Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction

Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, Patrick Traynor

USENIX Security 2022
August 11, 2022

DeepFakes

DeepFakes

DeepFakes

DeepFakes

TREND MICRO | Business For Home

Products Solutions Why Trend Micro Research Support Partners Company

Security News > Cyber Attacks > Unusual CEO Fraud via Deepfake Audio Steals US$243,000 From UK Company

Unusual CEO Fraud via Deepfake Audio Steals US$243,000 From UK Company

September 05, 2019

DeepFakes

The Deepfake problem is not just about authentication, fundamentally, it's about

"Is this speaker even human?"

Hypothesis

- ML accurately capture macro-acoustic features of audio

Hypothesis

- ML accurately capture macro-acoustic features of audio
  - But does it capture all of the important features?

Hypothesis

- ML accurately capture macro-acoustic features of audio
  - But does it capture all of the important features?
- **Hypothesis:** We can apply techniques from fluid-dynamics *to extract biological structure from audio.*

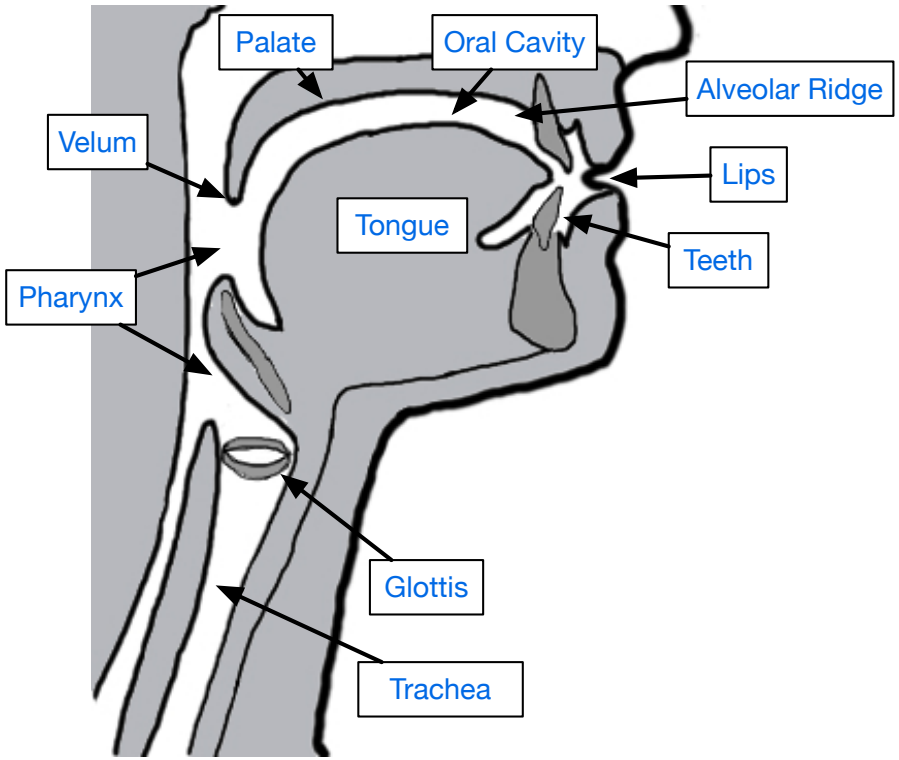
Hypothesis

- ML accurately capture macro-acoustic features of audio
  - But does it capture all of the important features?
- **Hypothesis:** We can apply techniques from fluid-dynamics *to extract biological structure from audio.*
  - Humans have limited configurations of their vocal tract.

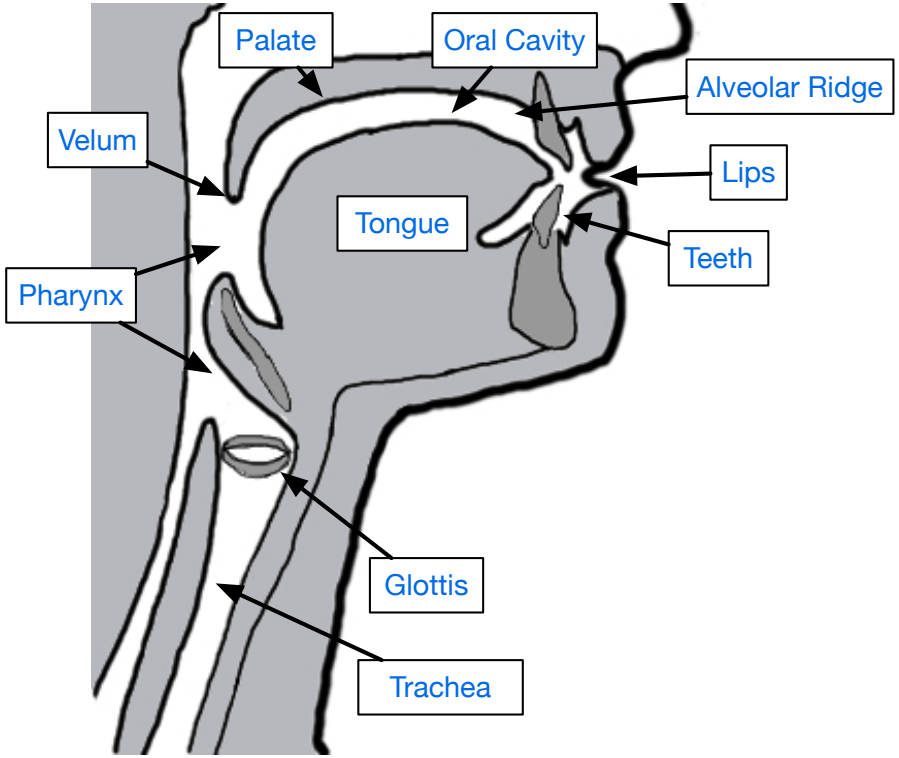
Hypothesis

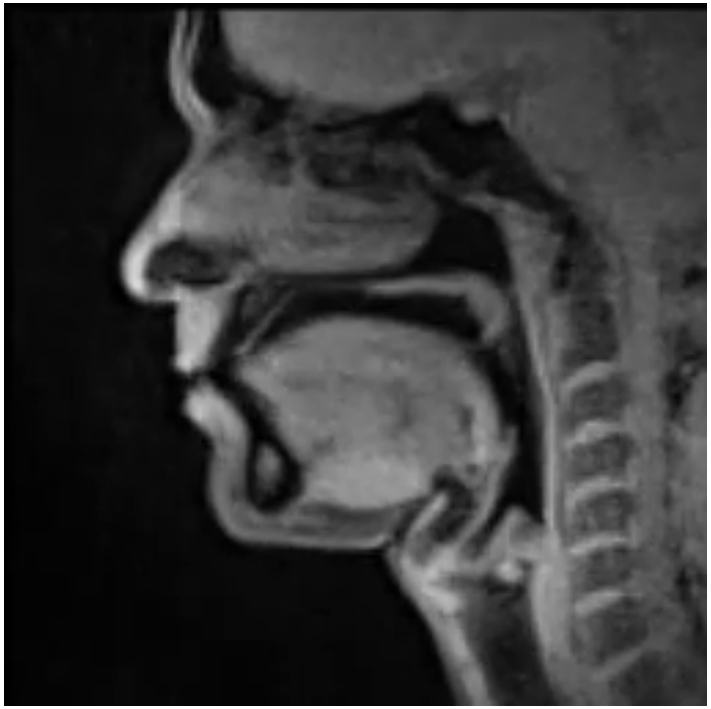
- ML accurately capture macro-acoustic features of audio
  - But does it capture all of the important features?
- **Hypothesis:** We can apply techniques from fluid-dynamics *to extract biological structure from audio.*
  - Humans have limited configurations of their vocal tract.
  - ML does not necessarily abide by these same constraints.

