

機械学習ライブラリを利用した超音波画像からの舌輪郭抽出の試み

的場 瞳¹ 北村 達也² 孫 静³ 林 良子³

^{1,2} 甲南大学知能情報学部 〒 658-8501 兵庫県神戸市東灘区岡本 8-9-1

³ 神戸大学大学院国際文化学研究科 〒 657-8501 兵庫県神戸市灘区鶴甲 1-2-1

E-mail: ²t-kitamu@konan-u.ac.jp

あらまし 超音波診断装置により撮影した舌運動の動画から機械学習を利用して舌輪郭の抽出を試みた。本研究では, Takemoto ら (2019) が MRI 動画の音声器官の輪郭抽出に用いている機械学習ライブラリ dlib のアンサンブル回帰木に基づく学習器を超音波画像に適用した。超音波診断装置の画面をキャプチャした動画からフレームを抽出し, 目視にてトレースした。10 フレーム分のデータを用いて学習を試みたところ, おおむね良好な結果が得られ, dlib が超音波画像にも有効であることを確認できた。

キーワード 超音波画像, 舌輪郭, 輪郭抽出, アンサンブル回帰木

Tongue Contour Tracking in Ultrasound Images using Machine Learning Library

Hitomi Matoba¹, Tatsuya Kitamura² SUN Jing³ Ryoko Hayashi³

^{1,2}Faculty of Intelligence and Informatics, Konan University 8-9-1, Okamoto, Higashinada, Kobe, Hyogo, 658-8501, Japan

³Graduate School of Intercultural Studies, Kobe University 1-2-1, Tsurukabuto, Nada, Kobe, Hyogo, 657-8501, Japan

E-mail: ²t-kitamu@konan-u.ac.jp

Abstract This preliminary study attempts to detect the edge of the tongue dorsum from an ultrasound movie by a machine learning method. The authors adopt dlib, a machine learning library that Takemoto *et al.* (2019) have used for extracting the edge of the speech organs from real-time MRI movies. The training data is the edge of the tongue dorsum traced manually from 10 frames of the ultrasound movie. The results show that dlib can extract the edge of the tongue dorsum almost accurately even for indistinct ultrasound images.

Keywords Ultrasound images, Tongue contour, Edge extraction, Ensemble of Regression Trees

1 はじめに

著者らは, 延伸や途切れ [1] を含む発話における発話運動 [2] や音声障害患者の発話運動の調査のために汎用超音波診断装置を導入した。現在は観測のノウハウを蓄積しつつ予備的な実験を行っている段階である。超音波診断装置は, 口蓋が造影されないなどの制約はあるものの, MRI と比較して導入や維持のコストが低く, 時間分解能が比較的高いという利点がある [3]。海外では, 以前より音声言語や音声障害の分野における超音波診断装置の利用が進んでおり, この領域の国際会議 (UltraFest) が定期的に開催されている。

音声分野にて超音波診断装置を利用する際には, 舌の運動を時間分解能は 60 fps から 100 fps の動画として記録することが一般的である。得られた動画を分析するためには, 舌の輪郭のトレースや調音点の特定が必要であるが, これらを全て手作業で行うと人的コストが大きくなる上, 再現性の問題も生じる [3]。そこで, この作業を自動化あるいは半自動化する試みが行われてきた。例えば, Li ら [4] は動的輪郭モデル Snakes [5] に基づく EdgeTrak¹を開発した。EdgeTrak は, ユーザーが動画のうちの 1 フレームをトレースするだけで, 後続のフレームの輪郭の候補

¹EdgeTrak のコンパイルについては付録 A を参照のこと。

が示される。また、Zhu ら [6] は深層学習 (Convolutional neural networks) に基づくツール MTracker²を公開しているが、残念ながらユーザーが自分のデータを学習させて精度を向上させることはできない。

