PhISANet: PHONETICALLY INFORMED SPEECH ANIMATION NETWORK

Carnegie Mellon University

Salvador Medina Sarah L. Taylor Carsten Stoll Gareth Edwards
Alex Hauptmann Shinji Watanabe Iain Matthews

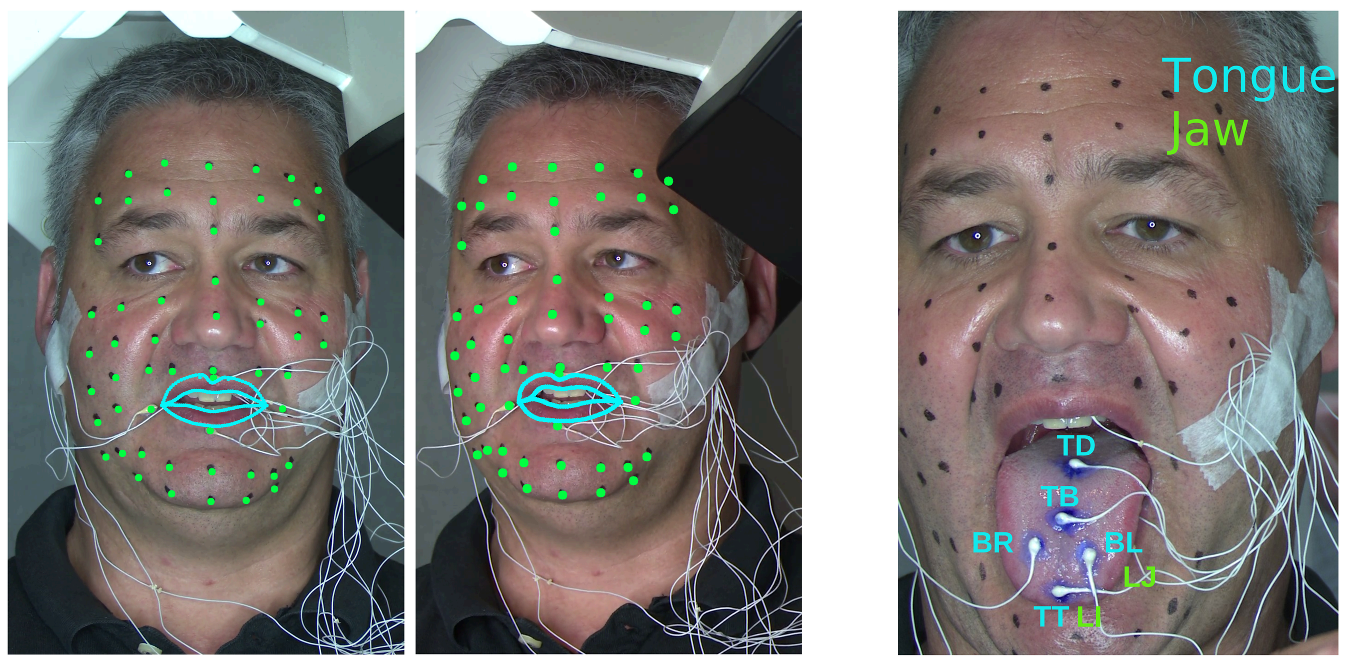
EPIC GAMES

We present **PhISANet**, an end-to-end model that fully **animates the lower face, including the jaw and tongue** from speech.
In this work we:

- Compare the use of **WavLM** [1], **Whisper** [2], and **Wav2Vec** [3] audio features for speech-to-animation.
- Improve articulation animation by regularizing the speech-animation model through **multi-task learning (MTL)** with a **Connectionist Temporal Classification (CTC)** [4] task.

