修士論文

歌唱により嚥下・構音機能を評価する手法の実装と評価

37-206502 平井 雄太

指導教員：矢谷 浩司 准教授

東京大学大学院
工学系研究科 電気系工学専攻

令和 3 年 1 月 27 日 提出

Acknowledgements

本研究を進めるにあたり，厚いご指導をいただきました指導教員の矢谷浩司准教授に感謝いたします．研究テーマを模索する中でいただいた的確なアドバイスのおかげで，ここまで研究を成し遂げることができました．また，日々の 1on1 ミーティングにおいてその時点で取り組むべき事柄を適切に指導していただき，自分の研究のリズムを保つことができました．修士課程における 2 年間，ご指導いただき誠にありがとうございました．

工学系研究科の小野寺宏教授にも大変お世話になりました．先生には，本研究を始める前に別のテーマで試行錯誤していた時から検査機器を貸していただいた他，医学的な観点から有益なアドバイス等のおかげで研究をスムーズに進めることができ，心から感謝しております．

東京大学医学部附属病院の長島優助教，代田悠一郎講師からは，医学的なアドバイスから倫理審査書類の作成まで大変お世話になりました．また，東京医科歯科大学の戸原玄教授にもさまざまな医学的アドバイスを頂きました．同大学の柳田陵介先生，山田大志先生，米田早織先生には，往診に来られる患者さんのデータをご収集いただき，大変感謝しております．特に柳田先生には，倫理審査書類の作成いただき，大変お手数おかけしました．皆様に厚くお礼を申し上げます．

所属する矢谷研究室のメンバーにも大変お世話になりました．同期である佐野翔子さん，高島諒君，林裕嵩君には研究の方向性に関する議論はもちろん，プライベートに関しても様々な場面で協力してもらいました．互いに切磋琢磨することのできる良き仲間だったと感じています．また，私と同じく口腔機能をテーマに研究を進めていた耿世嫻さんには，研究内容をディスカッションしたり共同でデータ収集を行うなど，様々な面で力を貸していただきました．Arissa Janejera Sato さんや Zefan Sramek さんには英語で書くべき書類の添削，スライド内容のアドバイスなど非常に幅広い場面で助けていただきました．様々な企画や呼びかけによって研究室の雰囲気を良いものにしていただいたことも感謝しています．下島銀士君，桝田拓磨君，吉川諒君をはじめとする後輩のみなさんにも発表後のフィードバックや論文の添削など様々な場面でお世話になりました．特に下島君にはデータ収集を手伝っていただいたおかげで，スムーズに実験を終えることができました．研究室の皆様に感謝の意を表します．

秘書の元岡みさ子さんには実験などの各種手続きの際に常に迅速な対応をしていただき，大変お世話になりました．また，私の不手際で面倒になったしまった支払い手続き等にも対処していただき，感謝の言葉もありません．

最後に大学生活を金銭的にも精神的にも支えてくださった家族に感謝して謝辞の結びとしたいと思います．本当にありがとうございました．

Abstract

A recent study has shown that a minor decline in oral function increases future health risks. Therefore it is crucial to evaluate oral function regularly, but existing methods are not suitable for patients to repeat these methods voluntarily and independently. To solve this problem, we designed a karaoke-like system that scores articulation and swallowing capabilities by singing. Toward the implementation, we collected data on oral function and singing from people including the elderly and patients with swallowing disabilities. We extracted acoustic and image features from the obtained singing data, and classified articulation and swallowing disabilities using multiple stepwise logistic regression, resulting in a maximum accuracy of 97%. In addition, we adopted a recent machine learning method to create highly interpretable decision trees. Finally, we present the limitations of this work and conclude with our future plans.

Abstract

近年，オーラルフレイルと呼ばれる高齢者の口腔機能の軽微な低下が将来の健康リスクを増加させることが示され，口腔機能の定期的な評価の重要性はますます高まっている．しかしながら既存の口腔機能の評価手法は，特別な測定機器の必要性や測定自体の単調さ等の理由から，ユーザーが自発的に繰り返し取り組み難いという課題がある．そこでカラオケに含まれるゲーム性が測定に対する心理的負担を軽減することに着目し，モバイル端末による歌唱を通じた嚥下・構音機能を評価手法を提案する．提案手法の実現に向けて，我々は非高齢者 110 名（65 歳未満）と高齢者 76 名（65 歳以上）の実験参加者から歌唱と構音・嚥下機能のデータを収集した．それらに加えて，嚥下機能に障害を持ち，通院している者からも同様にデータの収集を行った．得られた歌唱データから音響・画像特徴量を抽出し，変数増減法によるロジスティック回帰を用いた構音・嚥下障害の分類を行った結果，最大で 97% の正解率を実現した．また，機械学習手法を用いた決定木による分類を行うことで，解釈性の高いモデルを作成した．最後に，最終的なアプリケーションの開発に向けた今後の課題について検討した．

Table of contents

List of figures

List of tables

Chapter 1

はじめに

1.1 背景

近年の医療の発展に伴い，先進国を中心に平均寿命は年々延伸し，我が国では現在，男性で 81 歳，女性で 87 歳と世界トップレベルを誇る [41]．一方で，自立して健康的に過ごすことのできる期間を表す健康寿命は平均寿命を 10 年程度下回る [45]．高齢化社会における健康寿命の低下は，社会保障費の増大をはじめとする様々な課題の原因となることから，健康寿命の延伸，すなわち高齢期であっても可能な限り長く自立した生活を送ることが期待されている．

フレイルとは健康状態と要介護状態の中間の状態にあることを示す概念で，適切な介入によって再び健常な状態に戻ることのできる可逆性がその大きな特徴である．しかしながらフレイルの進行を阻止できない場合，社会的・心理的要素を含めた様々な要因が絡み合い，負の連鎖を起こしながら自立度が低下していき，やがて要介護状態に至ることが知られている [38]．

近年，口腔機能の軽微な低下に伴う食習慣の悪化がその後の要介護状態や筋量低下などのリスクを増加させ，死亡率に至っては健康な場合と比較して約 2 倍にも達することが示された [32]．そこで口腔機能の維持の重要性を啓発するため，「オーラルフレイル」と呼ばれる新概念が提唱され，注目を集めている [43]．オーラルフレイルは，健康状態とフレイルの中間にある状態のことである．フレイルと同様に，オーラルフレイルも適切な介入を行えば再び健康な状態に戻ることができるため，口腔機能の低下を早期に発見し，リハビリテーションを行なっていくことは極めて重要である．

口腔機能を評価する既存手法として，/pa/, /ta/, /ka/，/ra/をそれぞれ 5 秒間に発音できた回数で口腔機能を評価するオーラルディアドコキネシス [46] や，何回連続で唾液を嚥

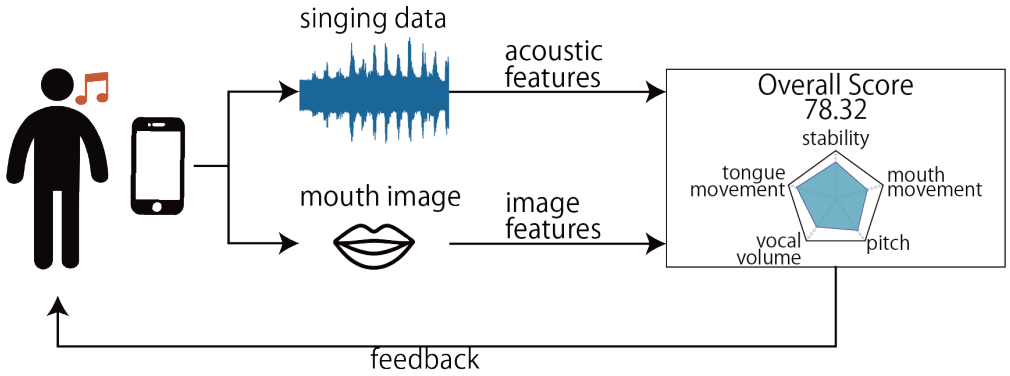

図 1.1: 提案するシステムのコンセプト図．ユーザーはスマートフォンから流れる音源に従って歌唱を行い，その様子の録画を行う．歌唱が終了すると，録画の結果から音響・画像特徴量が抽出され，採点結果を表示する．

下できたかによって口腔機能を評価する反復唾液嚥下テスト [42] が挙げられる．こうした既存の口腔機能の評価手法は手軽である一方，医療従事者による観察を必要とする点や，手法自体が単調であるため，自発的・継続的に取り組むには心理的な負担が大きい点が課題として挙げられる．また，舌圧の測定等の評価手法は体内に測定器を挿入するなど侵襲的であるため，被測定者に身体的な負担が生じることがある．

そこで我々は上記の課題を解決するため，モバイル端末を用いた，歌唱による口腔機能の定量的評価手法を提案する．提案手法のコンセプトを図1.1に示す．これはいわば，カラオケの歌唱力採点機能に口腔機能の採点機能を加えたシステムである．利用者はスマートフォンなどのモバイル端末から流れる伴奏に合わせて歌唱を行い，その様子をモバイル端末に搭載されたマイクとカメラによって記録する．曲が終了すると，我々の提案するアルゴリズムに従って，口腔機能の点数が表示される．提案手法は，医療従事者の助けを必要とせずに，そして非侵襲的に口腔機能を評価することができるだけでなく，歌唱とその採点というゲーム性を含めることによって，利用者の自発的・継続的な活用が期待できる．

提案手法の実現に向けて，本稿では我々が収集した非高齢者及び高齢者の歌唱時の音響・画像分析を行い，嚥下・構音機能との関連性を調査する．我々が収集したデータには，歌唱時の音声・動画に加えて，アンケート調査による嚥下機能の評価手法である EAT-10 [1]と定量的な構音機能の評価手法であるオーラルディアドコキネシス試験時の音声・動画が含まれている．

収集した歌唱データから音響・画像特徴量を計算したのち，変数増減法によるロジスティック回帰分析によって嚥下・構音障害の分類に有用な特徴量の選択を行った．そして，

[1]https://www.nestlenutrition-institute.org/resources/nutrition-tools/details/swallowing-assessment-tool

選択された特徴量を用いたロジスティック回帰による嚥下・構音障害の分類性能について評価を行い，比較的高い分類性能を確認した．一方で，再現率が低いなどの課題も見られた．そこで，それまでの分析方法の改善策を検討し，さらに嚥下機能に問題を抱え通院している患者のデータを追加して，改めて変数増減法によるロジスティック回帰分析を行った．

さらに，単純なロジスティック回帰による分析だけでなく，より複雑なモデルに対応した機械学習手法として，決定木を活用した LightGBM を用い，さらなる精度の向上や分析の可視化を試みた．その結果，高い分類性能を実現したほか，各特徴量の決定木による可視化を実現することができた．

1.2 貢献

本研究の最終的な目的は，歌唱による口腔機能の評価手法を確立することである．本研究の貢献は以下の通りである．

- 歌唱中による口腔機能の評価手法の提案
- 歌唱中の音声・画像から構音・嚥下障害を分類する手法の提案
- 歌唱中の音声・画像から構音・嚥下障害を分類する有効性の確認

Chapter 2

関連研究

本章では，本研究と関連する既存研究について説明し，次章以降の内容理解に役立つ概念等を説明する．また，それらの研究と比較した本研究の立ち位置を明確にする．

2.1 オーラルフレイル

フレイルとは特に高齢期において身体機能が低下した状態のことで，将来の転倒，障害，要介護，死亡などのリスクと大きく関連する一方で，健康状態と要介護状態の間の中間的な状態にあり，適切な介入により心身の機能を回復させることができるという特徴を持つ [10]．

近年，咀嚼能力の低下や天然歯の減少などの軽微な口腔状態の低下が，将来のフレイルや要介護状態，死亡率などのリスクの増加と関連していることが示された [32]．そこで，歯科口腔系の軽微な機能低下を医療従事者や一般国民が見逃さないようにするため，「オーラルフレイル」という概念が提案された [47]．オーラルフレイルもフレイルと同様に，適切な措置を取ることで健康な状態に戻ることができるため，早期にこれを発見して適切な措置を取り，将来の健康リスクを減らすことが社会的課題となる．

オーラルフレイルの概念はまだ生まれたばかりであり，今後のさらなる研究が期待されているが，既に日本歯科医師会や東京大学高齢社会総合研究機構が中心となってこれを啓発を進めている．特に日本歯科医師会は，80 歳における天然歯の数を 20 本以上維持することを目指した「8020 運動」が国民運動として広まった結果大きな成果を挙げたことを例に挙げ，オーラルフレイルも同様に国民全体に普及させていくことを目指している [1]．このような状況を踏まえ，今後ますますオーラルフレイルの重要性は高まっていくと予想される．

[1] https://www.jda.or.jp/enlightenment/oral/about.html

本研究でも，オーラルフレイルの概念の今後の浸透に伴い，口腔機能の評価の重要性がより一層高まる可能性に注目している．明確な病的状態ではないオーラルフレイルは，その状態にある高齢者にも見過ごされてしまう危険性が高い．このように自身のオーラルフレイルを見過ごしている場合には，自ら病院に赴いて医師から助言を受けることも難しいと考えられる．そのため，医者にかかることなく単独で，定期的・自発的に口腔機能のチェックを行うことのできるアプリケーションが必要だと我々は考えている．そこで本研究では，モバイル端末を用いた評価手法によって，直接の医師の診断を介さずにオーラルフレイルの可能性に気づけることを目指す．また，定期的・自発的な利用を促進するため，カラオケのようなゲーム性を取り入れることで心理的な負担の軽減を狙っている．

2.2 口腔機能の臨床的評価手法

口腔機能の定量的な評価手法はこれまで数多く提案されており，評価項目として口腔衛生状態，舌圧，吻合力，舌口唇運動機能，咀嚼・嚥下機能等が一般的である [43]．それらの中でも本研究との関連が深い舌口唇運動機能と嚥下機能の評価手法について本節で説明する．

舌口唇運動機能の評価手法として広く用いられているオーラルディアドコキネシスは，/pa/，/ta/，/ka/，/ra/を一定時間内に繰り返し発音できた回数を指標とする [46]．発音回数の基準値は各医学会，医師会などが独自に定めていることが多いが，5章以降では，先行研究を参考に 5 秒間で 20 回以下の発音を構音障害であると定めた [44]．発音回数は医師が手動で計測するほか，計算機を用いて自動で回数を測定する方法も提案されている [40, 26]．

嚥下機能の評価手法の 1 つとして知られているのが反復唾液嚥下テスト (the Repetitive Saliva Swallowing Test，RSST) である [42]．この手法では，30 秒間の唾液の嚥下回数によって嚥下機能を評価する．また，嚥下機能のスクリーニングツールとして用いられているのが EAT-10 である [7]．EAT-10 は，例えば「飲み込みの問題が原因で、体重が減少した」などの計 10 個の質問に対して 0 から 4 点の 5 段階で主観的評価を記入する質問用紙であり，合計点数が 3 点以上の場合には結果を専門医に相談することが推奨される．日本を含む世界各国で嚥下機能を評価する手法として有効であることが示されている [25, 8]．

本研究では，口腔機能の評価データとしてオーラルディアドコキネシス試験 (/pa/，/ta/，/ka/，/ra/の 5 秒間の発音回数) と主観的な嚥下機能のアンケート調査である EAT-10 を採用した．なぜなら，これらの手法は患者の体への接触や機器の挿入がないため評価手法

が簡便であり，病的な状態にある患者を対象としたデータ収集において安全性が高いと考えたからである．

2.3 口腔機能の訓練を目指したアプリケーション

口腔機能低下予防に向けた取り組みとして「口腔体操」が各自治体で採用され，日本医師会によるオーラルフレイル対策にも取り上げられている [2]．

これまで様々な口腔機能向上トレーニングが提案されている一方で，高齢者自身が自発的・継続的に取り組むには課題が多い．そこでゲームの要素を組み入れて自発的・継続的なトレーニングの促進を目指したのが Squach である [4]．Squach は口及び舌の動きを使って操作するビデオゲームで，介護施設入居者を対象とした実験では，オーラルディアドコキネシスなどいくつかの口腔機能の評価において改善が報告されている．

その他に，舌をカメラでトラッキングするゲームを開発し，構音・嚥下障害を持つ患者と健康な研究参加者との間で舌の動きの違いを調べた研究 [18] や，舌に装着した器具を用いて操作を行うゲームを開発し，リハビリテーションに活かそうとする研究 [13] も存在する．

これらの計算機を用いた手法は主に，口腔機能のトレーニングに対するゲーミフィケーションを提案している．一方で，本研究では，まだあまり行われていない口腔機能の評価のゲーミフィケーションに焦点を当てている．

