# Speech Driven Tongue Animation

Salvador Medina[1,2], Denis Tome[2], Carsten Stoll[2],
Mark Tiede[3], Kevin Munhall[4], Alex Hauptmann[1], Iain Matthews[2]

1 Carnegie Mellon University
2 EPIC GAMES
3 Haskins Laboratories
4 Queen's University

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

## Goal

Animate the tongue and jaw from only speech signal to add realism to facial animations.
Accurately animating the tongue is difficult since:

- Performance capture is not reliable as tongue and teeth are partially visible.
- Manually animating the tongue is nearly impossible.

## EMA Tongue Motion Dataset

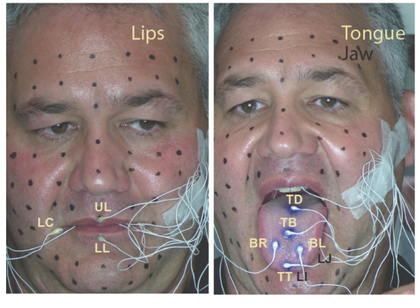

We captured the first large scale electromagnetic articulography (EMA) tongue dataset with parasagittal sensors for animation purposes.

Carstens AG501[1]
- Sample Rate: 250 Hz
- Capture Error < 1mm
- 10 sensors on tongue and lips
- 3 sensors for bite plane

Total samples: 2160
- One English speaker
- 720 Harvard Sentences
- 1440 TIMIT
- 2.55 hours

| EMA Sensor | Placement |
|---|---|
| TD | Tongue Dorsum |
| TB | Tongue Blade |
| BR | Tongue Blade Right |
| BL | Tongue Blade Left |
| TT | Tongue Tip |
| UL | Upper Lip |
| LC | Right Lip Corner |
| LL | Lower Lip |
| LI | Jaw, medial incisors |
| LJ | Jaw, canine & first premolar |

Sensor Placement

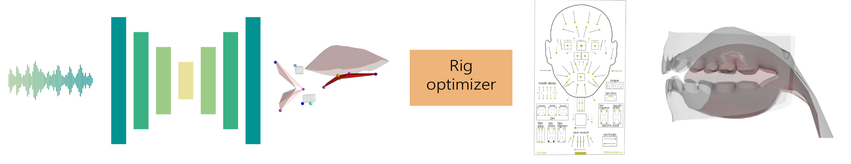



We placed 5 sensors on the tongue, 2 on the jaw and 3 on the lips.

Data available for download at *https://salmedina.github.io/tongue-anim*

## Our Approach

We first estimate tongue landmark positions from speech through an auto-encoder model, and then solve for rig parameters frame by frame to animate a character.




### Speech-to-Tongue Prediction

We evaluate combinations of audio feature encoders and articulation decoders.

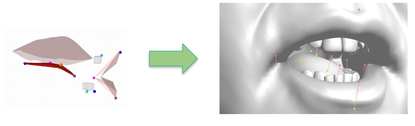
| Encoder | Decoder |
|---|---|
| Phone | MLP |
| MFCC | LSTM |
| DeepSpeech2 | GRU |
| Wav2Vec-C | Transformer |
| Wav2Vec-Z | |

**Audio Features Visualization**

*"Orange juice tastes funny after toothpaste"*


### Rig Fitting

We match the predicted landmark positions with their corresponding points on the mesh by fitting the rig using L-BFGS. We add symmetry constraints on the jaw and lips.

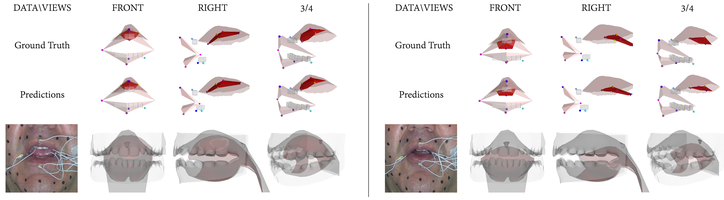


$$\arg\min_{\Theta_i} \; e_{pose}(\Theta_i) + \alpha e_{prior}(\Theta_i)$$

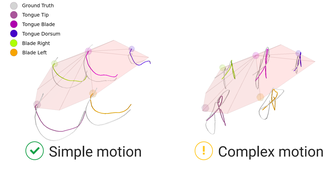
$$e_{pose}(\Theta_i) = \sum_{l=1}^{10+2} \| c(\Theta_i)_l - (c(n)_k + \delta_l^{(i)}) \|^2 \qquad e_{prior}(\Theta_i) = \sum_{l=1}^{L} \| \Theta_l^{(i)} \|$$

## Evaluation

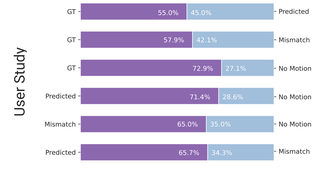
Our method produces realistic tongue animations due to low error inner-mouth pose estimation.




