PHISANET: PHONETICALLY INFORMED SPEECH ANIMATION NETWORK

Salvador Medina[⋆†] *Sarah L. Taylor*[†] *Carsten Stoll*[†] *Gareth Edwards*[†]
Alex Hauptmann[⋆] *Shinji Watanabe*[⋆] *Iain Matthews*[†]

[⋆]Carnegie Mellon University, Pittsburgh, PA, USA
[†]Epic Games, Pittsburgh, PA, USA

ABSTRACT

Realistic animation is crucial for immersive and seamless human-avatar interactions as digital avatars become more prevalent. This work presents PhISANet, an encoder-decoder model that realistically animates the face and tongue solely from speech. PhISANet leverages neural audio representations trained on vast amounts of speech to map the speech signal into animation parameters that control the lower face and tongue of realistic 3D models. By integrating a novel multi-task learning strategy during the training phase, PhISANet reincorporates the phonetic information from the input speech, improving articulation in the generated animations. A thorough quantitative and qualitative study validates this improvement, and it determines that WavLM and Whisper features are ideal for training a generalizable speech-animation model regardless of gender, age, and language.

Index Terms— Speech Animation, Multi-task Learning, CTC, Tongue, EMA

1. INTRODUCTION

Animating realistic-looking digital avatars has become an important task due to its wide range of applications in entertainment, productivity, and healthcare, to mention a few. The filmmaking and video game industry highly benefits from lifelike 3D characters that speak naturally for a more immersive experience. The recent development of large language models and the outstanding results of automatic speech recognition, speech-to-text, and 3D graphics are fuelling the research in interactive agents. Hence, it is important to generate animations in an automated way that look as natural as possible to avoid any disruption to the user interaction. In the healthcare sector, a group of researchers has provided a voice to a patient with aphasia through a MetaHuman [1] 3D character driven by audio generated from estimating a sequence of phonemes from brain signals [2]. Generating a natural articulation animation of these characters could improve how patients with similar pathologies communicate.

In this paper, we introduce PhISANet, a solution that animates the face and tongue of a 3D character from speech audio input, allowing us to efficiently generate realistic speech animation. PhISANet is a PHonetically Informed Speech Animation encoder-decoder Network trained on high-quality articulation data. It uses a pre-trained audio encoder together with an animation decoder to generate animations that generalize across gender, age, and language. We additionally introduce a novel multi-task learning framework that uses a Connectionist Temporal Classification (CTC) [3] auxiliary task to provide a phone alignment constraint.

The main contributions of this work are as follows: (1) We present PhISANet, an end-to-end model that fully animates the lower face, including the jaw and tongue, on a frame-by-frame basis with temporal consistency solely from speech signal. (2) We introduce the use of WavLM and Whisper audio features into the speech-to-animation field of study. (3) We introduce the regularization of a speech-animation model through multi-task learning with a CTC task, which we demonstrate improves the generated animations through a quantitative and qualitative analysis.

2. RELATED WORK

Parke introduced the first parametric face model [4], which was animated by keyframing particular poses in accordance with the phonemes [5]. The clustering of phonemes based on their visual similarity was later termed ”visemes” [6]. Over the years, several studies have built upon this foundational work, focusing on animating 3D models using visemes via procedural methods [7]. However, one limitation of these rule-based methods is their demand for extensive manual intervention. This methodology requires meticulous rule design to yield desired outcomes, notably evident in the JALI framework [8].

To overcome these problems, the research community explored deep learning-based solutions like those introduced by Zhou *et al.* [9] with VisemeNet, which learns to produce animation curves from JALI-generated data. Taylor *et al.* [10] directly learns coarticulation motions from data based on an Active Appearance Model. Richard *et al.* [11] presented MeshTalk to learn a categorical latent space of facial animations from 4D scan data and disentangle audio correlated and non-correlated face motions. Fan *et al.* [12] introduced the Transformer architecture to speech animation as an autoregressive model through Faceformer. Furthermore, Xing *et al.* [13] proposed the CodeTalker model, which maps audio input to facial motions through a self-reconstruction method that consists of learning a codebook in tandem with a decoder to capture realistic facial motion priors. While these models represent significant advancements, a limitation is their constrained generalization due to the targeted face parametric model. In contrast, our solution is compatible with the MetaHuman rig model [1], enabling the seamless transfer of animations between different MetaHuman-based character models.

