

超音波・MRI調音データの輪 郭自動抽出の試み —DeepLabCutを用いて—

©孫静（神戸大），北村達也（甲南大），林良子（神戸大）

研究背景

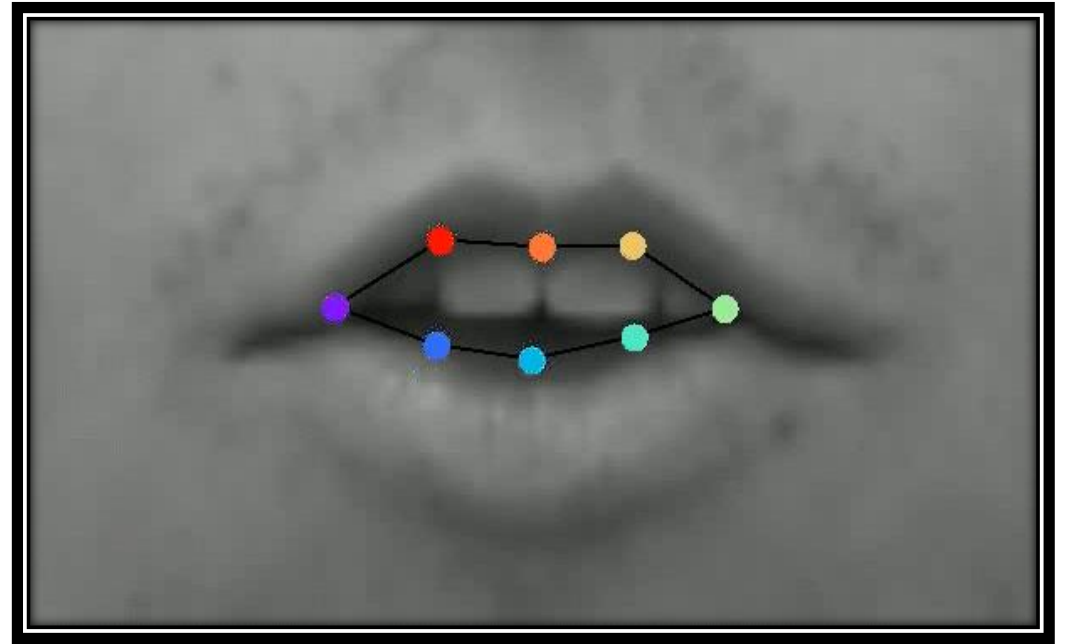
- 超音波やMRIを用いた調音データは，舌の詳細な動きを捉えるために広く利用されている
- 従来の手法は，時間と労力がかかり，精度にも課題がある（Liら2005，Chenら2020，Tiede2021）。
- 機械学習ライブラリを用いたrtMRI動画の輪郭抽出法（Takemotoら2019）は高精度だが，使用が難しい
- Wrenchら（2022）では，1つの超音波装置と少数の話者を対象に，鮮明な舌輪郭データを用いてモデルを作成している
 - このため，不鮮明な輪郭の抽出が困難である

目的

- 本研究では、動物の行動解析のために開発された深層学習ベースのツールであるDeepLabCut ver. 2.3.10（Mathis ら）を用いて、動画から輪郭抽出を試みた結果を報告する
- Wrenchらの手法を参考に、さまざまなデータを追加して新たなモデルをトレーニングし、より多様な超音波データからの輪郭抽出を目指している

DeepLabCutの歴史および利用例

- DeepLabCut (DLC) は、ディープラーニング技術を用いて動物や人の動きを追跡・解析するソフトウェアで、2018年にMathisらによって発表された
 - ネズミや馬の行動解析，唇の輪郭抽出



DLCのアルゴリズム

- 転移学習とディープニューラルネットワーク（特にResNet）を利用することで、最小限のトレーニングデータでも高い精度で動物の体の部位を追跡できる
 - 具体的には、DeeperCutと呼ばれるポーズ推定アルゴリズムの一部を活用
 - 事前にImageNetで学習された非常に深いニューラルネットワークを使用しており、転移学習を通じて迅速かつ効果的に特定の実験条件に適応させることが可能

