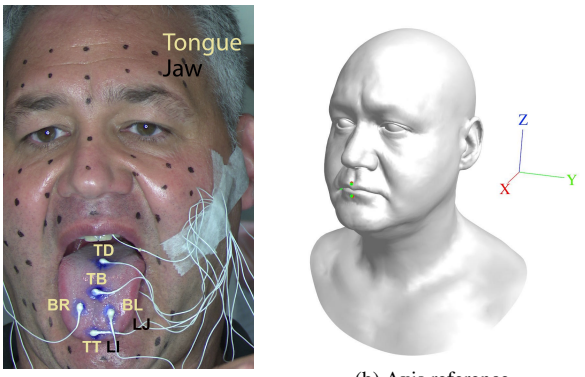
Speech Driven Tongue Animation - Supplementary Material

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]Queens University
{salvadom, alex}@cs.cmu.edu, tiede@haskins.yale.edu, munhallk@queensu.ca
{denis.tome, carsten.stoll, iain.matthews}@epicgames.com

1. Further Experiments

We repeated the same set of experiments as described on Section 5, but with a larger input window of 1000 ms instead of 300 ms. As we can see in Table 1, the results across all the models and features follow the same pattern with an overall improvement, at the cost of an increase in the number of parameters and inference time per model. The inference time of these models make them **non-practical to use in real-time applications**, such as interactive avatars in video games or telecommunications.

We also found in both experimental setups, training the models with an input of 300 ms and 1000 ms that traditional audio feature representations such as phonemes and MFCC did not generalize as well as deep-learning based representations on out-of-domain speech audio. A comparison of the resulting animations is shown in the supplementary video.



(a) Sensor Placement (b) Axis reference

Figure 1. Reference images. (a) Sensor placement on tongue, lips and jaw. (b) Head-axis reference: X-axis describes the anteroposterior direction (back to front), Y-axis describes the mediolateral direction (right to left), and Z-axis describes the vertical direction (bottom to top).

2. Jaw Motion Analysis

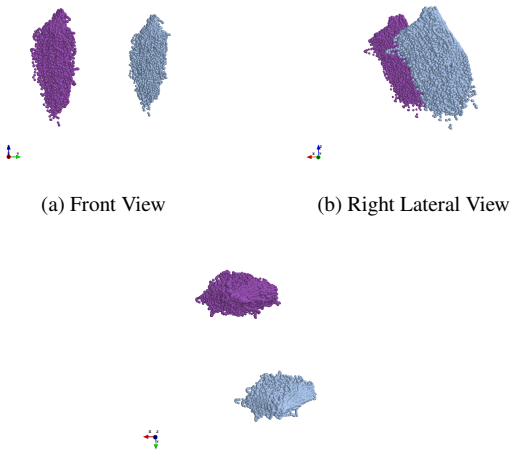
In Figure 2 we visualized all the samples of both jaw sensors LI (midsagittal) and LJ (parasagittal) in a 3D scatter plot from three different views. As in can be seen, our captured data follows the Posselt's Envelope of Motion (PEM) [1]. The PEM describes the range under which the jaw can move due to the physiological constraints based on the bones, muscles and tendons that form the jaw. For context, we can see in Figure 1b that the reference frame of our captured data is the following: the X-axis describes the anteroposterior direction, the Y-axis the mediolateral direction, and the Z-axis the vertical direction. As carefully described in [2], the jaw motion from a frontal view usually follows a shield appearance as seen in Figure 2a. From a sagittal or lateral view, it follows a prolonged fang shape as shown in Figure 2b. While from a top view, the jaw motion follows a diamond shape as the one displayed on Figure 2c. Our visualization shows that our data follows such shapes with undefined edges. The main reason being that the actor did not do any extreme articulation during the capture session, as he only uttered regular sentences at a normal pace and fast pace in a neutral emotion. This visualization supports that it would be appropriate to increase the variability of the gesticulations during a capture session in future work.

3. Predicted Tongue Motion Analysis

We traced the predicted motion and compared it against the ground truth on test data samples for a better understanding of our best model's behavior beyond the temporal mean sample error. The trace visualization in Figure 3a shows how the long motions are close to the ground truth. However, the sensor's positions have an overall shift to the front of the mouth, as these depend on the predicted initial position of the tongue at the beginning of the sequence. We also found that the articulatory decoder network learned to generate shortcuts on fast and complex motions which take place within a small space. In Figure 3b, we can see a clear shortcut on the motions of the *Blade Left* and *Tongue Tip* sensors. These results are reasonable since the models were trained only on a Mean Square Error loss leading the pre-

Table 1. Model architecture evaluation using different audio feature representations. Models were trained with 1 s input windows of audio. The error is the temporal MSE in mm calculated through the validation split. The number of parameters reported is the amount of trainable parameters per architecture design. The inference time is the mean time over the validation split measured as ms per second of audio input.