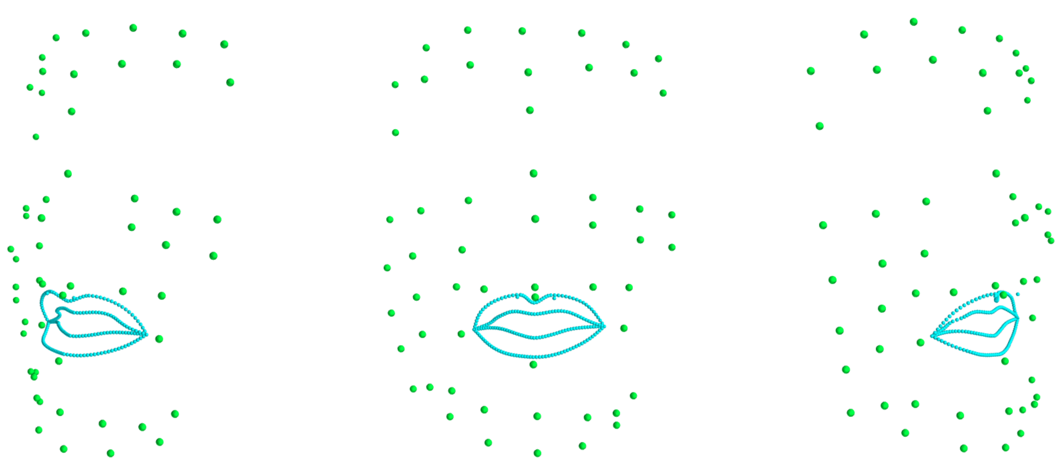
Data

Data Capture



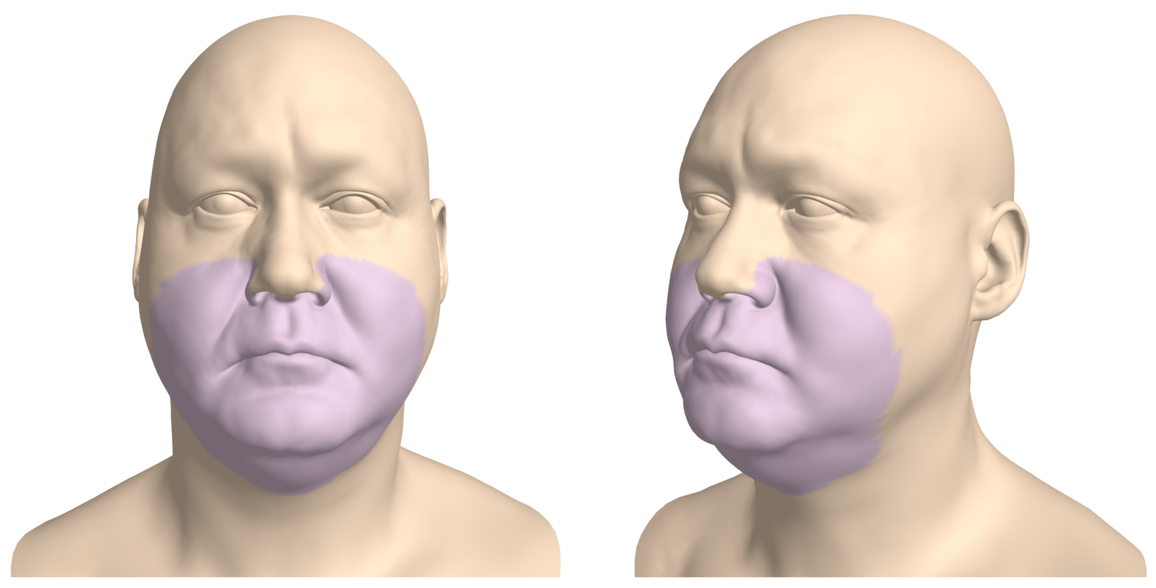
Stereo Video EMA

3D Reconstruction



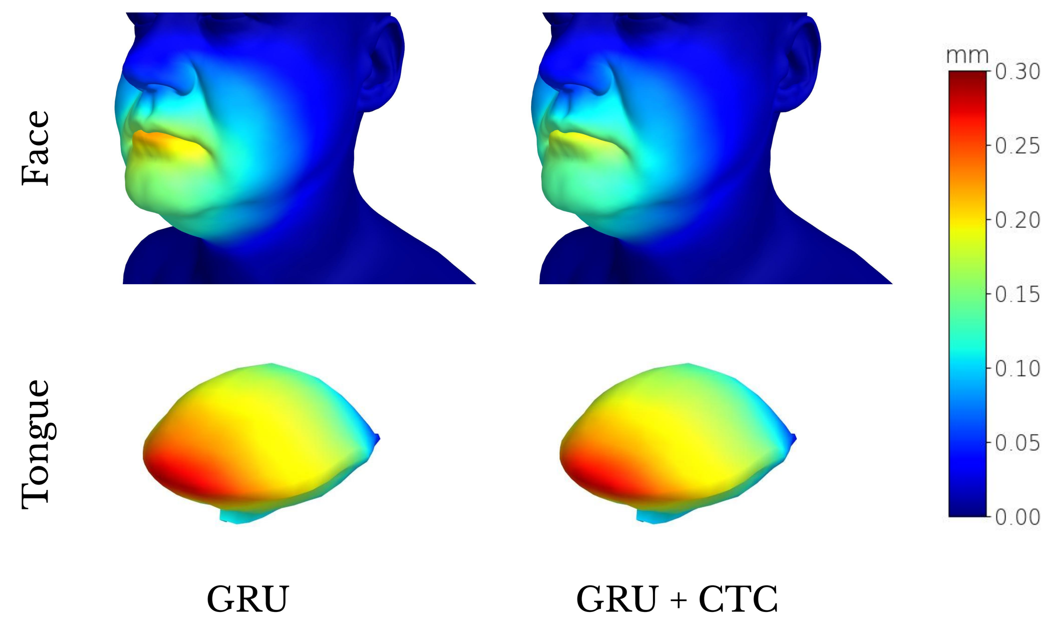
L-BFGS

Mesh Fitting



Evaluation