2.4 構音障害の音響分析

構音障害とは「構音障害とは、言葉を正常にはっきり発音する能力が失われる障害」[3]である．構音機能の評価には，患者の発声の分析が用いられることが多い．例えば，セラピストが患者の発声を評価する際の基準として GRBAS 尺度などが知られ，一定の信頼性が認められている [9]．また，標準ディサースリア検査 (Assessment of Motor Speech for Dysarthria，AMSD) を用いた構音障害の検査も臨床の現場では行われている [46]．AMSD では，舌口唇運動機能などに加えて，発話の明瞭度や発話の異常性を第三者が主観的に評価し，点数化を行うことによってディサースリア (神経・筋系の病変に起因する発話の障害) の検査を行う．一方で，より客観的でコストのかからない計算機による評価指標の確立に向けて，これまで様々な音響特徴量が提案されてきた．

[2] https://www.jda.or.jp/oral_flail/gymnastics/
[3] https://www.msdmanuals.com/ja-jp/ホーム/09-脳、脊髄、末梢神経の病気/脳の機能障害/構音障害

構音障害を評価する古典的な音響特徴量として，jitter，shimmer，harmonic-to-noise ratio (HNR)，メル周波数ケプストラム係数 (Mel Frequency Cepstral Coefficient, MFCC) などが挙げられる．jitter，shimmer, HNR は，発声中の音声の様々な要素が健常者では周期的であるのに対し，構音障害者では非周期的であることに着目した特徴量である．jitter は発声中の基本周波数の偏差を，shimmer は振幅の偏差を定量化し，HNR は周期成分と非周期成分の比率を表す．実際に jitter，shimmer，HNR によって健常者と構音障害者を分類可能であることが示されている [22]．一方で，jitter, shimmer, HNR の算出に必要となる基本周波数は，性質が健常者と異なる構音障害者を対象とした場合に正しく計算できない可能性があるため，基本周波数を必要としない MFCC などの特徴量を用いて，構音障害を検出する手法も提案されている [15]．こうした手法に加え，非周期成分に着目した Reccurent Period Density Entropy (RPDE) [22]，detrended fluctuation analysis (DFA) [16] などこれまで様々な特徴量が提案され，構音障害の分類に有用であることが示されている．本研究でもこれらの音響特徴量を分析に利用した．

その他にも，これまで様々な音響特徴量が提案されており，その種類は膨大である．音声中のさまざまな特徴量の計算を可能とする DisVoice [4]には，これまで提案されてきた音響特徴量がまとめられている．このライブラリでは，音響特徴量の種類が glottal, phonation, articulation, prosody, phonological, Representation learning の 6 つに分類されており，本研究でもその一部を5章以降の音響分析で使用した．

その他にも，/pa/,/ta/,/ka/,/ra/を連続して発音するオーラルディアドコキネシス試験中の音声に対して音響的分析を適用する試みもある．オーラルディアドコキネシスの音声データを計算機によって分析するため，音声データを各音節ごとに分割し，分割された音声内で initial burst (空気流の生成の開始時点)，vowel onset (無声から有声に移り変わる時点)，occlusion (有音から無声に移り変わる時点) の位置を推定し，それらを利用した音響特徴量を計算することで，構音障害者と健常者を分類することができることが示されている [26]．本研究でも，4.2節において収集したオーラルディアドコキネシスのデータに対して音響分析を適用した．

2.5 嚥下障害の音響・画像分析

嚥下障害とは，食べ物をうまく飲み込めない状態のことである．

単純な発声を利用した音響分析によって嚥下障害を分類する手法はいくつか提案されている．例えば，嚥下後に/a/と持続して発音した際の jitter，shimmer，SNR といった音響

[4]https://github.com/jcvasquezc/DisVoice

特徴量が嚥下障害の予測に有効である可能性が示唆されている [36]．しかしながら，嚥下障害を発声のみから判断する研究例はまだ少なく，精密な検査をしばしば必要とする．その他に，嚥下中の音声に基づいて嚥下障害を分類する研究 [31] もいくつかあるが，本研究との関連性は低い．

2.6 高齢者とカラオケ

カラオケは機械から流れる伴奏に合わせて歌唱を行う娯楽活動で，我が国のみならず世界中で楽しまれている．

音読が認知機能の向上に寄与することや肺活量を向上させることに着目して，介護施設を利用している高齢者への歌唱トレーニングの効果を調査した研究では，認知機能が向上したと報告されている [23]．本研究では口腔機能の評価のみに限定して分析を行ったが，提案手法の活用により，認知機能の向上などの副次的な効果がある可能性がある．

2.7 まとめ

近年，特に高齢期において口腔機能が僅かに低下した状態，すなわちオーラルフレイルが将来の健康リスクを増加させることが示され，口腔機能の評価の重要性はますます高まっている．口腔機能のうち，本研究が着目する構音・嚥下障害を分類するために，これまで様々な音響特徴量が提案されており，本研究でもそれらを利用した．口腔機能の訓練とゲームを組み合わせた研究はこれまでにもいくつか存在しているが，本研究で提案する歌唱による評価手法はこれまでに存在していなかった．次章以降で本研究について詳しく説明する．

Chapter 3

歌唱データと構音・嚥下機能のデータ収集

筆者が調査した範囲では，歌唱時の音声・画像から口腔機能を評価する既存研究を確認することはできなかった．そこで提案手法の実現に向けて，歌唱時の録画データから口腔機能を評価するため，我々はまず，歌唱中のデータと口腔機能のデータを収集することにした．正確な口腔機能の評価のためには，機器を患者の口の中に挿入するなど侵襲的な手段が必要である．しかしながら，将来的には口腔機能に何かしらの障害を持つ患者のデータを収集する可能性があり，そうした方への侵襲的な口腔機能の検査はリスクが大きいため，我々は実験参加者への負担が少ないデータの取得方法を検討した．その結果，口腔機能のデータとして，飲み込みの機能である嚥下機能と，発音の機能に関連する構音機能を含めることとし，簡便に収集可能なデータとして，EAT-10 とオーラルディアドコキネシスを採用した．どちらの手法も，患者の体に直接触れることなく口腔機能を評価することが可能である．また，広い年代層が知っているであろう曲として，歌唱してもらう曲には童謡の「ふるさと」を選択した．我々が募った実験参加者の中にこの曲を知らない人はいなかった．本章では，収集したデータの詳細及び収集の手順について詳しく説明し，次章以降で詳細な分析結果を報告する．

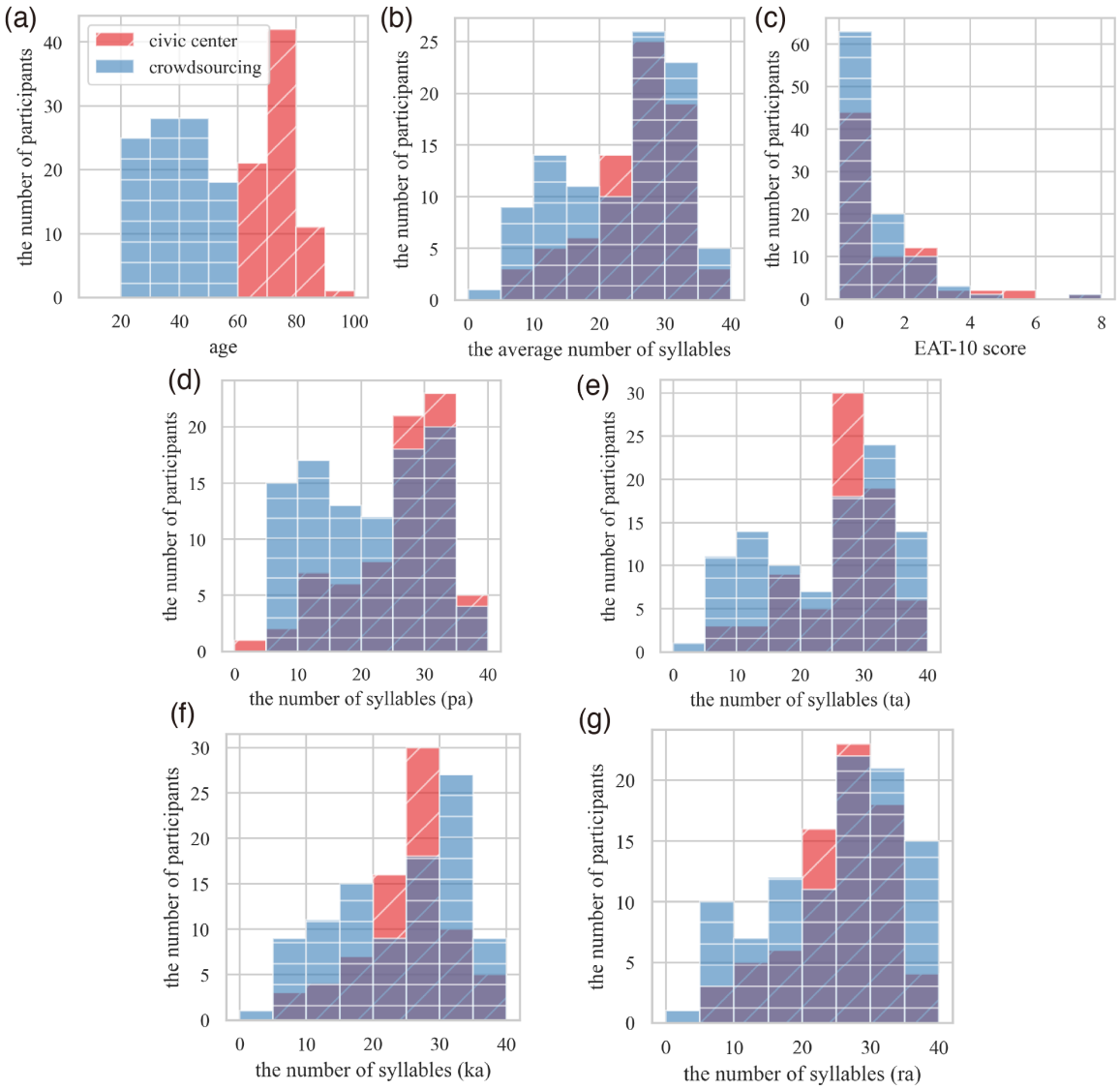

図 3.1: 参加者の分布を示したヒストグラム．青がクラウドソーシングの，赤が東京都文京区の高齢者の参加者を表す．左から順に，(a) 年齢，(b)/pa/，/ta/，/ka/，/ra/の 5 秒間の発音回数の平均，(c)EAT-10 スコアを表す．(d)，(e)，(f)，(g) それぞれ/pa/，/ta/，/ka/，/ra/の発音回数を示す．

3.1 実験参加者の募集

歌による嚥下・構音機能の定量的評価手法の実現に向けて，我々はまず，国内のクラウドソーシングサービスの CrowdWorks [1]と Craudia [2]で実験参加者を募集し，118 件のデータを得た．各実験参加者には謝金として一律で 220 円を支払った．

クラウドソーシングで募集した参加者の年齢を調べた結果，図 3.1 (a) に示すようにほとんどが 60 代未満であった．そこで，我々がめざすアプリケーションの主な対象者である高齢者のデータを収集するため，我々は東京都文京区で 60 歳から 90 歳の計 76 名の実験参加者を募集した．各実験参加者には謝金として 1000 円を支払った．なお，その際のデータ収集はオンラインでは実施せず，実験は全て同一施設で行い，モバイル端末の操作は主に著者らが行った．また，モバイル端末には Apple 社の iPhone SE (第二世代) を使用した．撮影時には三脚を用いてスマートフォンを固定し，撮影中に揺れ動かないようにした．

その後，さらに嚥下等に何らかの病的な症状を持つ高齢者のデータ収集を行うため，嚥下リハビリテーション科に往診に来る患者のデータ収集を行った．収集方法は区民センターの時と同様とし，リハビリテーション科の先生方にスマートフォンの操作をお願いした．使用端末に関しても，iPhone SE 及び三脚を貸与し，同様の撮影環境下でデータ収集が行われるようにした．その結果，合計で 9 件のデータを取得した．これらのデータは4章では使用せず，5章以降で使用した．

3.2 実験の実施手順

表 3.1: EAT-10 の質問内容

質問内容
質問 1: 飲み込みの問題が原因で、体重が減少した
質問 2: 飲み込みの問題が外食に行くための障害になっている
質問 3: 液体を飲み込む時に、余分な努力が必要だ
質問 4: 固形物を飲み込む時に、余分な努力が必要だ
質問 5: 錠剤を飲み込む時に、余分な努力が必要だ
質問 6: 飲み込むことが苦痛だ
質問 7: 食べる喜びが飲み込みによって影響を受けている
質問 8: 飲み込む時に食べ物がのどに引っかかる
質問 9: 食べる時に咳が出る
質問 10: 飲み込むことはストレスが多い

[1]https://crowdworks.jp/
[2]https://www.craudia.com/

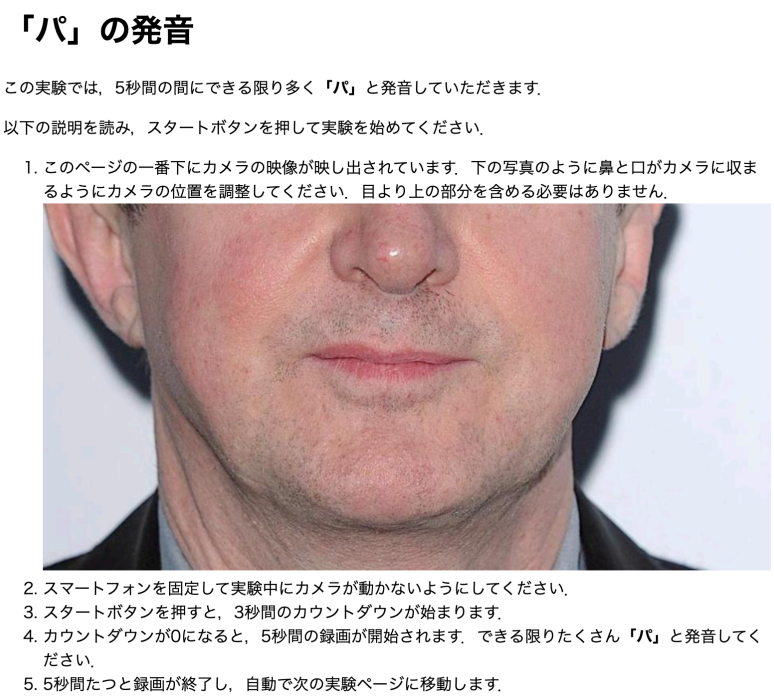

図 3.2: オーラルディアドコキネシス試験のために作成したウェブページ．画像は/pa/の発音を行うページである．実際のウェブページでは，この下にスマートフォンのカメラの映像が映し出されており，実験参加者はそれを見てカメラの位置を調整することになる．

実験は以下の手順で実施した．まず，Google Forms [3]を用いたアンケート調査を行った．アンケートでは，性別・年齢を記入したあと，「食べ物を飲み込む際にむせこむことはありますか?」「話すときにろれつが回らないと感じたことはありますか?」「電話で話す時、ろれつが回らないなどで意思疎通に問題がありますか?」の 3 つの質問について，「よくある」「たまにある」「ほとんどない」「全くない」の 4 段階で回答をお願いした．次に，EAT-10 による嚥下機能の評価を行った．EAT-10 は，例えば「飲み込みの問題が原因で、体重が減少した」などの合計 10 個の質問に対して，「0= 問題なし」から「4= ひどく問題」の 5 段階で主観的に答えるアンケートで，合計点数が 3 点以上の場合には結果を専門医に相談することが推奨される．日本を含む世界各国で嚥下機能を評価する手法として有効であることが示されている [25, 8]．質問内容を表 3.1に示す．Google Forms 上での回答が終了すると，これらのデータとその後に収集するデータを紐づけるため，被験者ごとの ID が出力されるようにした．

その後，我々が作成したウェブサイト上で以下の実験を実施した．実験を開始する前に，Google Forms 上で出力された先述の ID を入力することで，被験者ごとのデータを紐づけることができるようにした．

[3]https://www.google.com/forms/about/

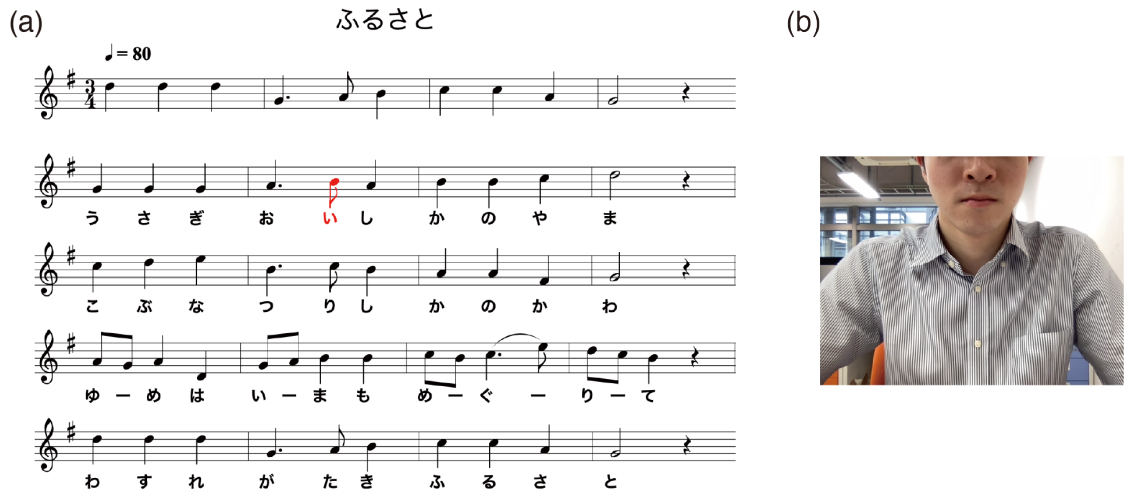

図 3.3: 「ふるさと」の歌唱の実験のため使用したウェブページ (抜粋)．(a) 歌唱中は「ふるさと」の伴奏が流れ，歌うべき箇所が楽譜上で赤く表示される．また，スマートフォンのフロントカメラにより，(b) のように映像が表示される．プライバシーを考慮し，被験者には鼻を含めた鼻より下の範囲が映るようにカメラの位置を固定するようお願いをした．実際に実験に使用したウェブサイトでは説明文が書かれているほか，一部レイアウトが図とは異なる．

