議事録自動作成システムのための
画像および音声情報を用いた
発話者判別手法に関する研究

2022

中　村　　悦　郎

内容梗概

近年，働き方改革の実現に伴い，多くの企業で会議の効率化が注目されている．例えば，会議において議事録が取得され，利活用されている．議事録は議論した内容や取り決めの共有に有用であり，かつその後の会議の質を向上させるために役立てることができる．一方，音声認識技術に基づいた議事録自動作成システムを導入することで議事録作成におけるヒューマンエラーの低減や，工数の削減を行い，業務を効率化することが可能である．特に，議事録自動作成システムにおいて，発言ごとに発話者を自動判別する技術は，議事録自動作成手法における重要な要素技術である．議事録自動作成システムによって作成されたテキストデータに対して，発言者を自動で割り振ることは，議事録作成後に話者を割り振るための作業の簡略化を可能にするため，業務の効率化に寄与すると考える．既存の発話者判別手法は，会議参加者の人数と同じ数のマイクが必要な点や，事前の声紋登録が必要な点が課題である．これらの課題は，画像における口唇の動きと音声情報の類似性を評価し，発話者を判別することで解決可能であると考える．画像情報と音声情報を用いた発話者判別手法を構築するためには，3 つの要素技術が必要である．具体的には，①口唇形状自動抽出手法，②画像情報と音声情報を用いた発話区間抽出手法，ならびに③画像情報と音声情報を用いた発話者判別手法である．これらの要素技術について，それぞれ研究が進められているが，①肌の色情報の個人差や顔の状態に起因して口唇抽出精度が低下する点，②発話者判別を目的とした発話区間抽出法は検討されていない点，③学習のために多くのデータを必要とする点などの課題を有している．本論文では，議事録自動作成システムにおける利便性の高い発話者判別手法の構築のために上記課題の解決を行い，工学上の進歩に寄与することを目的とする．すなわち，人物が発話している動画データを対象とし，①口唇形状自動抽出法に関する検討，②発話区間の抽出手法に関する検討，ならびに③発話者の判別手法について検討し，工学上の進歩に寄与することを目的とする．本論文は全 5 章により構成されている．

第 1 章を緒論とし，本論文の主題である会議の効率化の必要性，実用化されている議事録自動作成システム，発話者判別手法の関連研究，ならびに本研究の目的および本研究に対する筆者の立場を述べるとともに，本論文の内容について述べている．

第 2 章では，口唇形状自動抽出法について検討を加えている．具体的には，口唇と肌の赤味に着目して口唇領域と肌領域をクラス分類する手法を提案した．5 名の被験者の動画データを使用して評価を行ったところ，提案手法は肌の色の個人差や陰影の影響による精度低下を低減可能であること，および顔の他の部位が

遮蔽されている場合においても口唇形状を抽出可能であることを明らかにした.

第 3 章では，発話区間の抽出手法について検討を加えた．具体的には，発話に伴う口唇の動きと音声が同時に発生したフレームを発話区間として抽出する手法に関して検討を加えた. 14 名の被験者の動画データを使用して評価を行ったところ，提案手法は機械学習を使用しないシンプルな手法を用いて，発話区間を判別することが可能であることを明らかにした.

第 4 章では，発話者の判別手法に関して検討を加えた．具体的には，音声を用いて推定された口唇の動きと実際の口唇の動きとの類似度を評価し，最も類似度が高い人物を発話者として判別する手法を検討した. 14 名の被験者の動画データを使用して評価を行ったところ，提案手法は従来の手法と比較して，単位学習データ当たりの学習効率が高いため，少ない学習データを用いて発話者判別のためのモデル構築が可能であることを明らかにした.

第 5 章は結論で，本研究で得られた主な成果と本論文の工学的意義および今後に残された課題について述べている.

目 次

第 1 章 緒論

1.1 本研究の背景

近年，働き方改革[1]の実現に伴い，多くの企業が業務の効率化に取り組んでいる．企業における長時間労働の改善策の実施割合[2]に関してまとめたグラフを図 1.1 に示す．作業の効率化に取り組んでいる企業数が最も多いことがわかる．作業の効率化における具体的な取り組みの事例としては，業務や会議の効率化，柔軟な働き方，ならびに労働時間管理の見直しなどが挙げられる(表 1.1 参照[2])．これらの取り組みは，新型コロナウイルスに伴う働き方の変化とも密接に関連する．具体的には，リモートワークやリモート会議に伴う業務や会議の効率化，出社とリモートワークを両立した柔軟な働き方，ならびにリモートワークにおける労働時間の管理の見直しなどが挙げられる．このような取り組みの中，会議の効率化が注目されている．会議は業務の方針などを決めるために有用な手段であるが，会議を行なっている間は業務の進捗が滞るため，会議に過剰な時間を割くことは業務の効率化に必ずしも有効ではないと考える．これを改善するために議事録が取得され，利活用される．議事録は議論した内容や取り決めの共有に有用であり，かつその後の会議の質を向上させるために役立てることができる．しかしながら，議事録を作成するためには最低でも 1 名の担当者が必要であり，コストが発生する．また，議事録作成中に発言内容の聞き漏らしや聞き間違いなどのヒューマンエラーが発生した場合，議事録作成がスムーズに行われず，かつ議論した内容が正しく共有されない場合がある．

上記の問題を解決するために，企業などは議事録自動作成システムが導入している．議事録自動作成システムは，音声認識技術を用いて会議参加者の発言内容を記録し，自動的にテキスト情報に書き起こすことが可能である[3, 4]．音声認識の技術を応用し，議事録を自動作成することは，議事録作成におけるヒューマンエラーの低減や，工数の削減に寄与できるため，会議および議事録の作成業務の効率化に寄与すると考える．また，新型コロナウイルスへの対応を想定した「働き方の新しいスタイル」の実現にも貢献できると考える．特に，発言ごとに発話者を自動判別する技術は，議事録自動作成手法における重要な要素技術である．議事録自動作成システムによって作成されたテキストデータに対して発言者を自動で割り振ることは，議事録作成後に話者を割り振るための作業を簡略化するため，業務の効率化に寄与すると考える．

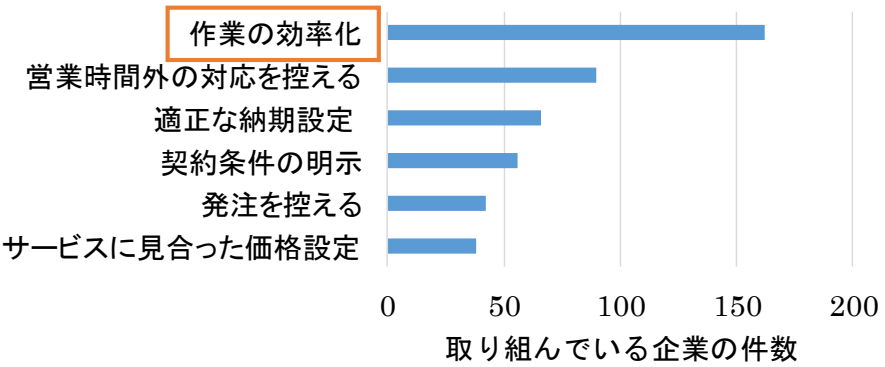


図 1.1　企業における長時間労働の改善策の実施割合[2]

表 1.1　取り組みの具体的事例[2]

業務や会議の効率化	・ 業務フローの見直しや RPA による業務改善・削減 ・ タブレット端末を活用してペーパーレス化 ・ 必要以上に資料を装飾しない／契約書等のフォーマットの定型化 ・ ウェブ，テレビ，電話による会議 ・ 必要な参加者のみの招集，制限時間の設定，目的の明確化など，会議のルールの見直し
柔軟な働き方	・ 海外顧客に対応したフレックスタイム制／テレワーク，サテライトオフィスの活用 ・ 裁量労働制の導入／計画的な有給休暇取得促進
労働時間管理の見直し	・ 勤務時間外のメールや電話の禁止／終業時刻，総労働時間管理，振替休日取得の厳格な管理等適正な勤務管理 ・ 部下の長時間労働の状況を管理監督者の人事考課に反映 ・ 定時にパソコンをシャットダウン

1.2 議事録自動作成システムの現状

実用化されている議事録自動作成システムに関する製品はいくつかあり，企業で取り入れられている事例もある．例えば，音声認識 AI クラウドサービスの「VoXT Voice Texting」[3]は，取得した会議や講演会などの音声を対象とし， AI による音声認識を行うことで，文字起こしの補助をすることができるアプリケーションである．しかしながら，この製品に発話者判別の機能は搭載されていないため，完全に議事録作成作業を自動化することは難しい．一方，NEC が開発した「音声認識・議事録作成支援ソリューション VoiceGraphy」[4]は，音声をテキスト情報として書き起こす機能と発話者を認識する機能が搭載されている．この機能は，取得した議事録に対して発話者を自動で割り振ることを可能にする．しかしながら，発話者を識別するためには会議参加者の人数と同じ数のマイクを事前に準備し，各マイクに会議参加者を割り振る必要があるため，設備への投資や事前の準備が必要であることが課題である．この課題に対する解決策として，Microsoft は「Speaker Recognition」[5]という API を実用化している．この製品は，マイクが会議参加者の人数と同じ数用意できない場合においても発話者を識別することが可能である．具体的には，各参加者の声紋を事前に登録し，これを用いて音声から発話者を識別する．しかしながら会議参加前に声紋を登録する必要があるため，事前準備が必要であるという点で課題が残されている．

1.3 本研究が想定する議事録自動作成システムの概観

議事録自動作成システムにおける発話者判別機能の課題を解決するために，画像情報から取得可能な口唇の動きおよび音声情報を併用することが有用であると考える．口唇の動きは発話した単語や文章特有の動きを有していることが知られている．このため，口唇の動きと音声の類似性を評価することで，発話者の判別が可能であると考える．口唇の動きと音声に相関があると仮定した場合，事前の声紋登録が不要である発話者判別手法の構築を行うことができる．また，音声を数台のマイクを使用して取得し，発話者の判別が可能である．

本研究では，議事録自動作成システムにおける発話者判別機能の課題を解決するために，画像情報と音声情報を併用した発話者判別手法の構築を行う．本研究が想定している議事録自動作成システムの発話者判別処理における概観を図 1.2 に示す．はじめに，カメラを使用して撮影した動画を対象として，顔認識・口唇抽出を行い，口唇の動きを取得する．次に，動画を対象として音声データを取得し，音声の特徴量を算出する．さらに，口唇の動きおよび音声の特徴量の発生タイミングに基づき，発話区間の抽出を行う．ここで，発話区間の生じた人物が 1 名である場合，その人物を発話者として判別する．最後に，複数の人物において発話区間が生じた場合，機械学習の 1 手法である Long Short Term Memory (以降，LSTM と表記する)[6]を用い，口唇の動きと音声の類似度を判別することで，類似度の最も高い人物を発話者として判別する．

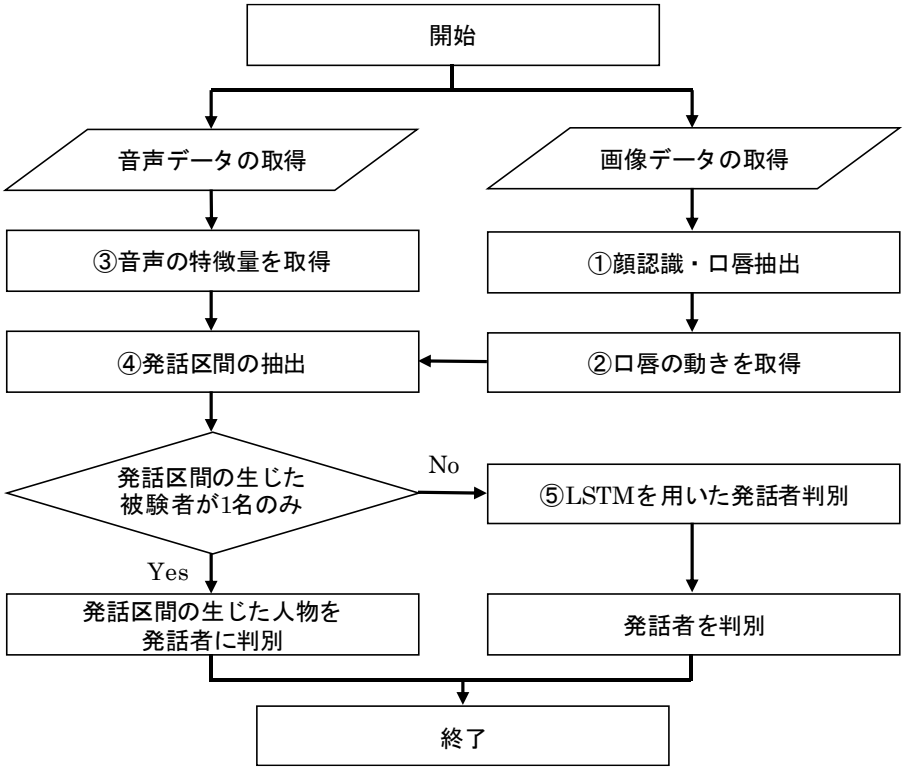


図 1.2　議事録自動作成システムにおける発話者判別手法の概観

1.4 関連研究

図 1.2 に示す発話者判別手法を構築するためには，①画像から口唇形状を抽出する手法，②発話区間を抽出する手法，③画像情報と音声情報の類似性を用いて発話者を判別する手法の 3 つの要素技術が必要である．各手法における関連研究を 1.4.1～1.4.3 項にまとめる．

1.4.1 口唇形状自動抽出法に関する関連研究

従来，佐藤氏らは[7]，蛍光灯照明環境下における人物の口唇領域自動抽出を目的として，ファジィ推論と口唇の形状情報を用いた手法(以降，ファジィ推論に基づいた従来手法と表記する)に関する検討を行った．ファジィ推論に基づいた従来手法では，顔画像における画素の色情報を L*a*b*色空間[8]に変換し，画素の赤味を示す a*値および黄味を示す b*を特徴量として使用した．これらの特徴量およびファジィ推論に基づいた 2 種類の関数を使用することで，口唇領域と肌領域のクラス分類を行い，口唇形状を抽出した．しかしながら，ファジィ推論に基づいた従来手法は，肌の色の個人差や陰影の影響で肌領域を過剰抽出する場合を認めた．一方，顔画像から顔の特徴点を抽出する研究[9]が行われており，特徴点を利用して顔の部位や形状を抽出する研究[10-13]が行われている．文献[10]では，深度画像および可視画像を用いて顔の特徴点を抽出し，口唇の位置を取得している．しかしながら，この手法は深度画像を用いているため，深度センサの搭載されていない一般的な Web カメラから取得された動画像には適用できない．文献[11]では，畳み込みニューラルネットワーク(以降，CNN と表記する)を用いて，可視画像から顔器官を抽出している．また，文献[12]では CNN に加えて，リカレントニューラルネットワークを用いて顔器官を抽出している．これらの手法では，あらかじめネットワークを学習させる必要があるため，多くの学習データを用意する必要がある．文献[13]では，顔の形状を表現したベクトルデータと，入力画像，ならびに学習済みの関数を用いて顔器官の特徴点抽出を行っている．この手法は，高精度かつ高速で口唇領域の抽出が可能であるが，前述した文献と同様に，多くの学習データが必要である．さらに，画像中の顔領域の一部(目や鼻など)が隠れている場合，顔器官の抽出に影響を及ぼし，口唇領域抽出が困難になる場合がある．

1.4.2 発話区間抽出における関連手法

発話区間抽出の類似手法として，音声区間検出[14]が挙げられる．音声区間検出とは，一般的にマイクロフォンから入力された音から，音声が存在する区間のみを抽出する手法として定義される．音声区間検出の目的としては，①音声の生じた区間のみを対象として処理時間を削減すること，②音声区間以外の領域に生じた雑音を除去し音声認識精度の向上を図ることが挙げられる．音声区間検出には「ゼロ交差率(zero crossing tare)」，「音声パワー(speech power)」，ならびに「信号対雑音比(signal-to-noise ratio；SNR)」が特徴量として一般的に使用される．ゼロ交差率とは，ある一定時間の間に音声波形がゼロレベルを交差する回数の割合を示す．ゼロ交差率が一定の閾値を超えた場合に，音声区間であると検出することが可能である．音声パワーは，音声の大きさを検出することで音声区間を検出するための特徴量である．信号対雑音比は雑音のパワーレベルに対する音声パワーのレベル比のことを示しており，この値と閾値を用いて音声区間の検出が行われる．これらの手法は，少ない計算量で音声区間検出が可能であるが，音声の有無が検出されるため，発話者の情報は検出することができない．

1.4.3 発話者判別に関する関連研究

発話者判別に関する研究事例として，声紋を使用した発話者判別手法[15, 16]が挙げられる．これらの研究では，音声認識で一般的に使用される特徴量の Mel-frequency cepstral coefficient (以降，MFCC と表記する)[14]を使用している．1 種類または複数種類の機械学習モデルへ MFCC を入力し，発話者を判別することが可能である．しかしながら，声紋を使用した発話者判別手法は，発話者判別対象の人物における声紋の情報を事前に登録または学習し，使用する必要がある．

一方，事前の声紋登録を行わない場合においても，発話者の判別が可能である．Peri 氏ら[17]は，ノイズの多い環境下において発話者を判別するためのノイズ除去手法を提案した．発話者判別の評価では k-means 法を適用し，既知のクラス数を設定して発話者の判別が行われた．Peri らの調査によって，ノイズのある環境下で声紋登録を行わずとも，発話者判別が可能であることが明らかにされた．しかしながら，クラス数を発話している人物の数と同じ数に設定した場合においても，一部の人物で発話者判別が正しく行われない場合がある．これは，類似した声を有した人物における発話者判別が困難であったためであると考える．

複数のマイクを使用することで，音声の到来方向を検出し，発話者を判別する手法が提案されている[18]．この方法では声紋の事前登録を必要とせずに発話者の判別が可能である．しかしながら，音声の到来方向を検出するためには複数台のマイクで構成された特殊な装置が必要である．また，音声の到来方向のみを用い

て発話者を識別する場合，発話者の追跡が困難になる場合がある．具体的には，発話者判別対象の人物が発声しない状態で移動した場合，音声の到来方向に基づいて発話者をラベル付けすることが困難になると考える．この場合，声紋の登録を用いた発話者判別を行う必要である．

Chung らの研究[19, 20]では，画像情報と音声情報を併用した発話者判別手法が提案されている．具体的には，機械学習の 1 手法である Convolutional neural networks（以降，CNN と表記する）および Long short-term memory（以降，LSTM と表記する)を用いて，画像中の口唇の動きと音声の類似度を評価し，発話者判別を行った．画像中の口唇の動きおよび MFCC を CNN へ入力し，算出された特徴量および LSTM を用いることで類似度の算出を行うことが可能である．これらの研究では，CNN および LSTM の学習のために約 606 時間のデータを使用しているため，教師データの収集，作成，ならびに学習に多くの時間が必要である点が課題である．

1.5 本研究の目的

会議における議事録の作成業務を効率化するために，音声認識に基づいた議事録自動作成システムの開発が行われ，実用化されている技術も存在する．特に，システムを用いて自動作成された議事録に対して，発話者を自動で割り振る技術は，議事録作成業務の効率化に大きく寄与するため，重要な要素技術である．具体的には，①マイクに人物を割り当てることで発話者を識別する方法，②声紋を用いて発話者を識別する方法，ならびに③画像情報と音声情報を用いて発話者を判別する方法などが提案されている．しかしながら，これらの技術は，①会議参加者の人数に応じて，マイクの数を調整することや，②事前に声紋の登録が必要である点が課題である．また，③教師データとして数百時間の動画データを必要とする点も課題である．これらの課題は，画像における口唇の動きおよび音声情報を使用することで解決可能であると考える．

そこで本論文では，発話者判別機能の搭載された利便性の高い議事録自動作成システムの開発に関する上記課題について研究を行い，工学上の進歩に寄与することを目的とする．すなわち，①口唇形状自動抽出手法，②発話区間抽出手法，ならびに③発話者判別手法を提案し，工学上の進捗に寄与することを目的とする．

会議中は異なる人物，かつ異なる照明環境下において行われ，かつ眼鏡をかけている人物や，前髪などの影響によって顔の部位(目など)が隠れる人物を対象とする場合がある．このため，画像中において顔の状態は多様になると考える．すなわち，画像における口唇形状と音声情報を用いた発話者の判別を行うためには，多様な環境下においても柔軟に口唇の形状情報を取得可能な手法を構築する必要がある．そこで本論文では，1 つ目の検討事項として，口唇形状の自動抽出手法について検討を加えた．

次に，発話に伴う口唇の動きと音声の発生タイミングに基づき，発話区間の抽出が可能であると考える．このとき，発話者が 1 名のみである場合は，発話区間の生じた人物を発話者として判別し，複数の人物が発話している場合は，発話者の候補者リストを算出することが可能であると考える．そこで本論文では，2 つ目の検討事項として，発話者判別を目的とした発話区間抽出手法について検討を加えた．

一方，発話区間抽出手法において複数の人物で口唇の動きが生じた場合，発話者の判別が困難になる場合があると考える．発話に伴う口唇の動きは発話内容特有の特徴を有していることが知られている．すなわち，発話に伴う口唇の動きと音声における類似性を評価することで，発話者の判別が可能であると考える．そこで本論文では，3 つ目の検討事項として，口唇の動きと音声情報の類似性に基づいた発話者判別手法に関する検討を加えた．

1.6 本論文の内容

本論文は全 5 章により構成され，第 1 章を緒論とした．

第 2 章では，口唇形状自動抽出手法について検討を行った．提案する口唇形状自動抽出手法は，L*a*b*色空間に基づき，口唇と肌の赤味に着目して口唇領域と肌領域をクラス分類する手法である．クラス分類には Feedforward neural network を使用した．被験者 5 名が眼鏡をかけていない状態，眼鏡をかけている状態，ならびにサングラスを掛けている状態で発話している様子を撮影し，これを対象として評価指標の Intersection over Union(以降，IoU と表記する)を算出した．この結果，提案手法は，最大で 0.9286 の高い IoU を得た．また，特徴点抽出手法と比較して，提案した口唇形状自動抽出法は，眼鏡やサングラスを掛けている場合においても平均で 0.9000 以上の高い IoU の値を得た．

第 3 章では，発話区間抽出手法に関して検討を行った．具体的には，動画の音声が生じた区間において，口唇の動きが生じた区間を人物ごとに抽出する手法の検討を行った．はじめに，特徴点抽出手法を使用して口内領域の縦幅を口唇の動き特徴量として算出した. 次に, MFCC を音声の特徴量として算出した. さらに，口唇の動き特徴量に基づいて抽出された発話区間のうち，音声の特徴量に基づいて抽出された発話区間を含む場合を最終的な抽出結果として取得した．最後に，被験者 14 名の動画データを対象として評価指標(F-measure)を算出した結果, 提案した発話区間抽出手法は，最大で 0.99 の高い F-measure の値を得た．また，比較手法[19, 20]と比較して提案した発話区間抽出手法は，14 名中 13 名で F-measure の向上を認めた．

第 4 章では，発話者判別手法について検討を行った．具体的には，口唇の動きと音声の類似性を評価し，最も類似した口唇の動きを有する人物を発話者として判別する手法について検討した．はじめに，口唇の縦幅と横幅の動きを口唇の動き特徴量として算出した．次に，MFCC の値を音声の特徴量として算出した．さらに，LSTM を用いて，音声の特徴量から口唇の動き特徴量を推定し，実際の口唇の動きが推定値と最も類似した人物を発話者として判別した．最後に，被験者 14 名が 11 文を発話している動画データを対象として発話者判別成功率算出した．この結果，比較手法[19, 20]に対する約 0.05%の量の学習データを使用し，最大で 93.0%，平均で 87.2%の判別成功率が得られた．

第 5 章では，本研究で得られた主な成果と本論文の工学的意義および今後に残された課題について述べている．

1.7 本論文で用いる主な用語

本論文で使用する用語について，以下に解説を加える．

- **口唇**[21]

  人体顔面の口部において，表皮の角化度(ケラチンの生成によって細胞が角質化している度合)が低く，血管が透けて赤く見える領域．

- **CIELAB 色空間(L*a*b*表色系, CIE 1976 L*a*b*色空間)**[22, 23]

  1976 年に CIE（国際照明委員会）が勧告した均等色空間であり，色の変化が知覚的に均等となる色空間である．均等色空間におけるユークリッド距離は，色の知覚的な違いを定量的に表す指標であり，「色差」と呼ばれている．

- Feedforward neural network (FFNN)[24]

  FFNN は，最も基本的かつ応用範囲の広いネットワークである．入力層，複数の中間層，ならびに出力層から構成され，各ユニットが全結合で次の層の各ユニットと連結している．入力層から入力された情報は出力層に向かって 1 方向に伝播し，入力値に対応した数値が出力される．

- Mel-frequency cepstrum coefficient (MFCC)[14]

  MFCC は音声認識に有用な特徴量を有しており、低次元成分に含まれるスペクトル包絡の特徴は、声道の音響特性や口腔の形状に起因して変化する．すなわち，MFCC の低次元成分の数値は，人間の口の形状に伴う特徴を有している．人間の聴覚が高周波になるにつれて，分解能が低くなる特性を表した特徴量である．

- Long short-term memory (LSTM)[6]

  時系列データの処理に特化したネットワークの 1 つである回帰結合型ニューラルネットワーク(Recurrent neural networks: RNN)を改良したモデル．RNN の回帰に加え，1 つのセル内で完結する内部回帰を有する構造をしている. また LSTM の内部には 3 つのゲートが情報を制御しているため, RNN と比較して勾配消失問題が起こりにくい特徴を有している．

1.8 本論文における個人情報を有するデータの取り扱い

本論文に関する各データ（人物画像，動画，評価者の主観評価・アンケートなど）は「秋田大学手形地区における人を対象とした研究に関する倫理規程第 6 条 2 項」に基づいて倫理審査の申請を行い，承認された研究計画の下に，被験者本人の了承を得て取得し，これを解析および実験に使用している．

第1章 参考文献

[1] 厚生労働省：「働き方改革の実現に向けて」, https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000148322.html (Access 2021/12/15)

[2] 日本経済団体連合会：「2019 年労働時間等実態調査 集計結果」, https://www.keidanren.or.jp/policy/2019/076.pdf (Access 2021/12/15)

[3] VoXT Voice Texting：「VoXT とは」, https://voxt.jp/outline/ (Access 2021/12/15)

[4] NEC ソリューションイノベータ：「Voice Graphy」, https://www.nec-solutioninnovators.co.jp/ss/smartwork/products/voicegraphy/ (Access 2021/12/15)

[5] Microsoft Azure：“Speaker Recognition”, https://azure.microsoft.com/ja-jp/services/cognitive-services/speaker-recognition/ (Access 2021/12/15)

[6] 斎藤康毅：「ゼロから作る Deep Learning –自然言語処理編-」, O’Reilly Japan, Inc. (2018)

[7] 佐藤慶幸, 成田純一, 景山陽一, 西田眞：「口唇の形状情報を用いた口唇領域自動抽出処理の改善」, 電学論 C, Vol.130, No.5, pp.873-881 (2010)

[8] 高木幹雄, 下田陽久監修：「新編 画像解析ハンドブック」, 東京大学出版会 (2004)

[9] A. Azeem, M. Sharif, J.H. Shah, and M. Raza：“Hexagonal Scale invariant feature transform (H-SIFT) for facial feature extraction”, Journal of Applied Research and Technology, Vol.13, No.3, pp.402–408 (2015)

[10] S. Jahanbin, A.C. Bovik, and H. Choi：“Automated facial feature detection from portrait and range images”. In Image analysis and interpretation, SSIAI.2008 IEEE southwest symposium on Image Analysis and Interpretation, DOI:10.1109 / SSIAI.2008.4512276 (2008)

[11] A. Jackson, M. Valstar, and G. Tzimiropoulos：“A CNN Cascade for Landmark Guided Semantic Part Segmentation”, ECCV 2016 Workshops, arXiv:1609.09642 (2016)

[12] U. Güçlü, Y. Güçlütürk, M. Madadi, S. Escalera, X. Baró, J. González, R. van Lier, and J.A.M van Gerven：“End-to-end semantic face

segmentation with conditional random fields as convolutional", recurrent and adversarial network. arXiv:1703.03305 (2017)

[13] V. Kazemi, and J. Sullivan："One millisecond face alignment with anensemble of regression trees", 2014 IEEE Conference on Computer Vision and Pattern Recognition, DOI:10.1109/CVPR.2014.241, Columbus, OH, USA (2014)

[14] 篠田浩一：「音声認識 Speech Recognition」，講談社 (2017)

[15] Z. Meng, M. Umair Bin Altaf, and B. Juang："Active voice authentication", Digital Signal Processing, Vol. 101, 102672 (2020)

[16] X. Wang, F. Xue, W. Wang, and A. Liu："A network model of speaker identification with new feature extraction methods and asymmetric BLSTM", Neurocomputing, Vol. 403, pp. 167−181 (2020)

[17] R. Peri, M. Pal, A. Jati, K. Somandepalli, and S. Narayanan："Robust speaker recognition using unsupervised adversarial invariance", 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, 10.1109/ICASSP40776.2020.9054601 (2020)

[18] F. Grondin, and F. Michaud："Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations", Robotics and Autonomous Systems, Vol. 113, pp. 63−80 (2019)

[19] J.S. Chung, and A. Zisserman："Out of time: automated lip sync in the wild", Workshop on Multi-view Lip-reading, ACCV (2016)

[20] J.S. Chung, and A. Zisserman："Learning to lip read words by watching videos", Computer Vision and Image Understanding, Vol. 173, pp. 76−85 (2018)

[21] 小野尊睦，飯塚忠彦，吉武一貞：「口腔外科学 改訂 7 版」，金芳堂 (2010)

[22] 日本色彩学会：「新編 色彩科学ハンドブック 第 3 版」，東京大学出版会 (2011)

[23] 高木幹雄，下田陽久監修：「新編　画像解析ハンドブック」，東京大学出版会 (2004)

[24] 岡谷貴之：「深層学習」，講談社 (2015)

第 2 章　口唇形状自動抽出法に関する検討

2.1 はじめに

会議は複数の人物，かつ不定の環境下において行われる．また，会議の参加者には眼鏡をかけている人物がいたり，前髪などの影響によって顔の部位(目など)が隠れたりする場合がある．このため，議事録自動作成システムを構築する場合，上記状況下で利用されることを想定したシステムを開発することが好ましいと考える．したがって，会議中の画像情報を用いて口唇形状を抽出し，発話者判別に活用するためには，会議参加者の顔が撮影される環境の違いに対してロバストな口唇形状抽出手法を構築し，使用することが求められる．

従来，佐藤氏ら[1]は，蛍光灯照明環境下における人物の口唇領域自動抽出を目的として，ファジィ推論と口唇の形状情報を用いた手法(以降，ファジィ推論に基づいた従来手法と表記する)に関する検討を行った．しかしながら，ファジィ推論に基づいた従来手法は，肌の色の個人差や陰影の影響を受け，肌領域を過剰抽出する場合を認めた．

一方，顔画像から顔の特徴点を抽出する研究[2]が行われており，特徴点を利用して顔の部位や形状を抽出する研究[3-6]が行われている．Kazemi 氏ら[6]は，顔の形状を表現したベクトルデータと，入力画像，ならびに学習済みの関数を用いて顔器官の特徴点抽出を行っている(以降，輝度勾配に基づいた従来手法と表記する)．この手法は，高精度かつ高速で口唇領域の抽出が可能であるが，多くの学習データが必要である．さらに，画像中の顔領域の一部(目や鼻など)が隠れている場合，顔器官の抽出に影響を及ぼし，口唇領域抽出が困難になる場合がある．

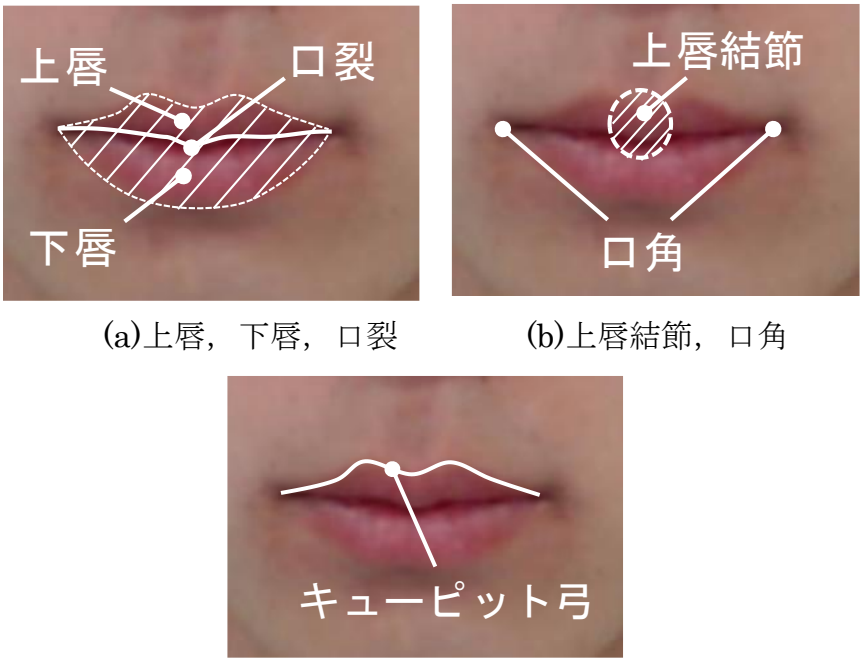
そこで本章では，上記の手法における①肌の色の個人差や陰影の影響によって肌領域を過剰抽出する課題，②多くの学習データを必要とする課題，ならびに③顔の部位が一部隠れている場合に正しく口唇領域を抽出できない課題の解決を目的とし，順伝播型ニューラルネットワーク[7]を用いた口唇形状自動抽出手法を提案する(以降，提案手法と表記する)．具体的には，処理対象動画データの人物における口唇領域抽出モデルを自動で学習することで，人物における肌の色の違いおよび撮影環境の違いに対してロバスト性の高い手法の構築を行った．また，学習データには 1 フレームのみの画像を使用し，かつ口唇領域周辺のみの画像を処理対象に設定することで，②多くの学習データを必要とする課題，および③顔の部位が一部隠れていた場合に口唇領域の抽出が困難となる課題の解決を行った．提案手法，ファジィ推論に基づいた従来手法，ならびに輝度勾配に基づいた従来手法を用いて，口唇領域を自動抽出し，抽出精度の評価を行った．

2.2 口唇の特徴と色空間

2.2.1 口唇の形状

本章において，表皮の角化度が低く血管が透けて赤色に見える赤唇領域[8]を口唇と定義した．口唇は部位毎に異なる特徴を有している．各部位の名称と，対応する各領域について以下にまとめる．また，口唇における各領域の位置を図 2.1 に示す．

- 口裂：上唇と下唇の境界領域
- 上唇：口唇における口裂の上部領域
- 下唇：口唇における口裂の下部領域
- 口角：口唇の左右端の領域
- 上唇結節：上唇中央部で膨らみが見られる領域
- キューピット弓：上唇の弓状形状



(a)上唇，下唇，口裂　　(b)上唇結節，口角

(c)キューピット弓

図 2.1　口唇における各部位の位置[8]

2.2.2 口唇の有する色彩情報の概要

本論文では，赤唇領域[8]と呼ばれる領域を口唇と定義している．このため，肌と比較して赤味を有していることが口唇の色彩情報における顕著な特徴であると考える．このように，色彩情報を人間の感覚に近い形で扱うためには，色を定量的・定性的に把握し，かつ 2 色の色差を定量的に評価可能な均等色空間[9,10]を使用する必要があると考える．