Quantitative Evaluation



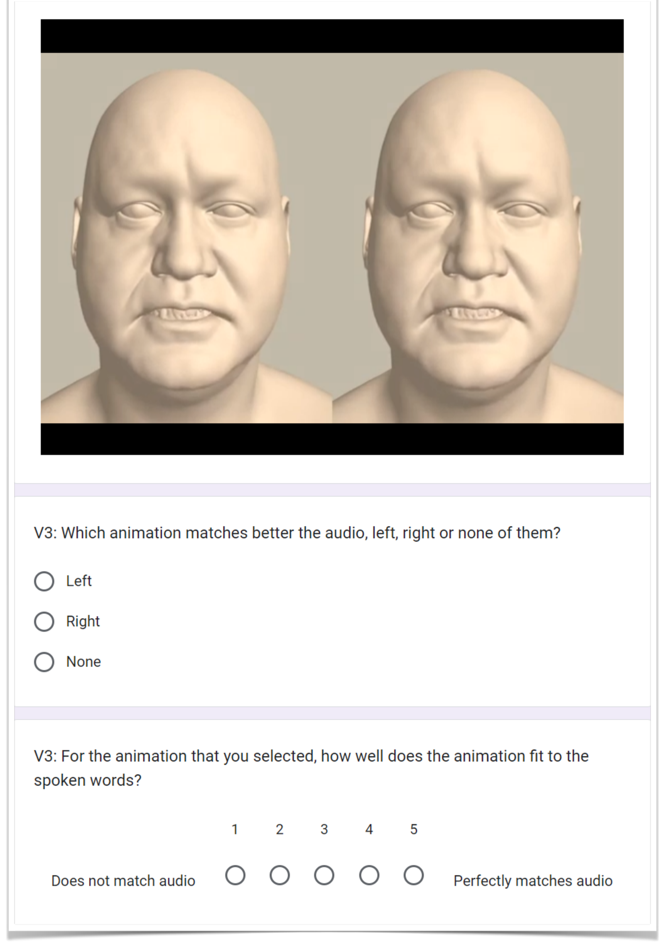
Mean Temporal Vertex Error

$$\mathrm{MTVE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\frac{1}{K}\sum_{k=1}^{K}(\hat{u}_{itk} - u_{itk})^2$$

