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

Anonymous CVPR submission

Paper ID 0006

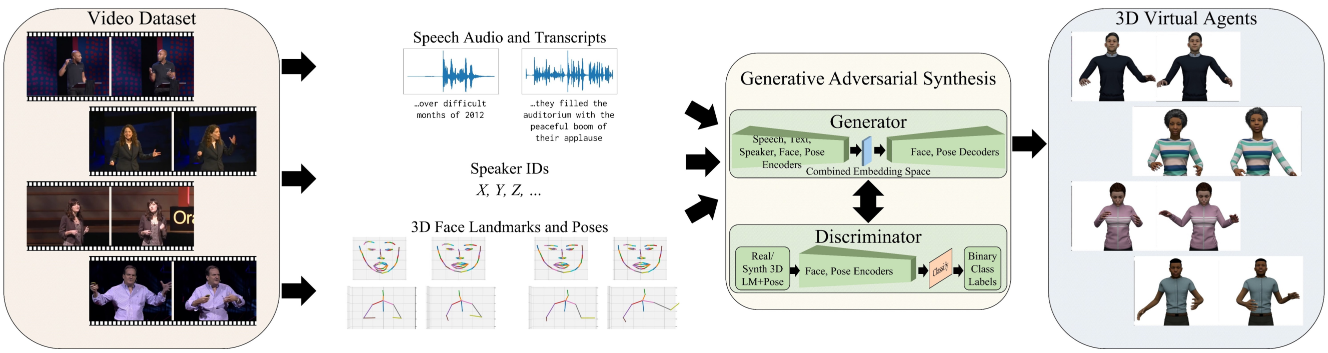


Figure 1. **Synthesizing unified co-speech 3D face and pose expressions.** Our method uses the speech audio, the corresponding text transcripts, the speaker's unique IDs, and their sparse 3D face landmarks and pose sequences computed from RGB video data. It learns a combined embedding space that captures the correlations between all these inputs, and leverages them to generate synchronous affective expressions for faces and poses in a continuous motion space.

Abstract

We present a multimodal learning-based method to simultaneously synthesize co-speech facial expressions and upper-body gestures for digital characters using RGB video data captured using commodity cameras. Our approach learns from sparse face landmarks and upper-body joints, estimated directly from video data, to generate plausible emotive character motions. Given a speech audio waveform and a token sequence of the speaker's face landmark motion and body-joint motion computed from a video, our method synthesizes the full sequence of motions for the speaker's face landmarks and body joints that match the content and the affect of the speech. To this end, we design a generator consisting of a set of encoders to transform all the inputs into a multimodal embedding space capturing their correlations, followed by a pair of decoders to synthesize the desired face and pose motions. To enhance the plausibility of our synthesized motions, we use an adversarial discriminator that learns to differentiate between the face and pose motions computed from the original videos and our synthesized motions based on their affective expressions. To evaluate our approach, we extend the TED Gesture Dataset to include view-normalized, co-speech face landmarks in addition to body gestures. We demonstrate the performance of our method through thorough quantitative and qualitative experiments on multiple evaluation metrics and via a user study, and observe that our method results in low reconstruction error and produces synthesized samples with diverse facial expressions and body gestures for digital characters. We will release the extended dataset as the TED Gesture+Face Dataset consisting of 250K samples and the relevant source code.

1. Introduction

Spoken communications are a significant component of everyday human-human interactions. Human communications through digital platforms and virtual spaces are prevalent in many applications, including online learning [27, 29, 44], virtual interviewing [6], counseling [14], social robotics [50], automated character designing [33], storyboard visualizing for consumer media [24, 48], and creating large-scale metaverse worlds [38]. Simulating immersive experiences in such digital applications necessitates the development of plausible human avatars with expressive face and body motions. This is a challenging problem to approach at scale, given the diversity in human expressions and their importance in human-human interactions [35, 39]. The problem becomes even harder given that humans express simultaneously through multiple cues or *modalities*, such as their speech, facial expressions, and body gestures [36]. The emotional expressions from these different modalities are also synchronous, *i.e.*, they follow the same rhythm of communication and complement each other to convey a sense of presence [25].