ファジィ推論に基づいた従来研究[1]において，顕色系であるマンセル表色系に写像可能な$L^*a^*b^*$色空間[9,10]に着目し，これを特徴量に用いた口唇抽出手法の検討が行われた．$L^*a^*b^*$色空間を図 2.2 に示す．$L^*a^*b^*$色空間は，均等色空間の一つであり，色の分布を人間の感覚に近い形で捉えることが可能である．このため，本研究においても$L^*a^*b^*$色空間を用いて特徴量を算出し，口唇抽出処理に用いた．

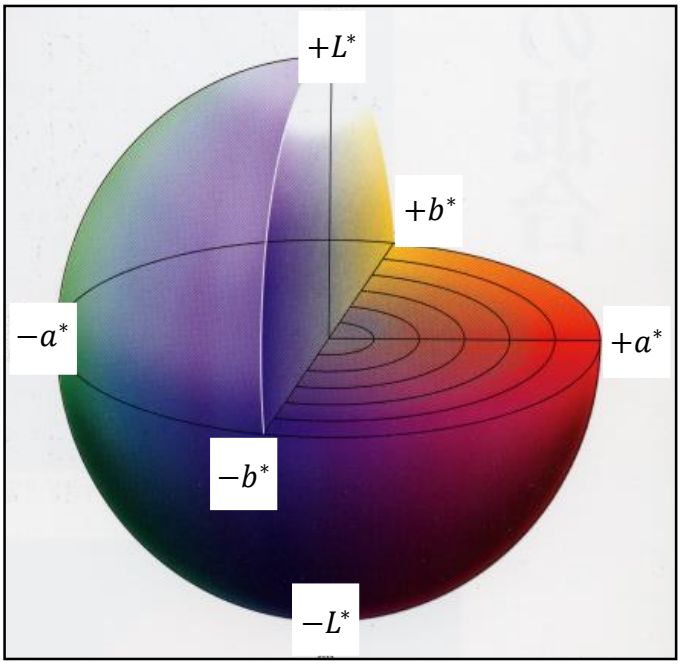


図 2.2 $L^*a^*b^*$色空間[9,10]

2.2.3 L*a*b*色空間

$L^*a^*b^*$色空間は，以下に示す(2.1)式から(2.3)式によって定義される[9,10]．ここで，L^*は明度指数である．また，a^*およびb^*は色相と彩度に関係する量であり，a^*は正に大きい程赤，負に大きい程緑であること，b^*は正に大きい程黄，負に大きいほど青であることをそれぞれ表している．

$$L^* = \begin{cases} 116.0\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - 16.0 & if \quad \left(\frac{Y}{Y_n}\right) > 0.008856 \\ 903.29\left(\frac{Y}{Y_n}\right) & if \quad \left(\frac{Y}{Y_n}\right) \leq 0.008856 \end{cases} \tag{2.1}$$

$$a^* = 504.3\left(\left(\frac{X}{X_n}\right)^{\frac{1}{3}} - \left(\frac{Y}{Y_n}\right)^{\frac{1}{3}}\right) \tag{2.2}$$

$$b^* = 201.7\left(\left(\frac{Y}{Y_n}\right)^{\frac{1}{3}} - \left(\frac{Z}{Z_n}\right)^{\frac{1}{3}}\right) \tag{2.3}$$

ここで，XYZ 表色系[9, 10]における三刺激値X，Y，Zは，RGB 表色系[9, 10]における三刺激値R，G，Bを用いて(2.4)式のように定義される．また，標準の光における完全拡散面の三刺激値X_n，Y_n，Z_nは(2.5)式のように与えられる．

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.4124 & 0.3576 & 0.1805 \\ 0.2126 & 0.7152 & 0.0722 \\ 0.0193 & 0.1192 & 0.9505 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.4}$$

$$(X_n \quad Y_n \quad Z_n) = (98.072 \quad 100.00 \quad 118.225) \tag{2.5}$$

2.3 使用データ

顔の部位の一部が遮蔽されている場合における口唇領域抽出手法の検討と評価を行うことを目的として，被験者が発話している様子を動画データとして取得した．データ取得では，眼鏡や前髪などによって目が隠れる事例を想定し，被験者ごとに目周辺の状態を 3 パターンずつ設定し撮影した．具体的には，眼鏡やサングラスなどをかけていない状態(以降，眼鏡無しデータと表記する)，眼鏡をかけている状態(以降，眼鏡ありデータと表記する)，ならびにサングラスを掛けている状態(以降，サングラスありデータ)を取得した．本節では，データ取得環境，データセットの設定方法，ならびに正解の口唇領域を示すマスク画像の作成方法について記述する．

2.3.1 データ取得環境

被験者 5 名(a～e)が「眼鏡をかけていない状態（眼鏡無しデータ)」,「眼鏡をかけている状態（眼鏡ありデータ)」，ならびに「サングラスをかけている状態（サングラスありデータ)」において，自分自身の氏名を 1 回ずつ発話した動画像を Web カメラ(Logicool 社製; C922 PRO STREAM WEBCAM)[11, 12]を使用して正面から撮影した．データ取得環境を図 2.3 に示す．また，データ取得条件を①～⑥に示す．なお，被験者にとって本人の氏名が最も発話しやすく，自然な口の動きとなることから自分自身の氏名をコマンドとして選定した．さらに，顔領域の一部が隠れている画像に対する評価を行うために，眼鏡およびサングラスをかけた場合における顔画像を取得している．

＜データ取得条件＞

① 日常一般的と考えられる蛍光灯による照明下(500～900lx)
② 被験者は 20 代日本人（男性：4 名，女性：1 名）
③ 被験者とカメラ間の距離は約 40cm
④ 発話の前後は口を閉じる
⑤ 口紅を塗布しない
⑥ 発話区間はオペレータ 1 名が目視で判定する

データ取得において，発話に伴い，口唇の形状が変化しても取得画像中に口唇形状が全て含まれる必要がある．そこで，データ取得画面内に 75×40 画素の矩形を設定し，被験者はディスプレイに表示された自身の状況をリアルタイムで確認しながら，矩形内に自身の口唇が収まるように微調整を行った．この位置調整は個人差があるものの，数 cm 程度である．さらに，矩形を設定することで発話中の無意識な頭部の前後移動を抑制する効果がある．取得した動画像を 30fps で分割し，得られた静止画像(24bit カラー，640×480 画素)を検討に使用した．

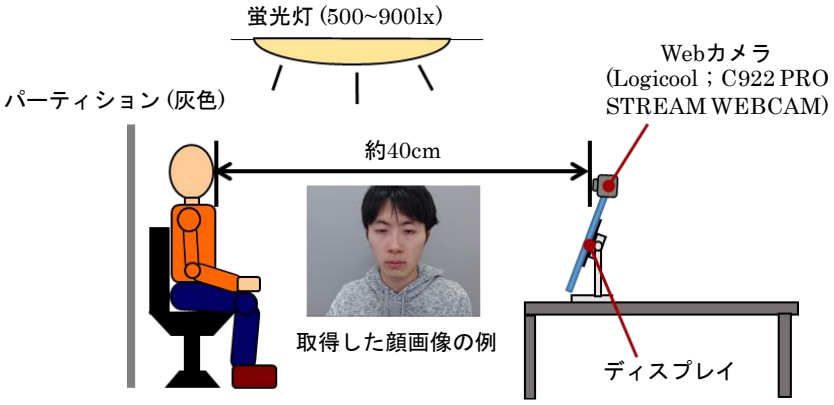


図 2.3　データ取得環境

2.3.2 使用機材について

データ取得には，Web カメラ(Logicool 社製; C922 PRO STREAM WEBCAM)を使用した．カメラの外観を図 2.4，とカメラの仕様を表 2.1 に示す．なお，撮影時はオートフォーカス機能を有効にし，ホワイトバランスの調整を自動に設定した上でデータを取得した．

図 2.4　Logicool 社製; C922 PRO STREAM WEBCAM の外観[11]

表 2.1　Logicool 社製; C922 PRO STREAM WEBCAM の仕様[12]

項目	仕様
製品寸法	高さ：29mm 長さ：95mm 奥行き：24mm 重量：162g
録画画像サイズ，フレームレート※	1080p30fps、720p60fps、720p30fps
フォーカスタイプ	20 段階オートフォーカス
対角視野（FOV）	78°
水平視野(FOV)	70.42°
垂直視野(FOV)	43.3°

※撮影用プログラムを用いて 720p30fps を 640×480 画素 30fps にトリミングして撮影している．

2.3.3 データセット

本研究では各種パラメータの検討範囲を増やし，より適切なパラメータを設定する必要がある．また，2.5 節で後述する口裂抽出処理の評価はオペレータの目視により行われる．このため，検討対象のデータを選定して使用することで，検討時間の効率化を図った．具体的には，表 2.2 に示すすべてのデータを対象として，以下に示す 3 パターンのデータセット(i～iii)を設定し各検討に使用した．

i) 口裂抽出処理評価用データセット:

被験者 5 名(a～e)の眼鏡を掛けていない状態の発話データにおける「1 から 10 フレーム目までのフレーム」，および「最後のフレームから 10 フレーム前までのフレーム」

ii) 学習条件検討用データセット:

被験者 a が眼鏡をかけていない状態で取得したデータ

iii) 口唇抽出精度評価用データセット:

被験者 5 名の「眼鏡無しデータ」，「眼鏡ありデータ」，ならびに「サングラスありデータ」

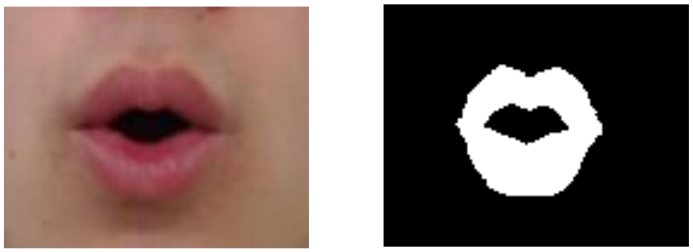
表 2.2　取得した各動画データの総フレーム数 (30 fps)

被験者	眼鏡無しデータ	眼鏡ありデータ	サングラスありデータ	合計
a	111	94	98	303
b	97	92	100	289
c	80	73	91	244
d	72	69	80	221
e	71	69	88	228
合計	431	397	457	1285

2.3.4 正解の口唇領域を示すマスク画像の作成

提案手法の評価を行うために，使用データにおける口唇周辺の画素を対象とし，マスク画像を作成した(図 2.5 参照)．マスク画像は，2.2.1 項「口唇の形状」における各部位の定義に基づき，以下の①～③の定義に一致した画素を口唇領域とし，オペレータ 1 名が作成している．なお，口内の画素は検討対象外とした．

① 顔面の口部に存在し，肌と比較して赤みを帯びている領域に属する画素
② 口角に属する画素
③ 上唇と下唇が接触している領域および口裂に属する画素



(a)元画像　　(b)マスク画像

図 2.5　口唇領域におけるマスク画像例

2.4 提案手法

2.4.1 提案手法の概要

提案手法は，以下の①～⑥の手順で発話 1 回分のデータから口唇領域を抽出する．フローチャートを図 2.6 に示す．

① 発話 1 回分の動画像から 1 フレーム目の画像を取得する．
② 1 フレーム目の画像に対して，口裂抽出処理を施す．
③ 抽出された口裂を基準に，特徴量を取得する画素を算出する．
④ 特徴量を取得する画素の位置を再設定する．
⑤ 再設定した「特徴量を取得する画素」の位置および 1 フレーム目の画像を用いて，FFNN の学習を行う．
⑥ FFNN を用いて，すべてのフレームの口唇領域を抽出する．

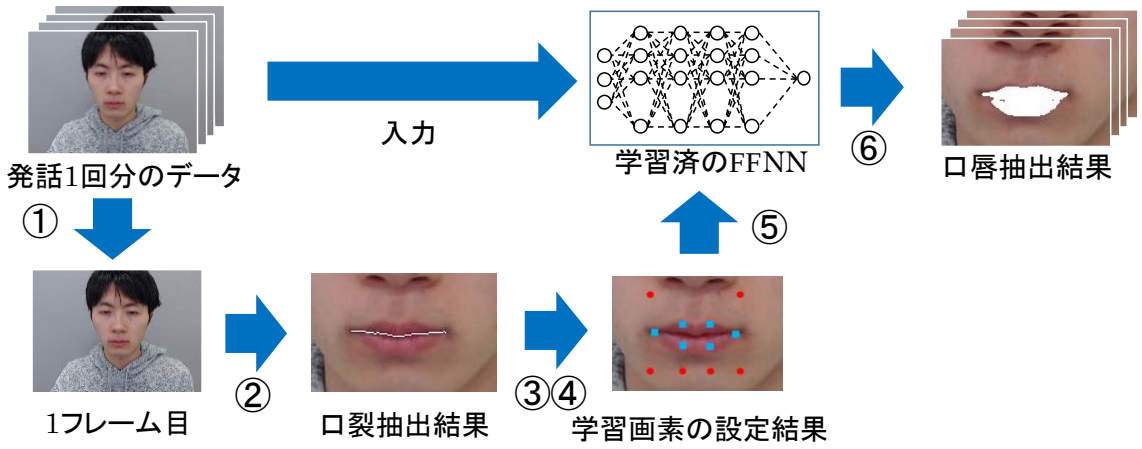


図 2.6　口唇形状自動抽出法(提案手法)の概要

2.4.2 順伝播型ニューラルネットワーク（FFNN）

本論文で使用する順伝播型ニューラルネットワーク(FFNN)[7]は，最も基本的かつ応用範囲の広いネットワークである．入力層，複数の中間層，ならびに出力層から構成され，各ユニットが全結合で次の層の各ユニットと連結している(図 2.7 参照)．入力層から入力された情報は出力層に向かって 1 方向に伝播し，入力値に対応した数値が出力される．本論文において，FFNN の学習には確率的勾配降下法[7]を用い，各ユニットにおける活性化関数は，(2.6)式のロジスティク関数[7]を用いた．関数の形状を図 2.8 に示す．

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.6}$$

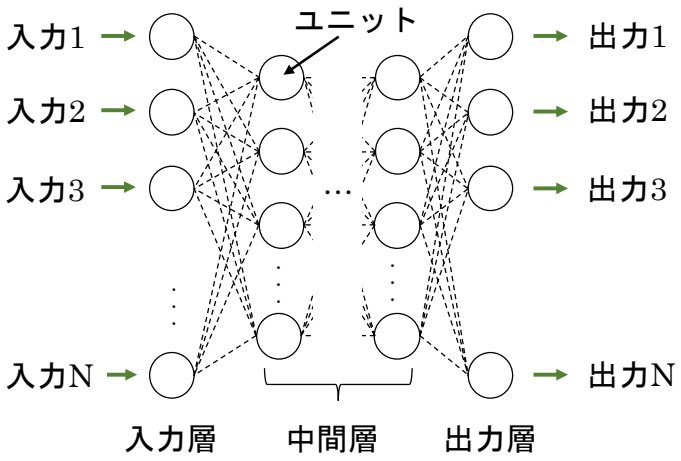


図 2.7　順伝播型ニューラルネットワーク(FFNN)の例

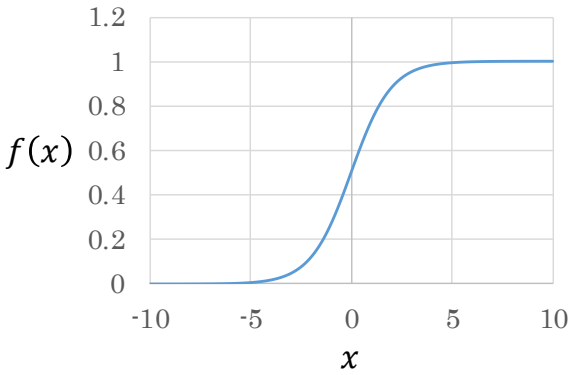


図 2.8　ロジスティック関数

2.4.3 口裂抽出処理

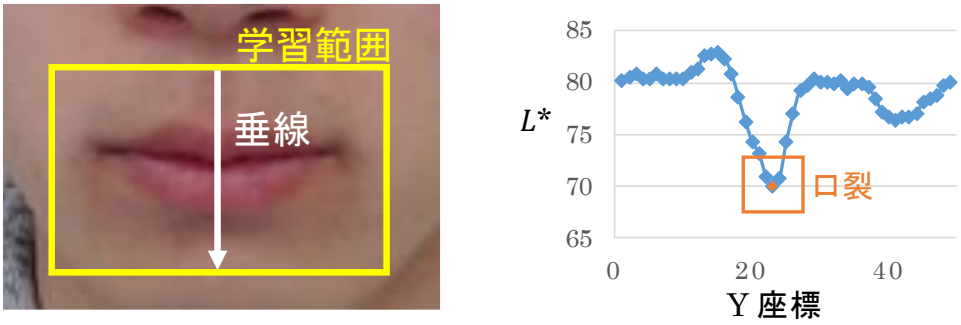
提案手法では発話 1 回分のデータにおける 1 フレーム目の画像から特徴量を取得し，FFNN の学習を行う．そこで，特徴量の取得位置に関する基準を設けるため，①～④の手順で口裂を抽出した．

① 学習範囲の設定

被験者が口を閉じている画像(1 フレーム目)に対して学習範囲を設定した．具体的には，被験者の口唇領域を包含し，かつ顔の領域に収まる矩形を設定し，これを学習範囲とした．

② 明度分布に着目した口裂抽出処理

口唇領域において口裂は，上唇と下唇の境界であることから陰影が生じ，周辺の領域と比較してL^*値(明度値)が低い．そこで，口唇領域におけるL^*値に着目して明度分布に谷が生じる画素の抽出を行い，口裂を抽出した．はじめに，学習範囲の上端から下端に向かって任意の位置に垂線を 1 本設定し，垂線上の明度分布をグラフ化した(図 2.9 参照)．次に，垂線上における任意の画素 A とその上下の画素 B(上)，C(下)を設定した(図 2.10 参照)．最後に，画素 A，B，C のL^*値を比較し，表 2.3 に示す谷の条件を満たした場合，画素 A は谷，山の条件を満たした場合，画素 A は山と判定し，谷を口裂として抽出した．最後に，谷の抽出を垂線上におけるすべての画素に対して実施した．



(a)垂線の設定例　　(b)垂線上の明度分布例

図 2.9　垂線上における明度分布の取得例

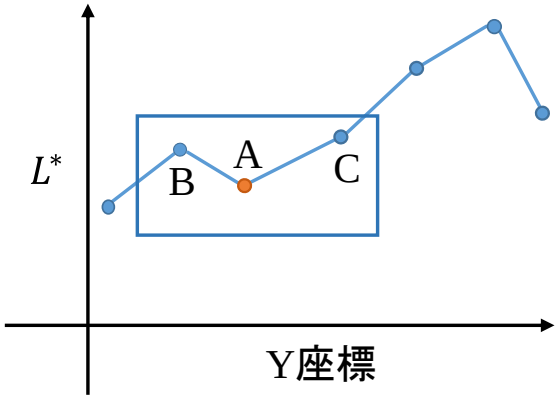


図 2.10　任意の連続した 3 点の設定例

表 2.3　口裂抽出処理のための条件式

	条件式
谷の条件	$(B > A = C) \vee (A < B \wedge A < C) \vee (B = A < C)$
山の条件	$(B < A = C) \vee (A > B \wedge A > C) \vee (B = A > C)$

※A，B，C：図 2.10 における任意の連続した 3 点の明度値(L^*)を示す.

③ ノイズ除去処理

谷と判定された画素には，口裂ではない画素がノイズとして含まれている場合がある．このため，谷と判定された画素群から口裂に対応しない画素を除去する処理を実施した．具体的には，図 2.11 に示すように，垂線の上端から下端に向かって明度分布をグラフ化し，谷と判定された画素とその前後に存在する山とのL^*値の差をt_1(前)，t_2(後)，最も深い谷とその前後に存在する山とのL^*値の差をT_1(前)，T_2(後)とした場合に，表 2.4 に示す条件 1，2 を満たした谷をノイズと判定した．なお，表 2.4 の閾値は，条件 1 では 0.0～10.0 まで 1.0 刻みで，条件 2 では 0.0～0.4 まで 0.1 刻みでそれぞれ検討し，最も良好に口裂を抽出可能な値を設定している．

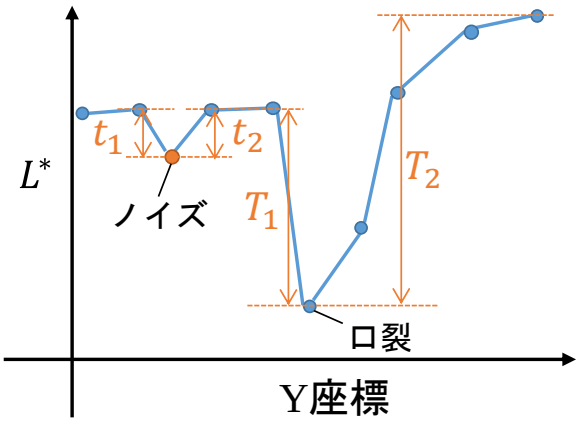


図 2.11　ノイズ除去処理に使用する 4 つの数値

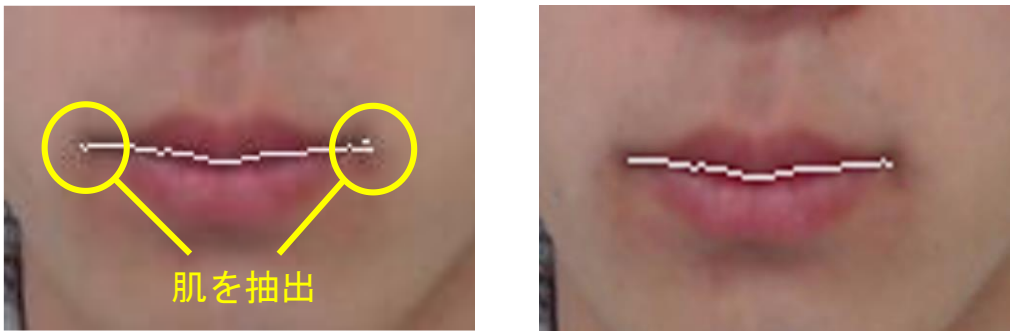
表 2.4　ノイズ除去のための条件式

	条件式
条件 1	$t_1 < 6.0 \ \wedge \ t_2 < 6.0$
条件 2	$\dfrac{min(t_1, t_2)}{min(T_1, T_2)} < 0.2$

※$min(\alpha, \beta)$：αとβのうち，小さい方の値を使用する意味を示す．

④ 口裂の補正処理

学習範囲の左端から右端にかけて本項②「明度分布に着目した口裂抽出処理」および③「ノイズ除去処理」を実施した．この結果，口裂を白い線として抽出する画像を得た．しかしながら，図 2.12 (a)に示すように，明度分布の谷が肌領域に生じることに起因し，適切に口裂が抽出されていない事例を認めた．そこで，肌の領域と比較して口裂の領域におけるL^*値が低いことに着目し，谷の画素群から肌領域を除外した．具体的には，すべての垂線において谷と判定された画素のL^*値を取得し，その平均値を算出する．さらに，「平均値以上のL^*値を有する谷は口裂でない」と判定し，谷の画素群から除外した．口裂の補正処理結果例を図 2.12 (b)に示す．



(a)補正前の口裂　　(b)補正後の口裂

図 2.12　口裂の補正処理結果例（白線：抽出した口裂）

2.4.4 口裂を基準とした学習画素の初期設定

1 フレーム目の画像における口唇領域を対象とし，図 2.13 に示すように，6 点の画素(以下，口唇の学習画素と表記)および，肌領域に 6 点の画素(以下，肌の学習画素と表記)をそれぞれ設定した(以下，学習画素と表記)．なお，すべての被験者における口唇領域内に“口唇の学習画素”が収まるように，口裂と学習画素の距離 m (図 2.13 参照)を 3 画素に設定した．

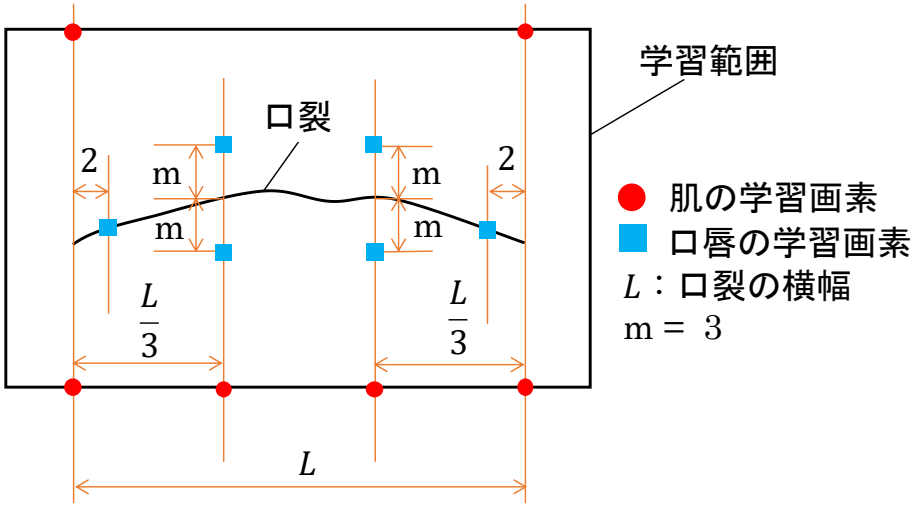


図 2.13　学習画素の初期設定

2.4.5 特徴量の取得

ファジィ推論に基づいた従来手法[1]において，$L^*a^*b^*$色空間[9, 10]で定義される知覚色度指数a^*が口唇領域の抽出に有用であることを明らかにしている．このため，本論文においても同様に$L^*a^*b^*$色空間のa^*値に着目し，口唇の特徴を取得した．具体的には，顔画像中の着目画素が口唇に属するか，肌に属するかを判定するために，着目画素とその周辺における画素のa^*値を取得し，以下の①および②に示す特徴量として検討に用いた．特徴量の例を図 2.14 に示す．なお，a^*値は(2.7)式を用いて－1.0～＋1.0 の範囲に正規化して使用した．

① SQ_n：着目画素および，これを中心として$n \times n$の領域から取得したa^*値（n: 3,5,7）

② CR_n：着目画素および，これを中心とした$n \times n$の領域におけるa^*値のうち，着目画素と x 座標または y 座標が同じ数値である画素のa^*値（n: 3,5,7）

$$a^{*\prime} = \frac{2 \times (a^* - a^*_{min})}{(a^*_{max} - a^*_{min})} - 1.0 \qquad (2.7)$$

ここで，

a^*：正規化前のa^*値，

$a^{*\prime}$：正規化後のa^*値，

a^*_{max}：a^*値の最大値，

a^*_{min}：a^*値の最小値である．

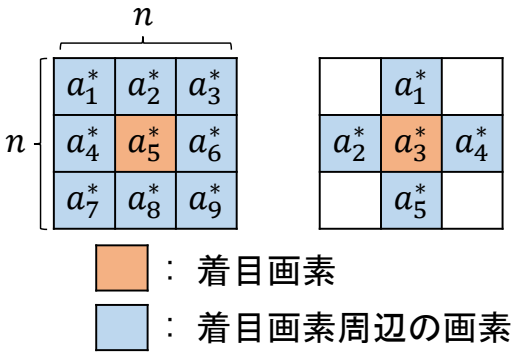


図 2.14　特徴量の例（n =3 の場合）

2.4.6 学習画素の再設定

初期設定の学習画素を着目画素とし，特徴量を取得した場合，上唇および下唇の中央付近の画素ではなく，口裂付近の陰影を含む画素を取得してしまう場合を認めた．このとき，本来の口唇の色情報は取得できないと考える．これは，口唇の厚さの個人差に起因するため，被験者ごとに適切な m (図 2.13 参照)を再設定する必要がある．そこで，以下の①～③の手順に従い，口唇の学習画素における m の再設定を行った．

① FFNN の学習処理

口唇領域を抽出するために，1 フレーム目の顔画像を用いて FFNN の学習処理を実施した．このとき，図 2.15 に示すように，各学習画素を着目画素として，2.4.5 項で示した特徴量を取得したものをそれぞれ教師データとして使用した．学習には確率的勾配降下法[7]を用い，口唇の学習画素から取得した特徴量を入力した場合に 1.0 を出力し，肌の場合には 0.0 を出力するよう FFNN を学習した．なお，図 2.15 では説明のために学習画素の位置を拡大表示している．

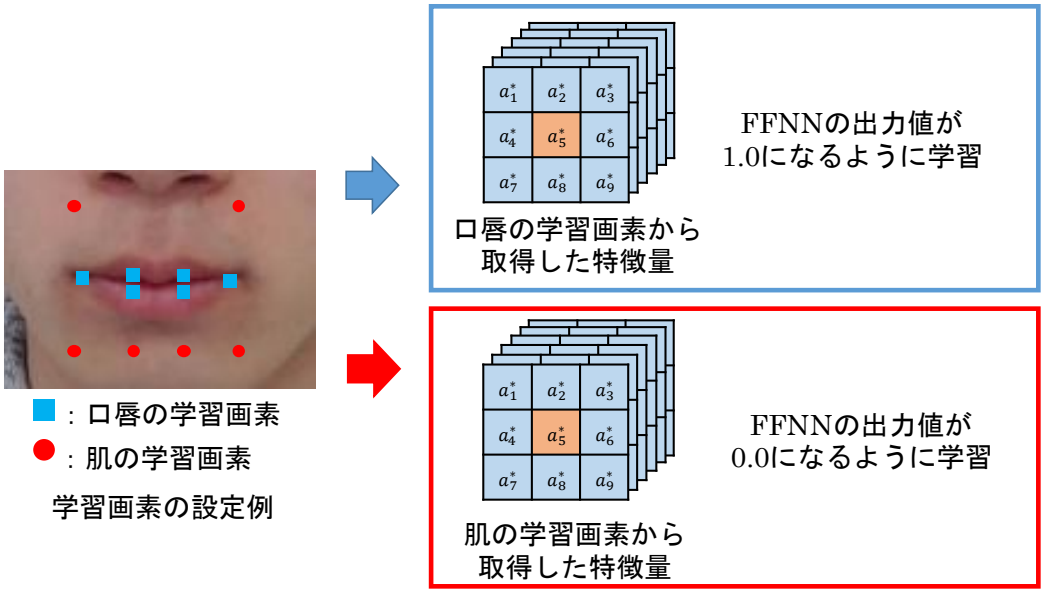


図 2.15　FFNN の学習処理例（特徴量SQ_3の場合）

② 口唇抽出処理

図 2.16 に示すように，画像中の学習範囲内における任意の画素を着目画素として「学習に用いたものと同じ特徴量」を取得し，学習済みの FFNN に入力した．これを学習範囲内のすべての画素に対して行った．本論文では，FFNN の出力値が 0.99 以上となった場合，着目画素を「口唇領域に属する画素」と判定して口唇領域を抽出した．また，判定した画素にはノイズが含まれている場合があるため，ノイズ除去処理を施した．具体的には，はじめに抽出した口唇領域に対して膨張処理[10]を行った．次に，膨張処理を施した画像に対して，ラベリング処理[10]を用いて最大領域を抽出し，マスク画像を作成した．最後に，マスク画像と口唇領域の論理積を算出し，ノイズを除去した．

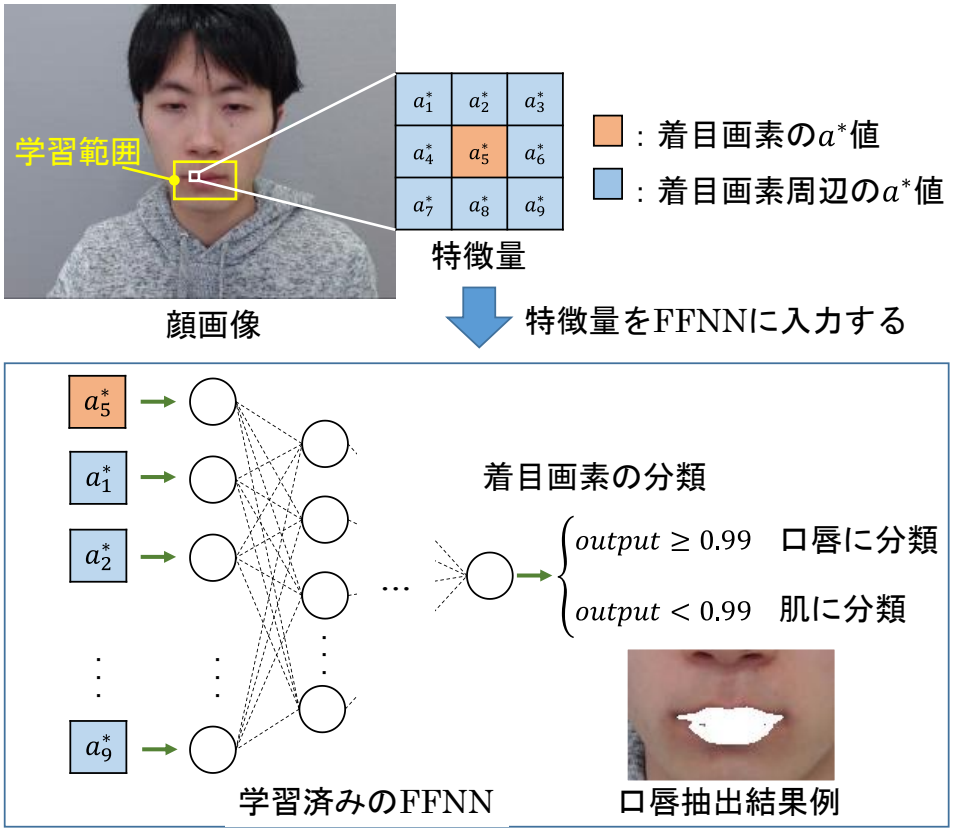


図 2.16 口唇抽出処理（特徴量SQ_3の場合）

③　学習画素の再設定処理

口唇抽出処理を 1 フレーム目の画像に対して実施し，抽出された口唇領域の上端および下端の位置を取得した．これらを用いて「口唇の学習画素と口裂間の距離」mを再設定した．具体的には，図 2.17 に示すように，口裂と口唇上端の距離(M_1, M_2)，ならびに口裂と口唇下端の距離(M_3, M_4)の値を取得した．さらに，(2.8)式を用い，口唇の学習画素と口裂の距離$(m_1 \sim m_4)$を再設定した．再設定前の学習画素と再設定後の学習画素の例を図 2.18 に示す．なお，実際のデータにおける学習画素は 1 画素であるが，図 2.18 では説明のために学習画素の位置を拡大表示している．

$$m_k = \frac{2}{3} M_k \quad (k: 1,2,3,4) \tag{2.8}$$

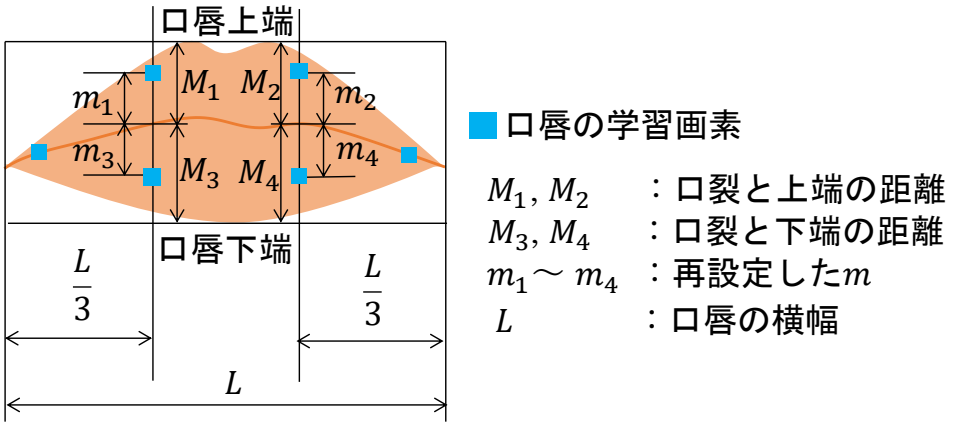


図 2.17　学習画素の再設定

(a)初期設定の学習画素例　　(b)再設定後の学習画素例

図 2.18　学習画素の再設定結果例

(青：口唇の学習画素，赤：肌の学習画素)

2.4.7 FFNNの学習および口唇抽出処理

再設定後の学習画素および 1 フレーム目の画像を用いて，図 2.19 に示すように，FFNN の学習および口唇領域の抽出を行った．はじめに，発話 1 回分のデータにおいて口唇領域がすべて含まれるように学習範囲を設定した．次に，再設定後の学習画素および 1 フレーム目の画像を用いて，2.4.6 項 ①と同様の方法で FFNN の学習を行った．最後に，学習済みの FFNN を用いて，発話 1 回分のすべてのフレームに対して 2.4.6 項 ②と同様の処理を実施し，口唇領域を抽出した．

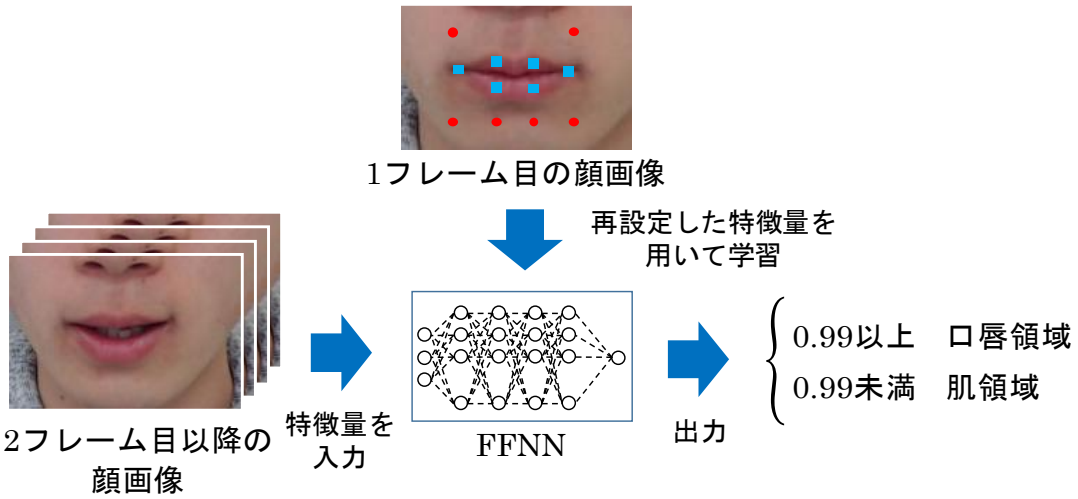


図 2.19　FFNN の学習および口唇抽出

2.5 口裂抽出処理の評価

2.5.1 評価方法

2.3.3 項で前述した「口裂抽出処理評価用データセット」100 枚を対象として，口裂抽出処理の評価を行った．具体的には，各画像に対して口裂抽出処理を実施し，オペレータ 5 名(オペレータ a～e)が口裂の抽出に関して評価を行った．このとき，2.2.1 項の内容に基づいて口裂を「左口角から右口角の間における，上唇と下唇の境界部分の領域」と定義し，以下の評価項目①，②に対して，A)～E)の 5 段階で評価した．なお，口裂が 100%抽出されていないとオペレータが判断した場合には，当該フレーム番号を記録するように指示した．

評価項目①：抽出された領域のうち，実際の口裂の領域は約何%あるか
評価項目②：実際の口裂のうち，抽出された領域は約何%あるか

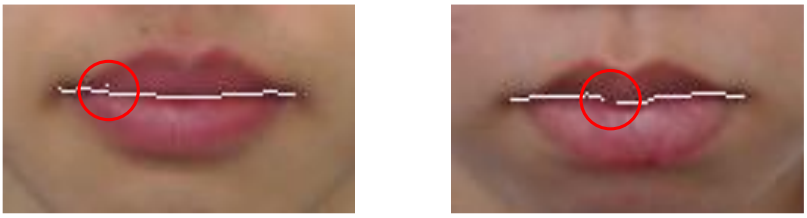
＜5 段階評価の内容＞
A) 約 80%～約 100%
B) 約 60%～約 80%未満
C) 約 40%～約 60%未満
D) 約 20%～約 40%未満
E) 約 0%～約 20%未満

2.5.2 評価結果

オペレータ 5 名による評価の結果，評価項目①はすべてのデータにおいて評価 A)が得られた．また，過剰抽出が生じたフレームにおいて，過剰抽出された画素数は数画素程度と，無視できる範囲内であることを確認した(図 2.20 (a)参照).