Evaluation of different Encoder-Decoder models

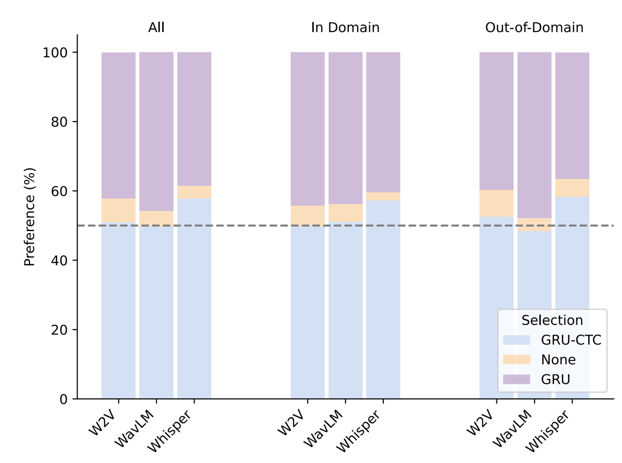
Encoder	Decoder	MTVE	Tongue MTVE	Lips MTVE
Wav2Vec	GRU	0.072 ± 0.014	0.180 ± 0.036	0.164 ± 0.036
	GRU+CTC	0.070 ± 0.014	0.173 ± 0.033	0.160 ± 0.037
	p-value	3.48×10^{-6}	9.505×10^{-18}	5.745×10^{-6}
WavLM	GRU	0.076 ± 0.013	0.166 ± 0.034	0.175 ± 0.035
	GRU+CTC	$\mathbf{0.068 \pm 0.014}$	$\mathbf{0.160 \pm 0.030}$	$\mathbf{0.157 \pm 0.035}$
	p-value	1.426×10^{-52}	9.729×10^{-18}	1.628×10^{-45}
Whisper	GRU	0.077 ± 0.015	0.190 ± 0.034	0.173 ± 0.034
	GRU+CTC	0.075 ± 0.015	0.172 ± 0.031	0.168 ± 0.036
	p-value	3.679×10^{-11}	1.159×10^{-57}	1.753×10^{-6}

