Speech2UnifiedExpressions: Synchronous Synthesis of Co-Speech Affective Face and Body Expressions from Affordable Inputs –Appendix–

Uttaran Bhattacharya*
Adobe Inc.
San Jose, CA, USA
ubhattac@adobe.com

Aniket Bera
Purdue University
West Lafayette, IN, USA
aniketbera@purdue.edu

Dinesh Manocha
University of Maryland
College Park, MD, USA
dmanocha@umd.edu

A. Face and Pose Preprocessing from Video

Given a video, we use Multi-Task Cascaded CNNs [7] to extract the 3D face landmarks. Since the faces can be arbitrarily oriented w.r.t. the camera, we rigidly transform the face landmarks per frame to a reference frame in the normalized view, where the face looks towards the camera. For each frame in the input video, we use the rotation and the translation given by the Umeyama method [5] to map the face landmarks in that frame to the face landmarks in the reference frame. We also use similarly view-normalized 3D poses. View normalization is helpful for two key reasons. First, it eliminates relative camera movements across the video frames and prevents a learning-based method from confusing camera movements with face and pose expression changes. Second, a frontal view offers maximal visibility of the faces and the poses, and minimizes errors in detecting the 3D face landmarks and body joints.

B. Phoneme Predictor

We train a separate network to learn the positions of the lip landmarks for the different phonemes in the audio. Our synthesis network separately learns the motions of the lip corners denoting the different facial expressions, and we superpose them to the phoneme-based lip shapes to complete the lip motions. Our phoneme predictor predicts the 3D positions of all the landmarks on the inner and the boundaries of the lips over all the T prediction time steps, which we denote as $p_{1:T}\ in \mathbb{R}^{T \times L_{\text{lip}} \times 3}$. Following prior approaches [4], we design a CNN backbone connected to fully connected blocks to predict the lip landmarks from the spectrograms of the speech inputs. Specifically, given the speech audio waveform a, we compute

$$p_{1:T} = \text{PhonemePred}\left(a; \theta_{\text{PhonemePred}}\right), \tag{B.1}$$

where $\theta_{\text{PhonemePred}}$ represents the trainable parameters.

*Work partially done while Uttaran was a Ph.D. student at UMD

C. Training Details

We train our phoneme predictor network using reconstruction losses for the lip shapes. We train our synthesis network using a combination of reconstruction losses for the face and the pose motions, the cross-speaker diversity loss to enforce visual differences in expressions across speakers, and the generative adversarial loss for added regularization. We describe these loss functions and our training and testing procedures.

C.1. Phoneme Predictor Losses

We represent our phoneme predictor loss as the robust ℓ_1-norm reconstruction loss between the ground truth and the synthesized lip landmark positions and velocities over the prediction time steps T as

$$\mathcal{L}_{\text{ph}} = \sum_{t=1}^{T} \left\| p_t^{(\text{GT})} - p_t^{(\text{sn})} \right\|_1 + \left\| \Delta_t p_t^{(\text{GT})} - \Delta_t p_t^{(\text{sn})} \right\|_1, \tag{C.1}$$

where the superscripts (GT) and (sn), respectively, denote the ground-truth and the synthesized data. Δ_t denotes the discrete forward difference between adjacent time steps t and $t-1$.

C.2. Synchronous Synthesis Network Losses

We use reconstruction losses to robustly align the outputs of our generator with the corresponding ground-truth face and pose motions. We use the generative adversarial loss to ensure that the synthesized motions are plausible, the affective expressions match the corresponding ground truths, and prevent the mode collapse of only synthesizing singular expressions.

C.2.1 Reconstruction Losses

We write our reconstruction losses as the ℓ_1-norm difference between the ground truth and the synthesized face and

pose positions and motions over the T prediction time steps as

$$\mathcal{L}_{\text{Rec}} = \sum_{t=1}^{T} \left\| \mathcal{F}_t^{(\text{GT})} - \mathcal{F}_t^{(\text{sn})} \right\|_1 + \left\| \mathcal{P}_t^{(\text{GT})} - \mathcal{P}_t^{(\text{sn})} \right\|_1 + \lambda_{\text{vel}} \left(\left\| f_t^{(\text{GT})} - f_t^{(\text{sn})} \right\|_1 + \left\| u_t^{(\text{GT})} - u_t^{(\text{sn})} \right\|_1 \right) + \lambda_{\text{acc}} \left(\left\| \Delta_t f_t^{(\text{GT})} - \Delta_t f_t^{(\text{sn})} \right\|_1 + \left\| \Delta_t u_t^{(\text{GT})} - \Delta_t u_t^{(\text{sn})} \right\|_1 \right), \tag{C.2}$$

where λ_{vel} and λ_{acc} are the relative weighting factors. We use the velocity and acceleration losses to enforce smoothness in the synthesized motions by reducing jitters.

C.2.2 Cross-Speaker Diversity Loss

Our cross-speaker diversity loss $\mathcal{L}_{\text{CSD}}$ follows that of Yoon et al. [6], consisting of a ranking loss between the ground-truth face and pose motions, and the synthesized face and pose motions using the same speaker as the ground-truth (positive example) and a randomly chosen different speaker (negative example).