一方で、リアルタイム MRI 動画を対象とした研究では、Takemoto ら [7] が機械学習ライブラリ dlib [8] を用いて音声器官の輪郭抽出を試みて大きな成果を得ている。彼らは、訓練データを適切に与えることによって、少量の学習データから訓練された作業者のトレースと同レベルの精度で輪郭を自動抽出できることを報告した。また、抽出した輪郭から平均声道を生成も試みている [9]。そこで、本研究では、超音波動画に対しても dlib が有効か否かを検討する。なお、dlib には様々な機能があり、Takemoto らが利用しているのは Kazemi and Sullivan [10] の顔特徴点抽出法を実装した機能であるが、本稿では簡単のためにこの機能を dlib と呼ぶ。

2 方法

2.1 データ収集

本研究では、General Electric (GE) Healthcare 社のポータブル型超音波診断装置 LOGIQ e Premium にマイクロコンベックスプローブ (8C-Rs) を接続し、B モード [11] にて撮影を行った。超音波ゼリー [12] を塗布したプローブを実験協力者自身が持ち、下顎に下から垂直にあてた状態で矢状面の舌運動を記録した。撮影時のフレームレートは 68 fps であった。本研究で用いた超音波診断装置ではフレームレートを直接設定することはできず、撮像範囲などのいくつかのパラメータの設定の結果としてフレームレートが定まる。そこで、舌の広い範囲を捉え、舌表面の輝度が高くなり、なおかつフレームレートが 60 fps を越える設定を探索して決定した。

この超音波診断装置には音声を同期収録する機能がないため、ビデオキャプチャー (IO Data GV-HUVC) を用いて、超音波診断装置の映像出力とマイク収録した音声を同期させて PC に保存した。PC 側では、OBS Studio³にて映像と音声を収録し、60 fps の MP4 形式の動画として保存した⁴。キャプチャした画像のサイズは 1,280 pixel×720 pixel であった。OBS Studio では音声は圧縮されてしまうため、同じ PC にて Audacity⁵を用いて音声を録音し、分析用に標準化周波数 44.1 kHz、量子化 16 bit の WAV 形式にて保存した。

²MTracker の使用法については付録 B 参照のこと。

³<https://obsproject.com/>

⁴本来は超音波診断装置と OBS Studio のフレームレートは一致させた方が良く、今後その方向で検討する。

⁵<https://www.audacityteam.org/>

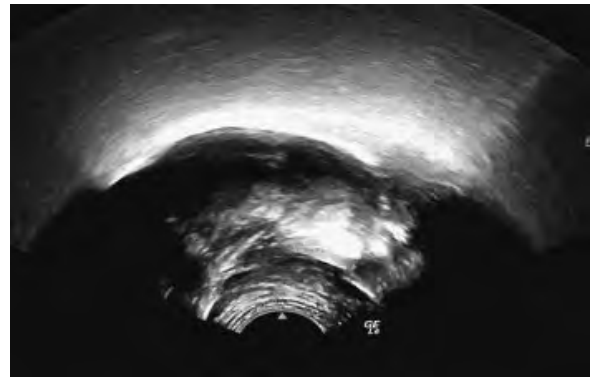


図 1: 超音波画像の例

本研究では特定の発話タスクは設けず、話者 1 名が日本語の音節や文を約 6 分間発話した映像を利用した。

2.2 画像データと舌輪郭データの作成

超音波診断装置のキャプチャ画像はメニューやサムネイルなどを含む。そこで、これらの不要な領域を除き、観測画像の領域 (700 pixel×450 pixel) を抽出した。図 1 に一例を示す。MRI 画像とは大きく異なり、舌の上面がぼんやりと白く造影され、口蓋や咽頭壁などは確認できない。本研究では、上記の動画の冒頭 2 分間を 1 秒に 2 枚 (30 フレームに 1 枚) の割合で抽出し、240 枚の画像ファイルとして保存した。

これらの画像について 4 枚に 1 枚の割合で 10 枚の画像の舌輪郭のトレースを行った。このようにフレームをスキップさせたのは、できるだけ変化の大きい画像を抽出するためである。トレースは、画像編集ソフト GIMP⁶ のペン機能を用いて行った。図 1 の画像をトレースした例を図 2 に示す。