User Studies

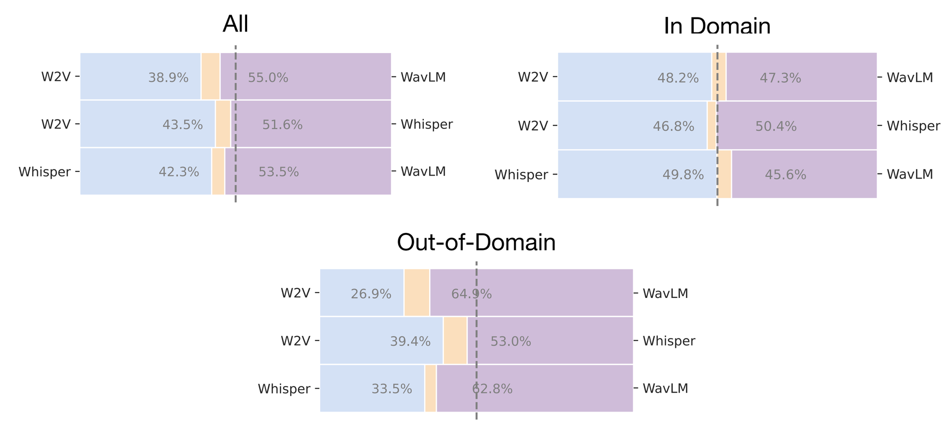


Pairwise Preference Test

Comparison of models for each feature vs. MTL CTC counterpart

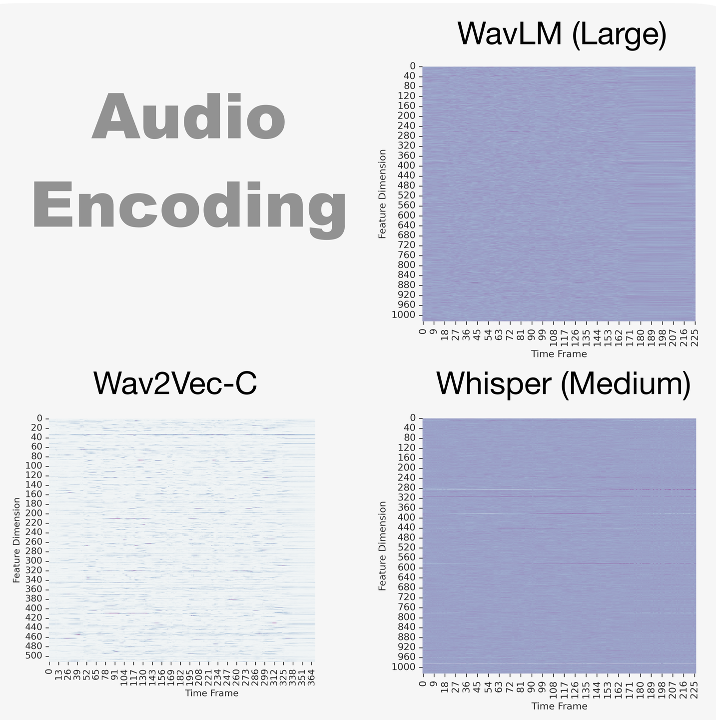


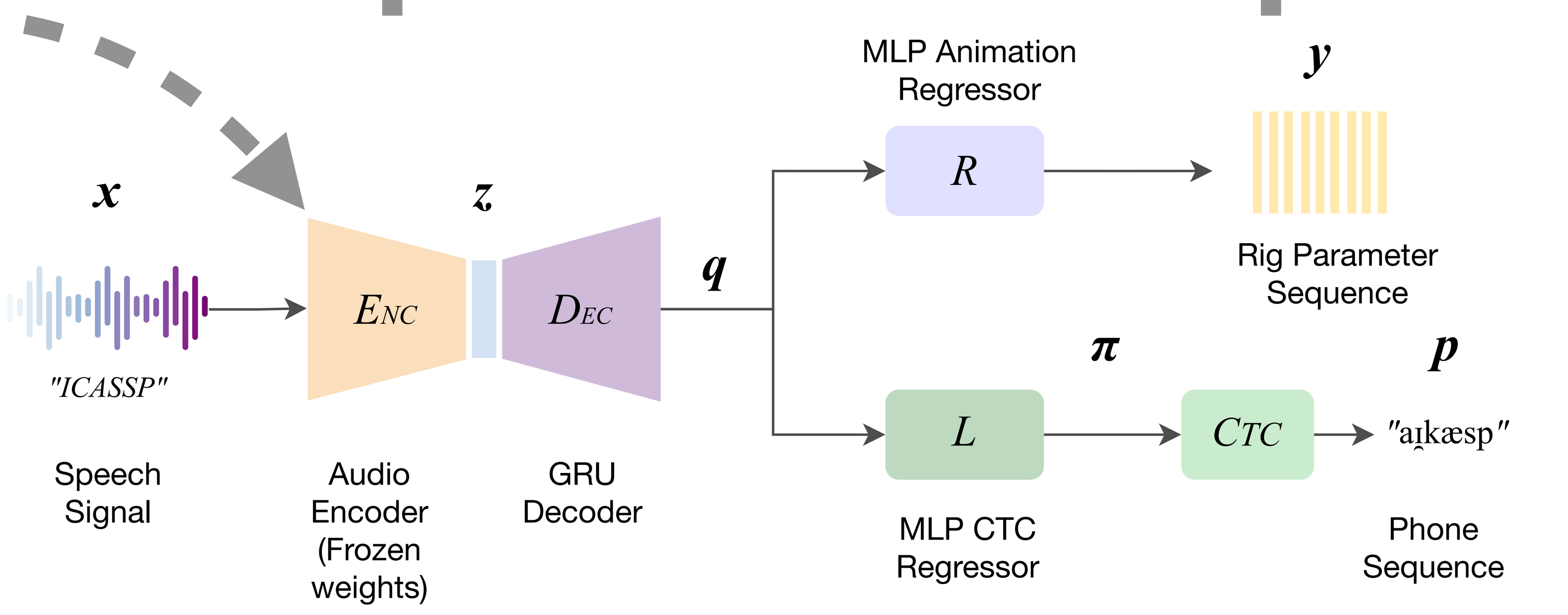
Comparison of MTL CTC models across features



