Synthesis of Realistic Tongue Movements for Speech Animation Using Gated Recurrent Units

Jake Mwangi-Powell[1], Robert Kosk[2], Tianxiang Yang[2],
Cathair Kerrigan[2], Marco Volino[1]

[1]University of Surrey, [2]Humain Studios

1 Introduction

Speech animation has long focused on lip-sync and facial movements, neglecting the tongue's role in articulation, which is critical for achieving realistic character animations [2]. This project addressed this gap by proposing a machine learning (ML)-based approach to synthesize realistic tongue movements from speech audio inputs, resulting in more lifelike digital speech. Previous methods primarily focused on lips and facial movements. ML, particularly Recurrent Neural Networks (RNNs) and LSTM networks, have shown potential in speech animation [3]. However, these techniques have not been fully utilized for tongue synthesis. Transformers have shown success in natural language tasks but are computationally intensive. This project explored these models to find the most efficient and accurate architecture for tongue movement synthesis.

This paper presents a method for synthesizing realistic tongue movements from speech inputs using machine learning, specifically Gated Recurrent Units (GRUs). By leveraging a dataset of magnetic Articulography (EMA), various neural architectures such as Long Short-Term Memory (LSTM) networks, transformers, and GRUs were evaluated. After comprehensive hyper-parameter optimization, the GRU model achieved the best performance with a Mean Squared Error (MSE) loss of 1.473 cm. The proposed approach improves speech animation by dynamically generating tongue movements that enhance the realism of digital character animations.

2 Methodology

Pre-existing tongue position data that was captured using Electromagnetic Articulography (EMA), which tracks tongue and lip positions in 3D space [1] was utilized. This dataset, combined with synchronised speech audio, was pre-processed to ensure uniform length for batch processing. A custom forward kinematics (FK) rig of the tongue was created for the project, as shown in Figure 1, which allowed precise control over key parts of the tongue. Joints placement overlaps with positioning of EMA sensor coils, allowing for seamless transfer from sensors data onto tongue mesh animation. Several neural architectures were evaluated, including RNNs, LSTMs and transformers. The GRU model, with bidirectionality enabled, was selected for its simplicity, efficiency, and ability to capture sequential data dependencies. Hyperparameter tuning identified an optimal learning rate of 0.001, 256 hidden dimensions, 4 layers, and a dropout rate of 0.3. The Adam optimizer was used due to its superior performance over alternatives like RMSprop and SGD. Cross-validation was employed to ensure the model generalized well across unseen data, with training, validation, and test sets split to measure robustness.

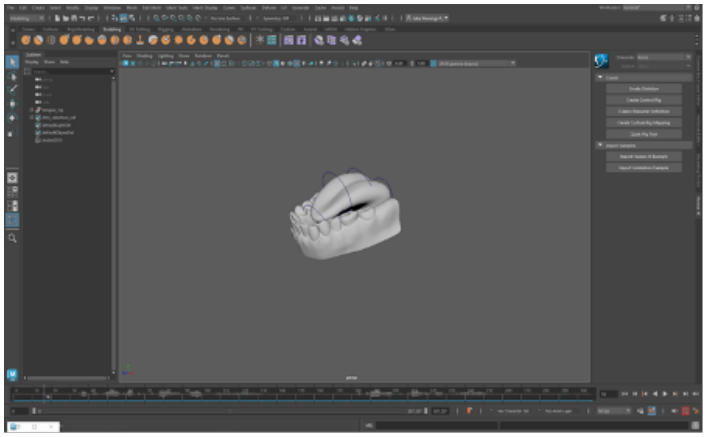

Figure 1: Tongue rig rendered in Autodesk Maya

3 Results

The GRU model achieved an MSE loss of 1.473 cm, outperforming LSTM and transformer models in both accuracy and computational efficiency which approximately achieved MSE values near 2.2. A batch size of 8 and early stopping with a patience value of 5 further optimized performance. Models without bidirectionality performed worse, demonstrating the importance of capturing context from both past and future frames in speech data. Evaluation on unseen data confirmed the model's ability to generalize well, with accurate predictions of tongue movements closely aligned with speech inputs, as shown in Figure 2. Objective evaluations using MSE, combined with subjective assessments of the generated animations, indicated a high level of realism in the synthesized tongue movements. Figure 2 shows that the model's predictions closely follow the actual values throughout most of the time steps, with minimal deviations. Despite some fluctuations, particularly in the earlier time steps, the overall alignment between predictions and actual values suggests the model captures the underlying pattern effectively.

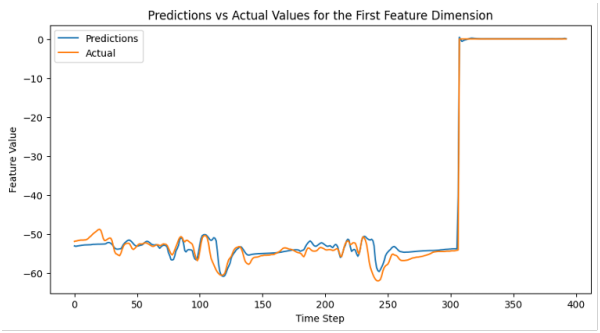

Figure 2: X-coordinate position for feature 1 at each frame

4 Conclusion

This work successfully developed a GRU-based approach for synthesizing realistic tongue movements from speech inputs. The proposed method enhances the realism of speech animations in digital characters, with applications in fields such as gaming, film, and VR. By demonstrating the effectiveness of ML in handling complex articulatory movements, this project opens new avenues for more immersive and lifelike digital animations. Future work could focus on expanding the dataset to include a wider range of speakers and speech variations. Integrating this tongue animation system with facial animation frameworks would create a more cohesive speech animation system. Additionally, optimizing the model for real-time performance would enable its use in interactive applications like virtual reality (VR) or live animation environments.

References

[1] Salvador Medina, Denis Tome, Carsten Stoll, Mark Tiede, Kevin Munhall, Alex Hauptmann, and Iain Matthews. Speech driven tongue animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, 2022.

[2] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov. Motion representations for articulated animation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13648–13657, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.01344.

[3] Pengcheng Zhu, Lei Xie, and Yunlin Chen. Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings. In *Interspeech 2015*, pages 2192–2196, 2015.