Organic Voice Creation

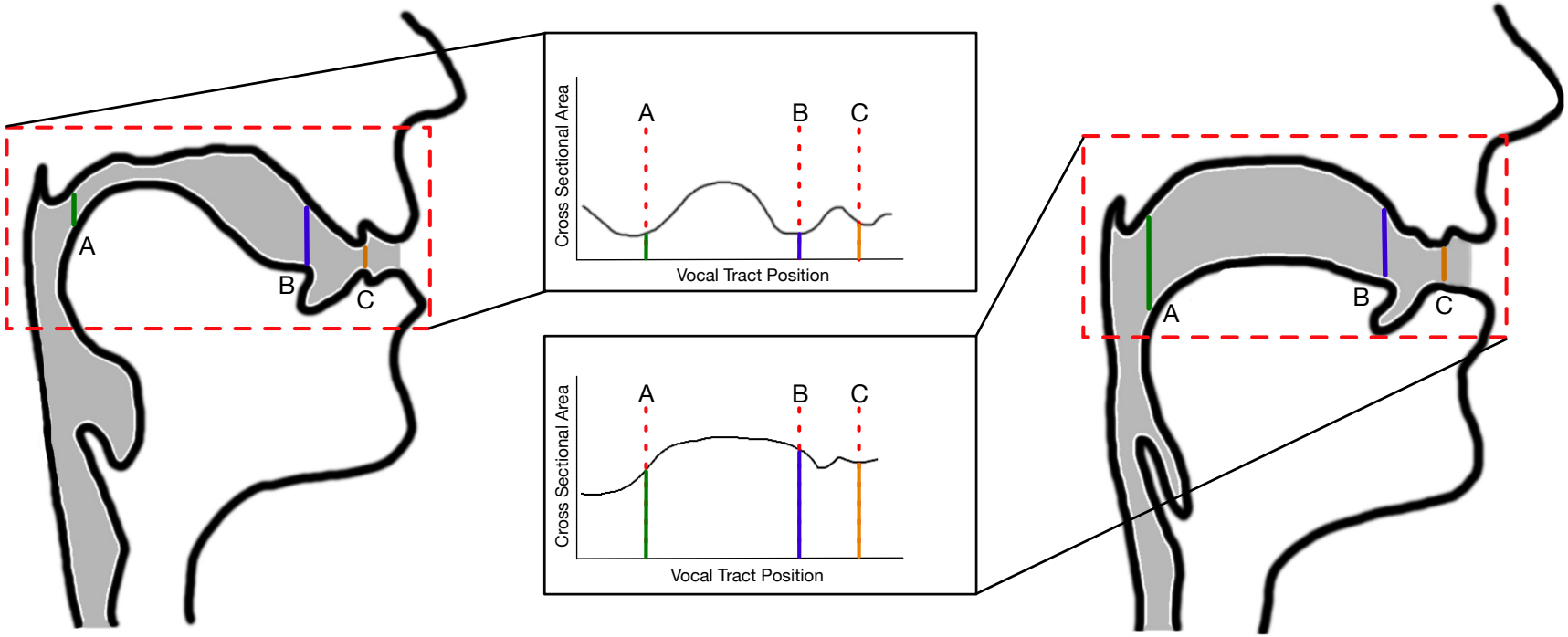


Organic Voice Creation





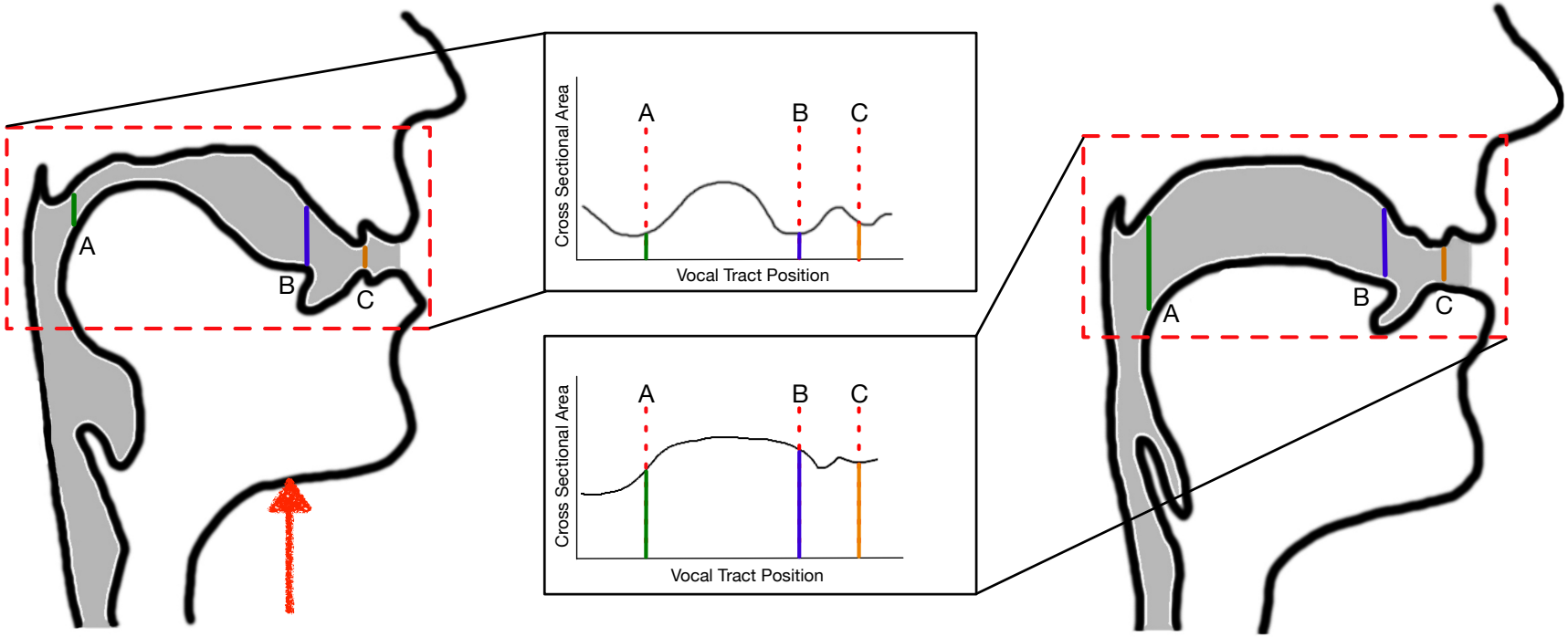
Organic Voice Creation



Who
/hu/

Has
/hæz/

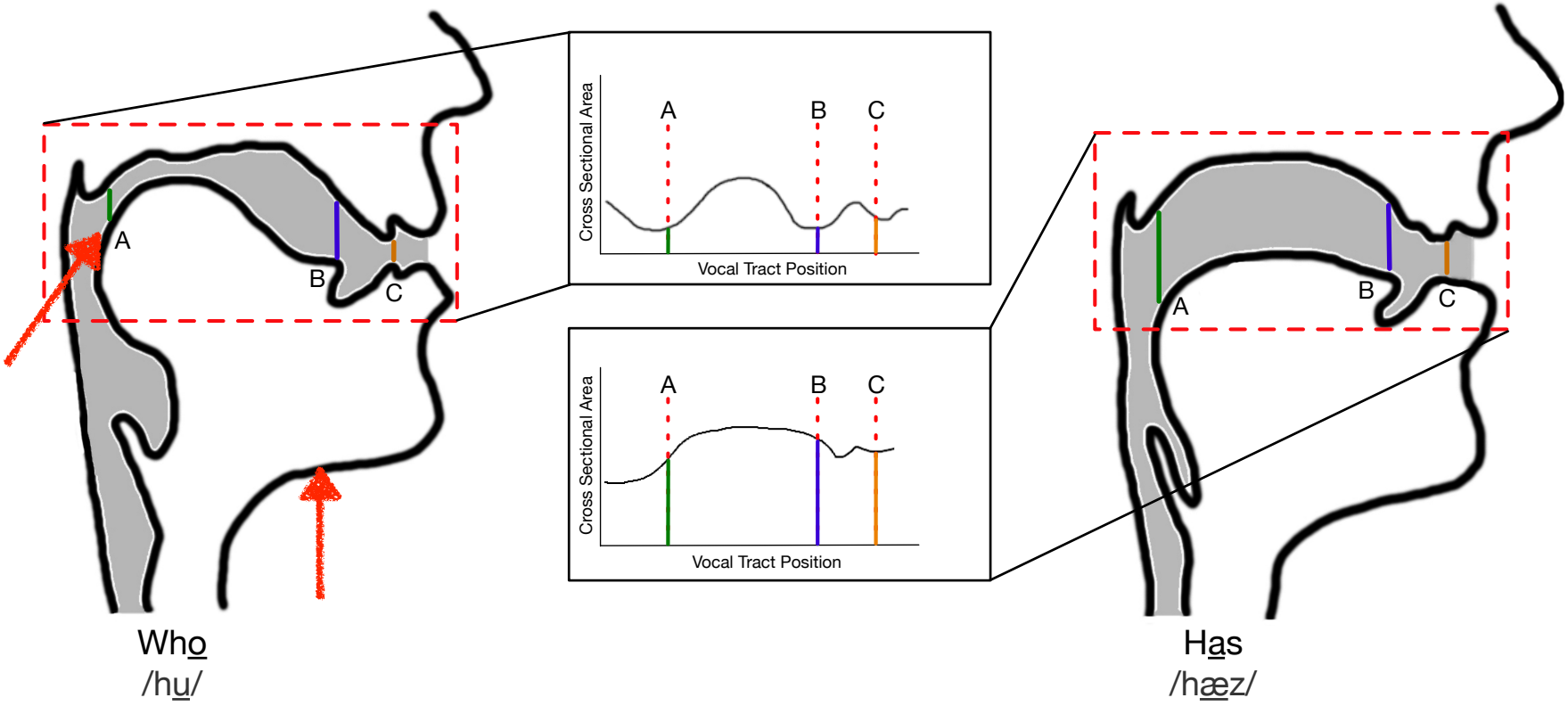
Organic Voice Creation



Who
/hu/

Has
/hæz/

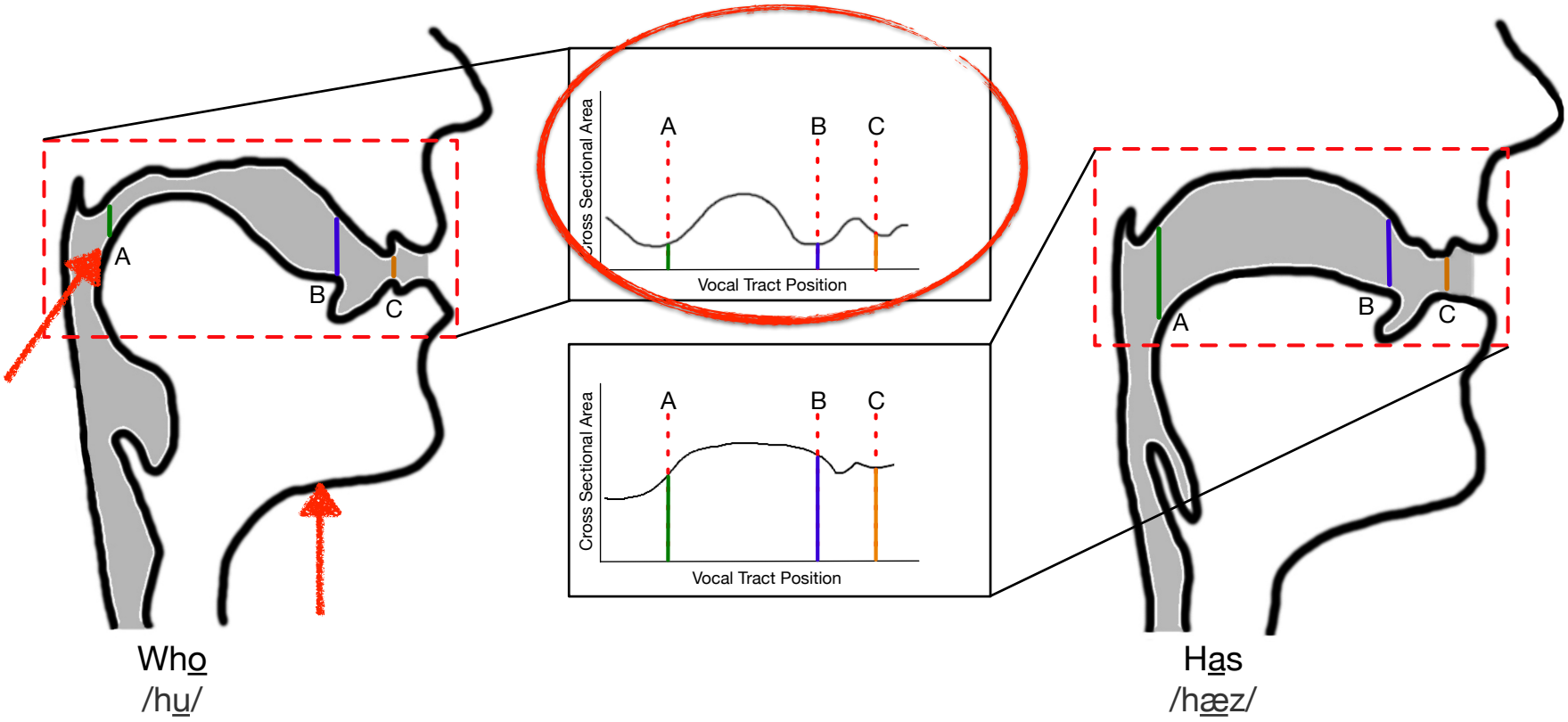
Organic Voice Creation



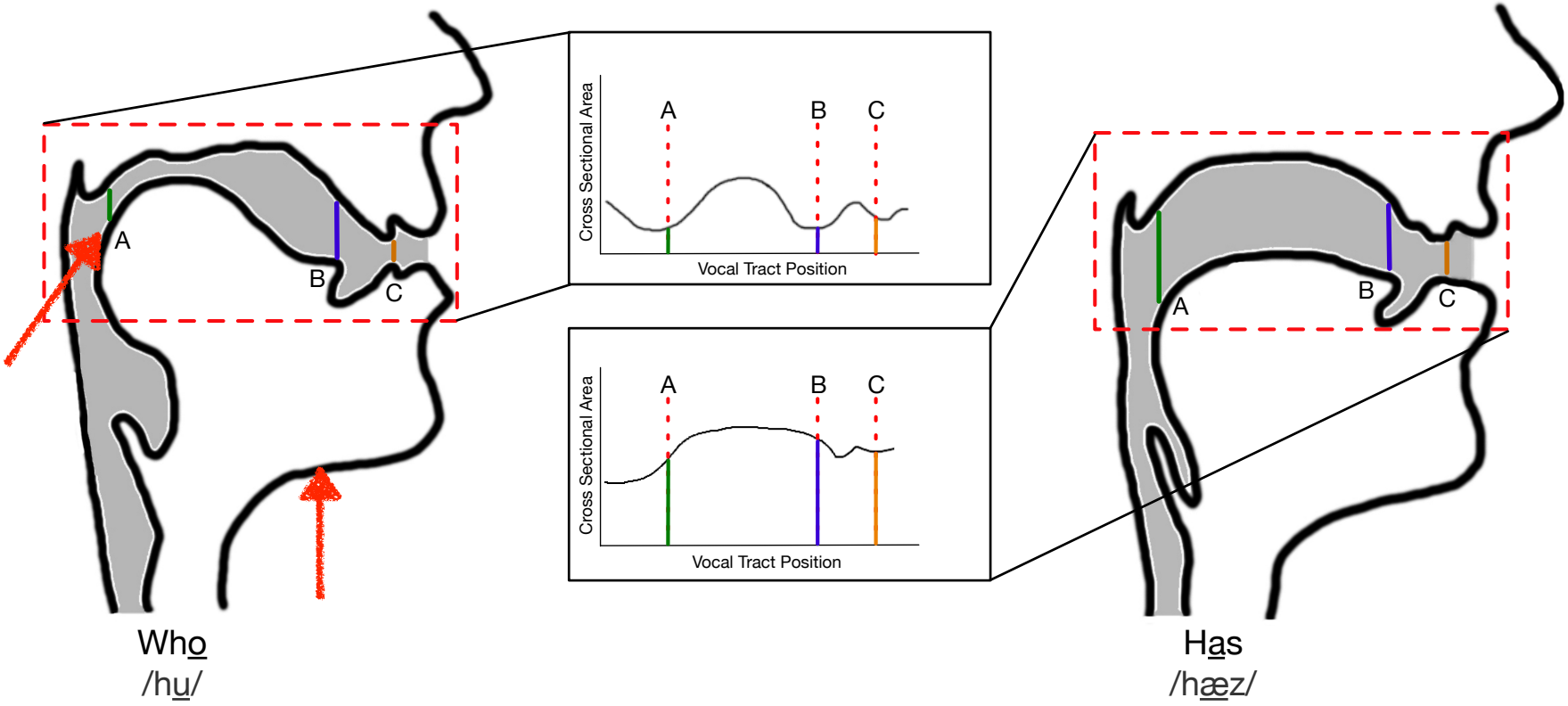
Who
/hu/

Has
/hæz/

Organic Voice Creation



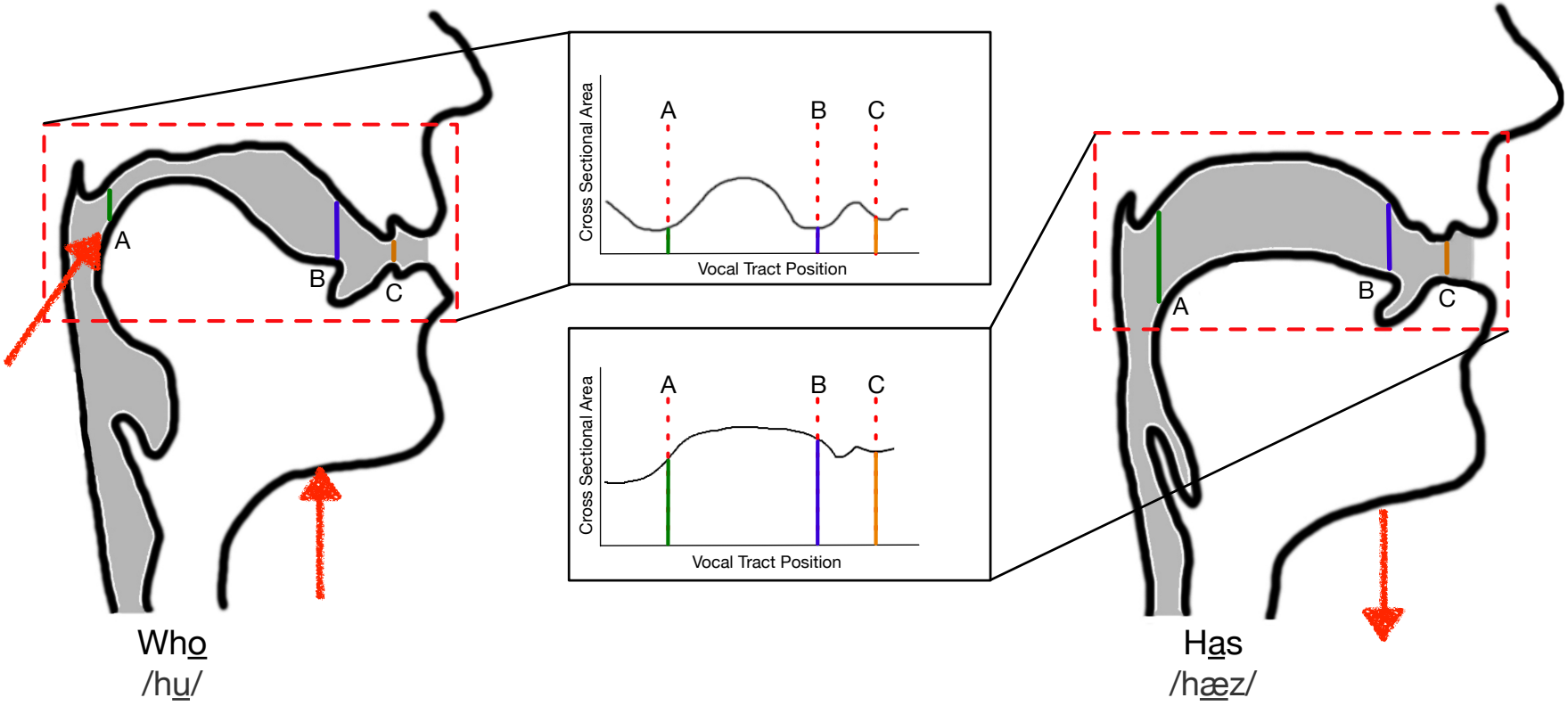
Organic Voice Creation



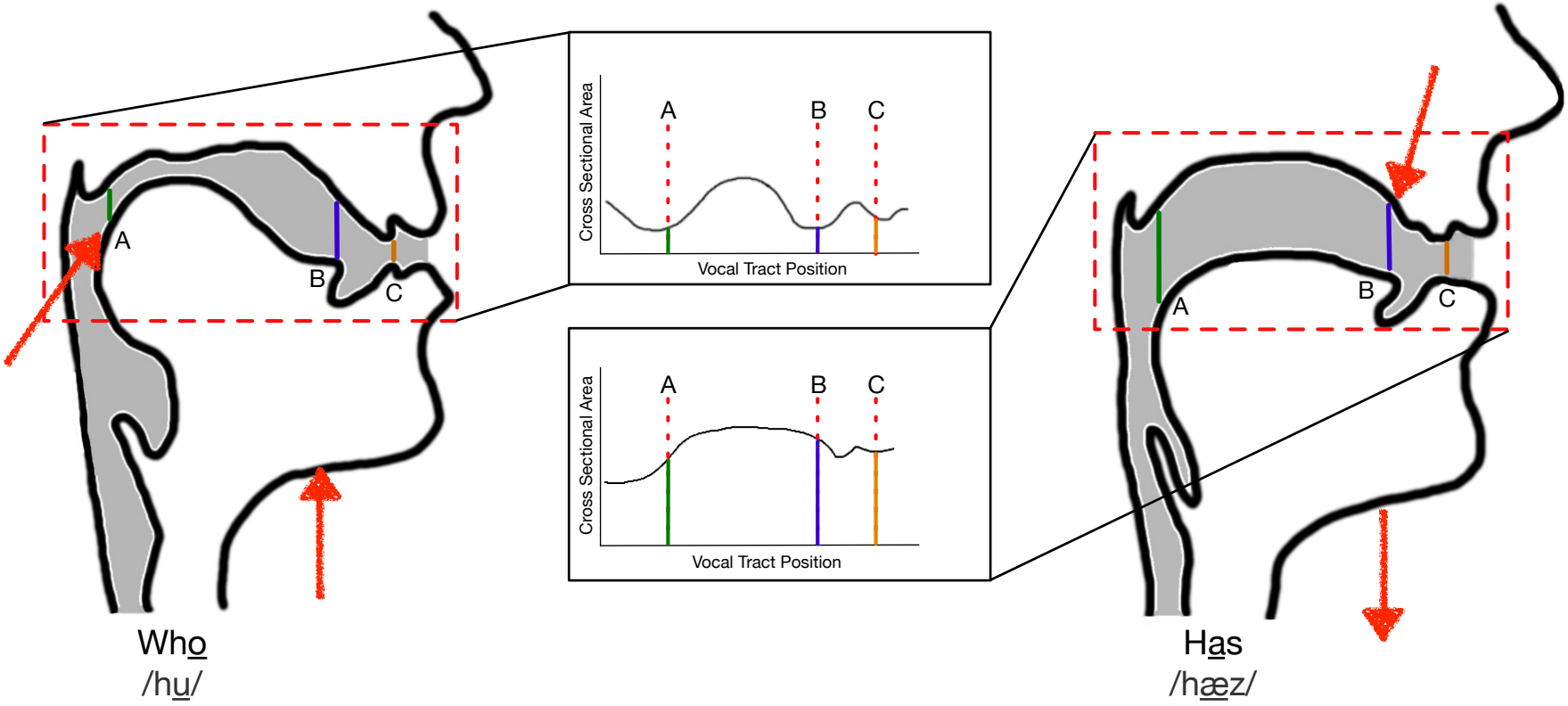
Who
/hu/

Has
/hæz/