まず，/pa/，/ta/，/ka/，/ra/の語をそれぞれ 5 秒間できるだけ多く繰り返し発音するオーラルディアドコキネシス試験を行い，その様子の録音と録画を行った．/pa/の発音を行う際に使用したウェブページを図 3.2に示す．なお，プライバシーを考慮し，実験中は図 3.2のように，鼻と口が映り，鼻より上の部分が写らないようにカメラの位置を調整するようお願いした．各語の発音後，撮影時の録画データが，ID と紐づけられてサーバー上にアップロードされるようにした．

次に，これらの嚥下・構音機能と歌唱時との関連性を調べるため，歌唱とその様子の録音・録画を行った．撮影範囲はオーラルディアドコキネシスの時同様に，鼻より下の範囲とした．歌唱する歌には，認知度の高い曲として「ふるさと」 (高野辰之作詞・岡野貞一作曲) を選択した．実験を開始すると，スマートフォンから伴奏が流れ始め，図 3.3 (a) に示すように，その時点で歌うべき箇所が赤色で強調して表示される．

システムの不具合によりファイルサイズが0 Byte であるものを含む参加者のデータ，いくつかの実験を正しく行わなかった参加者のデータ，正しく録音・録画ができなかったと申告のあった参加者のデータを除外し，クラウドソーシングで 99，シビックセンターで 75 件のデータを得た．ただし，音声には問題がないものの，録画がうまく行っていないデータが 1 件，クラウドソーシングで収集したデータに含まれていた．「ふるさと」の歌唱の際，実際の歌唱前に数秒間ラグが入る場合があったため，このラグをあらかじめ消去し，どのデータの歌唱時間も 45 秒間になるように調整した．

3.3 収集したデータの分布

図 3.1 (a) に実験参加者の年齢分布を示す．クラウドソーシングで収集したデータのほとんどは 60 歳未満であったが，区民センターでのデータで新たに実験参加者の募集を行なったことで，幅広い年代層のデータが収集できていることがわかる．

図 3.1 (b) にオーラルディアドコキネシス試験時の/pa/，/ta/，/ka/，/ra/の発音回数の平均の分布を示す．回数の測定は筆者が手動で行い，記録した．図 3.1 (d) (e) (f) (g) に，/pa/，/ta/，/ka/，/ra/のそれぞれの発音回数の分布を示す．クラウドソーシングの実験参加者の方が区民センターでの実験参加者と比較して年齢層が低いにも関わらず，発音回数は比較的少なくなってしまっている．この点については，5章で考察する．

図 3.1 (c) に EAT-10 の合計点の分布を示す．ほとんどの実験参加者は 3 点未満となっているが，一部の参加者は医師への診断が推奨される 3 点以上となっている．

3.4 まとめ

提案手法の実現に向けて，我々は 65 歳以上の高齢者を含む幅広い年代層の歌唱と口腔機能のデータを収集し，計 183 件の有効なデータを得た．次章では，収集したデータを用いて，歌唱から構音・嚥下機能を分類することを試みる．

Chapter 4

変数増減法によるロジスティック回帰分析

前章では，筆者らが実施したデータ収集の手順について詳述した．本章では，収集した歌唱中の音声・画像データから，収集した口腔機能のデータである構音・嚥下機能を，変数増減法によるロジスティック回帰によって分析した．なお，音響分析および画像分析はそれぞれ個別に行った．

4.1 嚥下・構音障害の基準

機械学習による嚥下・構音障害の分類を行うにあたって，嚥下・構音機能障害の基準を定める必要がある．本研究では，医師への相談が推奨される，EAT-10 スコアの合計が 3 点以上を嚥下障害と定義した．また，構音障害はオーラルディアドコキネシス試験の際の/pa/，/ta/，/ka/，/ra/の 5 秒間の発音回数のいずれかが 15 回以下であることとした．その結果，174 件のデータのうち，52 件が構音障害に，15 件が嚥下障害に該当した．

4.2 オーラルディアドコキネシスの音響分析

まず，オーラルディアドコキネシスに対する音響分析の結果が先行研究と一致するかどうかを確かめるため，各発音を音節ごとに分割して，特徴量を計算することで健常者と構音障害者を計算機によって区別する先行研究を再実装し，評価を行った [26]．図 4.1に示すように，この手法では音声信号を 1 音節ごとに分割したのち，initial burst (空気流に起因する発音の開始地点), vowel onset (声帯の振動に起因する周期的な音声信号の開始地点), occlusion (緩やかな音声信号の減少) を求めて，特徴量を計算する．なお，再実装にあ

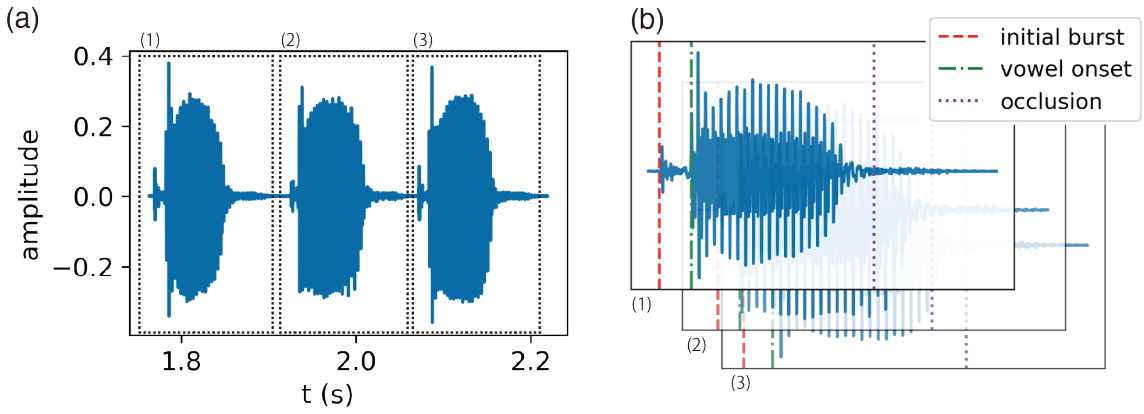

図 4.1: オーラルディアドコキネシス時の連続した発音の分析手法 [15]．(a) 同じ語を繰り返し発音した時の音声を音節ごとに分割する．(b) (a) において分割された各音節ごとに，initial burst, vowel onset, occlusion の箇所を計算する．

表 4.1: /pa/の連続発音から構音・嚥下障害を予測した時の混同行列

	構音障害	Predicted P	Predicted N
Actual	P	34	18
	N	4	118

	嚥下障害	Predicted P	Predicted N
Actual	P	0	15
	N	0	159

たっては，著者らが提案したベイズ変化点検知手法の代わりに，より実装の容易な 2 元変化点検知を用いた [12]．特徴量には，/pa/の連続発音時の vowel variability quotient (vvq, 1 音節中における振幅の分散)，vowel onset time (vot, initial burst から vowel onset までの時間)，consonant spectral trend (cst，スペクトラムの線形回帰係数) の 3 つを選択し，これを説明変数とした．また，目的変数は嚥下障害および構音障害とした．そして，サポートベクターマシン (support vector machine，SVM) による 5 分割交差検証によって音響特徴量の有用性を評価した．

SVM による 5 分割交差検証による構音・嚥下機能の予測結果を4.1に示す．構音障害の予測では，正解率 85%，適合率 93%，再現率 65% となり，構音障害を概ね分類可能であることが分かった．これは先行研究と一致する [26]．一方で，嚥下障害の予測では全てが陰性だと判定され，適合率，再現率がともに 0% となり，今回用いた特徴量は嚥下障害の予測には有用でないことが分かった．

表 4.2: 歌唱データの音響分析に用いた特徴量

名称	説明
stdF0	基本周波数の標準偏差
hnr	音声の周期性の比率
jitter	基本周波数の変動する割合
shimmer	振幅の変動する割合
cpp	ケプストラムのピーク値の卓立度
dfa	DFA よって計算された Hurst 係数 [16]
dfa_norm	dfa をシグモイド関数に代入した値

4.3 使用した歌唱中の音響特徴量

我々はまず，前節と同様の計算手法により，initial burst，vowel onset，occlusion の検出を行い，特徴量の計算を試みようとした．しかしながら，/pa/などの特定の語を連続で発音するオーラルディアドコキネシスと違い，歌唱中は音節ごとの境界が明確ではないため，この手法はうまく機能しないことが分かった．

そこで，歌唱データが伴奏に合わせて歌われたものであるため，開始から何秒の時点でどの部分を歌唱したのかをおおよそ知ることができることに着目した．音節を分割する様子を図 4.2に示す．まず，図 3.3 (a) の楽譜の歌詞がある音符ごとに録画データを分割し，これらを音節とした．その際，実際の歌唱は，本来歌うべきパートからずれることがあることを考慮して，境界の前後に100 ms のマージンを加え，本来の時間から200 ms 短くなる用にした．次に，分割された各音声に対して，4.2に示す特徴量を計算し，特徴量ごとに標準化を行った [1]．

4.4 使用した画像特徴量

構音・嚥下機能の評価に向けた画像分析は未開拓の領域であり，どのような特徴量がそれらの予測に適しているのかも不明である．そこで，我々は撮影画像から口の高さ・幅と口の大きさに対して，時系列データ一般に適用される特徴量を歌唱中の動画データに対して適用し，有用な特徴量を選ぶことにした．

まず，口の高さ・幅と口の大きさを歌唱中の録画データから推定する手順を図 4.3に示す．まず，図 4.3 (a) のように，収集した歌唱中の動画データをフレームごとに分割して画像データとし，haar cascade を用いて口の範囲を抽出し，その範囲を中心に縦と横の大きさを 3 倍に拡大した後，縦128 px, 横 256 px になるようにリサイズした [35]．ただし，haar cascade では正しく口の範囲を推定できていないことが多かった．本章では，完全なシステムの構築ではなく，歌唱と嚥下・構音機能の関係の調査を目的とするため，うまく口の

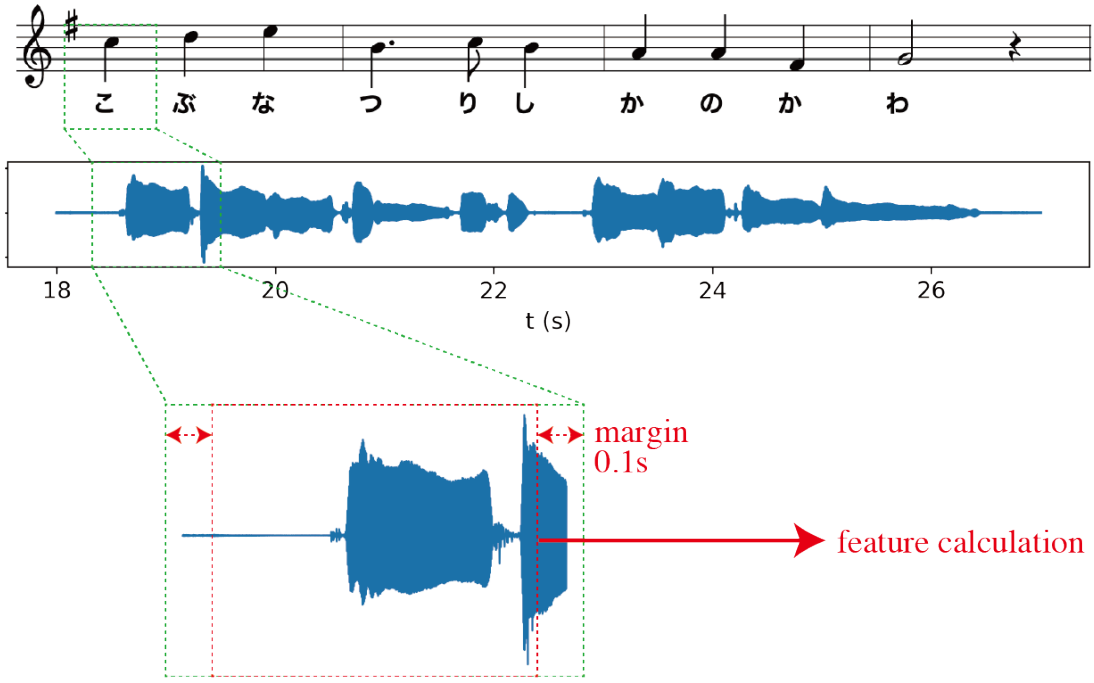

図 4.2: 音節の分割を行う様子．楽譜データをもとに各音節を分割し，それぞれの区間に対して音響特徴量の計算を行う．ただし，歌い遅れが存在することを考慮して前後に100 ms，合計200 ms のマージンを加えている．

範囲が切り出せていない画像は手動でラベリングし，その画像の口の範囲は，正しい前後の切り出し範囲と同じとした．

次に，haar cascade によって切り出された口の画像のどのピクセルが口に対応しているかを調べるため，近年提案された MaskGan を用い，顔画像の各ピクセルを顔の部位ごとに分類する face segmentation を行った [20]．MaskGan では，顔画像と手動でラベリングされた正解データをディープラーニングにより学習することで face segmentation を行う．MaskGan に用いられたデータセットである CelebA-HQ には，正面を向いた人の顔画像が含まれており，下半分のみを切り出すことで，鼻より下の範囲のみが概ね映るようにすることができる．そこで，鼻より下の部分のみを含む我々のデータに対して MaskGan を適用するため，CelebA-HQ の各画像の下半分のみを用いて再学習を行った．MaskGan による顔の部位の分類結果を図 4.3 (b) に示す．本稿では，MaskGan による face-segmentation の結果出力される顔の部位のうち，上唇，下唇，口の中の部分を口とみなした．図 4.3 (c) のように，口に分類された部分のみを白く表示すると，本来口ではないピクセルの一部が誤検出されていることがわかる．誤検出を排除するため，隣り合う各ピクセルを連結した

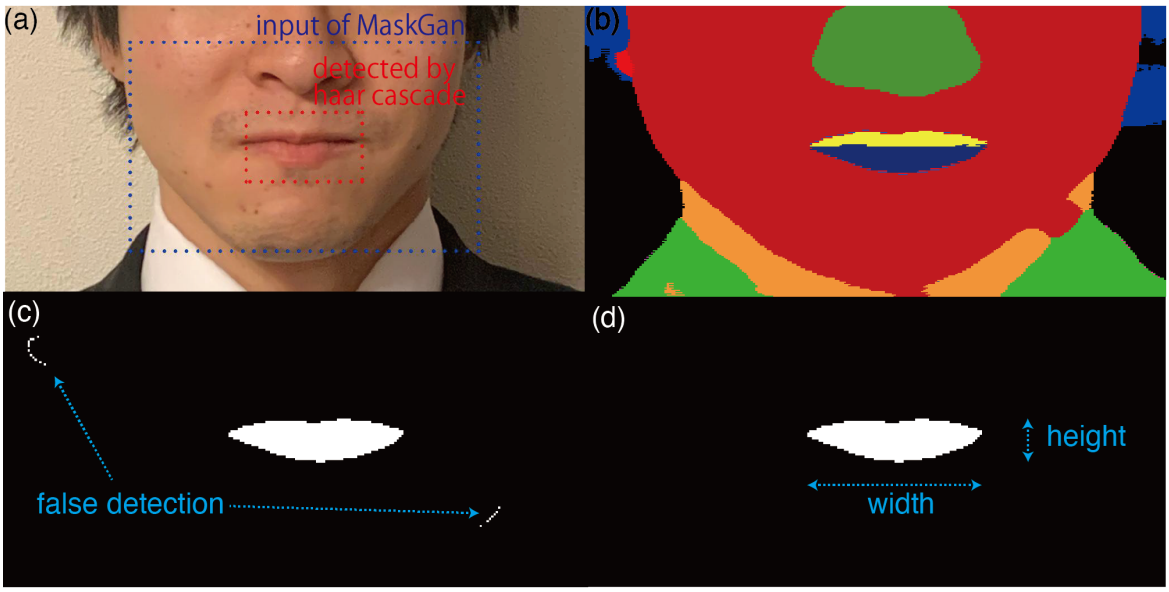

図 4.3: 画像中の口の部分を推定手順．(a) まず，haar cascade を用いて口の範囲を推定し，縦横比が 1:2 になるように調整する (図中赤点線) [35]．その後，縦・横にそれぞれ 3 倍に拡大した範囲を MaskGan の入力とする (図中青点線) [20]．(b)MaskGan によって出力された face segmentation の結果．顔の部位ごとに色分けをしている．(c)(b) の内，上唇・下唇・口の中のみを白で表示した画像．一部がご検出されていることが確認できる．(d) 口に分類されたピクセルのうち，隣り合う各ピクセルを連結したときに総ピクセル数が最大になるもののみを口とし，白で表示した画像．

