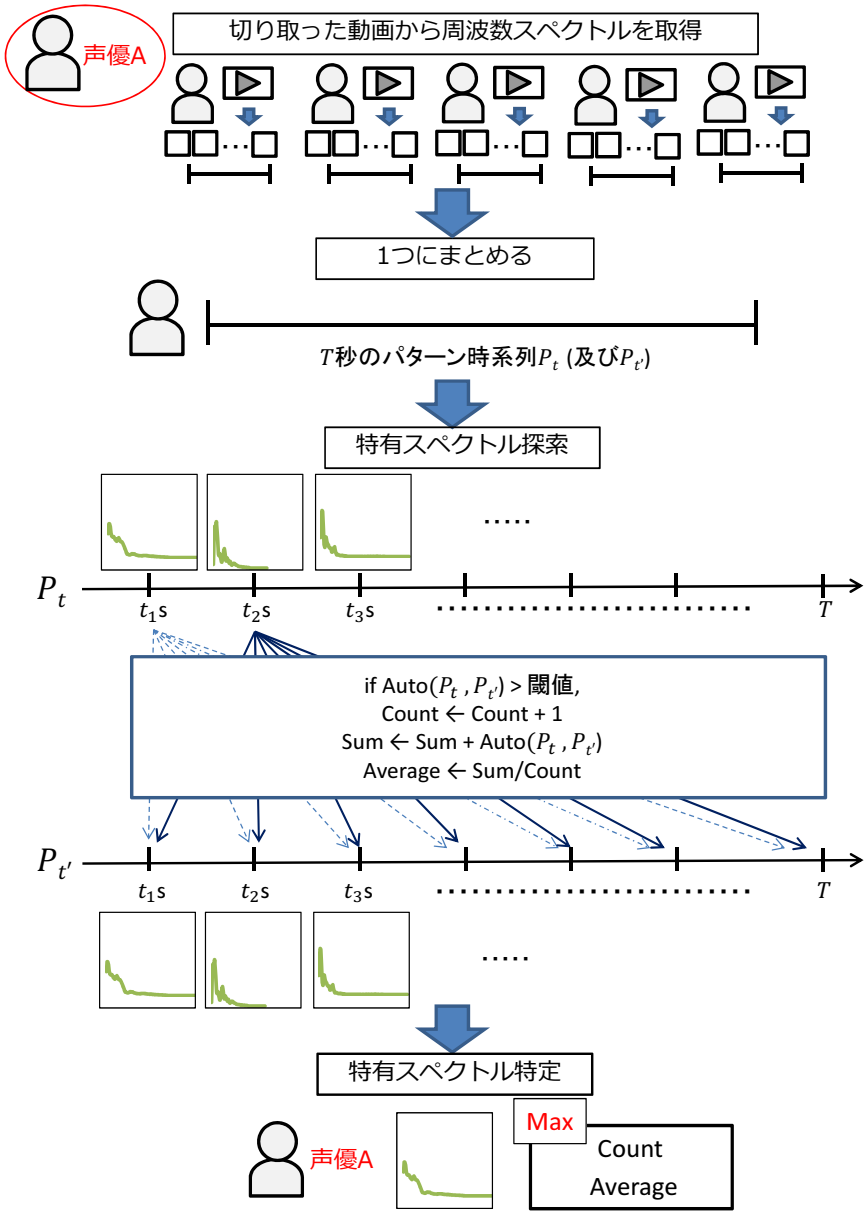


図 7.1: 特有スペクトル探索アルゴリズム

7.2　声優認識アルゴリズム

本節では，実際にアニメ動画を視聴しているユーザに対して，アニメ動画内に出て来た音声の持ち主は誰であるか声優名を提供する根幹に関する認識手法の説明を行う．声優認識する為の提案システムの詳細を図 7.2 に示す．図 7.2 のようにアニメ動画が流れる時系列毎のその瞬間の周波数スペクトルを取得した $\boldsymbol{v_t}$ と，声優データベースに予め用意されている各声優 i の特有スペクトル $\boldsymbol{a_i}$ との，それら 2 つの情報を用いて算出された類似度を利用して，その動画内に流れた音声の持ち主の声優を判定する．ある時系列の時点の時に取得された $\boldsymbol{v_t}$ の数値を $v_{t,1}, v_{t,2}, \cdots, v_{t,B}$ と置き直す．本研究では，声優認識する為に，周波数スペクトル $\boldsymbol{v_t}$ と特有スペクトル $\boldsymbol{a_i}$ がどのくらい類似しているかを算出する計算式として，以下のコサイン類似度の数式を用いる．

$$\boldsymbol{v_t} = (v_{t,1}, \cdots, v_{t,B}), \quad \boldsymbol{a_i} = (a_{i,1}, \cdots, a_{i,B})$$

$$\mathrm{sim}(\boldsymbol{v_t}, \boldsymbol{a_i}) = \frac{\sum_{j=1}^{B} v_{t,j} \cdot a_{i,j}}{\sqrt{\sum_{j=1}^{B} v_{t,j}} \sqrt{\sum_{j=1}^{B} a_{i,j}}}$$

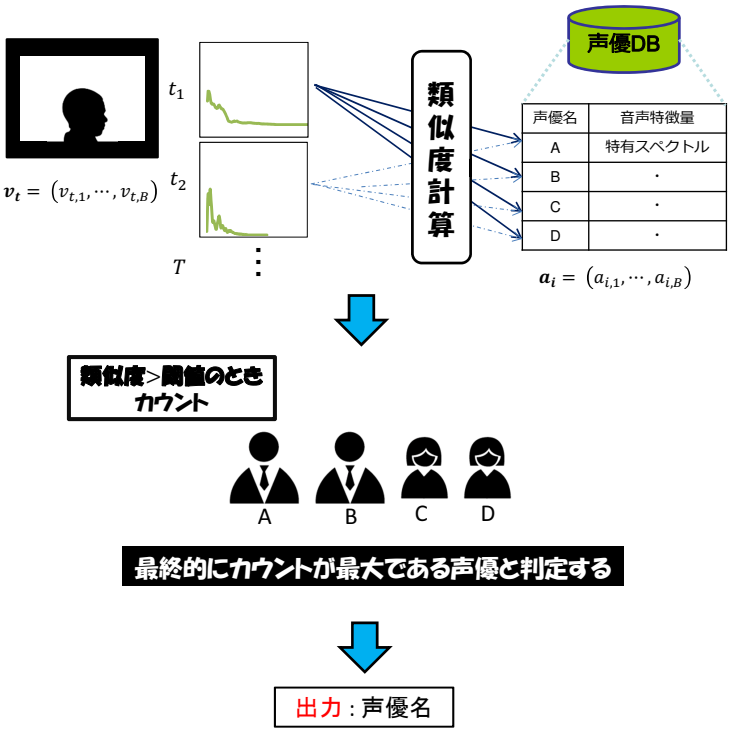


図 7.2: 声優認識システムイメージ

分析対象であるアニメ動画から流れた音声の再生が終了した後，声優を判定する処理に入る．再生中のアニメ動画から取得される時系列毎の周波数スペクトル $\boldsymbol{v_t}$（$t \in \{t_1, t_2, \cdots, T\}$）と各声優 i の特有スペクトル $\boldsymbol{a_i}$ との類似度に対する閾値のパラメータを予め定めておき，その定めた閾値を上回った時の $\boldsymbol{a_i}$ の声優 i に対してのカウント数を $+1$ 加算する．分析対象の音声データが最後まで流れた後，流れている間に加算された各声優 i の総計のカウント数同士で比較を行い，閾値を上回ったカウント数を一番多く獲得した声優 i をそのアニメ動画から流れた音声の持ち主であると判定する．しかし，声優データベースに登録する特有スペクトルを特定する時と同様に，閾値を上回ったカウント数が同数になることがある．その場合には，特有スペクトルを特定する時と同様に，閾値を上回った時だけの類似度の値をそれぞれ加算した平均で比較を行い，類似度の平均が最も大きかった $\boldsymbol{a_i}$ を持つ声優 i を，そのアニメ動画から流れた音声の声優であると認識する．