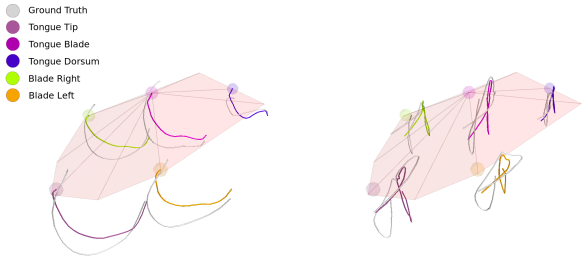
Decoder \ Feature	Phone	MFCC	DeepSpeech2	W2V-C	W2V-Z	Num. Parameters	Inference [ms]
MLP 50:15	2.315	2.157	2.619	1.957	1.928	3.29×10^8	0.309
LSTM-1L	2.657	2.299	2.350	2.048	4.219	3.17×10^6	1.150
LSTM-2L	4.216	2.219	2.342	2.016	2.026	5.27×10^6	2.238
LSTM-5L	2.609	2.133	2.331	2.014	1.994	1.16×10^7	5.432
Bi-LSTM-1L	3.355	2.074	2.204	1.977	2.272	6.33×10^6	2.229
Bi-LSTM-2L	2.268	1.874	2.096	1.825	1.781	1.26×10^7	4.512
Bi-LSTM-5L	2.247	1.732	1.987	1.754	1.708	3.15×10^7	11.000
GRU-1L	4.195	2.213	2.283	1.943	2.001	2.38×10^6	1.144
GRU-2L	2.559	2.098	2.248	1.905	1.945	3.95×10^6	2.193
GRU-5L	2.570	2.003	2.235	1.908	1.943	8.68×10^6	5.339
Bi-GRU-1L	4.304	1.964	2.094	1.828	1.910	4.76×10^6	2.290
Bi-GRU-2L	2.206	1.784	2.091	1.744	1.714	9.48×10^6	4.439
Bi-GRU-5L	**2.179**	**1.660**	**1.935**	**1.684**	**1.648**	2.37×10^7	10.955
Transformer	2.349	2.393	2.139	1.926	2.044	5.049×10^7	3.552



(a) Front View (b) Right Lateral View

(c) Top View

Figure 2. Different views for Posselt's Envelope of Motion for both jaw sensors LI and LJ: midsagittal (purple) and parasagittal (blue) respectively.



(a) Long Motion (b) Prediction Shortcut

Figure 3. Visualization of 300 [ms] of predicted motion on a test sample vs. ground truth. Ground truth is in gray and the predicted motion is in color according to the sensor. (a) Shows an example of a predicted long motion. (b) Shows an example of motion shortcuts on prediction visible on *Blade Left* and *Tongue Tip* sensors.

dicted sensor position to a local minimum. Future work should address these issues by adding more constraints to the learning loss.

3.1. Landmark Pose Prediction Error Analysis

We analyzed the landmark position error of the model predictions across all audio representations on the and summarize the results in Figure 4.

Overall, the model trained with phonetic representation shows the highest error followed by the model trained with DeepSpeech2 audio features. The models trained using MFCC and both Wav2Vec features follow a similar pattern across all the landmark predictions.

The most accurate landmark prediction is the upper lip (UL), with a mean error of 1.16 mm across all audio features. The least accurate landmark is the tongue tip (TT) with a mean error of 2.26 mm. Most of the errors occur while predicting landmark positions for the tongue and lower lip (LL) as these points have the most significant motion during an utterance. Midsagittal and parasagittal jaw landmark (LI, LJ) predictions have lower error, which

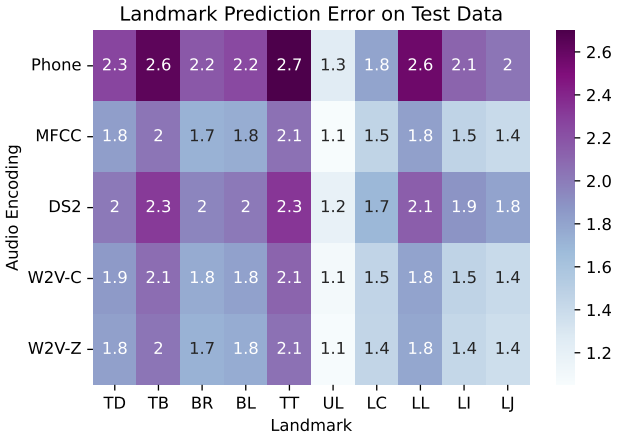


Figure 4. Landmark Prediction Error on the test set of the bidirectional 5-layered GRU over all audio encodings. The error values shown are the mean sample error in mm of each landmark.