時に，総ピクセル数が最大になる部分のみを口であるとした．最終的に口に分類された部分を白く表示した画像を図 4.3 (d) に示す．この時の高さ・幅・口の面積 (口の総ピクセル数) を動画の各フレームに対して計算した．

その後，前節と同様に歌詞ごとにデータを分割したのち，時系列データ分析のための Python のライブラリである tsfresh[1] を用い，高さ・幅・口の面積に対する特徴量の列挙を行った．そして，前節の音響特徴量の選択手法と同様の変数増減法により，有用な特徴量を選択し，ロジスティック回帰による分類性能を確かめた．

4.5 変数増減法に基づくロジスティック回帰分析

得られた特徴量に対して以下の手順で変数増減法によるロジスティック回帰分析を行い，有用な特徴量の選択を行った．まず，選択された特徴量がゼロの状態からスタートし，特徴量を加えたときに赤池情報量規準 (AIC) が減少する特徴量を有用な特徴量として選択する．ただし，多重共線性を回避するため，分散拡大係数が 5 より大きくなる特徴量は除いた．これを，AIC がそれ以上小さくならなくなるか，選択された特徴量の数が 10 個

[1] https://github.com/blue-yonder/tsfresh

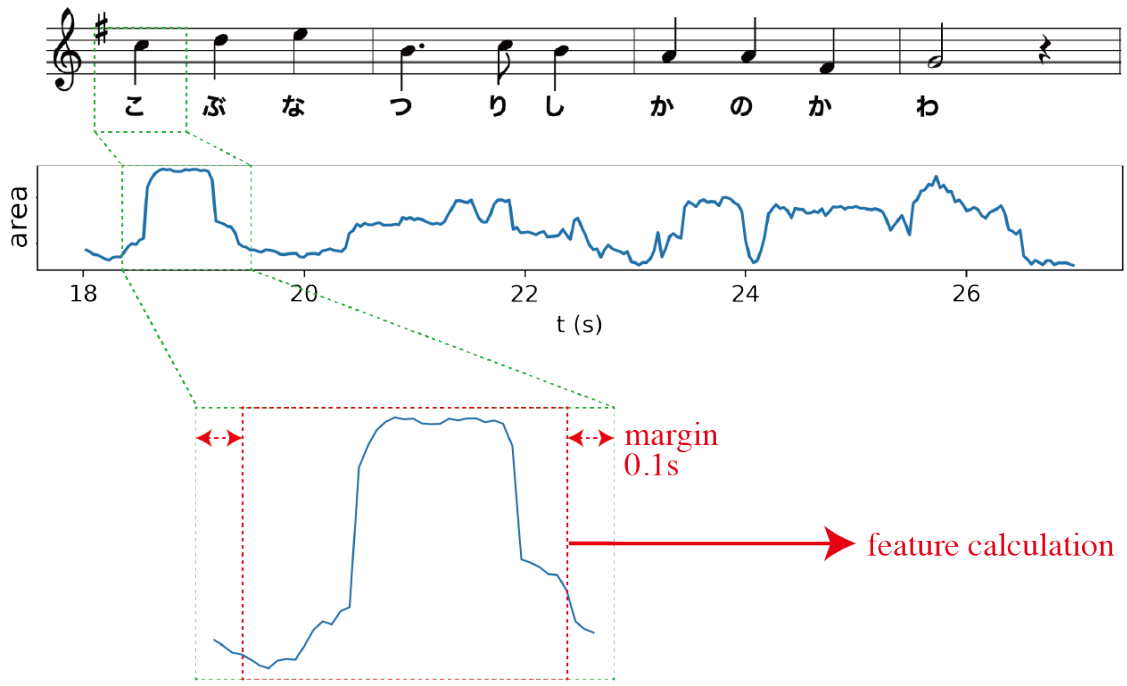

図 4.4: 図 4.2と同様に，各音節ごとに分割された区間に対して特徴量の計算を行う．

に達するまで繰り返した．その後，得られた特徴量による嚥下・構音障害の分類性能をロジスティック回帰によって確かめた．

変数増減法によって得られた音響特徴量を4.3，4.4に示す．また，これらの特徴量を用いたロジスティック回帰分析によって得られた混同行列を4.5に示す．嚥下障害では正解率 92%，適合度 67%, 再現率 13% となり，特に再現率の低さが目立つ．一方で構音障害の分類の場合，正解率 76%，適合度 67%, 再現率 42% となり，再現率を比較して，嚥下障害より高い分類性能が得られた．

音響分析と同様に，変数増減法によって得られた画像特徴量を表 4.6，表 4.7に示す．表 4.6の height_autocorrelation は口の高さの自己相関を，height_large_std は口の高さの最大と最小の差と標準偏差との比率と表す．また，表 4.6の width_lempel_complexity は変化の複雑度を定量化した値であり [21]，area_change_quantiles は口の面積の四分位範囲の変化の度合いを定量化した値である．これらの特徴量を用いたロジスティック回帰分析によって得られた混同行列を表 4.8に示す．嚥下障害では，正解率 94%，適合率 100%，再現率 29% となっており，音響分析の場合と同様にほとんどが嚥下障害者ではないと分

表 4.3: 変数増減法により選択された構音障害を予測する音響特徴量．括弧内は音符の番号と歌詞を表す (以下同様)．

feature	coef	std err	$P > \|z\|$
const	-1.1820	0.218	0.000
hnr(16_shi)	1.1958	0.312	0.000
jitter(25_tu)	0.4506	0.207	0.030
hnr(13_gi)	-0.9857	0.339	0.004
cpp(23_bu)	-0.5643	0.246	0.022
cpp(20_ma)	0.7434	0.242	0.002
cpp(53_ta)	-0.5836	0.224	0.009
hnr(12_sa)	0.6716	0.326	0.040
jitter(39_ma)	-0.6282	0.257	0.015
shimmer(50_su)	0.4384	0.208	0.035
cpp(44_u)	-0.4147	0.233	0.075

表 4.4: 変数増減法により抽出された嚥下障害を予測する音響特徴量

feature	coef	std err	$P > \|z\|$
const	-6.6844	1.406	0.000
dfa_alpha(36_ha)	-1.7405	0.643	0.007
dfa_alpha_norm(45_ri)	4.1206	1.639	0.012
cpp(38_i)	2.5413	0.702	0.000
cpp(20_ma)	-3.0874	0.904	0.001
cpp(13_gi)	1.8324	0.581	0.002
dfa_alpha(47_te)	-2.1185	0.752	0.005
dfa_alpha_norm(14_o)	2.5197	0.872	0.004
dfa_alpha(43_gu)	-2.8262	1.103	0.010
hnr(16_shi)	-1.4880	0.571	0.009
hnr(46_i)	1.0126	0.499	0.042

表 4.5: 歌唱中の音響特徴量による構音・嚥下障害予測時の混同行列

		Predicted	
	構音障害	P	N
Actual	P	22	30
	N	11	111

		Predicted	
	嚥下障害	P	N
Actual	P	2	13
	N	1	158

類されている．一方構音障害では，正解率 78%，適合率 77%，再現率 40% となり，嚥下障害と比較すると，特に再現率に関して良い結果が得られた．

表 4.6: 変数増減法により抽出された嚥下障害を予測する画像特徴量

feature	coef	std err	$P > \|z\|$
const	-4.1361	0.705	0.000
height__autocorrelation(14__o)	1.8895	0.480	0.000
height__large__std(50__su)	-1.4336	0.392	0.000

表 4.7: 変数増減法により抽出された構音障害を予測する画像特徴量

feature	coef	std err	$P > \|z\|$
const	-1.0912	0.202	0.000
width__lempel__complexity(26__ri)	1.3178	0.313	0.000
area__change__quantiles(43__gu)	6.2685	2.526	0.013

表 4.8: 歌唱中の画像特徴量による構音・嚥下障害予測時の混同行列

	構音障害	Predicted P	N
Actual	P	20	31
	N	6	114

	嚥下障害	Predicted P	N
Actual	P	4	10
	N	0	157

4.6 まとめ

本章では，歌唱中の音声・画像を音節ごとに分割し，音響・画像特徴量を抽出して変数増減法によるロジスティック回帰分析を行った．その結果，ある程度の分類を行うことができたものの，再現率が高く，偽陽性が多くなっているという結果が得られた．本章の結果を踏まえて，次章で改善を行った変数増減法によるロジスティック回帰分析を行う．

Chapter 5

ロジスティック回帰分析の改良

5.1 前章での分析の問題点と改善策

前章までにおこなった手法では，構音・嚥下障害ともに再現率が低くなっており，満足できる分類性能が得られているとは言い難い．そこで我々は，前章での分析にはいくつか問題があると考え，それらの対策を加えて再度分析を行うことにした．以下で問題点とその改善策を列挙する．

問題点の 1 つ目は楽譜データに基づいて行なった音節の分割である．実際の歌唱においては，歌い遅れるなどして本来の歌唱よりも時間差を生じて歌ってしまう例がしばしば見受けられた．その対策として，200 ms のマージンを加えて対処したものの，これによって全ての歌い遅れを考慮できるわけでは無いため，より正確な音節の分割が実際には必要である．音声分析において，音声とそれに対応するトランスクリプトから，その音声のどの時点で何を話しているかを推定する分析は forced word alignment と呼ばれ，広く研究されている．本章では，この既存技術を用いることで音節のより正確な分析を試みる．

また，音響特徴量が少ないという点も問題点として挙げられる．今回は，これまでの研究で有用であり，かつ有名である音響特徴量をもとにロジスティック回帰を行ったが，これらの音響特徴量の他にもこれまで様々な音響特徴量が提案されてきた．これらを組み合わせることによって，更なる分類精度の向上が期待できる．追加する音響特徴量については次節以降で説明を行う．

さらに，前章での分析に用いたデータには，何らかの病気の診断を受け，通院している者のデータは含まれていなかった．そこで本章以降では，摂食嚥下リハビリテーション科に通院している患者のデータを加えることで分類性能の向上を図る．

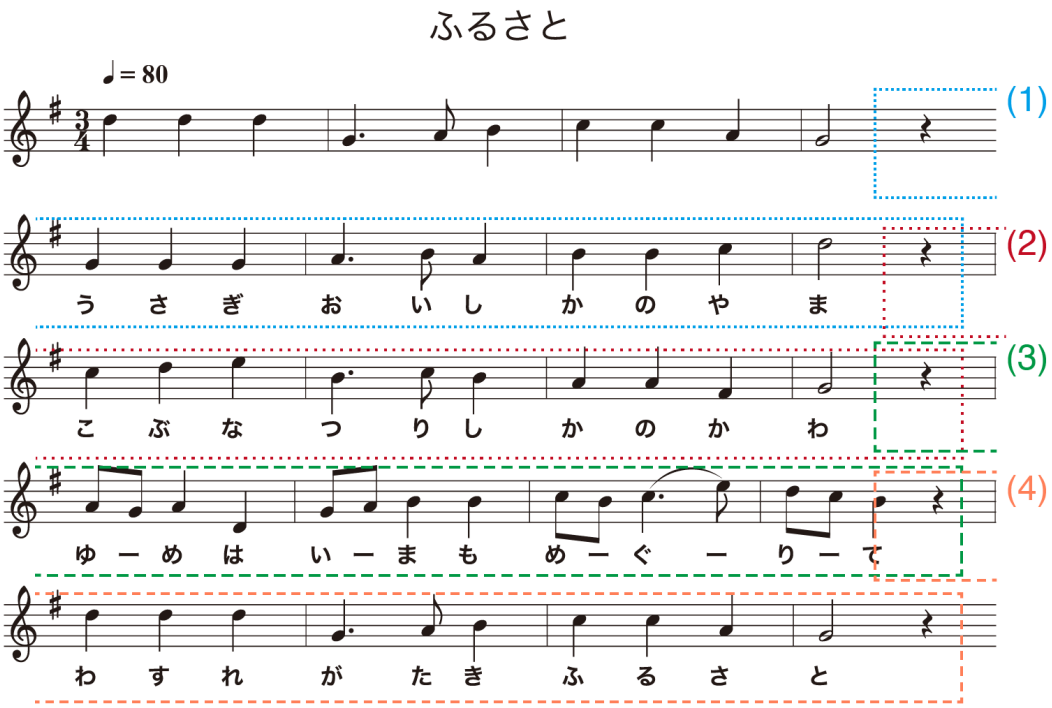

図 5.1: forced word alignment を行う際の楽譜の分割方法．4 小節を 1 まとめにして，前後の休符を含めた時間で音声ファイルを分割した．分割した際の開始・終了時刻を表 5.1に示す．

その他にも，図 3.1 (b) では，総じて年齢層の若い被験者が含まれるクラウドソーシングの被験者のオーラルディアドコキネシスの発音回数が，高齢者のデータを主に含む区民センターの被験者の発音回数よりも少なくなっている場合が多い．より若い被験者の方が構音障害を持つ可能性が低いことを踏まえると，これは奇妙な現象である．考えうる原因として，クラウドソーシングの被験者はオンライン上で実験を実施したために，オーラルディアドコキネシスで行うべき，なるべく速い連続した発音の趣旨を理解できなかった可能性が高い．従って，クラウドソーシングの参加者は，発音回数が低い場合であってもこれを構音障害として分類するのは問題があると考えられる．構音障害については基準値の再考が必要である．

5.2 forced word alignment による音節の分割

音節の正確な分割には，先述の forced word alignment の技術を用いた．現状，日本語に対応した forced word alignment のライブラリは少ないが，本研究では日本語の音声認識

表 5.1: 各グループの開始時間と終了時間

歌詞	start (s)	end (s)
うさぎおいしかのやま	9.0	18.0
こぶなつりしかのかわ	17.25	27.0
ゆめはいまもめぐりて	26.25	36.0
わすれがたきふるさと	35.25	45.0

でしばしば利用されるプログラムである julius [1]を使用した [19]．julius には，あらかじめ与えられた言語モデルに基づいて，単語ごとに音声区間の検出を行う機能 (forced word alignment) を備えている．言語モデルには，語の遷移の規則が記述され，ここでは曲の歌詞がこれに相当する．julius は言語モデルに合致する語の遷移の中で最も尤度が高いものを計算し，その語の音声区間を出力する．本研究では以下のように言語モデルの作成を行なった．

歌詞に基づいた言語モデルを作成するに先立って，まず，図 5.1に示すように，歌詞のある部分を 4 小節ごとにひとまとめにして「うさぎおいしかのやま」,「こぶなつりかのかわ」,「ゆめはいまもめぐりて」,「わすれがたきふるさと」の 4 つに分割した．さらに，歌い遅れを考慮して，それぞれの前後に存在する休符を含めた区間に分割する．このように分割するのは，forced word alignment の精度を向上させるためである．すなわち，できるだけ音声とそれに対応するスクリプトが短いほど forced word alignment の精度は向上すると考えられるが，実際の歌唱においてどれだけ歌い遅れが生じているかはわからないため，短すぎる区間で分割するとスクリプトにずれが生じるかもしれない．そこで，休符を跨いだ歌い遅れが起こる可能性は低いと仮定し，このような分割を行っている．参考までに，表 5.1に各グループの開始時刻と終了時刻を示した．この時刻に基づいて各音声ファイルを分割する．

次に言語モデルの作成を行なった．言語モデルには，入力された音声に含まれる語のありうる遷移の規則を入力する．julius における言語モデルを作成するためには，grammar ファイルと voca ファイルを作成する必要がある．各ファイルの記述方法は julius のドキュメントに書かれている[2]．文法規則などはそちらのドキュメントを参照されたいが，ここでは工夫点などを説明するため，grammar ファイルと voca ファイルの内容を簡単に解説する．grammar ファイルには構文制約を単語のカテゴリを終端規則として記述し，voca ファイルにはカテゴリごとに単語の表記と読み（音素列）を登録する．以下に本研究で用いた grammar ファイルと voca ファイルの内容を示す．なお，後述するように本研究では julius の Gausian Mixture Model (GMM) 版と Deep Neural Network (DNN) 版の両方を用

[1]https://so-zou.jp/software/tech/library/julius/
[2]https://julius.osdn.jp/juliusbook/ja/desc_lm.html