C.2.3 Generative Adversarial Loss

The generative adversarial loss consists of opposing losses $\mathcal{L}_{\text{Gen}}$ for the generator and $\mathcal{L}_{\text{Dis}}$ for the discriminator, following a min-max optimization strategy [2]. We write these losses as

$$\mathcal{L}_{Gen} = -\mathbb{E}\left[\log\left(c_{\text{disc}}^{\text{GT}}\right)\right], \tag{C.3}$$

$$\mathcal{L}_{Dis} = -\mathbb{E}\left[\log\left(c_{\text{disc}}^{\text{GT}}\right)\right] - \mathbb{E}\left[\log\left(1 - c_{\text{disc}}^{\text{sn}}\right)\right], \tag{C.4}$$

where c_{disc} denotes the output of our discriminator network (Eq. 15). This loss adds plausibility to our synthesized samples by forcing them to have affective expressions similar to those of the corresponding ground-truth samples.

C.3. Training Procedure

We train our phoneme predictor network using the Adam optimizer [3] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, a batch size of 1024, and a learning rate of 10^{-3} for 500 epochs. We train our synthesis network using the Adam optimizer [3] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, a batch size of 256, and learning rates of 10^{-4} for our generator and 5×10^{-5} for our discriminator, both decayed by a factor of 0.999 per epoch, for 1000 epochs. We train both our phoneme detector network and our synthesis network on an NVIDIA GeForce RTX 2080 Ti GPU, which takes 3 seconds and 7 seconds per epoch, respectively.

D. Testing and Rendering

We provide the details of the testing procedure of our network and the rendering of our synthesized outputs in a 3D environment.

D.1. Testing Procedure and Mapping to Digital Characters

Each test sample for our network consists of a speech audio waveform, the corresponding text transcript, a speaker ID, and the speaker's seed face and pose motions. Our phoneme predictor network provides the lip sync for the given speech audio, and the generator of our synthesis network provides the required face and pose motions. We superpose the lip landmarks given by our phoneme predictor network with the lip corner landmarks given by our generator at each prediction time step to obtain the complete lip motions of the speaker. We map these motions to a rigged 3D human upper-body mesh in Blender. For mapping the face motions, we set a one-to-one mapping between our face landmarks and the landmarks on the face of the human mesh, and use them as control points for the facial motions of the mesh. For mapping the pose motions, we use FABRIK [1] to obtain the joint rotations given our predicted joint positions and use those rotations to animate the rigged human mesh.

D.2. Rendering and Visualization

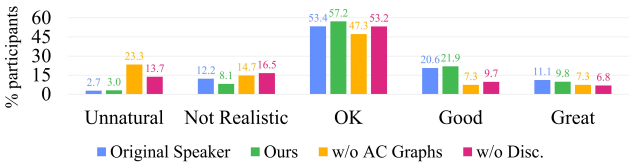
Given an input speech audio, we can synthesize the motions for our pre-rigged digital characters at an interactive rate of about 250 frames per second on an NVIDIA GeForce RTX 2080 Ti GPU. We design our digital environment using Blender. For each of our digital characters, we place them on a stage and position the camera such that it looks front and center at the agent. As the character narrates the input speech audio using our synthesized face and upper-body expressions, we slowly pan the camera in to get a more focused view of those expressions. Since we do not synthesize any lower-body motions, our digital characters stay standing at their initial positions during the entire narration. The full video demos are available with our supplementary material.

E. User Study

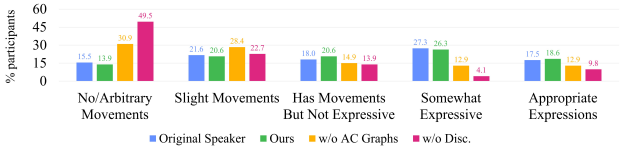
We provide all the details of our user study, including setup, evaluation process, and results.

E.1. Setup

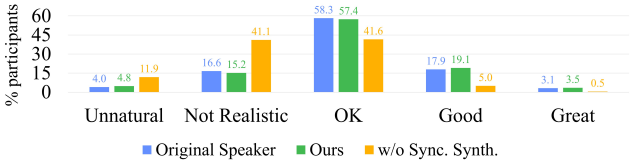
A total of 90 participants participated in our user study. All participants were aged 18 years or older, and had normal or corrected to normal vision and hearing. Each participant observed two sets of character motions. There were eight groups of motions in each set, and each group had a unique speech input. In the first set, there were four types of motions in each group corresponding to the same speech: the original speaker motions rendered using their face landmarks and the poses extracted from the video, and motions rendered using the face landmarks and poses synthesized by our network and two of its ablated versions. One ab-



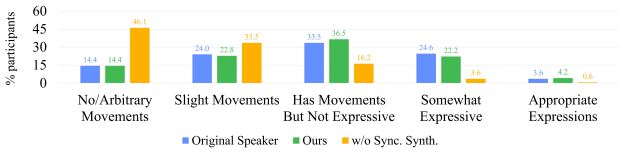
(a) **Set 1: Motion plausibility.** Compared to the ablated versions, we observe a higher distribution of "OK" or better for the motions of the original speakers and our synthesized agents. The modes of all the distributions are on "OK", implying that the corresponding participants found the visual qualities of all the motions to be reasonable.