Deep learning has also benefited other research areas, such as automatic speech recognition (ASR). A common practice to improve the performance of ASR models is through multi-task learning (MTL) by adding an auxiliary task handled by a CTC. For instance, Kim *et al.* [14] introduced a joint CTC attention-based end-to-end ASR system that utilized multi-task learning for sequence labeling and output sequence prediction, thereby significantly enhancing system performance. Heba *et al.* [15] addressed character-level speech recognition through multi-task learning, employing Consonant-

Vowel (CV) recognition as an auxiliary task via CTC. Moreover, Chen *et al.* [16] improved multilingual ASR by incorporating hierarchical CTC objectives into an encoder-decoder model, postulating that language identification assists model convergence. Inspired by this work, we explored the introduction of a CTC into an end-to-end model designed to predict a sequence of animation parameters from a speech signal by incorporating the phonetic information into the training of the model through an MTL CTC, which aligns the corresponding phones of the speech signal.

In this work, we also examine the efficacy of pre-trained audio representation models in the development of generalizable speech-to-animation frameworks. Building upon the findings of Medina *et al.*[17], the investigation reveals that training a speech animation decoder on features derived from the Wav2Vec audio encoder yields robust generalized performance, even when exclusively trained on a single actor. Wav2Vec, a causal self-supervised convolutional network tailored for general speech audio representation, is compared against WavLM and Whisper audio encoders in terms of speech animation regressions. WavLM[18], designed within the principles of HuBERT [19] and Wav2Vec 2.0 [20], integrates masked speech prediction as a denoising component within a Transformer-based architecture. WavLM outperforms contemporary models in the SUPERB challenge [21]. Whisper [22], trained on an extensive dataset comprising 680,000 hours of undisclosed multilingual audio sources, shares an architectural resemblance with WavLM and employs an end-to-end encoder-decoder Transformer framework. Despite lacking specific fine-tuning for any particular dataset, Whisper demonstrates remarkable robustness in zero-shot scenarios across various tasks.

3. DATA

We build upon the IMT'22 dataset introduced by [17], which captures the tongue, lips, and jaw motion through an electromagnetic articulography (EMA) device from a single English actor. This dataset consists of synchronized speech at 16 kHz, transcripts, and 3D landmarks from 10 EMA sensors, three on the lips, two on the jaw, and five on the tongue. In this work, we add the rotation of the EMA sensors and track 51 facial dots along with 17 lip contour landmarks from the multi-view videos captured during the same session with an industrial-grade visual tracker [23]. The 3D facial landmarks are obtained using a stereo reconstruction of the 2D landmarks [24]. The landmarks are mapped into the 3D target head mesh space using a similarity transformation, and we subsequently fit the mesh to our data (EMA + landmarks) on a frame-by-frame basis, adopting the method from [17] using an L-BFGS optimizer [25]. Lip contours are integrated into the optimizer as 2D constraints, as the reconstructed depth was unreliable due to the difficulties in placing these landmarks consistently across views.

The target head mesh was created from a high-quality 3D capture of the actor and it is controlled by 173 rig parameters, as outlined by the Epic Games MetaHuman model [1]. Given our emphasis on producing realistic coarticulation animations, we narrowed our focus to 67 parameters controlling only the lower face and inner mouth. A significant benefit of the MetaHuman model is its ability to transfer predicted motions to other MetaHuman characters, making our solution well-suited for practical, industrial applications.

To train the phone CTC, we extract the allophones from the audio samples with their corresponding transcripts using the Montreal Forced Aligner [26]. Our dataset revealed a total of 88 allophones, including the silence token.

Our enhanced dataset encompasses 1700 samples, equivalent to 2.28 hours of audio, each paired with sequences of 67-D rig parameters at 50 FPS to train our model and their corresponding timestamped phone sequences.