いているが，voca ファイルの記述方法は両者でやや異なる．ここでは GMM 版の voca ファイルを示しているが，DNN 版も本質的には同様の内容である．

grammar ファイルには文法規則が記述され，voca ファイルには各語の音素が記述される．NS_B, NS_E は最初と最後の無音区間を示している．また，歌唱時に各小節ごとに息継ぎなどに起因する無音区間が入る可能性を考慮して，NOISE (voca ファイルでは sp) を挿入している．さらに，通常の発音時と異なり，歌唱時には音を伸ばして発音する可能性があることを踏まえ (例：うーさーぎおーいしー)，voca ファイルに記述する音素には，通常の発音 (例：k o) に加えて，語尾を伸ばす発音 (例：k o:) を加えている．音声認識時には，どちらかのうち尤度の高い方が選択される．

コード 5.1 grammar ファイルの例

```
1 S:NS_B KOBUNA TURISHI KANOKAWA NS_E
2 KOBUNA:KO BU NA
3 KOBUNA:KO BU NA NOISE
4 TURISHI:TU RI SHI
5 TURISHI:TU RI SHI NOISE
6 KANOKAWA:KA NO KA WA
7 KANOKAWA:KA NO KA WA NOISE
```

コード 5.2 voca ファイルの例 (GMM 版)

```
%KO
こ k o
こ k o:
%BU
ぶ b u
ぶ b u:
%NA
な n a
な n a:
%TU
つ t u
つ t u:
%RI
り r i
り r i:
%SHI
し sh i
し sh i:
%KA
か k a
か k a:
%NO
の n o
の n o:
%WA
わ w a
わ w a:
%NOISE
sp sp
% NS_B
<s> silB
% NS_E
</s> silE
```

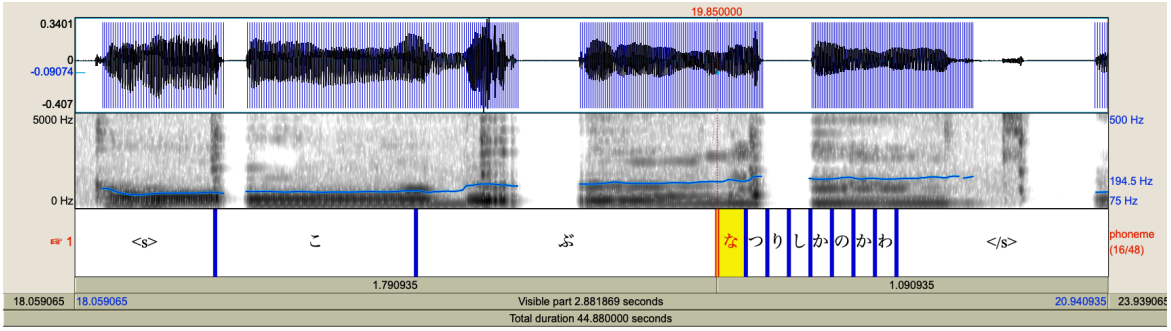

図 5.2: julius の DNN 版を用いて forced word alignment を行った結果．「こぶなつりしかのかわ」にあたる部分を表示している．「こ」及び「ぶ」は概ね正しい区間が出力されているものの，その他の語では出力された区間が本来の区間よりも大幅に短くなっている．

先述したように，julius には Gausian Mixture Model (GMM) 版と，Deep Neural Network (DNN) 版の 2 種類が存在する．筆者が GMM 版と DNN 版の両者を収集した歌唱データに適用し，音声ファイルを確認したところ，どちらを使った場合においてもいくつかの部分で正しく区間を出力できていないことがわかった．参考までに，図 5.2に julius の DNN 版によって出力された区間を可視化した図を示す．それぞれの出力区間を確認したところ，正しく出力されなかった区間の長さは，本来出力されるべき時間よりも大幅に短いかあるいは大幅に長いことが確認された．また，DNN 版で誤った区間が出力されている部分でも，GMM 版では正しく出力されている区間があり，その逆も確認された．そこで，DNN 版と GMM 版，及び楽譜データをそれぞれ使い分けることによってより正確な forced word alignment の実現を目指すことにした．

まず，DNN 版によって forced word alignment の結果を出力し，各区間の長さを楽譜データと比較する．その際，楽譜データと比較して，本来歌うべき区間よりも 50% 長く，あるいは短く区間が出力されてしまっている場合，その区間は正しくないとして棄却した．次に GMM 版によって同様に word alignment の結果を出力した．この時，DNN 版によっては正しく結果を出力できなかった区間において，GMM 版によって出力された区間の長さが楽譜データとしても 50% 長く，あるいは短く区間が出力されてしまっていない場合，それを正しい区間とし，それ以外は棄却した．最後に，DNN 版と GMM 版のどちらによっても正しく出力されなかった区間に対しては，楽譜データを正しい区間として採用することにした．

5.3 音響特徴量の追加

前章では，jitter や shimmer など，比較的単純で古くから有効性が確認されてきた特徴量を用いて分析を行った．一方で，これらの特徴量の他にも，構音障害等の計算機による

検出に向けてこれまで様々な音響特徴量が提案されてきた．こうした特徴量をさらに追加することで，分類精度の向上が期待できる．

新たな特徴量の計算のために，我々はこれまで提案されてきた音響特徴量をまとめたライブラリである DisVoice [3]を用いることにした．このライブラリでは，音響特徴量の種類を glottal, phonation, articulation, prosody, phonological, Representation learning の 6 つに分類している．本研究では，技術的な問題で使用できなかった phonological を除いた glottal, phonation, articulation, prosody, phonological, Representation learning を使用することにした．これらによって計算される音響特徴量の種類は多岐にわたるため，追加された全ての音響特徴量については説明しないこととし，後の分析で重要となった音響特徴量についてのみ適宜説明を行う．

以上の DisVoice による音響特徴量に加えて，非周期成分に着目した音響特徴量で，構音障害の分類に有用であることが示されている Reccurent Period Density Entropy (RPDE) を追加した [22]．

また，嚥下機能低下には性差が関わっており，これを考慮したして性別を特徴量として追加した [24]．年齢も重要な説明因子であると考えらるが，今回のデータ収集では東京都文京区の区民センターでの研究参加者を 60 歳以上に限定するなど，特に年齢に関してデータの偏りがあるため，年齢だけでも構音・嚥下障害を簡単に予測できてしまうため，年齢は特徴量として追加しなかった．

5.4 嚥下機能に障害を持つ患者のデータの追加

リハビリテーション科に通院する 65 歳以上の患者を対象に，これまでと同様のデータ収集を行い，計 9 件のデータを追加した．この 9 名の全てが，医師への診断が推奨される EAT-10 スコア 3 点以上に該当していた．

5.5 構音障害の定義の再考

オーラルディアドコキネシスの発音回数の基準値は各病院等が独自に定めていることが多く，統一した基準値が存在しないのが現状ではあるが，ここでは先行研究を参考に 5 秒間で 20 回以下の発音を構音障害であると定めた [44]．

だが，このように構音障害を定義することによって，以下のような問題が生じる．図 3.1 (a) に示されるように，クラウドソーシングで募った実験参加者の年齢は概ね 60 歳未満であり，区民センターでは概ね 60 歳以上である．しかしながら，図 3.1 (b) から読み

[3]https://github.com/jcvasquezc/DisVoice

取れるように，より若年層を含むクラウドソーシングの実験参加者の方が，オーラルディアドコキネシスの発音回数が少ない参加者が多くなっている．加齢に伴い口唇機能が低下することが自然であることを踏まえると，これは奇妙である．この原因の一つとして，クラウドソーシングの実験参加者の多くがオーラルディアドコキネシスの趣旨を理解できなかったのではないかと推察される．区民センターでのデータ収集では，筆者らが口頭で実験参加者に実験の趣旨を説明したものの，クラウドソーシングではウェブ上の文章に趣旨を記載するにとどまっていたため，この記載をよく読まなかった実験参加者が多かったのではないかと考えられる．これらの参加者を構音障害者と定義するには問題がある．従って，本章以降ではクラウドソーシングの実験参加者は，オーラルディアドコキネシスの発音回数に関わらず構音障害に該当しないようにした．

5.6 変数増減法によるロジスティック分析 (音響)

前章と同様に，変数増減法によるロジスティック回帰分析を行った．ただし，有用だとして追加すると p 値が 1 または 0 に収束する特徴量があったため，全ての p 値が 0.1 未満となる特徴量の組み合わせのみを採用することにした．選択された特徴量を表 5.2，5.3に示す．選択された各特徴量について以下で説明を行う．

まず，各特徴量に記載されている on 及び off は，音声が無音状態から発生状態に移り変わる際 (onset) 及び発生中の声が無声化していく過渡期 (offset) における区間で計算された特徴量であることを示す．また，一部の特徴量の接頭辞に記載されている avg，std，skewness，kurtosis はそれぞれ，平均，標準偏差，歪度，尖度を計算した特徴量であることを示している．なお，各特徴量は，変数増減法の各施行回数で AIC が最小になるものから選ばれており，表 5.2， 5.3に示されている特徴量は，定数である const を除き，上位に存在する特徴量の方が一般的により有用である．そのため，下位に示されている特徴量は実際にはそれほど有用でない可能性があることに注意する必要がある．

BBE は，バーク尺度に基づく周波数帯域で計算されたエネルギー (Bark Band Energy)，定性的には声帯の動きの開始や終了を適切に行うことができるかどうかに関わる [34]．バーク尺度とは，聴覚に基づいて周波数帯域を 24 個に分割した尺度のことである [39]．例えば，BBEoff_13 という特徴量は，offset 時に計算された，バーク尺度における 13 番目の周波数帯域におけるエネルギーであることを示す．

MFCC はメル周波数ケプストラム係数 (Mel Frequency Cepstral Coefficient) であり，メル周波数に基づいて計算されたケプストラムを表す．メル周波数とは，人間の聴覚が低い音の変化に敏感で，高い音に対して鈍感であるという事実に基づいて変換された周波数で

表 5.2: 変数増減法により選択された構音障害を予測する音響特徴量．括弧内は音符の番号と歌詞を表す (以下同様)．

feature	coef	std err	$P > \|z\|$
const	-52.96	26.22	0.04
avg BBEoff_13_1_u	-32.68	15.96	0.04
skewness BBEoff_3_27_me	-30.22	14.99	0.04
skewness MFCCon_12_4_o	-16.88	9.05	0.06
std MFCCoff_4_35_ta	18.02	9.08	0.05
std MFCCoff_10_18_no	18.55	9.45	0.05
lastF0skw_10_ma	-12.51	6.50	0.05
kurtosis BBEoff_20_4_o	10.65	5.40	0.05
skewness MFCCoff_2_32_su	-4.43	2.65	0.09

ある．また，ケプストラムとは，音声のパワースペクトルの対数に逆フーリエ変換を施したものである．DMFCC は Mel Frequency Cepstral Coefficient (MFCC) の微分値を取ることにより計算され，音声周波数がどれくらい滑らかに変化するかを定量化した特徴量であるとされる [14]．DMFCC_on は onset 時の，DMFCC_off は offset 時の区間で計算された MFCC であることを示す．

lastEunvoiced は，無音区間の最終区間におけるエネルギーを表す [28]．BBE と同様に，声帯振動の開始及び停止をうまく行えているかどうかを特徴づける変数であり，構音障害の分類にも有用である [27]．

F0tiltavg は与えられた音声の基本周波数 (F0) の回帰直線の傾きを表す．

これらの音響特徴量を用いた際のロジスティック回帰による分類結果の混同行列を表 5.4に示す．前章で表 4.5に示した結果よりも大幅に良い結果が得られた．特に前章での分析手法で課題となっていた再現率に関しては，構音障害では 42% から 93%，嚥下障害では 13% から 92% へと上昇した．

ここまでの分析では，データを全て学習に用いて分類を行ったが，モデルの汎化性能を調べるためには交差検証を行うことが重要である．そこで，選択された有用な特徴量をもとに，5 分割交差検証を行なった際の混同行列を表 5.5に示す．なお，交差検証における訓練データとテストデータの分割の際にデータの偏りが生じないようにするために，層化抽出法による交差検証を行った．表 5.4と比較すると，分類性能が多少下がるものの，概ね正しく分類されていることが確認できる．

図 5.3，5.4に各特徴量の分布を示す．特徴量として選ばれた順に，上位 4 つのみを示している．

表 5.3: 変数増減法により選択された嚥下障害を予測する音響特徴量．括弧内は音符の番号と歌詞を表す (以下同様)．

feature	coef	std err	$P > \|z\|$
const	-34.23	15.86	0.03
stdlastEunvoiced_27_me	9.33	4.95	0.06
skewness BBEoff_4_18_no	24.78	11.43	0.03
skewness MFCCon_2_32_su	10.19	4.86	0.04
kurtosis DMFCCoff_4_5_i	21.14	10.65	0.05
skewness MFCCon_3_10_ma	-7.44	3.67	0.04
avg MFCCoff_5_27_me	11.15	5.30	0.04
std MFCCon_12_4_o	6.39	3.11	0.04
std MFCCon_3_14_tu	5.64	2.76	0.04

表 5.4: 歌唱中の音響特徴量による構音・嚥下障害予測時の混同行列

	構音障害	Predicted P	N
Actual	P	25	2
	N	1	155

	嚥下障害	Predicted P	N
Actual	P	22	2
	N	6	153

5.7 変数増減法によるロジスティック分析 (画像)

画像分析の手法は前章と同様である．前章との違いは，音節の分割をより正確に行なったこととリハビリテーション科に通院する患者のデータを追加したこと，及び構音障害の定義の変更である．

音響分析の時と同様，表 5.6， 5.7に示されている特徴量は，定数である const を除き，上位の存在されている特徴量の方が一般的により有用である．そのため，下位に示されている特徴量は実際にはそれほど有用でない可能性がある．

構音・嚥下障害の分類のために抽出された特徴量を表 5.6，5.7に示す．また，分類結果を表 5.8に示す．前章の表 4.8と比較して，分類性能が向上していることが確認できる．

音響分析の時と同様，選択された各特徴量について以下で説明を行う．なお，各特徴量に接頭辞として記載されている area，width，height はそれぞれ，口の大きさ (面積)，幅，高さを表す．

表 5.5: 5 分割交差検証による構音・嚥下障害予測時の混同行列 (音響)

	構音障害	Predicted P	N
Actual	P	24	3
	N	7	149

	嚥下障害	Predicted P	N
Actual	P	19	5
	N	6	153

表 5.6: 変数増減法により選択された構音障害を予測する画像特徴量
括弧内は音符の番号と歌詞を表す (以下同様).

feature	coef	std err	$P > \|z\|$
const	-112.56	60.85	0.06
area__mean__abs__change__16__shi	65.93	36.47	0.07
width__change__quantiles__"mean"__qh__0.6__ql__0.0__38__ru	48.86	26.66	0.07
width__change__quantiles__"mean"__qh__0.8__ql__0.6__8__no	57.99	31.26	0.06
width__fourier__entropy___bins__3__37__hu	-37.04	19.11	0.05
width__longest__strike__below__mean__33__re	-35.03	19.46	0.07
area__minimum__10__ma	29.21	16.06	0.07
height__has__duplicate__max__30__te	25.87	14.06	0.07
area__augmented__dickey__fuller__27__me	-25.15	13.10	0.06

5.7.1 構音障害を分類するの有用だとされた画像特徴量

まず，表 5.6に記載されている特徴量の説明から行う．

mean__abs__change は，時系列データの前後の差の絶対値に対して平均を取ったものである．

change__quantiles__"mean"__qh__0.6__ql__0.0 について説明する．まず，qh__0.6__ql__0.0 は，時系列データのうち，値が最小値から最大値の 60% のもののみが計算に使われることを意味している．この範囲に含まれるデータに対して，各データの変化の平均を取ったのがこの特徴量である change__quantiles__"mean"__qh__0.8__ql__0.6 も同様である．

fourier__entropy__bins__3 は，welch 法に基づいて計算されたパワースペクトル密度に対して，binned entropy を計算する [37, 6]．パワースペクトル密度は，角周波数成分についてどの程度の周期性が含まれているかを定量化したものである．一方で binned entropy では，データを任意の個数ごとに分割し，順番に 1ñ までラベル付けした後，さらにデータを昇順に並び替える．例えば，データを 3 個ずつに分割する今回の場合，(3,2,1) や (2,3,1) などが出現することになる．この出現パターンに対して計算された平均情報量 (エントロピー) が最終的な出力となる．permutation entorpy とも呼ばれるこの手法では，データの複雑さを定量化できるとされている [6].

longest__strike__below__mean は，データの連続値のうち，平均を下回るデータが連続で出現する回数の最大値を表す特徴量である．

minimum は，その言葉が示す通り，データの最小値を示す．

has__duplicate__max は最大値が観測された回数を示す．

augmented__dickey__fuller は，拡張ディッキー-フラー検定の結果を示す [30]．拡張ディッキー-フラー検定は，時系列データが単位根を持つかどうかの検定である．単位根を持つ場合，その確率過程は非定常である．

表 5.7: 変数増減法により選択された嚥下障害を予測する画像特徴量
括弧内は音符の番号と歌詞を表す (以下同様).