第 8 章

評価実験

本章では，まず初めに，キャラクターデータベースにキャラクターの性別を自動登録する為の，Web テキストを用いたキャラクター性別判定のアルゴリズムの精度評価を行う．次に，アニメ動画の声優認識の精度向上を狙って新たに導入した性別判定の精度に関する評価実験を行い，最後に声優認識の精度向上を狙って性別判定を導入した声優認識手法の精度に関する評価実験を行う．また，声優認識を行う前に，視聴中のアニメ動画のタイトルが特定されていることで，そのタイトルに基づいて Web 検索されたキャスト情報によって声優をキャスト陣のみに絞り込んでいる状態を想定している為，実験対象のアニメの話数に応じて声優データベースに登録される声優は変動する．

8.1　キャラクター性別判定の評価

5 章で提案したキャラクター性別判定のアルゴリズムの精度評価を行う為に，Wikipedia，ピクシブ百科事典，ニコニコ大百科の 3 つの Web サイトを対象に実験を行った．20 タイトルのアニメを 1 タイトルにつき男女 5 人ずつキャラクターを選出して，男性 100 人，女性 100 人の説明が載っている Web テキストをそれぞれ収集した．これらの Web テキストを対象に，5.1 節で用意した 5 種類の辞書を用いて，5.2 節で提案したキャラクターの性別判定アルゴリズムの精度評価を行った．図 8.1 から図 8.6 は，20 タイトルの男女キャラクター 100 人ずつを性別判定した結果，正解した数の割合を減衰率 δ の変動毎に求めた結果を表している．