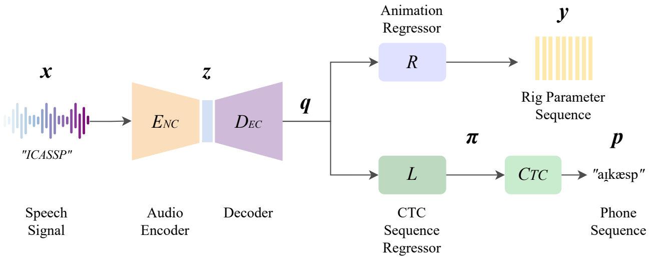

Fig. 1: PhISANet is an encoder-decoder model that is trained through multi-task learning with a phone CTC task to predict a sequence of rig parameters from an arbitrary speech signal.

4. MODEL

PhISANet is an encoder-decoder model that aims to generate 3D speech animations given a speech signal through a multi-task learning approach. Inspired by how ASR has improved by adopting a multi-task learning CTC [14, 15, 16] strategy while training, we seek to explore its effect on the speech animation task by introducing an auxiliary phone CTC task during training.

The PhISANet model architecture is shown in Figure 1. It considers an audio signal $\mathbf{x} = [x_1, ..., x_T]$ to be mapped into a latent space $\mathbf{z} = [z_1, ..., z_T]$ through a pre-trained audio encoder $\textsc{Enc}$, where T is the number of frames in the output sequence, x_t is the corresponding audio window for frame t and z_t is the corresponding audio embedding. Then $\mathbf{z}$ is mapped by a decoder $\textsc{Dec}$ into a shared space $\mathbf{q} = [q_1, ..., q_T]$ which serves as input to the animation regressor R to predict the sequence of rig parameters $\mathbf{y} = [y_1, ..., y_T]$, and also as input to the CTC Sequence Regressor L whose output serves a CTC to align the audio's corresponding phones $\mathbf{p} = [p_1, ..., p_S]$ with a sequence length $S \leq T$.

We trained the model with a weighted loss described in Eq. 1, which consists of a reconstruction term based on the MSE of the rig parameters $\mathcal{L}_{\text{rec}}$, a velocity term $\mathcal{L}_{\text{vel}} = \frac{1}{|T|}\sum_{t=1}^{T-1}(\hat{\mathbf{v}}_t - \mathbf{v}_t)^2$ of the rig parameters in contiguous timesteps $\mathbf{v}_t = \mathbf{y}_{t+1} - \mathbf{y}_t$ and an acceleration term $\mathcal{L}_{\text{accel}} = \frac{1}{|T|}\sum_{t=1}^{T-2}(\hat{\mathbf{a}}_t - \mathbf{a}_t)^2$ as the difference of velocities in contiguous timesteps $\mathbf{a}_t = \mathbf{v}_{t+1} - \mathbf{v}_t$. Each loss term is weighted through an independent $\lambda_{(.)}$ coefficient.

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{accel}}\mathcal{L}_{\text{accel}} + \lambda_{\text{CTC}}\mathcal{L}_{\text{CTC}} \tag{1}$$

To compute the multi-task CTC loss $\mathcal{L}_{\text{CTC}}$, the decoder features $\mathbf{q}$ are linearly mapped by L into phone probability representations $P(\boldsymbol{\pi}|\mathbf{q}) = \text{Softmax}(L(\mathbf{q}))$, where an alignment sequence π allows repetitions of the phones found in our dataset plus blank symbols $\{-\}$. Through training the CTC branch, we seek to maximize the probability distribution over all possible alignment sequences $\pi \in \Psi(\mathbf{p})$ as described in Eq. 2. The CTC loss is the negative log-likelihood of such a probability as described in Eq. 3.

$$P(\mathbf{p}|\mathbf{q}) = \sum_{\boldsymbol{\pi} \in \Psi(\mathbf{p})} P(\boldsymbol{\pi}|\mathbf{q}) \tag{2}$$

$$\mathcal{L}_{\text{CTC}} \overset{\text{def}}{=} -\ln P(\mathbf{p}|\mathbf{q}) \tag{3}$$

5. EXPERIMENTS AND RESULTS

We compare PhiSANet with an equivalent baseline model trained without the CTC task by removing the CTC term in Eq.1. We additionally seek to determine which pre-trained model between Wav2Vec, WavLM, and Whisper functions as the best audio encoder within PhiSANet. No fine-tuning is applied to the audio encoders.