feature	coef	std err	$P > \|z\|$
const	-84.10	43.83	0.06
area_index_mass_quantile_q_0.6_5_i	46.73	25.74	0.07
area_permutation_entropy_dimension_3_tau_1_23_ha	81.15	43.50	0.06
height_change_quantiles_qh_0.6_ql_0.2_1_u	18.04	10.25	0.08
area_minimum_8_no	-48.65	25.77	0.06
width_mean_12_bu	25.39	13.11	0.05
width_symmetry_looking___r_0.05_2_sa	-13.15	7.33	0.07
area_linear_trend_attr_"stderr"_35_ta	3.95	2.23	0.08
height_change_quantiles_qh_1.0_ql_0.8_26_mo	3.41	1.97	0.08

表 5.8: 歌唱中の音響特徴量による構音・嚥下障害予測時の混同行列

	構音障害	Predicted P	N
Actual	P	154	1
	N	1	26

	嚥下障害	Predicted P	N
Actual	P	26	1
	N	1	154

5.7.2 嚥下障害を分類するの有用だとされた画像特徴量

次に，表 5.7の示される特徴量のうち，まだ説明されていない特徴量の説明を行う．

permutation_entropy___dimension_3_tau_1 は，先述した permutation entropy と根本的に等価な特徴量で，dimension (今回は 3) はデータの分割個数，tau (今回は 1) は何個ずつずらして分割していくかを示す [6].

mean は文字通り平均を表す．

symmetry_looking_r_0.05 は対象の時系列データが対称になっているかどうかを表す特徴量である．$(max(X) - min(X))/|mean(X) - median(X)|$ が基準値 r (今回は 0.05) を下回るかどうかを示す．

linear_trend___attr_"stderr" は，時系列データに対して計算された回帰直線とデータとの間の標準誤差を示す．

表 5.9: 5 分割交差検証による構音・嚥下障害予測時の混同行列 (画像)

	構音障害	Predicted P	N
Actual	P	22	5
	N	6	149

	嚥下障害	Predicted P	N
Actual	P	18	5
	N	3	156

5.8 考察

本章では，4章の分析結果をもとにした課題点に対処するため，word alignment による音声のより正確な分割，及び音響特徴量の追加，嚥下障害を持つ患者のデータの追加，構音障害の定義の再考を行い，再び変数増減法によるロジスティック回帰分析を行った．その結果，音響・画像分析ともに，4章よりも，特に再現率において大幅に高い分類性能を実現することができた．

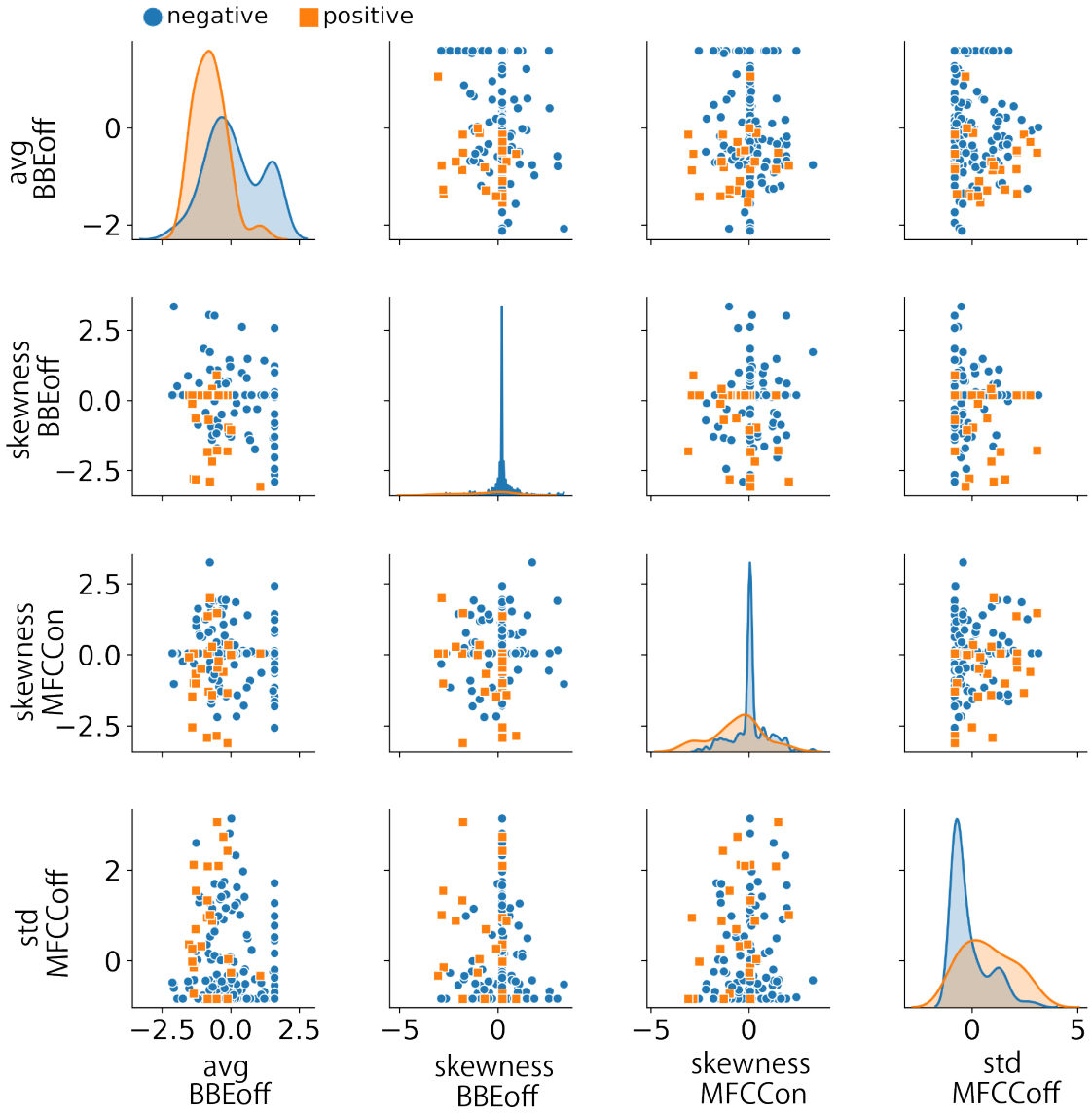

図 5.3: 構音障害の分類において有用な音響特徴量の分布及び散布図 (上位 4 つのみ)．変数名は簡略化されているので，詳しくは表 5.2を参照されたい．

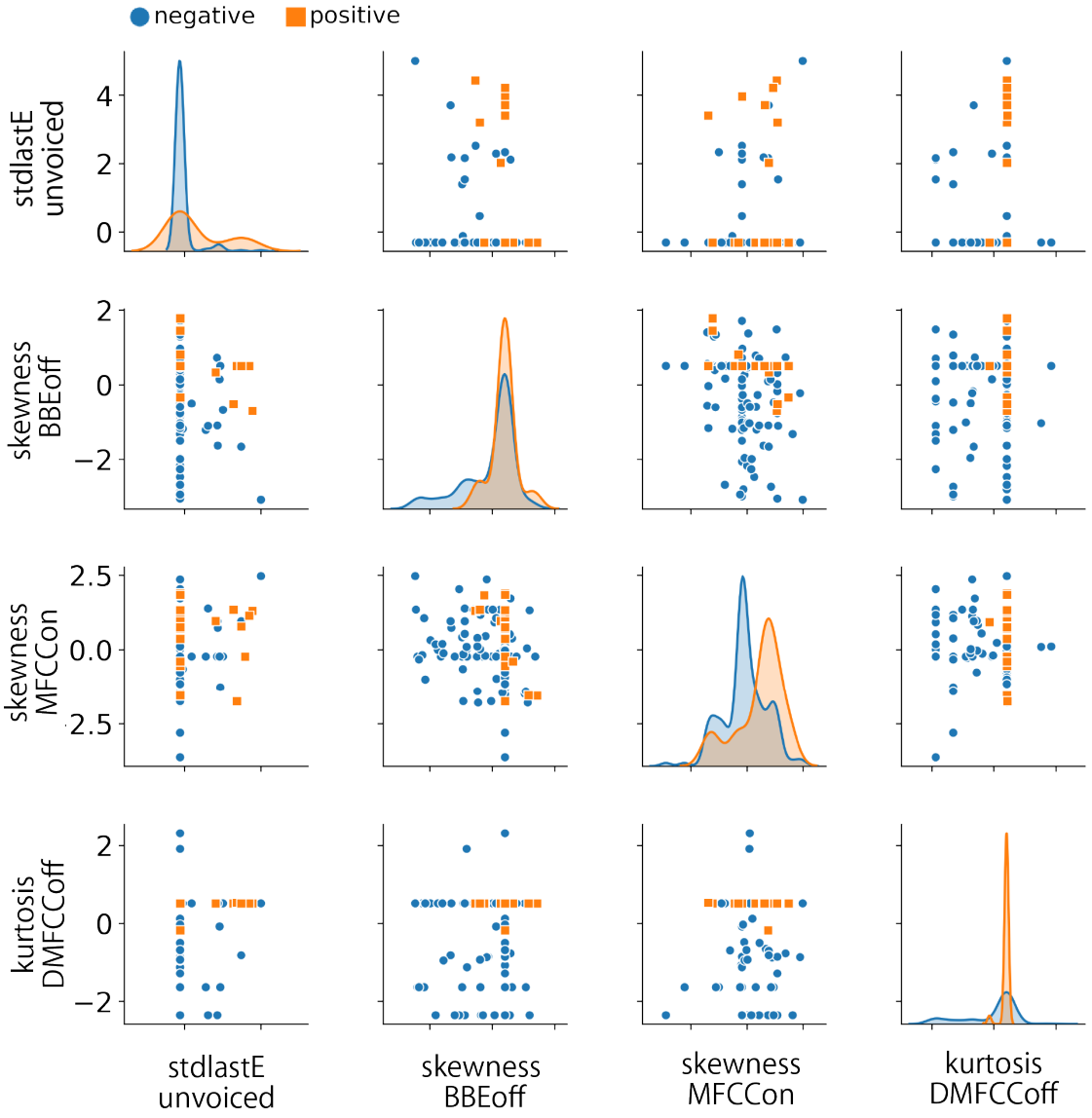

図 5.4: 嚥下障害の分類において有用な音響特徴量の分布及び散布図 (上位 4 つのみ)．変数名は簡略化されているので，詳しくは表 5.3を参照されたい．

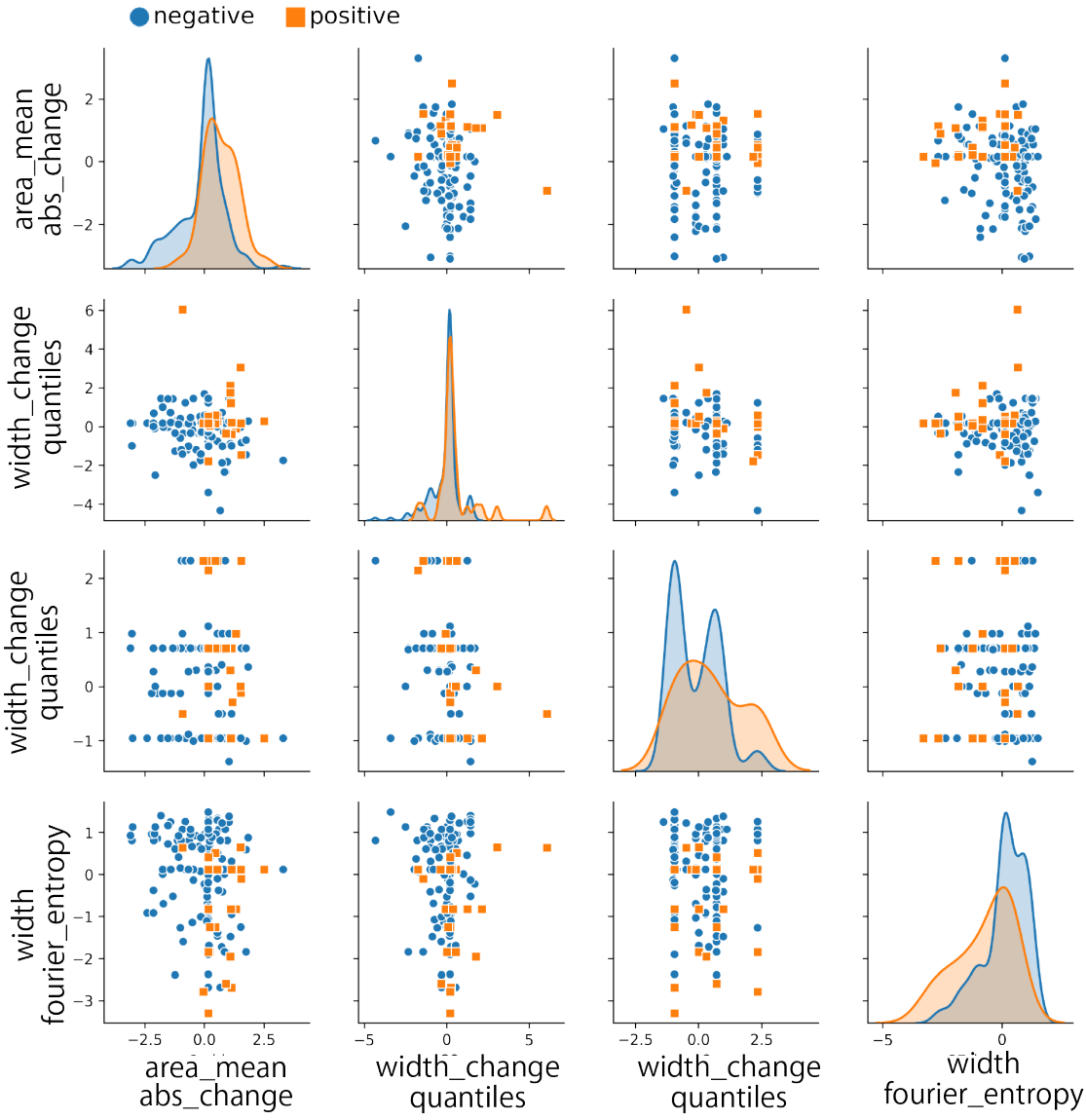

図 5.5: 構音障害の分類において有用な画像特徴量の分布及び散布図 (上位 4 つのみ)．変数名は簡略化されているので，詳しくは表 5.6を参照されたい．

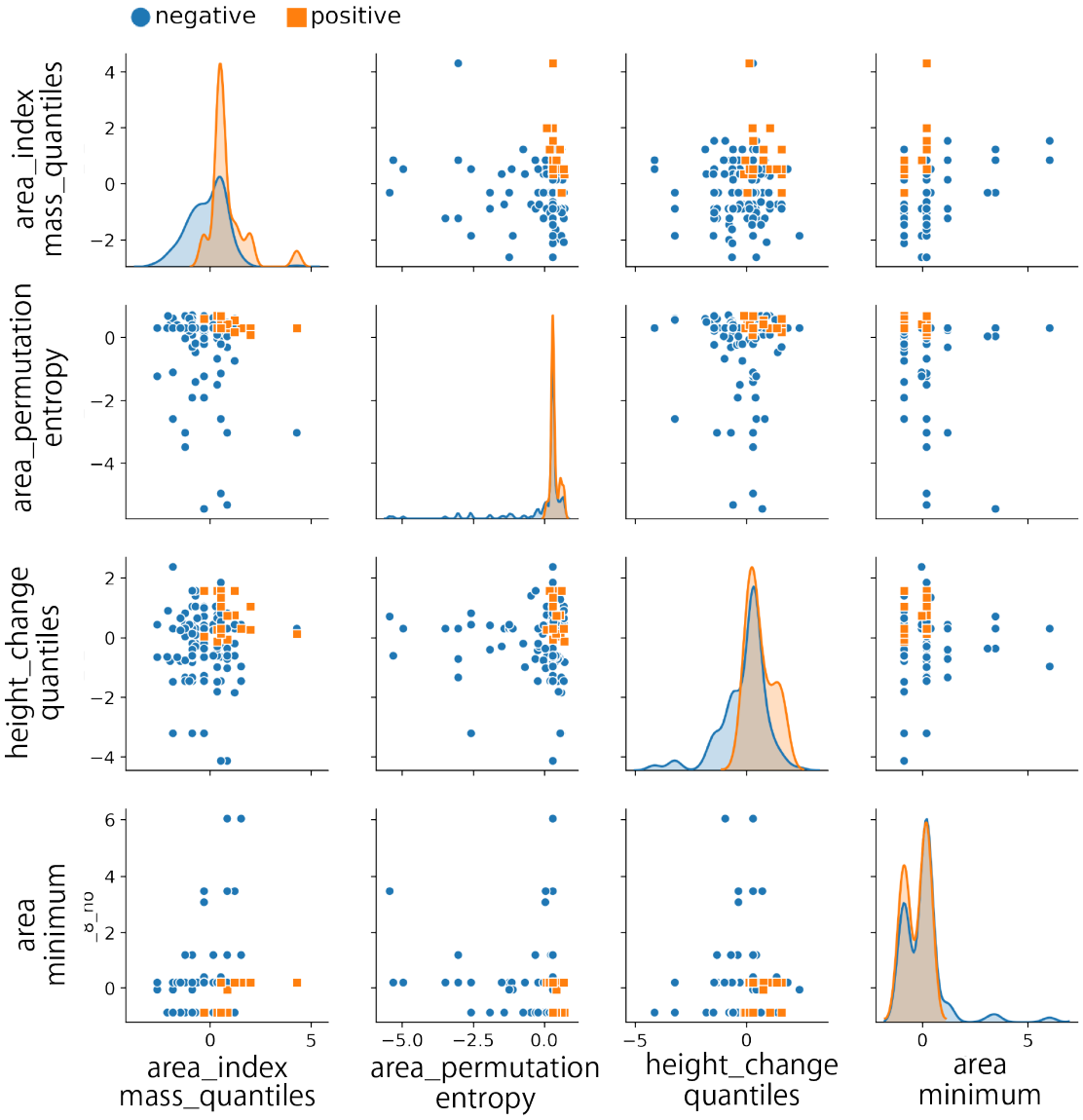

図 5.6: 嚥下障害の分類において有用な画像特徴量の分布及び散布図 (上位 4 つのみ)．変数名は簡略化されているので，詳しくは表 5.7を参照されたい．

Chapter 6

LightGBM を用いた分析

4，5章では，歌唱から構音・嚥下障害を分類するための初期検討として，変数増減法によるロジスティック回帰分析を行った．ロジスティック回帰は単純であるというメリットを有する一方で，一般線形モデルの一種であり複雑なモデルを表現をしづらいという欠点がある．そこで本章では，分類性能の向上を目指して近年提案されている LightGBM を用いた分類を行い，その性能を評価する．LightGBM では高速に学習を行い，高い分類性能を実現できる他，図 6.1のように決定木を作成するという点で結果の解釈性が高いため，これを採用した．解釈性が高い手法を採用することによって，医学的な専門知識を有する研究者からのフィードバックを得やすくなると考えている．