トレースの線の表現方法としてはいくつか考えられるが、本研究では、プローブの位置 (扇状の撮像領域の要の位置) を原点に 6 度 ($=\pi/30$ [rad]) 間隔で 20 本の放射状グリッドをはり、そのグリッドとトレースの交点 (x_i, y_i) ($i = 1, \dots, 20$) を用いる方法を試みた。図 2 に対して放射状グリッドをはり、トレースとの交点を求めた例を図 3 に示す。この図では左右各 2 本のグリッドがトレースと交点を持たない。このような場合には、当該のグリッドに対応する交点として (0,0) を指定した。

2.3 学習と舌輪郭の自動抽出

上述のように、dlib に実装されているのは Kazemi and Sullivan [10] のアルゴリズムである。この手法はアンサ

⁶<https://www.gimp.org/>

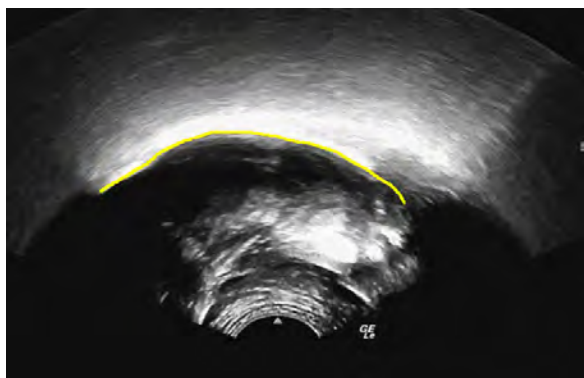
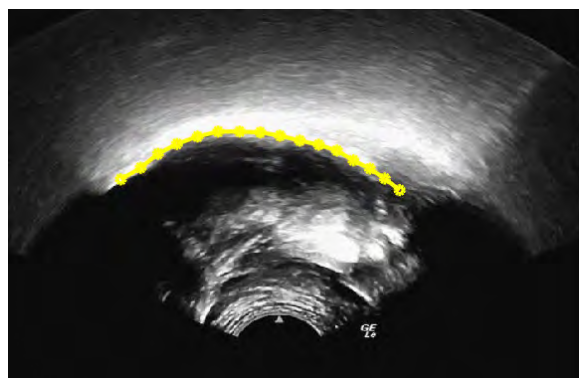


図 2: 図 1 をトレースした例 (トレースは黄色で示す)



(a)

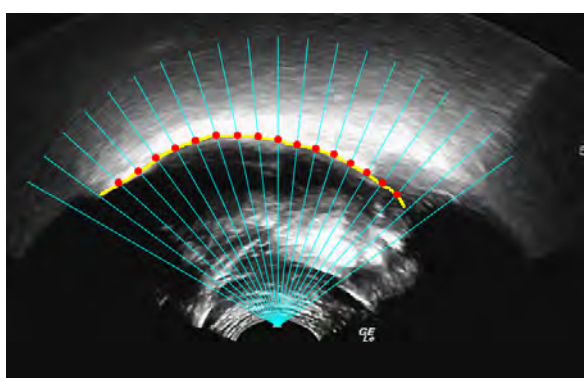
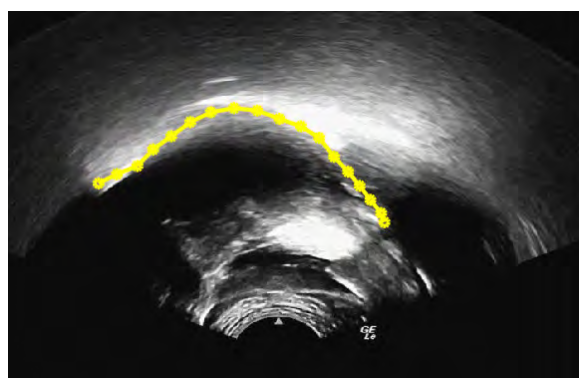


図 3: 図 2 の画像に放射状グリッド (シアン) をはり、トレースとの交点 (赤丸) を求めた例



(b)

図 4: 学習に用いていないフレームの自動抽出例

ンブル回帰木により画像中の特徴点を抽出する。このアルゴリズムが顔画像に限らず適用可能であることはすでに Takemoto ら [7], 白勢と北村 [13] が示している。本研究では、彼らと同様に dlib に同梱されている顔特徴点抽出器のサンプルスクリプトと学習用の XML ファイルを改変して舌輪郭を抽出した (付録 C 参照)。