一方，評価項目②では，オペレータ d が被験者 b のデータを評価した場合に，10 枚の画像が B)に評価されたことを除き，すべての画像が A)に評価された．また，欠損が生じたフレームを確認したところ，欠損は数画素程度と，無視できる範囲内であることが明らかとなった(図 2.20(b)参照).

上記結果は，口裂抽出処理が精度良く口裂を抽出可能であることを示唆している.



(a)過剰抽出例　　(b)欠損例

図 2.20　評価画像例

2.6 提案手法の学習係数に関する予備検討

2.6.1 評価指標および交差検証に関して

提案手法は，1 フレーム目を学習データに用い，2 フレーム目以降の未知のデータに対して口唇抽出を行っている．このため，発話 1 回分のデータを対象として，フレーム間におけるモデルの汎用性を評価するために交差検証[13]を行った．具体的には，被験者 1 名の発話 1 回分のデータ(nフレーム)における，1 フレームを学習データとし，残りの$n-1$フレームをテストデータとしてn回の検証を行った．また，マスク画像と交差検証によって抽出した口唇領域を用い，Intersection-over-Union (IoU) [14]を算出し，1 つ目の評価指標とした．加えて，精度(Precision)および再現率(Recall)を算出し，これらを用いて F 値(F-measure)[13]を算出することで 2 つ目の評価指標とした．算出式を(2.9)式～(2.12)式に示す．

$$IoU = \frac{Area\ \ of\ \ overlap}{Area\ \ of\ \ union} \tag{2.9}$$

$$Precision = \frac{TP}{TP\ \ +\ \ FP} \tag{2.10}$$

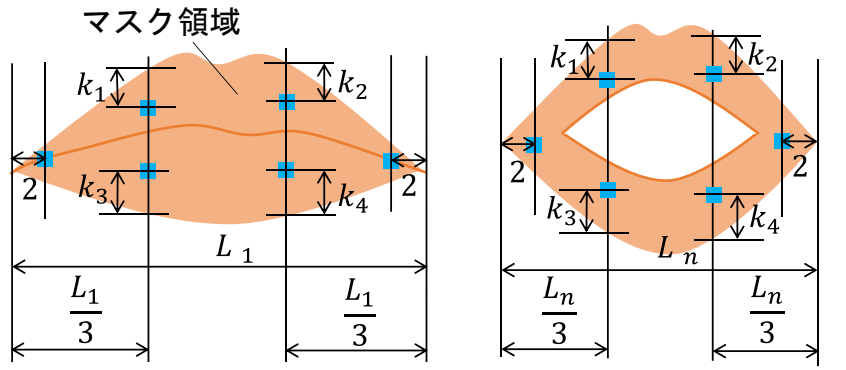
$$Recall = \frac{TP}{TP\ \ +\ \ FN} \tag{2.11}$$

$$F\text{-}measure = \frac{2\ \ \times\ \ Precision\ \ \times\ \ Recall}{Precision\ \ +\ \ Recall} \tag{2.12}$$

ここで，“$Area\ \ of\ \ overlap$”はマスク画像における口唇領域に各辺が接する矩形の領域(以下，マスク画像の矩形と表記)と抽出された口唇領域に各辺が接する矩形の領域(以下，口唇抽出画像の矩形と表記)の論理積の面積である．“$Area\ \ of\ \ union$”はマスク画像の矩形と口唇抽出画像の矩形の論理和の面積である．また，TP：口唇領域が正しく口唇領域に分類された画素数，FP：肌領域が誤って口唇領域に分類された画素数，FN：口唇領域が誤って肌領域に分類された画素数を示している．

交差検証では，1 フレーム目以外のフレームから口唇の学習画素を取得する必

要があるため，口裂抽出処理に依存しない方法で学習画素を設定した．具体的には，はじめに 1 フレーム目に 2.4.4 項および 2.4.6 項の手法を用いて学習画素を設定した．次に，設定した口唇の学習画素における中央の 4 点と 1 フレーム目のマスク画像を用いて，図 2.21(a)に示すようにk_1～k_4を算出した．最後に，n フレーム目に対して n フレーム目のマスク画像とk_1～k_4を用い，図 2.21 (b)に示すように学習画素を設定した．



(a)1フレームのマスク画像　(b) nフレーム目のマスク画像

■：口唇の着目画素

k_1, k_2　：着目画素と口唇上端の間の画素数

k_3, k_4　：着目画素と口唇下端の間の画素数

L_n　：nフレーム目のマスク画像における口唇の横幅

図 2.21　交差検証における学習画素の設定方法

2.6.2 検討内容

学習条件検討用データセットから 10 フレームおきに取得した画像(12 枚)を対象として交差検証を実施した．なお，検証では特徴量CR_3，中間層の次元数 9 次元，FFNN の層数 3 層，学習回数 25000 回を学習条件とし，学習係数を 0.10 から 0.01 刻みで減少させ，口唇を抽出した．抽出された口唇領域の IoU および F-measure の算出を行い，学習係数ごとに全フレームの平均を算出した．

2.6.3 検討結果

得られた結果を表 2.5 に示す．学習係数を減少させることで精度が良くなる傾向がみられたが，0.07 から 0.06 に変化させた場合では，IoU および F-measure が減少した．このため，0.07 が適切であると判断し，以降の検討にこの値を用いた．

表 2.5　各学習係数における IoU および F-measure の平均値

学習係数	0.06	0.07	0.08	0.09	0.10
IoU	0.8682	0.8685	0.8670	0.8669	0.8654
F-measure	0.9345	0.9350	0.9350	0.9349	0.9344

2.7 提案手法の学習回数に関する予備検討

2.7.1 検討方法

学習条件によって適切な学習回数が異なる．このため，学習条件検討用データセットから 10 フレームおきに取得した画像(12 枚)を対象とし，表 2.6 に示す学習条件におけるすべてのパターンを用いて，交差検証[13]を実施した．このとき，各学習条件における IoU および F-measure の平均値を算出した．さらに，IoU および F-measure に対して有意水準 0.05 の分散分析[15]を実施し，異なる学習回数において評価指標の平均に有意な差があるか否かを分析した．なお，有意であった場合，最も評価指標の平均が高い学習回数，有意でなかった場合は最も回数の少ない学習回数を各学習条件に適用し，これを以降の検討に用いた．

表 2.6　パラメータの検討範囲

学習係数	0.07
特徴量	SQ_n，CR_n (n: 3, 5, 7) 合計 6 パターン
レイヤー数	3 層, 4 層，ならびに 5 層
中間層の次元数	“入力層と同じ次元数”, “入力層の次元数+100 次元, “入力層の次元数 + 200 次元
出力層の次元数	1
学習回数	10000, 25000, 50000, and 75000

2.7.2 検討結果

表 2.7 に各パラメータにおいて設定された学習回数を示す．以降の検討において，パラメータごとに選定した学習回数のモデルを使用した．

表 2.7　各パラメータにおける適切な学習回数

特徴量	層数	中間層の次元数		
		入力層と同じ	入力層+100	入力層+200
SQ_3	3	25000	25000	25000
	4	10000	10000	10000
	5	25000	25000	25000
SQ_5	3	10000	10000	10000
	4	10000	10000	10000
	5	10000	10000	10000
SQ_7	3	10000	10000	10000
	4	10000	10000	10000
	5	10000	10000	10000
CR_3	3	50000	50000	50000
	4	25000	10000	10000
	5	50000	25000	25000
CR_5	3	10000	10000	10000
	4	10000	10000	10000
	5	25000	10000	10000
CR_7	3	10000	10000	10000
	4	10000	10000	10000
	5	25000	10000	10000

2.8 提案手法の特徴量と FFNN の層数に関する検討

2.8.1 検討方法

2.6 節および 2.7 節の検討結果を踏まえ，学習条件検討用データセット(111 枚)を対象に表 2.8 に示す学習条件におけるすべての組み合わせを用いて，交差検証[13]を実施した．このとき，学習条件ごとに IoU および F-measure の平均値を算出した．

表 2.8　パラメータの検討範囲

学習係数	0.07
特徴量	SQ_n，CR_n (n: 3, 5, 7) 合計 6 パターン
レイヤー数	3 層, 4 層，ならびに 5 層
中間層の次元数	“入力層と同じ次元数”, “入力層の次元数+100 次元, “入力層の次元数 + 200 次元
出力層の次元数	1
学習回数	選定された学習回数(表 2.7 参照)

2.8.2 検討結果

評価指標の平均値を比較した結果を表 2.9 および表 2.10 に示す．特徴量SQ_3における FFNN の層数が 3 層の場合に F-measure の平均値が最も高く，また，FFNN の層数が 4 層の場合に IoU の平均値が最も高い結果を得た．

表 2.9　各パラメータにおける F-measure の平均値算出結果

特徴量	層の数	中間層の次元数			平均値
		入力層と同じ	入力層+100	入力層+200	
SQ_3	3	0.9351	0.9353	0.9353	0.9353
	4	0.9305	0.9337	0.9338	0.9326
	5	0.9306	0.9298	0.9298	0.9301
SQ_5	3	0.8959	0.8952	0.8952	0.8954
	4	0.8849	0.8834	0.8830	0.8838
	5	0.8816	0.8832	0.8823	0.8824
SQ_7	3	0.8436	0.8428	0.8428	0.8430
	4	0.8190	0.8173	0.8174	0.8179
	5	0.8167	0.8100	0.8096	0.8121
CR_3	3	0.9311	0.9309	0.9309	0.9310
	4	0.9304	0.9247	0.9251	0.9267
	5	0.9154	0.9282	0.9282	0.9239
CR_5	3	0.8930	0.8934	0.8935	0.8933
	4	0.8912	0.8886	0.8882	0.8893
	5	0.8642	0.7225	0.7748	0.7872
CR_7	3	0.8482	0.8462	0.8460	0.8468
	4	0.8301	0.8200	0.8190	0.8230
	5	0.7630	0.8235	0.8431	0.8099

※赤字：平均値が最大のセル，橙色セル：選定された特徴量と層数

表 2.10　各パラメータにおける IoU の平均値算出結果

特徴量	層の数	中間層の次元数			平均値
		入力層と同じ	入力層+100	入力層+200	
SQ_3	3	0.8743	0.8744	0.8743	0.8743
	4	0.8766	0.8760	0.8760	0.8762
	5	0.8613	0.8590	0.8587	0.8597
SQ_5	3	0.8122	0.8115	0.8114	0.8117
	4	0.7959	0.7928	0.7921	0.7936
	5	0.0843	0.0425	0.0442	0.0570
SQ_7	3	0.7347	0.7333	0.7331	0.7337
	4	0.6889	0.6860	0.6858	0.6869
	5	0.6867	0.6749	0.6741	0.6786
CR_3	3	0.8625	0.8622	0.8621	0.8622
	4	0.8608	0.8733	0.8732	0.8691
	5	0.8333	0.8539	0.8540	0.8471
CR_5	3	0.8063	0.8066	0.8067	0.8066
	4	0.8035	0.7980	0.7973	0.7996
	5	0.7322	0.6654	0.7106	0.7028
CR_7	3	0.7401	0.7357	0.7352	0.7370
	4	0.6983	0.6761	0.6746	0.6830
	5	0.5965	0.7177	0.7355	0.6832

※赤字：平均値が最大のセル，橙色セル：選定された特徴量と層数

2.9 提案手法の中間層の次元と学習条件に関する検討

2.9.1 検討方法

2.8 節で評価指標の値が高かった学習条件の IoU および F-measure の平均値に対して有意水準 0.05 の分散分析[15]を実施し，異なる中間層の次元において評価指標の平均に有意な差があるか否かを分析した．この結果，層の数が 3 の場合において有意ではなかったため，最も次元数の少ない 9 次元を採用した．また，層の数が 4 層の場合においては，F-measure に有意な差が認められたため，F-measure が最も高い 209 次元を採用した．学習条件を定めるために，上記の 2 種類の学習条件を用いて，学習条件検討用データセット(111 枚)に対して提案手法を実施し，口唇領域を抽出した．抽出結果画像から IoU および F-measure の平均値を算出した．

2.9.2 検討結果

検討結果を表 2.11 に示す．特徴量SQ_3，FFNN の層数が 3 層，中間層の次元数が 9 次元，学習回数 25000 回，学習係数 0.07 の場合において，IoU および F-measure は最大の値を得たため，この学習条件が提案手法において有用であると判断した．

表 2.11　2 種類の学習条件における IoU と F-measure の比較結果
(被験者: a，特徴量 ：SQ_3，学習係数 ：0.07).

	学習条件①	学習条件②
IoU	0.8808	0.8773
F-measure	0.9324	0.9254

学習条件①：中間層の次元：9, FFNN の層数：3, 学習回数：25000 回

学習条件②：中間層の次元：209, FFNN の総数：4, 学習回数：10000 回

2.10 提案手法における口唇抽出精度の比較評価

2.10.1 検討方法

口唇抽出精度評価用データセットを対象として提案手法，ファジィ推論に基づいた従来手法[1]，ならびに輝度勾配に基づいた従来手法[6]を適用し，口唇領域を抽出した．このとき，従来手法と提案手法は各被験者における IoU および F-measure の平均値を算出し，評価を行った．また，輝度勾配に基づいた従来手法の場合は，IoU の平均値のみ算出を行った．なお，輝度勾配に基づいた従来手法は，機械学習のオープンソースライブラリである Dlib[6]を用い，公式で配布されている顔器官検出の学習済みモデル(IBUG 300-W データセット[16-18]で学習されたもの) [6]を使用して特徴点を抽出した．

2.10.2 比較結果および考察

提案手法およびファジィ推論に基づいた従来手法[1]の F-measure の平均値を被験者ごとに比較した結果を表 2.12～表 2.14 に示す．ここで，ファジィ推論に基づいた従来手法を従来手法と表記する．提案手法は従来手法と比較してすべての被験者において F-measure が向上し，最大で 0.1456，平均で 0.0809 向上する結果を得た．提案手法，ファジィ推論に基づいた従来手法，ならびに Dlib を用いた手法における IoU の平均値を算出した結果を表 2.15～表 2.17 に，提案手法と Dlib の口唇抽出結果例を図 2.22 に示す．「眼鏡無しデータ」を用いた場合，表 2.15 に示すように提案手法は従来手法と比較して良好な結果を得た．また，提案手法と Dlib の結果を比較した場合，被験者によって IoU の大小関係が変動するが，全被験者の平均値を比較すると同程度の精度であることがわかる．また，提案手法と Dilb の口唇抽出結果画像(図 2.22 参照)を比較すると，Dlib では口唇下端の部分に抽出できていない領域を認めた．一方，提案手法は，同フレームにおいても精度良く口唇の形状を抽出していることがわかる．さらに，表 2.16 および表 2.17 に示す「眼鏡ありデータ」および，「サングラスありデータ」の IoU を比較すると，提案手法の IoU が最も高いことが明らかとなった．特に，表 2.17 に示す結果では，Dlib の IoU が他の手法と比較して，著しく低いことがわかる．これは，サングラスをかけることで顔画像中の特徴点が一部欠損し，口唇の特徴点抽出に影響したものと考える．これに対し，提案手法は，顔の状態に影響されることなく口唇領域を抽出可能であるため，安定した精度を得ている．

上記結果は，提案手法が少ない学習データで精度よく口唇形状を抽出可能であり，かつ他の手法と比較して同程度またはそれ以上の精度を有することを示唆している．

以上のことは，本論文で提案する口唇抽出手法が口唇領域の抽出に有用であることを示唆している．

表 2.12　提案手法と従来手法（ファジィ推論に基づいた従来手法)の F-measure 算出結果（眼鏡無しデータ）

被験者	a	b	c	d	e	平均
提案手法	0.9324	0.9478	0.9000	0.9426	0.9160	0.9278
従来手法	0.8119	0.8432	0.8736	0.7969	0.8539	0.8359
差分	0.1205	0.1046	0.0264	0.1456	0.0621	0.0919

※赤字：各被験者における最大値

※橙セル：表 2.12～2.14 のうち最大の差分

表 2.13　提案手法と従来手法(ファジィ推論に基づいた従来手法)の F-measure 算出結果（眼鏡ありデータ）

被験者	a	b	c	d	e	平均
提案手法	0.9165	0.9346	0.9195	0.9463	0.9064	0.9247
従来手法	0.8175	0.8661	0.8911	0.8081	0.8591	0.8484
差分	0.0990	0.0685	0.0284	0.1382	0.0474	0.0763

※赤字：各被験者における最大値

表 2.14　提案手法と従来手法(ファジィ推論に基づいた従来手法)の F-measure 算出結果（サングラスありデータ）

被験者	a	b	c	d	e	平均
提案手法	0.8928	0.9528	0.8928	0.8978	0.9148	0.9102
従来手法	0.8231	0.8549	0.8899	0.7744	0.8355	0.8356
差分	0.0698	0.0978	0.0029	0.1234	0.0793	0.0746

※赤字：各被験者における最大値

表 2.15　各手法における IoU 算出結果（眼鏡無しデータ）

被験者	a	b	c	d	e	平均
提案手法	0.8808	0.9184	0.9229	0.9067	0.8893	0.9036
Dlib	0.9228	0.8866	0.8385	0.9314	0.9336	0.9026
従来手法	0.7294	0.7577	0.8954	0.6595	0.7422	0.7568

※赤字：各被験者における最大値，青字：各被験者における最小値

表 2.16　各手法における IoU 算出結果（眼鏡ありデータ）

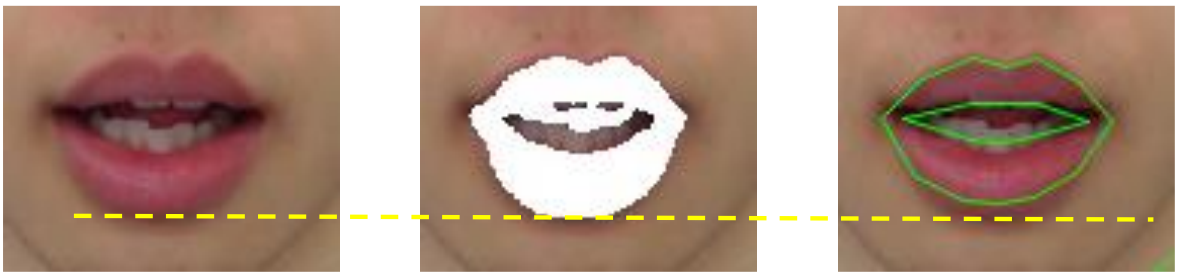
被験者	a	b	c	d	e	平均
提案手法	0.8913	0.9105	0.8805	0.9256	0.9126	0.9041
Dlib	0.8821	0.8595	0.8304	0.9165	0.9099	0.8797
従来手法	0.7353	0.8454	0.8968	0.6891	0.8163	0.7966

※赤字：各被験者における最大値，青字：各被験者における最小値

表 2.17　各手法における IoU 算出結果（サングラスありデータ）

被験者	a	b	c	d	e	平均
提案手法	0.8875	0.9286	0.8885	0.9153	0.9271	0.9094
Dlib	0.3842	0.1139	0.8030	0.0436	0.2934	0.3276
従来手法	0.6897	0.7953	0.8858	0.6432	0.6998	0.7428

※赤字：各被験者における最大値，青字：各被験者における最小値



(a)元画像　　(b)提案手法　　(c)Dlib

図 2.22　被験者 c の 30 フレーム目における口唇抽出結果(眼鏡無しデータ)

2.10.3 光源色と a*値の関連についての考察

本論文では，一般的な白色蛍光灯下(500～900lx)において取得した顔画像データを対象として検討を加えた．一方，実用面を考慮した場合，データ取得環境は照明条件や日光の光量などによって変化するため，同じ色の物体を撮影した場合であっても，画像中の色情報は異なる場合がある．本論文では，異なる光源色の照明環境下における口唇抽出精度の比較実験を行っていないため，入力特徴量の L*a*b*色空間の特性を踏まえて光源色に関する考察を行った．

提案手法は 1 フレーム目の画像から特徴量を複数取得し，これらを入力して学習した FFNN モデルを使用することで口唇領域の抽出を行う手法である．加えて，特徴量として 3×3 画素の領域から色情報を取得しているため，抽出対象の動画データにおける口唇領域と肌領域の間の「領域内における色情報のパターンの違い」および「色差」の特徴に基づいて口唇領域の抽出を行うことが可能である．「領域内における色情報のパターンの違い」に着目すると，光源色が異なる場合であっても，3×3 画素内の相対的な色情報のパターンは変化しにくいと推察される．

一方，口唇領域と肌領域の色差に関しては，光源色の影響を受けると考える．特に，提案手法は a*値を特徴量として使用しているため，a*値のみを用いて表現が困難な色(主に黄色系統と青色系統間)の違いを識別することが困難である．また，色の付いた光源色を使用することで口唇領域と肌領域の色差が小さくなる場合，口唇領域の抽出が困難になる．したがって，提案手法が利用可能な光源色は，白色の一般的な蛍光灯下，または口唇領域と肌領域の色差を識別可能な程度の明るさの下であると推察される．会議の環境において青色や赤色などという光源色が使用されることは少ないため，実用する上で光源色や環境光による影響は小さいと考える．なお，本論文において照明の色に関する検討は行われていないため，屋内で使用される複数種類の照明の下で撮影された画像を対象として口唇抽出精度を評価し，手法の改良を行うことで精度の向上を図ることが可能であると考える．

2.10.4 提案手法の処理時間に関する考察

2.10.2 項で前述した通り，Dlib[6]と比較して提案手法は同程度または高い精度で口唇領域を抽出可能である．特に，顔の部位が手などで遮蔽されている場合，Dlib よりも提案手法の口唇抽出精度は高い．一方，表 2.18 に示すように処理時間を比較すると，提案手法の処理速度は Dlib よりも遅いことがわかる．したがって，実用化を考えた場合，顔の部位が遮蔽されていることに起因して Dlib の使用が困難な条件下において，提案手法を使用し口唇の形状情報を良好に取得することが望ましいと考える．すなわち，提案手法と Dlib を併用し，顔の状態に応じて手法を使い分けることによって，高速かつ高精度な口唇の領域抽出が可能になると考える．

表 2.18　提案手法と Dlib の処理時間比較

手法	処理時間（フレームレート：30fps）
提案手法	1 フレーム目の学習時間：約 18.52 秒 2 フレーム目以降の口唇抽出時間：約 16.80 秒 / 30 フレーム
Dlib	約 1.26 秒 / 30 フレーム

2.11 まとめ

本論文では，被験者 5 名の顔画像を対象とし，明度分布の特徴を用いた口裂の抽出処理および機械学習の一手法である順伝播型ニューラルネットワーク(FFNN)を用いた口唇形状の抽出手法を提案した．得られた成果を以下にまとめる．

(1) 口唇領域における明度分布は，口裂の抽出に有用な特徴となり得ることを明らかにした．

(2) 顔画像における a*値は，FFNN を用いた口唇領域の抽出に有用な特徴量であることを明らかにした．

(3) 提案手法は，ファジィ推論に基づいた従来手法[1]と比較して口唇を精度よく抽出可能であることを明らかにした．

(4) 提案手法は，輝度勾配に基づいた従来手法(Dlib[6]を使用)と比較し，少ない学習データを用いて同等の精度もしくは，それ以上の精度で口唇を抽出可能であることを明らかにした．

第 2 章 参考文献

[1] 佐藤慶幸, 成田純一, 景山陽一, 西田眞:「口唇の形状情報を用いた口唇領域自動抽出処理の改善」, 電学論 C, Vol.130, No.5, pp.873−881 (2010)

[2] A. Azeem, M. Sharif, J.H. Shah, and M. Raza : "Hexagonal Scale invariant feature transform (H-SIFT) for facial feature extraction", Journal of Applied Research and Technology, Vol.13, No.3, pp.402−408 (2015)

[3] S. Jahanbin, A.C. Bovik, and H. Choi : "Automated facial feature detection from portrait and range images". In Image analysis and interpretation, SSIAI.2008 IEEE southwest symposium on Image Analysis and Interpretation, DOI:10.1109 / SSIAI.2008.4512276 (2008)

[4] A. Jackson, M. Valstar, and G. Tzimiropoulos : "A CNN Cascade for Landmark Guided Semantic Part Segmentation", ECCV 2016 Workshops, arXiv:1609.09642 (2016)

[5] U. Güçlü, Y. Güçlütürk, M. Madadi, S. Escalera, X. Baró, J. González, R. van Lier, and J.A.M. van Gerven : "End-to-end semantic face segmentation with conditional random fields as convolutional", recurrent and adversarial network. arXiv:1703.03305 (2017)

[6] V. Kazemi, and J. Sullivan : "One millisecond face alignment with anensemble of regression trees", 2014 IEEE Conference on Computer Vision and Pattern Recognition, DOI:10.1109/CVPR.2014.241, Columbus, OH, USA (2014)

[7] 岡谷貴之 :「深層学習」, 講談社 (2015)

[8] 小野尊睦, 飯塚忠彦, 吉武一貞 :「口腔外科学 改訂 7 版」, 金芳堂 (2010)

[9] 日本色彩学会 :「新編 色彩科学ハンドブック 第 3 版」, 東京大学出版会 (2011)

[10] 高木幹雄, 下田陽久監修 :「新編 画像解析ハンドブック」, 東京大学出版会 (2004)

[11] Logicool :「C922 PRO HD ストリームウェブカメラ」, https://www.logicool.co.jp/ja-jp/products/webcams/c922n-pro-stream-webcam.960-001262.html (Access 2021/12/15)

[12] Logicool サポート :「C922 Pro Stream Webcam 技術仕様」, https://support.logi.com/hc/ja/articles/360023462473 (Access 2021/12/15)

[13] 竹村彰通監訳他 :「機械学習 データを読み解くアルゴリズムの技法」, 朝

倉書店（2017）

[14] 株式会社フォワードネットワーク監修，藤田一弥，高原歩：「実装ディープラーニング」， オーム社（2016）

[15] 星野満博，西崎雅仁，「数理統計の探求　-経営的問題解決能力の開発と論理的思考の展開-」，晃洋書房（2012）

[16] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic：“300 Faces In-The-Wild Challenge: Database and results”, Image and Vision Computing (IMAVIS), Vol.47, pp.3–18 (2016).

[17] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic：“A Semi-automatic Methodology for Facial Landmark Annotation”, Proceedings of IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR-W), 5th Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2013). Oregon, DOI: 10.1109/CVPRW.2013.132 (2013)

[18] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic：“300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge”, Proceedings of IEEE Int'l Conf. on Computer Vision (ICCVW), 300 Faces in-the-Wild Challenge (300W), DOI: 10.1109/ICCVW.2013.59 (2013)

第 3 章　発話区間の抽出手法に関する検討

3.1 はじめに

議事録自動作成システム[1-3]は，音声認識技術を用いて議事録を自動作成する．このとき，音声区間を検出(以降，音声区間検出と表記する)し，これに対して音声認識処理を行う．音声区間検出とは，一般的にマイクロフォンから入力された音から，音声が存在する区間のみを抽出する手法として定義される[4]．すなわち，音声区間の検出は，音声波形の情報を用いて「ノイズまたは無音」と「人の音声」の区間を検出する処理を示す．音声区間検出の目的としては，①音声の生じた区間のみを対象として処理時間を削減すること，②音声区間以外の領域に生じた雑音を除去し，音声認識精度の向上を図ることが挙げられる．一般的な音声区間検出法は，音声の情報のみを使用して音声区間を検出しているため，検出した音声区間の情報と発話者を対応させることが困難であると考える．音声区間と発話者を対応させる場合，声門の登録が必要[5, 6]であるため，議事録自動作成システムにおける「事前の準備が必要である課題」を解決することは難しい．

一方，音声情報に加えて画像情報を使用することは，シンプルな処理で発話者の判別を可能にすると考える．すなわち，音声と画像中における唇の動きの発生タイミングに基づき，各人物の発話区間を抽出する．このため，画像中の唇が動いている人物における唇の座標と，音声の発生タイミングを紐付けることが可能である．したがって，人物の声門を事前に登録することなく発話区間を抽出し，かつ位置情報から発話者を判別することができる．

類似研究として，画像情報と音声情報を用いた発話者の判別手法が提案されている[7, 8]．この手法は，CNN および LSTM を用いることで，発話者の判別が可能である．さらに，被験者ごとの発話者度合をフレームごとに算出することが可能であるため，発話区間の抽出にも利活用可能であると考える．しかしながら，発話区間の抽出に関する精度の評価は行われていない．また，CNN や LSTM の学習には約 606 時間の動画データが使用されているため，膨大な量の学習データを必要とする課題を有している．

そこで本論文では，議事録自動作成システムにおける発話者判別手法の構築を目的とし，その前処理として発話区間抽出手法に関する検討を行った．具体的には，画像情報と音声情報を使用して発話区間をそれぞれ抽出し，抽出された 2 つの発話区間の発生タイミングに基づいて，最終的な発話区間の決定を行った．本発話区間抽出手法(以降，提案手法と表記する)は，教師データが不要であるため，類似研究[7, 8]における課題を解決可能である．本章の最後に，提案手法と類似研究[7, 8]における発話区間抽出精度の比較および評価を行った．

3.2 使用データ

本論文では対面における会議を想定し，全方位カメラを使用してデータの取得を行った．全方位カメラを使用することで，360°全方位の画像を取得可能であるため，会議などのように向かい合って会話をする環境において，効率良く顔画像が取得できる．

3.2.1 データ取得環境

データ取得環境を図 3.1 に示す．蛍光灯による照明の下(顔周辺照度：450～900lx)，全方位カメラ(THETA V : RICOH 社製) [9]およびマイクロフォン(TA-1 : RICOH 社製) [10]を用い，被験者 14 名が表 3.1 に示す 11 種類の文章(以下，文章 1～文章 11 と表記する)を 2 回ずつ発話する様子を撮影・録音し，発話動画データを取得した．各文章の発話動画データ取得前に，被験者には 1 回程度の発話の練習をしてもらった．また，被験者に発話内容を指示するために，11 種類の文章を 1 文ずつディスプレイに表示した. 取得した発話動画データは, RICHO THITA 基本アプリ[11]を用いて全天球の動画(3,840×2,160，30fps)に変換し，これを検討に使用した(図 3.2 参照)．なお，11 種類の文章は，Web のニュース記事から抜粋したものを使用している．

＜データ取得条件＞

- 被験者は 20 代の健常者 14 名（A～N，男性：8 名，女性 6 名，イーストアジアン)
- 被験者とカメラおよびマイク間の距離は約 50cm
- 発話開始時および終了時には，口を閉じることを指示

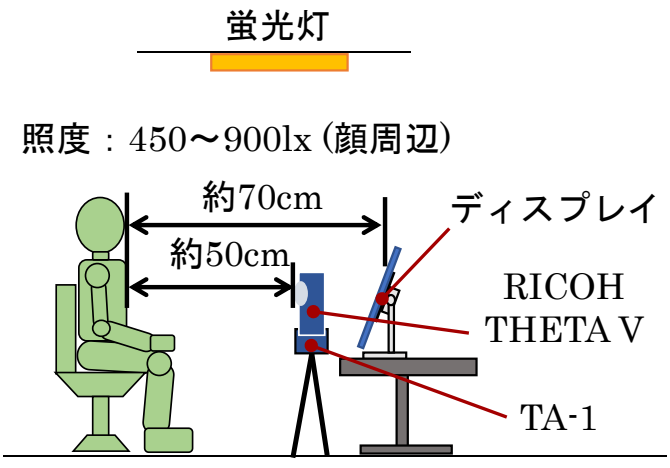


図 3.1　データ取得環境

表 3.1　使用した 11 種類の文章一覧[12-17]

文章名	内容
文章 1	現在，一般的な AI は，ディープラーニング形式といい，学習精度を高めるためにネットにつないで学習させ，エンジニアによる調整作業が必要で，通信コストと人件費がかかっていた.
文章 2	ユーザー側で使っていたデバイスごとに学習したものを，クラウドで吸い上げて統合すると，平均化された賢い AI ができる．それをわたしたちはストックしていき，使えるようにしていくのが，最終的な目標.
文章 3	少子化や人口減少などについて最新技術による支援を受け，2030 年を見据えて持続可能な社会の構築を目指す.
文章 4	両者は今後，キャッシュレス化による市域内経済の活性化，困っている人と助ける人をつなぐマッチングシステムの開発，AI やビッグデータを使ったインフラの監視，IT スキル育成による定住促進などに取り組む.
文章 5	実はそんな人ほど用心すべきです．特に，IT リテラシーに自信があるひとり暮らしの人は，怪しいサイトを見つけても相談する相手がおらず，思わぬ落とし穴にはまる可能性があります.
文章 6	物理世界に対する活用サービスのためには，データ取得のための IoT 技術と，実際に動いて仕事をするロボット技術との連携が重要になる.
文章 7	ロボットに求められる役割も変わっており，完全自動化は難しい．生産現場でも結局人手集約が必要になっているのが現状だ.
文章 8	iPS 細胞から様々な細胞を作り，血液提供者の遺伝子や健康に関する情報と組み合わせて，病気の解析や個別化医療の開発に役立てる.
文章 9	「iPS 細胞を使って何千人規模の研究が培養皿の中でできる可能性がある．これまでできなかった個別化医療の実現につながれば」と話した.
文章 10	プラスチック資源循環や，PET ボトルの 100%有効利用に向けベクトルを合わせ，業界を挙げて取り組んでいくことを決めている.
文章 11	タバコのポイ捨て禁止条例のようなものが必要なのか．生活者の意識の変化を促すとともに，業界としての抜本的な対策も必要になってくるかもしれない.