Our model is trained to generate animation at 50 FPS, which matches the sampling rate of WavLM and Whisper. For the Wav2Vec *Large* model output, the features are subsampled from 100 FPS by concatenating the embeddings from two contiguous audio frames as instructed in [17]. We further select the WavLM *Large* and the Whisper *Medium* models as the number of parameters is similar between these models.

The decoder in all experiments is formed by a 5-layered bidirectional GRU [27] with 1,024 hidden units, whose output is linearly projected into the 67-D rig parameter space for all the baseline models. For the CTC models, we added to the baseline model a two-layer MLP as the CTC Sequence Regressor with a 512-D hidden size. The batch size is set to 32 and the learning rate to 5×10^{-6}, and the maximum number of epochs is limited to 100. The RAdam optimizer [28] is employed during training and an early stop strategy is applied at ten epochs. A grid search determined the best coefficients to be $\lambda_{\text{vel}} = 0.5$, $\lambda_{\text{accel}} = 0.7$, and $\lambda_{\text{CTC}} = 0.008$.

The data was split in an 80/20 fashion for the training and test sets. The training samples were subsampled from the original data through overlapping windows of 45 audio frames (900 ms) and a stride of 1 frame (20 ms). The output is subjected to min-max normalization, as the MetaHuman rig parameters ranges are either $[0, 1]$ and $[-1, 1]$. For the linear layers and MLPs within the architecture, a consistent dropout rate of 0.1 is applied.

The performance of the models is evaluated through the mean temporal vertex error (MTVE) (Eq.4) of the animated facial mesh. The MTVE is computed across all the 3D vertices of the lower face, including those in the inner mouth, and also over the isolated subregions of the lips and tongue to gain a deeper understanding of the model's behavior.

$$\text{MTVE} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\frac{1}{K}\sum_{k=1}^{K}\left(\hat{u}_{itk} - u_{itk}\right)^2, \qquad (4)$$

where u_{itk} represents the k^{th} ground truth 3D vertex of the j-th timestep in the i^{th} sample, while $\hat{u}_{itk}$ denotes the vertex corresponding to the prediction of the model being evaluated. Furthermore, N indicates the total number of samples in the test set. T_i is the number of frames of the i^{th} sample, and K corresponds to the number of vertices to be evaluated.

5.1. Quantitative Results

Our findings are summarized in Table 1. We observe that utilizing multi-task learning CTC for phone alignment significantly reduces the MTVE in the generated animations across all audio encoders, maintaining the integrity of the lips and tongue regions. To validate these outcomes, we conducted a pairwise t-test of the MTVE per test sample, comparing the baselines of each audio encoder with their CTC variant, yielding a significant $p < 10^{-5}$. Notably, the model with the best performance employed the WavLM for audio encoding and was trained using a phone CTC. Figure 2 shows a vertex-level visualization of MTVE for models using WavLM audio representations, highlighting the benefits of integrating phonetic data into the network. Improvements are evident on the lower lip and its surrounding area, and the region surrounding the tongue tip.

Table 1: Effect of different encoder-decoder combinations on Lower Face, Tongue, and Lip MTVE. The WavLM encoder with the GRU+CTC decoder yielded the lowest MTVE.

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}

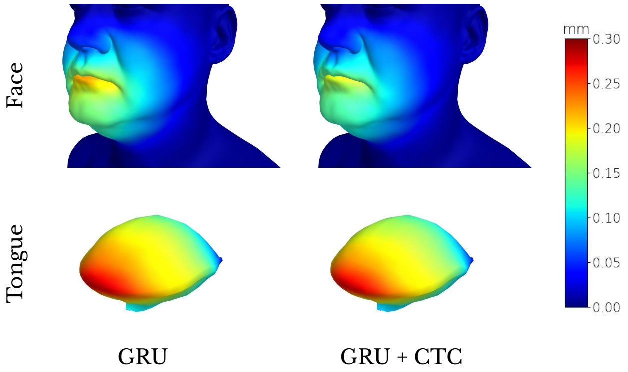

Fig. 2: Visualization of the MTVE for the WavLM-based model showing how incorporating a CTC multi-task learning strategy to align phones reduces the overall error of the face, lips, and tongue.