学習データは、2.2 節にて述べた 10 フレーム分の画像と、その画像のトレースから得られた各 20 点の座標である。これらを用いて学習を行い、全ての画像 (240 枚) から舌輪郭を自動抽出した。自動抽出により得られる舌輪郭のデータは、学習データと同様に画像上の 20 点の座標である。学習および自動抽出は、Intel Core i5-8265U (1.60 GHz), メモリ 8 GiB 搭載のノート PC を用いて約 30 秒で完了した。

3 結果

2 つのフレームの結果を図 4 に示す。いずれも学習には使用していないフレームの結果である。図 4(a) は図 1,

図 2 と同一のフレームであり、図 4(b) は無作為に抽出したフレームである。学習にわずか 10 フレームしか用いていないにもかかわらず、図 4(a), (b) とともに舌輪郭を抽出できており、今後に期待が持てる。また、MRI のように高品質な画像でなくても舌輪郭が抽出できることが確認できた意義も大きい。アンサンブル回帰木に基づく本手法は、深層学習に基づく手法 [6, 14, 15, 16] と比較して計算コストが大幅に小さいことも利点である。

4 おわりに

本研究では、機械学習ライブラリを用いて超音波動画から舌輪郭を自動的に抽出することを試みた。まだ予備的な段階であり、検討対象のデータ数がわずかであるため結論を出すことはできないが、今回対象としたデータに関してはおおむね良好な結果が得られた。今後、より大規模なデータで精度評価を行うとともに、効率的な学習データの作成法や、話者や画質の違いに対するロバストネスについて検討する。

謝辞

本研究は JSPS 科研費基盤 (S)(20H05630), 基盤 (A)(20H00291, 20H01265, 21H04348), 基盤 (C)(17K02769) の支援を受けて行われた. EdgeTrak に関してご教示いただいた Chandra Kambhamettu 教授および Rohit Venkata Sai Dulam 氏 (University of Delaware) に感謝します.

参考文献

- [1] 定延, 記述言語学の観点から見た非流暢性, 第 35 回日本音声学会全国大会予稿集 (発表予定).
- [2] 林, 北村, 孫, 定延, 発話の途切れ・延伸時の調音運動に関する一検討, 音講論 (秋) (発表予定).
- [3] 北村, 動画で見る音声生成系の観測手法, 音響誌, 76(12), 700–705 (2020).
- [4] Li, Kambhamettu, and Stone, Automatic contour tracking in ultrasound images, *Clin. Linguist. Phon.*, 19(6-7), 545–554 (2005).
- [5] Kass, Witkin, and Terzopoulos, Snakes: Active contour model, *Int. J. Computer Vision*, 1(4), 321–331 (1988).
- [6] Zhu, Styler, and Calloway, A CNN-based tool for automatic tongue contour tracking in ultrasound images, *arXiv preprint*, arXiv:1907.10210, (2019).
- [7] Takemoto, Goto, Hagihara, Hamanaka, Kitamura, Nota, and Maekawa, Speech organ contour extraction using real-time MRI and machine learning method, *Proc. of Interspeech 2019*, 904–908 (2019).
- [8] King, Dlib-ml: A Machine Learning Toolkit, *Journal of Machine Learning Research*, 10, 1755–1758 (2009).
- [9] 後藤, 天野, 竹本, 北村, 能田, 前川, rtMRI 動画から抽出した発話器官の輪郭データに基づく平均声道の生成と分析, 音講論 (秋), 655–666 (2020).
- [10] Kazemi and SullivanOne, Millisecond Face Alignment with an Ensemble of Regression Trees, *IEEE Conference on Computer Vision and Pattern Recognition* (2014).
- [11] 長谷川, 医用超音波, 電子情報通信学会知識ベース知識の森, 1 群, 10 編, 5 章 (2014). https://www.ieice-hbkb.org/files/01/01gun_10hen_05.pdf (2021 年 8 月 24 日参照).
- [12] 山田, 日本音響学会 音の何でもコーナー Q&A (118), <https://acoustics.jp/qanda/answer/118.html> (2021 年 8 月 24 日参照).
- [13] 白勢, 北村, 児童の日本語母音発声時の口唇運動に関する動画記録による検討, 音講論 (秋), 675–676 (2020).
- [14] Fasel and Berry, Deep belief networks for real-time extraction of tongue contours from ultrasound during speech, *Proc. 2010 20th International Conference on Pattern Recognition*, 1493–1496 (2010).
- [15] Xu, Roussel, Csapó, and Denby, Convolutional neural network-based automatic classification of mid-sagittal tongue gestural targets using B-mode ultrasound images, *JASA*, 141(6), EL541–EL537 (2017).
- [16] Mozaffari, Wen, Wang, and Lee, Real-time automatic tongue contour tracking in ultrasound video for guided pronunciation training, *Proc. VISI-GRAPP 2019*, 302–309 (2019).