In this paper, we consider the problem of synthesizing 3D digital human motions with synchronous facial expressions and upper-body gestures aligned with given speech audio inputs. Given the speech audio, existing approaches in computer vision and graphics tackle the sub-problems of "talking heads" [23] – synthesizing lip movements and facial expressions given the speech audio, and co-speech gesture synthesis [51] – synthesizing poses for upper-body gestures, including head motions. However, these approaches synthesize only one modality, either facial expressions or body gestures. More recent approaches consider head and body motions simultaneously [20, 49], but are confined to a limted set of speakers and their expressions. The inherent difficulty in synthesizing expressions synchronized across diverse speakers is to under the correlations between the modalities for both the expressions and the individual styles [2]. In other words, not only is the combined space of the multimodal expressions very high-dimensional, but only a small fraction of that space corresponds to *valid* expressions for different speakers. Moreover, existing approaches generally require specialized data such dense 3D face scans [13] and motion-captured gestures [10, 11] to provide meaningful results. By contrast, our goal is to leverage large-scale video datasets [50] to develop synchronous co-speech face and pose expressions, with the aim of synthesizing fully expressive 3D digital humans for democratized use in various social environments.

Main Contributions. We present a multimodal learning method to synthesize animated 3D digital characters with synchronous face and upper-body pose sequences for different affective expressions given speech audio. We also consider both intra- and inter-speaker variability by introducing random sampling on a latent space for speakers. Our main contributions include:

- **Synchronous co-speech face and pose expression synthesis.** Our method simultaneously synthesizes face and upper-body pose expressions given speech audio through a generative multimodal embedding space and an affective discriminator. Our method reduces the mean absolute errors on the face landmarks by 30%, and the body poses by 21%, compared to the baseline talking head and co-speech gesture syntheses methods, thereby indicating measurable benefits over asynchronously combining the synthesized outputs of the two modalities.
- **Using data from affordable commodity cameras.** In contrast to facial expression synthesis using dense 3D face scans or gesture synthesis from expensive motion-captured data, our method only relies on face landmarks and pose joints obtainable from commodity hardware such as video cameras. As a result, our method scales affordably to large datasets and is applicable in large-scale social applications.
- **Plausible motions and proposed evaluation metric for facial expressions.** Through quantitative evaluations and user studies, we verify that our synthesized synchronous expressions have low reconstruction errors and are satisfactory to human observers. We also propose the Fréchet Landmark Distance to evaluate the quality of the synthesized face landmarks.
- **TED Gesture+Face Dataset.** We extend the TED Gesture Dataset to include 3D face landmarks extracted from the raw videos that we denoise and align with the poses. We release this multimodal dataset of speech audio, 3D face landmarks, and 3D body pose joints with our paper and the associated source code.

2. Related Work

We briefly review the body of work on perceiving multimodal affective expressions, particularly from faces, speech, and gestures, and also the synthesis of digital characters with co-speech face and pose expressions.

Perceiving Multimodal Affective Expressions. Studies in psychology and affective computing indicate that humans express emotions simultaneously through multiple modalities, including facial expressions, prosody and intonations of the voice, and body gestures [36, 46]. Methods for detecting facial expressions [17] generally depend on facial action units [52]. Methods for detecting various affective vocal patterns commonly use Mel-Frequency Cepstral Coefficients (MFCCs) [37]. Methods to detect emotions from body gestures use physiological features, such as arm swings, spine posture, and head motions that are either predefined [5, 7] or learned automatically from the gestures [8]. The emotions themselves can be represented either as discrete categories such as the Ekman emotions [15] or as combinations of continuous dimensions, such as the Valence-Arousal-Dominance (VAD) model [34]. In our work, we leverage the current approaches for detecting facial, vocal, and pose expressions to design our co-speech face and gesture synthesis method. While we do not explicitly consider specific emotions, our representation implicitly considers emotions in the continuous VAD space, leading to appropriately expressive face and pose synthesis.