6.1 LightGBM

LightGBM (Light Gradient Boosting Machine) は，機械学習アルゴリズムの 1 つである勾配ブースティング決定木 (Gradient Boosting Decision Tree，GBDT) に基づいて高速に最適化を行う機械学習アルゴリズムである [17]．以下でその手法について詳述する．

まず，LightGBM が大きく依存している GBDT について説明する．GBDT は勾配降下法 (Gradient), アンサンブル学習の 1 つであるブースティング (Boosting), 決定木 (Decision Tree) の 3 つを組み合わせた手法である [11]．アンサンブル学習は，それ自体では性能の低い弱学習器を組み合わせることで性能の高い学習器を実現する．GBDT では決定木が弱学習器として用いられる．この弱学習器を直列に繋げて精度を高めるのがブースティングである．勾配降下法に基づいて損失関数の誤差を小さくすることにより，精度を高めていく．最終的な結果を，例えば後述の図 6.1のように決定木として表現することにより，解釈性の高い分類モデルを構築することができる．

GBDT は高い精度を実現したものの，データの数や特徴量が増加した際の効率性には不満があった．そこで，Gradient-based One-Side Sampling (GOSS) 及び Exclusive Feature Bundling (EFB) によってこの課題を解決を図ったのが LightGBM である [17].

ロジスティック回帰分析と比較して，LightGBM を用いた分類手法には幾つかのメリットが存在する．まず，ロジスティック回帰ではデータの特徴量が正規分布をしており，尚且つ超平面によって線形に分類できることを前提としているが，LightGBM にはそのような制約は存在せず，より複雑なモデルにも対応可能である．また，4，5では，非数や無限大が含まれる特徴量は計算に使用しなかったが，実際には特徴量が非数や無限大になっていること自体に，分類に有用な何らかの意味が含まれている可能性がある．LightGBM では，これらの値を削除することなくそのまま使用して最適化を実行できる．

6.2 LightGBM による分類方法

まず，前章と同様に音響・画像特徴量をそれぞれ抽出し，各特徴量に対して標準化を行った．その後，データ全体の 80% を訓練データに，20% を評価データに分割した．そして，構音障害・嚥下障害ごとに LightGBM による学習を行った．

なお，LightGBM ではハイパーパラメータの調整が重要である．次節でその詳細を説明する．

6.3 LightGBM のパラメータの決定

表 6.1: パラメータチューニングを行なった LightGBM のハイパーパラメータ

hyperparameter	description	range
learning_rate	学習率	0.001 ～ 0.5
num_leaves	決定木に含まれる分岐の数	1 ～ 128
min_data_in_leaf	葉に含まれるデータの最小数	1 ～ 32
reg_alpha	L1 正則化項の係数	0.0001 ～ 0.5
reg_lambda	L1 正則化項の係数	0.0001 ～ 0.5

表 6.2: optuna による最適化の結果得られたハイパーパラメータ

	learning_rate	num_leaves	min_data_in_leaf	reg_alpha	reg_lambda
構音 (音響)	0.31	103	6	0.087	0.027
嚥下 (音響)	0.38	76	27	0.032	0.038
構音 (画像)	0.31	39	19	0.043	0.014
嚥下 (画像)	0.14	96	8	0.038	0.037

LightGBM には，決定木の深さや葉の数など幾つかのハイパーパラメータが存在し，これらのパラメータチューニングを行うことが高性能な分類を実現する上で重要となる．今回は，表 6.1のパラメータのチューニングをハイパーパラメータ最適化フレームワークである optuna を用いて行う [2]．取りうるパラメータの全ての組み合わせを試行して最適なパラメータを決定するグリッドサーチと比較して，optuna はベイズ最適化を用いることで，より少ない試行回数でハイパーパラメータを決定することができる．最適化を行うハイパーパラメータを表 6.1に示す．その他のパラメータとしてにはデフォルトのものを採用した [1]．

optuna によるハイパーパラメータ探索の試行回数は 100 回とし，目的関数は LightGBM による出力結果の交差エントロピーとした．得られたパラメータを表 6.2に示す．

6.4 音響分析の結果

図 6.1，6.2に作成された決定木を示す．

5章と同様に，一部の特徴量の接頭辞に記載されている avg，std，skewness，kurtosis はそれぞれ，平均，標準偏差，歪度，尖度を計算した特徴量であることを示している．

まず各音響特徴量の概要について説明を行う．まず，図 6.1に示された構音障害の分類のための音響特徴量から説明を行う．

mean_bottleneck_149 は，先行研究において，畳み込み及び回帰型ニューラルネットワークのボトルネック層から抽出された 149 番目の特徴量を平均化した変数であることを示す [33]．

meanF0 は，基本周波数 F_0 の平均値を表す．

skewness_BBEon は，5.6節に示されているものと同様であるため，説明を省略する．

mean_error は，基本周波数の回帰直線と実際の基本周波数の誤差の平均を取ったものである．

DF0 は，基本周波数に対して 1 階微分を適用した特徴量である．基本周波数の変動は構音障害の分類のための音響特徴量として用いられるほか，加齢に伴い増加すると報告されている [5]．

次に，図 6.2に示された嚥下障害の分類のための音響特徴量の説明を行う．

NAQ (Normalized Amplitude Quotient) は，声門の閉鎖を定量化した特徴量であり，ノイズにロバストな特徴を有する [3]．

MFCCon は，5.6章で説明したものと同様のため，ここでの説明は省略する．

[1] https://lightgbm.readthedocs.io/en/latest/Parameters.html

LightGBM による分類の結果を表 6.3に示す．構音・嚥下障害の場合でともに，概ね高い性能で分類可能であることが読み取れる．

表 6.3: LightGBM による構音・嚥下障害予測時の混同行列 (音響)

	構音障害	Predicted P	N
Actual	P	23	4
	N	1	155

	嚥下障害	Predicted P	N
Actual	P	19	5
	N	0	159

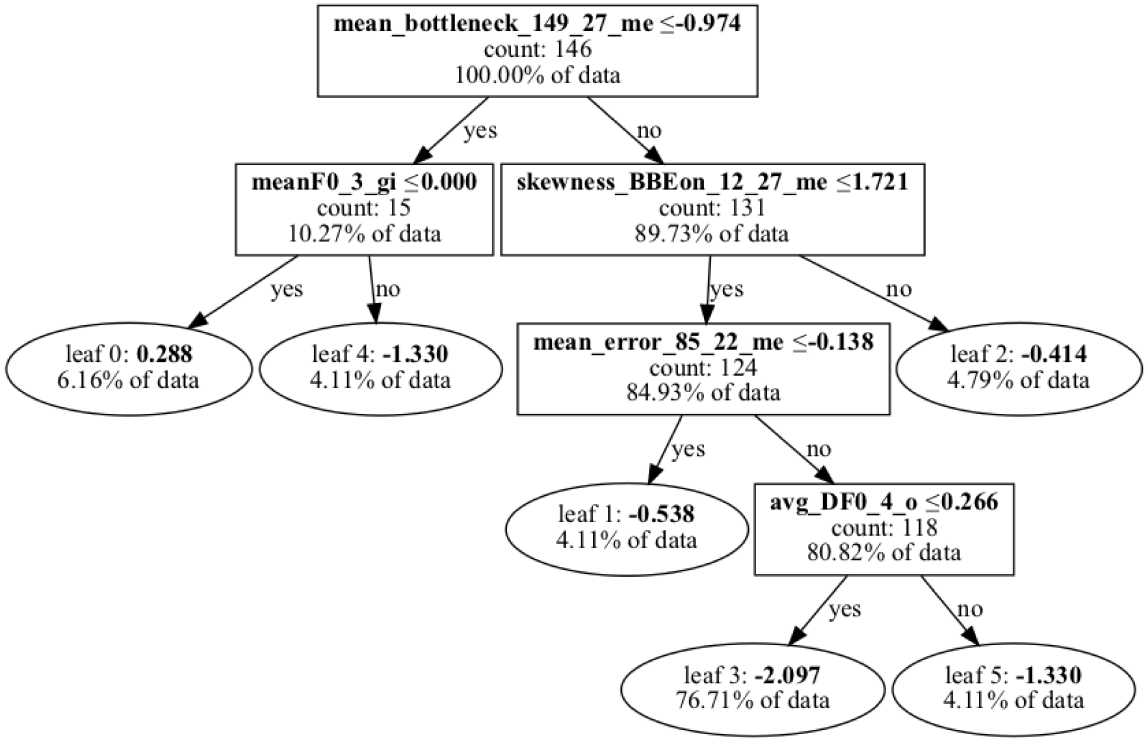

図 6.1: LightGBM によって作成された構音障害分類のための決定木 (音響).

6.5 画像分析の結果

同様に画像特徴量の説明を行う．

approximate_entropy は，配列の規則性を定量化した特徴量である [29]．計算にあたっては，まず整数 m と正の実数 r を定数として定める．approximate_entropy_m_2_r_01 と記載されている今回の場合は，m は 2，r は 0.1 で固定されている．m は計算において何個ずつ数列をとって比較していくかを表し，r はそれらによる配列の最大値を比較するときの基準値を表す．

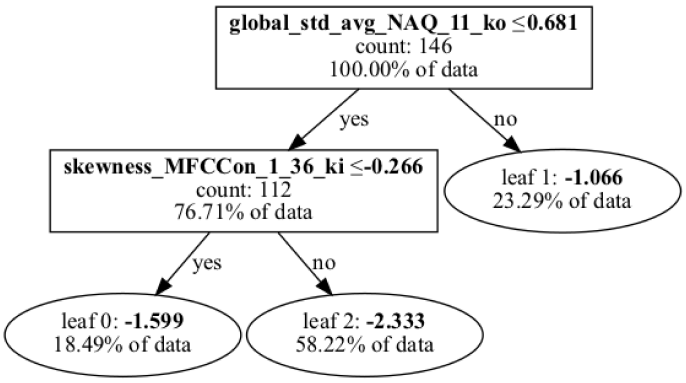

図 6.2: LightGBM によって作成された嚥下障害分類のための決定木 (音響).

fft_coefficient___attr_angle は，高速離散フーリエ変換をした際の位相を表す．

linear_trend は，5.7.1章で示したものと同様なので，説明を省略する．

続いて図 6.4に示される嚥下障害の方の特徴量について説明を行う．

cwt_coefficients は，連続ウェーブレット変換 (continuous wavelet transform，cwt) によって得られる特徴量であり，以下の式によって計算される．

$$\frac{2}{\sqrt{3a}\pi^{\frac{1}{4}}}(1-\frac{x^2}{a^2})exp(-\frac{x^2}{2a^2})$$

ar_coefficient は，自己回帰 (autoregressive model, AR) モデルの無条件最尤推定に関わる特徴量で，時系列データに対して以下の式をフィッティングすることで得られる．

$$X_t = \varphi_0 + \sum_{i=1}^{k} \varphi_i X_{t-i} + \varepsilon_t$$

ここでは，固定したパラメータ k に応じて，φ_0 から φ_k までの値が決定される．今回の場合，使用されている特徴量は ar_coefficient___coeff_9___k_10 であり，これは，$k = 10$ の時の φ_9 が特徴量となっていることを示す．

fourier_entropy_bins_10 は5.7.1節で説明したものと同様であるので，ここでは説明を省略する．

表 6.4: LightGBM による構音・嚥下障害予測時の混同行列 (画像)

	構音障害	Predicted P	N
Actual	P	22	5
	N	1	154

	嚥下障害	Predicted P	N
Actual	P	18	5
	N	0	159

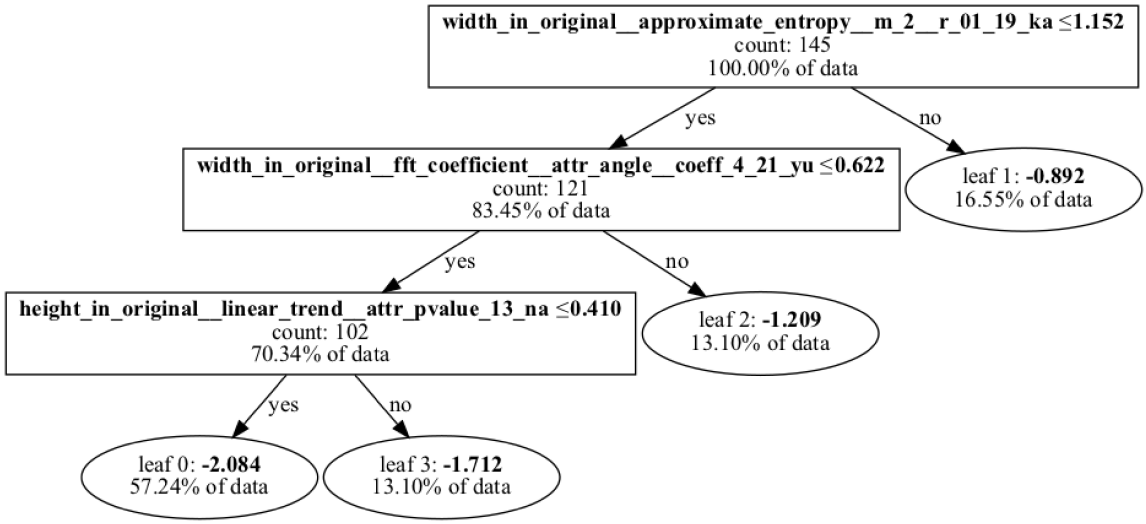

図 6.3: LightGBM によって作成された構音障害分類のための決定木 (画像).

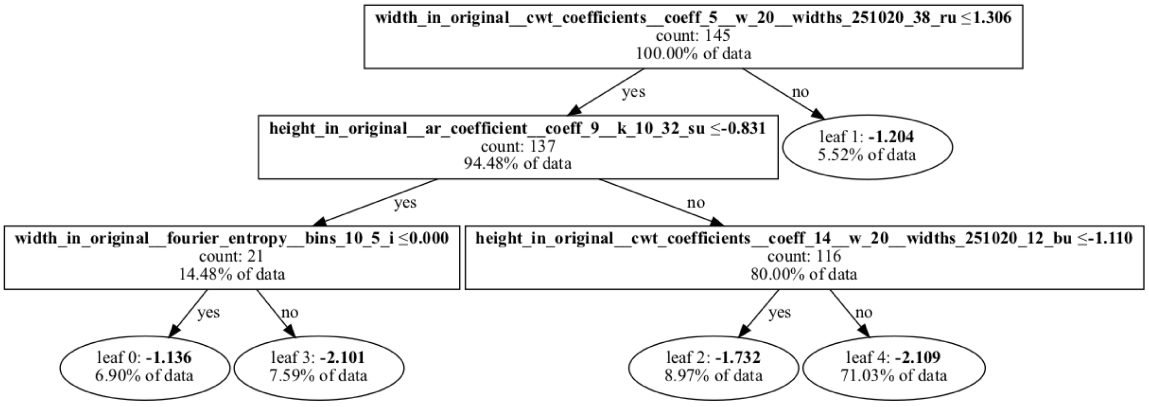

図 6.4: LightGBM によって作成された嚥下障害分類のための決定木 (画像).