5.2. Qualitative Results

Visual representations of the study's findings are accessible via the project's website[1]. Upon examining the animations produced by the proposed models, it was observed that the tongue's behavior aligns with theoretical expectations: during the articulation of open vowels, it is positioned on the floor of the mouth, and for dental consonants, it is located between the anterior teeth. A comparative analysis of the leading models across different audio encodings indicated that the animations derived from in-domain samples maintained a consistent level of plausibility. Notably, the model incorporating WavLM encoding with an MTL phone CTC demonstrated superior performance, as evidenced in a *silence* audio sample. Moreover, the methodology displayed generalization capabilities as it could animate female voices and articulations in languages such as German, Spanish, Japanese, and Mandarin despite the absence of these voices in the training data.

6. USER STUDIES

Since the ultimate goal of our work is to create animations that are perceived as realistic by humans, evaluating the model through a human perception study is an essential step. We conducted three user studies to evaluate our proposed models' effectiveness. Firstly, we compared the animations generated by the highest-performing model per audio encoder against its ground truth counterpart. Then,

[1] https://github.com/salmedina/PhISANet

Table 2: User study results from comparing ground-truth vs. MTL CTC models with different audio features.

Comparison	Audio Encoder	Preference	Rating
1	**Ground truth**	**55.6**	**4.03±0.78**
	None	6.7	2.58±1.38
	Wav2Vec	37.8	3.97±0.75
2	**Ground truth**	**58.9**	**3.99±0.77**
	None	5	3.00±1.15
	WavLM	36.1	3.95±0.69
3	**Ground truth**	**52.8**	**4.04±0.85**
	None	5.6	3.10±1.45
	Whisper	41.6	3.91±0.81

we compared the baseline model for each audio encoder against its multi-task learning version with a CTC. Lastly, we identified the highest-performing model per audio encoder and compared them against each other to ascertain which was more favorably received by users. As our ground-truth only has English speech, we limited the participation to native English speakers residing in English-speaking countries. This strategic selection of participants ensured a more informed and precise evaluation of our animations.

The user studies were conducted as pairwise preference tests, where participants were presented with two videos from a sample from all possible combinations required by the study. Participants were instructed to select the preferred video or indicate *neither* or *both*. They were also asked to rate the congruence of the selected animation with the audio on a 5-star scale ranging from *"Does not match the audio"* to *"Perfectly matches the audio"*. We presented the users with 10 videos per study. The first study showed animations from 10 *in-domain* audio samples from our test data. The second and third studies used 5 *in domain* and 5 *out-of-domain* samples, selected from outside sources varying the intonation and range of motion.

The design of our user studies drew inspiration from those conducted for the GENEA challenge [29], with the specific aim of obtaining results that are statistically robust to enable a comprehensive evaluation of our findings. The first three videos shown to participants were to acclimate participants to the evaluation process, and responses were discarded. To avoid potential bias, the sequence order and left-right positioning of the videos was randomized. Furthermore, three control videos featuring non-matching audio were added to ensure participants were attentive throughout each video. They were asked to evaluate such videos by selecting the *"None"* option and set the rating to one star.

6.1. User Study Results

Table 2 summarizes our first user study comparing CTC-based models with the ground truth. The ground-truth animations predominantly received a rating of 4.0. Predicted animations from all audio encoders achieved ratings close to ground truth, indicating that all were deemed to match the audio for this *in domain* test data. Tables 3 and 4 further support this, showing that our CTC models align closely with the ground truth, as seen in the *in domain* section of the results when evaluating the models with our test data, where the models received ratings near 4.0. The Whisper-based animations were most commonly favored over the ground truth animations with a preference rate of 41.6% over the preference of 37.8% for Wav2Vec and 36.1% for WavLM.

Table 3: User study results from comparing baseline models for each feature vs. their MTL CTC counterparts.