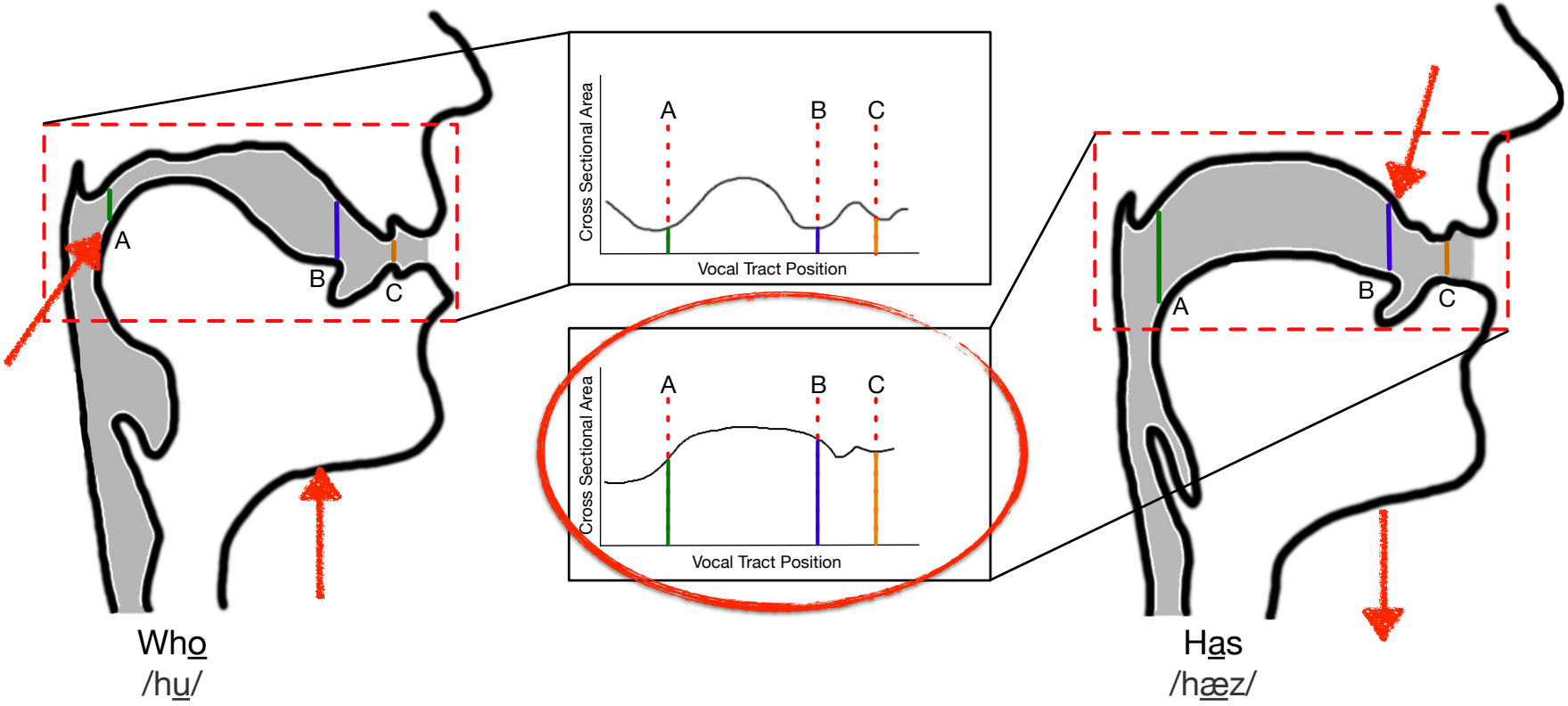
Organic Voice Creation



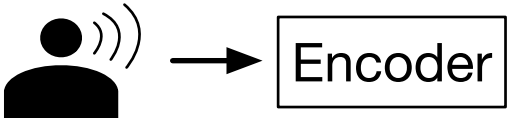
Organic Voice Creation



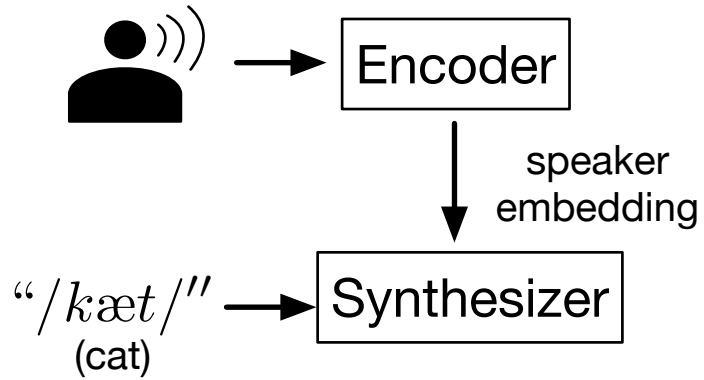
Organic Voice Creation



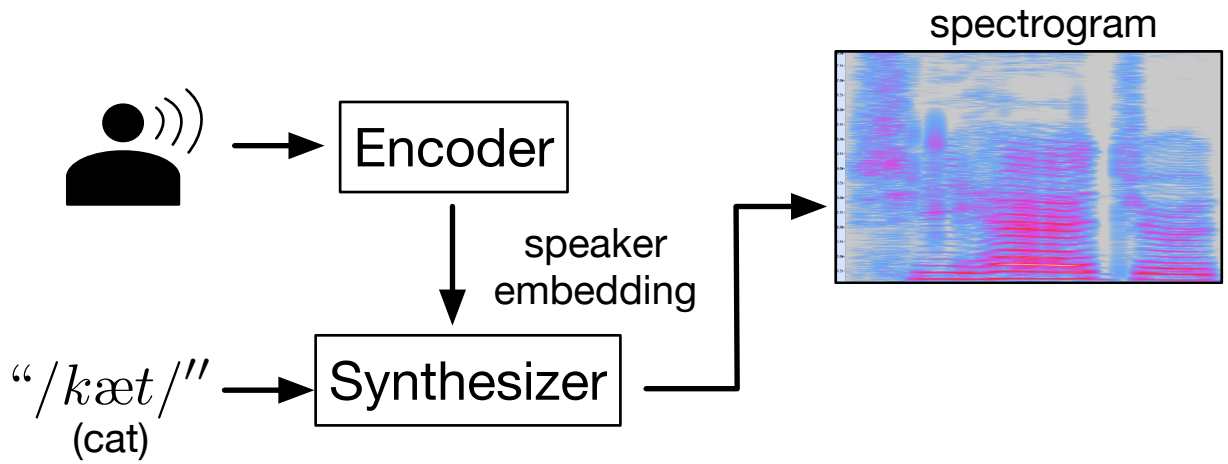
DeepFake Creation



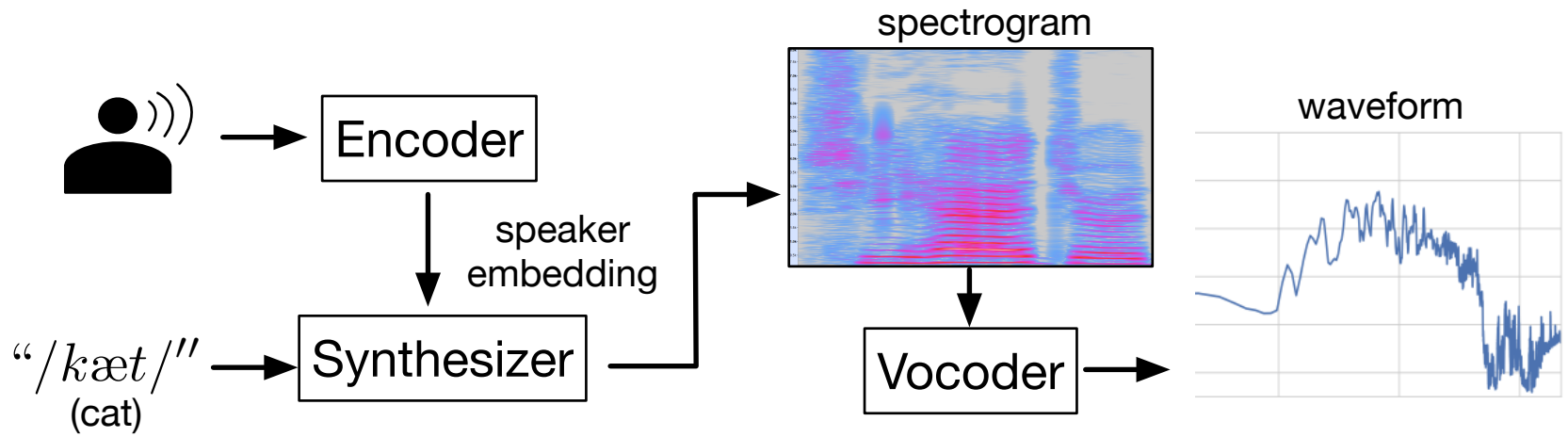
DeepFake Creation



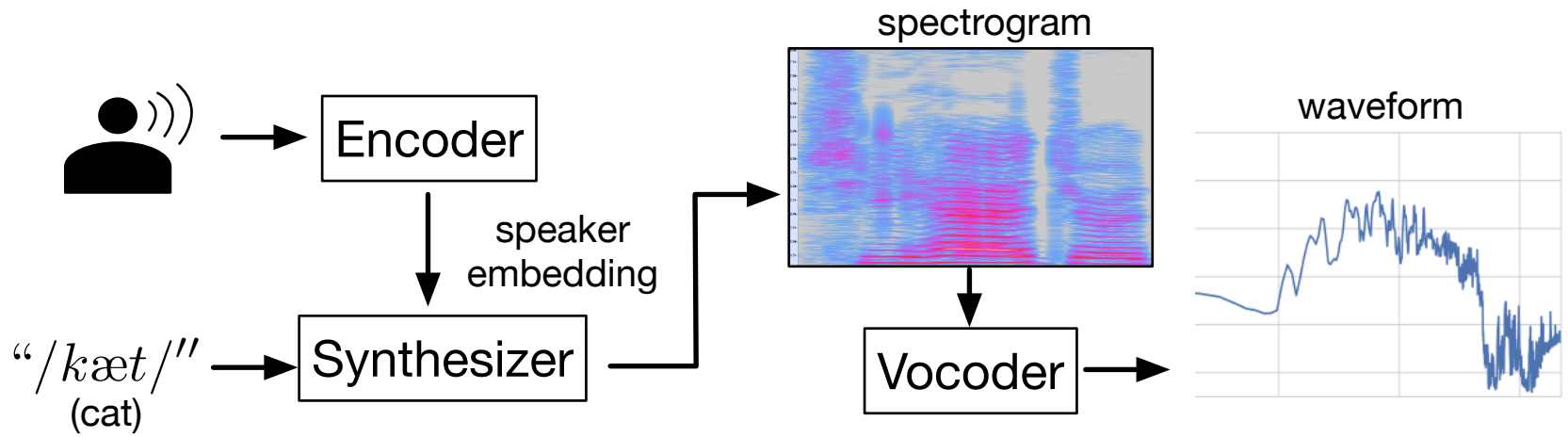
DeepFake Creation



DeepFake Creation

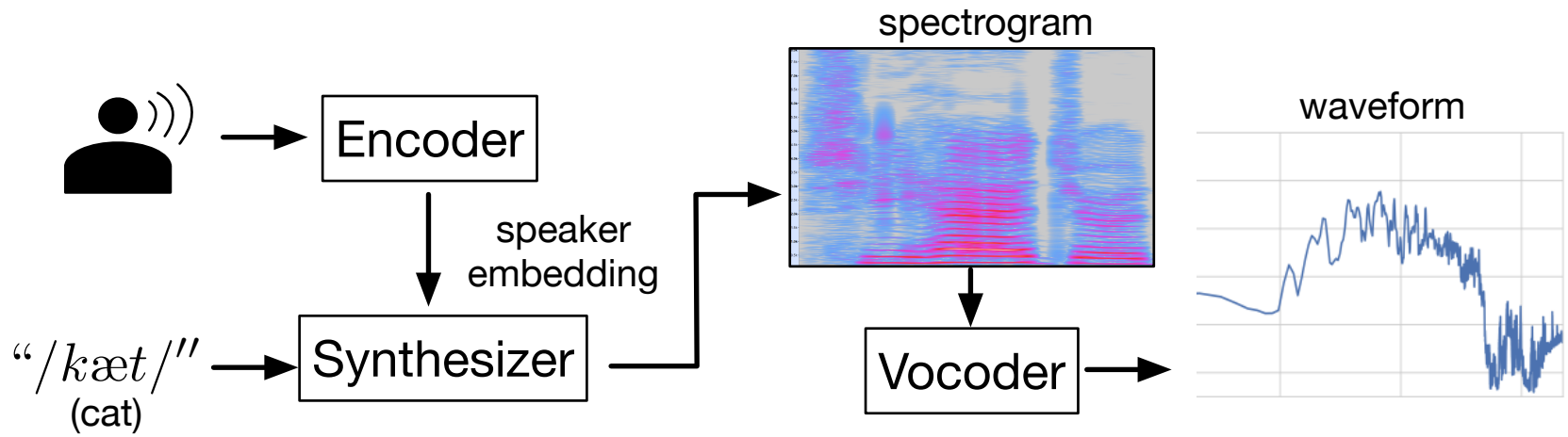


DeepFake Creation



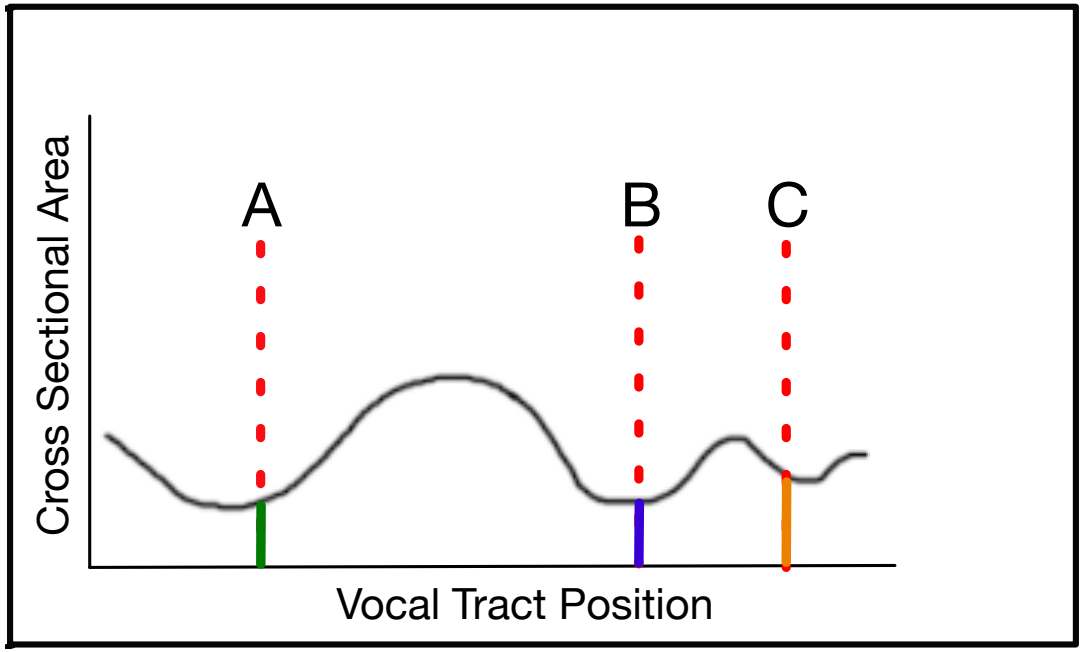
- Fundamentally not capturing the fluid dynamic behavior of the vocal tract