実験方法

- 超音波データの収集

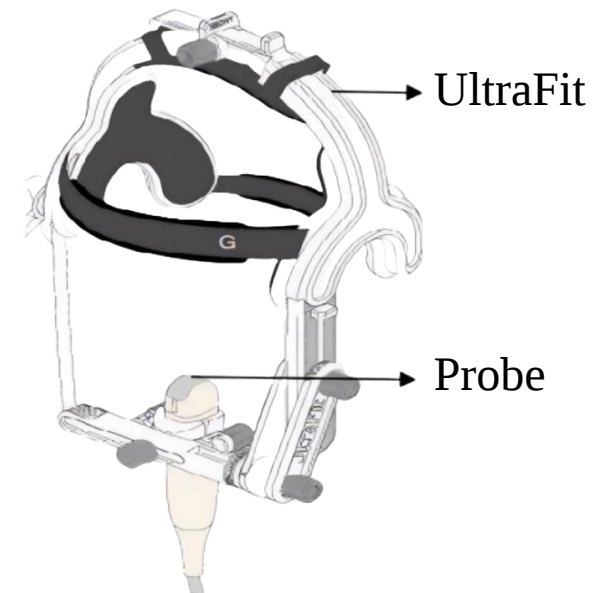
- 超音波断層撮像装置 (GE Healthcare LOGIQ e Premium)
- マイクロコンベックスプローブ (8C-Rs)
- Headset (UltraFit)
- Rubix 22 USB AUDIO INTERFACE
- Audacity (録音)
- IO Data GV-HUVC
- **O**pen **B**roadcaster **S**oftware (動画キャプチャ)

- rtMRIデータ

- 本研究では『リアルタイムMRI日本語調音運動データベース』を使用 (Maekawa 2024)