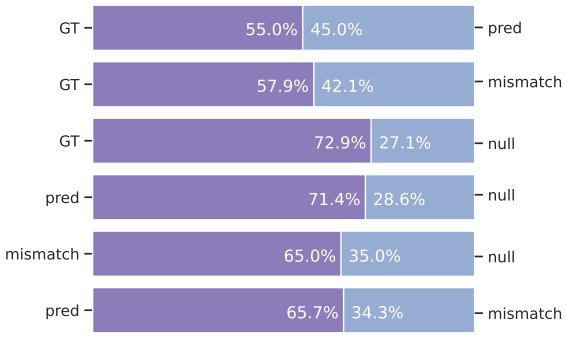


Figure 5. Pairwise preference perception user-study results. *GT*: Ground-truth animation; *pred*: predicted animation; *mismatch*: tongue mismatched animation; *null*: nullified tongue animation

seems reasonable since the jaw moves slower than the tongue and lower lip in general.

4. Perceptual User Study

We conducted a pairwise perceptual user-study to evaluate our generated animations using ten test sentences. The range of jaw and tongue motion go from tongue visible tongue motions to almost closed teeth. For each test sentence, we generated four animations using: *a)* ground truth (*GT*), *b)* prediction of our best model (*pred*), *c)* an animation taking the lips from *GT* and injecting the tongue animation from another sample (*mismatch*), and *d)* by removing the tongue motion from *GT* (*null*). We generated six pairwise comparison videos per sample, presenting a total of 60 videos to 15 users and asked them to select the *most realistic animation* from each pair.

The results shown in Figure 5 present the percentage of users that preferred one animation over the other. As we can see, most users prefer *GT* over the others (rows 1-3). Moreover, an interesting finding from this user study reveals that users prefer an animated tongue over a nullified tongue motion, even when the tongue motion is not matching the spoken sentence (rows 3-5). Finally, the fact that users prefer *GT* over *mismatch* and *pred* over *mismatch* (row 2, 6) proves the consistency of our estimations as most of the users prefer *pred* over *mismatch* (row 6).

5. External Asset License Details

In this work, we used public tools and pre-trained models to encode the speech audio. We used the Montreal Forced Aligner[1] to extract the phoneme representations by using the English acoustic model, which has an MIT license. To obtain the MFCC features, we used the Python package Librosa[2] released under the ISC license. For DeepSpeech2 features, we used publicly available code[3] and its LibriSpeech pre-trained model[4] released under the MIT license. Finally, we used the the *large* Wav2Vec pre-trained model[5] from the Fairseq[6] package to encode the input audio, both the package and the model are released under an MIT license. As a final remark, LibriSpeech is a corpus which comprises 1000 hours of diverse speech audio, and is released under the CC BY 4.0 license.

The high quality animations shown in the supplementary video were created using the MetaHuman Creator[7] tool and rendered in Unreal Engine with permission from Epic Games.

References

[1] Ulf Posselt et al. Range of movement of the mandiblew. *The Journal of the American Dental Association*, 56(1):10–13, 1958. 1

[2] Gaspard Zoss, Derek Bradley, Pascal Bérard, and Thabo Beeler. An empirical rig for jaw animation. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 1