DeepFake Creation

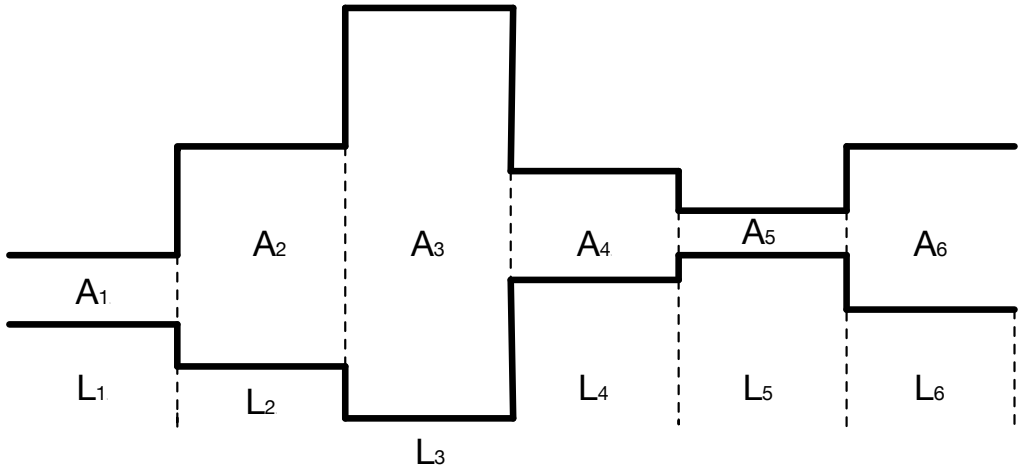


- Fundamentally not capturing the fluid dynamic behavior of the vocal tract
- Relying on ML to learn the fluid dynamics

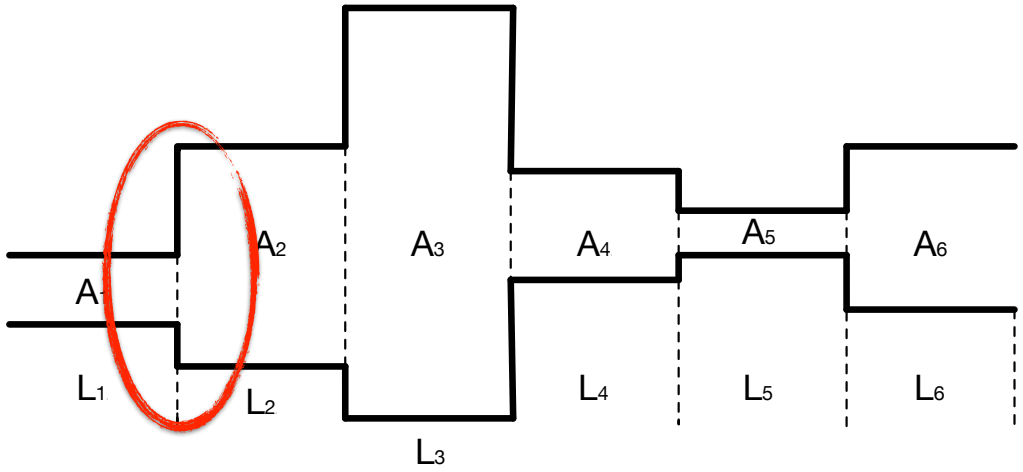
Modeling Human Voice Creation



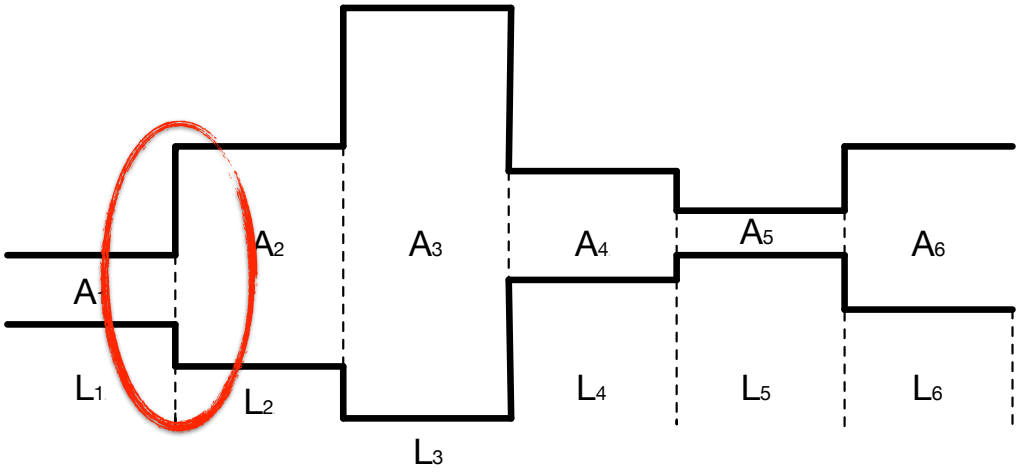
Modeling Human Voice Creation

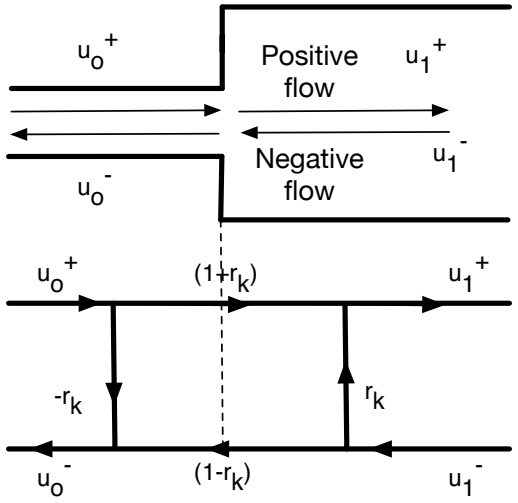


Modeling Human Voice Creation

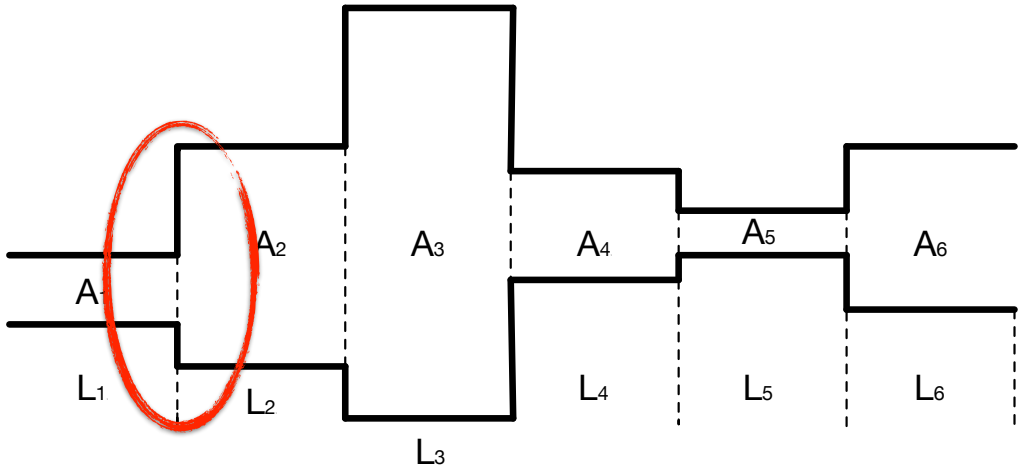


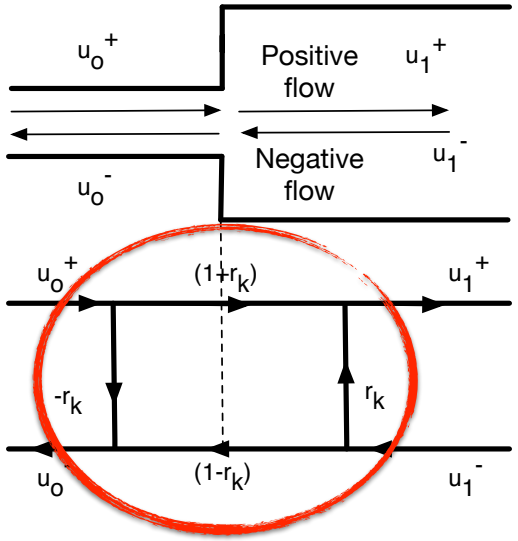
Modeling Human Voice Creation





Modeling Human Voice Creation



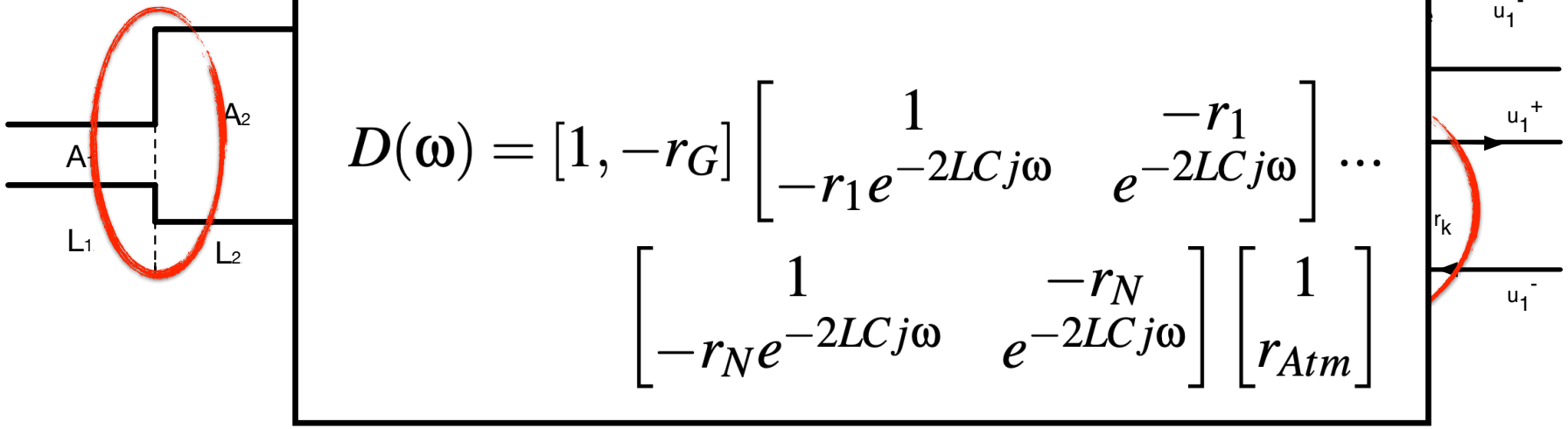


Modeling Human Voice Creation

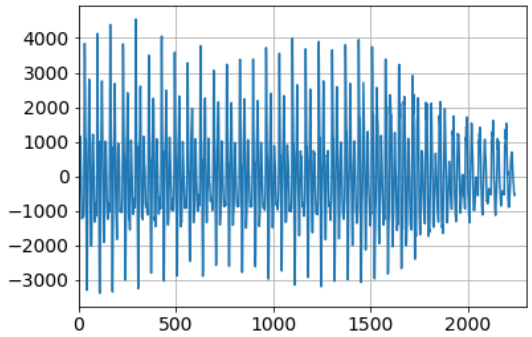
$$V(\omega) = \frac{0.5(1+r_G)\prod_{k=1}^{N}(1+r_k)e^{-LCNj\omega}}{D(\omega)}$$

$$D(\omega) = [1, -r_G]\begin{bmatrix} 1 & -r_1 \\ -r_1 e^{-2LCj\omega} & e^{-2LCj\omega} \end{bmatrix}\cdots$$

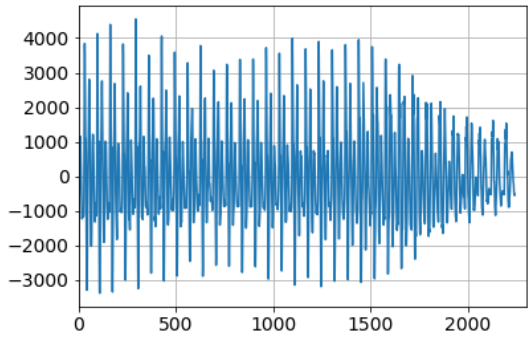
$$\begin{bmatrix} 1 & -r_N \\ -r_N e^{-2LCj\omega} & e^{-2LCj\omega} \end{bmatrix}\begin{bmatrix} 1 \\ r_{Atm} \end{bmatrix}$$