付録 A. EdgeTrak のコンパイル

EdgeTrak [4] はその Web ページ⁷よりソースコードを入手し, Windows の Visual Studio Community 2019 の C++にてコンパイルして実行する. EdgeTrak のコンパイルのためには, Visual Studio Community 2019 のインストールの際, v142 のビルドツールと MFC も導入する必要がある.

付録 B. MTracker の実行手順

ここでは, Windows に Anaconda (<https://www.anaconda.com/>) が導入されていることを想定して説明するが, macOS でも基本的には同様の手順で実行できる.

まず, github の MTracker [6] のページ⁸よりコード一式をダウンロードして, ZIP ファイルを展開する (git を使ってもよいがここでは略). 展開して得られたフォルダ (ここでは `mtracker` とする) の中に `model` というフォルダを作り, 上記ページの Downloading the trained model というパートから `dense_aug.hdf5` を `model` フォルダにダウンロードする (Google drive 経由のダウンロード).

次に, `mtracker\src\cnn_model.py` というファイルをエディタ (メモ帳など) で開き, 14 行目の先頭に以下のように `#` を付けてコメントアウトする (パッケージのバージョンの違いに対応するため).

```
#from keras.utils import plot_model
```

次に, 実行に必要な Python のライブラリをインストールする. スタートメニューから Anaconda Powershell

⁷<https://www.eecis.udel.edu/wiki/vims/index.php/Main/EdgeTrak>

⁸<https://github.com/lingjzhu/mtracker.github.io>

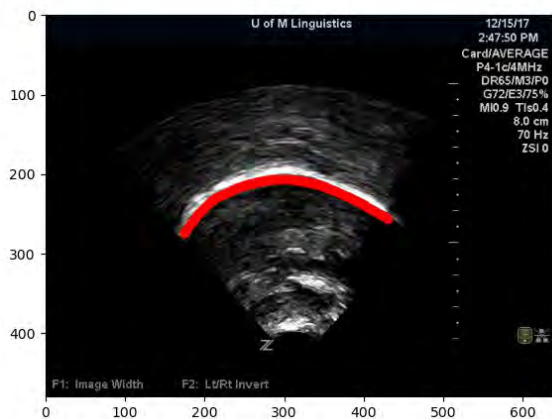


図 5: MTracker の処理例 (パッケージに同梱された画像を利用)

Prompt を起動し、プロンプトに以下のコマンドを 1 行ずつ入力する (実行する)。これによって、MTracker の実行に必要なパッケージがインストールされる。

```
pip install keras
pip install scikit-image
pip install imageio
pip install imageio-ffmpeg
pip install praatio
pip install tqdm
pip install tensorflow
```

以上で準備ができたので、MTracker を実行する。ここでは `mtracker\demo\demo_video.mp4` を処理してみる。Anaconda Powershell Prompt にて `cd` コマンドを用いて `mtracker` フォルダに移動してから、以下のコマンドを実行する (紙面の制約から複数行に書いているが、1 行で入力する)。

```
python track_video.py
-v ./demo/demo_video.mp4 -t du
-m ./model/dense_aug.hdf5
-o ./demo/out.csv -f ./demo -n 5
```