図 3.2　取得した発話動画データのフレーム例

3.2.2 使用機材について

本論文では，全方位カメラ(THETA V：RICOH 社製)[9]およびマイクロフォン(TA-1：RICOH 社製)[10]を用いてデータを取得している．全方位カメラには，2 つの超広角レンズが搭載されており，カメラの上下左右 360°の空間を撮影することができる．また，最大で 4K(3840×1920 画素)の大きさの動画像を取得することができる．

一方，RICOH THETA V 専用のマイクロフォン TA-1 は，RICOH THETA V 内臓のマイクロフォンと比較し，録音可能な音の周波数帯が広い特徴を有している．また，ウインドスクリーンを装着することで，風切り音などのノイズを低減し，録音することができる．各機材の外観を図 3.3 に示す．また，各機材の主な仕様を表 3.2 および表 3.3 にそれぞれまとめる．

(a)　THETA V　　(b)　TA-1

図 3.3　使用機材の外観[9, 10]

表 3.2　THETA V の主な仕様[9]

外形・寸法	約 45.2mm(幅) × 130.6mm(高さ) × 22.9mm (奥行き)
質量	約 1210g
静止画解像度	5376 × 2688
動画解像度 / フレームレート	4K：3840×1920 / 29.97fps 2K：1920×960 / 29.97fps

表 3.3　TA-1 の主な仕様[10]

外形・寸法	約 45.3mm(幅) × 105.2mm(高さ) × 37mm(奥行き)
質量	約 63g
録音形式	バックエレクトレット・コンデンサ型
指向特性	単一指向性×4ch，（アンビソニックス）

3.2.3 データセット

被験者 14 名が 11 文を 2 回ずつ発話した動画データ(308 データ)を用いて，2 つのデータセットを作成した．データセットの定義を以下に示す．

i) 検討用データセット:

被験者 A～F が 11 文を 1 回ずつ発話した動画データ(66 データ)．これらのデータは，パラメータの検討に使用した．

ii) 評価用データセット:

被験者 14 名が 11 文を 2 回ずつ発話した動画データ(308 データ)．これらのデータを用いて，提案手法の評価を行った．

3.2.4 正解の発話区間の設定

発話動画データの各フレームを発話フレームまたは無発話フレームに分類し，正解ラベルデータを作成した．具体的には，発話開始フレームおよび発話終了フレームを以下の定義に基づいて，目視で設定した．

＜定義＞

・発話開始フレーム：

発話開始時のフレームにおいて，口を開き始める直前の口を閉じたフレーム．

・発話終了フレーム：

発話終了時，口を閉じたフレーム．

次に，発話開始フレームおよび発話終了フレームの間のフレームを正解の発話区間とし，発話区間内のフレームを発話フレームとした．また，発話区間以外の区間を無発話区間として設定し，無発話区間内のフレームを無発話フレームとした．正解の発話区間の設定例を図 3.4 に示す．

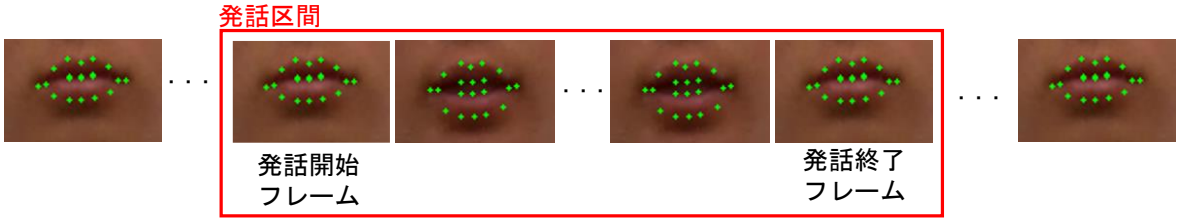


図 3.4　正解の発話区間の設定例

※緑点：口唇における特徴点（3.3.2 項で後述する）

3.3 提案手法

3.3.1 提案手法の概要

音声が生じている，かつ口唇が動いているフレームを発話区間として抽出した．発話区間抽出手法の手順を①～⑨に示す．また，発話区間抽出手法(提案手法)の流れを図 3.5 に示す．

① 取得した発話動画データの各フレームに対して，顔器官の特徴点検出を行い，顔領域を検出する．
② 検出した顔器官の特徴点における口唇の特徴点を用いて，口内領域の縦幅を抽出する．さらに，発話動画データにおける口内領域の縦幅の時系列変化を算出する．
③ ノイズの除去を目的とし，口内領域の縦幅の時系列変化に対して平滑化処理を行う．
④ 鼻の特徴点を使用して発話区間抽出のための閾値を算出する．
⑤ ③の処理で算出した時系列変化および④の閾値を使用し，口内領域の縦幅が大きく動いている区間を発話区間として抽出する．
⑥ ⑤の処理において，抽出されない発話区間を再抽出する．
⑦ 発話動画データを対象とし，音声の特徴量の時系列変化を取得する．
⑧ 音声の特徴量の時系列変化を用いて，発話区間を抽出する．
⑨ 口唇の動きを用いて抽出された発話区間のうち，音声を用いて抽出された発話区間を含む区間を発話区間として抽出する．

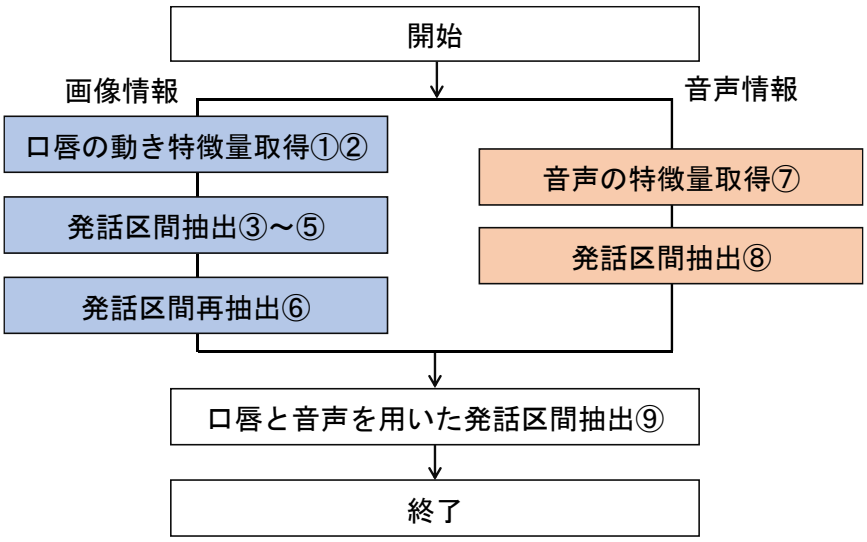


図 3.5　発話区間抽出手法(提案手法)の概要

3.3.2 顔器官検出

本論文では処理のリアルタイム性を考慮し，オープンソースライブラリの Dlib[18]による顔器官検出を行い，口唇の形状を取得した．このとき，Dlib に搭載されている顔器官検出機能および，公式で配布されている学習済みのモデル (IBGU　300-W データセット[19-21]を用いて学習されたモデル) [22]を使用した．図 3.6 に示すように，Dlib を用いることで，顔の各器官を 68 点の特徴点として検出可能である．また，口唇の形状を 20 点の特徴点として取得することができる．

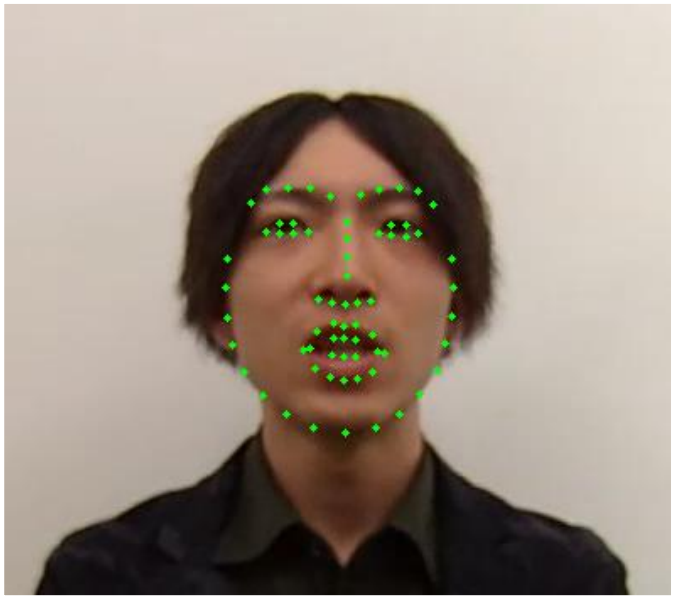


図 3.6　Dlib を用いて取得された顔器官における 68 点の特徴点例

3.3.3 各フレームに対する顔器官検出処理

取得した発話動画の各フレームを対象とし，Dlib を用いて顔領域の検出処理を行った．このとき，一般的な Web カメラによって取得されるフレームの大きさ(640×480 画素)と比較して，使用データにおける各フレームの画像サイズ(3840×2160 画素)が大きいことに起因し，顔器官検出の処理速度が低下する．このため，顔器官検出処理を行う際に，ひとつ前のフレームにおける顔の位置情報を使用して，処理の効率化を図った．処理手順を①～③に示す．

① 1 フレーム目の画像に対して，Dlib を用いた顔器官検出を行った．
② 図 3.7 に示すように，顔器官検出結果における顔左端，顔右端，ならびに顔下端の特徴点間距離(L1，L2)を用いて，顔周辺の領域(以下，トリミング領域と表記する)をトリミングした．このとき，トリミング領域の各頂点の座標を記録した．
③ 2 フレーム目以降は，前のフレームにおいて記録したトリミング領域の各頂点の座標を用いて，顔器官検出対象のフレームをトリミングした(図 3.8 参照)．さらに，トリミング領域に対して Dlib を用いた顔器官検出処理を行い，顔器官の各特徴点を検出した．

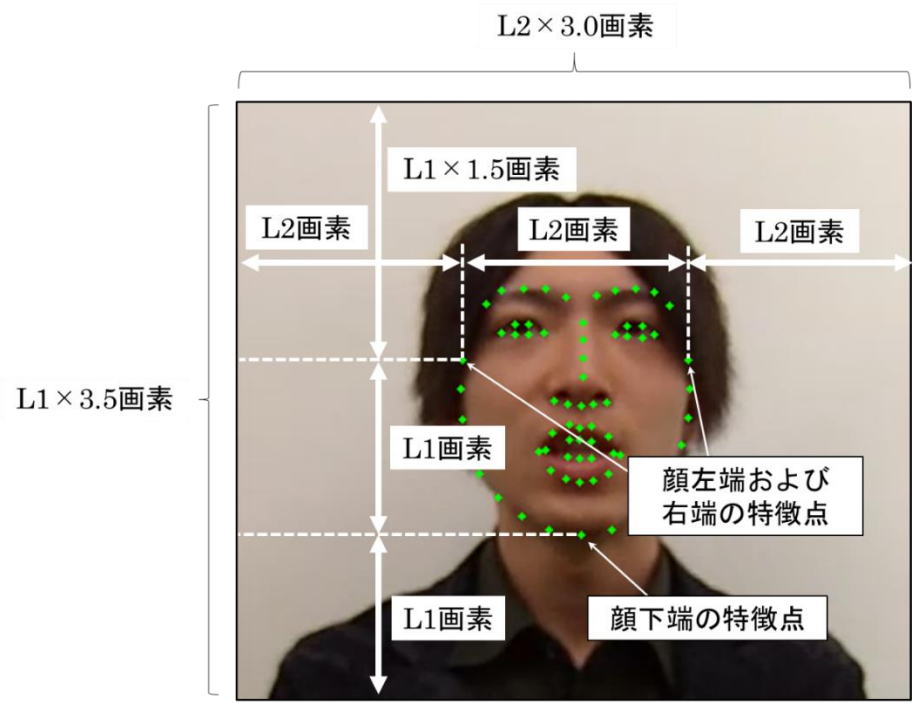


図 3.7　トリミング領域の設定例

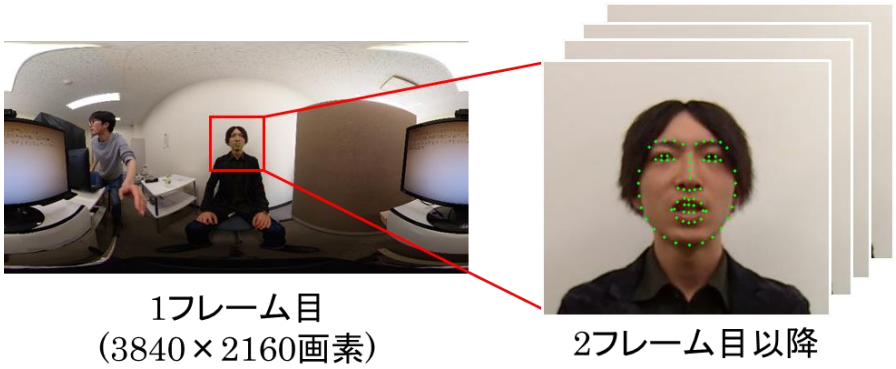


図 3.8　顔検出結果例

3.3.4 口唇の動き特徴量の取得

発話に伴って口内領域の縦幅に生じた動きは，発話区間抽出に有用であると考える．このため，Dlib を用いて検出した顔の特徴点のうち，口内領域上端および口内領域下端の特徴点を用いて，口内領域の縦幅を抽出した(図 3.9 参照)．さらに，すべてのフレームにおける口内領域の縦幅を用いて，口内領域の縦幅の時系列変化(以降，口唇の動き特徴量と表記する)を取得した．取得結果例を図 3.10 に示す．

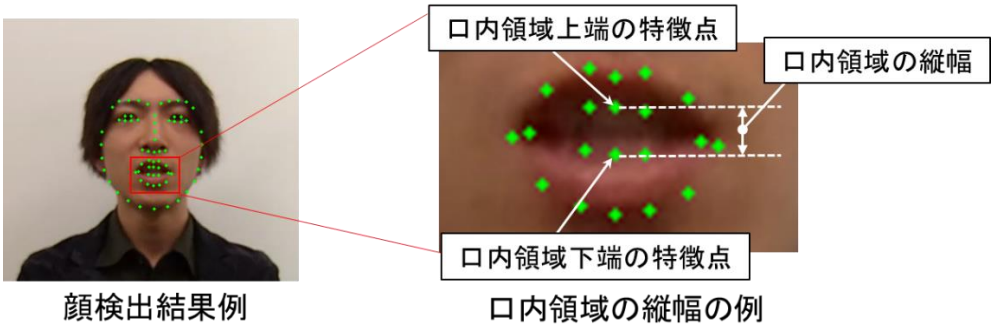


図 3.9　口内領域の縦幅抽出例

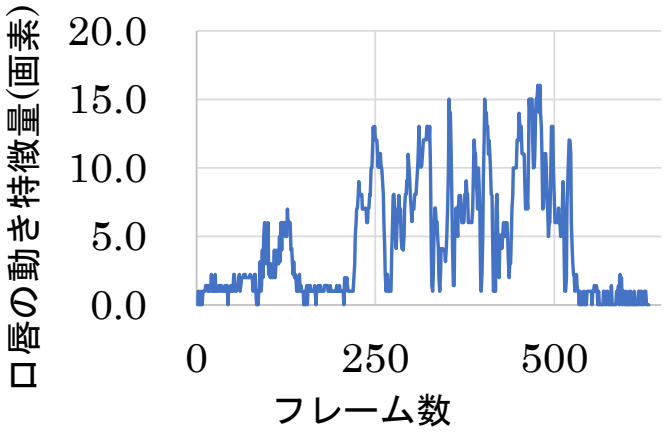
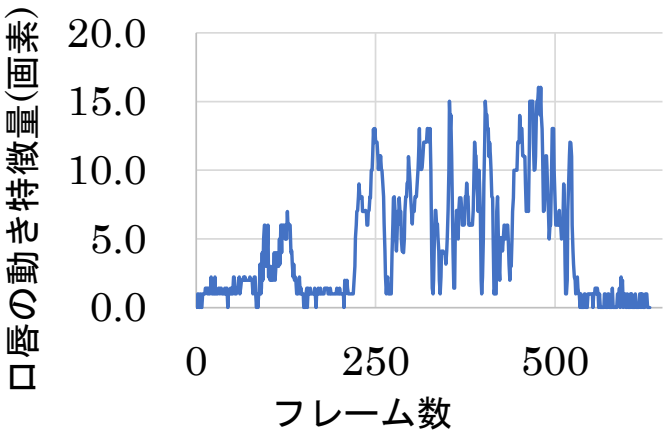


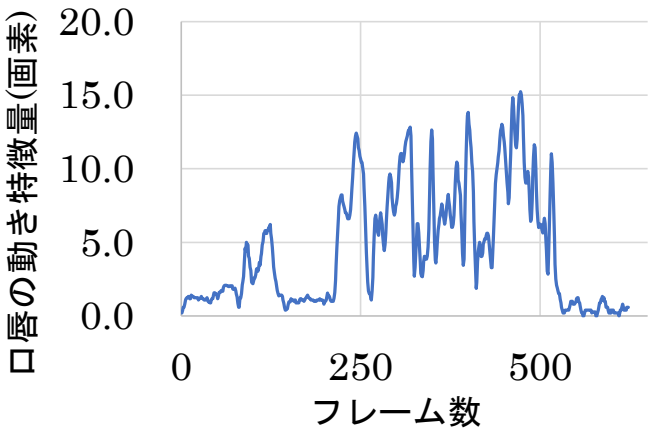
図 3.10　口内領域の縦幅の時系列変化(口唇の動き特徴量)取得結果例

3.3.5 平滑化処理

撮影時における照明の微細な変動に起因し，口唇の動き特徴量にノイズが生じる．そこで，ノイズの除去を目的として，口唇の動き特徴量に対してフィルタサイズ 5 の移動平均フィルタを適用し，平滑化処理を行った．平滑化処理前後の口唇の動き特徴量を図 3.11 に示す．なお，フィルタサイズに関しては 3.4 節の検討結果に基づいて良好な値を設定している．



(a) 平滑化処理前



(b) 平滑化処理後

図 3.11　平滑化処理結果例

3.3.6 口唇の動き特徴量を用いた発話区間の抽出

無発話区間と比較して発話区間は口唇の動きが大きくなると考える．このため，発話動画データの任意の区間における口内領域の縦幅の動きの大きさを用いて，発話区間の判別を行った．具体的には，はじめに任意のnフレーム目およびその前後 7 フレーム(合計 15 フレーム)を対象とし，口内領域縦幅の最大値および最小値を取得した．次に，任意のnフレーム目における最大値から最小値を減算した値(Lf)を算出し，Lfの時系列変化を取得した．Lfの算出方法の概要を図 3.12 に示す．ここで，最大値と最小値の取得範囲(15 フレーム)は，3.4 節の検討結果に基づいて良好な値を設定した．さらに，図 3.13 に示す鼻の横幅および(3.1)式を用いて発話区間抽出のための閾値Thを算出した．ここで，Nwは発話区間抽出対象の被験者における 1 つの動画データの鼻の横幅の中央値を表す．なお，閾値Thの算出式における係数は，3.5 節の検討結果に基づいて設定した．最後に，図 3.14 に示すように，Lfが閾値Th以上のフレームを発話区間のフレームとして抽出した．

$$Th = 0.05 \times Nw - 0.04 \tag{3.1}$$

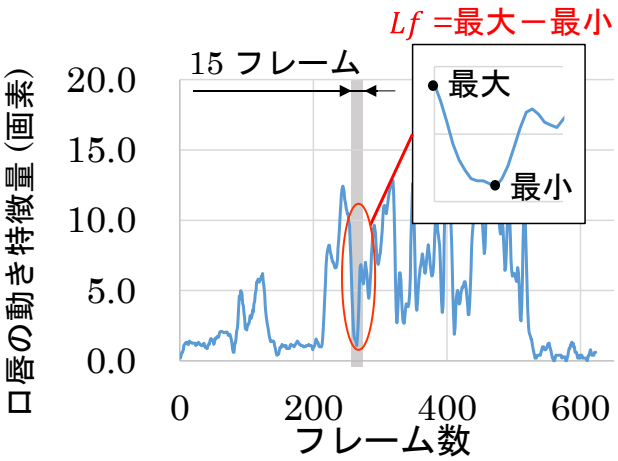


図 3.12　Lfの算出方法の概要

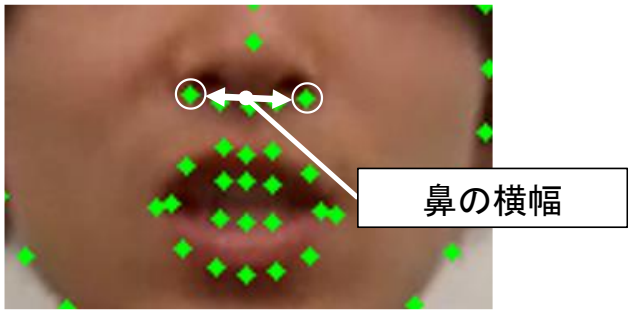


図 3.13　鼻の横幅の算出結果例

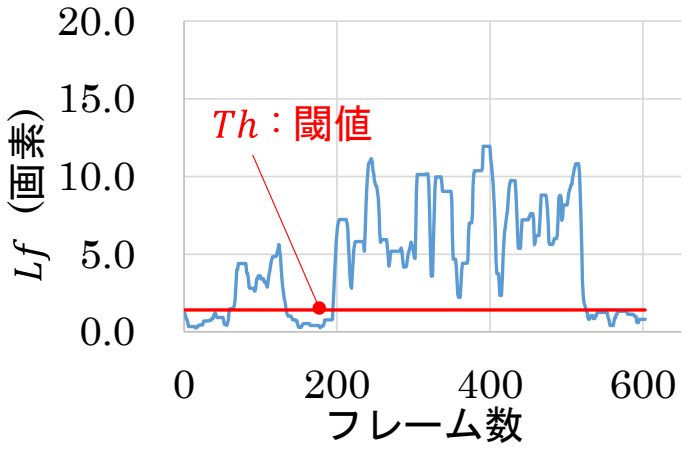


図 3.14　　Lfの算出結果例

3.3.7 口唇の動き特徴量を用いた発話区間の再抽出

図 3.15 に示すように，発話区間のフレームにおいて，口内領域の縦幅の動きが小さいことに起因して，数フレームが無発話フレームに判別される結果を認めた．このため，発話区間を再抽出する処理を行った．具体的には，処理対象の無発話フレームから 25 フレーム先までの間に発話フレームが存在する場合，処理対象の無発話フレームを発話フレームとして再抽出した．なお，再抽出のフレーム幅(25 フレーム)は 3.4 節の検討結果に基づいて良好な値を設定した．

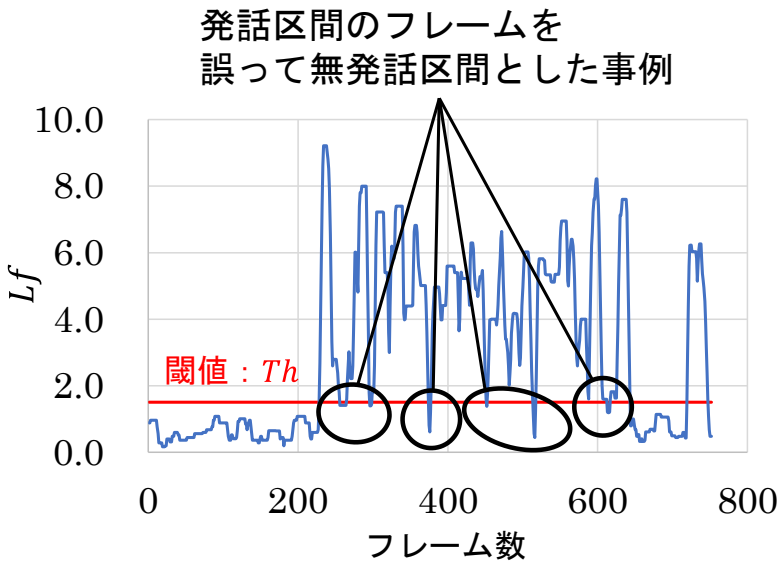


図 3.15　無発話区間のフレームとして誤って抽出されたフレームの例

3.3.8 Mel-frequency cepstral coefficients を用いた音声の特徴量取得処理

本論文では，音声の特徴量として MFCC [4]を使用した．MFCC は音声認識に有用な特徴量を有しており，低次元成分に含まれるスペクトル包絡の特徴は，声道の音響特性や口腔の形状に起因して変化する．すなわち，低次元成分を使用することで人物間における声の特徴の違いではなく，口の形状によって変化する音声の特徴を取得することが可能である．はじめに，音声データに対して短時間フーリエ変換(STFT: short-time Fourier transform) [4]を行い，周波数成分を取得した．次に，周波数成分をデシベル単位に変換し，メルフィルタバンクを適用した．人間の聴覚は高周波になるにつれて分解能が低くなる特性を有している[4]．メルフィルタバンクは，周波数が高い領域において分解能が低くなるような構造のフィルタであるため，人間の聴覚の特性を音声データに反映させることが可能である．最後に，メルフィルタバンク処理後の周波数成分に対し，離散コサイン変換を用いて MFCC の各係数における値を取得した．MFCC の横軸の単位はケフレンシーと呼ばれ，時間軸に相当する単位である．特に，0 次元目は音声の直流成分(音声レベル)を表現可能である．一般的に，MFCC の低いほうから連続した 10～15 次元程度が音声認識に使用される[4]．

本論文では，図 3.16 に示すように，音声データに対してフレーム長 20ms，フレーム間隔 10ms の窓関数を適応し，抽出された局所的な音声データに対して MFCC を算出した．MFCC の低次元成分は声道の音響特性や口腔の形状に起因して変化する特徴を有していること，0 次元目は音声レベル(直流成分)を表現可能であることを踏まえ，最も低い次元である 0 次元目の時系列変化を音声の特徴量として検討に使用した．音声の特徴量の時系列変化取得結果例を図 3.17 に示す．

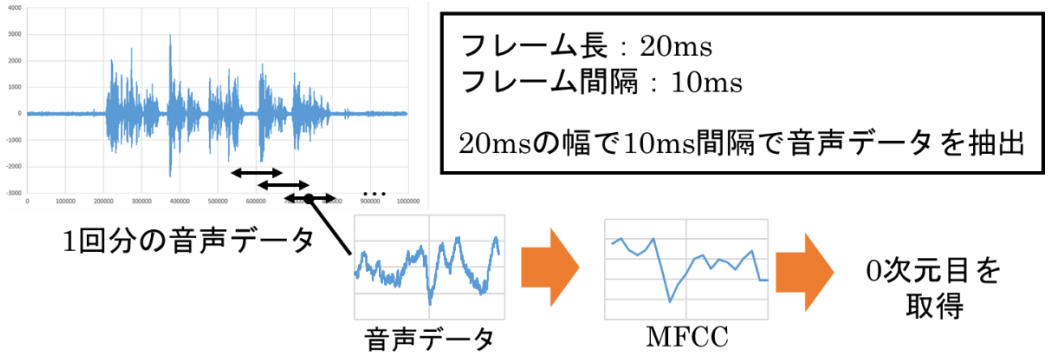


図 3.16　音声の特徴量取得処理例

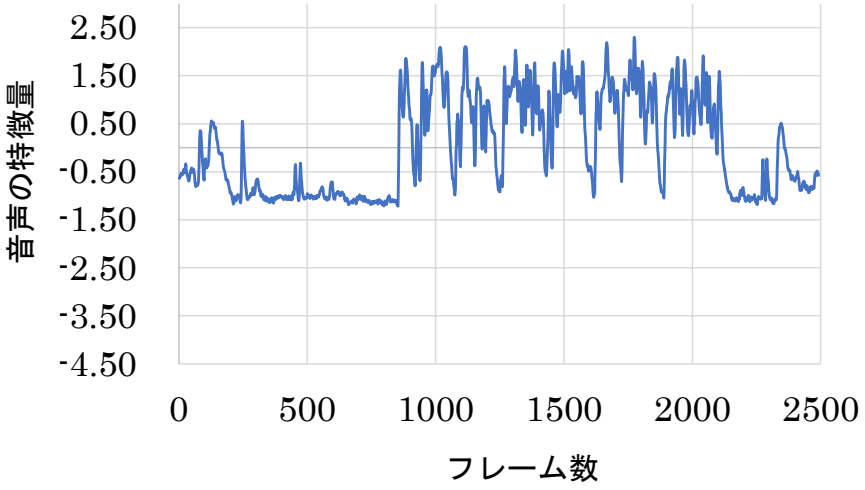


図 3.17　音声の特徴量の時系列変化取得結果例

3.3.9 音声の特徴量を使用した発話区間の抽出処理

図 3.18 に示すように，無発話区間と比較して発話区間における音声の特徴量は，単位時間あたりにおける変動が大きいこと，高い値を維持する傾向があることを認めた．このため，音声の特徴量の時系列変化における「高い値かつ安定した値を維持する区間」を発話区間として抽出する処理を行った．処理の流れを①～③に示す．なお，3.4 節の検討結果に基づき，音声の特徴量は平滑化を行わずに使用した．

① 音声の特徴量の時系列変化における任意のm番目のフレームを着目フレームとし，着目フレームおよびその前後 14 フレーム(合計 29 フレーム)を対象として，音声の特徴量の最大値および最小値を算出した．これをすべてのフレームに対して実施した．なお，フレーム幅は 3.4 節の検討結果に基づいて良好な値を設定した．

② ①で算出したm番目のフレームにおける最大値から最小値を減算した値(dm)を算出し，dmの時系列変化を取得した(図 3.19 参照)．

③ 音声の特徴量の値から同じフレーム番号のdmの数値を減算し，これをVfとして算出した．これをすべてのフレームに対して実施し，Vfの時系列変化を算出した．

検討用データセットにおけるすべての被験者の発話区間において，正の値のVfが含まれる結果を認めた．このため，Vfの値が閾値 0.00 以上のフレームを発話区間のフレームとして抽出した(図 3.20 参照)．

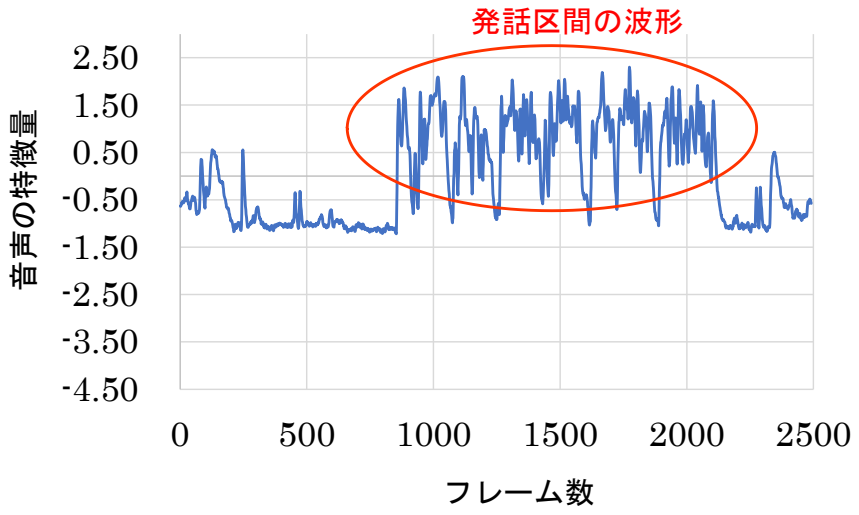


図 3.18　発話区間における音声の特徴量

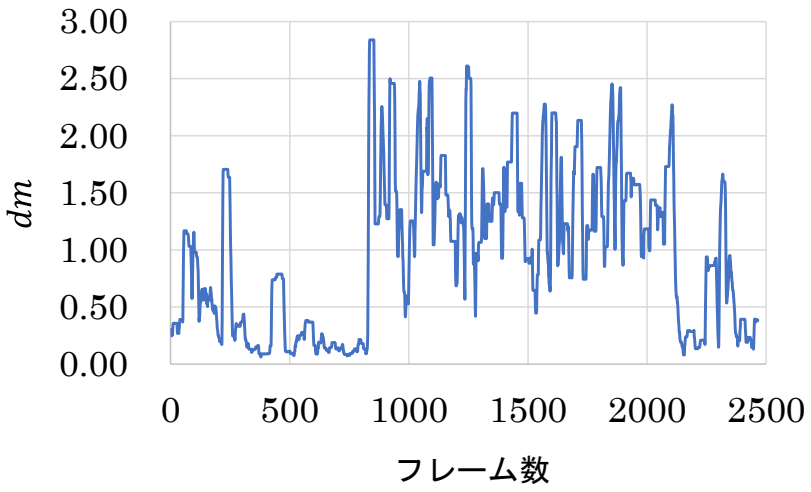


図 3.19　dmの算出結果例

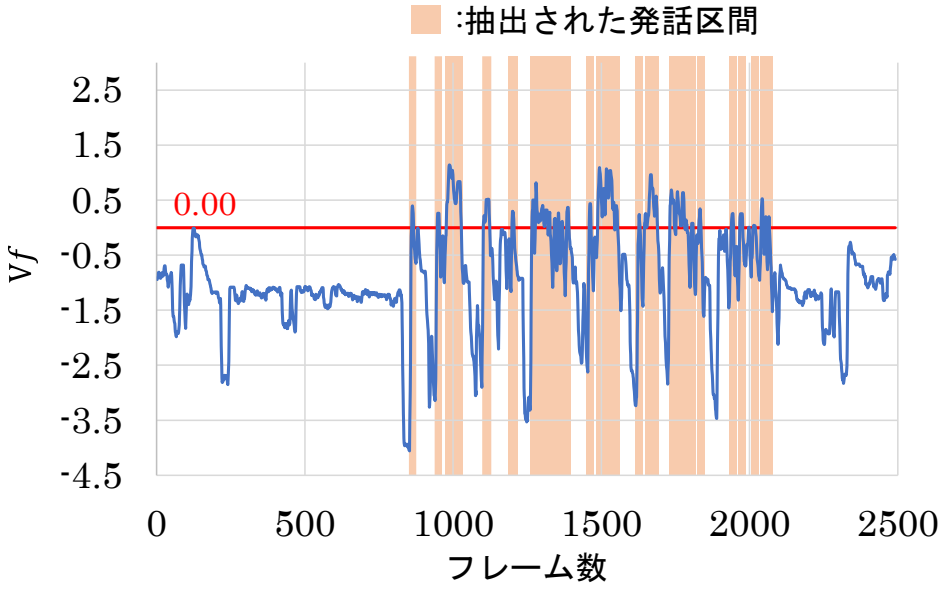
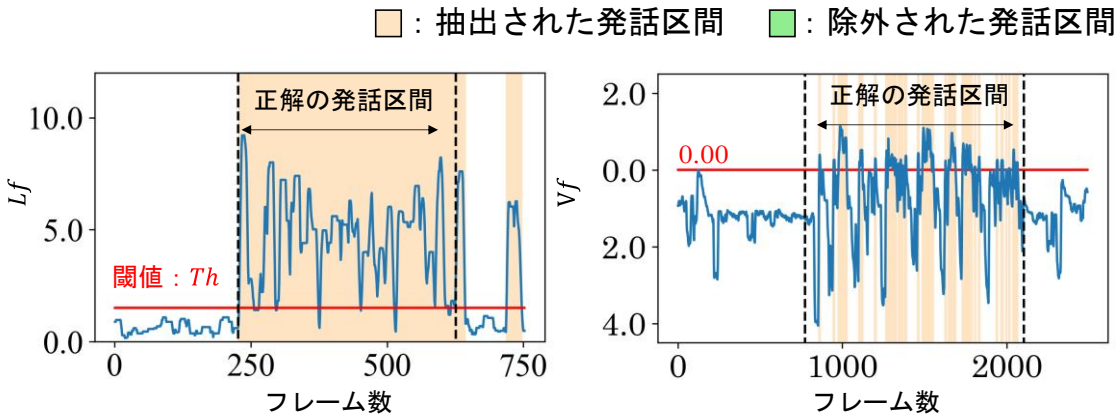