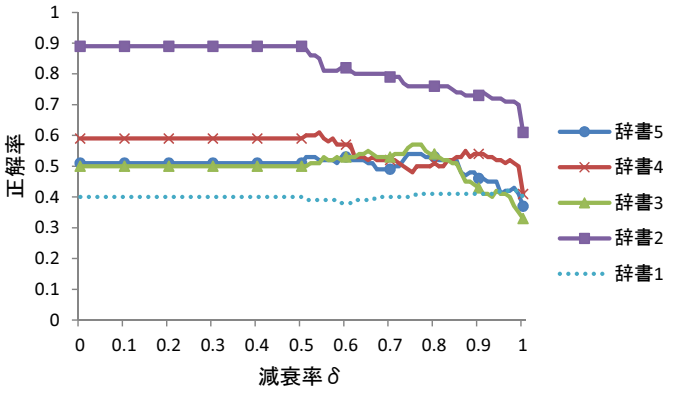


図 8.1: Wikipedia（女性キャラ）

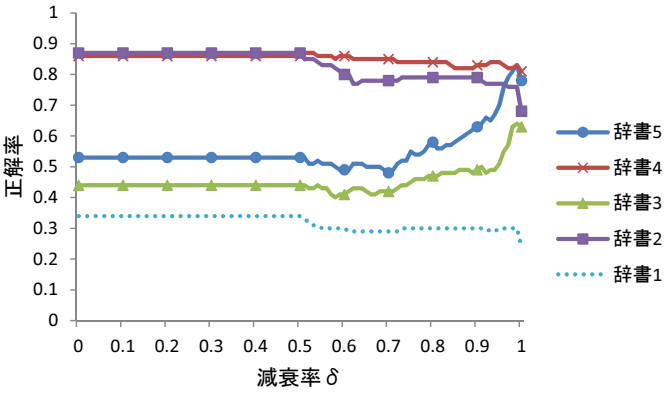


図 8.2: Wikipedia（男性キャラ）

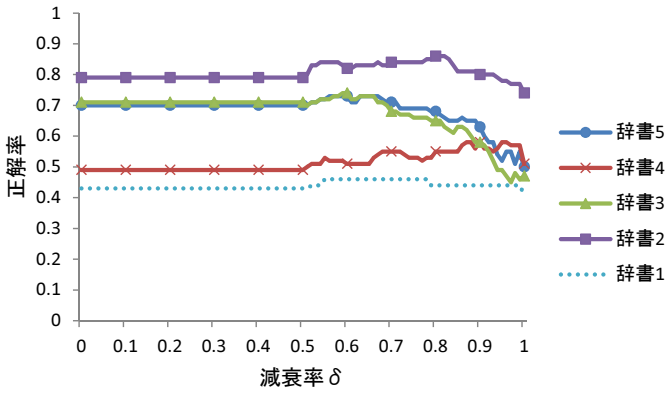


図 8.3: ピクシブ百科事典（女性キャラ）

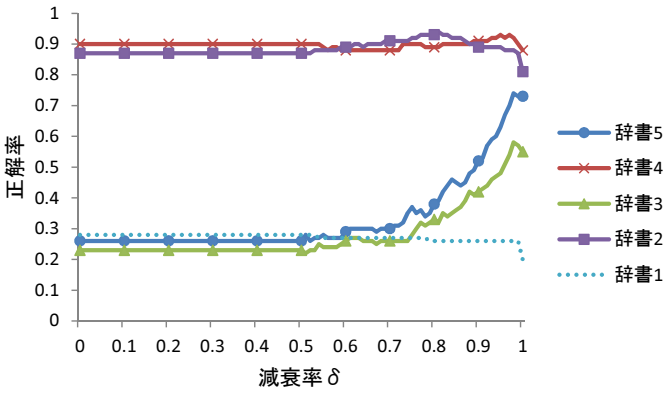


図 8.4: ピクシブ百科事典（男性キャラ）

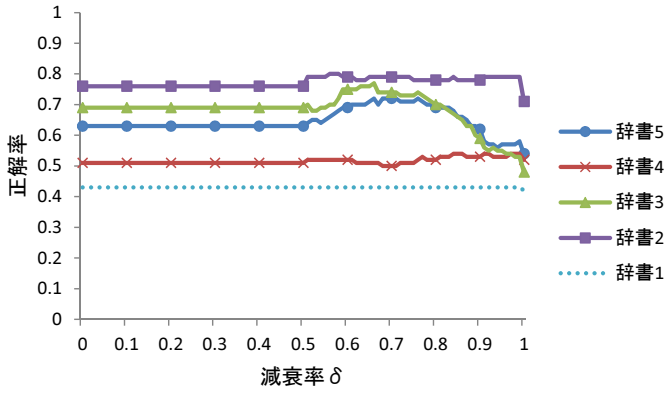


図 8.5: ニコニコ大百科（女性キャラ）

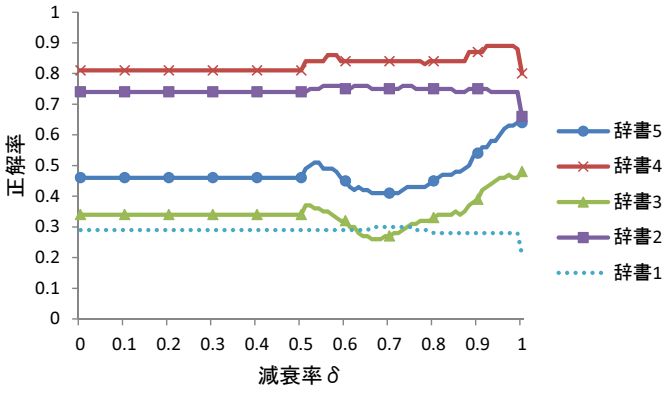


図 8.6: ニコニコ大百科（男性キャラ）

まず初めに，5 種類の辞書においての精度を比較すると，著者が主観で決めた辞書 2 の精度が良い印象を受ける．しかし，全ての Web サイトの男性キャラクターの正解率に注目してみると，日本語 WordNet を利用して作成した辞書 4 の精度が辞書 2 を上回っている．日本語シソーラス連想類語辞典で作成された辞書 3 に関しては，女性キャラクターに対して精度は良いが，男性キャラクターに対しては精度は悪い．辞書 3 が辞書 4 よりも全体的に精度が悪いのは，まず，ノイズが含まれている量が多いことが考えられる．一部例を挙げると，男性の辞書に「ルックス」「愛人」といった男女のどちらでも意味が取れる単語が見つかった．また，「お嬢様」「皇女」など，明らかに女性を意味する単語が含まれていることも分かった．日本語シソーラス連想類語辞典は，連想語の単語も取得出来ることから，ノイズが含まれるといった弊害が起きたと考えられる．また，辞書間の単語数の違いも原因の一つとして考える．図 8.3，8.5 を見ると，女性キャラクターを対象にした解析で，男性辞書よりも単語数が少ない女性辞書の場合，正解率が高い．図 8.2，8.4，8.6 でも，男性キャラクターを対象にした解析で，女性辞書よりも単語数が少ない男性辞書の場合，正解率は高い．これらの考察から，辞書の単語数を男女共に出来るだけ少なく均等にして，また，重要な単語を選別出来れば，精度向上を期待出来る．

最後に，減衰率 δ が及ぼしている影響を考察すると，どの図も減衰率が 0.5 の辺りで変動し始めていることが分かる．その後の変動は上昇下降とそれぞれ動きが異なるが，一番精度が良かった辞書だけを対象に，安定した精度を確保するならば 0.9 がベストであると考察する．

8.2　正規分布に基づく音声性別判定の評価

本研究で提案した音声の F0 を用いた正規分布に基づく性別判定の評価を行う．実験対象である音声は 3 種類のアニメタイトルを用いて，そのアニメ動画のある話数に限定を掛けて，そこから取得出来たキャラクターボイス，総計男性 85 個，女性 87 個をそれぞれ用意した．それらは，BGM や効果音といったノイズが出来るだけ含まないものを実験対象の音声として選定した．尚，用意したセリフ一つずつの長さは多様である．また，声優データベースには，実験対象音声の話数に出演するキャスト陣の基本周波数が含まれている．タイトル A のキャスト情報は 8 人，タイトル B のキャスト情報は 2 人，タイトル C のキャスト情報は 5 人であった．声優データベースも実験対象音声と同様に，ノイズを出来るだけ含まないものを選択した（あるいは著者がその部分を除去した）．声優データベースの音声は，声優が所属している事務所で公開しているサンプル音声や，ある声優に特化したセリフ集などから収集したものである．

表 8.1: 性別判定の評価（正解数/実験音声の数）

声優	男性 1	男性 2	女性 1	女性 2	女性 3
タイトル A	16/24	9/9	28/28	—	—
タイトル B	6/6	—	4/5	—	—
タイトル C	43/46	—	33/34	10/10	10/10

※タイトルに依ってキャスト情報が異なる為，声優はそれぞれ別人である

総計 ： 正解率（男性）0.870， 正解率（女性）0.970

表 8.1 の実験結果を考察すると，全体的に男性，女性と共に高い精度であると言える．しかし，タイトル A の男性 1 の精度が比較的悪い．その要因として考えられるのは，タイトル A の男性 1 は，声質が他の男性の声優と比べると，主観ではあるが比較的高い部類である．そこで，男性の正規分布の確率密度関数によって得られた確率値が，女性の正規分布の確率密度関数で得られた確率値よりも小さくなり不正解が多くなってしまったと考えられる．明確に男女の声質が高低に分かれているならば，声優データベースに登録されている F0 値のデータによって作成される正規分布は男女で重ならないことが予測されて，性別判定が上手くいくと考えられる．

8.3　声優認識の評価

本節では，声優認識の精度向上を狙って性別判定を導入した声優認識手法の精度に関する評価実験を行う．本評価実験では，Web からアニメ動画のキャスト情報を取得して，元々の声優データベースの声優の人数を各話毎に絞り込めた状態であると仮定する．特有スペクトルを特定する時の閾値を t_1 ($\in\{0.50,0.51,\cdots,1.00\}$) と表し，声優認識に使う閾値を t_2 ($\in\{0.50,0.51,\cdots,1.00\}$) と表す．また，周波数スペクトル同士の類似度計算式は特有スペクトルを特定する時，声優認識を行う時と共にコサイン類似度を用いる．音声信号処理には Python のライブラリの一つである scipy を用いた．評価実験を行う時に設定した内容を以下に示す．

- 音声信号処理
  - サンプリング周波数：16000Hz
  - 窓関数：ハミング窓
  - サンプル数：$2^N (N \in \{9,10,11\})$（3 種類）
  - バンド幅：0〜500Hz，0〜8000Hz（2 種類）
- 声優データベース
  - 声優のセリフの数，長さは多様なデータを収集
  - 公開しているサンプル音声やセリフ集から収集
  - BGM を出来るだけ含まないもの
  - 実験動画のアニメの各話に対応するキャスト情報の分だけを登録
    - 1 話：男性 1 人，女性 5 人
    - 2 話：男性 1 人，女性 4 人
    - 3 話：男性 2 人，女性 2 人
    - 4 話：男性 1 人，女性 4 人
    - 5 話：男性 2 人，女性 4 人
- 実験動画
  - BGM を出来るだけ含まないもの
  - あるアニメタイトルの 5 話分から音声を収集
    - 1 話：男性 26 音声，女性 26 音声
    - 2 話：男性 14 音声，女性 21 音声
    - 3 話：男性 17 音声，女性 19 音声
    - 4 話：男性 20 音声，女性 39 音声
    - 5 話：男性 36 音声，女性 39 音声
    - 総計：男性 113 音声，女性 144 音声