Ultrasound Device



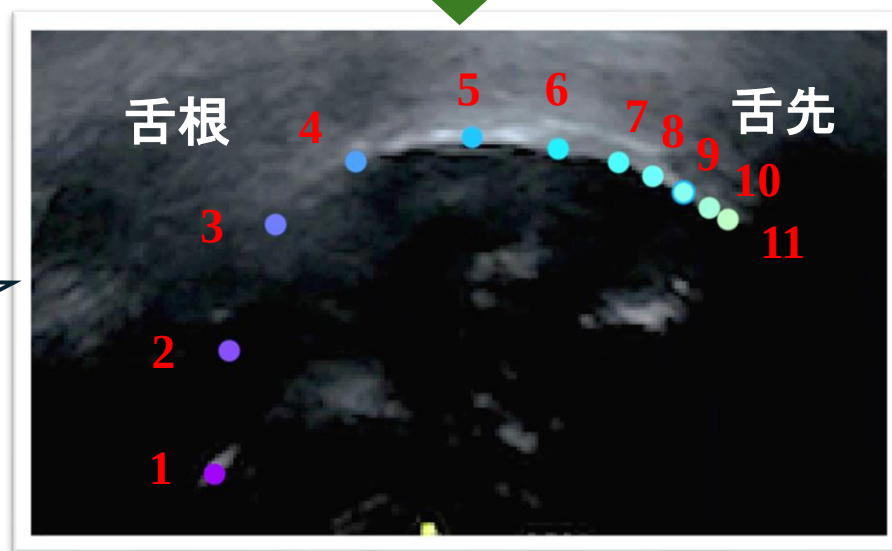
学習データの作成とモデルの学習

キーフレーム
を選択

手動でラベ
ル付け

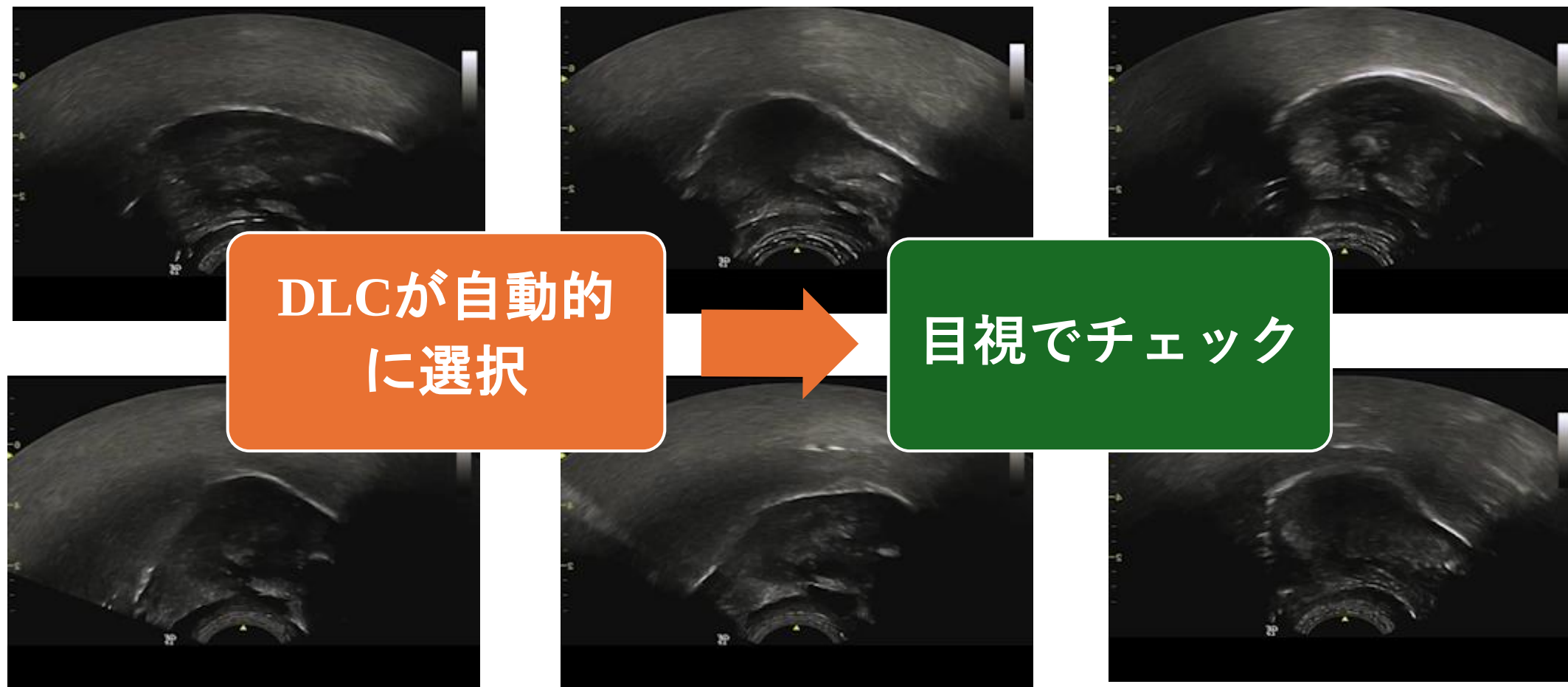
モデルの学
習

輪郭線上の11点
を指定



話者数:25
フレーム数 : 3000

キーフレームの選び方



舌の位置や極端な形状をとるフレームを選ぶ

キーフレーム抽出のアルゴリズム

- 反復的なキーフレーム選択と微調整
 - 初期のキーフレームをラベル付けした後，そのデータを用いてモデルを訓練する
 - その後にエラーの大きいフレームやモデルの予測が不確かなフレームを新たなキーフレームとして選び出す反復的なプロセス
 - これにより，モデルの性能が向上するにつれて，キーフレームを徐々に追加していくことができる