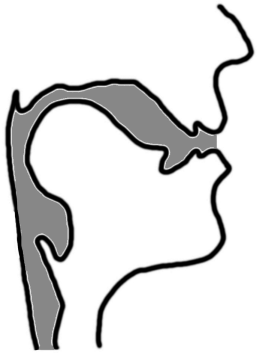


Audio to Biology

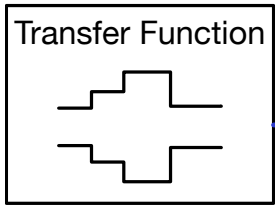


Audio to Biology

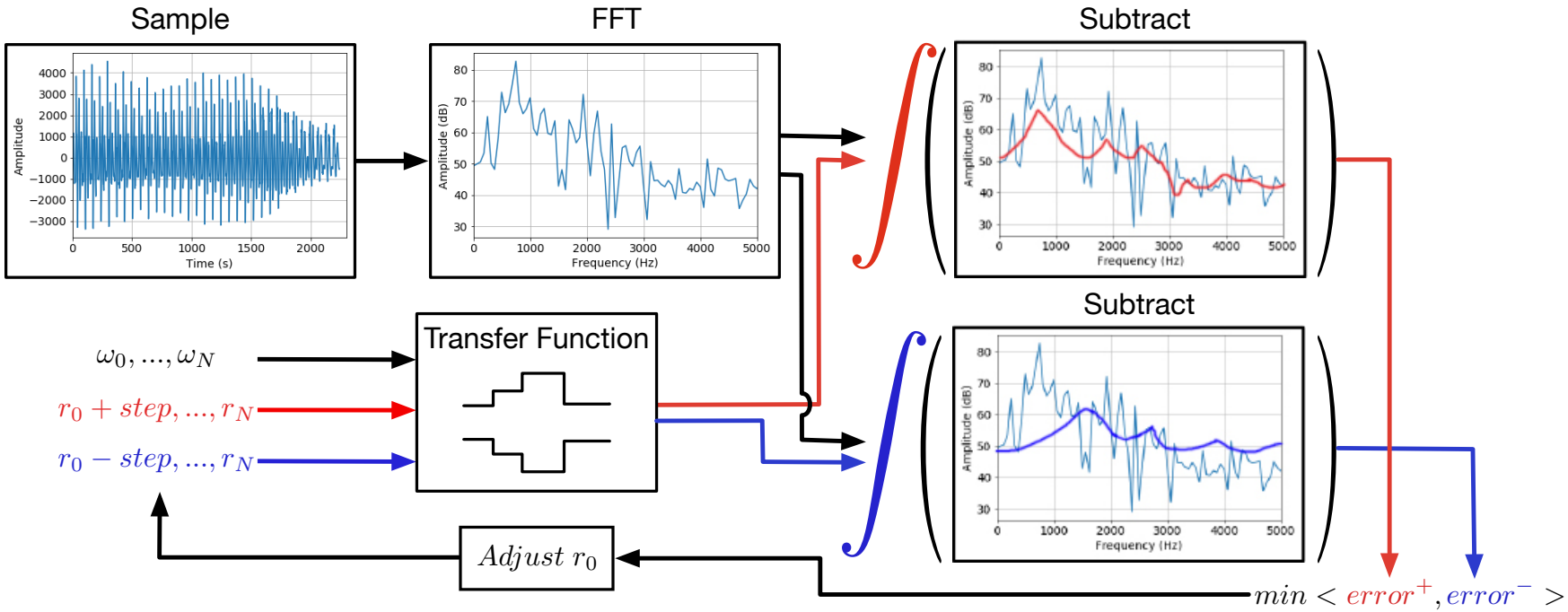




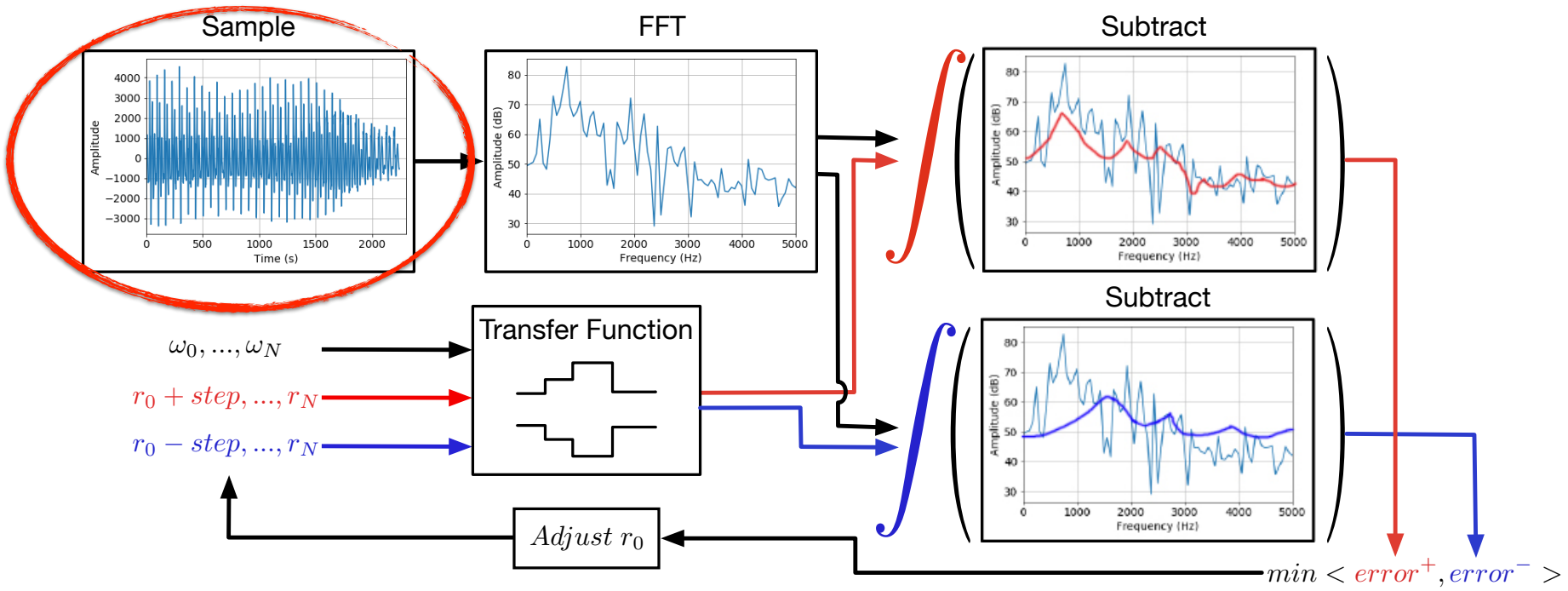
Audio to Biology



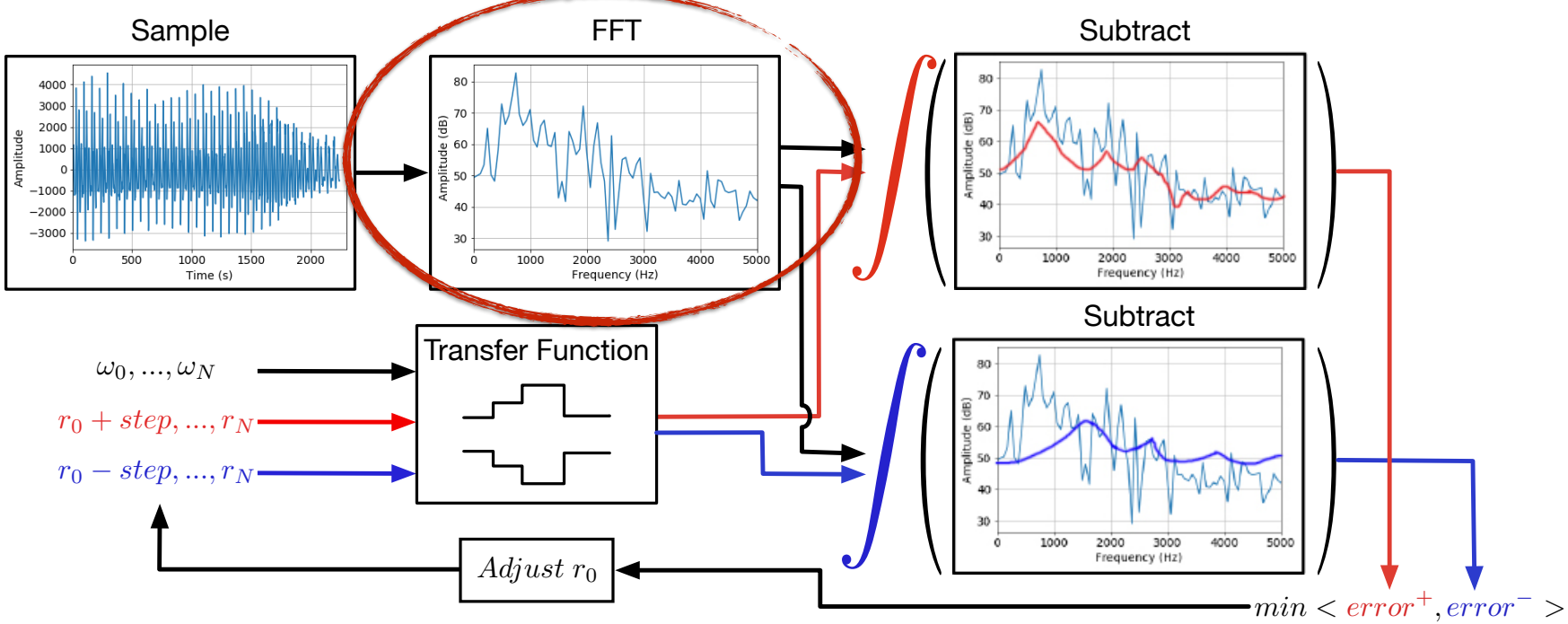
Audio to Biology



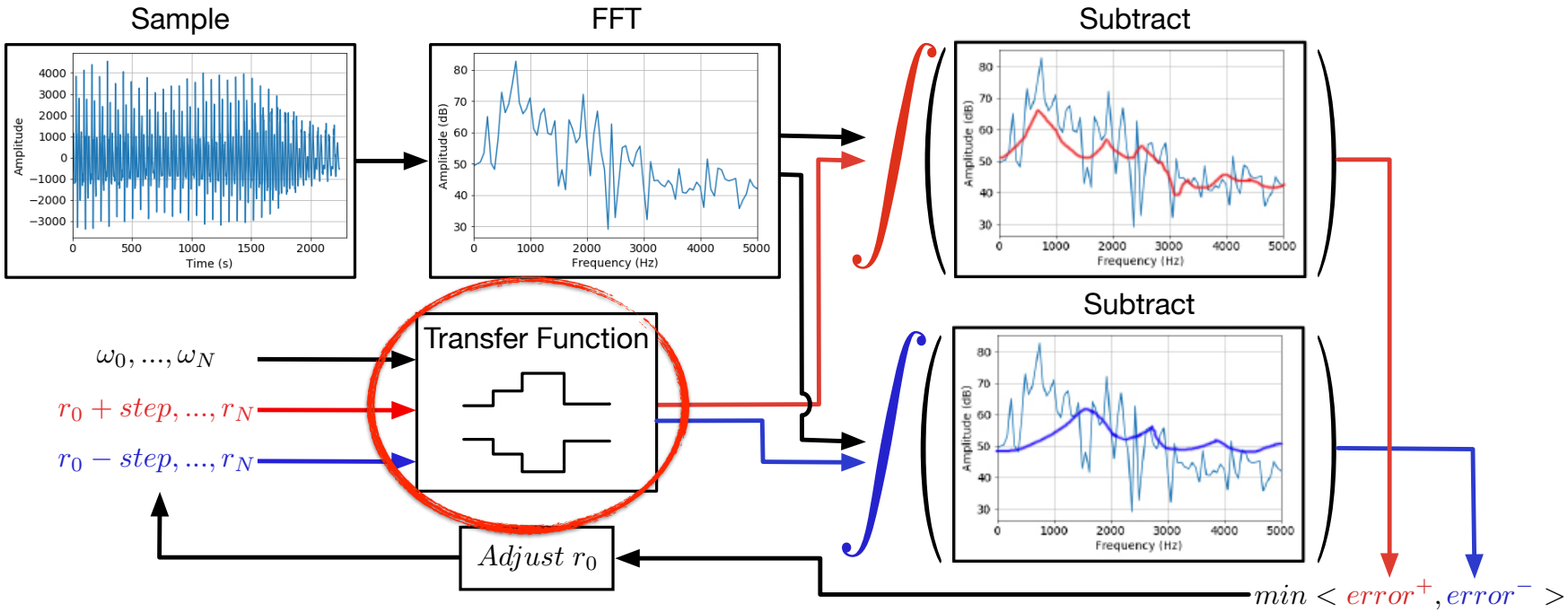
Audio to Biology



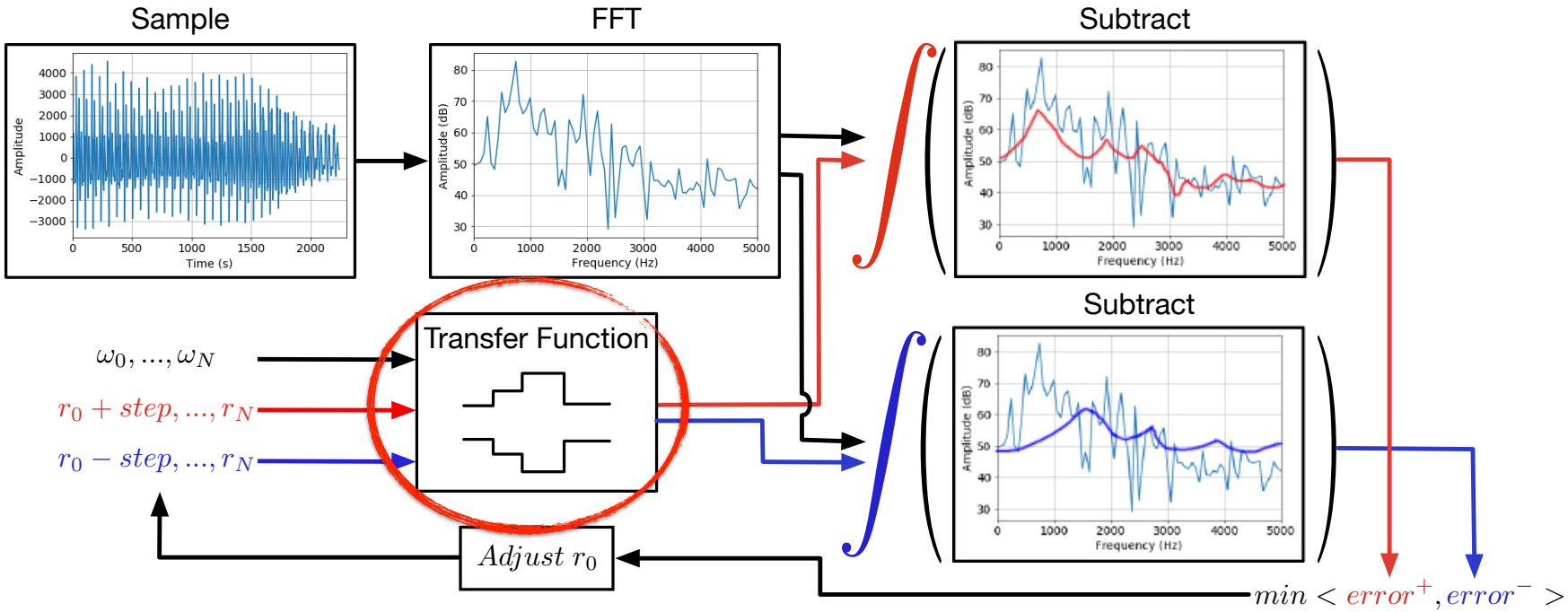
Audio to Biology



Audio to Biology

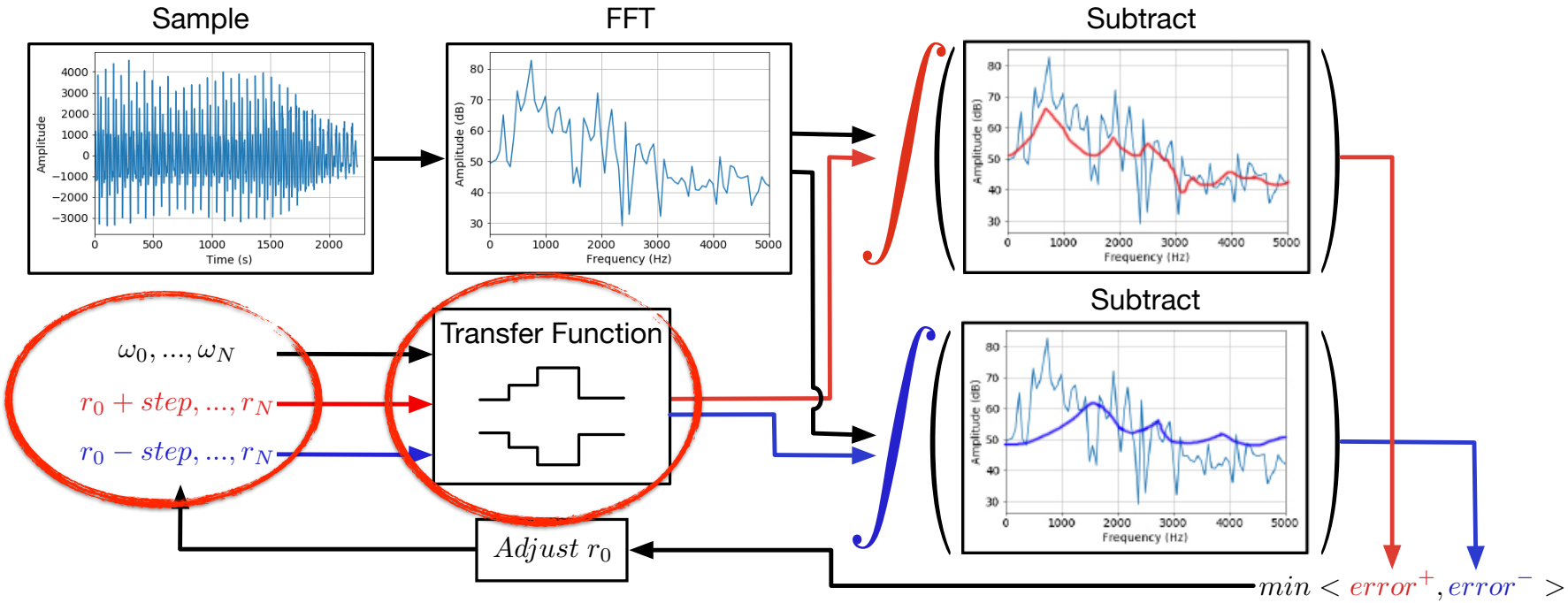


Audio to Biology



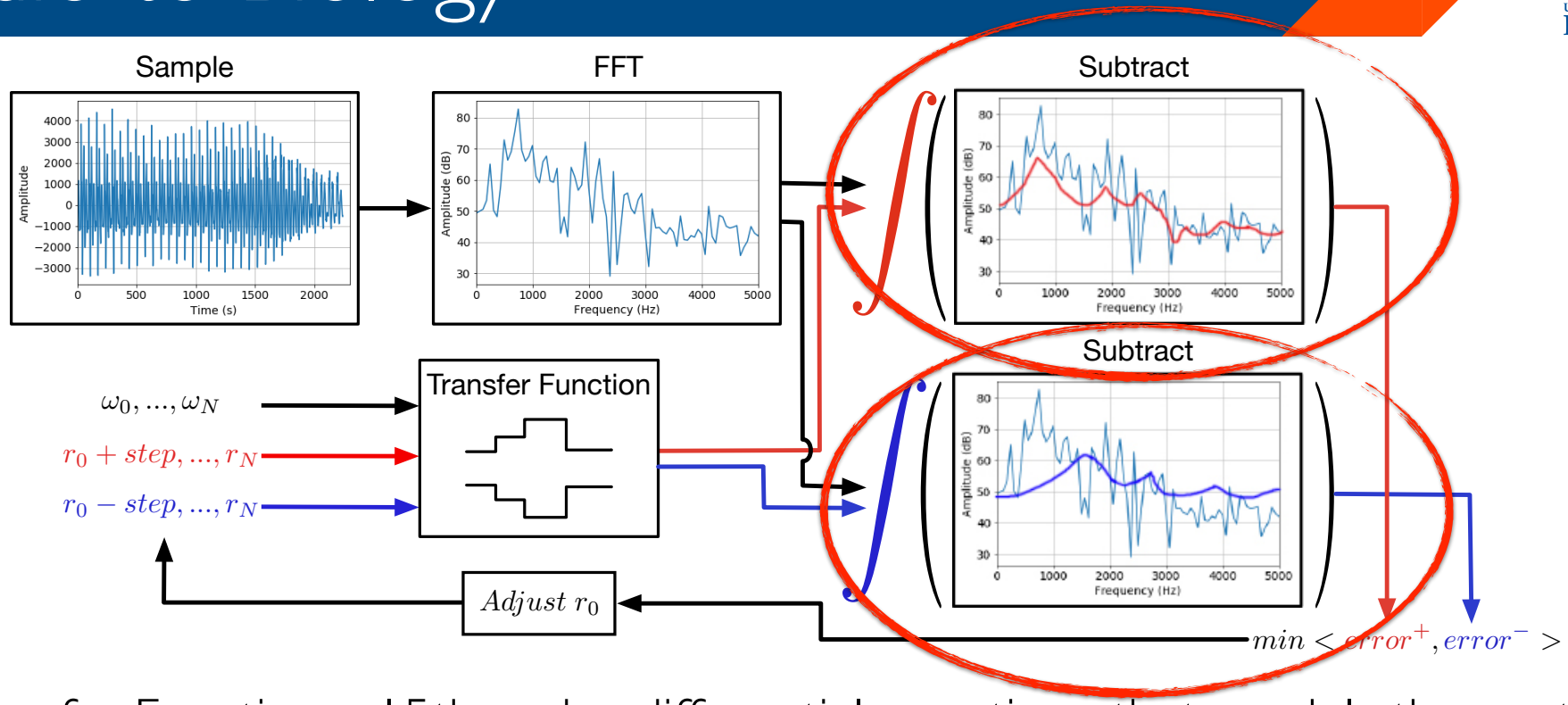
- Transfer Function - 15th order differential equations that models the anatomy
  - Without knowing appropriate boundary conditions, there exists an infinite number of possible solutions

Audio to Biology



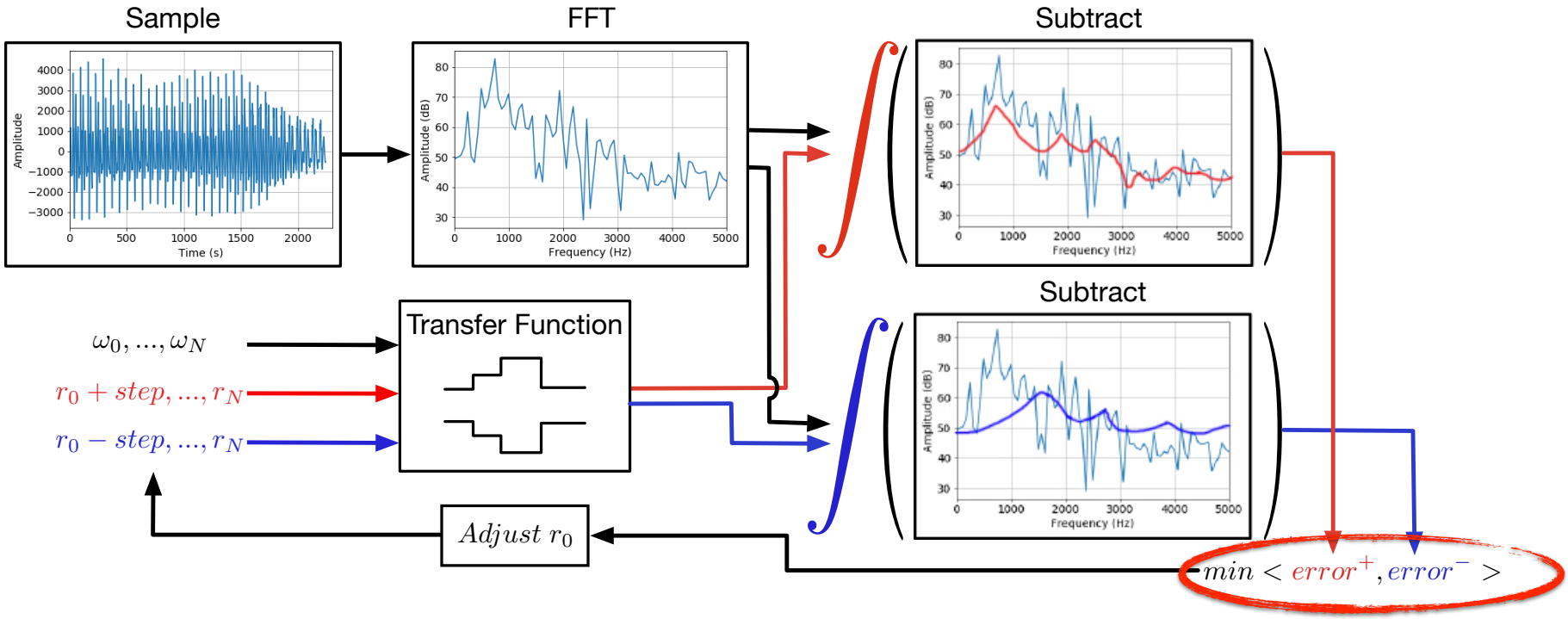
- Transfer Function - 15th order differential equations that models the anatomy
  - Without knowing appropriate boundary conditions, there exists an infinite number of possible solutions