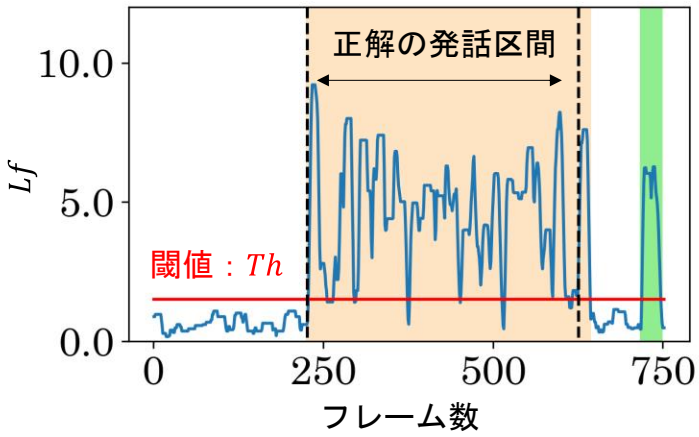
図 3.20　Vfの算出と発話区間の抽出処理結果例

3.3.10 口唇と音声の特徴量を使用した発話区間の抽出処理

図 3.21 に示すように，口内領域の縦幅を用いて抽出した発話区間(図 3.21(a)参照)および音声の特徴量を用いて取得した発話区間(図 3.21(b)参照)を用いて，発話区間の判別を行う．具体的には，口内領域の動きを用いて抽出した発話区間であり，かつ音声の特徴量を用いて抽出した発話区間が含まれている場合，この区間に含まれるフレームを発話区間のフレームとして判別(図 3.21(c)参照)した．



(a)口唇の動き特徴量を用いた結果例　　(b)音声の特徴量を用いた結果例



(c)口唇の動き特徴量および音声の特徴量を使用した結果例

図 3.21　口唇の動き特徴量と音声の特徴量を使用した発話区間抽出結果例

3.4 パラメータの選定に関する検討

3.4.1 概要

発話区間抽出手法における 5 種類のパラメータの選定を行った．具体的には，①口唇の動き特徴量における平滑化フィルタのサイズ：p (3.3.5 項参照)，②Lfを算出するための最大値と最小値の算出範囲：q (3.3.6 項参照)，③発話区間再抽出の範囲：r (3.3.7 項参照)，④音声の特徴量の平滑化フィルタサイズ：t (3.3.9 項参照)，ならびに⑤dmを算出するための最大値と最小値の算出範囲：s (3.3.9 項参照)について検討した．

3.4.2 評価指標

各パラメータの最適値を決定するために，(3.2)～(3.4)式を用いて F-measure [23]を算出し，評価指標として使用した．F-measure は 0.0～1.0 の範囲をとり，値が 1.0 に近いほど発話区間抽出成功率が高いことを示す．ここで，TP：発話フレームが正しく発話フレームとして判別されたフレーム数，FP：無発話フレームが誤って発話フレームとして分類されたフレーム数，FN：発話フレームが誤って無発話フレームとして分類されたフレーム数，TN：無発話フレームが正しく無発話フレームに判別されたフレーム数をそれぞれ示す．判別成功率算出のために使用した数値を表 3.4 にまとめる．

$$precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$recall = \frac{TP}{TP + FN} \tag{3.3}$$

$$F\text{-}measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{3.4}$$

表 3.4　判別成功率判定のためのフレーム数　(単位：フレーム数)

		実際の発話区間	
		発話フレーム	無発話フレーム
抽出結果	発話フレーム	TP	FP
	無発話フレーム	FN	TN

3.4.3 口唇を用いた発話区間のパラメータに関する検討

検討用データセットを使用し，パラメータ p，q，ならびに r(3.4.1 項参照)について検討を行った(以降，検討 1 と表記する)．なお，パラメータの検討パターンが膨大となるため，音声の特徴量を用いた発話区間の抽出に関するパラメータ t，および s(3.4.1 項参照)については固定の値を使用して検討した．各パラメータの検討範囲を表 3.5 に示す．各パラメータの組み合わせに対して F-measure の平均値を算出し，得られた結果を比較した．

表 3.5　検討 1 における各パラメータの検討範囲

パラメータ名	検討範囲
口唇の動き特徴量における平滑化フィルタのサイズ：p	5.0～50.0 まで 5.0 刻み
Lfを算出するための最大値と最小値の算出範囲：q	5.0～50.0 まで 5.0 刻み
発話区間再抽出の範囲：r	5.0～50.0 まで 5.0 刻み
音声の特徴量の平滑化フィルタサイズ：t	100.0
dmを算出するための最大値と最小値の算出範囲：s	10.0

3.4.4 口唇を用いた発話区間のパラメータに関する検討結果

検討 1 において算出されたパラメータの検討結果を表 3.6 に示す．p：5.0，q：5.0，r：30.0 において，最大の F-measure[23]の平均値が得られた．次に，これらの結果を踏まえ，未検討のパラメータの範囲について検討を行った(以降，検討 2 と表記する)．検討 2 におけるパラメータの検討範囲を表 3.7 に示す．この結果，p：5.0，q：7.0，r：25.0 において最大の F-measure の平均値が得られた(表 3.8 参照)．

以上の結果から，各パラメータの最適値を p：5.0，q：7.0，r：25.0 と判断し，以降の検討に使用した．

表 3.6　検討 1 において最大の F-measure の平均値が得られたパラメータのパターン

p	q	r	F-measure の平均値
5.0	5.0	30.0	0.977

表 3.7　検討 2 における各パラメータの検討範囲

パラメータ名	検討範囲
口唇の動き特徴量における平滑化フィルタのサイズ：p	1.0～11.0 まで 1.0 刻み
Lf を算出するための最大値と最小値の算出範囲：q	1.0～11.0 まで 1.0 刻み
発話区間再抽出の範囲：r	25.0～36.0 まで 1.0 刻み
音声の特徴量の平滑化フィルタサイズ：t	100
dm を算出するための最大値と最小値の算出範囲：s	10

表 3.8　検討 2 において最大の F-measure の平均値が得られたパラメータのパターン(最終的なパラメータの選定結果)

p	q	r	F-measure の平均値
5.0	7.0	25.0	0.981

3.4.5 音声の特徴量を用いた発話区間抽出のパラメータに関する検討

検討用データセットを使用し，パラメータ t，および s(3.4.1 項参照)について検討を行った(以降，検討 3 と表記する)．なお，口唇の動きを用いた発話区間抽出に使用したパラメータ p，q，ならびに r(3.4.1 項参照)については，3.4.4 項において選定したパラメータを使用した. 各パラメータの検討範囲を表 3.9 に示す. 各パラメータの組み合わせに対して F-measure[23]の平均値を算出し, 比較した.

表 3.9　検討 3 における各パラメータの検討範囲

パラメータ名	検討範囲
口唇の動き特徴量における平滑化フィルタのサイズ：p	5.0
Lf を算出するための最大値と最小値の算出範囲：q	7.0
発話区間再抽出の範囲：r	25.0
音声の特徴量の平滑化フィルタサイズ：t	10.0～150.0 まで 10 刻み
*dm*を算出するための最大値と最小値の算出範囲：s	10.0～150.0 まで 10 刻み

3.4.6 音声の特徴量を用いた発話区間抽出のパラメータに関する検討結果

検討 3 において算出されたパラメータの検討結果を表 3.10 に示す. t : 10.0, s : 20.0 において最大の F-measure[23]の平均値が得られた. これらの結果を踏まえ, 未検討のパラメータの範囲について検討を行った(以降, 検討 4 と表記する). 検討 4 におけるパラメータの検討範囲を表 3.11 に示す. この結果, 複数のパラメータの組み合わせにおいて F-measure の平均値は最大の値を得た. パラメータの値が小さいほど, 波形の処理区間の範囲が狭くなるため, 処理前の波形の時系列変化を維持することが可能である. このため, F-measure の平均値が最大であるパターンのうち, パラメータの値が最も低い組み合わせを選定した(表 3.12 参照).

以上の結果から, 各パラメータの最適値を t : 1.0(平滑化処理なし), s : 14.0 と判断し, 以降の検討に使用した.

表 3.10 検討 3 において最大の F-measure の平均値が得られたパラメータのパターン

t	s	F-measure の平均値
10.0	20.0	0.981

表 3.11 検討 4 における各パラメータの検討範囲

パラメータ名	検討範囲
口唇の動き特徴量における平滑化フィルタのサイズ : p	5.0
Lf を算出するための最大値と最小値の算出範囲 : q	7.0
発話区間再抽出の範囲 : r	25.0
音声の特徴量の平滑化フィルタサイズ : t	1.0～20.0 まで 1.0 刻み
dm を算出するための最大値と最小値の算出範囲 : s	10.0～30.0 まで 1.0 刻み

表 3.12 検討 4 において最大の F-measure の平均値が得られたパラメータのパターン(最終的なパラメータの選定結果)

t	S	F-measure の平均値
1.0（平滑化なし）	14.0	0.981

3.5 閾値の自動算出処理に関する検討

3.5.1 概要

会議において人物の位置は変動する．したがって，被験者とカメラ間距離は常に変動する可能性があると考える．提案手法は口唇の動きおよび閾値を使用して発話区間の抽出を行っているため，カメラと被験者間の距離に応じて適切な閾値を設定することで，精度良く発話区間抽出が可能であると考える．本節では，被験者とカメラ間距離の変動に伴い，画像中における顔の部位のサイズが変動すると仮定し，これに基づいた閾値の設定に関する 2 つの検討を行った．具体的には，①顔の特徴点間距離を用いた閾値算出処理のための説明変数に関する検討(3.5.2 および 3.5.3 項)，および②説明変数を用いた閾値算出式に関する検討(3.5.4 および 3.5.5 項)を行った．

3.5.2 閾値算出のための説明変数に関する検討

説明変数に使用する特徴点間距離は，2 つの条件を満たす必要がある．具体的には，①距離変動以外の要因(発話や表情の変化)に伴って長さが変化しにくいこと，および②すべてのフレームにおいて安定的に取得可能であることである．①については，鼻の特徴点間距離を使用することで条件を満たすことが可能である．一方，鼻の特徴点のうち，②の条件を満たす特徴点間距離に関しては選定の余地がある．そこで，検討用データセットを使用して②の条件に関する検討を行った．はじめに，図 3.22 に示すように，すべてのフレームから鼻の横幅と縦幅を算出した．次に，各被験者の各動画データにおける鼻の横幅と縦幅の標準偏差をそれぞれ算出した．最後に，各被験者の鼻の横幅と縦幅の標準偏差における平均値をそれぞれ算出して比較した．

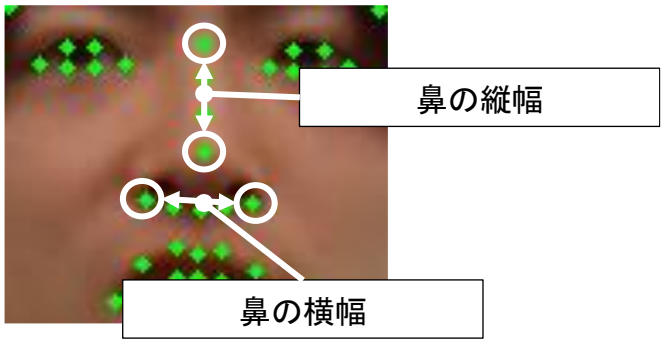


図 3.22　鼻の横幅と縦幅の算出例

3.5.3 説明変数の検討結果

表 3.13 に各被験者における鼻の横幅と縦幅の標準偏差算出結果を示す．被験者 6 名中 5 名で鼻の縦幅よりも横幅の標準偏差は小さい値を得た．また，平均値を比較した場合においても，鼻の縦幅よりも横幅の標準偏差は小さい．これらの結果は，鼻の縦幅と比較して鼻の横幅は，1 つの動画データにおいて安定的に取得可能であることを示唆している．

以上の結果は，鼻の横幅を説明変数として使用することは，閾値算出のために有用である可能性を示している．

表 3.13　鼻の横幅と縦幅における標準偏差の平均値

被験者	A	B	C	D	E	F	平均値
鼻の横幅	1.18	0.66	0.60	0.93	0.67	0.91	0.82
鼻の縦幅	1.16	0.71	0.96	1.99	1.07	2.20	1.35

※赤字：各列における最小値

3.5.4 閾値の算出式に関する検討

検討用データセットを対象として閾値の算出式に関する検討を行った．閾値の算出式の検討概要を図 3.23 に示す．はじめに，各発話動画データの鼻の横幅の中央値(*Nw*)を閾値計算の説明変数としてそれぞれ算出した．次に，被験者とカメラ間距離を疑似的に変更するために，口唇の動き特徴量と*Nw*に対して重み(0.1～5.0 まで 0.1 刻みで変更)を乗算した．ここで，重みの値が小さいほどカメラと被験者間距離が遠いことを示す．さらに，閾値*Th*の値を 0.0～5.0 まで 0.1 刻みで変更し，各重みと閾値の組み合わせにおける F-measure[23]の値を各被験者の文章ごとに算出した．加えて，各閾値において 0.90 以上の F-measure の平均値を有するデータ(以降，高精度事例と表記する)を抽出した．最後に，高精度事例における*Nw*の中央値(*M*-*Nw*)を各被験者の文章における閾値ごとに算出し，プロットした．プロットした点群に対して最小二乗法を適用し，算出された 1 次関数を閾値算出式とした．

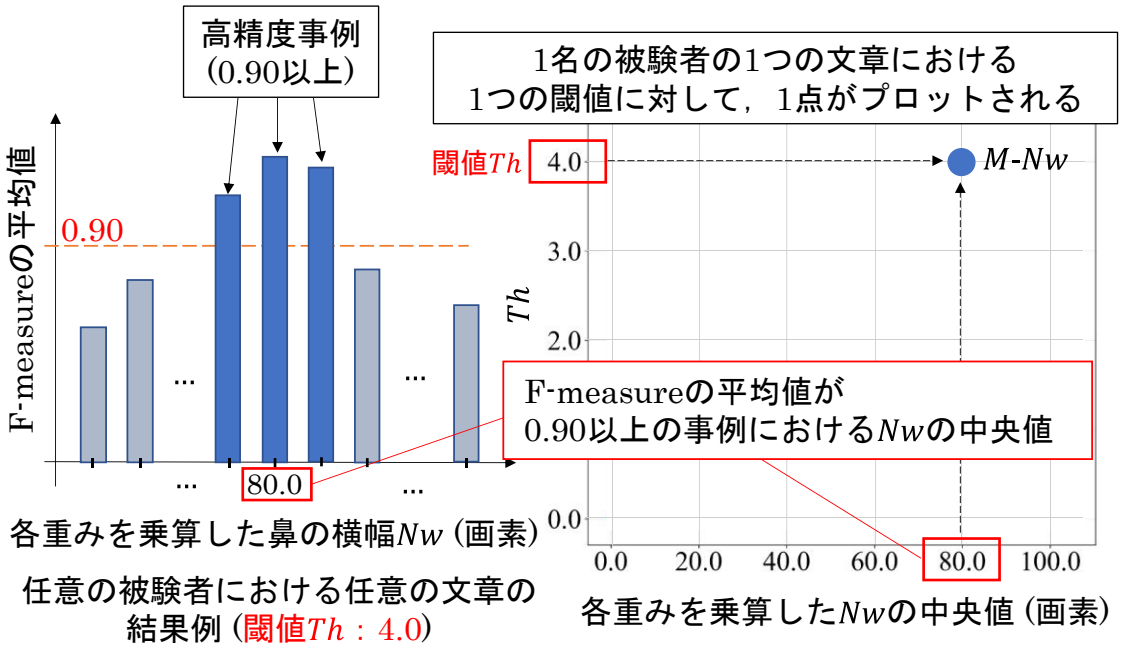


図 3.23　閾値の算出式の検討概要

3.5.5 閾値の算出式に関する検討結果

図 3.24 に検討用データセットにおける$M\text{-}Nw$の算出結果とその近似直線を示す．閾値ThとNwの値が大きくなるのにしたがって，$M\text{-}Nw$の分布範囲は広くなる傾向がある．最小二乗法を用い，$M\text{-}Nw$の分布の中央を通るような近似直線を算出した結果を(3.5)式および図 3.24 の赤線として示す．なお，(3.5)式における係数は小数点以下第 3 位を四捨五入して算出した．本論文では，閾値の算出式として(3.5)式が有用であると考え，以降の検討にした．

$$Th = 0.05 \times Nw - 0.04 \tag{3.5}$$

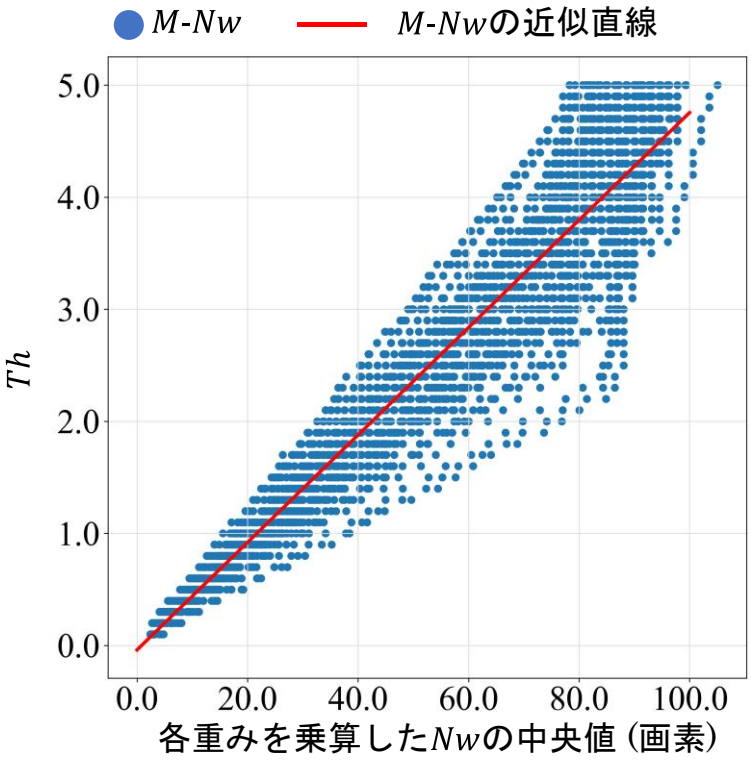


図 3.24　$M\text{-}Nw$の算出結果と近似直線

3.6 閾値の算出式に対する評価

3.6.1 評価方法

評価用データセットを用いて閾値の算出式に関する評価を行った．具体的には，閾値を自動設定した場合と固定値を用いた場合とを比較した．検討用データセットを使用し，閾値Thを 0.0～5.0 まで 0.1 刻みで検討し，Thが 1.6 および 1.7 の場合に最大の F-measure[23]を得た．このため，閾値を固定した方法は 1.65(1.6 と 1.7 の平均値)を閾値Thの値として設定した．なお，被験者とカメラ間距離を疑似的に変動させるために，3.5.4 項と同じ方法を用いて口唇の動き特徴量および鼻の横幅の中央値Nwに重みを乗算し，得られた値を本評価に使用した．

3.6.2 評価結果

図 3.25 に閾値を自動設定した場合と閾値を固定した場合における F-measure の平均値算出結果を示す．閾値を自動設定することで，重みが変動した場合であっても F-measure の平均値は 0.95 以上の値を維持している．一方，閾値に固定値を使用した場合，重みの変動に伴って F-measure の平均値が低下していることがわかる．表 3.14 に各手法における F-measure の最大値，最小値，ならびに平均値を示す．閾値に固定値を使用した場合と比較して，閾値を自動設定した場合に高い F-measure の平均値が得られた．

以上の結果は，提案する閾値の自動設定手法は被験者とカメラ間距離が変動した場合であっても，高い精度で発話区間を抽出可能であることを示唆している．

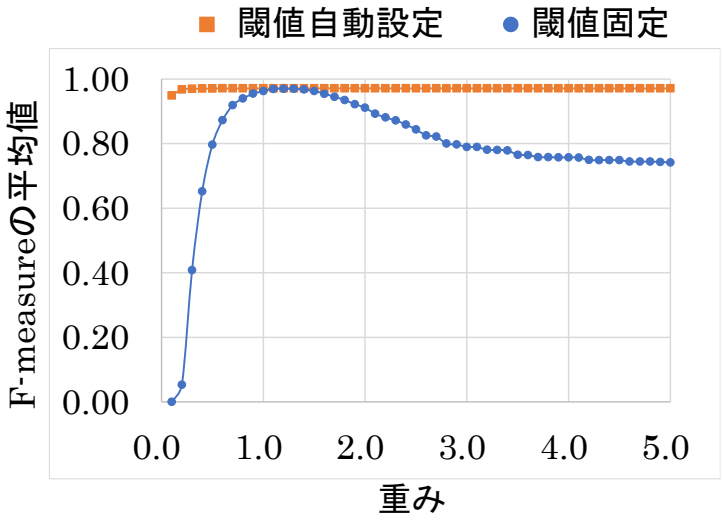


図 3.25　各重みの値における F-measure の平均値

表 3.14　各手法における F-measure の平均値

	閾値自動設定	閾値固定
最大値	0.97	0.97
最小値	0.95	0.00
平均値	0.97	0.79

3.7 提案手法における発話区間抽出精度の比較評価

3.7.1 評価方法

評価用データセットを対象とし，提案手法および比較手法[7, 8]の比較評価を行った．このとき，評価指標として 3.4.2 項に示す F-measure を使用した．各手法における被験者ごとの F-measure の平均値を算出し，得られた値を比較した．

3.7.2 比較手法

比較手法として，画像情報と音声情報を用いた発話者推定手法[7, 8]を使用した．比較手法は，画像中における口唇の動きと音声の類似性を評価し，各フレームにおける話者を識別する手法である．本評価では，比較手法を発話区間抽出に適用するために非発話動画を作成した(図 3.26 参照)．非発話動画は，発話動画データにおける 1 フレーム目の画像のみで構成された動画データに対して，音声を付加した動画である．したがって，非発話動画は静止画に対して音声情報が付加されている．非発話動画データにおいて，口唇の動きと音声間における類似性は，口唇に動きのある動画データ(以降，発話動画と表記する)よりも低くなると考える．すなわち，発話動画と非発話動画を対象とし，同時刻のフレーム間における発話者度合を比較することで，発話者判別が可能であると考える．

以上を踏まえ，本評価では比較手法を用い，以下に示す 2 つの方法(以降，比較手法 1，比較手法 2 とそれぞれ表記する)で発話区間の抽出を行った．

＜比較手法 1＞

はじめに，発話動画と非発話動画における同じフレームの発話者度合の差分($Diff$)を算出した．さらに，$Diff$が 0.0 以上の場合，発話フレームとして抽出した．すなわち，発話動画と非発話動画における同じフレームの発話者度合を比較し，発話動画の発話者度合が高い場合，発話フレームとして抽出した．

＜比較手法 2＞

はじめに，発話動画と非発話動画における同じフレームの発話者度合の差分($Diff$)を算出した．次に，発話動画開始時の約 1.0 秒間と終了時の約 1.0 秒間における無発話区間の$Diff$を取得し，これらの区間における$Diff$の最大値を閾値として算出した．最後に，$Diff$の値が閾値を超える区間を発話区間として抽出した．なお，閾値は動画データごとに算出して使用した．閾値の設定例を図 3.27 に示す．

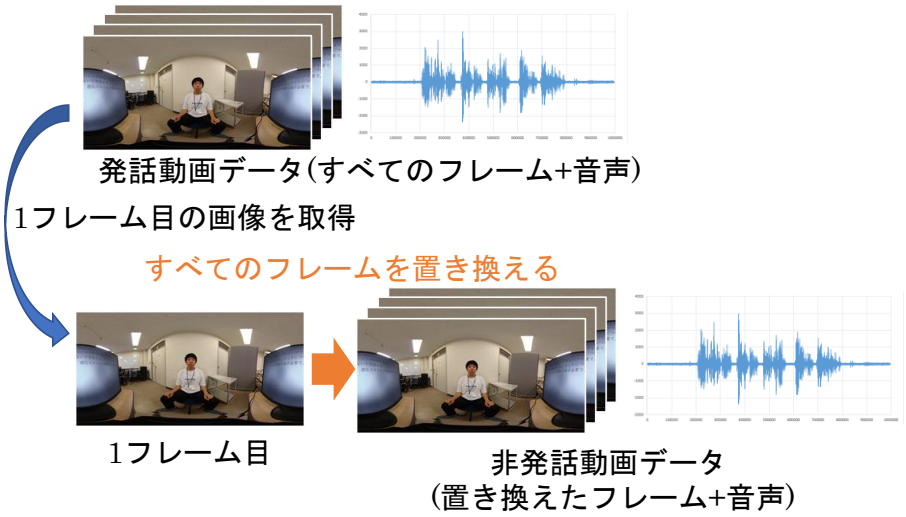


図 3.26　発話動画と非発話動画の定義

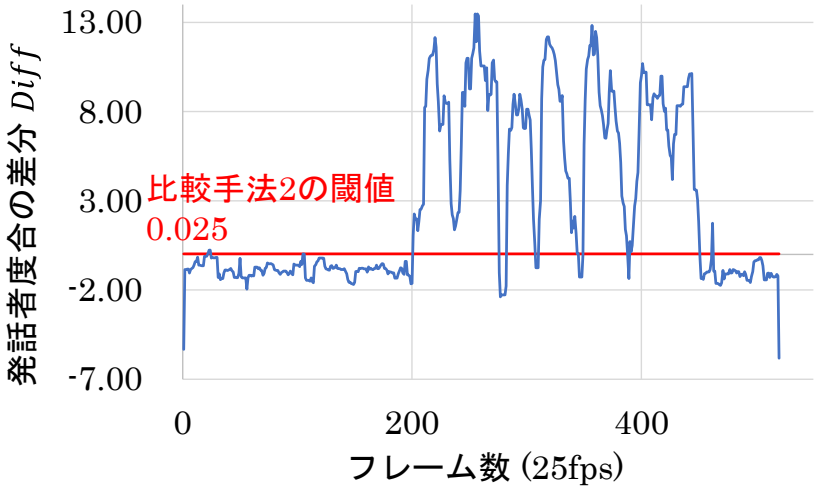


図 3.27　被験者 B の文章 6 における比較手法 2 の閾値設定例
(閾値は対象の動画データごとに異なる値を設定した)

3.7.3 評価結果

表 3.15 に各手法における各被験者の F-measure の平均値を示す．比較手法 1 および 2 と比較して，提案手法は 14 名中 13 名で F-measure の値が向上した．特に，被験者 C において提案手法の F-measure は 0.99 となり，最大の値を得た．図 3.28 および図 3.29 に発話区間抽出結果例を示す．正解の発話区間を包含するように発話区間が抽出されており，かつ誤って抽出された発話区間を除外して抽出が行われていることがわかる．これらの結果は，提案手法は精度よく発話区間を抽出可能であり，かつ比較手法よりも高い精度が得られることを示唆している．

表 3.15　各手法における各被験者の F-measure の平均値

被験者	提案手法	比較手法 1	比較手法 2
A	0.98	0.61	0.77
B	0.98	0.90	0.93
C	0.99	0.89	0.92
D	0.96	0.79	0.86
E	0.97	0.90	0.90
F	0.98	0.94	0.94
G	0.97	0.96	0.96
H	0.97	0.92	0.92
I	0.97	0.81	0.89
J	0.97	0.92	0.93
K	0.96	0.94	0.94
L	0.96	0.93	0.93
M	0.97	0.93	0.93
N	0.95	0.96	0.96
平均	0.97	0.88	0.91

※赤字：各行における最大値

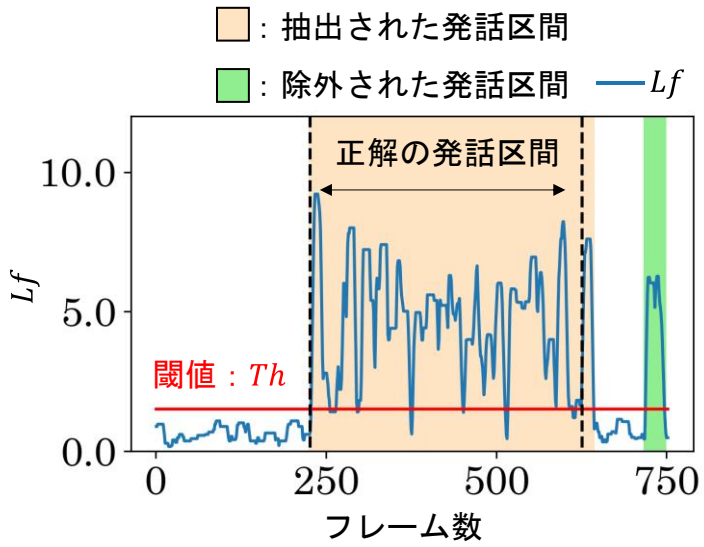


図 3.28　被験者 E の発話区間抽出結果例
(文章 1，1 回目，F-measure：0.98)

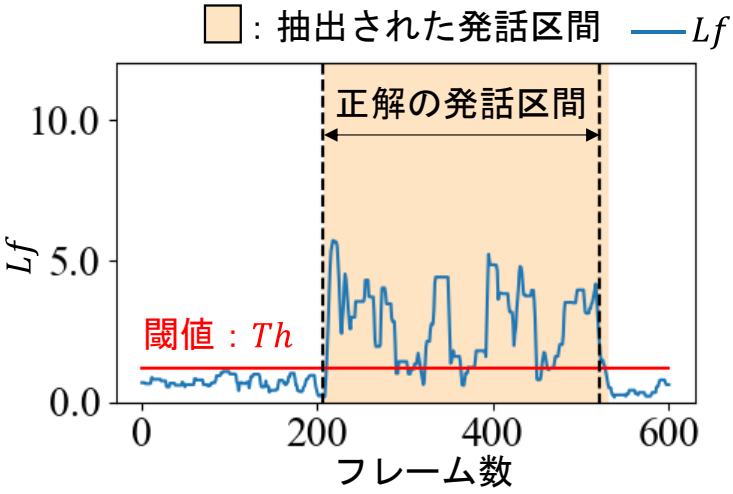


図 3.29　被験者 C の発話区間抽出結果例
(文章 7，1 回目，F-measure：0.99)

3.7.4 精度の低下した被験者 N に対する考察

被験者 N における F-measure の平均値は，比較手法よりも 0.01 低い結果を得た．これは発話終了直後に口唇の動きが生じたことに起因すると考える．被験者 N の発話区間抽出結果例を図 3.30 および図 3.31 にそれぞれ示す．発話終了直後(以降，発話終了直後の区間と表記する)のLfの値は，閾値を超えていることがわかる(図 3.30 参照)．一方，図 3.31 に示すVfの結果では，発話終了直後における区間(破線の丸で示す区間)では，発話区間が抽出されていないことがわかる．このように，口唇の動きと音声が同時に生じていない区間であるため，3.3.10 項「口唇と音声の特徴量を使用した発話区間の抽出処理」によって除外することが可能な発話区間であると考える．この区間に対して 3.3.7 項で前述した「口唇の動き特徴量を用いた発話区間の再抽出」の処理を行った場合，直前の発話区間と連続して抽出される．このため，発話終了直後の区間に音声が生じていないにも関わらず，発話区間として誤って抽出されると考える．したがって，「口唇の動き特徴量を用いた発話区間の再抽出」における処理について再検討し，発話終了直後の区間における過剰な抽出を改善することで，精度の向上が可能になると考える．

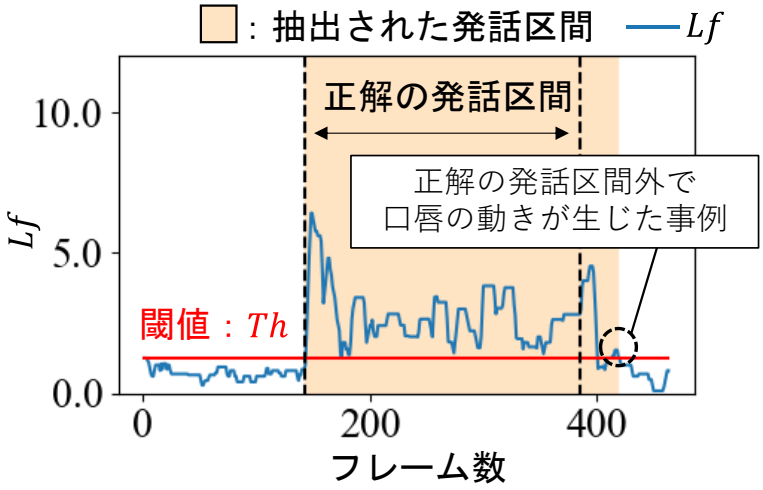


図 3.30　被験者 N の発話区間抽出結果例
(文章 7，1 回目，F-measure：0.94)

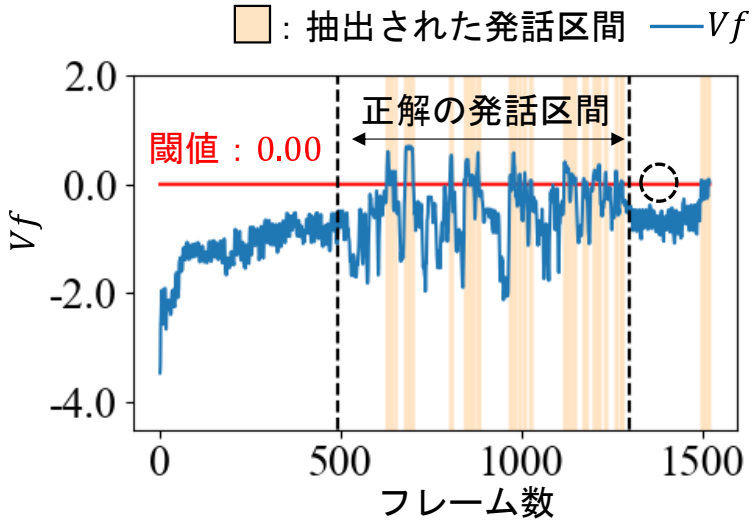


図 3.31　被験者 N の発話区間抽出における音声の特徴量の結果例
（文章 7，1 回目，F-measure：0.94）

3.8 まとめ

本論文では，被験者 14 名の発話動画を対象とし，画像中の口唇の動きと音声情報を用いた発話区間抽出に関する検討を行った．得られた成果を以下にまとめる．

(1) 提案手法の閾値自動設定処理において，閾値の算出のための説明変数として鼻の横幅を使用することが有用であることを明らかにした．

(2) 提案手法の閾値自動設定処理は，被験者とカメラ間距離の変動に伴う発話区間抽出精度の低下を低減可能である．また，最大で 0.99 の F-measure の平均値が得られ，高い精度で発話区間の抽出が可能であることを明らかにした．

(3) 比較手法 1 および 2[7, 8]と比較して，提案手法は高い精度で発話区間の抽出が可能であることを明らかにした．

第 3 章 参考文献

[1] VoXT Voice Texting：「VoXT とは」, https://voxt.jp/outline/ (Access 2021/12/15)

[2] NEC ソリューションイノベータ：「Voice Graphy」, https://www.nec-solutioninnovators.co.jp/ss/smartwork/products/voicegraphy/ (Access 2021/12/15)

[3] Microsoft Azure, “Speaker Recognition”, https://azure.microsoft.com/ja-jp/services/cognitive-services/speaker-recognition/ (Access 2021/12/15)

[4] 篠田浩一：「音声認識 Speech Recognition」, 講談社 (2017)

[5] Z. Meng, M. Umair Bin Altaf, and B. Juang：“Active voice authentication”, Digital Signal Processing, Vol. 101, 102672 (2020)

[6] X. Wang, F. Xue, W. Wang, and A. Liu：“A network model of speaker identification with new feature extraction methods and asymmetric BLSTM”, Neurocomputing, Vol. 403, pp. 167−181 (2020)

[7] J.S. Chung, and A. Zisserman：“Out of time: automated lip sync in the wild”, Workshop on Multi-view Lip-reading, ACCV (2016)

[8] J.S. Chung, and A. Zisserman：“Learning to lip read words by watching videos”, Computer Vision and Image Understanding, Vol. 173, pp. 76−85 (2018)

[9] THETA：「製品 RICOH THETA V」, https://theta360.com/ja/about/theta/v.html (Accessed: 2021/12/15)

[10] THETA：「製品 アクセサリー」, https://theta360.com/ja/about/theta/accessory.html (Accessed: 2021/12/15)

[11] THETA：「アプリケーション」, https://theta360.com/ja/about/application/ (Accessed: 2021/12/15)

[12] YahooJapan ニュース：「世界初 自律学習する AI の可能性」, https://headlines.yahoo.co.jp/videonews/fnn?a=20190412-00416026-fnn-soci (Accessed: 2018/3/10)

[13] YahooJapan ニュース：「先端技術で地方の課題解決 京都・舞鶴市とオムロン子会社協定」, https://headlines.yahoo.co.jp/hl?a=20190414-00000009-kyt-bus_all (Accessed: 2018/3/10)

[14] YahooJapan ニュース：「新元号に便乗したフィッシング詐欺も　巧妙化する「インターネット詐欺」のいま」,
https://headlines.yahoo.co.jp/hl?a=20190405-00010001-danro-life
(Accessed: 2018/3/10)

[15] YahooJapan ニュース：「産総研，人・機械協調 AI を研究するサイバーフィジカルシステム研究棟を公開」,
https://headlines.yahoo.co.jp/hl?a=20190416-00000120-impress-sci
(Accessed: 2018/3/10)

[16] YahooJapan ニュース：「保存血液，ｉＰＳ細胞研究に利用へ」,
https://headlines.yahoo.co.jp/hl?a=20190411-00050258-yom-sci
(Accessed: 2018/3/10)

[17] YahooJapan ニュース：「PET ボトルの 100%有効利用に本気 飲料業界のプラスチック資源循環宣言」,
https://headlines.yahoo.co.jp/hl?a=20190329-00010001-ssnp-bus_all
(Accessed: 2018/3/10)

[18] “Dlib C++ Liblary”,　http://dlib.net/ (Accessed 2021/12/15)

[19] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic：“300 Faces In-The-Wild Challenge: Database and results”, Image and Vision Computing (IMAVIS), Vol.47, pp.3–18 (2016).