8.3.1　2 種類の閾値の評価

7 章で説明した声優認識の手法では，特有スペクトルを特定する時の閾値 t_1 と声優認識のアルゴリズムで用いる閾値 t_2 を設けた．本節の閾値の評価実験では，t_1，t_2 の最適な組み合わせのパラメータを，サンプル数の 3 パターンとバンド幅の 2 パターンの組み合わせにおいて，それぞれ求める．実験動画の総計数である男性 113 音声，女性 144 音声に対して，著者が 6 章で提案した，声域分割を行わない性別判定，声域分割を行う性別判定，性別判定を行わなかった場合，性別判定が 100% 成功すると仮定した場合の 4 パターンによる正解数をまとめた割合で正解率を算出した．表 8.2 にサンプル数とバンド幅の組み合わせ毎に，最も精度が良かった t_1，t_2 の組み合わせを示す．

まず初めに表 8.2 を参考に，サンプル数が同じ値でバンド幅同士を比較してみると，余り精度が変わらないように見える．次に，バンド幅が同じ条件で，サンプル数の変動を見てみると，こちらも同様に精度の変化が余り見られない．この結果から，それぞれのパラメータの組み合わせにおいて，適切な閾値を定めておくことにより精度の誤差を最小限に抑えることが出来ると言える．

また，t_1，t_2 の閾値の組み合わせについて考察を行うと，サンプル数とバンド幅のどの組み合わせにおいても，t_1 より，t_2 の値の方が大きい．この結果は，特有スペクトルを定める場合の閾値よりも，声優認識の閾値を厳しめに設定することにより精度が良くなることを示している．考察すると，特有スペクトル探索の場合は，声優自身の様々な音声データから抽出された周波数スペクトル同士，比較しているのに対して，声優認識の場合，未知の声優（または，未知のアニメ声）の分析対象の音声データから抽出された周波数スペクトルと各声優の特有スペクトルとの比較を行うので，閾値を厳しめに設定することで，より良く似ているスペクトルを見つけ出す必要があることが分かる．

表 8.2: 閾値の組み合わせに依る正解率の評価

		バンド幅	
		0〜500Hz	0〜8000Hz
サンプル数	$N = 9$	0.465 $(t_1 = 0.59, t_2 = 0.80)$	0.481 $(t_1 = 0.58, t_2 = 0.65)$
	$N = 10$	0.458 $(t_1 = 0.50, t_2 = 0.74)$	0.485 $(t_1 = 0.62, t_2 = 0.65)$
	$N = 11$	0.469 $(t_1 = 0.59, t_2 = 0.68)$	0.475 $(t_1 = 0.51, t_2 = 0.63)$

閾値 t_1: 7.1 節の特有スペクトル探索アルゴリズムで用いる
閾値 t_2: 7.2 節の声優認識アルゴリズムで用いる

8.3.2　性別判定による声優認識の評価

本節では本研究で提案した性別判定を導入することで声優認識の精度が向上すると考えて精度評価を行う．その為に，以下の 4 パターンの性別判定を組み込んだ声優認識システムの精度を比較する．

- 性別判定を行わなかった場合
- 声域分割を行わない場合の性別判定
- 声域分割を行う場合の性別判定
- もしも性別判定の精度が 100% の場合

また，特有スペクトル探索手法や，実験動画から取得される周波数スペクトルと声優データベースに登録されている特有スペクトルとの類似性の計算に影響を及ぼすと考えられるバンド幅の長さの違いや，サンプル数の違いに依る影響度にも注目して同時に比較実験を行う．サンプル数とバンド幅の組み合わせにおける閾値 t_1，t_2 に関しては，前節で求めた最適値にそれぞれ固定して実験を行う．図 8.7 から図 8.12 は実験動画対象の男女声優の音声の男性 113 音声，女性 144 音声ずつを声優認識した結果の正解した数の割合を性別判定手法の種類毎に求めた結果を表している．また，表 8.3 から表 8.7 は実験動画対象の男女声優の音声の性別判定を各話毎に行った実験結果を表している．その各話毎の実験動画対象の音声に性別判定を行った時の，各話毎のキャスト情報によって絞られている声優データベースに登録されている男女それぞれの F0 値のヒストグラムを図 8.13 から図 8.17 まで表している．図 8.18 に関しては，8.2 節の表 8.1 で評価したタイトル C に対応した，声優データベースに登録されている男女それぞれの F0 値のヒストグラムを表しており，上手く性別判定が行われている声優データベースの F0 値分布の例として，性別判定が上手く行われていない声優データベースの F0 値分布との比較を行うために，併せて載せている．ヒストグラムにおいて縦軸のデータ数は出現確率を表している．表 8.8 は，実験動画対象の音声に性別判定を行った時の，音声の高低に基づく判定クラスタにおける，それぞれの平均値 μ と分散 σ^2 を表している．6 章で述べた手法により，分散の値によって声域分割を行うか，行わないか各話毎に判定している．