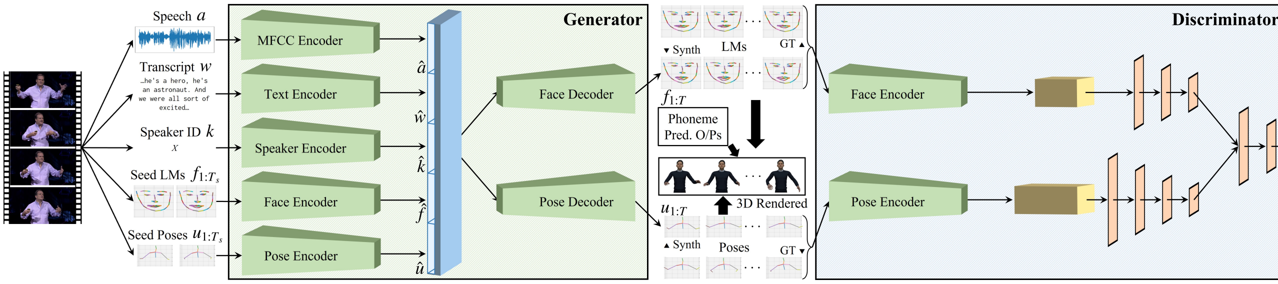


Figure 2. **Network architecture for synchronous synthesis of co-speech face and pose expressions.** Our generator encodes all the inputs: the speech audio, the corresponding test transcript, the speaker ID, and the seed 3D face landmarks and the seed 3D poses into a multimodal embedding space. It decodes variables from this space to produce the synchronized sequences of co-speech 3D face landmarks and poses. Our discriminator classifies these synthesized sequences and the corresponding ground-truths (3D motions of the original speakers), computed directly from the videos, into two different classes based both on their plausibility and their synchronous expressions. To obtain our rendered 3D character motions, we combine the outputs of our generator with our phoneme predictor network and map them to 3D meshes.

Synthesizing Co-Speech Expressions. We consider digital characters with faces and body gestures.

Co-Speech Facial Expressions. Wang and Soong [47] compute controllable parameters for synthesizing talking heads with desired facial expressions using a Hidden Markov Model and MFCCs of the speech audio. Recent techniques automate the facial motions for large-scale synthesis, using generative paradigms such as VAEs [19] and GANs [42]. Karras et al. [23] train a DNN to map speech audio to 3D face vertices conditioned on learned latent features corresponding to different facial expressions. Zhou et al. [53], learn sequences of predefined visemes using LSTM networks from audio. Cudeiro et al. [13] propose a dataset of 4D face scans and learn per-vertex offsets to synthesize the face motions from audio. Richard et al. [41] learn co-speech facial motions using dense face meshes by disentangling speech-correlated and speech-uncorrelated facial features. Sinha et al. [45] focus on adding emotional expressions to the faces. Lahiri et al. [26] focus on the accuracy of the lip movements and use an autoregressive approach to synthesize 3D vertex sequences for the lips that are synced with the speech audio. In contrast to these approaches, our facial expression synthesis method uses much sparser 3D face landmarks detected from real-world videos with arbitrary orientations and lighting conditions of the faces w.r.t. the cameras, and synthesizes facial and pose expressions that are mutually coherent.

Co-Speech Gestures. We can consider co-speech gesture synthesis to be a special case of gesture stylization, where the style refers to the pose expressions that are inferred from and aligned with the speech. This line of work has been richly explored [3, 12, 21, 28, 30–32, 40]. Ginosar et al. [18] propose a method to synthesize speaker-specific co-speech gestures by training a neural network given their identities and individual gesticulation patterns. Ferstl et al. [16] additionally propose using adversarial losses in the training process to improve the fidelity of the synthesized gestures. Yoon et al. [51] extend the concept of individualized gestures to a continuous space of speakers to incorporate natural variability in the synthesized gestures even for the same speaker. Bhattacharya et al. [9] build on top of [51] to improve the affective expressions in the co-speech gestures. More recent methods have also explored diffusion-based approaches for editability [4]. Our method conditions the gesture synthesis on both the input speech and the synthesized facial expressions.

Co-Speech Multimodal Expressions. Co-speech face and upper-body generation has gained particular interest recently, primarily due to the availability of rich 3D datasets of famous speakers [20]. Current approaches train adversarial encoder-decoder architecture on datasets of one speaker at a time [20] and use vector quantization for tokenized generation using a transformer [49]. These approaches are limited to a fixed set of speakers and lose fine-grained expressions when using quantization. In our work, we consider the full continuous space of affective face and body expressions and develop a network that is generalizable to multiple speakers.

3. Synchronous Face and Pose Synthesis

Given a speech audio waveform a, the corresponding text transcript w, the speaker's unique ID k in a set of speakers K, and the associated seed face landmark deltas $f_{1:T_s}$ and seed pose unit vectors $u_{1:T_s}$, T_s being the number of seed time steps, we synthesize the synchronous sequences of face landmark deltas $f_{1:T}$ and pose unit vectors $u_{1:T}$ for the speaker for the T prediction time steps ($T \gg T_s$), matching the content and the affect in their speech. We describe our end-to-end pipeline, including a detailed description of our inputs and outputs and their usage. We provide the details of obtaining these facial and landmarks and poses from input videos in the appendix (Sec. A).