[20] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic：“A Semi-automatic Methodology for Facial Landmark Annotation”, Proceedings of IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR-W), 5th Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2013). Oregon, DOI: 10.1109/CVPRW.2013.132 (2013)

[21] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic：“300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge”, Proceedings of IEEE Int'l Conf. on Computer Vision (ICCVW), 300 Faces in-the-Wild Challenge (300W), DOI: 10.1109/ICCVW.2013.59 (2013)

[22] V. Kazemi, and J. Sullivan：“One millisecond face alignment with anensemble of regression trees”, 2014 IEEE Conference on Computer Vision and Pattern Recognition, DOI:10.1109/CVPR.2014.241, Columbus, OH, USA (2014)

[23] 竹村彰通監訳他：「機械学習　データを読み解くアルゴリズムの技法」, 朝倉書店（2017）

第 4 章　発話者の判別手法に関する検討

4.1 はじめに

議事録自動作成システムには，発話者の判別機能が搭載されている．具体的には，各会議参加者にマイクを割り当てる方法[1]や声紋を事前に登録する方法[2-4]などが発話者判別のために使用される．しかしながら，これらの方法は会議参加者の人数と同じ台数のマイクが必要な点や事前に声紋登録が必要な点など，事前準備を必要とする点が課題である．また，声紋を使用した発話者判別手法の場合，登録した音声データの管理が必要であることや，登録者数が多い場合に管理対象のデータ量が膨大になる可能性がある．

k-means 法を適用し，既知のクラス数を設定することで，声紋の登録が不要な発話者の判別が可能である[5]．しかしながら，類似した声を有した人物間の発話者判別が困難である課題を有している．また，音声の到来方向を検出して発話者を判別する手法が提案されている[6]．しかしながら，この手法では複数台のマイクで構成された特殊な装置が必要である．また，会議中に発声することなく人物が移動する場合があるため，音声の到来方向のみを用いて発話者を判別する手法では，発話者の追跡が困難になる場合がある．

一方，画像情報と音声情報を併用することは，上記課題の解決を可能にすると考える．具体的には，画像から取得可能な口唇の動きと音声情報の類似性を評価し，最も類似した人物を発話者として判別することが可能であると考える．類似研究として，顔画像と音声情報を用いた発話者判別手法が提案されている[7, 8]．これらの手法は，CNN や LSTM [9]の学習のために，膨大な量の学習データが必要であるという課題を有している．

一般的に口唇の大まかな動きは，機械学習を使用せずに音声のみの情報から推定することが可能である [10]．具体的には，声の第一フォルマント周波数[10]の動きによって顎の開閉を大まかに推定可能である．これに対して機械学習を応用することで，口唇の動きの推定精度の向上が可能であると考える．音声情報と口唇の動きの関連性に基づき，Suwajanakorn 氏ら[11]は，音声から口唇の動きを推定するための手法を提案した．この研究では，音声情報と LSTM を用いて口唇における 18 点の特徴点の動きを推定した．しかしながら，LSTM の学習に使用した人物は 1 名であることから，特定人物の口の動きに特化したモデルが構築されていると考える．複数の人物への汎用性を向上させ，かつ学習データ量を削減するためには, 18 点よりも少ない口唇の特徴点を検討対象とすることが有効であると考える．この結果，推定する動きのパターンが少なくなるため，少ない学習データを用いた汎用的な口唇の動き推定が可能になると考える．

上記課題を解決した発話者判別手法は，本論文における 3 章の「発話区間の抽出手法」において複数名の発話区間が抽出された場合に，真の発話者を判別することが可能であると考える．

本論文では，議事録自動作成システムにおける発話者判別機能の構築を目的とし，画像情報と音声情報を用いた発話者判別手法(以降，提案手法と表記する)に関して検討を加えた．具体的には，音声から口唇の動きを推定し，推定値と実際の口唇の動きが最も類似している人物を発話者として判別する手法について検討を加えた．なお，口唇の動きの推定には，LSTM[9]を使用した．提案手法と複数の比較手法[8, 10]における発話者判別成功率を比較し，有用性を評価した．

4.2 使用データ

4.2.1 データ取得および発話区間の設定に関して

本論文では対面における会議を想定し，全方位カメラを使用してデータの取得を行った．全方位カメラは 360°全方位の画像を取得可能であるため，会議などのように向かい合って会話をする環境下において，効率良く顔画像が取得できる．本章では，3 章の 3.2.2 項と同じデータにおける 1 回分の発話動画データをそれぞれ使用した．すなわち，被験者 14 名が 11 文を 1 回ずつ発話している発話動画データを使用した．また，使用機材も 3 章の 3.2.3 項と同様である．なお，本論文では発話区間のみを処理の対象とするために，3 章の 3.2.4 項と同様の手順で発話区間を設定して，検討に使用した．

4.2.2 データセット

被験者 14 名が 11 文を 1 回ずつ発話した動画データ(154 データ)を用いて，2 つのデータセットを作成した．データセットの定義を以下に示す．

i) 予備検討用テストデータセット:

被験者 A～F のデータ(66 データ)を表 4.1 に示すパターンで 6 分割し，テストデータとして使用した．このとき，教師データに関してはテストデータと被験者および文章が重複しないように設定した．たとえば，表 4.1 の PITD 1.1 をテストデータとして使用した場合，PITD 2.2 および PITD 3.2 を教師データとして使用した．なお，表 4.1 では，各文章における発話時間の平均値に基づき，各データ量間の差が小さくなるように文章の組み合わせを選定している．

ii) 検討・評価用テストデータセット:

被験者 A～N のデータ(154 データ)を用いて表 4.2 の 8 パターンのデータの組み合わせを作成し，テストデータとして使用した．このとき，教師データに関しては,テストデータと被験者および文章が重複しないように設定した．たとえば，表 4.2 の ITD 1.1 をテストデータとして使用した場合，被験者 F～N の文章 2, 4, 5, 6, 9 における動画データを教師データとして使用した．なお，表 4.2 では各文章における発話時間の平均値に基づき，各データ量間の差が小さくなるように文章の組み合わせを選定している．

表 4.1　予備検討用テストデータセット

被験者	文章 3, 6, 7, 9, 10, 11	文章 1, 2, 4, 5, 8
A, B	PITD 1.1	PITD 1.2
C, D	PITD 2.1	PITD 2.2
E, F	PITD 3.1	PITD 3.2

表 4.2　検討・評価用テストデータセット

被験者	文章 1, 3, 7, 8, 10, 11	文章 2, 4, 5, 6, 9
A, B, C, D, E	ITD 1.1	ITD 1.2
D, E, F, G, H	ITD 2.1	ITD 2.2
G, H, I, J, K	ITD 3.1	ITD 3.2
J, K, L, M, N	ITD 4.1	ITD 4.2

4.3 提案手法

4.3.1 提案手法の概要

時系列データの特徴を分類可能な LSTM [9]を用いて，発話者を判別した．はじめに，音声から得られた特徴量を用いて，口唇の動きの推定値を算出した．さらに，推定した口唇の動きと最も類似した動きを有する人物を，発話者として判別した. 発話者判別手法の手順を①～⑦に示す. また発話者判別手法の流れを図 4.1 に示す.

① 発話動画データの各フレームに対して，顔器官の特徴点検出を行い，顔領域を検出した.
② 検出した顔器官の特徴点における口唇の特徴点を用いて，口唇の縦幅および横幅を算出した．さらに，発話動画データにおける口唇の縦幅と横幅の時系列変化を算出した.
③ 口唇の縦幅と横幅の時系列変化に対して補正処理を行い，被験者とカメラ間距離の変動に伴う口唇の縦幅と横幅の変化を低減した.
④ 口唇の縦幅および横幅を用いて，口唇の動き特徴量を算出した.
⑤ 発話動画データを対象とし，音声の特徴量の時系列変化を取得した.
⑥ 音声の特徴量を入力し，口唇の動き特徴量の推定値が出力可能な LSTM モデルを構築し，これを用いて口唇の動き特徴量の推定値を算出した.
⑦ 口唇の動き特徴量の推定値と実際の口唇の動き特徴量の類似度を算出し，最も類似度が高い被験者を発話者として判別した.

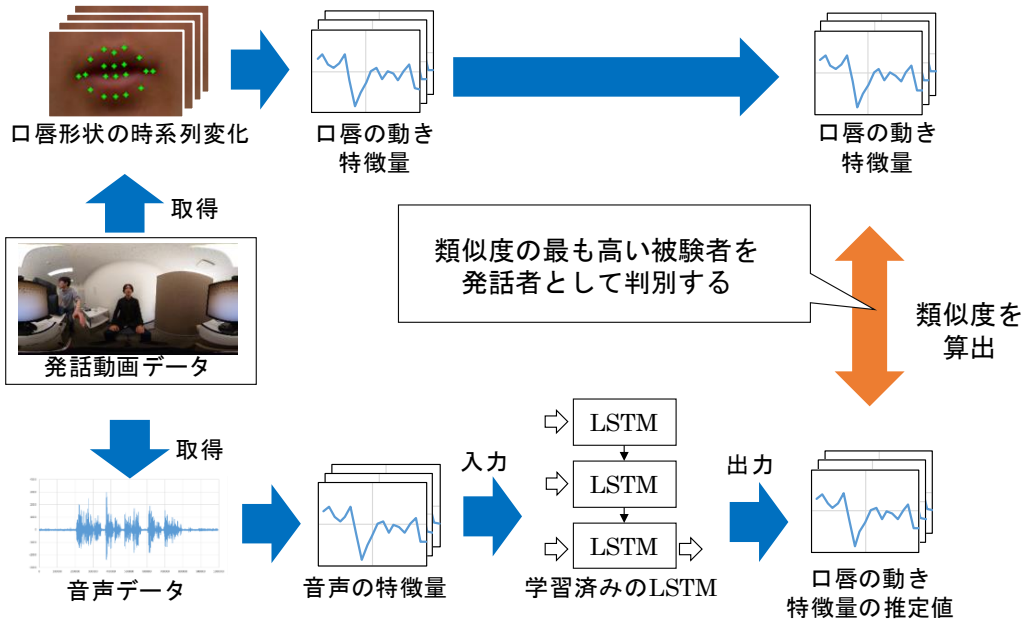


図 4.1　発話者判別手法（提案手法)の概要

4.3.2 口唇の特徴点の取得

本論文では処理のリアルタイム性を考慮し，オープンソースライブラリの Dlib による顔器官検出を行い，口唇の形状を取得した．このとき，Dlib[12]に搭載されている顔器官検出機能および，公式で配布されている学習済みのモデル(IBGU 300-W データセット[13-15]を用いて学習されたモデル)[16]を使用した．図 4.2 に示すように，Dlib を用いることで，顔の各器官を 68 点の特徴点として検出可能である．また，口唇の形状を 20 点の特徴点として取得することができる．なお，画像サイズが大きいことに起因して顔器官検出の処理速度が低下することを考慮し，3 章の 3.3.3 項における処理を行った．

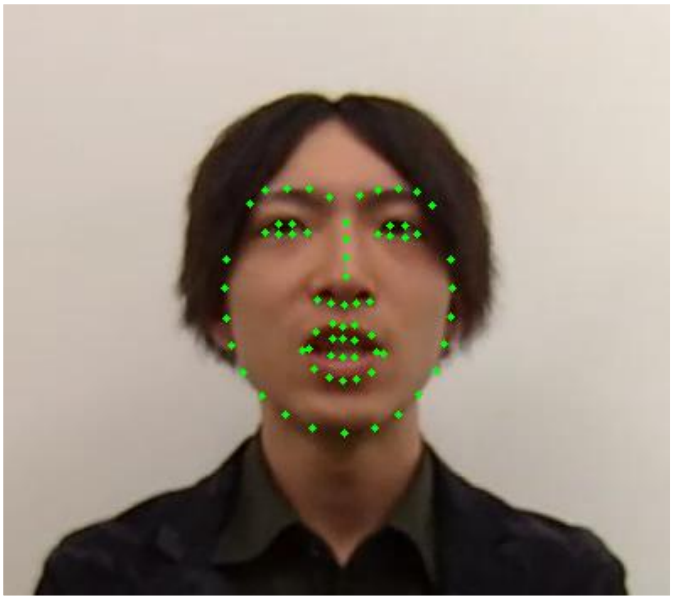


図 4.2　Dlib を用いて取得された顔器官における 68 点の特徴点例

4.3.3 口唇の縦幅および横幅の時系列変化算出処理

Dlib[12]により検出された顔器官の特徴点を用いて，発話動画データにおける口唇の縦幅および横幅の時系列変化を算出した．はじめに，図 4.3 に示すように，口唇上端および口唇下端の特徴点の距離を口唇の縦幅として算出した．次に，口唇左端および右端の特徴点の距離を口唇の横幅として算出した．最後に，発話動画データにおけるすべてのフレームに対して，口唇の縦幅および横幅を算出し，これらの時系列変化を取得した(図 4.4 参照)．

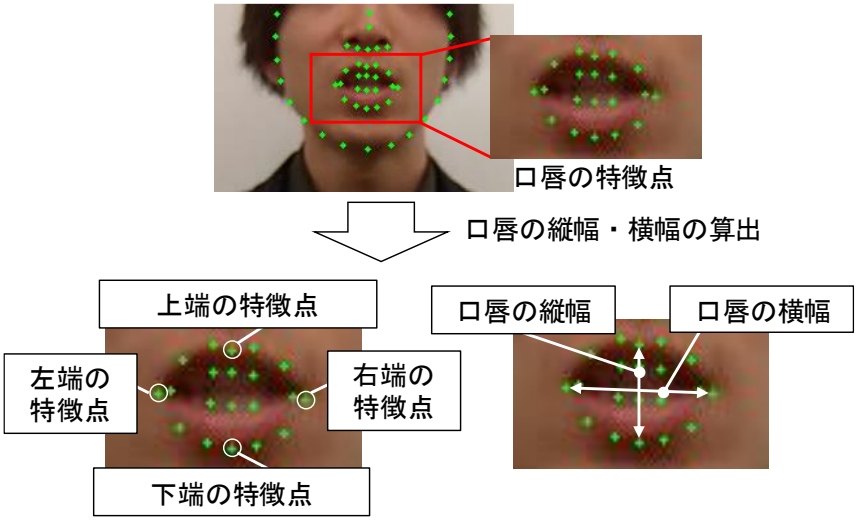


図 4.3　口唇縦幅および横幅の算出処理例

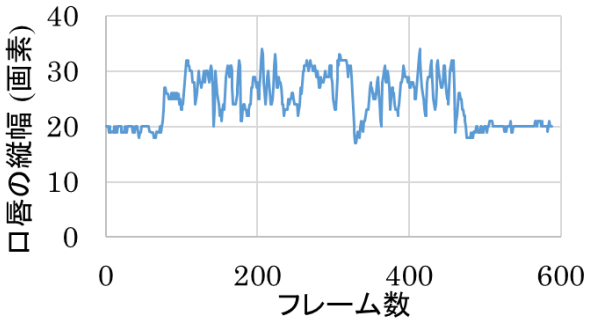


図 4.4　口唇縦幅の時系列変化算出結果例

4.3.4 口唇の縦幅および横幅の補正処理

発話時に被験者とカメラ間の距離が変動する場合があるため，発話以外に起因した口唇形状の変動が生じる．これは，発話者判別手法において判別精度低下の要因になる可能性があると考える．このため，口唇の縦幅および横幅の補正処理を行った．はじめに，図 4.5 に示すように鼻の特徴点を用いて，鼻の長さおよび鼻の横幅を算出した．次に，発話動画データにおける鼻の長さおよび鼻の横幅の時系列変化を取得した．最後に，式(4.1)および式(4.2)を用いて，各フレームにおける口唇の縦幅および横幅を補正し，以降の検討に使用した．補正結果例を図 4.6 に示す．なお，補正後の口唇の縦幅および横幅をそれぞれ縦幅の特徴量および横幅の特徴量と表記する．

$$\text{縦幅の特徴量} = \frac{\text{口唇の縦幅}}{\text{鼻の長さ}} \tag{4.1}$$

$$\text{横幅の特徴量} = \frac{\text{口唇の横幅}}{\text{鼻の横幅}} \tag{4.2}$$

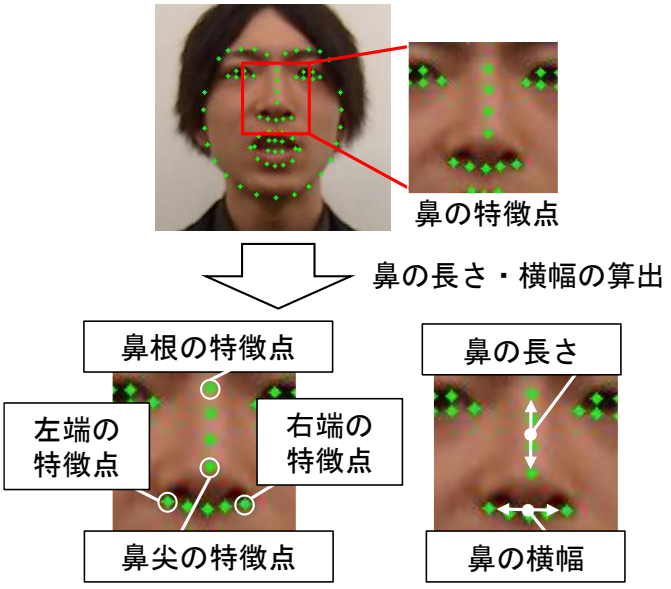


図 4.5　鼻の長さおよび横幅の算出例

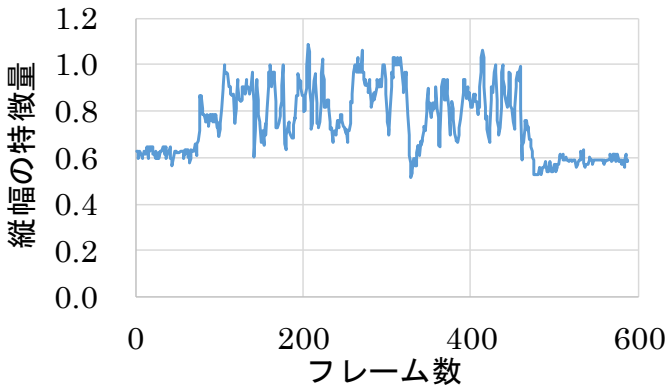


図 4.6　縦幅の特徴量の時系列変化算出結果例

4.3.5 口唇の動き特徴量算出処理

提案手法は，音声を用いて口唇の動きを推定し，これを発話者判別のために使用している．発話に伴う口唇の縦幅および横幅の動きはそれぞれ独立しており，発話した内容に依存したパターンで動作する．したがって，口唇の縦幅と横幅の時系列変化をそれぞれ単独で推定するのではなく，縦幅と横幅を複合した特徴量の時系列変化を推定することは，良好な発話者の判別を可能にすると考える．本論文では，(4.3)式を用いて縦幅と横幅を複合した特徴量を算出し，この時系列変化を口唇の動き特徴量として使用した．なお，(4.3)式は 4.4 節に後述する検討の結果に基づいて設定した．

$$\text{口唇の動き特徴量} = \text{横幅の特徴量} + \text{縦幅の特徴量} \tag{4.3}$$

4.3.6　Mel-frequency cepstral coefficients を用いた音声の特徴量取得処理

3 章の 3.3.8 項と同様に，音声の特徴量取得のために MFCC[17]を使用した．一般的に，MFCC の低いほうから連続した 10～15 次元程度が音声認識に使用される[17]．提案手法では，4.5 節の検討結果に基づき，MFCC の 0～13 次元の数値を音声の特徴量として使用した(図 4.7 参照)．

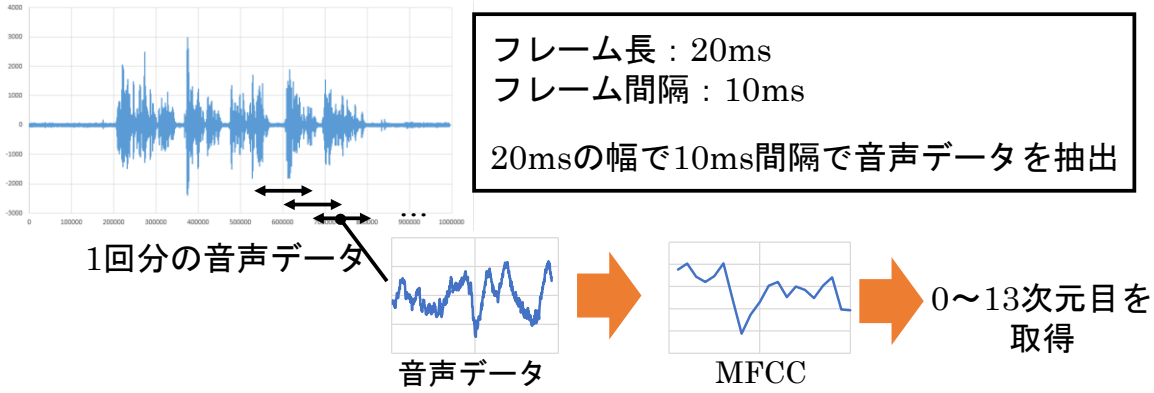


図 4.7　音声の特徴量取得処理例

4.3.7 特徴量の線形補間処理

発話動画データにおける画像のフレームレートと音声のサンプリングレートが異なるため，同時刻の区間における口唇の動き特徴量と音声の特徴量はデータ量が異なる．具体的には，1.0 秒間に口唇の動き特徴量は約 30 データ，音声の特徴量は約 100 データ含まれる．音声の特徴量のデータ量と口唇の動き特徴量のデータ量を等しくするために，式(4.4)および式(4.5)を用いて口唇の動き特徴量を線形補間した．線形補間の処理例を図 4.8 に示す．

$$F_n = \frac{lip\_total}{voice\_total} n \tag{4.4}$$

$$comp\_lip_n = (lip_{k+1} - lip_k) \times (F_n - k) + lip_k \tag{4.5}$$

ここで，

$comp\_lip_n$：補間後nフレーム目における口唇の動き特徴量，

lip_total：口唇の動き特徴量のフレーム数，

$voice\_total$：音声の特徴量のフレーム数，

F_n：n番目のフレームに対応する補間前のフレーム番号，

k：F_nの小数点を切り捨てた数値，

lip_k：kフレーム目の口唇の動き特徴量，

を表す．

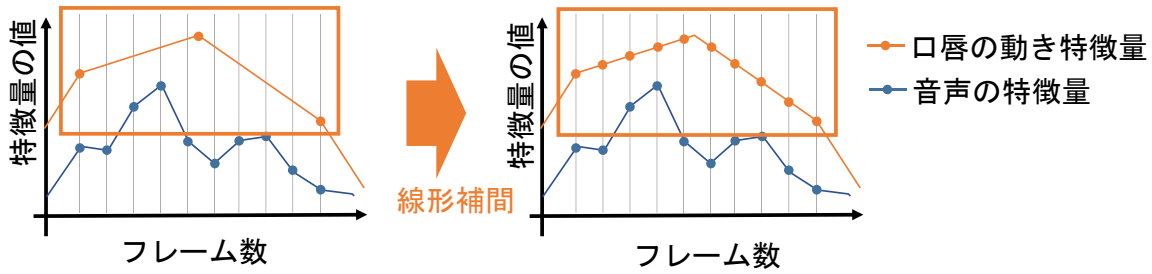


図 4.8　線形補間の処理例

4.3.8 平滑化処理および標準化処理

口唇の動き特徴量および音声の特徴量の時系列データにおけるノイズ除去を目的として，平滑化処理を実施した．このとき，平滑化後の特徴量の時系列変化は，可能な限り平滑化前の情報を保持していることが好ましい．このため，平滑化に使用する移動平均フィルタ[18]のサイズを適切に設定する必要がある．各文章の発話は，日本語の音節(「あ」や「も」などの平仮名 1 文字で表現される音)の連続した発話で構成される．したがって，平滑化前の情報を可能な限り保持した状態で，ノイズ除去処理を実施するためには，フィルタサイズを 1 音節以上，かつ 2 音節未満の長さにする必要があると考える．そこで，4.2.2 項で前述した予備検討用データを用いて, 1 音節当たりの平均発話時間を算出したところ, 約 130 ミリ秒である結果を得た．したがって，本平滑化処理では，1 音節以上かつ 2 音節未満の長さである 20 フレーム(1 フレーム当たり約 10 ミリ秒)を移動平均フィルタのサイズに設定した．

被験者間における口唇の動き特徴量および音声の特徴量の差異を低減するために，(4.6)式を用いて各特徴量の標準化を行った．なお，標準化は発話動画データごとに行った．

$$x' = \frac{x - \bar{x}}{s} \tag{4.6}$$

ここで,

x'：標準化後の特徴量,

x：標準化対象の発話動画データにおける任意のフレームの特徴量,

$\bar{x}$：特徴量の平均値（標準化対象の発話動画データごとに算出),

s：特徴量の標準偏差（標準化対象の発話動画データごとに算出),

を表す.

4.3.9 Long short-term memory (LSTM)

回帰結合型ニューラルネットワーク(Recurrent neural networks: RNN) [9]は時系列データを処理に特化したネットワークの 1 つである．また，RNN は可変長な時系列データに対する処理も可能であるという大きな特徴を有している．

一方，LSTM [9]は，RNN の回帰に加え，1 つのセル内で完結する内部回帰を有する(図 4.9 参照)．さらに，図 4.10 に示すように LSTM は，入力ゲート，忘却ゲート，ならびに出力ゲートの 3 つのゲートを有し，これらを用いて内部回帰の情報を制御している．入力ゲートは，内部回帰に対して新しい情報を追加するか否かを制御するゲートである．忘却ゲートは，内部回帰における情報を忘却するか否かを制御する．出力ゲートは，内部回帰における情報をどの程度，次の時刻に伝播するかを制御するゲートである．これら 3 つのゲートが内部回帰の情報を制御するため，RNN と比較して LSTM は，勾配消失問題が起こりにくい特徴を有している[9]．

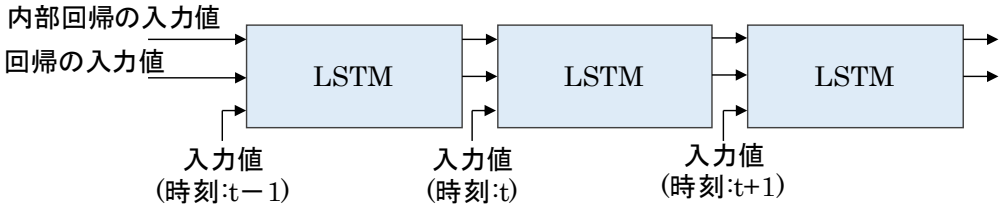


図 4.9　LSTM を用いた時系列データの処理例

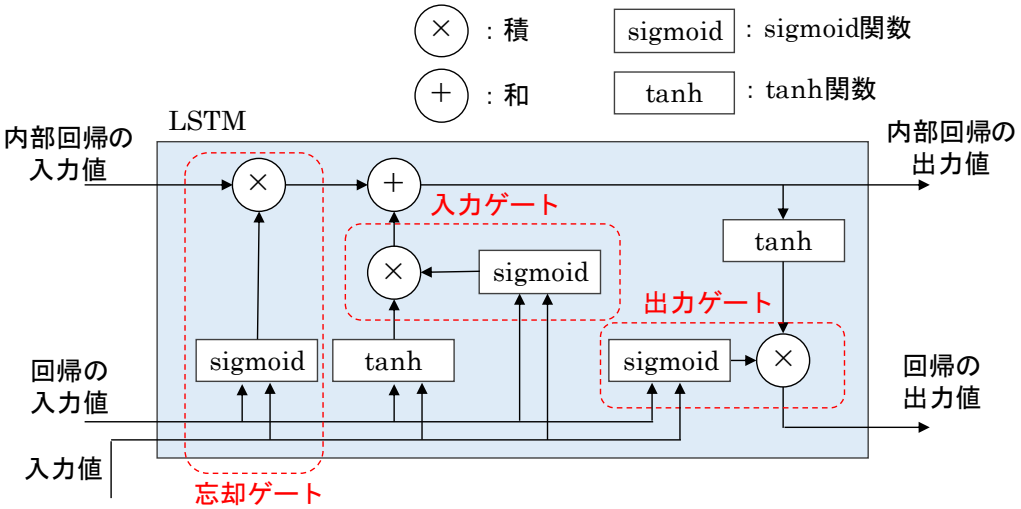


図 4.10　LSTM セルにおける 3 つのゲートと内部構造

4.3.10 学習処理

S. Suwajanakorn 氏ら[11]は，音声から口唇の動きを推定する場合，過去と未来の音声情報を使用することで，精度よく口唇の動きが推定可能であることを明らかにしている．そこで，本論文では図 4.11 に示すように，音声の特徴量 100 フレーム(約 1.0 秒間)を用いて，50 フレーム目の口唇の動き特徴量の推定値を出力する処理を行った．具体的には図 4.12 に示すように，入力層 14 次元，LSTM 層 125 次元，出力層 1 次元の 3 層構造の LSTM に対して，連続した 100 フレームの音声の特徴量を時系列順に入力した. 学習には勾配法として Adam [19]を使用し，100 フレーム目の音声の特徴量入力後における出力値と，50 フレーム目の口唇の動き特徴量との平均二乗誤差が最小になるように学習を行った．なお，中間層にあたる LSTM 層の次元数は，4.6 節に後述する検討の結果に基づいて良好な値を設定した.

LSTM の学習における勾配法には, Adam[19]を使用した. Adam は, Momentum [20]および AdaGrad [20]の利点を融合した手法である．Momentum は，物体が勾配方向に力を受け，速度が加速するような物理法則をベースとした重みの更新方法である．このため，ボールが傾斜を転がるように滑らかに重みが更新される利点を有する．AdaGrad は，個々の重みごとに学習係数を調整しながら学習を行うことが可能な勾配法である．また，大きく更新された重みの学習係数が小さくなるように設定し，更新が行われにくくすることが可能である．このため，重みの値を効率よく更新することが可能であり，最適解への収束速度が速いという特徴を有する.