Experimental Setup

Audio Encoding





Multi-Task Learning Framework

Loss

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{vel}}\mathcal{L}_{\mathrm{vel}} + \lambda_{\mathrm{accel}}\mathcal{L}_{\mathrm{accel}} + \lambda_{\mathrm{CTC}}\mathcal{L}_{\mathrm{CTC}}$$

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\mathbf{y}}_t - \mathbf{y}_t)^2$$

$$\mathcal{L}_{\mathrm{vel}} = \frac{1}{T-1}\sum_{t=1}^{T-1}(\hat{\mathbf{v}}_t - \mathbf{v}_t)^2$$

$$\mathcal{L}_{\mathrm{accel}} = \frac{1}{T-2}\sum_{t=1}^{T-2}(\hat{\mathbf{a}}_t - \mathbf{a}_t)^2$$

$$P(\boldsymbol{\pi}|\mathbf{q}) = \mathrm{Softmax}(L(\mathbf{q}))$$

$$P(\mathbf{p}|\mathbf{q}) = \sum_{\boldsymbol{\pi}\in\Psi(\mathbf{p})} P(\boldsymbol{\pi}|\mathbf{q})$$

$$\mathcal{L}_{\mathrm{CTC}} \stackrel{\mathrm{def}}{=} -\ln P(\mathbf{p}|\mathbf{q})$$

Conclusions

- PhISANet is an encoder-decoder model that generates realistic speech animation by training on data derived from high-quality 3D capture of facial and tongue movements.
- PhISANet delivers realistic and high-quality animations regardless of the gender, age, or language, leveraging on robust audio encodings.
- Incorporating Connectionist Temporal Classification multi-task learning, enhances the realism of the generated speech animations.
- State-of-the-art speech audio encoders, such as Wav2Vec, WavLM, or Whisper, can effectively drive plausible speech animation generation.
- Animations generated by the WavLM-based model were preferred by users due to their "*natural and lifelike motion*".

[1] S. Chen, C. Wang, Z. Chen, Y. Wu, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," IEEE Journal of Selected Topics in Signal Processing, vol. 16, pp. 1505–1518, 2021.
[2] A. Radford, J.W. Kim, T. Xu, et al., "Robust speech recognition via large-scale weak supervision," ArXiv preprint, vol. abs/2212.04356, 2022.
[3] S. Schneider, A. Baevski, R. Collobert, M. Auli, "Wav2Vec: Unsupervised pre-training for speech recognition," in INTERSPEECH, pp 1–9, 2019.
[4] A. Graves, S. Fernandez, F.J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in Proc. of ICML, vol. 148, pp. 369–376, 2006.