Audio to Biology



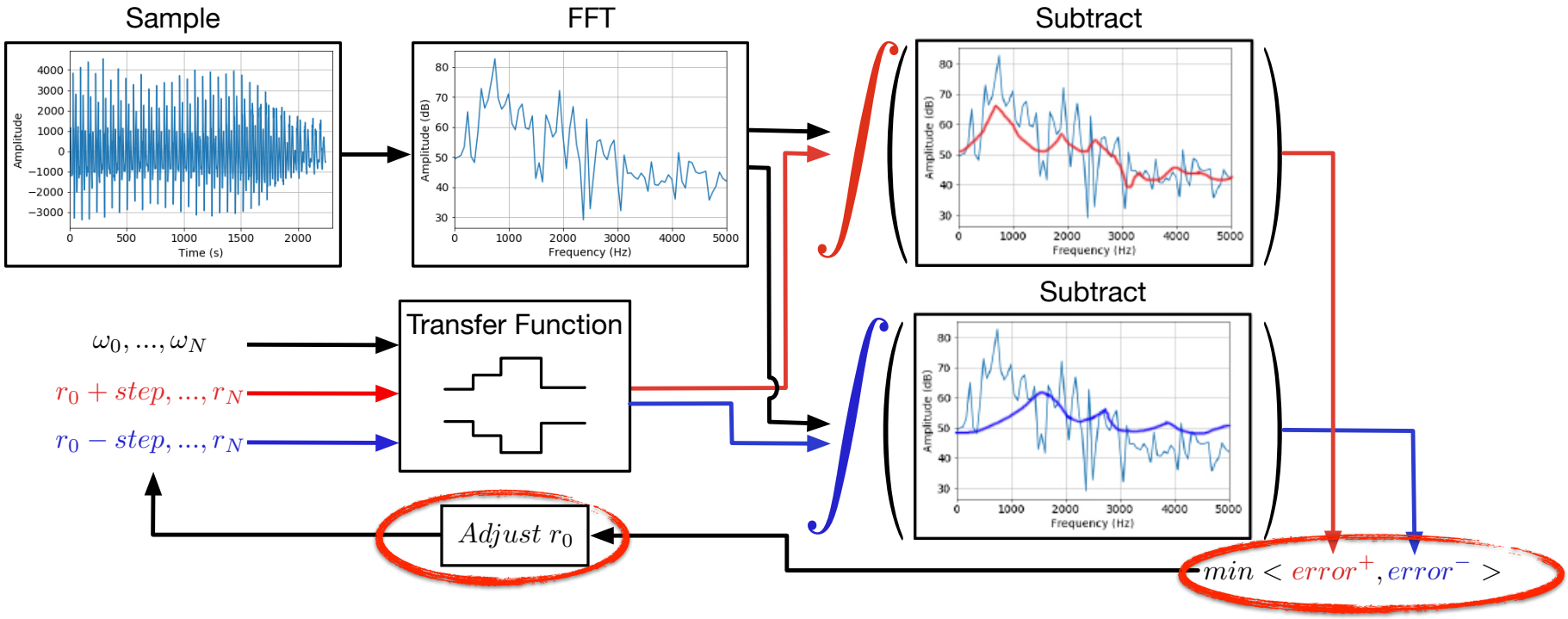
- Transfer Function - 15th order differential equations that models the anatomy
  - Without knowing appropriate boundary conditions, there exists an infinite number of possible solutions

Audio to Biology



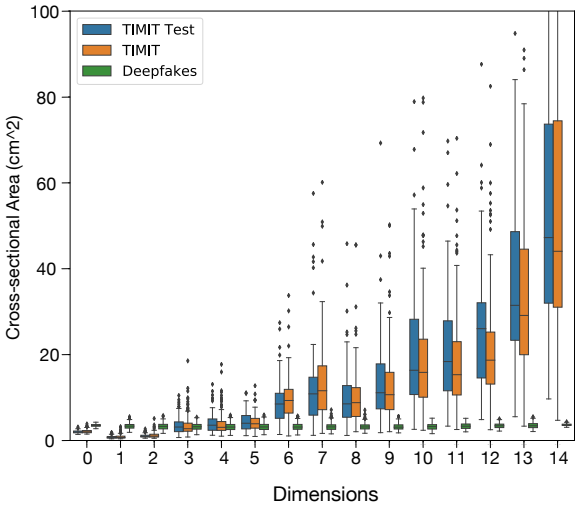
- Transfer Function - 15th order differential equations that models the anatomy
  - Without knowing appropriate boundary conditions, there exists an infinite number of possible solutions

Audio to Biology



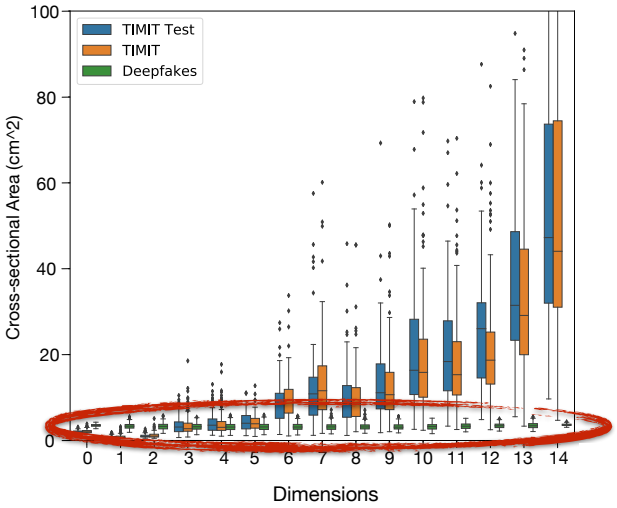
- Transfer Function - 15th order differential equations that models the anatomy
  - Without knowing appropriate boundary conditions, there exists an infinite number of possible solutions

The Straw-man



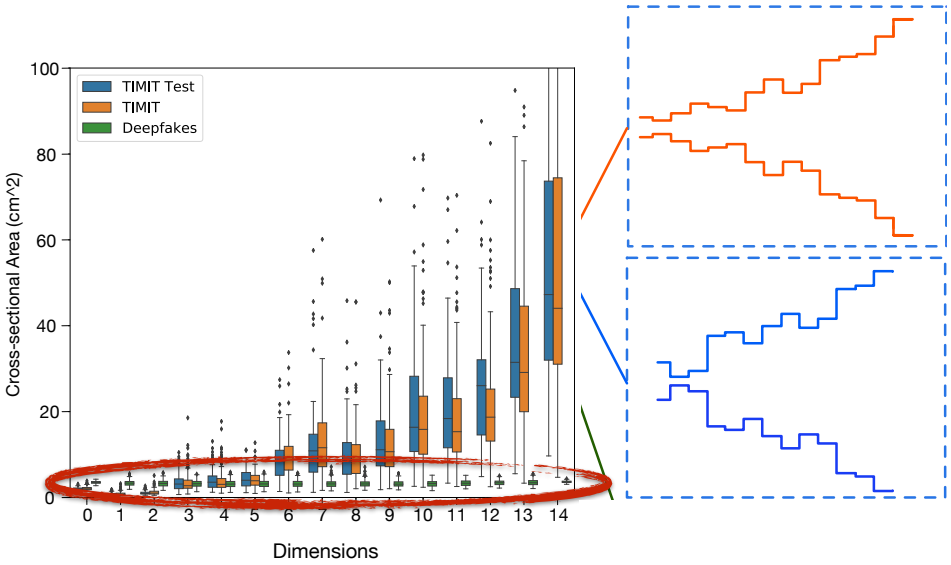
a)

The Straw-man



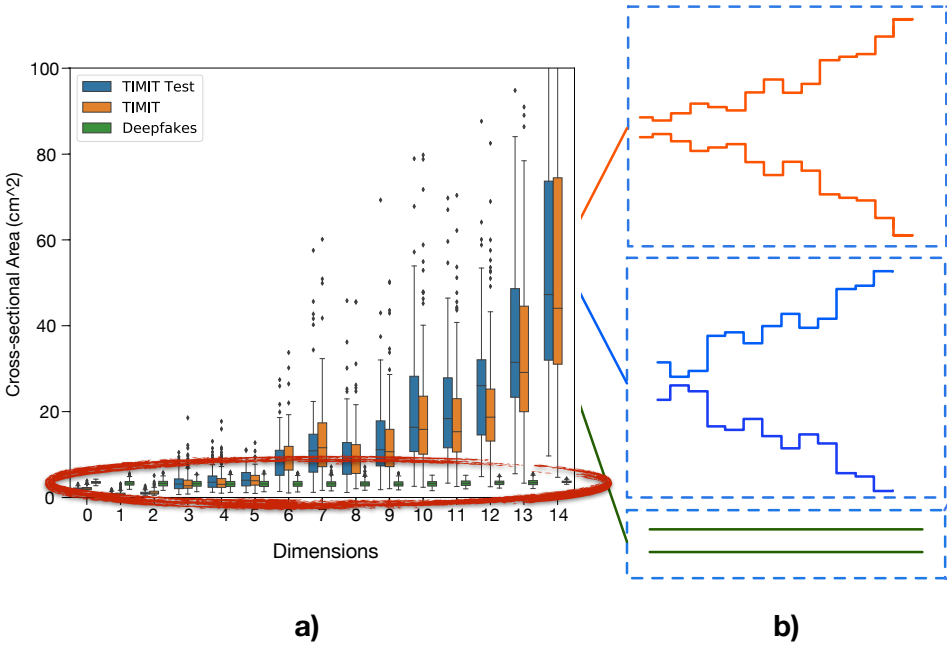
a)