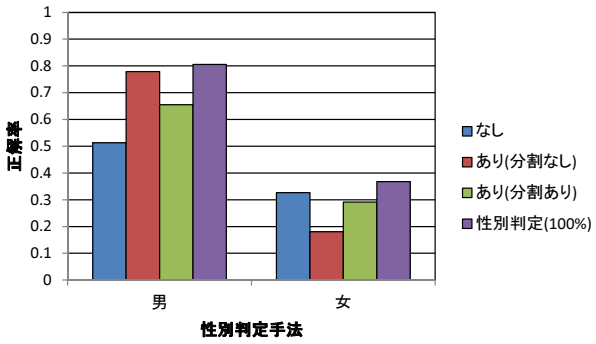


図 8.7: N=9, バンド幅 0～500Hz

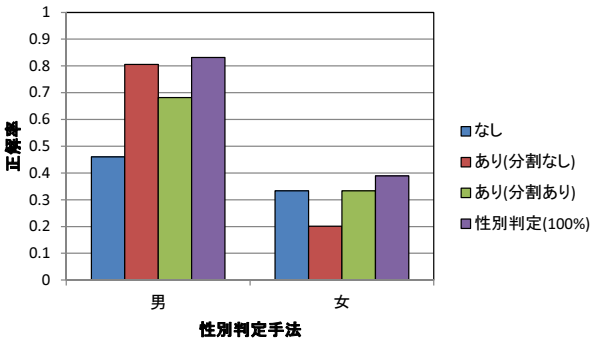


図 8.8: N=9, バンド幅 0～8000Hz

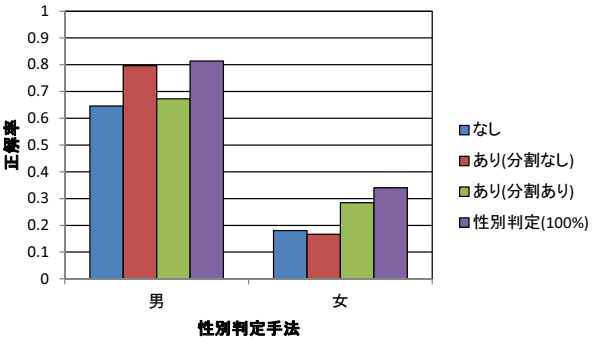


図 8.9: N=10, バンド幅 0～500Hz

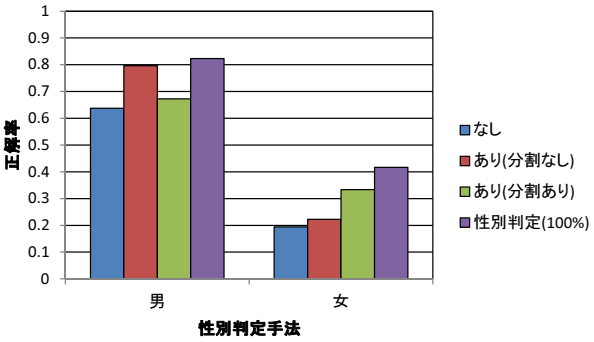


図 8.10: N=10, バンド幅 0～8000Hz

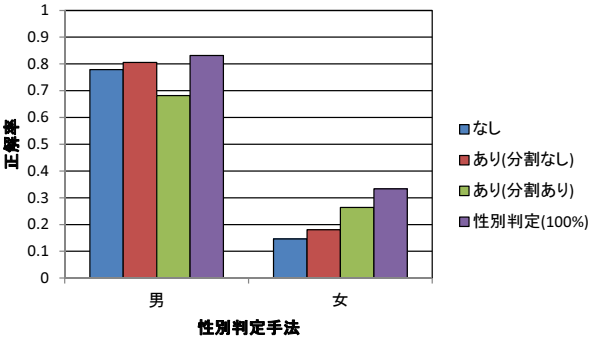


図 8.11: N=11, バンド幅 0～500Hz

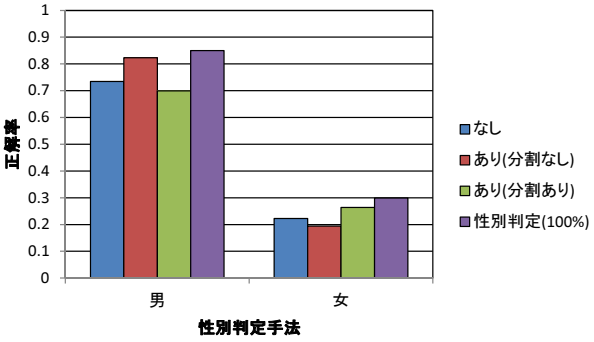


図 8.12: N=11, バンド幅 0～8000Hz

表 8.3: 1 話の性別判定の評価（正解数/実験音声の数）

声優	男性 1	女性 1	女性 2	女性 3
1 話（分割なし）	26/26	4/4	14/15	0/7
1 話（分割あり）	15/26	3/4	12/15	7/7

表 8.4: 2 話の性別判定の評価（正解数/実験音声の数）

声優	男性 1	女性 1	女性 2	女性 3
2 話（分割なし）	13/14	0/2	2/9	9/10
2 話（分割あり）	13/14	2/2	3/9	10/10

表 8.5: 3 話の性別判定の評価（正解数/実験音声の数）

声優	男性 1	男性 2	女性 1	女性 2
3 話（分割なし）	13/14	3/3	6/10	3/9
3 話（分割あり）	11/14	3/3	9/10	9/9

表 8.6: 4 話の性別判定の評価（正解数/実験音声の数）

声優	男性 1	女性 1	女性 2	女性 3	女性 4
4 話（分割なし）	20/20	5/5	0/1	22/25	4/8
4 話（分割あり）	19/20	5/5	0/1	21/25	8/8

表 8.7: 5 話の性別判定の評価（正解数/実験音声の数）

声優	男性 1	男性 2	女性 1	女性 2	女性 3
5 話（分割なし）	19/20	16/16	3/3	13/20	14/16
5 話（分割あり）	19/20	16/16	3/3	15/20	15/16