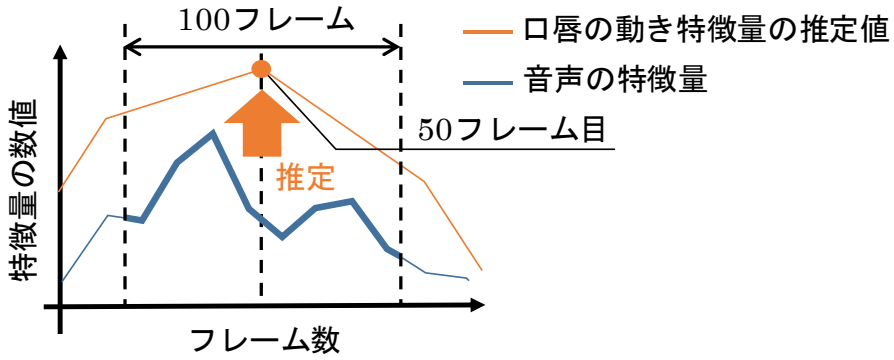


図 4.11　口唇の動き特徴量の推定処理例

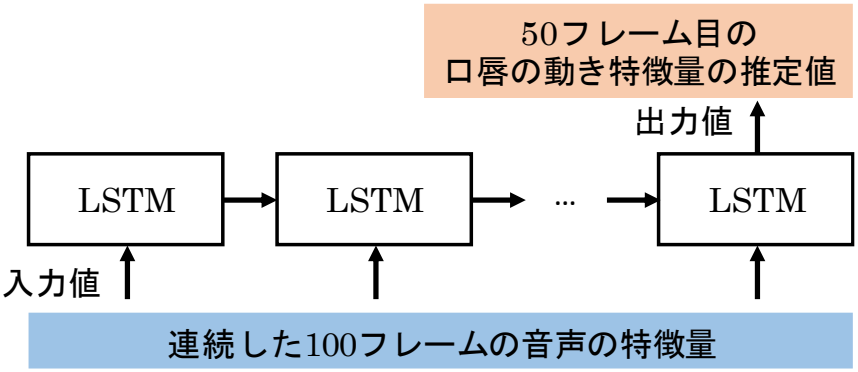


図 4.12　LSTM の入力および出力の例

4.3.11 発話者判別処理

提案手法では，口唇の動き特徴量の推定値と実際の値との類似度を算出し，最も類似度の高い被験者を発話者として判別した．発話者判別処理の流れを①～⑥に示す．

① 学習済みの LSTM を用いて，口唇の動き特徴量の推定値を算出した．
② 推定値および実際の口唇の動き特徴量から，同時刻における連続した 100 フレーム(以降，相関係数算出フィルタと表記する)の数値をそれぞれ取得した．なお，相関係数算出フィルタサイズに関しては，4.6 節の結果に基づき，良好な値を設定した．
③ ②で取得したフレームにおける推定値および実際の口唇の動き特徴量の時系列変化を算出し，これらの時系列変化間の相関係数を算出した．
④ ②で指定した相関係数算出フィルタを 1 フレームずつシフトさせ，すべてのフレームに対して②および③の処理を実施した．図 4.13 に，口唇の動き特徴量の推定結果例および相関係数[21]の算出結果例を示す．
⑤ 1 つの動画データにおいて，「算出した相関係数の総数」に対する「相関係数が 0.20 以上である事例の数」の割合を算出し，これを類似度とした．相関係数の閾値 0.20 に関しては，4.6 節の結果に基づいて良好な値を設定した．
⑥ 類似度の最も高い被験者を発話者として判別した．

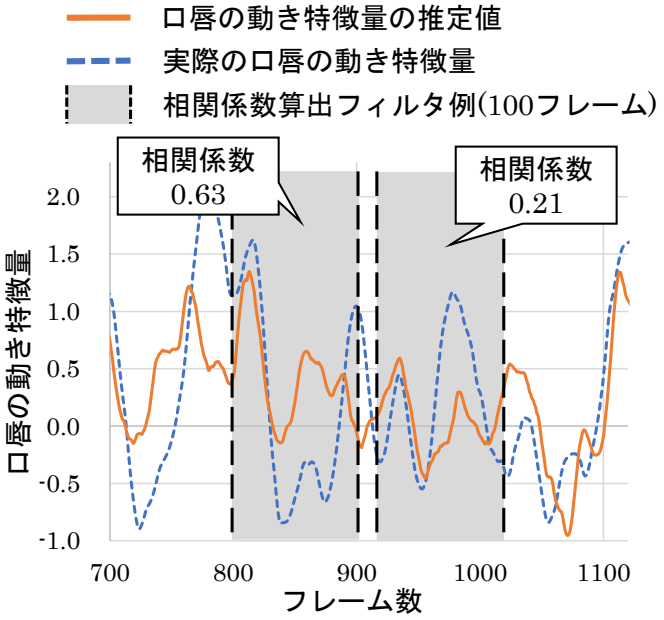


図 4.13　口唇の動き特徴量の推定および相関係数の算出結果例
(被験者 A，文章 1)

4.4 口唇の動き特徴量の選定

本節では，口唇の特徴点を用いて 7 種類の口唇の動き特徴量を算出し，発話者判別に有用な口唇の動き特徴量の検討を行った．具体的には，口唇の特徴点を 20 点使用して算出された口唇の動き特徴量(4.4.1 項参照)，および口唇の縦幅と横幅に基づいて算出された 6 種類の特徴量(4.4.2 項参照)を用いて，発話者判別を行い，良好な結果が得られた特徴量を選定した．

4.4.1 口唇の特徴点を 20 点使用した口唇の動き特徴量の算出

口唇の特徴点を使用した口唇の動き特徴量(以降，CV_nと表記する)の算出手順を①～⑥に示す．なお，4.3.7 項および 4.3.8 項と同様の処理を用いて，CV_n(n:1～20)ごとに線形補間，平滑化，ならびに標準化を行って，検討に使用した．

① オープンソースライブラリの Dlib における顔の特徴点検出機能を用いて，口唇における 20 点の特徴点および鼻の中央の特徴点を算出した．
② 鼻の中央の特徴点と口唇の特徴点間のベクトル(以降，V_nと表記する)をそれぞれ算出した(図 4.14 参照)．ここで，V_nはn番目の口唇の特徴点を使用して算出されたベクトルを示す．
③ 鼻の横幅と鼻の長さのベクトルをそれぞれ算出した(図 4.14 参照)．
④ 鼻の横幅のベクトルをx軸とし，V_nをx成分のベクトル(V_{nx})とy成分のベクトル(V_{ny})に分離した（図 4.15 参照)．
⑤ 分離したV_{nx}およびV_{ny}に対して(4.7)式および(4.8)式を適用し，各ベクトルを補正した．ここで，補正後のベクトルをそれぞれCV_{nx}およびCV_{ny}と定義する．
⑥ 補正後のベクトル(CV_{nx}およびCV_{ny})を合成し，合成後のベクトルの長さを補正後の口唇の動き特徴量CV_nとして算出した(図 4.15 参照)．CV_nを以降の検討に使用した．

$$CVn_x = \frac{Vn_x}{\text{鼻の横幅}} \tag{4.7}$$

$$CVn_y = \frac{Vn_y}{\text{鼻の長さ}} \tag{4.8}$$

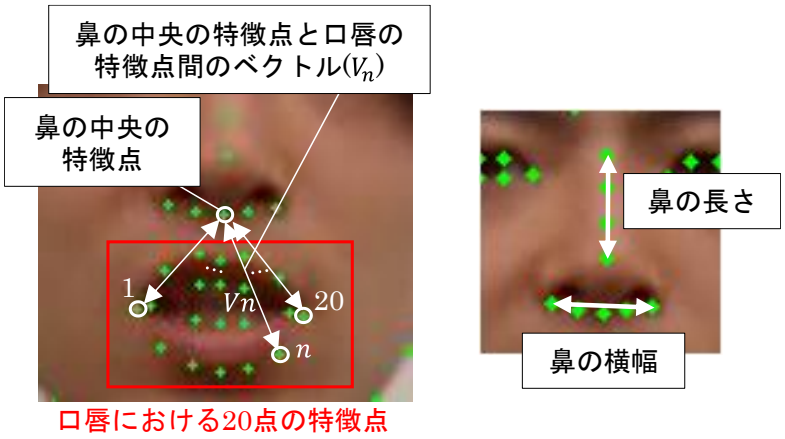
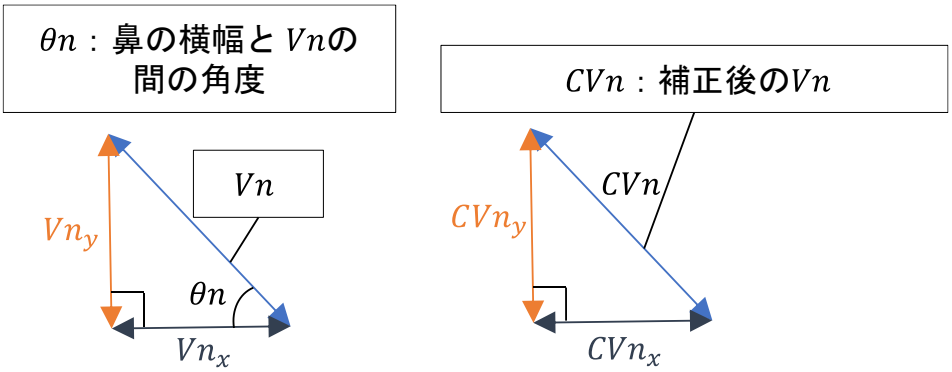


図 4.14　V_n，鼻の長さ，ならびに鼻の横幅の算出例



(a)ベクトルの分離　　(b)補正後のベクトルの合成

図 4.15　ベクトルの分離と補正
(V_n：鼻の中央の特徴点と口唇の特徴点間のベクトル)

4.4.2 口唇の縦幅と横幅に基づいた口唇の動き特徴量の算出

発話に伴う口唇の縦幅と横幅の動きに基づいて，6 種類の特徴量を算出し，検討に使用した．はじめに，口唇の縦幅の特徴量および横幅の特徴量を 4.3.4 項の手順で算出した．次に，(4.9)式～(4.12)式を用いて和の特徴量～商の特徴量を算出し，これらの時系列変化を取得した．本論文では，横幅，縦幅，和，差，積，ならびに商の特徴量を検討に使用した．なお，4.3.7 項および 4.3.8 項と同様の処理を用いて，特徴量ごとに線形補間，平滑化，ならびに標準化を行って，検討に使用した．

$$\text{和の特徴量} = \text{横幅の特徴量} + \text{縦幅の特徴量} \tag{4.9}$$

$$\text{差の特徴量} = \text{横幅の特徴量} - \text{縦幅の特徴量} \tag{4.10}$$

$$\text{積の特徴量} = \text{横幅の特徴量} \times \text{縦幅の特徴量} \tag{4.11}$$

$$\text{商の特徴量} = \frac{\text{横幅の特徴量}}{\text{縦幅の特徴量}} \tag{4.12}$$

4.4.3 口唇の動き特徴量の選定手順

4.4.1 項および 4.4.2 項で前述したCV_n，横幅，縦幅，和，差，積，ならびに商の特徴量のうち，発話者判別に最も有用な特徴量の選定を行った．このとき，7 種類の口唇の動き特徴量，4 パターンの相関係数算出フィルタ(4.3.11 項参照)のサイズ，ならびに音声の特徴量を 2 パターン設定し，合計で 56 パターンのパラメータの組み合わせに関して検討を加えた．本論文における音声の特徴量(MFCC [17])は，低い次元から連続した 10～15 次元程度が音声認識に使用される[17]．また，0 次元目の成分は直流成分を表していることから，他の次元と比べて物理的な意味が異なる．したがって，MFCC の「0～10 次元目」，および「1～10 次元目」の 2 パターンを設定し，発話者判別における MFCC の 0 次元目の有用性を検討した．検討には，予備検討用テストデータセットを用い，各パラメータのパターンにおける発話者判別成功率(4.4.4 項参照)の平均値を算出して比較した．なお，学習回数は 10～100 まで 10 回刻みで変更して検討し，最も良好な結果が得られたモデルを検討に使用した．パラメータの検討パターンを表 4.3 に示す．

表 4.3　口唇の動き特徴量の選定における検討範囲

特徴量とパラメータ名	検討範囲
音声の特徴量	• MFCC の 0～10 次元目 • MFCC の 1～10 次元目
口唇の動き特徴量	• CV_n • 横幅の特徴量 • 縦幅の特徴量 • 和の特徴量 • 差の特徴量 • 積の特徴量 • 商の特徴量
LSTM の入力層の次元数	音声の特徴量と同じ次元数
LSTM の中間層の次元数	50
LSTM の出力層の次元数	口唇の動き特徴量と同じ次元数
学習回数	10～100 回まで 10 回刻みで検討
相関係数算出フィルタのサイズ（W）	10, 50, 100, 150 フレーム
相関係数の閾値（Th）	0.40

4.4.4 発話者判別成功率の算出手順

本論文における発話者判別成功率の算出手順を①～⑥に示す．また，発話者判別成功率の算出手順の概要を図 4.16 に示す．ここで，テストデータセットとは「予備検討用テストデータセット」または「検討・評価用テストデータセット」を示し，テストデータとはテストデータセット内の 1 データセットを示す(例：PITD 1.1)．なお，本項における発話者判別成功率算出手順は，「予備検討用テストデータセット」および「検討・評価用テストデータセット」の両方に共通の手順である．

① テストデータセットのうち，評価対象のテストデータと異なる被験者および異なる文章を有する発話動画データを教師データに設定した．
② 教師データを用いて LSTM[9]モデルを学習した．
③ 評価対象のテストデータにおける任意の被験者と文章のデータを発話者データとして定義し，発話者データとは異なる被験者かつ異なる文章を有するデータを非発話者データとして定義した(図 4.16 参照)．
④ 発話者データおよび 1 つの非発話者データを対象として，発話者の判別を行った．はじめに，発話者データから口唇の動き特徴量と音声の特徴量を取得し，非発話者データから口唇の動き特徴量を取得した．次に，発話者データにおける音声の特徴量を学習済みの LSTM へ入力し，口唇の動き特徴量の推定値を算出した．最後に，発話者データおよび非発話者データの口唇の動き特徴量と推定値の類似度を算出し，発話者データにおける類似度が高い場合，発話者判別成功とした(図 4.16 参照)．
⑤ ②～④の手順をすべてのデータの組み合わせに対して実施し，テストデータごとに発話者判別成功率の平均値を算出した．なお，テストデータごとに発話者判別成功率の平均値が最も高い学習回数のモデルを検討に使用した．
⑥ ⑤において算出された各テストデータにおける発話者判別成功率の平均値を対象として，平均値を算出した．

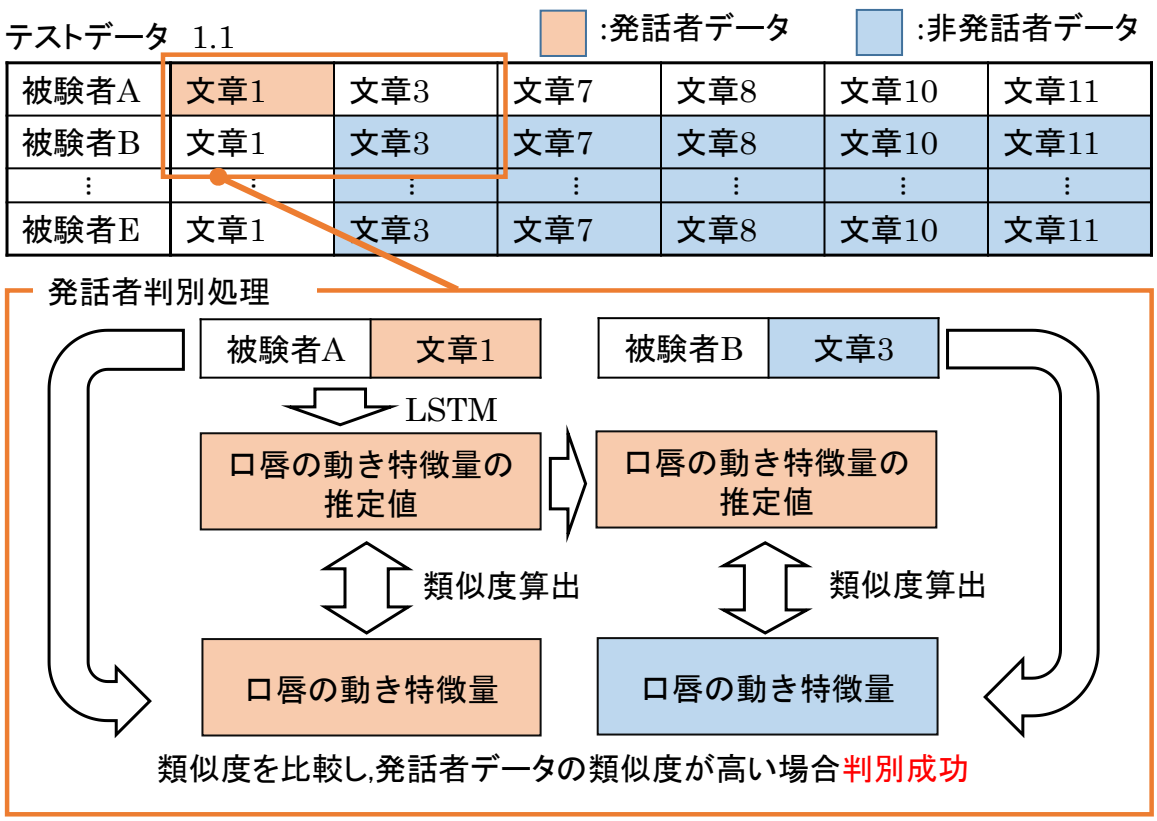


図 4.16　発話者判別成功率の算出手順

4.4.5 口唇の動き特徴量の選定結果

各パラメータのパターンにおける発話者判別成功率の平均値を表 4.4 および表 4.5 にまとめる．相関係数のフィルタサイズを 150，口唇の動き特徴量に和の特徴量，音声の特徴量として MFCC の 0～10 次元目を使用した場合において，最大の発話者判別成功率の平均値を得た．

以上の結果に基づき，口唇の動き特徴量として「和の特徴量」を使用し，かつ音声の特徴量は 0 次元目の値を含む範囲を選定して，以降の検討に使用した．

表 4.4　各パラメータのパターンにおける発話者判別成功率の平均値
（音声の特徴量として MFCC の 0～10 次元目を使用）　[%]

	口唇の動き特徴量						
相関係数算出フィルタ(W)	縦幅	横幅	和	差	積	商	CVn
10	72.5	66.9	75.3	63.4	78.3	69.8	64.8
50	72.3	71.3	77.4	61.6	81.3	65.6	69.2
100	72.2	71.0	81.9	62.5	80.9	64.1	71.1
150	75.1	68.2	**82.9**	63.1	78.1	62.3	75.1
平均	73.0	69.4	79.4	62.6	79.6	65.4	70.1

※赤字：表 4.4 および 4.5 の中で最大の値

表 4.5　各パラメータのパターンにおける発話者判別成功率の平均値
（音声の特徴量として MFCC の 1～10 次元目を使用）　[%]

	口唇の動き特徴量						
相関係数算出フィルタ(W)	縦幅	横幅	和	差	積	商	CVn
10	69.7	69.2	72.6	70.5	75.7	66.9	70.9
50	75.2	70.3	80.7	68.4	77.5	62.6	72.7
100	76.4	68.6	81.6	68.5	79.9	66.5	77.6
150	73.2	70.8	81.6	65.2	81.1	65.3	78.2
平均	73.6	69.7	79.1	68.2	78.5	65.3	74.9

4.5 音声の特徴量の選定

4.5.1 音声の特徴量の選定手順

検討・評価用テストデータセットを使用し，音声の特徴量における MFCC [17] の範囲について検討を行った．具体的には，MFCC の 0～n次元目(nは 10～19 まで 1 刻みで変更)までの範囲を検討した．評価に使用したパラメータのパターンを表 4.6 に示す．また，4.4.4 項と同じ手順で発話者判別成功率を算出し，音声の特徴量のパターンごとに発話者判別成功率の平均値を算出して，これらの結果を比較した．

表 4.6　音声の特徴量の選定における検討範囲

特徴量とパラメータ名	検討範囲
音声の特徴量	MFCC の 0～n 次元目 (n は 10～19 まで 1 刻み)
口唇の動き特徴量	和の特徴量
LSTM の入力層の次元数	音声の特徴量と同じ次元数
LSTM の中間層の次元数	50
LSTM の出力層の次元数	1
学習回数	10～100 回まで 10 回刻みで検討
相関係数算出フィルタのサイズ（W）	150 フレーム
相関係数の閾値（Th）	0.40

4.5.2 音声の特徴量の選定結果

図 4.17 および表 4.7 に各音声の特徴量における発話者判別成功率の平均値を示す．検討の結果，MFCC の 0～13 次元目までの数値を音声の特徴量として使用することで，最大の発話者判別成功率の平均値を得た．したがって，音声の特徴量として MFCC の 0～13 次元目を使用することで，発話者判別成功率が向上すると判断し，これを以降の検討で使用した．

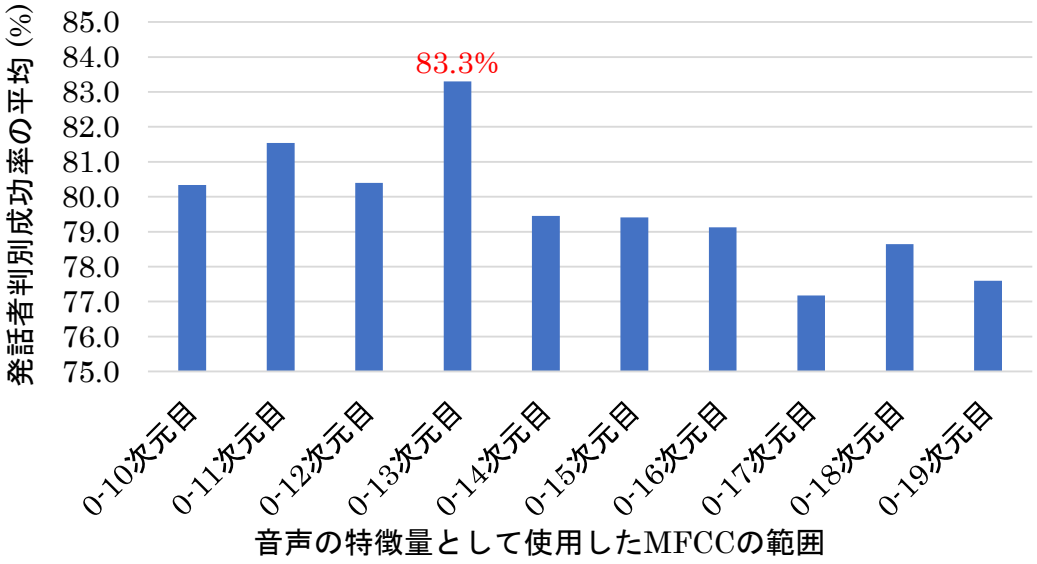


図 4.17　各音声の特徴量における発話者判別成功率の平均値

表 4.7　各音声の特徴量における発話者判別成功率の平均値

MFCC の次元数	発話者判別成功率の平均値［%］
0 次元目～10 次元目	80.3
0 次元目～11 次元目	81.5
0 次元目～12 次元目	80.4
0 次元目～13 次元目	83.3
0 次元目～14 次元目	79.4
0 次元目～15 次元目	79.4
0 次元目～16 次元目	79.1
0 次元目～17 次元目	77.2
0 次元目～18 次元目	78.6
0 次元目～19 次元目	77.6

4.6 パラメータの選定に関する検討

4.6.1 パラメータの選定手順

4.3.11 項で前述した「相関係数算出フィルタ」のサイズ，「相関係数の閾値」の適切な値，ならびに「LSTM[9]の中間層の次元数」に関して検討を行った．検討範囲を表 4.8 にまとめる．また，4.4.4 項と同じ手順を用いて発話者判別成功率を算出し，パラメータのパターンごとに発話者判別成功率の平均値を算出して，これらの結果を比較した．

表 4.8　音声の特徴量の選定における検討範囲

特徴量とパラメータ名	検討範囲
音声の特徴量	MFCC の 0～13 次元目
口唇の動き特徴量	和の特徴量
LSTM の入力層の次元数	音声の特徴量と同じ次元数
LSTM の中間層の次元数	25，50，75，100，ならびに 125
LSTM の出力層の次元数	1
学習回数	10～100 回まで 10 回刻みで検討
相関係数算出フィルタのサイズ（W）	10，50，100，ならびに 150
相関係数の閾値（Th）	0.10～0.90 まで 0.10 刻みで検討

4.6.2 パラメータの選定結果

表 4.9 に各相関係数の閾値における発話者判別成功率の平均値を示す．相関係数の閾値を 0.20 に設定した場合において，発話者判別成功率の平均値は 80.2%となり，最大の値を得た．このため，以降の検討では相関係数の閾値を 0.20 に設定した場合について評価を行った．

表 4.10 に相関係数の閾値を 0.20 に設定した場合において，得られた各パラメータの話者判別成功率を示す．相関係数算出フィルタを 100 とし，かつ LSTM の中間層の次元数を 125 に設定した場合において，87.2%の発話者判別成功率の平均値となり，最大の値を得た．

以上の結果から，提案手法におけるパラメータは，相関係数の閾値：0.20，相関係数算出フィルタ：100，ならびに LSTM の中間層の次元数：125 が適切であると判断し，得られた値を用いて以降の検討を行った．

表 4.9　各相関係数の閾値における発話者判別成功率の平均値

相関係数の閾値	発話者判別成功率の平均値 [%]
0.10	80.1
0.20	80.2
0.30	79.9
0.40	79.6
0.50	78.9
0.60	77.6
0.70	73.9
0.80	66.1
0.90	49.5

表 4.10　相関係数の閾値を 0.20 に設定した場合における
発話者判別成功率の平均値 [%]

相関係数算出フィルタ(W)	LSTM の中間層の次元数				
	25	50	75	100	125
10	66.5	68.9	77.8	80.9	79.8
50	72.8	76.2	82.2	83.9	84.8
100	76.8	81.1	84.9	85.5	87.2
150	77.2	81.2	84.7	84.9	86.6

4.7 各動画データを対象とした発話者判別成功率の評価

4.7.1 概要

提案手法と 4 つの既存手法(以降，比較手法 i～v と表記する)の発話者判別成功率を比較した．比較手法として，第 1 フォルマントに基づいて口唇の動きを推定する 3 つの手法(比較手法 i～iii) [10]および画像情報と音声情報を用いた発話者判別手法(比較手法 v) [8]を使用した．各比較手法を使用した発話者判別の手順を 4.7.2 項および 4.7.3 項に示す．

4.7.2 第 1 フォルマントに基づいた比較手法 i～iii について

音声信号の周波数成分は，声帯の振動や個人の声の高さによって変化するスペクトル微細構造，および発話に伴う声道や鼻腔の形状変化によって変化するスペクトル包絡構造の 2 つが含まれる[10]．比較手法 i～iii では，異なる方法をそれぞれ用いてスペクトル包絡構造を抽出し，評価に使用した． スペクトル包絡構造におけるピークを抽出したとき，最も低い周波数に位置するピークは，第 1 フォルマントと定義される[10]. 第 1 フォルマントの動きと顎の開閉には関連があることが知られているため[10], 第 1 フォルマントの動きから大まかな口唇の動きを推定することが可能であると考える．具体的には，第 1 フォルマントの周波数が高いほど，顎が開いていることを示す．

比較手法 i および ii では，MFCC [17]の値に基づいて第 1 フォルマントを算出し，口唇の動きの推定に使用した. はじめに，4.3.6 項と同じ手順を用いて MFCC の 0～19 次元の数値を算出した．次に，比較手法 i においては，MFCC の 1～13 次元の数値，比較手法 ii では MFCC の 1～19 次元目の数値すべてを対象とし，逆離散コサイン変換を行うことで，スペクトル包絡構造を抽出した．使用する MFCC の次元の範囲は，①MFCC の低次元成分はスペクトル包絡構造を表すこと[10]，②提案手法において良好な結果が得られたのは 13 次元目までの範囲であること，ならびに③本論文で使用した最大の次元数は 19 次元目までの範囲であることに基づいて設定した．横軸がケプストラム(時間成分と同義)である MFCC に対して逆離散コサイン変換を行うことで，横軸が周波数成分である波形に変換することが可能である．最後に，得られたスペクトル包絡構造から第 1 フォルマントを抽出し，第 1 フォルマントの周波数の時系列変化をそれぞれ算出した．

比較手法 iii では，線形予測分析(Linear Predictive Coefficient：LPC) [10]を用いてスペクトル包絡構造を算出し，検討に使用した．LPC とは，過去の波形値の組み合わせを用いて現在の波形値を予測することを目的とした手法である．

(4.13)式[10]に示すように，発話に伴う音声は，スペクトル微細構造$E(z)$，およびスペクトル包絡構造$H(z)$を組み合わせた波形$X(z)$として出力される．

$$X(z) = \frac{E(z)}{1 + \sum_{i=1}^{p} \alpha_i z^{-i}} = E(z)H(z) \tag{4.13}$$

ここで，

$X(z)$：口から発声される音声，

$E(z)$：声帯の音声信号（スペクトル微細構造），

$H(z)$：声道や口腔の形状による音声信号（スペクトル包絡構造），

α_i：線形予測係数，

z^{-i}：信号の複素数にあたる成分，

p：声道や口腔の形状を表現する音響管の個数（本論文では 12 に設定），

を表す．

LPC を用いて，声帯によるスペクトル微細構造を最小化するように線形予測係数α_iの値を算出することで，スペクトル包絡構造を算出することが可能である[10]．比較手法 iii では，LPC を用いて算出したスペクトル包絡構造から第 1 フォルマントを抽出し，第 1 フォルマントの周波数の時系列変化をそれぞれ算出した．

以上の手順を用いて算出した比較手法 i～iii における第 1 フォルマントの周波数の時系列変化は，発話動画データごとに標準化し，口唇の動き特徴量の推定値として使用した(図 4.18 参照)．また，比較手法 i～iii において使用する口唇の動き特徴量およびパラメータは，4.7.4 項に後述する手順で検討し，最も良好なものを使用した．なお，発話者判別成功率の算出方法に関しては 4.4.4 項と同様の方法を用いた．

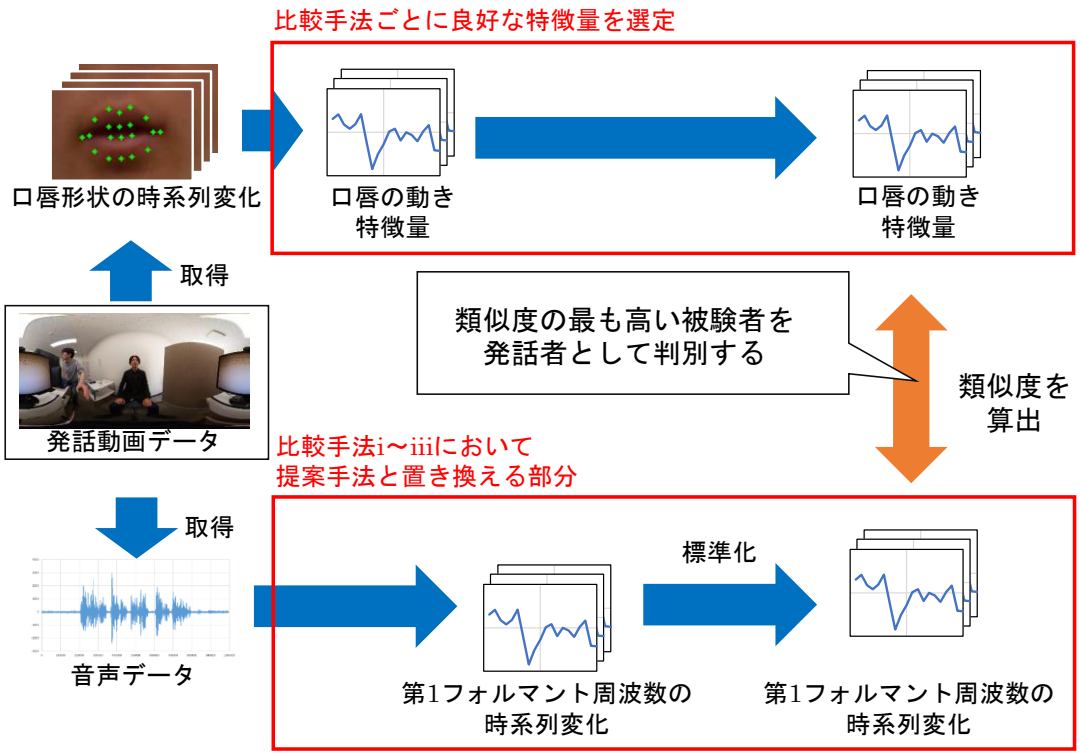


図 4.18　比較手法 i～iii における発話者判別の概要

4.7.3 画像情報と音声情報を用いた比較手法 v について

画像情報と音声情報を使用した発話者判別手法[8]を比較手法 v として使用した．この手法は，口唇の領域の画像と MFCC の 13 次元の数値との類似性を CNN および LSTM を用いて判別する手法である．比較手法 v を用いることで，発話動画データにおける各人物の口唇の動きと音声の類似度を，発話者度合としてフレームごとに算出することが可能である．本論文では，比較手法 v においても 4.4.4 項で前述した方法と同様にデータセットを設定して検討を行った(図 4.19 参照)．また，図 4.19 に示すように，比較手法 v を用いて発話者データと非発話者データの発話者度合を比較し，発話者データの発話者度合が高い場合は，判別成功フレームと定義した．さらに，判別成功フレームが 50.0%以上の場合を判別成功と定義し，発話者判別成功率を算出した．

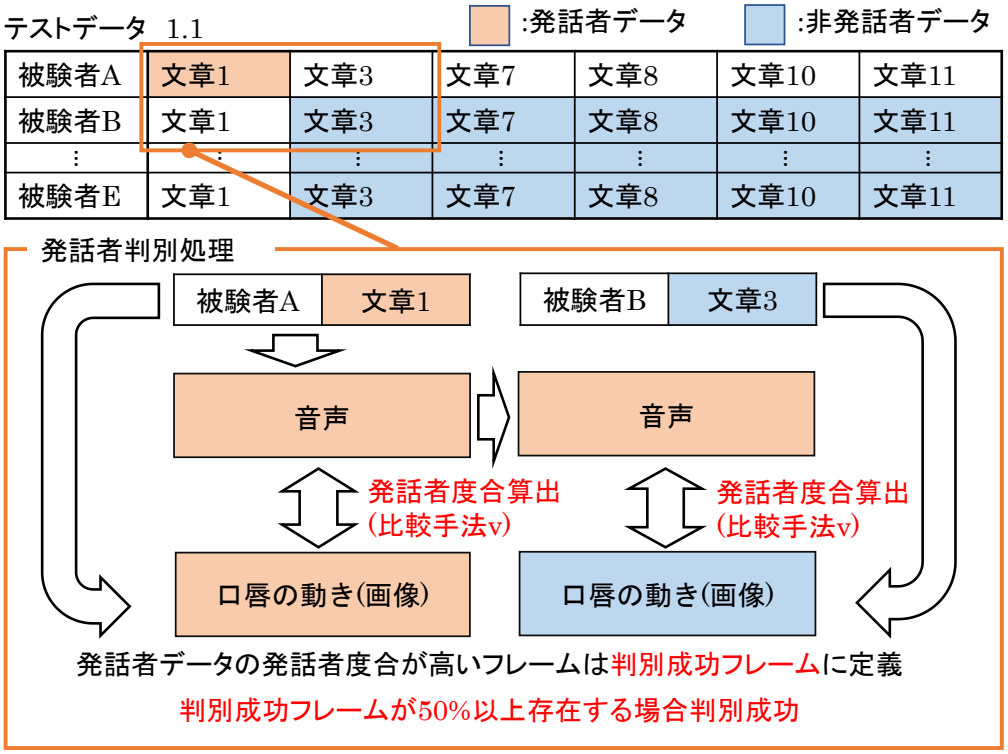


図 4.19　比較手法 v における発話者判別成功の判定方法

4.7.4 評価方法

検討・評価用テストデータセットを用いて，提案手法および比較手法 i～v における発話者判別成功率の平均値を算出して，比較した．このとき，比較手法 i～iii においては，口唇の動き特徴量，相関係数算出フィルタのサイズ，ならびに相関係数の閾値に関して検討を加え，発話者判別成功率の最も高いパターンを評価に使用した．各パラメータの検討範囲を表 4.11 に示す．

表 4.11　比較手法 i～v における特徴量とパラメータの検討範囲

特徴量とパラメータ名	検討範囲
口唇の動き特徴量	• 横幅の特徴量 • 縦幅の特徴量 • 和の特徴量 • 差の特徴量 • 積の特徴量 • 商の特徴量
相関係数算出フィルタのサイズ（W）	10, 50, 100, 150 フレーム
相関係数の閾値（Th）	0.10～0.90 まで 0.10 刻みで変更

4.7.5 評価結果

表 4.12 に比較手法 i～iii[10]において最も良好な結果が得られたパラメータのパターンを示す. 比較手法 i～iii では表 4.12 に示した特徴量とパラメータが有用であると判断し，以降の検討に使用した.

表 4.13 に提案手法および比較手法 i～v における発話者判別成功率の平均値をまとめる．比較手法 i～iii と比較し，提案手法はすべてのテストデータセットにおいて，発話者判別成功率が高い結果を得た．特に，テストデータ 1.2 において最大の発話者判別成功率となり，93.0%の値を得た．これらの結果は，提案手法は第 1 フォルマントを用いた手法[10]と比較して, 発話者判別に有用であることを示唆している.