The Straw-man



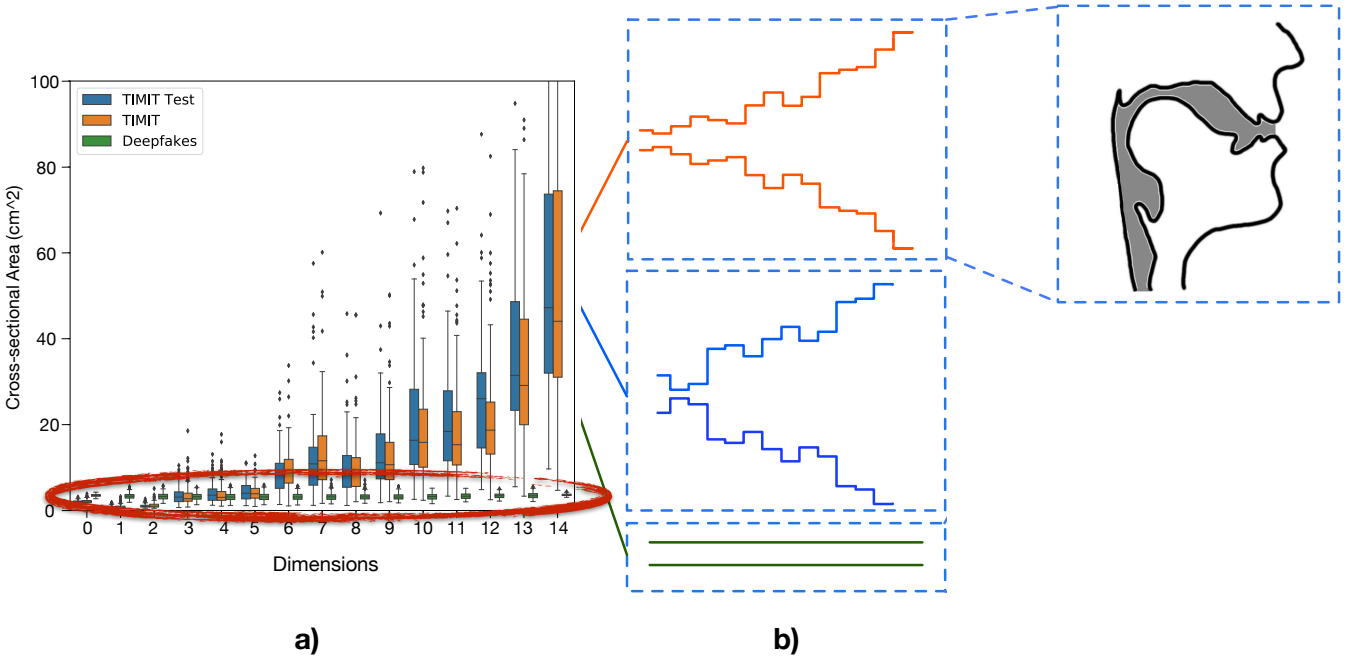
a)

The Straw-man

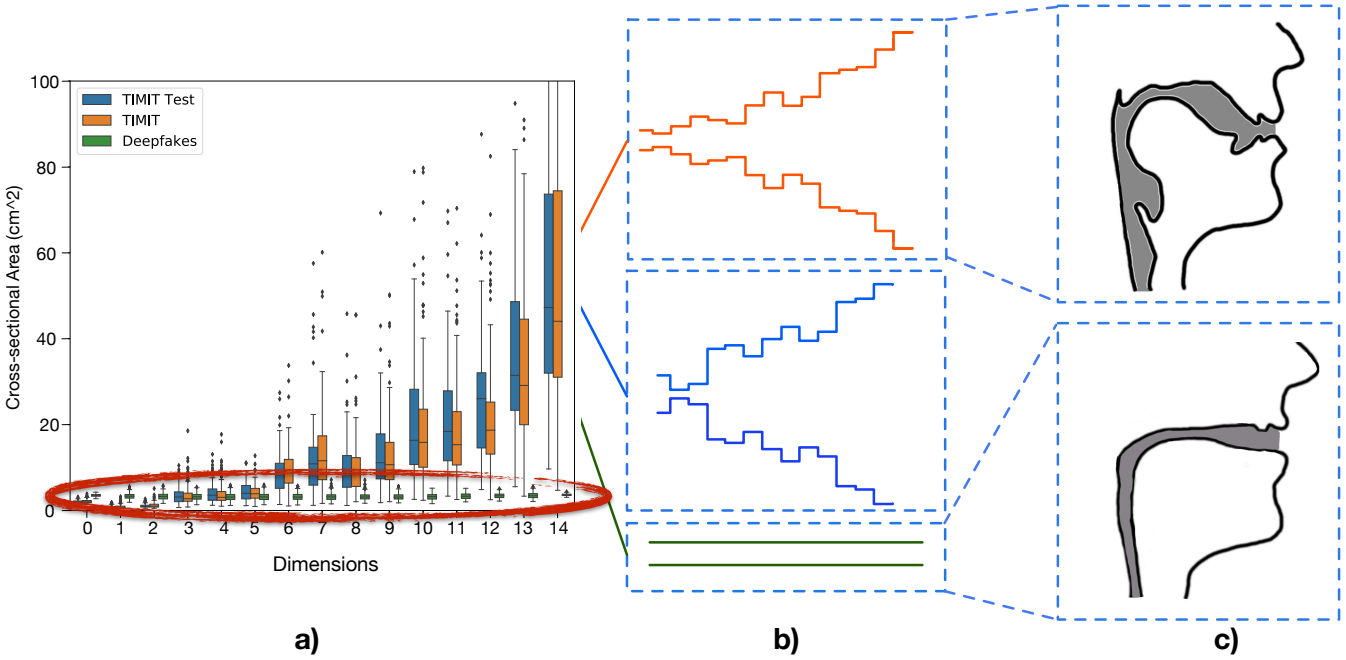


a) b)

