4279

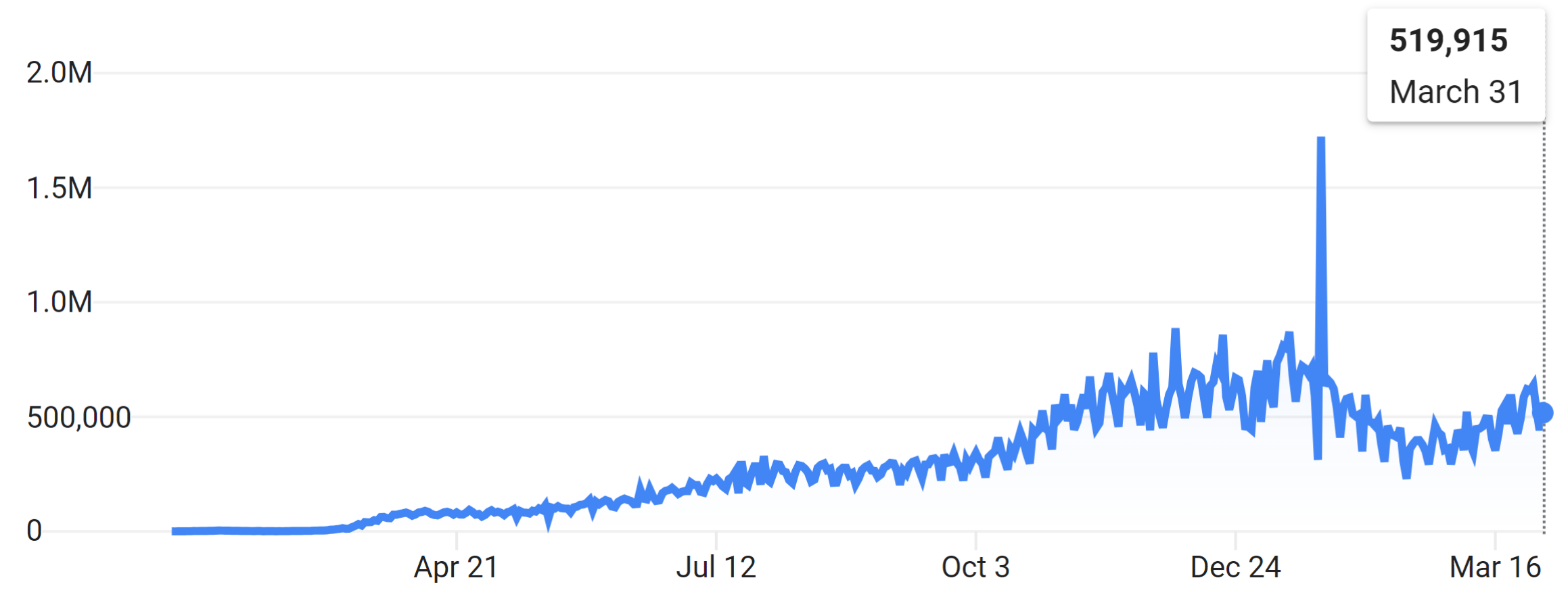
Interpreting glottal flow dynamics for detecting COVID-19 from voice

Soham Deskmukh[1], Mahmoud Al Ismail[1], Rita Singh[2]
[1]Microsoft, [2]Carnegie Mellon University

* This work was done at Carnegie Mellon University

COVID-19

- COVID-19 has resulted in more than 100 million infections, and more than *2 million casualties* spanning across 200 countries.
- For treatment, it's critical to identify people who need care in a timely fashion with subsequent isolation steps.
- This necessitates a scalable method accessible to millions of people and provide results within reasonable time range for effective treatment and isolation.