(b) **Set 1: Synchronization between the face and the pose expressions given the speech.** Compared to the ablated versions, we obverse clear preferences for the motions of the original speakers and our synthesized agents. The modes of the distributions for these two types of motions are on "somewhat expressive" while the modes of the two ablated versions are on "no/arbitrary movements".



(c) **Set 2: Motion plausibility.** Compared to the ablated version without synchronous synthesis, we observe a higher distribution of "OK" or better for the motions of the original speakers and our synthesized agents. Similar to the motion plausibility in set 1, we observe modes of all the distributions on "OK".



(d) **Set 2: Synchronization between the face and the pose expressions given the speech.** We again obverse clear preferences for the motions of the original speakers and our synthesized agents compared to the ablated version without synchronous synthesis. However, in contrast to the same study in set 1, we notice the modes of the distributions for the first two types of motions are one point lower on the Likert scale, whereas the mode for the ablated version remains on "no/arbitrary movements". We hypothesize this to be the consequence of removing the other ablated versions from the participants' cognitive window. In the absence of other variants, the participants focused more closely on the relative qualities of asynchronous vs. synchronous motions and assessed them more critically.

Figure E.1. **Distributions of the user study responses.** Likert-scale response distributions to the two sets of motions rendered using the five different types of face landmark and pose data (Sec. E). We show the distributions of each of the five Likert-scale points for each type of motion as a percentage of the total responses across all the groups in each set.

Table E.1. **Likert-scale score statistics.** We compute the mean and the standard deviation of the Likert-scale scores across all the motions. For the mean scores, higher values are better, bold indicates **best**, and underline indicates second-best.

	Synthesis type	Plausibility		Synchronization	
		Mean	St. Dev.	Mean	St. Dev.
Set 1	Original Speaker	3.25	0.90	3.10	1.34
	Ours	**3.27**	0.86	**3.15**	1.32
	w/o AC Graphs	2.61	1.14	2.48	1.38
	w/o Disc.	2.79	1.02	2.02	1.30
Set 2	Original Speaker	2.99	0.80	**2.79**	1.08
	Ours	**3.01**	0.82	**2.79**	1.07
	w/o Synchronous Synthesis	2.41	0.78	1.79	0.88

lated version was without using the face and pose anatomical component (AC) graphs for training, and one without our discriminator. In the second set, there were three types of motions in each group corresponding to the same speech: the original speaker motions, motions rendered using the face landmarks and poses synthesized by our network, and the ablated version using asynchronously synthesized faces and poses. Our motivation to separately compare with the asynchronously synthesized motions was to eliminate distractors from other motions and enable our participants to focus more closely on the synchronization between the face and the pose expressions. We randomized the order of these motions in each group in each set and kept the order unknown to the participants. We did not present our other ablated versions to the participants as they had insufficient motion and were visually inferior in obvious ways.

E.2. Evaluation Process

Our aim in the user study is to evaluate our synthesized motions on two key aspects: *(i)* how plausible they appear to human observers compared to the motions of the original speakers and the ablated versions, and *(ii)* whether synchronous synthesis of face and pose expressions produces perceptible improvements over asynchronous synthesis. To evaluate plausibility, we ask the participants to rate each motion in each group in each set on "how natural the motion looks" on a five-point Likert scale, with the options "very unnatural" (worst), "not realistic", "looks OK" (average), "looks good", and "looks great" (best). To evaluate the effect of synchronous synthesis, we ask the participants to observe the face and the pose movements in each motion in each group in each set and rate them on "how the face and the pose sync with the speech" on a five-point Likert scale, with the options "no/arbitrary movements" (worst),

"slight movements", "has movements, but are not expressive" (average), "somewhat expressive movements", and "have movements with appropriate expressions" (best).

E.3. Results

Since we randomly selected the speech for each of the eight groups of motions each participant watched and randomized the order of the motions in each group in each set, we can consider the participants' responses in each group to be independent of all the other groups. Thus, we aggregate their responses to each type of motion across all the groups within a set to obtain the overall distributions of the Likert-scale scores of the motions for that set. We show these distributions for each of the two questions on plausibility and synchronization in each set in Fig. E.1. We also report the Likert-scale score statistics for each type of motion on the two questions in each set in Table E.1. Overall, in the two sets, 88.89% and 80.00% participants, respectively, marked our synchronously synthesized motions 3 or above on the first question, and 65.46% and 62.87% participants, respectively, marked 3 or above on the second question. This indicates that the majority of participants found the motions satisfactory.

References

[1] Andreas Aristidou and Joan Lasenby. Fabrik: A fast, iterative solver for the inverse kinematics problem. *Graphical Models*, 73(5):243–260, 2011. 2

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 2

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[4] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2755–2764, 2021. 1

[5] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, pages 376–380, 1991. 1

[6] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics*, 39(6), 2020. 2

[7] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 1