6.6 考察

音響分析では，構音障害の分類では正解率 97%，再現率 85%，嚥下障害の分類では正解率 98%，再現率 70% という結果となった．また，画像分析では，構音障害の分類では正解率 97%，再現率 81%，嚥下障害の分類では正解率 97%，再現率 78% となった．表 6.4，6.3に示された LightGBM による分類と，表 5.5， 5.9に示された 5 分割交差検証時の変数増減法によるロジスティック回帰分析とを比較すると，分類性能の向上はそれほど見られなかった．本研究のおいては，ロジスティック回帰でも十分に分類可能であることを示している．一方で，LightGBM を用いた決定木の作成を行うことで，より解釈性の高い結果を得ることができた．すなわち，どのような特徴量から先に注目するべきか等の情報が，作成された決定木から知ることができるという点でロジスティック回帰分析には無いメリットも存在する．

Chapter 7

おわりに

本研究では，歌唱を通じた口腔機能の評価手法の実現に向けて幅広い年代層の口腔機能および歌唱データを収集し，それらに対して音響・画像分析を行った．最後に，本研究とその課題点を簡単にまとめた後，今後の展望を述べる．

7.1 本研究のまとめ

近年，口腔機能の低下が将来の健康リスクを増加させることが示され，口腔機能の評価の重要性はますます高まっている．しかしながら，既存の口腔機能の評価手法は，手法自体が単調であることや機器の測定に身体的な負担があることなどが原因で，利用者が自発的・定期的に取り組みやすいとは言い難い．そこで，歌唱が利用者の心理的な負担を軽減することを期待して，図 7.1のような歌唱による口腔機能の評価手法を提案した．

提案手法の実現に向けて，我々はクラウドソーシングで非高齢者を，東京都文京区区民センターで高齢者を中心とする実験参加者の歌唱と口腔機能のデータを収集した．そして，歌唱中の音声・画像から構音・嚥下機能を分類できるかどうかを確かめるため，変数増減法によるロジスティック分析を行い，その分類性能を評価した．その結果，ある程度の分類性能を確認することができたものの，再現率が 10% から 40% 程度と低いことが確認された．

そこで，それまでの分析方法の改善策を検討し，さらに東京医科歯科大学で口腔機能に何かしらの障害を持つ方のデータを追加して，改めて変数増減法によるロジスティック回帰分析を行った結果，改善前よりも遥かに良い分類性能が得られた．

また，より複雑なモデルに対応すると同時に決定木による解釈性の高い分類を行うことを目指して LightGBM による分類を行った．その結果，分類性能が同様に高いことが確認された．

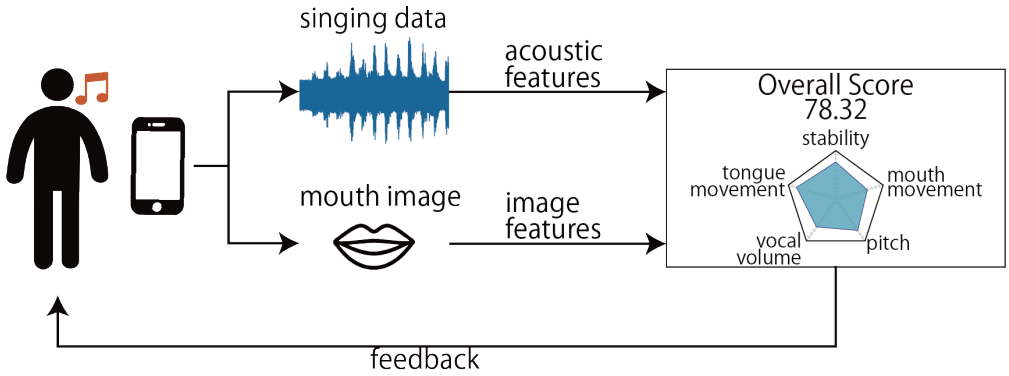

図 7.1: 提案するシステムのコンセプト図 (再掲)．ユーザーはスマートフォンから流れる音源に従って歌唱を行い，その様子の録画を行う．歌唱が終了すると，録画の結果から音響・画像特徴量が抽出され，採点結果を表示する．

7.2 本研究の課題

7.2.1 データ数の少なさ

本研究では，非高齢者 99 件，高齢者 75 件，嚥下機能に障害を持ち通院を行っている患者 9 件の計 183 件のデータを用いた分析を行った．しかしながら，これらのデータ件数は十分とは言えず，特に口腔機能に障害を持つ方やオーラルフレイルであると医師に診断された方のデータの収集が今後の分析の上で重要であると考えられる．

7.2.2 録音環境の違い

今回収集したデータのうち，区民センター及び東京医科歯科大学でのデータ収集では，スマートフォンを含め全て同じ機器を使用した．また，区民センターでのデータ収集は全て同一の部屋で行った．一方で，クラウドソーシングのデータ収集はオンライン上で実施したため，使用端末や録音環境は実験参加者によってそれぞれ異なり，録音環境の違いが分類性能に影響を与えた可能性がある．一方で，スマートフォンなどのモバイル端末の利用を想定した提案手法では，録音環境は利用者ごとに異なると考えられるため，そうした違いに堅牢な分析手法が必要である．

7.3 考察

7.3.1 歌の上手さとの関連

歌をうまく歌おうと試みた結果，ユーザーは通常とは異なる発声を行う可能性がある．その際，歌の上手さなどが分類性能に影響を与える可能性がある．今後の研究では，歌の

上手さなどの指標も併せて分析を行うことで，これらの影響を別々に考える必要があると考える．

7.4 今後の展望

7.4.1 口の切り出しの完全な自動化

本研究では，初期検討として歌唱中のデータから構音・嚥下機能の分類を行うため，画像中の口の範囲の切り出しは一部自動でおこなっていた．これは，haar cascade による口の範囲の切り出しの精度が低かったためである．口の範囲の切り出しを行う既存研究はこれまでいくつか存在していたが，それはは顔全体が画像中に含まれていることを前提としており，プライバシーを考慮して鼻より下の範囲のみを撮影した我々のデータに適用することができなかった．完全な自動化に向けては，この課題を解決し，口の範囲の切り出しを自動で行う手法を検討する必要がある．ただし実際の利用においては，顔全体を撮影することが許容される可能性があることも踏まえて検討を行う必要があるだろう．

7.4.2 最終的なカラオケアプリの開発

図 7.2: UltraStarPlay の歌唱中の動画．

提案手法の最終的な実現に向けて，カラオケアプリの作成も必要である．現在，オープンソースのカラオケアプリである Ultra Star Play [1]を用いて提案手法の実現を行うことを検討している．図 7.2に Ultra Star Play でカラオケの歌唱中の画像を示す．このアプリでは，歌唱中の音程の正確さなどをもとに採点を行う機能が備わっている．この採点機能に，我々が提案する採点手法を組み込むことによって，目指すカラオケアプリの作成ができるのではないかと考えている．

7.4.3 使用可能な曲の拡大

今回は歌唱からの構音・嚥下機能の分類を単純にするために，曲の歌詞を「ふるさと」のみに限定して分析を行った．しかしながら，カラオケに含まれるゲーム性に着目して利用者の心理的負担を軽減するという本研究の観点を考慮すると，さまざまな曲を通じて

7.4.4 アプリの有用性の評価

前節のアプリの完成後，オーラルフレイルの状態にあるような高齢者の方々に対して実際にアプリを使用してもらい，提案手法の有効性について調査を行う必要があると考える．特に，我々が着目していた自発的・定期的な利用の促進を評価するため，単純な既存の評価手法と比べた利用回数などを測定したいと考えている．

[1]https://github.com/UltraStar-Deluxe/Play

Publications

国内研究会

- 平井雄太，耿世嫻，下島銀士，小野寺宏，矢谷浩司，歌による嚥下・構音機能の定量的評価手法の実現に向けた歌唱データの音響・画像分析，情報処理学会　第 72 回ユビキタスコンピューティングシステム研究発表会.（2021 年 12 月）

References

[1] Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike, pages 199–213. Springer.

[2] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pages 2623–2631.

[3] Alku, P., Bäckström, T., and Vilkman, E. (2002). Normalized amplitude quotient for parametrization of the glottal flow. the Journal of the Acoustical Society of America, 112(2):701–710.

[4] Ando, T., Masaki, A., Liu, Q., Ooka, T., Sakurai, S., Hirota, K., and Nojima, T. (2018). Squachu: a training game to improve oral function via a non-contact tongue-mouth-motion detection system. In Proceedings of the 2018 International Conference on Advanced Visual Interfaces, pages 1–8.

[5] Arias-Vergara, T., Vásquez-Correa, J. C., and Orozco-Arroyave, J. R. (2017). Parkinson' s disease and aging: analysis of their effect in phonation and articulation of speech. Cognitive Computation, 9(6):731–748.

[6] Bandt, C. and Pompe, B. (2002). Permutation entropy: a natural complexity measure for time series. Physical review letters, 88(17):174102.

[7] Belafsky, P. C., Mouadeb, D. A., Rees, C. J., Pryor, J. C., Postma, G. N., Allen, J., and Leonard, R. J. (2008a). Validity and reliability of the eating assessment tool (eat-10). Annals of Otology, Rhinology & Laryngology, 117(12):919–924.

[8] Belafsky, P. C., Mouadeb, D. A., Rees, C. J., Pryor, J. C., Postma, G. N., Allen, J., and Leonard, R. J. (2008b). Validity and reliability of the eating assessment tool (eat-10). Annals of Otology, Rhinology & Laryngology, 117(12):919–924.

[9] Dejonckere, P., Remacle, M., Fresnel-Elbaz, E., Woisard, V., Crevier, L., and Millet, B. (1998). Reliability and clinical relevance of perceptual evaluation of pathological voices. Revue de laryngologie-otologie-rhinologie, 119(4):247–248.

[10] Fried, L. P., Tangen, C. M., Walston, J., Newman, A. B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W. J., Burke, G., et al. (2001). Frailty in older adults: evidence for a phenotype. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 56(3):M146–M157.

[11] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232.

[12] Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. The Annals of Statistics, 42(6):2243–2281.

[13] Furlan, R. M. M. M., Santana, G. A., Bischof, W. F., Motta, A. R., and de Las Casas, E. B. (2019). A new method for tongue rehabilitation with computer games: Pilot study. Journal of oral rehabilitation, 46(6):518–525.

[14] Godino-Llorente, J. I., Gomez-Vilda, P., and Blanco-Velasco, M. (2006). Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. IEEE transactions on biomedical engineering, 53(10):1943–1953.

[15] Godino-Llorente, J. I. and Gómez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. IEEE Transactions on Biomedical Engineering, 51(2):380–384.

[16] Hu, K., Ivanov, P. C., Chen, Z., Carpena, P., and Stanley, H. E. (2001). Effect of trends on detrended fluctuation analysis. Physical Review E, 64(1):011114.

[17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30:3146–3154.

[18] Kothari, M., Svensson, P., Jensen, J., Holm, T. D., Nielsen, M. S., Mosegaard, T., Nielsen, J. F., Ghovanloo, M., and Baad-Hansen, L. (2014). Tongue-controlled computer game: a new approach for rehabilitation of tongue motor function. Archives of physical medicine and rehabilitation, 95(3):524–530.

[19] Lee, A. and Kawahara, T. (2009). Recent development of open-source speech recognition engine julius. In Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, pages 131–137. Asia-Pacific Signal and Information Processing Association, 2009 Annual ⋯.

[20] Lee, C.-H., Liu, Z., Wu, L., and Luo, P. (2020). Maskgan: Towards diverse and interactive facial image manipulation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[21] Lempel, A. and Ziv, J. (1976). On the complexity of finite sequences. IEEE Transactions on information theory, 22(1):75–81.

[22] Little, M., McSharry, P., Roberts, S., Costello, D., and Moroz, I. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Nature Precedings, pages 1–1.

[23] Miyazaki, A. and Mori, H. (2020). Frequent karaoke training improves frontal executive cognitive skills, tongue pressure, and respiratory function in elderly people: pilot study from a randomized controlled trial. International journal of environmental research and public health, 17(4):1459.

[24] Molfenter, S. M. and Steele, C. M. (2013). Variation in temporal measures of swallowing: sex and volume effects. Dysphagia, 28(2):226–233.

[25] Nishida, T., Yamabe, K., Ide, Y., and Honda, S. (2020). Utility of the eating assessment tool-10 (eat-10) in evaluating self-reported dysphagia associated with oral frailty in japanese community-dwelling older people. The journal of nutrition, health & aging, 24(1):3–8.

[26] Novotnỳ, M., Rusz, J., Čmejla, R., and Růžička, E. (2014). Automatic evaluation of articulatory disorders in parkinson' s disease. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(9):1366–1378.

[27] Orozco-Arroyave, J. R. (2016). Analysis of speech of people with Parkinson's disease, volume 41. Logos Verlag Berlin GmbH.

[28] Orozco-Arroyave, J. R., Belalcazar-Bolanos, E. A., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Skodda, S., Rusz, J., Daqrouq, K., Hönig, F., and Nöth, E. (2015). Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases. IEEE journal of biomedical and health informatics, 19(6):1820–1828.

[29] Richman, J. S. and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. American Journal of Physiology-Heart and Circulatory Physiology, 278(6):H2039–H2049.

[30] Said, S. E. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika, 71(3):599–607.

[31] Santamato, A., Panza, F., Solfrizzi, V., Russo, A., Frisardi, V., Megna, M., Ranieri, M., and Fiore, P. (2009). Acoustic analysis of swallowing sounds: a new technique for assessing dysphagia. Journal of rehabilitation medicine, 41(8):639–645.

[32] Tanaka, T., Takahashi, K., Hirano, H., Kikutani, T., Watanabe, Y., Ohara, Y., Furuya, H., Tetsuo, T., Akishita, M., and Iijima, K. (2018). Oral frailty as a risk factor for physical frailty and mortality in community-dwelling elderly. The Journals of Gerontology: Series A, 73(12):1661–1667.

[33] Vasquez-Correa, J. C., Arias-Vergara, T., Schuster, M., Orozco-Arroyave, J. R., and Nöth, E. (2020). Parallel representation learning for the classification of pathological speech: Studies on parkinson' s disease and cleft lip and palate. Speech Communication, 122:56–67.

[34] Vásquez-Correa, J. C., Orozco-Arroyave, J., Bocklet, T., and Nöth, E. (2018). Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease. Journal of communication disorders, 76:21–36.

[35] Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, volume 1, pages I–I. Ieee.

[36] Waito, A., Bailey, G. L., Molfenter, S. M., Zoratto, D. C., and Steele, C. M. (2011). Voice-quality abnormalities as a sign of dysphagia: validation against acoustic and videofluoroscopic data. Dysphagia, 26(2):125–134.

[37] Welch, P. (1967). The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Transactions on audio and electroacoustics, 15(2):70–73.

[38] Xue, Q.-L., Bandeen-Roche, K., Varadhan, R., Zhou, J., and Fried, L. P. (2008). Initial manifestations of frailty criteria and the development of frailty phenotype in the women's health and aging study ii. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 63(9):984–990.

[39] Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). The Journal of the Acoustical Society of America, 33(2):248–248.

[40] 伊藤加代子 (2009). 新しい口腔機能測定器を用いたオーラルディアドコキネシスの測定. 新潟歯学会雑誌, 39(1):61–63.

[41] 厚生労働省 (2020). 令和 2 年簡易生命表.

[42] 小口和代, 才藤栄一, 水野雅康, 馬場尊, 奥井美枝, and 鈴木美保 (2000). 機能的えん下障害スクリーニングテスト「反復唾液えん下テスト」(the repetitive saliva swallowing test: Rsst) の検討 (1) 正常値の検討. リハビリテーション医学, 37(6):375–382.

[43] 日本歯科医師会, . (2019). 歯科診療所におけるオーラルフレイル対応マニュアル 2019 年版.

[44] 杉本智子, 葭原明弘, 伊藤加代子, and 宮﨑秀夫 (2012). オーラルディアドコキネシスを用いた構音機能の評価と発声発語器官障害との関連. 口腔衛生学会雑誌, 62(5):445–453.

[45] 橋本修二, 川戸美由紀, and 尾島俊之 (2013). 健康寿命における将来予測と生活習慣病対策の費用対効果に関する研究.

[46] 西尾正輝 and 新美成二 (2002). Dysarthria における音節の交互反復運動. 音声言語医学, 43(1):9–20.

[47] 飯島勝矢 (2015). 食 (栄養) および口腔機能に着目した加齢症候群の概念の確立と介護予防 (虚弱化予防) から要介護状態に至る口腔機能支援等の包括的対策の構築および検証を目的とした調査研究. 事業実施報告書. 平成 26 年度厚生労働省老人保健事業推進等補助金. 老人保健健康増進事業.