[1] https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

[2] https://github.com/librosa/librosa

[3] https://github.com/SeanNaren/deepspeech.pytorch

[4] https://github.com/SeanNaren/deepspeech.pytorch/releases

[5] https://github.com/pytorch/fairseq/tree/main/examples/wav2vec

[6] https://github.com/pytorch/fairseq

[7] https://www.unrealengine.com/en-US/digital-humans

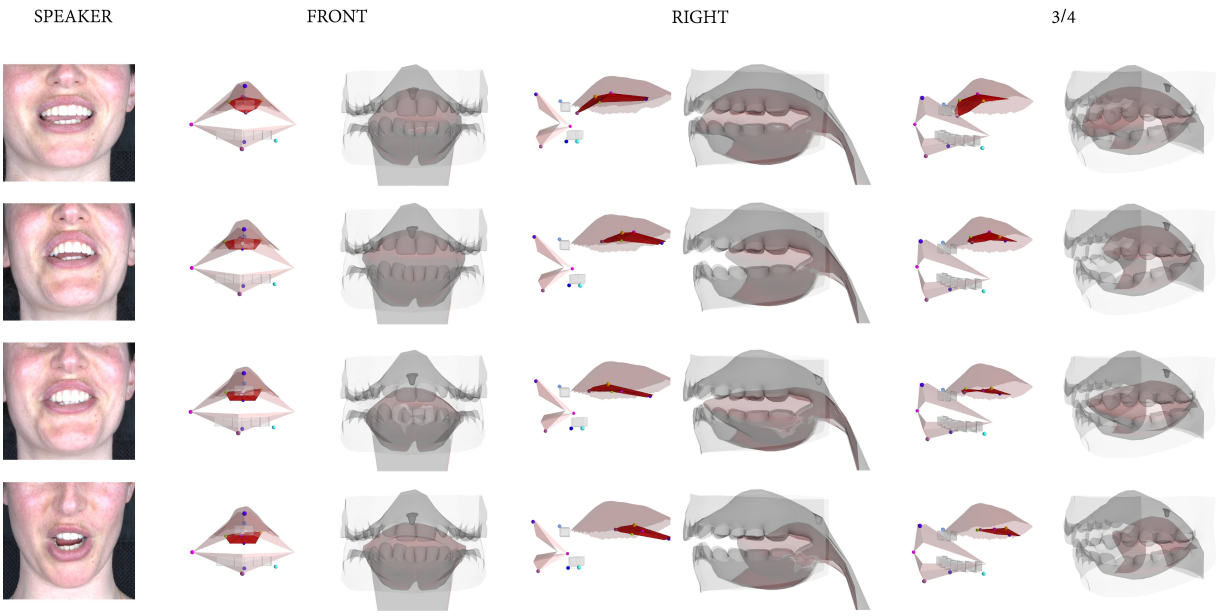


Figure 6. Frames from the generated animation at inference time for a speech input incoming from a speaker not seen during training. For each camera view, both predicted landmark locations and the solved animation rig outputs are shown.

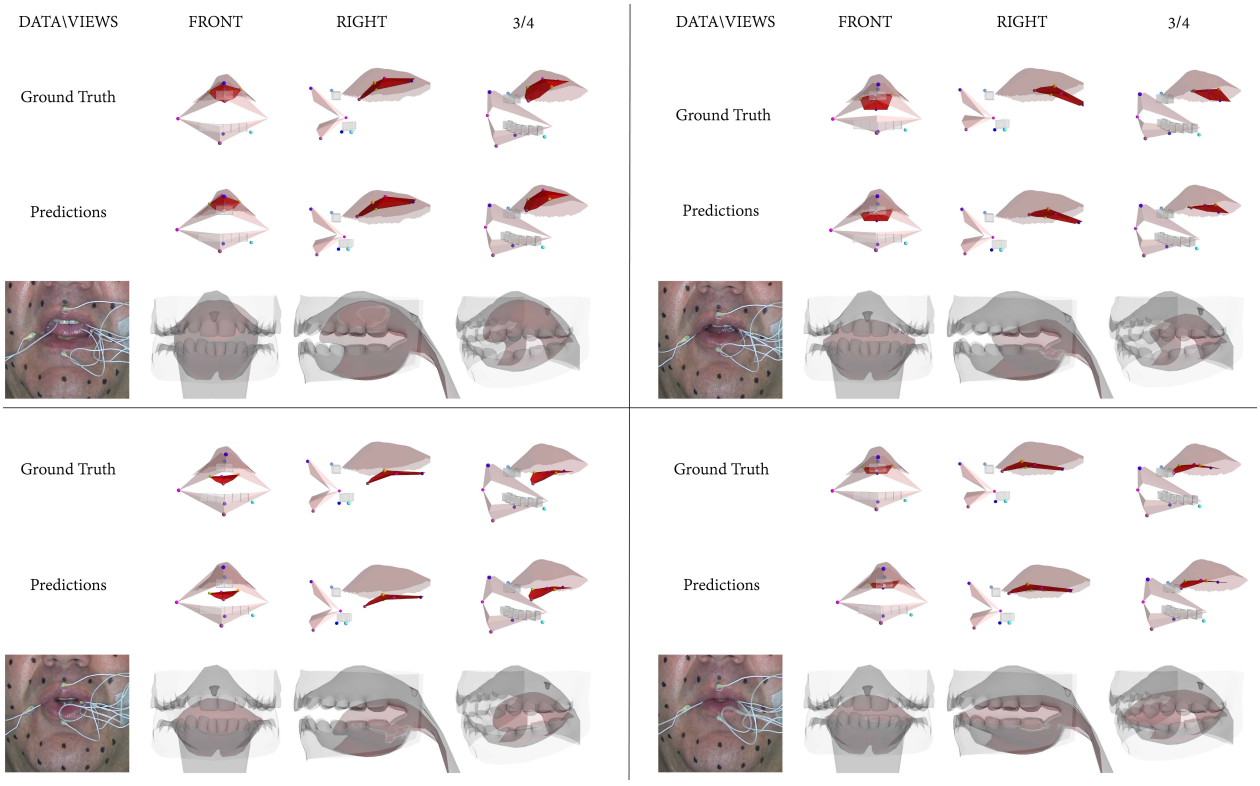


Figure 7. Visualization of a few samples from the training speaker on test speech samples not in the training set. For each camera view, both predicted landmark locations and the solved animation rig outputs are provided. As this is test data we can also show comparison to the ground truth EMA landmark locations.