総計（分割なし）　：正解率（男性）0.973（—），正解率（女性）0.687（—）

総計（分割あり）　：正解率（男性）0.849（↓），正解率（女性）0.847（↑）

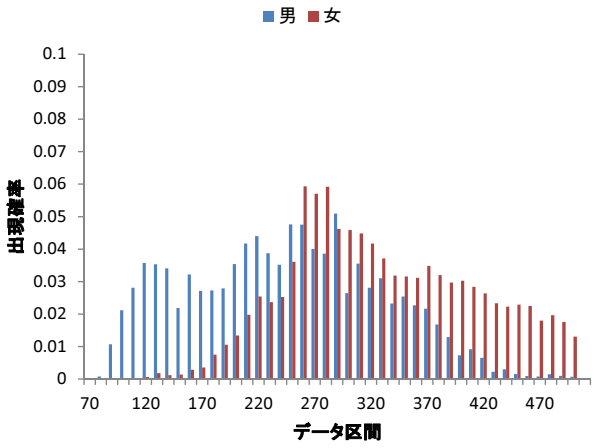


図 8.13: 声優 DB の F0 値出現確率（1 話）

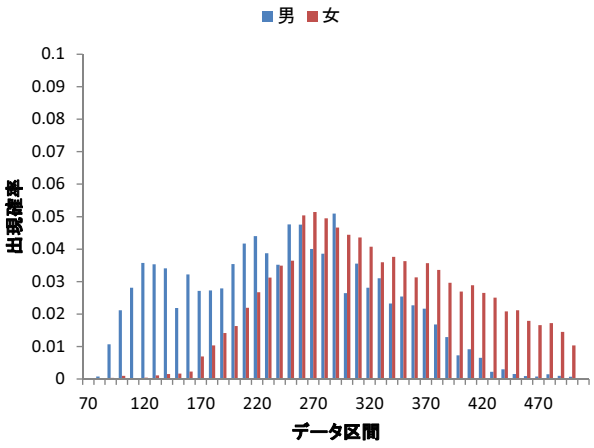


図 8.14: 声優 DB の F0 値出現確率（2 話）

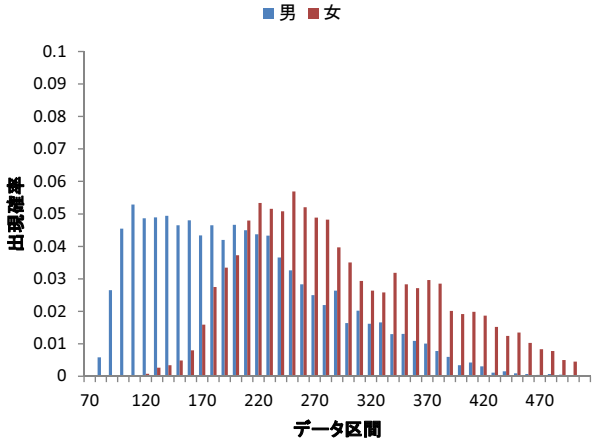


図 8.15: 声優 DB の F0 値出現確率（3 話）

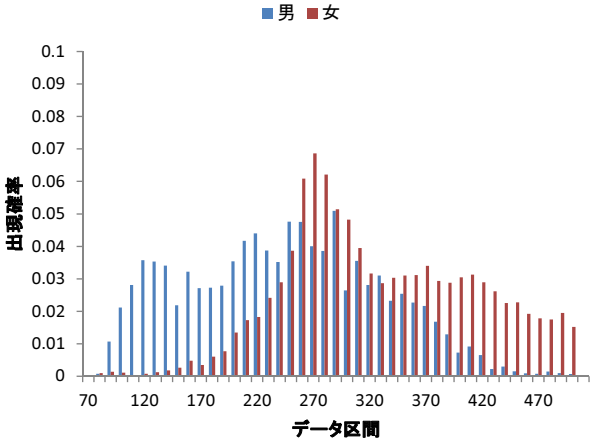


図 8.16: 声優 DB の F0 値出現確率（4 話）

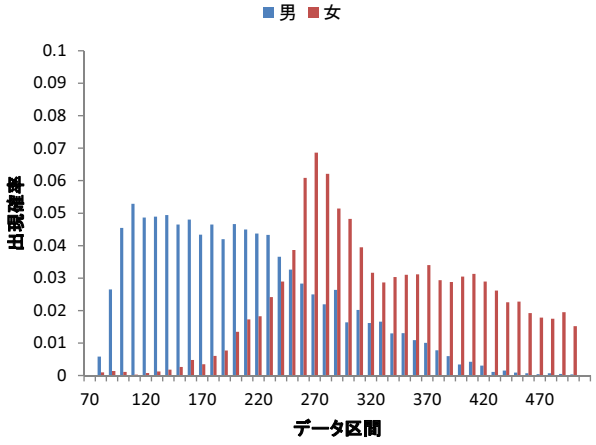


図 8.17: 声優 DB の F0 値出現確率（5 話）

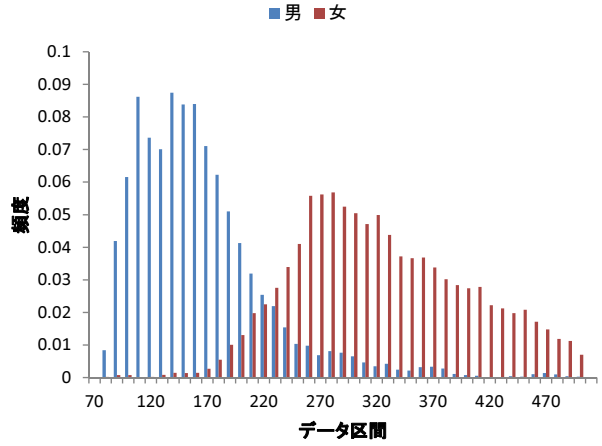


図 8.18: 声優 DB の F0 値出現確率（タイトル C）

表 8.8: 分割された声域の詳細

	男		男（声低）		男（声高）		分割判定
	平均	分散	平均	分散	平均	分散	
1 話	238.83	7212.77	159.55	3045.27	271.38	5284.31	○
2 話	238.83	7212.77	159.55	3045.27	271.38	5284.31	○
3 話	199.68	6390.82	153.99	2446.10	259.37	5254.31	○
4 話	238.83	7212.77	159.55	3045.27	271.38	5284.31	○
5 話	199.68	6390.82	153.99	2446.10	259.37	5254.31	○

	女		女（声低）		女（声高）		分割判定
	平均	分散	平均	分散	平均	分散	
1 話	325.13	6686.50	243.78	3767.64	341.38	5683.73	○
2 話	320.40	6733.99	268.30	4685.71	349.18	5537.86	○
3 話	286.66	6642.24	231.86	2810.92	320.80	5992.60	○
4 話	324.31	6941.53	264.31	5135.80	347.73	5692.44	○
5 話	324.31	6941.53	264.31	5135.80	347.73	5692.44	○

まず初めに，性別判定を組み込まない場合，性別判定（分割なし）を組み込んだ場合の声優認識の精度の比較を行う．図 8.7 から図 8.12 までの，性別判定を組み込まなかった場合と性別判定（分割なし）を組み込んだ場合との声優認識精度の比較をすると，性別判定を組み込むことで精度が向上するという仮説通りに，性別判定（分割なし）のシステムを組み込んだ声優認識の精度が全体的に上がったと見られる．男性声優に対する精度が，比較的良くなる傾向を示した実験結果になった．しかし，女性声優に対してはサンプル数とバンド幅の組み合わせに依って，精度が下がる傾向を示している結果も見られる．そこで，各話の性別判定の実験結果を表している，表 8.3 から表 8.7 までを見てみると，男性声優の性別判定とは対照に女性声優に対しての性別判定（分割なし）の手法による精度が良くない．その結果が声優認識に悪影響を及ぼして，女性声優の認識結果が悪くなったと考えられる．性別判定（分割なし）による手法が女性声優に対して，上手くいかなかった原因の分析として，図 8.13 から図 8.17，表 8.8 を見てみる．性別判定（分割なし）における女性声優の性別判定精度として，中でも良くなかった話数の図 8.13，図 8.14 に注目すると，男性声優の F0 値の出現確率と女性声優の F0 値の出現確率の分布が幅広く重なり合っている状態が見られる．加えて，表 8.8 の 1 話，2 話において，男性の分散 σ^2 が大きく，比較的精度が良いであろう 5 話と比べて男性と女性の平均 μ の差が狭い．この事象が，女性声優である音声の持ち主が，男性声優と誤判定された原因であると考察する．また，性別判定（分割なし）が比較的上手く行われている話数の，図 8.17 を見てみると，低周波数において男性声優の F0 値の出現確率が高いことが分かる．表 8.8 の 3 話，

5 話の平均と分散からも読み取れる．この現象が起きた原因として，キャスト情報によって絞り込みを掛けた声優データベースの男女比率の影響，また，絞り込まれた男性声優が持つ声域の高低範囲などの要因が考えられる．図 8.15，図 8.17 の話数のキャスト情報は男性声優が二人であるが，図 8.13，図 8.14，図 8.16 の話数のキャスト情報は男性声優が一人である．その一人の男性声優が持つ声域が広域であり，F0 値の出現確率が広がってしまった為に，女性声優に対して性別判定が上手くいかなかったと考えられる．

しかし，性別判定システムを声優認識を行う前に組み込むことで，男性声優の精度が格段に上昇した．また，女性声優に対しても，サンプル数とバンド幅の組み合わせにも依るが，わずかに声優認識精度が上がった．このことから性別判定を行わない声優認識のみの手法で，実験音声の持ち主の性別と正反対の性別の誰かと認識していたものが，性別判定を組み合わせることに依って，正しい音声の持ち主を認識出来た結果であると言える．性別判定手法の実験結果の表 8.3 から表 8.7 に示している通り，男性声優に対しては精度良く判定が行えている．このことから，性別判定手法によるシステムの精度が向上することが出来れば，声優認識の精度に大きく貢献出来ると言える．