		All		**In Domain**		**Out of Domain**	
Audio Encoder	**Decoder**	**(%) Preference**	**Rating**	**(%) Preference**	**Rating**	**(%) Preference**	**Rating**
W2V	GRU	42.1	3.76±0.89	44.2	3.88±0.90	39.7	**3.61±0.85**
	None	6.8	1.83±0.64	6.1	1.82±0.58	7.7	1.83±0.69
	GRU+CTC	**51.0**	**3.80±0.89**	**49.7**	**4.02±0.83**	**52.6**	3.56±0.89
WavLM	GRU	45.8	3.88±0.88	43.8	3.88±0.87	47.8	3.89±0.89
	None	4.5	1.50±0.50	5.1	1.56±0.50	3.8	1.43±0.50
	GRU+CTC	**49.7**	**4.07±0.83**	**51.1**	**4.14±0.79**	**48.4**	**4.00±0.86**
Whisper	GRU	38.5	**3.81±0.96**	40.4	**4.01±0.79**	36.5	3.56±1.08
	None	3.7	1.92±0.49	2.3	1.75±0.43	5.1	2.00±0.50
	GRU+CTC	**57.8**	3.78±0.86	**57.3**	3.98±0.71	**58.3**	**3.57±0.95**

Table 4: User study results from comparing MTL CTC models across different audio features.

		All		**In Domain**		**Out of Domain**	
Comparison	**Audio Encoder**	**(%) Preference**	**Rating**	**(%) Preference**	**Rating**	**(%) Preference**	**Rating**
1	Wav2Vec	38.9	3.78±0.82	**48.2**	3.89±0.80	26.9	3.54±0.80
	None	6.1	1.67±0.69	4.5	1.70±0.64	8.2	1.64±0.72
	WavLM	**55.0**	**3.84±0.87**	47.3	**3.89±0.84**	**64.9**	**3.80±0.90**
2	Wav2Vec	43.5	**3.94±0.86**	46.8	**4.17±0.73**	39.4	**3.60±0.92**
	None	4.9	1.79±0.61	2.8	2.00±0.58	7.6	1.69±0.60
	Whisper	**51.6**	3.81±0.84	**50.5**	4.02±0.77	**52.9**	3.56±0.85
3	**WavLM**	**53.6**	**3.90±0.75**	45.7	**3.95±0.82**	**62.8**	**3.86±0.69**
	None	4.2	1.65±0.59	4.6	1.60±0.49	3.7	1.71±0.70
	Whisper	42.3	3.85±0.76	**49.8**	3.85±0.81	33.5	3.84±0.65

Our second study highlights the benefits of using a phone CTC (*GRU+CTC*) to regularize the animation decoder. As Table 3 displays, users consistently favored these animations over baseline models without an auxiliary CTC. Specifically, 57.8% favored the improved quality of the Whisper animations using the CTC, compared to 38.5% for the standard model. This preference held across various audio encodings and both *in domain* and *out of domain* samples, aligning with the quantitative findings in Table 1.

In the final study, we examined preferences between CTC-based models with different audio encodings. While WavLM and Whisper were generally preferred over Wav2Vec, the difference became clearer for out-of-domain audio, where WavLM led by 12.9% against Wav2Vec and 21.8% over Whisper. These results suggest that PhISANet trained with the WavLM audio encoder generalizes better across speakers.

7. CONCLUSIONS

In this work, we introduced PhISANet, an encoder-decoder model for realistic speech animation trained on data that adds 3D face elements to the IMT'22 dataset, which consists of EMA capture of the inner mouth's speech motion to produce realistic articulation animations. PhISANet introduces to the field of speech animation a phonetic constraint through a multi-task learning CTC training strategy to improve the quality of the generated articulation animations. A quantitative evaluation using the mean temporal vertex error and a comprehensive user study confirm our findings.

We additionally explored three pre-trained audio encoders for the speech-animation task: Wav2Vec, WavLM, and Whisper. Our analysis showed that by combining any explored audio encoder model with our data can effectively reconstruct the original animation data. However, PhISANet performed best when paired with WavLM encodings, generalizing across diverse audio samples from different genders, ages, and languages, even when the animation decoder is trained solely on a single actor's voice.

8. REFERENCES

[1] Epic Games, "Metahuman creator," 2023, Accessed: September 6, 2023.