COVID-19 Daily Cases Count

Motivation

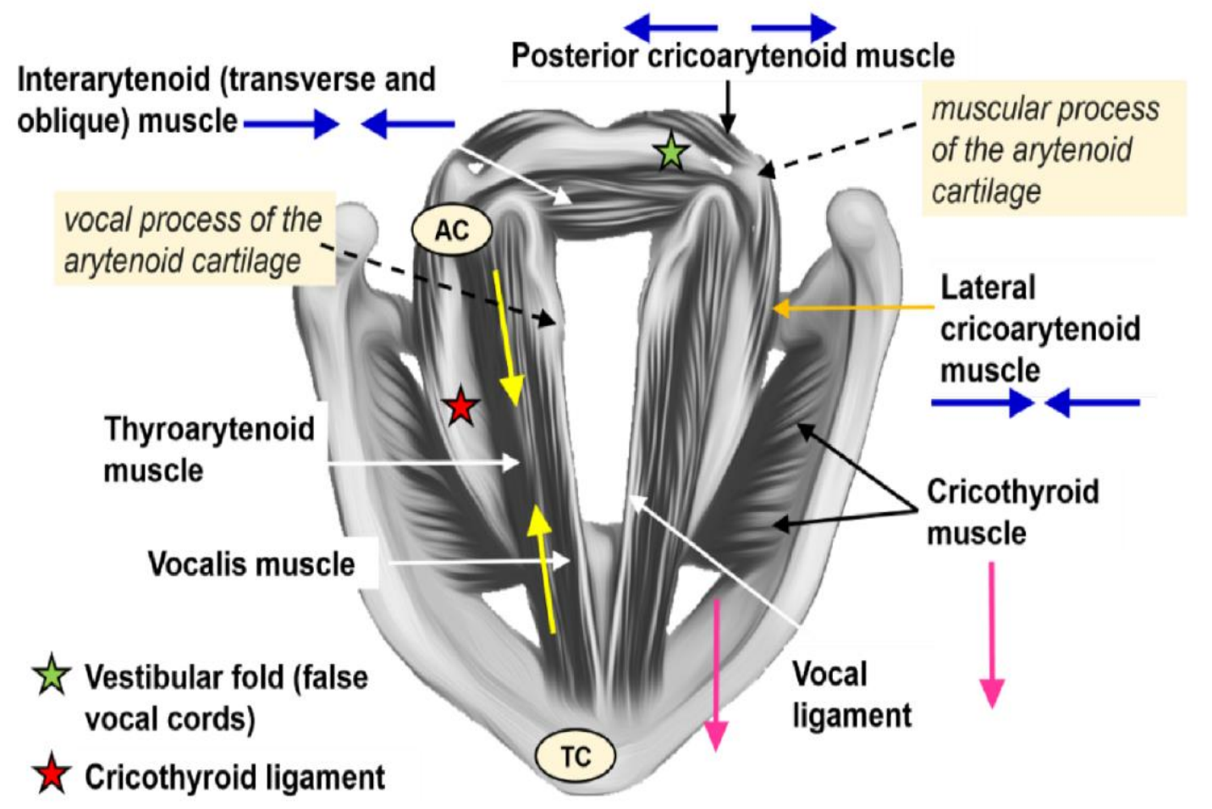
- Recent studies have shown vocal fold motion is adversely affected in symptomatic patients
- However, these studies can identify broad-level anomalies by visual comparisons between oscillation patters of healthy and *symptomatic* COVID-19 positive people.

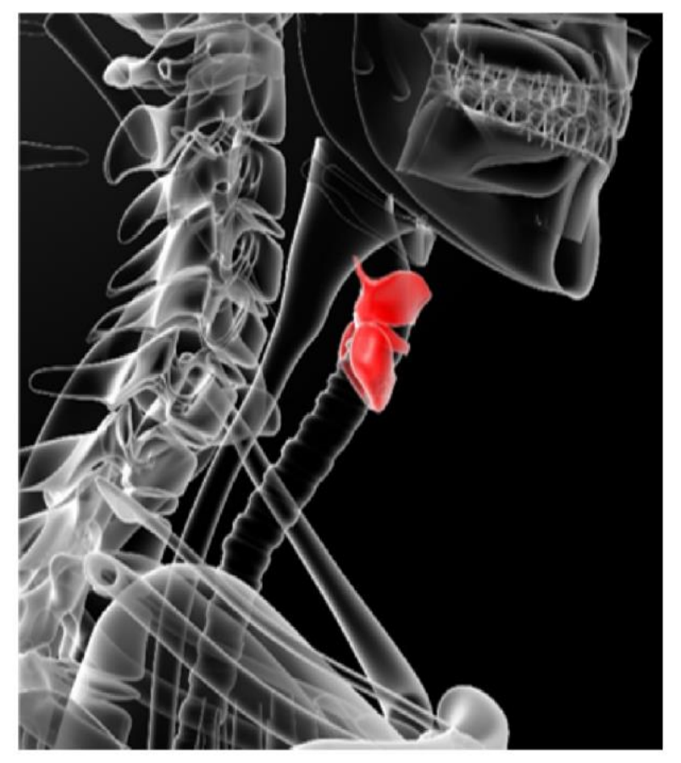
In this work:

- We focus on developing a ***non-invasive scalable way*** to detect and analyze vocal fold motion during voice production and utilize it for downstream tasks like COVID-19 detection