また，その根拠を更に明確にする為に，図 8.7 から図 8.12 に示している，性別判定がもしも 100% 成功すると仮定した場合の声優認識結果に表れている．サンプル数とバンド幅の組み合わせに依って精度の変動はあるが，どの組み合わせにおいても精度の上昇が見られる．男性において，正解率の上昇値が一番高い場合，約 30%，女性においては，約 20% の精度の向上が見られる．

次に声域分割を行う性別判定を組み込んだ声優認識精度の考察を行う．図 8.7 から図 8.12 までの，声域分割を行う性別判定手法を組み込んだ声優認識結果を見てみると，女性声優に対して良い精度が出ている反面，男性声優に関しては，性別判定を行わない声優認識精度と比較すると，図 8.7 や図 8.8 のように，精度向上が見られたり，精度が大きく変動しなかったりするが，性別判定（分割なし）の声優認識精度と比較すると，精度が下がっている傾向が見られた．性別判定（分割なし）と比較する為に，表 8.3 から表 8.7 までを見ると，全体的に男性声優に対して精度が悪くなる傾向が見られる．しかし，女性声優に対しては精度の大きな向上が見られた．表 8.8 に注目すると，どの話数においても声域分割を行うことで，男女共に声域分割を行う前より，声域の高低共に分散 σ^2 がより小さくなっていることが分かる．分割が行われた結果，広がっている分布の分散が上手く狭まって，女性声優であるが男性声優と判定されていたミスが覆されることで，女性声優の認識精度の上昇が見られたと考えられる．しかし，男性声優に対して性別判定（分割なし）より精度が下がった理由として，その傾向が顕著である表 8.8 の 1 話に注目して考察すると，男（声高）と女（声低）の平均 μ の値が男性の方が大きく，女性の方が小さい値であり，また，分散 σ^2 の値が男性の方が大きく，女性の方が小さい値になったからである．その為，女（声低）の出現確率の分布が男（声高）の分布を含む形になった為に，分析対象の男性声優の音声から抽出された F0 値が約 260Hz 付近でも，女性の出現確率値が大きくなってしまったと考察する．

また，男性声優の声優認識精度が上がったもう一つの要因として，本実験で用いた実験動画のキャスト情報が全体的に，男性が一人，女性多数の場合が多いことが挙げられる．分析対象

の音声の持ち主が女性である場合に，性別判定で女性と判定され絞り込みを掛けても，女性の候補が多い為，得られる利点は小さく女性の精度は良くならなかった．しかし，分析対象の音声の持ち主が男性である場合に，性別判定で男性と判定することが出来れば，声優は一人に絞られる為，男性声優の精度に良くなる傾向が見られたと考えられる．

次にサンプル数とバンド幅の精度評価を行う．図 8.7 から図 8.12 を見ると，性別判定を行う場合の 2 つの手法は，性別判定の結果の重みが声優認識の精度に大きく傾き，サンプル数やバンド幅に大きく変動が見られない．しかし，性別判定を行わない場合の声優認識精度に注目すると，サンプル数やバンド幅が声優認識精度に影響を与えているのが分かる．男性声優の場合，サンプル数 2^N が大きいほど精度が良くなっていく傾向が見られるが，女性声優の場合は精度が下がる傾向が見られる．これは，サンプル数の変動により周波数分解能が上がることで，細かく Hz の間隔を解析することが出来ることに依って低周波数の値をより多く解析出来るようになる為，男性声優の精度が上がったと言える．逆に，粗く Hz の間隔を解析すると，女性声優の精度が上がると考えられる．また，バンド幅の違いに依る影響が小さいことが分かる．このことから，人の個人性を表す声質である重要な周波数の集まりは低周波数に集まっていると言える．低周波数の集まりから，個人性を表す特徴量を取得出来れば，声優認識精度が上がると考えられる．

最後に，8.2 節の性別判定（分割なし）の精度と本節の声域分割を行わない場合の性別判定の精度との比較を行う．8.2 節の実験結果である表 8.1 と本節の表 8.3 から表 8.7 までの実験結果を比較すると，女性声優の精度が約 0.30 ほど精度に差があることが分かる．また，8.2 節の性別判定（分割なし）の男女共に高精度である．そこで男女共に性別判定を精度良くする為に，8.2 節の実験結果である表 8.1 の，男女共に比較的精度が良かったタイトル C を代表に分析を行う．タイトル C の声優 DB の F0 値出現確率を図 8.18 に示している．この図 8.18 から，男女の F0 値の分布が明確に双峰型であり，男女の分布が被っている面積が小さいことが分かる．このように，互いの男女の平均値に大きな差があって，互いの分散を小さい分布の場合では，性別判定が精度良く行えると分析する．

第 9 章

まとめと今後の課題

ユーザがアニメ動画を視聴している時，「この声の人は誰だったかな」と不満が出て来る場合がある．ユーザが音声の持ち主が誰であるか調べようとするならば，インターネットで調べたり，エンディングのスタッフロールまで飛ばそうとするであろう．しかし，調べる手間が掛かったり，必ずしもネット状況が繋がっている保証があるとは限らない．また，エンディングまで飛ばそうとするならば，ロード時間が掛かるであろう．そこで本研究では，リアルタイムで音声の解析を行い，音声の持ち主の名前をユーザに提供出来るシステムを目指した．音声の持ち主を特定することが出来れば，名前を表示するだけでなく，その人物に関する様々な情報をユーザに提供することが出来る．例えば，音声の持ち主に関連した出演アニメや出演ゲーム，イベント情報などの娯楽，また，関連商品の推薦といった情報である．

本研究では，音声データから音声の持ち主を認識する為に，様々な人物の音声データから声質を表す特徴量の抽出を行い，その特徴量を予め声優データベースに登録しておくことで，分析したい音声と登録した特徴量との類似性を求めて，声優を特定する手法を提案した．声優データベースに登録されている各声優が持つ特徴量を「特有スペクトル」と定義した．分析対象である音声から抽出した周波数スペクトルの時系列パターンと，特有スペクトルとの類似度を求めて声優認識を行った．また，分析対象である音声の持ち主を特定する為に，音声から得られる特徴量だけでなく，Web 上から得られるキャスト情報を活用して，声優データベースの中から音声の持ち主の候補となる声優を絞り込む工程を行った．加えて，分析対象である音声に対して声優認識を行う前に，本研究で提案した性別判定システムを組み込むことで，声優認識の精度の向上を狙った．キャスト情報により既に音声の持ち主の候補が絞り込まれていた声優データベースに対して，性別判定処理を行うことで更に候補を絞り込む為である．本研究で提案した性別判定手法には，声の要素の一つである基本周波数（F0）の値を活用した．声優データベースに登録されている男女毎の F0 値から正規分布の作成を行い，分析対象の音声から抽出された F0 値のベクトル情報を正規分布の確率密度関数に代入して男女判定を行った．性別判定の実験を行った結果，精度は必ずしも良くはならなかった．そこで，性別判定手法の改善として男女毎の高低に基づいた声域分割を行った．声優の声質は広域であるという仮説から，男女毎の高低に音声データを上手く分類することで，より声質を細かく分析出来ると考えた．声域分割を行った場合の性別判定手法の実験を行った結果，男性，女性と共に精度良く判