3.1. Computing Face and Pose Expressions

We consider a reference neutral expression $\mathcal{F} \in \mathbb{R}^{L \times 3}$ for each user, L being the number of face landmarks. To synthesize facial expressions, we compute the relative motion of each landmark w.r.t. the reference expression. Specifically, we obtain the configuration $\mathcal{F}_t$ at time step t as

$$\mathcal{F}_t = \mathcal{F} + f_t, \tag{1}$$

where $f_t \in \mathbb{R}^L$ denotes the set of relative motions of the landmarks w.r.t. $\mathcal{F}$ at time step t.

On the other hand, we assume the body joints are rigidly connected by the bones. We represent each user's body joints as 3D point vectors $\mathcal{P} \in \mathbb{R}^{J \times 3}$ in a global coordinate space, where J is the number of joints. We consider directed line vectors connecting adjacent joints. The direction is along the path from the root (pelvis) joint to the end effectors (such as wrists). These 3D point vectors and line vectors collectively form a directed tree with J nodes and $J - 1$ edges. We assume that the magnitudes of these line vectors correspond to the bone lengths and that these magnitudes are known and fixed. To synthesize the users' body gestures, we compute the orientations of these line vectors at each time step t in the reference frame of the global coordinate space. Specifically, for each bone b with bone length (magnitude) $\|b\|$ and connecting the source joint $s_b(t)$ to the destination joint $d_b(t)$ at time step t, we compute a unit vector u_t such that

$$d_b = s_b + \frac{\|b\|}{\|u_t\|} u_t. \tag{2}$$

We do not assume any locomotion, *i.e.*, we consider the root joint is fixed at the global origin at all the time steps.

3.2. Synthesizing Faces and Poses

Our network architecture (Fig. 2) consists of a phoneme predictor to predict the lip shapes corresponding to the audio and a generator-discriminator pair to synthesize plausible co-speech face and pose expressions. We design our phoneme predictor following prior approaches [26] and provide its details in the appendix (Sec. B). Our generator follows a multimodal learning strategy. It consists of separate encoders to transform the speech audio, the text transcript, the speaker ID, the seed face landmark deltas, and the seed pose unit vectors into a latent embedding space representing their correlations. It subsequently synthesizes the appropriate face and pose motions from this multimodal embedding space. Our discriminator enforces our generator to synthesize plausible face and pose motions in terms of their affective expressions. To this end, we use the same encoder architecture for the faces and the poses as in our generator, but learned separately. We describe each of the components of our generator and discriminator.

3.2.1 Encoding Speech, Text, and Speaker IDs

We use the Mel-Frequency Cepstral Coefficients (MFCCs) for the speech audio to accurately capture the affective intonations in the speech, and use an MFCC encoder to obtain speech-based latent embeddings $\hat{a} \in \mathbb{R}^{T \times D_a}$ of dimension D_a as

$$\hat{a} = \text{MFCCEncoder}\left(a; \theta_{\text{MFCC}}\right), \tag{3}$$

where θ_{MFCC} represents the trainable parameters.

Similarly, we use the sentiment-aware FastText [43] embeddings of the words in the transcript and a convolution-based text encoder to obtain the text-based latent embeddings $\hat{w} \in \mathbb{R}^{T \times D_w}$ of dimensions D_w as

$$\hat{w} = \text{TextEncoder}\left(w; \theta_{\text{text}}\right), \tag{4}$$

where θ_{text} represents the trainable parameters.

We also represent the speaker IDs $k \in \{0, 1\}^K$ as one-hot vectors for a total of K speakers and use a speaker encoder to obtain the parameters $\mu_k \in \mathbb{R}^{D_k}$ and $\Sigma_k \in \mathbb{R}_+^{D_k \times D_k}$ of a latent distribution space of dimension D_k as