Phonation

- Voiced sounds are produced through the process of phonation, where the movement of vocal folds are self-sustained through the interaction of physical and aerodynamic forces across the glottis
- The actual movements are governed by various biomechanical properties of the vocal fold such as elasticity, resistance, Young's modulus, viscosity etc. We use asymmetric body cover model for emulating the motion of vocal fold


Posterior cricoarytenoid muscle
Interarytenoid (transverse and oblique) muscle
muscular process of the arytenoid cartilage
vocal process of the arytenoid cartilage
AC
Lateral cricoarytenoid muscle
Thyroarytenoid muscle
Cricothyroid muscle
Vocalis muscle
★ Vestibular fold (false vocal cords)
★ Cricothyroid ligament
TC
Vocal ligament



Laryngoscopic view of the vocal folds

Approach

- **Assumption:** The 1-mass asymmetric body mass physical model emulated by Van der poll oscillator can explain the vocal fold motions of an individual
- **Practically:** The 1d asymmetric body mass model can best *model healthy persons* with certain degree of asymmetry in vocal fold motion
- **Approach:** Capture the vocal fold oscillation impairment as discrepancy surfacing in the form of differences between glottal flow waveform obtained from inverse filtering and glottal flow waveform estimated from 1d asymmetrical body mass model

ADLES

$$\min \int_0^T (u_0(t) - u_0^m(t))^2 dt$$

$$\Leftrightarrow \min \int_0^T (\tilde{c}d(2x_0 + x_l(t) + x_r(t)) - \frac{A(0)}{\rho c}\mathcal{F}^{-1}(p_m(t)))^2$$

$$\begin{aligned} \text{s.t.} \quad & \ddot{x}_r + \beta(1 + x_r^2)\dot{x}_r + x_r - \frac{\Delta}{2}x_r = \alpha(\dot{x}_r + \dot{x}_l) \\ & \ddot{x}_l + \beta(1 + x_l^2)\dot{x}_l + x_l + \frac{\Delta}{2}x_l = \alpha(\dot{x}_r + \dot{x}_l) \\ & x_r(0) = C_r, x_l(0) = C_l, \dot{x}_r(0) = 0, \dot{x}_l(0) = 0 \end{aligned}$$

Notation

$u_0(t)$: Measured glottal flow
$u_0^m(t)$: Estimated glottal flow
$\tilde{c}$: Air particle velocity
A: Vocal tract area function
$\mathcal{F}^{-1}$: Inverse filter
α, β, Δ: Model parameters where

- α is the coupling coefficient between the supra- and sub-glottal pressure
- β incorporates the mass, spring and damping coefficients of the vocal folds
- Δ is an asymmetry coefficient.

Use ADLES to iteratively estimate the model parameters α, β, Δ

Multiple Instance Learning (MIL)

- In order to provide interpretability in classifier decisions and scalability we use Multiple Instance Learning (MIL)
- Specifically, we use a variant of two step attention as the pooling function in MIL

Two step attention pooling

$$\widehat{Z}_{a1} = \frac{e^{\sigma(\mathbf{Z}\mathbf{w}_{a1}^T + b_{a1})}}{\sum_{i=1}^F e^{\sigma(\mathbf{Z}\mathbf{w}_{a1}^T + b_{a1})}}, Z_{p1} = \sum_{i=0}^F (\mathbf{Z}\mathbf{w}_{c1}^T + b_{c1}) \cdot \widehat{Z}_{a1} \tag{6}$$

$$Z_x = f_3(Z_{p1}) \tag{7}$$

$$\widehat{Z}_{a2} = \frac{e^{\sigma(Z_x\mathbf{w}_{a2}^T + b_{a2})}}{\sum_{t=1}^T e^{\sigma(Z_x\mathbf{w}_{a2}^T + b_{a2})}}, Z_{p2} = \sum_{t=0}^T (Z_x\mathbf{w}_{c2}^T + b_{c2}) \cdot \widehat{Z}_{a2}$$