実験結果（超音波）

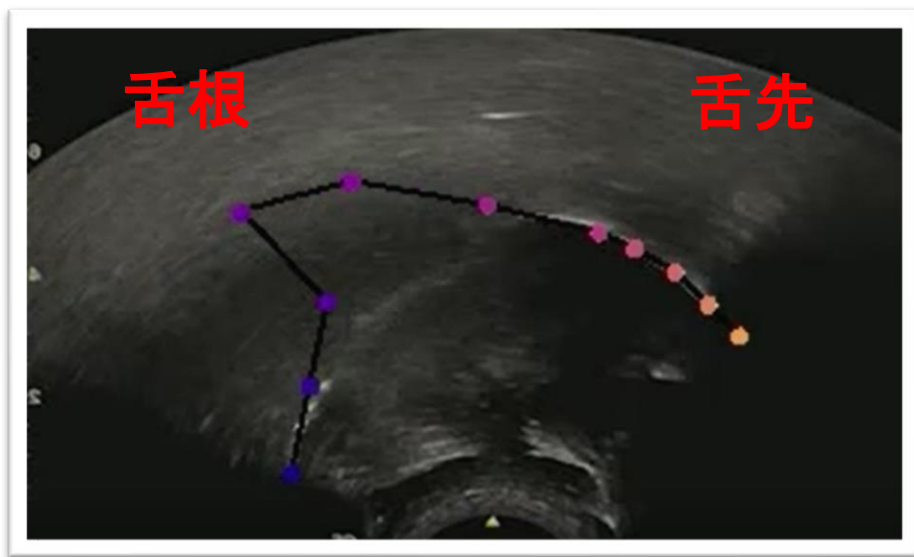


図1 Wrench (2022)のモデルで抽出したもの

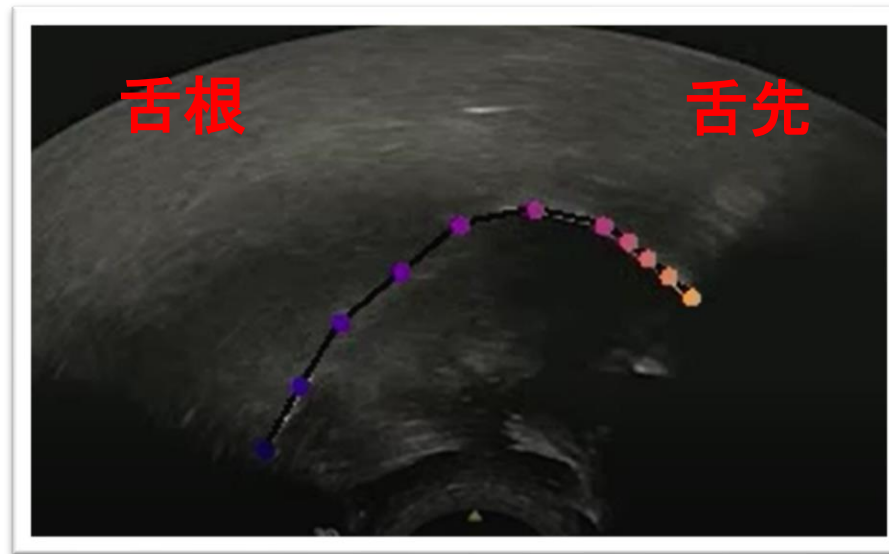


図2 発表者がトレーニングしたモデルで抽出したもの

- 図1→輪郭線上の点がいくつかずれている
- 図2→輪郭線上の点が修復された

実験結果 (rtMRI)



図3 発表者がトレーニングしたMRI用のモデルで抽出したもの

	超音波	rtMRI
Train error (px)	3.21	1.71
Test error (px)	3.34	3.24
Train error with p-cutoff	3.04	1.71
Test error with p-cutoff	3.22	2.60

表1 評価結果

- 各調音器官の輪郭抽出ができた
- 誤差は3pixels程度で高精度な抽出が実現された

学習および輪郭抽出の所要時間

- 学習時間はデータ量やネットワークの深さ（ResNet-50, ResNet-101など）、使用機器に依存する
 - 例えば、NVIDIA 1080-Ti GPUを使って数百～数千フレームのデータを学習させる場合、数時間から数十時間程度
 - CPUのみの環境では、計算時間が数倍から数十倍に増加
- 輪郭抽出は高性能なGPUでは1秒間に約200フレーム（ 320×240 ）の速度で処理可能
 - 解像度の高い動画や複雑なモデルを使用する場合、1フレームあたりの処理速度は遅くなる

まとめと考察

- Wrenchらの手法を基に，DLCを用い，**大量で多様なデータセット**を使用してモデルを訓練することにより，**不明瞭な輪郭**からも**高い精度**で抽出することが可能となった
- 同様の手法を用いて**ポイントを増やし**，MRIによる調音器官の輪郭抽出も実現された
- ただし，誤差が約3 pixelsであるため，Takemotoらの手法より若干低い

参考文献

Li *et al.*, *Clinical Linguistics and Phonetics*, 19(6-7), 545-554, 2005.

Maekawa, K. (2024). Real-time MRI articulatory movement database and its application to articulatory phonetics. *Acoustical Science and Technology*. Advance online publication.

Mathis *et al.*, *Nature Neuroscience*, 21(9), 1281-1289, 2018.

リアルタイムMRI日本語調音運動データベース 第2版, 2024.
<https://rtmridb.ninjal.ac.jp/>

Takemoto *et al.*, *Proc. INTERSPEECH 2019*, 904-908, 2019.

Tiede, GetContours Ver. 3.5, 2021. <https://github.com/mktiede/GetContours>

Chen *et al.*, *Proc. ISSP*, 2020.

Wrench, A., & Balch-Tomes, J. (2022). Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut. *Sensors*, 22(3), 1133.