定出来た．本研究で提案したシステム全体の評価実験を行った結果，声優認識において全体的には良い精度を出すことは出来なかった．

今後の課題として，本研究では声優データベースの中に声優一人につき「特有スペクトル」を一つ登録して声優認識を行うアルゴリズムを提案したが，新たな手法として，声優一人につき複数のセリフの特徴量を登録して声優認識を行うアルゴリズムの検討や，分析対象である音声の，時系列毎に取得される周波数スペクトルから代表的なスペクトルを一つ見つけて，声優データベースに登録されている特有スペクトルとの類似性を求めて声優を特定する手法の検討などが挙げられる．また，本研究で提案した声域分割する場合の性別判定手法における分類手法と，SVM や GMM といった既存の分類手法との精度比較を行う必要がある．また，本研究では声の周波数スペクトルに注目して人物認識を行ったが，フォルマントや抑揚，時間などといった声質から人物を特定することが出来るか検証する．

最後に，本研究における成果をまとめる．まず初めに，社会的な貢献として，本研究ではアニメ動画の音声に特化した人物認識を行ったが，研究の題材が生体認証の一部である為，本研究で提案した声優認識の精度を最大限まで高めることが出来れば，「オレオレ詐欺」といった犯罪の防止，「電話口の顧客対応の個人識別」の認証支援など，様々な場面で貢献出来るであろう．また，エンターテイメントとして，ある人物の声にどのくらい似ているかといった「声真似」の練習も出来ると言える．次に技術的な貢献として，声優に限定した特徴的なアニメ声における声優認識システム開発を行った．声優認識を行う為に用いる声の特徴量として，声優の音声データから取得された周波数スペクトルから，「特有スペクトル」という特徴量を抽出する手法を提案した．本研究で提案した「特有スペクトル」の特徴量と，著者の従来の研究で用いた特徴量との技術的な違いとして，従来では無作為に選出した音声データから，音声波形の振幅データを特徴量としていたが，本研究では，無作為に選出した音声データの数々から，それぞれの時系列毎に取得される周波数スペクトルを対象に，本研究で提案した探索アルゴリズムによって，頻出した（類似性がある）周波数スペクトルを特徴量として定めたことが技術的な違いである．また，本研究で提案した 2 種類の性別判定処理を声優認識処理を行う前に組み込んだシステム全体の評価実験を行った．その結果，性別判定を精度良く行うことが出来た為，性別判定を行わない声優認識の精度より，性別判定を行った声優認識の精度を上げることが出来た．また，性別判定の精度においても，新しく声域分割を行う性別判定手法を提案することで，性別判定精度がより良くなることを示した．結果から，まだ，声優認識システムの精度として，著者が満足出来る精度である，男女共，8 割の認識精度には至らなかったが，声優認識を行う前に性別判定処理を組み込むことで，声優認識の精度が良くなる傾向があることを示すことが出来た．

謝辞

本研究に際して，様々なご指導を頂きました服部峻助教を初めとして，服部研究室の皆様に感謝を致します．そして，小林洋介助教の信号処理のご指導に感謝を致します．また，実験に使った Python のライブラリ製作者の皆様にも感謝致します．そして，本研究で用いたフリーのオープンソースソフトウェアを提供している皆様に感謝致します．

参考文献

[1] 榮田 基希, 服部 峻, “アニメ動画の音声とキャスト情報を用いた声優認識,” 電子情報通信学会 情報ネットワーク研究会, 信学技報, Vol.115, No.405, pp.7–12 (2016).

[2] 榮田 基希, 服部 峻, “アニメ動画における音声の周波数スペクトルを用いた声優認識,” 電子情報通信学会 情報ネットワーク研究会, 信学技報, Vol.116, No.304, pp.25–30 (2016).

[3] 榮田 基希, 服部 峻, “アニメ動画の声優認識のためのコンテキストを意識した性別判定,” 電子情報通信学会 人工知能と知識処理研究会, 信学技報, Vol.117, No.326, pp.55–60 (2017).

[4] 榮田 基希, 服部 峻, “アニメ動画における性別判定を用いた声優認識のための音声の高低に基づく判定クラスタの細分化,” 電子情報通信学会 情報ネットワーク研究会, 信学技報, Vol.117, No.397, pp.85–90 (2018).

[5] 杉江 嘉昭, 小林 哲則, “Dempster–Shafer 理論を用いた音声・画像情報の統合による個人認識システム,” 電子情報通信学会, パターン認識・メディア理解研究会, 信学技報, Vol.101, No.423, pp.63–68 (2001).

[6] 徳田 恵一, “音声情報処理技術の最先端：1. 隠れマルコフモデルによる音声認識と音声合成,” 情報処理, Vol.45, No.10, pp.1005–1011 (2004).

[7] 森勢 将雅, “2 群 9 編 2 章 2-2 基本周波数推定（歌声に関する視点から）,” 電子情報通信学会「知識ベース」, pp.6–10 (2012).

[8] 山口 順一, “人の認識, 展望,” 精密工学会誌, Vol.71, No.2, pp.159–162 (2005).

[9] 櫻庭 京子, 今泉 敏, 峯松 信明, 田中 二郎, 堀川 直央, “女性と判定される声の特徴 –性同一性障害者の話声位–,” 音声言語医学, Vol.50, No.1, pp.14–20 (2009).

[10] 櫻庭 京子, 丸山 数孝, 峯松 信明, 広瀬 啓吉, 田中 二郎, 今泉 敏, 山内 俊雄, “話者認識技術を用いた性同一性症者（MtF）の音声に対する男声度・女声度の自動推定とその臨床応用,” 電子情報通信学会, 音声研究会 (SP) , 信学技報, Vol.105, No.686, pp.29–34 (2006).

[11] 大野 涼平, 森勢 将雅, 北原 鉄郎, “音声における「かわいらしさ」の知覚と聴取時間の関係性の検討,” 情報処理学会, 研究報告音楽情報科学（MUS）, Vol.111, No.50, pp.1–5 (2016).

[12] まうまう☆, `https://www.mau2.com/` (2017).

[13] Wikipedia, `https://www.wikipedia.org/` (2017).

[14] 早川 昭二, 板倉 文忠, “音声の高域に含まれる個人性情報を用いた話者認識,” 日本音響学会誌 51 巻 11 号, pp.861–868 (1995).

[15] 横山 雅夫, “音声に含まれる個人性情報,” 福島大学行政社会学会, 行政社会論集, 第 4 巻第 3 号, pp.96–113 (1992).

[16] Praat, http://www.fon.hum.uva.nl/praat/ (2017).

[17] ピクシブ百科事典, https://dic.pixiv.net/ (2017).

[18] ニコニコ大百科, http://dic.nicovideo.jp/ (2017).

[19] 日本語シソーラス連想類語辞典, http://renso-ruigo.com/ (2017).

[20] Francis Bond, Timothy Baldwin, Richard Fothergill, Kiyotaka Uchimoto, “Japanese SemCor: A Sense-tagged Corpus of Japanese,” Proceedings of the 6th International Conference of the Global WordNet Association (GWC’12), pp.56–63 (2012).

[21] CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer, http://taku910.github.io/cabocha/ (2017).