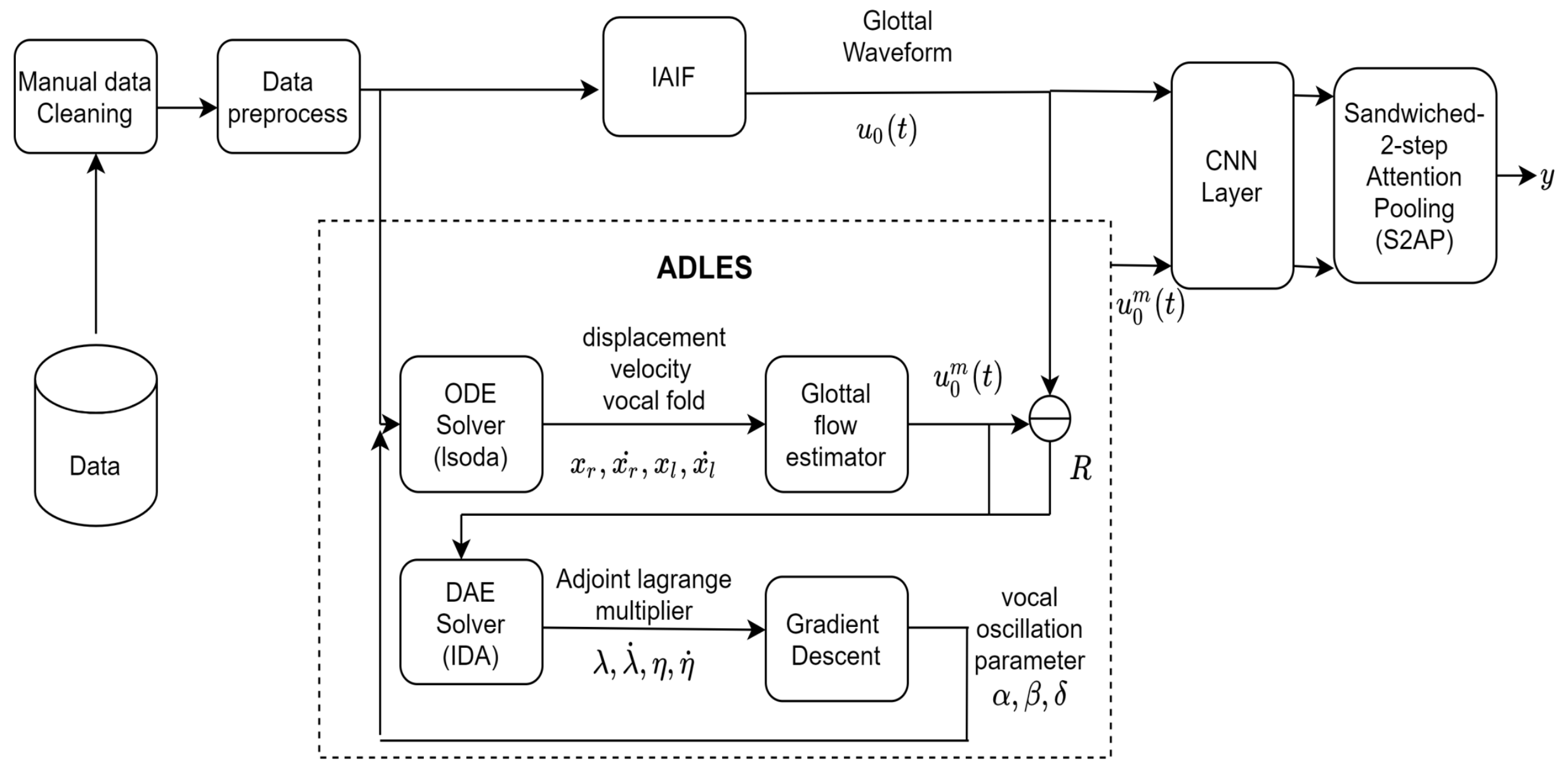
Experiment

We used extended recordings of vowels /a/, /u/, /i/ collected under expert clinical supervision

The dataset contains 19 individuals:

- 10 females - 5 diagnosed with COVID-19
- 9 males - 4 diagnosed with COVID-19

Each utterance is segmented using a window of 50ms with a shift of 25ms, resulting in a total of 3835 frames


Manual data Cleaning
Data preprocess
IAIF
Glottal Waveform
$u_0(t)$
CNN Layer
Sandwiched-2-step Attention Pooling (S2AP)
y
$u_0^m(t)$
Data
ADLES
ODE Solver (lsoda)
displacement velocity vocal fold
$x_r, \dot{x}_r, x_l, \dot{x}_l$
Glottal flow estimator
$u_0^m(t)$
R
DAE Solver (IDA)
Adjoint lagrange multiplier
$\lambda, \dot{\lambda}, \eta, \dot{\eta}$
Gradient Descent
vocal oscillation parameter
α, β, δ