[2] Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, et al., "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, pp. 1–10, 2023.

[3] Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. of ICML*, William W. Cohen and Andrew W. Moore, Eds. 2006, vol. 148 of *ACM International Conference Proceeding Series*, pp. 369–376, ACM.

[4] Frederic I Parke, "A model for human faces that allows speech synchronized animation," *Computers & Graphics*, vol. 1, no. 1, pp. 3–4, 1975.

[5] Andrew Pearce, Brian Wyvill, Geoff Wyvill, and David Hill, "Speech and expression: A computer solution to face animation," in *Graphics Interface*, 1986, vol. 86, pp. 136–140.

[6] Cletus G Fisher, "Confusions among visually perceived consonants," *Journal of speech and hearing research*, vol. 11, no. 4, pp. 796–804, 1968.

[7] Michael M Cohen and Dominic W Massaro, "Synthesis of visible speech," *Behavior Research Methods, Instruments, & Computers*, vol. 22, no. 2, pp. 260–263, 1990.

[8] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh, "Jali: an animator-centric viseme model for expressive lip synchronization," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–11, 2016.

[9] Yang Zhou, Zhan Xu, Chris Landreth, et al., "Visemenet: Audio-driven animator-centric speech animation," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–10, 2018.

[10] Sarah L. Taylor, Taehwan Kim, Yisong Yue, et al., "A deep learning approach for generalized speech animation," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017.

[11] Alexander Richard, Michael Zollhöfer, Yandong Wen, et al., "Meshtalk: 3d face animation from speech using cross-modality disentanglement," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. 2021, pp. 1153–1162, IEEE.

[12] Yingruo Fan, Zhaojiang Lin, Jun Saito, et al., "Faceformer: Speech-driven 3d facial animation with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. 2022, pp. 18749–18758, IEEE.

[13] Jinbo Xing, Menghan Xia, Yuechen Zhang, et al., "Codetalker: Speech-driven 3d facial animation with discrete motion prior," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12780–12790.

[14] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. 2017, pp. 4835–4839, IEEE.

[15] Abdelwahab Heba, Thomas Pellegrini, Jean-Pierre Lorré, and Régine André-Obrecht, "Char+cv-ctc: Combining graphemes and consonant/vowel units for ctc-based ASR using multitask learning," in *Proc. of INTERSPEECH*, Gernot Kubin and Zdravko Kacic, Eds. 2019, pp. 1611–1615, ISCA.

[16] William Chen, Brian Yan, Jiatong Shi, et al., "Improving massively multilingual asr with auxiliary ctc objectives," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[17] Salvador Medina, Denis Tome, Carsten Stoll, et al., "Speech driven tongue animation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20406–20416.

[18] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu 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.

[19] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, et al., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[20] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[21] Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, et al., "SUPERB: speech processing universal performance benchmark," in *Proc. of INTERSPEECH*, Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek, Eds. 2021, pp. 1194–1198, ISCA.

[22] Alec Radford, Jong Wook Kim, Tao Xu, et al., "Robust speech recognition via large-scale weak supervision," *ArXiv preprint*, vol. abs/2212.04356, 2022.

[23] Cubic Motion, "Animation services," 2023, Accessed: September 5, 2023.

[24] Robert Shapiro, "Direct linear transformation method for three-dimensional cinematography," *Research Quarterly. American Alliance for Health, Physical Education and Recreation*, vol. 49, no. 2, pp. 197–205, 1978.

[25] Jorge Nocedal and Stephen J Wright, *Numerical optimization*, Springer, 1999.

[26] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, et al., "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Proc. of INTERSPEECH*, Francisco Lacerda, Ed. 2017, pp. 498–502, ISCA.

[27] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. of EMNLP*, Doha, Qatar, 2014, pp. 1724–1734, Association for Computational Linguistics.

[28] Liyuan Liu, Haoming Jiang, Pengcheng He, et al., "On the variance of the adaptive learning rate and beyond," in *Proc. of ICLR*. 2020, OpenReview.net.

[29] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, et al., "The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation," in *ACM International Conference on Multimodal Interaction*, 2022, p. 736–747.