Tongue motions with more complexity are not modeled as accurately.



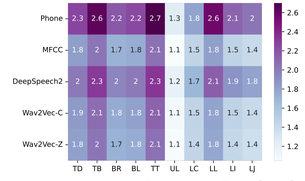

Animations produced from our method are preferred over a no tongue or mismatched animation, and confused with the GT.




Our best results combine Wav2Vec-C features with a bidirectional 5-layered GRU.

| Decoder \ Feature | Phone | MFCC | DS2 | W2V-C | W2V-Z | Num. Parameters | Inference [ms] | Latency [ms] |
|---|---|---|---|---|---|---|---|---|
| MLP 15:5 | 2.445 | 2.075 | 2.393 | 1.959 | 1.937 | $6.62 \times 10^7$ | 0.232 | 300 |
| LSTM-1L | 4.207 | 2.344 | 2.269 | 2.047 | 2.140 | $3.17 \times 10^6$ | 1.150 | 20 |
| LSTM-2L | 4.209 | 2.178 | 4.206 | 1.990 | 4.212 | $5.27 \times 10^6$ | 2.238 | 20 |
| LSTM-5L | 2.656 | 2.037 | 2.264 | 1.999 | 1.960 | $1.16 \times 10^7$ | 5.432 | 20 |
| Bi-LSTM-1L | 3.664 | 2.346 | 2.375 | 2.373 | 3.481 | $6.33 \times 10^6$ | 2.229 | 300 |
| Bi-LSTM-2L | 4.577 | 2.109 | 2.844 | 2.188 | 3.874 | $1.26 \times 10^7$ | 4.512 | 300 |
| Bi-LSTM-5L | 4.365 | **1.912** | 2.218 | 1.927 | 2.929 | $3.15 \times 10^7$ | 11.000 | 300 |
| GRU-1L | 4.150 | 2.200 | 2.250 | 1.949 | 2.071 | $2.38 \times 10^6$ | 1.144 | 20 |
| GRU-2L | 2.623 | 2.117 | 2.179 | 1.897 | 1.980 | $3.95 \times 10^6$ | 2.193 | 20 |
| GRU-5L | 2.661 | 2.006 | 2.184 | 1.916 | 1.954 | $8.68 \times 10^6$ | 5.339 | 20 |
| Bi-GRU-1L | 4.405 | 2.368 | 2.529 | 2.055 | 2.613 | $4.76 \times 10^6$ | 2.290 | 300 |
| Bi-GRU-2L | 3.143 | 1.953 | 2.947 | 1.932 | 2.513 | $9.48 \times 10^6$ | 4.439 | 300 |
| Bi-GRU-5L | **2.341** | 1.973 | **2.058** | **1.757** | **1.784** | $2.37 \times 10^7$ | 10.955 | 300 |
| Transformer | 2.368 | 2.283 | 2.168 | 1.935 | 1.942 | $5.045 \times 10^7$ | 3.515 | 300 |

Experimental Results. Error is temporal mean MSE [mm].




Landmark Prediction Error [mm]

## References

[1] Carstens Medizinelektronik GmbH. 3D electromagnetic articulograph. https://www.articulograph.de/
[2] Rothauser, E. H. "IEEE recommended practice for speech quality measurements." *IEEE Trans. on Audio and Electroacoustics* 17 (1969): 225-246.
[3] Zue, Victor, Stephanie Seneff, and James Glass. "Speech database development at MIT: TIMIT and beyond." *Speech communication* 9.4 (1990): 351-356.
[4] Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." *International conference on machine learning*. PMLR, 2016.
[5] Schneider, Steffen, et al. "wav2vec: Unsupervised pre-training for speech recognition." *arXiv preprint arXiv:1904.05862* (2019).

## Conclusions

- Our inner-mouth mocap dataset enables the training of data-driven models
- Deep Learning audio representations outperform traditional methods for speech-animation
- Simple-RNN based articulation decoders generalize across gender, age, and prosody
- Limited lip animation due to the sparsity of the sensors