Covid analysis setup

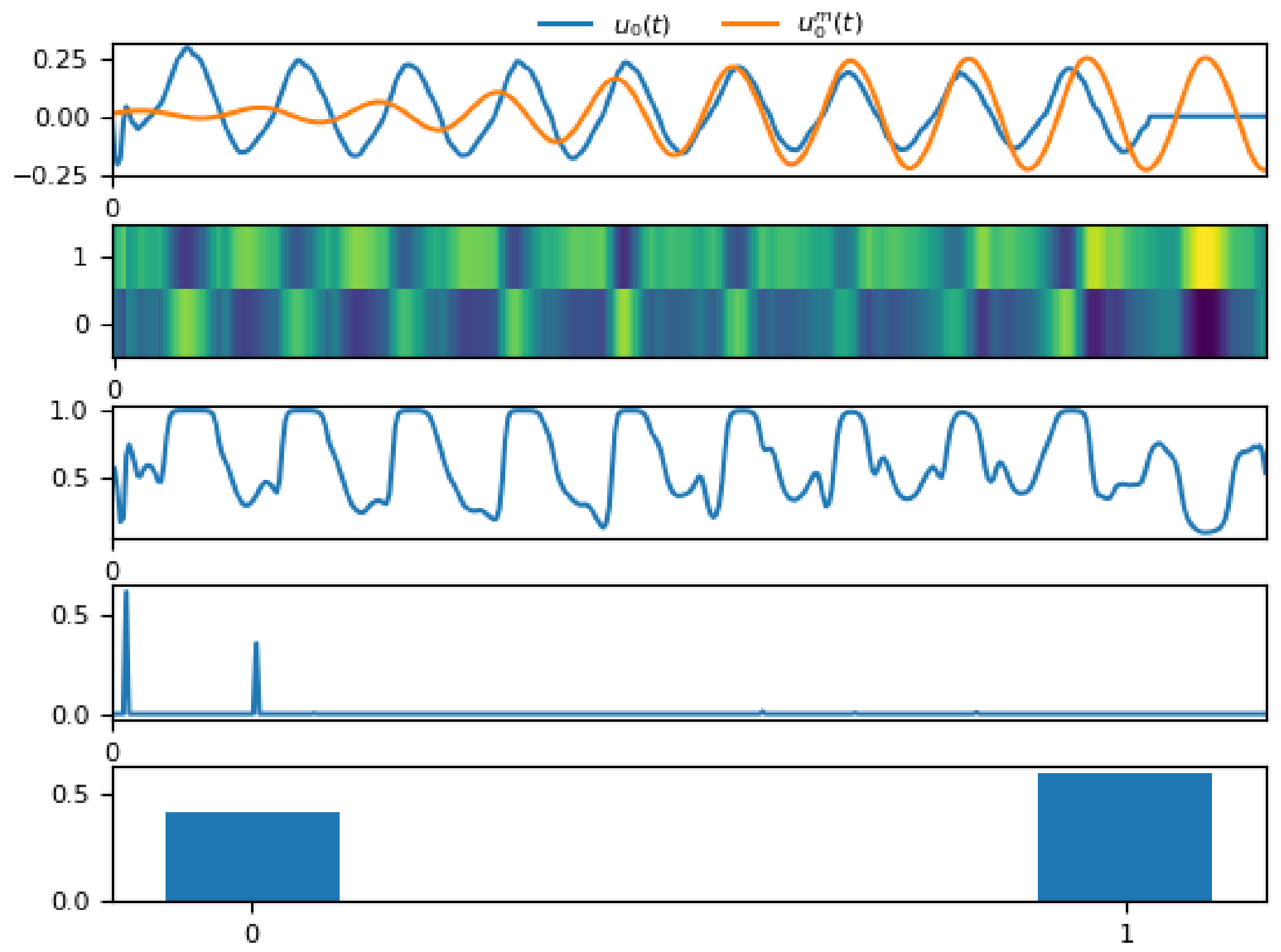
Results

Feature extractor	Pooling	ROC-AUC	STD
-	2AP	0.6611	0.0978
-	2SAP	0.7925	0.1073
CNN (1,3,32)	2AP	0.8009	0.1009
CNN (2,3,32)	2AP	0.8248	0.0790
CNN (2,3,32)	2SAP	0.8330	0.0745
CNN (2,5,64)	2SAP	**0.8520**	**0.0577**

	/a/	/i/	/u/	/a/+/i/	/a/+/u/	/i/+/u/
AUC	0.57	0.839	0.896	0.690	0.804	**0.900**
STD	0.119	0.102	0.067	0.064	0.074	**0.062**

- Without any feature extractor, ROC of 0.7925 is obtained, indicating the ADLES estimations themselves carry significant amounts of information with little task specific tuning.
- The best performing model is 2-layer CNN and variant of two-step attention pooling with highest performance of 0.8520
- A key point to note here is, the CNNs are shallow and simple and the main here is to show the *effectiveness of method* in detecting upper respiratory tract illness.

Visualization



Visualization of two-step attention pooling

Two step attention pooling enables:

- Interpretable results
- Expert human intervention
- Way to discard erroneous results