処理にはしばらく時間を要し、処理が終わると `demo0.jpg` などのトレース結果が作られる。この例では、`-n` の後に 100 を指定することによって、100 フレームに 1 枚の割合でトレース結果を保存している。また、`-o` にて指定した `demo\out.csv` というファイルには、全フレームの舌輪郭座標が記録される。処理例を図 5 に示す。

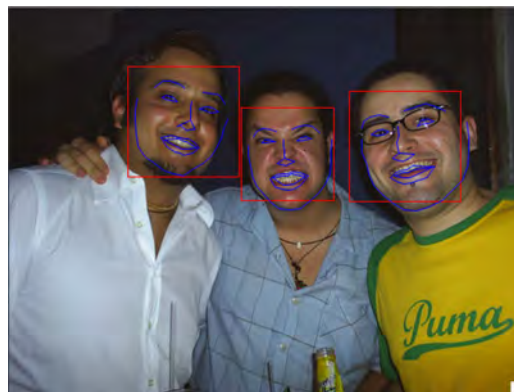


図 6: dlib の顔特徴点抽出器の処理例 (パッケージに同梱された画像を利用)

付録 C. dlib による学習と認識

dlib には顔特徴点抽出器 [10] の学習と認識を行う Python のサンプルスクリプトが提供されている。本研究ではこのスクリプトを改変して超音波画像の舌輪郭抽出に用いている。ここでは、付録 B と同様に Windows に Anaconda がインストールされていることを想定して、このサンプルスクリプトの実行法を解説する。

始めに、dlib の web ページ⁹の “Download dlib ver. 19.22” (2021 年 8 月 24 日時点) という青いアイコンをクリックし、一式 (`dlib-19.22.tar.bz2`) をダウンロードする。この形式のファイルは Windows 標準のファイル展開機能では展開できないため、7-zip (<https://sevenzip.osdn.jp/>) をインストールして使うとよい。展開すると、`dlib-19.22` というフォルダができる。

顔特徴点抽出器の学習に必要なデータは、`dlib-19.22\examples\faces` フォルダに保存されている。JPEG 形式の画像ファイルと、それぞれの画像の顔領域および各特徴点の座標が XML 形式にて `training_with_face_landmarks.xml` に保存されている。この XML ファイルの一部を以下に示す (`box` タグは 2 行に折り返して記載されている)。

```
<images>
<image file='2007_007763.jpg'>
  <box top='90' left='194' width='37'
  height='37'>
    <part name='00' x='201' y='107' />
    <part name='01' x='201' y='110' />
    <part name='02' x='201' y='113' />
```

⁹<http://dlib.net/>

image タグが学習用の画像ファイルを指定している。超音波画像の学習を行う場合には、このファイル名を変更する。box タグは顔領域 (四角形) を指定しており、超音波画像の場合には画像全体を指定する。part タグは特徴点を指定しており、ここにトレースで得られた舌輪郭の座標を指定すれば、舌輪郭の認識ができるようになる。なお、筆者らの経験ではこの XML ファイルを UTF-16 形式にて保存すると、後のスクリプトの実行時にエラーになり、UTF-8 (BOM なし) 等で保存する必要がある。

サンプルを用いて顔特徴点抽出器の学習を行うのであれば、dlib-19.22\examples\faces フォルダ以下のファイルに手を加える必要はない。Anaconda Powershell Prompt を起動し、まずは以下のコマンドによって dlib をインストールする。

```
pip install dlib
```

次に、dlib-19.22\python_examples に移動し、以下のコマンドにて顔特徴点抽出器の学習と認識を実行する (1 行で入力する)。

```
python ./train_shape_predictor.py  
../examples/faces
```

学習は 1 分程度で完了し、認識結果が順次表示される。処理例を図 6 に示す。dlib-19.22\examples\faces フォルダの学習データが少ないため精度は高くないが、学習済みのモデルをダウンロードすることによって高精度な認識が可能である。また、このスクリプトを超音波画像の舌輪郭抽出に用いる際には、顔領域検出のコードを取り除くなど若干の改変が必要である。