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

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

jakemwangipowell@gmail.com, {robert, cathair}@humain-studios.com, {y.tianxianq, m.volino}@surrey.ac.uk

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

Motivation

- 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].
- Machine learning has shown potential in speech animation [3]; however, these techniques have not been fully utilised for tongue synthesis.

Contribution

- This paper presents a method for synthesising realistic tongue movements from speech inputs using machine learning, specifically Gated Recurrent Units (GRUs).
- By leveraging a dataset of Electro-Magnetic Articulography (EMA) [1], various neural architectures, such as Long Short-Term Memory (LSTM) networks, transformers, and GRUs are evaluated.

Methodology

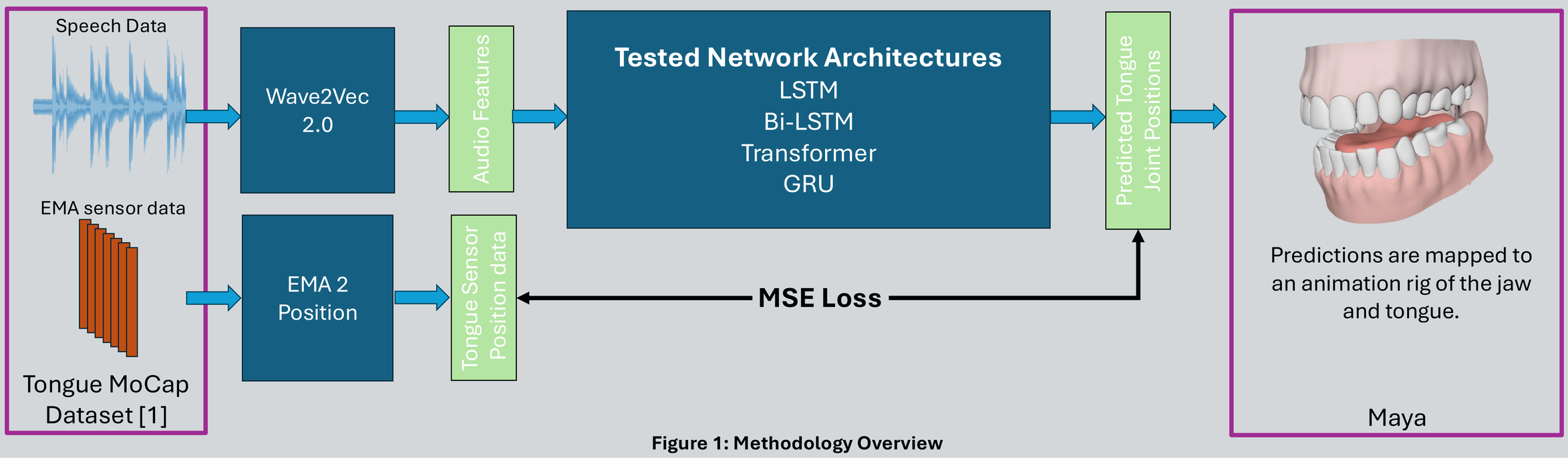


Figure 1: Methodology Overview

Results

Table 1: Comparing model architectures

Model Architecture	MSE
LSTM	2.267
Bi-LSTM	1.715
Transformer	2.216
GRU	**1.688**

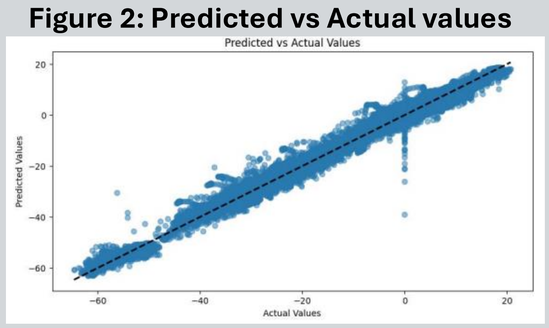
Table 2: GRU hidden dimension

Hidden Dimension	MSE
64	1.751
128	1.688
256	**1.598**
512	1.638

Table 3: GRU hidden dimension

Number of Layers	MSE
2	1.715
3	1.598
4	**1.559**
5	1.674

Figure 2: Predicted vs Actual values



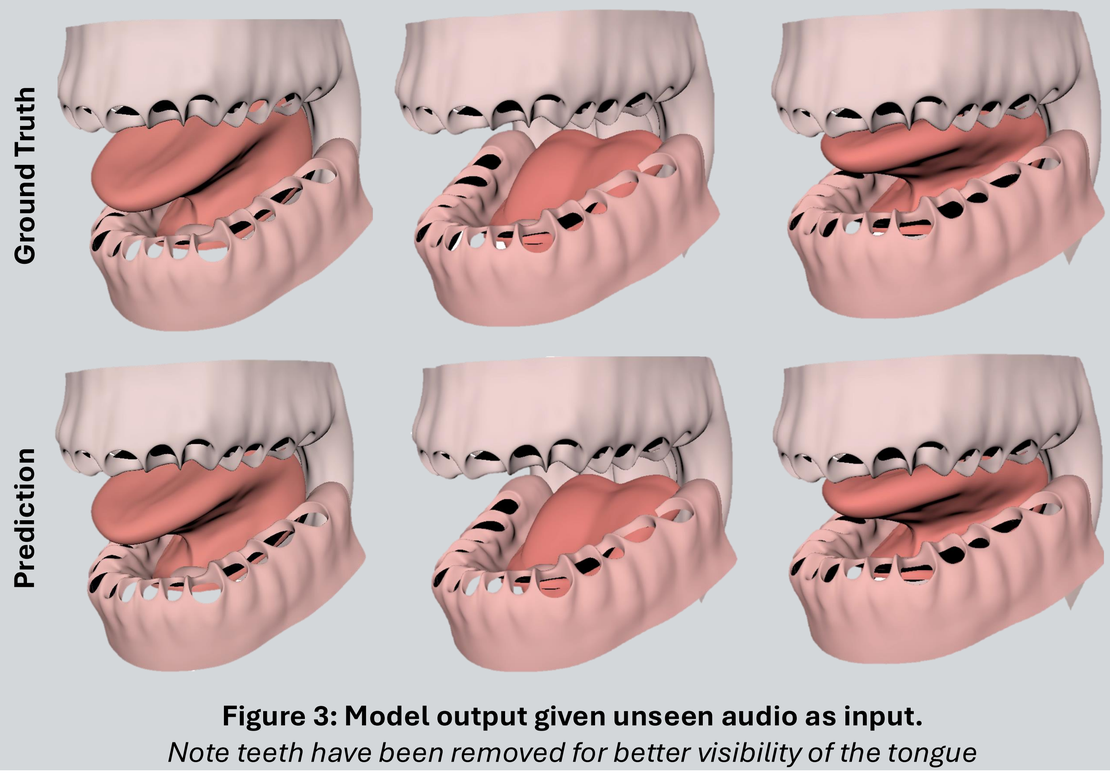


Figure 3: Model output given unseen audio as input.
Note teeth have been removed for better visibility of the tongue

Conclusions

- This work successfully developed a GRU-based approach for synthesising realistic tongue movements from speech inputs.
- The proposed method has the potential to enhance the realism of speech animations in digital characters, with applications in fields such as gaming, film, and VR.

References

1. S. Medina, D. Tome, C. Stoll, M. Tiede, K. Munhall, A. Hauptmann, and I. Matthews. Speech driven tongue animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 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).
3. P. Zhu, L. Xie, and Y. Chen. Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings. In Interspeech 2015.