一方，すべてのテストデータにおいて比較手法 v[8]の発話者判別成功率の値は 100.0%を得た. 比較手法 v はフレーム単位で発話者を判別することを前提とした手法である．そこで，4.7.3 項で前述した「判別成功フレーム」の割合に関して調査した．表 4.14 は各テストデータセット内の発話動画データに対して判別成功フレームの割合を算出し，これらの平均値，最大値，ならびに最小値を算出した結果である．比較手法 v における判別成功フレームの割合は，最小で 57.5%であるため，最大で 42.5%のフレームを誤って非発話者として判別していることがわかる．したがって，動画データ内のフレームごとに正しく発話者フレームを判別している区間と，誤って判別している区間が存在すると考える．このため，比較手法 v に対して発話者判別成功率の詳細な評価を行うためには，発話動画データにおける局所的な区間を複数設定し，これらに対して発話者判別性成功率の評価を行う必要があると考える．この点に関しては，4.8 項の検討で後述する.

表 4.12　比較手法 i～iii において最も良好な結果が得られた
パラメータのパターン

	比較手法 i	比較手法 ii	比較手法 iii
口唇の動き特徴量	縦幅	商	積
相関係数のフィルタサイズ（W）	50	50	10
相関係数の閾値（Th）	0.20	0.10	0.8
発話者判別成功率の平均値［%］	75.2	65.2	64.8

表 4.13　各手法の各テストデータセットにおける
発話者判別成功率の平均値 [%]

データセット	1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2	平均
提案手法	77.0	93.0	86.0	90.0	85.8	92.3	86.3	87.5	87.2
比較手法 i	66.3	75.3	76.0	78.5	78.2	73.8	77.2	76.8	75.2
比較手法 ii	55.2	45.0	56.5	66.8	82.2	85.8	72.0	58.5	65.2
比較手法 iii	54.8	64.5	59.8	70.8	62.0	68.5	62.7	75.0	64.8
比較手法 v	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

表 4.14　各テストデータセットにおける判別成功フレームの割合

データセット	1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2	平均
平均	85.1	85.4	81.0	81.2	85.5	85.3	83.0	83.3	83.7
最大	100.0	96.8	95.8	96.8	100.0	99.1	100.0	99.1	98.4
最小	62.0	65.4	60.8	58.5	57.5	66.3	58.3	58.0	60.8

4.8 データセットの作成コストと学習コストの評価

4.8.1 評価方法

提案手法と比較手法 v[8]を対象として，学習データのデータ量，データセット作成時間，ならびにモデルの学習時間について比較した．なお，比較手法 v のデータセット作成時間およびモデルの学習時間については，提案手法のデータセット作成時間およびモデルの学習時間に基づいて算出した．

4.8.2 評価結果

提案手法と比較手法 v における教師データのデータ量比較結果を表 4.15 に示す．比較手法 v に対する提案手法の教師データのデータ量は 0.05%であることがわかる．したがって，比較手法 v と比較して提案手法は，少ないデータ量でモデルの学習が可能であることが明らかになった．表 4.16 および表 4.17 に提案手法と比較手法 v における教師データの作成時間とモデルの学習時間をまとめる．提案手法における教師データの作成時間は平均で約 21.6 分，モデルの学習時間は最小で約 2.0 分，最大で約 17.2 分である．これらの結果は，比較手法 v と比較して，提案手法は教師データの作成コストとモデルの学習コストを大幅に削減可能であることを示唆している．

以上の結果は，提案手法は比較手法と比較して単位学習データあたりにかかる時間的コストが低いため，少ない学習時間でモデルの学習が可能であることを示唆している．

表 4.15　提案手法および比較手法 v における教師データのデータ量

提案手法の テストデータ	提案手法の 教師データ② [分]	比較手法 v の 教師データ① [分]	比率 [%] ① /② × 100
ITD 1.1	16.00	36360	0.04
ITD 1.2	16.66	36360	0.05
ITD 2.1	17.29	36360	0.05
ITD 2.2	17.44	36360	0.05
ITD 3.1	17.24	36360	0.05
ITD 3.2	18.09	36360	0.05
ITD 4.1	19.03	36360	0.05
ITD 4.2	19.06	36360	0.05
平均	17.60	36360	0.05

表 4.16　提案手法と比較手法 v における教師データ作成時間
(比較手法 v の教師データ作成時間は，提案手法の時間に基づいて算出)

	教師データの長さとフレームレート			1秒当たりのデータ長に対する処理時間 [秒]		教師データの平均作成時間	
	平均 [分]	画像 [fps]	音声 [kHz]	顔検出	MFCC	分	時間
提案手法	17.6	30.0	48.0	1.17	0.06	21.6	0.4
比較手法 v	36360.0	25.0	16.0	0.97	0.02	35996.4	599.9

表 4.17　提案手法と比較手法 v における学習回数の比較
(比較手法 v の学習時間は，提案手法の学習時間に基づいて算出)

	教師データの平均の長さ [分]	1 回あたりの学習時間 [分]	最小の学習回数		最大の学習回数	
			回数	学習時間 [分]	回数	学習時間 [分]
提案手法	17.6	0.2	10	2.0	100	17.2
比較手法 v	36360.0	354.6	1	354.6	20	7093.0

4.9 任意の長さの区間を対象とした発話者判別成功率の評価

4.9.1 評価方法

提案手法は 1 つの動画データの全区間を対象として，発話者判別成功率を算出する．しかしながら，実際の会議における発話動画データの長さは不定であるため，特定のサイズのフィルタを用いて任意の区間における発話動画データを抽出し，これに対して発話者判別を行う必要があると考える．また，4.7 節の検討結果を踏まえると，特定のサイズのフィルタを用いて抽出された区間を評価の対象とすることで，比較手法 v[8]の詳細な評価を行うことが可能であると考える．そこで，本評価では提案手法と比較手法 v に対して，同じ秒数のフィルタを適用し，発話動画データの任意の区間を抽出して発話者判別を行った．提案手法のフレームレートは 100fps，比較手法は 25fps であることに基づき，0.04 秒～2.00 秒の異なる大きさのフィルタを 10 パターン設定して検討を行った．図 4.20 に本評価の概要を示す．フィルタを用いて抽出した区間のうち，時刻が対応するデータ間に対して発話者判別処理を行った．なお，提案手法および比較手法 v における発話者判別成功の定義は，4.4.4 項および 4.7.3 項と同様である．

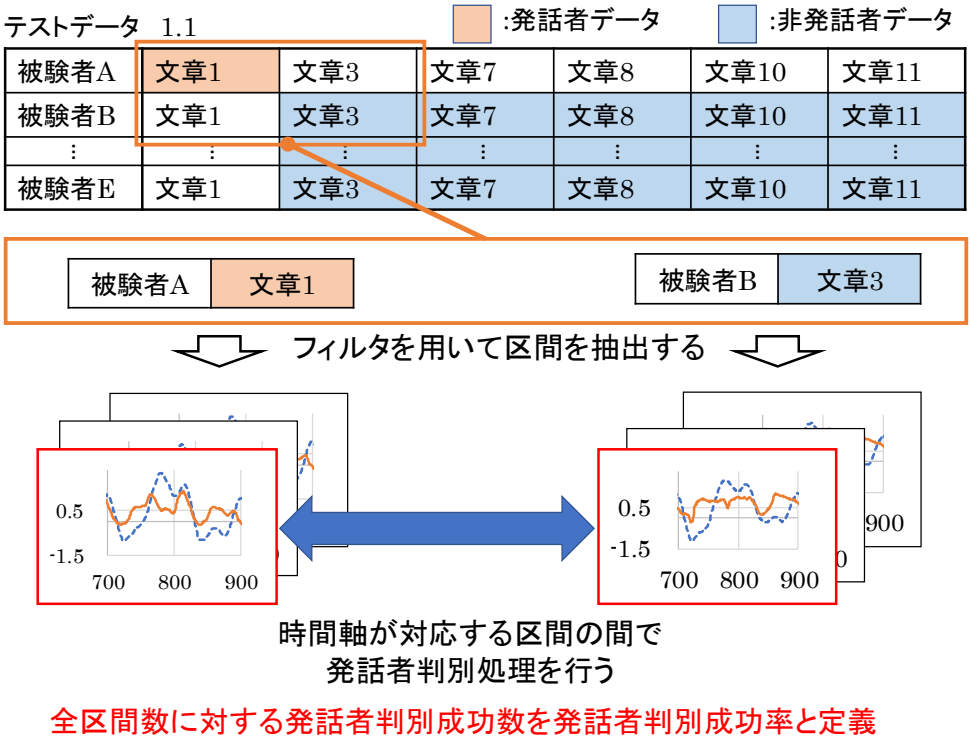


図 4.20　任意の長さの区間を対象とした発話者判別成功率の評価の概要

4.9.2 評価結果

表 4.18 に各フィルタサイズにおける提案手法と比較手法 v の発話者判別成功率をまとめる．提案手法の発話者判別成功率は，フィルタサイズ 1.80 秒で最大となり，73.41%の値を得た．同じフィルタサイズにおける提案手法と比較手法 v の比率に着目すると，提案手法は比較手法に対する 75.59%の発話者判別成功率を得たことがわかる．4.7 項の結果を踏まえると，提案手法と比較手法 v の教師データ量の比率が 0.05%にも関わらず，提案手法は比較手法に対する 75.59%の発話者判別成功率が得られる．

以上の結果は，提案手法は比較手法 v と比較して，少ない学習データを用いた発話者判別が可能であることを示唆している．

表 4.18　各フィルタサイズにおける発話者判別成功率の比較

フィルタサイズ			判別成功率［%］		比率
提案手法 [フレーム]	比較手法 v [フレーム]	秒	② 提案手法	②比較手法 v	①/② ×100 [%]
4.00	1.00	0.04	40.64	83.78	48.51
20.00	5.00	0.20	49.94	84.24	59.28
40.00	10.00	0.40	56.90	83.90	67.82
60.00	15.00	0.60	61.97	88.88	69.72
80.00	20.00	0.80	65.09	90.01	72.32
100.00	25.00	1.00	67.33	92.74	72.60
120.00	30.00	1.20	68.84	93.35	73.75
140.00	35.00	1.40	70.76	94.74	74.69
160.00	40.00	1.60	71.97	95.84	75.09
180.00	45.00	1.80	73.41	97.11	75.59
200.00	50.00	2.00	73.40	97.44	75.33

4.10 まとめ

本論文では，議事録自動作成システムにおける発話者判別機能の構築を目的とし，画像情報と音声情報を用いた発話者判別手法に関して検討を加えた．具体的には，音声から口唇の動きを推定し，推定値と実際の口唇の動きが最も類似している人物を発話者として判別する手法について検討を加えた．得られた成果を以下にまとめる．

(1) 提案手法は口唇の動き特徴量として「和の特徴量」，音声の特徴量として「MFCC の 0～13 次元目」を使用することで，良好な発話者判別成功率が得られることを明らかにした．

(2) 提案手法は第 1 フォルマントを用いた比較手法 i～iii [10]と比較して，発話者判別に有用であることを明らかにした．

(3) 提案手法は画像情報と音声情報を使用した比較手法 v[8]と比較して，教師データの量，教師データの作成時間，ならびに機械学習モデルの学習時間が少なく，1 つのモデル作成にかかる時間的コストが少ないことを明らかにした．

(4) 提案手法は画像情報と音声情報を使用した比較手法 v[8]に対する 0.05%の教師データを用いて，最大で 93.0%の発話者判別成功率を得た．また，1.80 秒間の動画データを使用した場合において，提案手法は比較手法 v に対する 75.59%の成功率で発話者判別が可能であることを明らかにした．

第 4 章 参考文献

[1] NEC ソリューションイノベータ：「Voice Graphy」，https://www.nec-solutioninnovators.co.jp/ss/smartwork/products/voicegraphy/ （Access 2021/12/15）

[2] Microsoft Azure, “Speaker Recognition”, https://azure.microsoft.com/ja-jp/services/cognitive-services/speaker-recognition/ （Access 2021/12/15）

[3] Z. Meng, M. Umair Bin Altaf, and B. Juang ： “Active voice authentication”, Digital Signal Processing, Vol. 101, 102672 (2020)

[4] X. Wang, F. Xue, W. Wang, and A. Liu：“A network model of speaker identification with new feature extraction methods and asymmetric BLSTM”, Neurocomputing, Vol. 403, pp. 167−181 (2020)

[5] R. Peri, M. Pal, A. Jati, K. Somandepalli, and S. Narayanan：“Robust speaker recognition using unsupervised adversarial invariance”, 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, 10.1109/ICASSP40776.2020.9054601 (2020)

[6] F. Grondin, and F. Michaud：“Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations”, Robotics and Autonomous Systems, Vol. 113, pp. 63−80 (2019)

[7] J.S. Chung, and A. Zisserman：“Out of time: automated lip sync in the wild”, Workshop on Multi-view Lip-reading, ACCV (2016)

[8] J.S. Chung, and A. Zisserman：“Learning to lip read words by watching videos”, Computer Vision and Image Understanding, Vol. 173, pp. 76−85 (2018)

[9] 斎藤康毅：「ゼロから作る Deep Learning -自然言語処理編-」,O’Reilly Japan, Inc. (2018)

[10] 古井貞熙：「ディジタル音声処理」，東海大学出版会（2001）

[11] S. Suwajanakorn, S.M. Seitz, and I. Kemelmacher-Shlizerman ： “Synthesizing Obama: Learning lip sync from audio”, ACM Transactions on Graphics, Vol. 36, No. 4, pp. 95:1–95:13 (2017)

[12] “Dlib C++ Liblary”， http://dlib.net/ (Accessed 2021/12/15)

[13] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic ： “300 Faces In-The-Wild Challenge: Database and results”,

Image and Vision Computing (IMAVIS), Vol.47, pp.3−18 (2016)

[14] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic : “A Semi-automatic Methodology for Facial Landmark Annotation”, Proceedings of IEEE Int’l Conf. Computer Vision and Pattern Recognition (CVPR-W), 5th Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2013). Oregon, DOI: 10.1109/CVPRW.2013.132 (2013)

[15] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic : “300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge”, Proceedings of IEEE Int’l Conf. on Computer Vision (ICCVW), 300 Faces in-the-Wild Challenge (300W), DOI: 10.1109/ICCVW.2013.59 (2013)

[16] V. Kazemi, and J. Sullivan : “One millisecond face alignment with anensemble of regression trees”, 2014 IEEE Conference on Computer Vision and Pattern Recognition, DOI:10.1109/CVPR.2014.241, Columbus, OH, USA (2014)

[17] 篠田 浩一 :「音声認識 Speech Recognition」, 講談社 (2017)

[18] 栗原 伸一 :「入門統計学ー検定から多変量解析・実験計画法までー」, オーム社 (2011)

[19] D.P. Kingma, and J. Ba : “Adam: A Method for Stochastic Optimization”, arXiv:1412.6980 (2017)

[20] 斎藤康毅 :「ゼロから作る Deep Learning - Python で学ぶディープラーニングの理論と実装 - 」, オーム社 (2018)

[21] 星野満博, 西崎雅仁 :「数理統計の探求 -経営的問題解決能力の開発と論理的思考の展開-」, 晃洋書房 (2012)

第 5 章 結論

働き方改革の実現や，新型コロナウイルスに伴う働き方の変化に伴い，多くの企業は長時間労働の改善策として作業の効率化に取り組んでいる．これらの取り組みの中で，会議の効率化が注目されている．会議における議事録は，議論した内容や取り決めの共有に有用であり，かつ会議の質を向上させるために役立てることができるため，会議の効率化に寄与すると考える．さらに，会議において，音声認識技術に基づいた議事録自動作成システムを利活用することは，議事録作成におけるヒューマンエラーの低減や議事録作成業務における工数の削減を可能にする．また，議事録自動作成システムによって作成されたテキストデータに対して，発言者を自動で割り振ることによって，議事録作成後に話者を割り振るための作業を簡略化することが可能であるため，業務の効率化に寄与すると考える．そこで本論文では，画像から取得可能な口唇の動きと音声情報を用いた発話者の判別手法の構築を目的とし，①口唇形状自動抽出法に関する検討，②発話区間の抽出手法に関する検討，ならびに③発話者の判別手法に関する検討を行った．以下に，本論文で得られた主な結果をまとめて記し，それに引き続いてこれらの工学的意義についてまとめる．

5.1 本論文により得られた主な知見

第 1 章では，本論文の主題である会議の効率化の必要性，実用化されている議事録自動作成システム，発話者判別手法の関連研究，ならびに本研究の目的および本研究に対する筆者の立場を述べるとともに，本論文の内容について述べた．

第 2 章では，画像情報と音声情報を用いた発話者判別手法の要素技術として，顔画像を用いた口唇形状自動抽出法ついて検討を加えた．具体的には，L*a*b* 色空間に基づき，口唇と肌の赤味に着目して口唇領域と肌領域をクラス分類する手法を提案した．さらに，提案手法の有用性について検討を加えたところ，次のような結果が得られた．

(1) 口唇領域における明度分布は，口裂の抽出に有用な特徴となり得ることを明らかにした．
(2) 顔画像における a*値は，FFNN を用いた口唇領域の抽出に有用な特徴量であることを明らかにした．
(3) 提案手法は，ファジィ推論に基づいた従来手法と比較して口唇を精度よく抽出可能であることを明らかにした．
(4) 提案手法は，輝度勾配に基づいた従来手法と比較し，少ない学習データを用いて同等の精度もしくは，それ以上の精度で口唇を抽出可能であることを明らかにした．

第 3 章では，画像情報と音声情報を用いた発話者判別手法の要素技術として，発話区間抽出手法に関して検討を行った．具体的には，動画データにおける口唇の動きが生じた区間，かつ音声の生じた区間を人物ごとに抽出することで，発話区間の抽出を行う手法を提案した．さらに，提案手法の有用性について検討を加えたところ，次のような結果が得られた．

(1) 提案手法の閾値自動設定処理において，閾値の算出のための説明変数として鼻の横幅を使用することが有用であることを明らかにした．

(2) 提案手法の閾値自動設定処理は，被験者とカメラ間距離の変動に伴う発話区間抽出精度の低下を低減可能である．また，最大で 0.99 の F-measure の平均値が得られ，高い精度で発話区間の抽出が可能であることを明らかにした．

(3) 画像情報と音声情報に基づいた既存の技術である比較手法と比較して，提案手法は高い精度で発話区間の抽出が可能であることを明らかにした．

4 章では，画像情報と音声情報を用いた発話者判別手法の要素技術として，発話者判別手法について検討を行った．具体的には，LSTM を用いて推定された口唇の動きを用いて，実際の口唇の動きと音声の類似性を評価し，最も類似した口唇の動きを有する人物を発話者として判別する手法について検討した．さらに，提案手法の有用性について検討を加えたところ，次のような結果が得られた．

(1) 提案手法は，口唇の動き特徴量として「和の特徴量」，音声の特徴量として「MFCC の 0～13 次元目」を使用することで，良好な発話者判別成功率が得られることを明らかにした．

(2) 提案手法は，第 1 フォルマントを用いた比較手法と比較して，発話者判別に有用であることを明らかにした．

(3) 提案手法は画像情報と音声情報を使用した比較手法と比較して，教師データの量，教師データの作成時間，ならびに機械学習モデルの学習時間が少なく，1 つのモデル作成にかかる時間的コストが少ないことを明らかにした．

(4) 提案手法は画像情報と音声情報を使用した比較手法 v に対する 0.05%の教師データを用いて，最大で 93.0%の発話者判別成功率を得た．また，1.80 秒間の動画データを使用した場合において，提案手法は比較手法 v に対する 75.59%の成功率で発話者判別が可能であることを明らかにした．

5.2 本論文の工学的意義

会議の効率化を目的とした利便性の高い議事録自動作成システムを構築するためには，発話者判別手法の開発が必要である．特に，事前に声紋の登録が不要であることや，少ない量の教師データを用いて発話者判別のためのモデル構築する技術の開発を行うことは，利活用しやすく，かつ利便性の高い議事録自動作成システムの構築に寄与すると考える．本論文では，①口唇領域自動抽出手法，②発話区間抽出手法，③発話者判別手法を提案することで，議事録自動作成システムにおける利便性の高い発話者判別手法の構築が可能であることを示した．以下に，本論文の工学的意義について述べる．

(1) 働き方改革の推進や新型コロナウイルスによる働き方の変化に伴い，新しい働き方のスタイルが普及する中，会議の効率化が重要視されている．すなわち，会議に関する業務の効率化を行うシステムの開発は最重要課題であるため，その工学的意義は大きい．

(2) 従来の口唇形状自動抽出手法は，肌の色の個人差や陰影の影響，ならびに顔の部位の一部が隠れていることに起因して，口唇領域を必ずしも正しく抽出できない課題がある．本論文では，取得した動画像における 1 フレーム目の情報を用いてモデルを学習する手法を提案し，提案手法は肌の色の個人差や陰影の影響による精度低下を低減可能であることを明らかにした．また，提案手法は口唇領域周辺の画像のみを用いて，口唇形状が自動抽出できるため，顔の他の部位が遮蔽されている場合においても口唇形状を抽出可能である．本研究で提案する「口唇の形状自動を抽出する技術」は発話者判別の他に，感情の検出技術や発話内容推定の技術などの幅広い分野で応用が可能であることから，画像処理やヒューマンセンシングの分野における工学的意義は大きい．また，提案手法は既存の手法と比較して少ない教師データを用いてモデルの構築が可能であるため，機械学習の分野に対する工学的意義も大きい．

(3) 発話区間の抽出方法として音声情報のみを用いて音声の有無を判別する手法(以降，音声区間検出と表記する)が提案されているものの，音声区間検出を用いて抽出された音声区間を発話者判別へ応用することは困難である．本論文では，画像情報と音声情報を併用した発話区間抽出手法を提案した．提案した発話区間抽出手法は，口唇の動きと音声区間が同時に生じているフレームを発話区間として抽出するため，発話している人物が 1 名の場合，その人物を発話者として判別することができる．また，提案手法は，機械学習を使用しないシンプルな手法を用いて，発話区間もしくは発話者を判別することが可能である．さらに，対面を想定した会議や，近

年普及が加速しているリモート会議などに応用可能であるため，その工学的意義は大きい．

(4) 従来の発話者判別手法において，会議参加者の声紋登録が必要な点やモデルの構築に膨大なデータが必要である点が課題である．これらの課題を解決することは，利便性の高い発話者判別手法の構築を可能にすると考える．本論文では，画像情報と音声情報を用いた発話者判別手法を提案した．提案手法は，音声情報を用いて推定した口唇の動きと，実際の口唇の動きの類似性に基づいて発話者の判別を行い，事前の声紋登録が不要な発話者判別を可能にする．また，発話者を判別する際に使用する特徴量の選定を行い，特徴量のパターンを減らすことで，モデルの構築に必要な教師データ量を削減した．さらに，声紋登録のような事前準備を必要としない発話者判別手法は，実用化されている議事録自動作成システムの利便性向上に寄与するため，その工学的意義は大きい．最後に，機械学習の分野では，大量の学習データを用いてモデルの学習を行うのが一般的である．少ない量の教師データを用いてモデルの学習が可能な点は，機械学習の関連分野に対する工学的意義が大きい．

5.3 今後に残された諸問題

最後に，今後に残された諸問題について述べる．

(1) 口唇形状自動抽出手法，発話区間抽出手法，発話者判別手法の統合

本論文において提案した「口唇形状自動抽出手法」,「発話区間抽出手法」，ならびに「発話者判別手法」を実用化に近い形で評価するためには，3 つの手法の統合が必要である．これらの手法を統合した上で，最終的な発話者判別成功率を評価し，改良を加えることで，提案手法の実用性と汎用性の向上を図る必要がある．

(2) 複数の人物が会話している動画データに対する有用性の評価

本論文において提案した「口唇形状自動抽出手法」,「発話区間抽出手法」，ならびに「発話者判別手法」の検討には，文章などを音読したデータを使用している．複数の人物が会話しているデータに対する有用性を検討することで，会話に対する提案手法の有用性や汎用性，および顔の状態や顔の角度が提案手法に与える影響などについて詳細な評価が可能であると考える．これらの評価結果を踏まえ，手法を改良することで，実用性の高い手法の構築を行う必要がある．

謝辞

本研究の遂行ならびに本論文の作成にあたって，終始懇切なるご指導とご鞭撻を賜りました秋田大学 教授 博士(工学) 景山 陽一 先生に心からお礼申し上げます.

本論文をまとめるにあたり，広い視野から数々の有益なご教示を頂きました秋田大学教授 博士（工学）有川 正俊 先生，ならびに同教授 博士（工学）水戸部 一孝 先生に深謝いたします.

本研究を進めるにあたり，多大なご協力，貴重な御助言を頂きました日本ビジネスシステムズ株式会社 廣瀬 聡 様，および関係各位に厚くお礼申し上げます.

本研究に関して貴重なご助言を頂きました放送大学秋田学習センター 所長 博士（工学）西田 眞 先生，秋田県立大学 博士（工学）石井 雅樹 先生，ならびに前秋田大学 技術専門員 博士（工学） 髙橋 毅 氏に謝意を表します.

本研究は，秋田大学 大学院理工学研究科 数理・電気電子情報学専攻 人間情報工学コース 景山研究室において行われたものです．本研究の遂行において適切な助言を与えて下さった秋田大学 准教授 博士（工学）横山 洋之 先生，同准教授 博士（工学）石沢 千佳子 先生，同助教 博士（工学）白井 光 先生，同助教 鄒 敏 先生，ならびに 伊藤 悠大 技術職員をはじめ，景山研究室の皆様，卒業生の皆様に心から感謝いたします.

本研究の一部は，JSPS 科研費 No. 15K00222，No. 19K12909，No. 21J15592 の助成を受けて行われたことを付記し，関係機関各位に厚くお礼申し上げます.

最後に，大学院博士後期課程への入学について理解を示し，在学中の支えとなってくれた家族と友人に心から感謝いたします.

本論文の第 2 章は, 電気学会論文誌 C 掲載論文「中村悦郎, 髙橋毅, 景山陽一, 石井雅樹, 西田眞：「順伝播型ニューラルネットワークを用いた口唇形状自動抽出法」, 電気学会論文誌 C, Vol.138, No. 12, pp. 1604−1612 (2018)」を基に執筆したものです. また, 本論文の第 3 章は, IEEJ Transactions on Electrical and Electronic Engineering「E. Nakamura, Y. Kageyama, and S. Hirose：“LSTM-based Japanese Speaker Identification Using an Omnidirectional Camera and Voice Information”, IEEJ Transactions on Electrical and Electronic Engineering, Vol. 17, No. 5, DOI:10.1002/tee.23555 (2022)」を基に執筆したものです. 加えて, 本論文の第 4 章は, Journal of Advanced Computational Intelligence and Intelligent Informatics へ投稿中の論文「E. Nakamura, Y. Kageyama, and S. Hirose：“Speech-Section Extraction Using Lip Movement and Voice Information in Japanese”, Journal of Advanced Computational Intelligence and Intelligent Informatics (投稿済み, 査読中)」を基に執筆したものです (本論文 146 頁, 本研究に関する発表論文 [学術論文誌] (1)～(3)参照).

- 中村悦郎, 髙橋毅, 景山陽一, 石井雅樹, 西田眞：「順伝播型ニューラルネットワークを用いた口唇形状自動抽出法」, 電気学会論文誌 C, Vol.138, No. 12, pp. 1604−1612 (2018)
- E. Nakamura, Y. Kageyama, and S. Hirose：“LSTM-based Japanese Speaker Identification Using an Omnidirectional Camera and Voice Information”, IEEJ Transactions on Electrical and Electronic Engineering, Vol. 17, No. 5, DOI:10.1002/tee.23555 (2022)
- E. Nakamura, Y. Kageyama, and S. Hirose：“Speech-Section Extraction Using Lip Movement and Voice Information in Japanese”, Journal of Advanced Computational Intelligence and Intelligent Informatics (投稿済み, 査読中)

本研究に関する発表論文

学術論文誌

レフェリー制のある学術雑誌

○(1) 中村悦郎，髙橋毅，景山陽一，石井雅樹，西田眞：「順伝播型ニューラルネットワークを用いた口唇形状自動抽出法」，電気学会論文誌 C，Vol.138, No. 12，pp. 1604−1612 (2018)

○(2) E. Nakamura, Y. Kageyama, and S. Hirose：“LSTM-based Japanese Speaker Identification Using an Omnidirectional Camera and Voice Information”, IEEJ Transactions on Electrical and Electronic Engineering, Vol. 17, No. 5, DOI:10.1002/tee.23555 (2022)

(3) E. Nakamura, Y. Kageyama, and S. Hirose：“Speech-Section Extraction Using Lip Movement and Voice Information in Japanese”, Journal of Advanced Computational Intelligence and Intelligent Informatics (投稿済み，査読中)

国際会議

(1) T. Takahashi, Y. Kageyama, M. Ishii, M. Nishida, E. Nakamura, and K. Fujisawa：“Study of Differences in Individual Behavior Regarding Not-so-good Conditions and Corresponding Changes in Lip Motion while Speaking”, The 6th IIAE International Conference on Intelligent Systems and Image Processing, PS-12, pp.395−400 (2018)

(2) E. Nakamura, T. Takahashi, Y. Kageyama, M. Ishii, M. Nishida, and M. Shirasu：“Shadow Removing Method for Lip Shape Extraction Using Feedforward Neural Network”, 2019 IEEE 1st Global Conference on Life Sciences and Technologies, POS-1.13, pp.87−88 (2019)

(3) E. Nakamura, T. Takahashi, Y. Kageyama, M. Ishii, M. Nishida, and M. Shirasu：“Improvement of Lip Extraction Method Using Interpolation Method”, The 7th International Conference on Intelligent Systems and Image Processing 2019, GS11-1, pp.286−291 (2019)

(4) E. Nakamura, Y. Kageyama, and M. Shirasu：“Speaker Identification Method Using Facial Image and Voice”, The 2020 IEEE 2nd Global Conference on Life Sciences and Technologies, No.RET2.4, pp. 411−412 (2020)

(5) M. Tanaka, E. Nakamura, Y. Kageyama, M. Shirasu, M. Ishii, and M. Nishida : “Identification of Utterance Content Using Lip Movement Features”, The 2020 IEEE 2nd Global Conference on Life Sciences and Technologies, pp.167−168 (2020)

(6) E. Nakamura, Y. Kageyama, and M. Shirasu : A Study on Feature Values as a Speaker Identification Method, Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems, #1570655520, pp.314−315 (2020)

(7) E. Nakamura, Y. Kageyama, and S. Hirose : “Speech Section Extraction Method Using Image and Voice Information”, The 9th IIAE International Conference on Industrial Application Engineering 2021, OS1-2, pp. 30−34 (2021)

(8) E. Nakamura, Y. Kageyama, and S. Hirose : “Application of Speech Section Extraction Method Using Lip and Voice”, The 9th International Conference on Materials Engineering for Resources, BP-5 (2021)

(9) E. Nakamura, Y. Honda, Y. Kageyama, and S. Hirose : “Facial Feature Points for Japanese Speech Content Estimation”, 2022 IEEE 4th Global Conference on Life Sciences and Technologies, 1570781476 (2022)

口頭発表

(1) 中村悦郎，髙橋毅，景山陽一，石井雅樹，西田眞：「心理・体調変化の解析を目的とした口唇の動き抽出法の改善」，平成 28 年度日本知能情報ファジィ学会東北支部研究会，B2-4 (2016)

(2) 中村悦郎，髙橋毅，景山陽一，石井雅樹，西田眞：「口唇の動きを用いた心理・体調変化の解析における特徴量抽出法の改善 」，平成 29 年度電気関係学会東北支部連合大会，1H16 (2017)

(3) 中村悦郎，髙橋毅，景山陽一，石井雅樹，西田眞：「室内環境下における発話に伴う口唇の動き抽出手法の改善」，平成 29 年度照明学会全国大会，10-04 (2017)

(4) 中村悦郎，髙橋毅，景山陽一，石井雅樹，西田眞：「心理・体調変化の解析を目的とした口唇の動き抽出法の提案」，第 18 回　計測自動制御学会システムインテグレーション部門講演会，1C5-08 (2017)

(5) 中村悦郎，髙橋毅，景山陽一，石井雅樹，西田眞，清水剛：「ニューラルネットワークを用いた口唇形状抽出法に関する検討」，平成 30 年度電気関係学会東北支部連合大会，1D01 (2018)

(6) 中村悦郎，髙橋毅，景山陽一，石井雅樹，白須礎成：「NN を用いた口唇形状抽出手法のネットワーク構造と学習条件に関する検討」，第 61 回自動制御連合講演会，4A5 (2018)

(7) 中村悦郎，髙橋毅，景山陽一，石井雅樹，西田眞，白須礎成：「照明条件にロバストなニューラルネットワークによる口唇抽出法に関する検討」，平成 30 年度 情報処理学会東北支部研究会，9 (2018)

(8) 中村悦郎，景山陽一，白須礎成：「全方位カメラを用いた発話区間判別手法に関する基礎検討」，2019 年度電気関係学会東北支部連合大会，2F07 (2019)

(9) 中村悦郎，景山陽一，白須礎成: 「全方位カメラを用いた発話者判別に関する基礎検討」，第 62 回自動制御連合講演会，1J4-04 (2019)

(10) 中村悦郎，景山陽一，白須礎成：「全方位カメラを用いた発話区間抽出手法の改善」，2020 年度電気関係学会東北支部連合大会，S02 (2020)

(11) 中村悦郎，景山陽一，白須礎成：「全方位カメラを用いた発話者判別手法の機械学習モデルに関する検討」，第 23 回 画像の認識・理解シンポジウム(MIRU2020)，IS2-2-18 (2020)

(12) 中村悦郎，景山陽一，白須礎成：「人物とカメラ間距離の変動を考慮した全方位カメラによる発話区間抽出法」，第 63 回 自動制御連合講演会，1I1-5 (2020)

(13) 中村悦郎，景山陽一，白須礎成：「画像情報と音声情報を用いた発話区間自動抽出手法」，映像情報メディア学会 創立 70 周年記念大会，31B-2 (2020)
(14) 中村悦郎，景山陽一，廣瀬聡：「画像情報と音声情報を用いた発話者判別手法における汎用性に関する検討」，情報処理学会 第 83 回全国大会，4C-03 (2021)
(15) 本田悠将，中村悦郎，景山陽一，廣瀬聡：「LSTM を用いた口唇画像による発話内容推定に関する基礎検討」，令和 2 年度日本知能情報ファジィ学会東北支部研究会，4-3 (2021)
(16) 中村悦郎，景山陽一，廣瀬聡：「全方位カメラを用いた発話者判別手法における音声特徴量に関する検討」，2021 年度電気関係学会東北支部連合大会，1D07 (2021)
(17) 中村悦郎，本田悠将，景山陽一，廣瀬聡：「発話内容の判別を目的とした口唇の動きの変化解析」，第 37 回 ファジィ システム シンポジウム，TE4-1 (2021)
(18) 本田悠将，中村悦郎，景山陽一，廣瀬聡：「口唇の動きおよび色情報を用いた発話内容推定に関する検討」，第 37 回 ファジィ システム シンポジウム，TE4-4 (2021)
(19) 中村悦郎，本田悠将，景山陽一，廣瀬聡：「発話に伴う口唇特徴点の動きと音素との関連性に関する解析」，映像情報メディア学会 2021 年冬季大会，13A-4 (2021)
(20) 本田悠将，中村悦郎，景山陽一，廣瀬聡：「口唇の動きおよび色情報を用いた発話内容推定に関する検討」，映像情報メディア学会 2021 年冬季大会，21A-3 (2021)
(21) 本田悠将，中村悦郎，景山陽一，廣瀬聡：「CNN-LSTM を用いた顔画像による発話内容推定に関する基礎検討」，情報処理学会第 84 回全国大会，2Q-07 (2022)