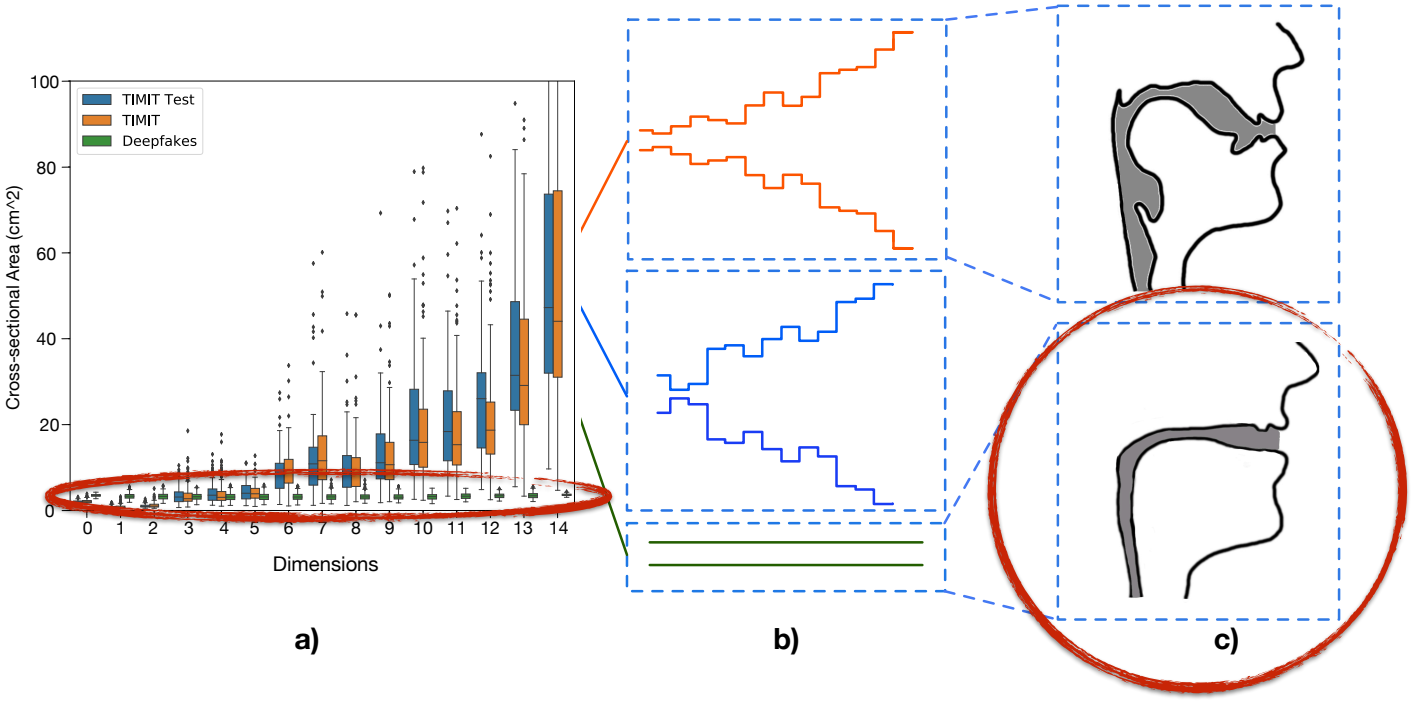
The Straw-man



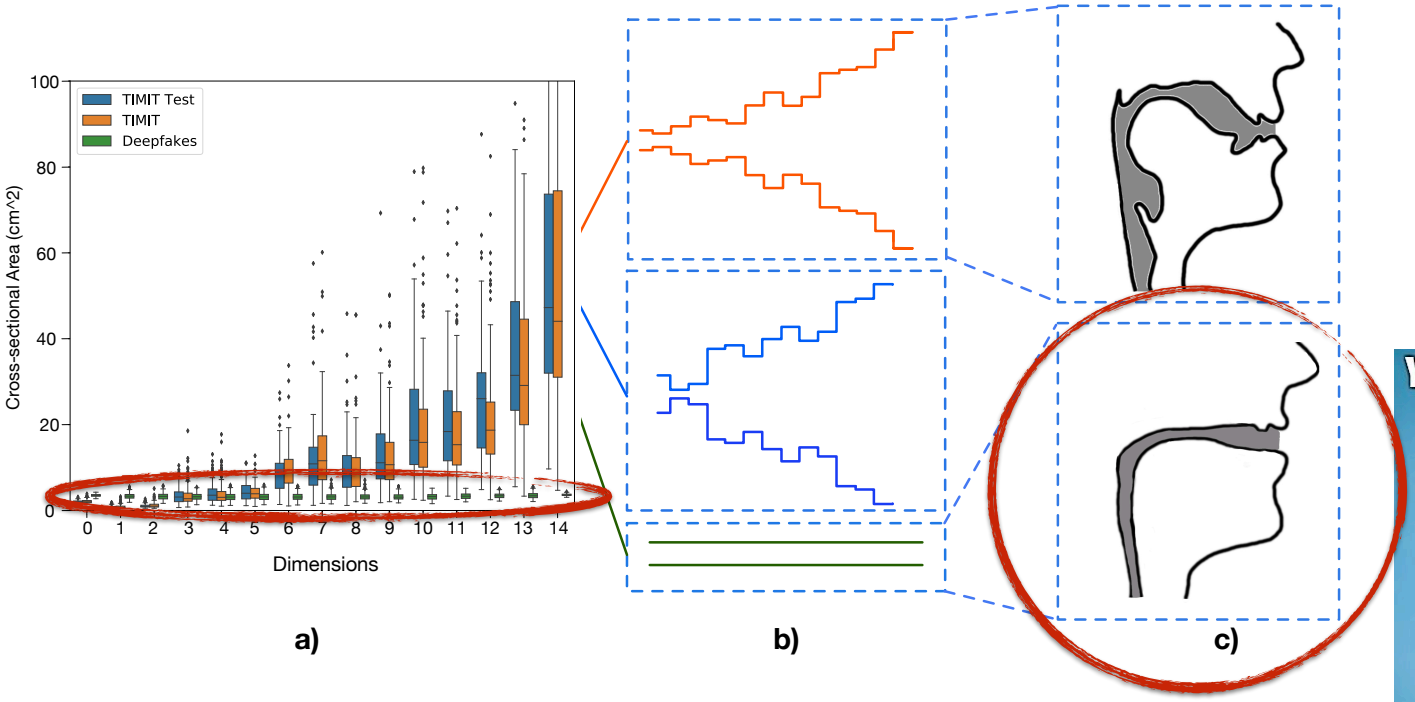
The Straw-man



The Straw-man

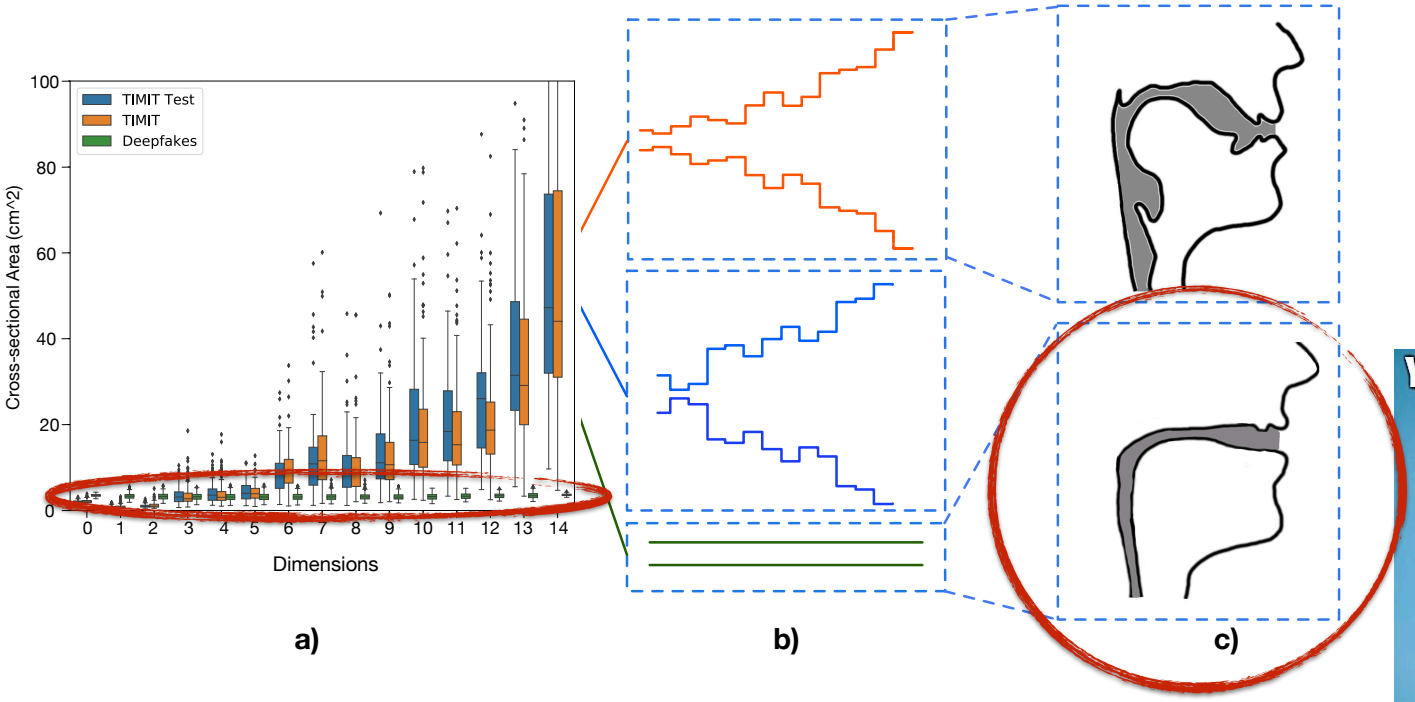


The Straw-man

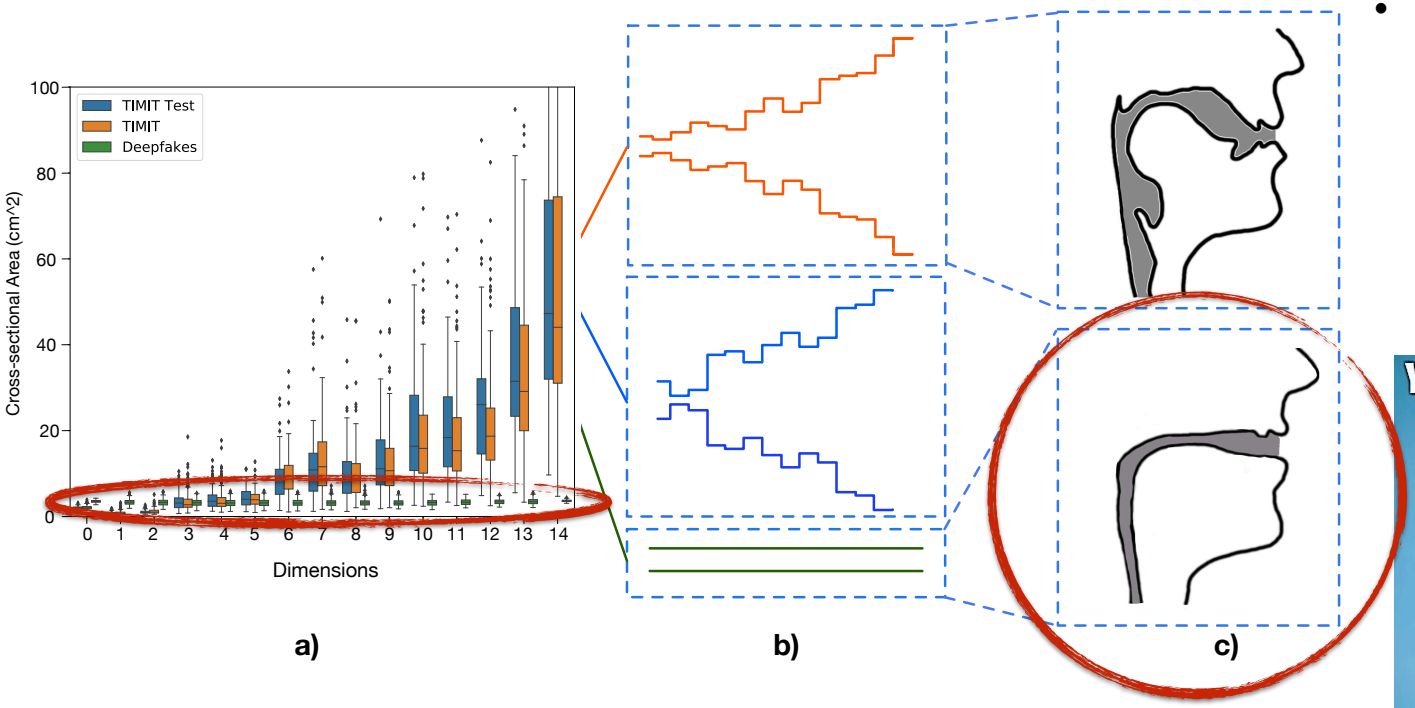


The Straw-man

- Impossible vocal tract

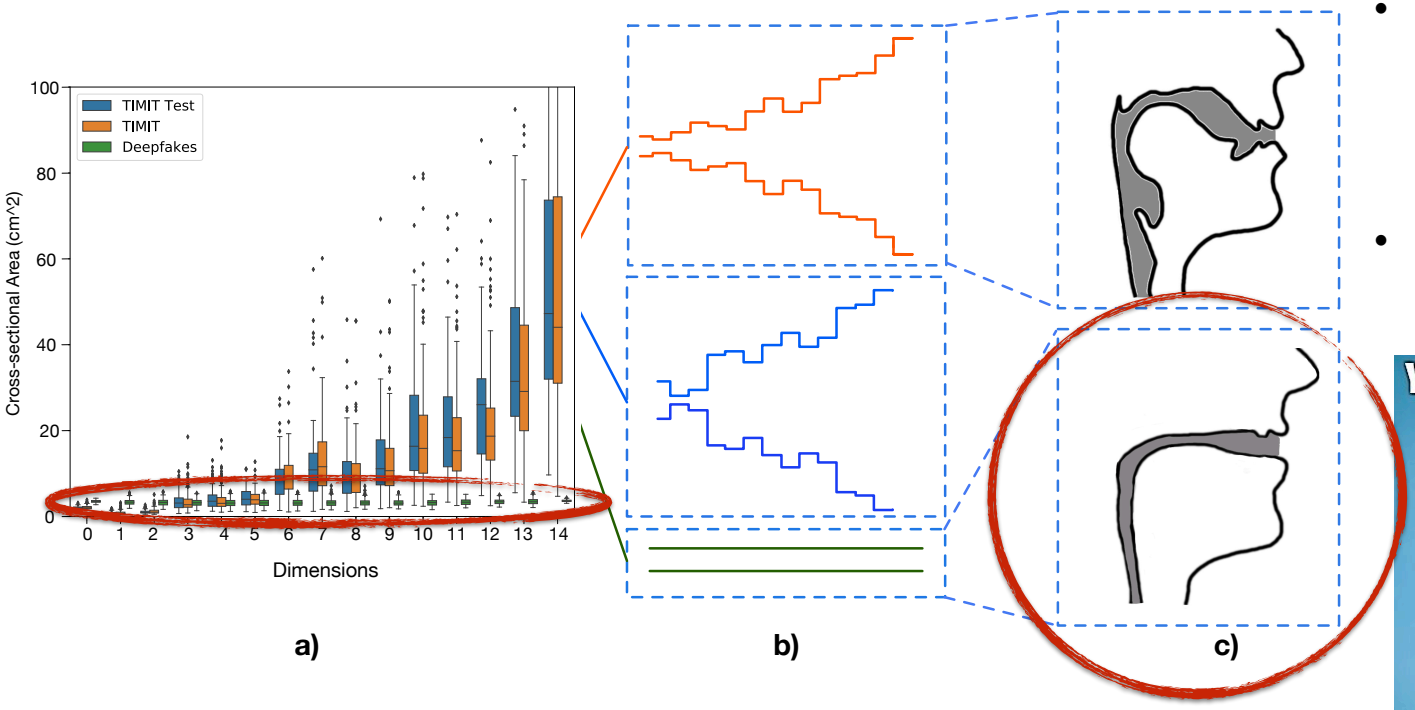


The Straw-man



- Impossible vocal tract
- Model correctly mimics macro-acoustical features

The Straw-man



- Impossible vocal tract
- Model correctly mimics macro-acoustical features
- Model fails to mimic micro-acoustical features

Detection Results

- Full sentence evaluation

Detection Results

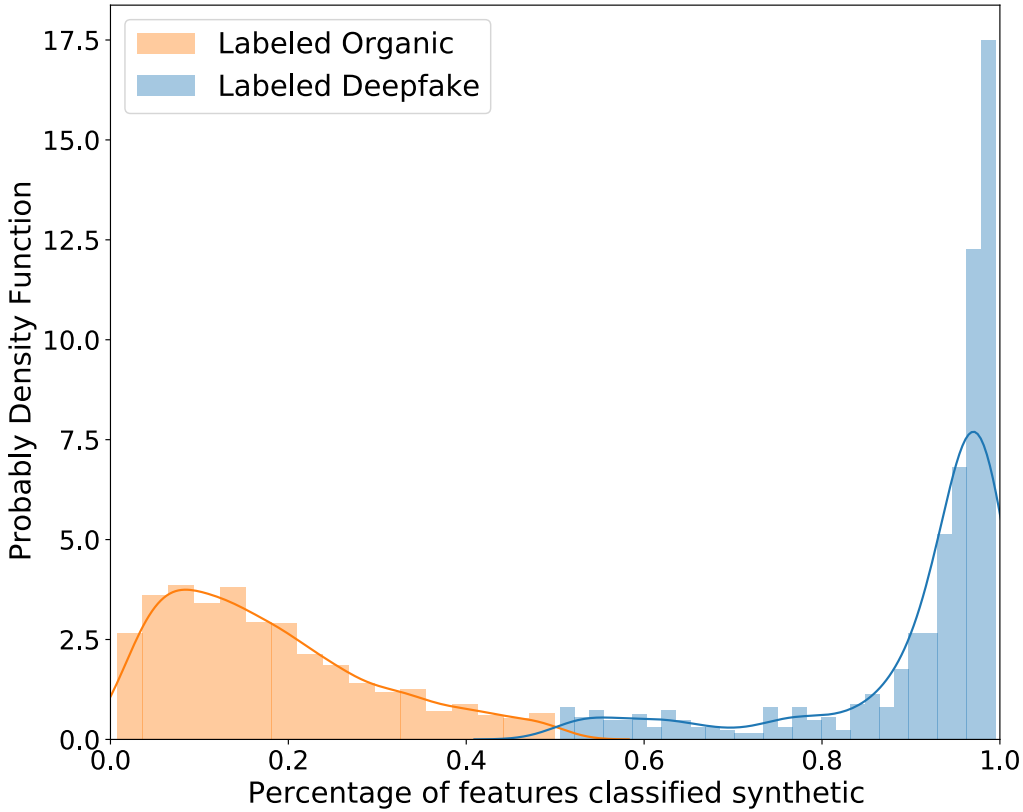
- Full sentence evaluation
  - 4,966 sentences examined

Detection Results

- Full sentence evaluation
  - 4,966 sentences examined
  - 12,525 Bigram-Feature examined

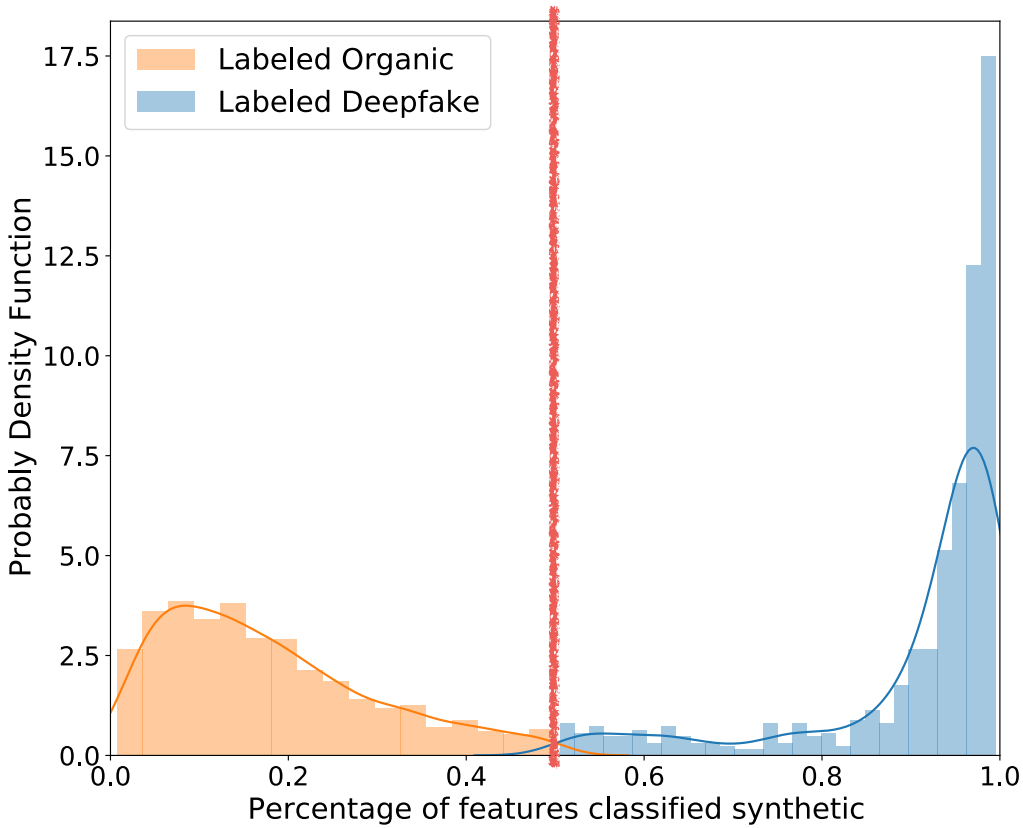
Detection Results

- Full sentence evaluation
  - 4,966 sentences examined
  - 12,525 Bigram-Feature examined



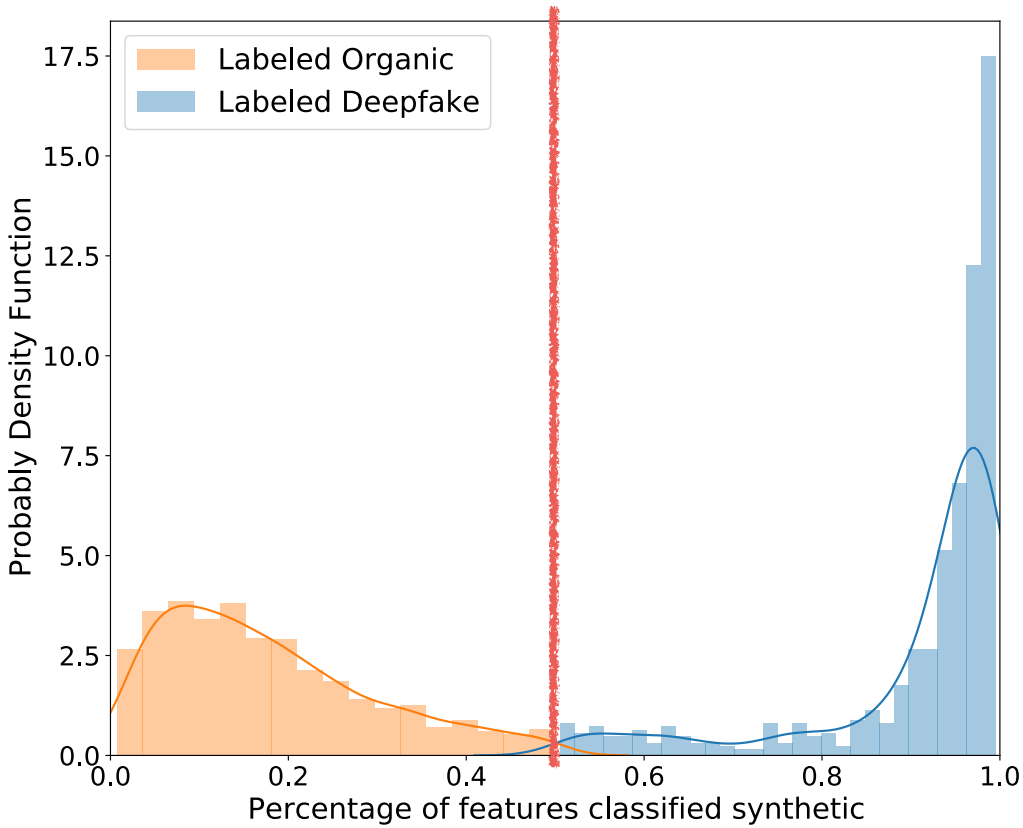
Detection Results

- Full sentence evaluation
  - 4,966 sentences examined
  - 12,525 Bigram-Feature examined



Detection Results

- Full sentence evaluation
  - 4,966 sentences examined
  - 12,525 Bigram-Feature examined
  - Precision: 99.9%
  - Recall: 99.5%



Take Away

- Fluid dynamics can be used to accurately recreate biology from audio
- Reconstruction of anatomical features can allow a defender to differentiate human generated audio from DeepFake audio
  - Are we even listening to a human? We can measure that!

Questions

Logan Blue
bluel@ufl.edu

This work was supported by the Office of Naval Research under grant number ONR-OTA N00014-21-1-2658