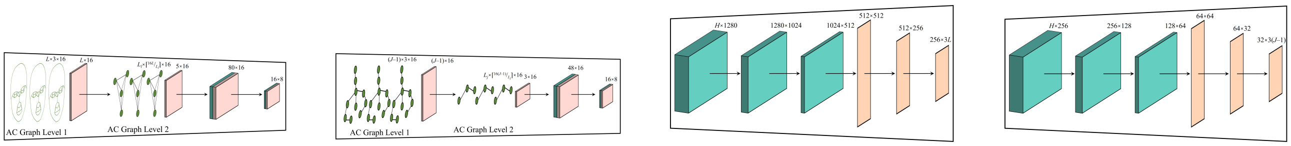
$$\mu_k, \Sigma_k = \text{SpeakerEncoder}\left(k; \theta_{\text{speaker}}\right), \tag{5}$$

where θ_{speaker} represents the trainable parameters. The latent distribution space enables us to sample a random vector $\hat{k}$ representing a speaker who is an arbitrary combination of the K speakers in the dataset. This allows for variations in the synthesized motions even for the same original speaker by slightly perturbing their speaker IDs in the latent distribution space, leading to more plausible results on multiple runs of our network. To learn faces and poses with appropriate expressions, we represent them as multi-scale graphs and encode them using graph convolutional networks.

3.2.2 Encoding Affective Expressions

The face landmarks we use are based on action units [52]. We represent the sequence of 3D landmarks $f_{1:T_s} \in \mathbb{R}^{T_s \times L \times 3}$ as a spatial-temporal anatomical component (AC) graph. Spatially, we consider landmarks belonging to the same anatomical component (Sec. 3.1) and nearest landmarks across different anatomical components to be adjacent. Temporally, all landmarks are adjacent to their temporal counterparts (same nodes at different time steps) within a predetermined time window. We consider the eyes, the nose, the lips, and the lower jaw as the anatomical components. We show the face landmarks graph in Fig. 3a with all the intra- and inter-anatomical-component adjacencies marked with lines. We apply a sequence of spatial-temporal graph convolutions on this graph to learn from the localized motions of the landmarks and obtain embeddings $\tilde{f} \in \mathbb{R}^{T_s \times L \times D_f}$ of feature dimension D_f as

$$\tilde{f} = \text{STGCN}_f\left(f_{1:T_s}; \theta_{\text{STGCN}_f}\right), \tag{6}$$



(a) Face encoder (b) Pose Encoder (c) Face Decoder (d) Pose Decoder

Figure 3. **Face and pose encoders and decoders.** We show their architectures with the layer sizes denoted (details in Sec. 3.2.2). Our architectures depend on the hierarchical anatomical component (AC) graphs for both faces and poses that efficiently learn their corresponding affect representations using spatial-temporal graph convolutions (green nodes and edges), 2D convolutions (teal blocks), 2D batch normalizations (pink blocks), and fully-connected layers (orange planes).

where θ_{STGCN_f} represents the trainable parameters. From the landmarks graph, we obtain a face anatomy graph, where we consider the nodes to represent entire anatomical components and the graph to be fully connected. To compute such a graph, we append the features of intra-anatomical-component nodes in the graph into collated features $l \in \mathbb{R}^{T_s \times L_l \times n_l D_f}$, where L_l denotes the number of anatomical components and n_l denotes the number of landmark nodes within each anatomical component. We take n_l to be the number of nodes in the anatomical component with the most landmarks and perform zero padding as appropriate to obtain the full collated features for the other components. This hierarchically pooled representation provides a "higher-level" view of the face and helps our network learn from the correlations between the motions of the different anatomical components. Specifically, we use another set of spatial-temporal graph convolutions to obtain the embeddings $\tilde{l} \in \mathbb{R}^{T_s \times L_l \times D_l}$ of feature dimension D_l as

$$\tilde{l} = \text{STGCN}_l \left(l; \theta_{\text{STGCN}_l} \right), \tag{7}$$

where θ_{STGCN_l} represents the trainable parameters. Collectively, the landmarks graph and the face anatomy graph provide complementary information to our network to encode and synthesize the required facial expressions at both the macro (anatomy) and the micro (landmark) levels. To complete our encoding, we flatten out the features of all the anatomical components in $\tilde{l}$, *i.e.*, reshaping such that $\tilde{l} \in \mathbb{R}^{T_s \times L_l D_l}$, and transform them using standard convolutional layers on the flattened feature channel and the temporal channel separately. This gives us our latent space embeddings $\hat{l} \in \mathbb{R}^{T \times D_{\tilde{l}}}$ as

$$\hat{l} = \text{ConvT}_{\tilde{l}} \left(\text{ConvS}_{\tilde{l}} \left(\tilde{l}; \theta_{\text{ConvS}_{\tilde{l}}} \right); \theta_{\text{ConvT}_{\tilde{l}}} \right), \tag{8}$$

where $\theta_{\text{ConvS}_{\tilde{l}}}$ and $\theta_{\text{ConvT}_{\tilde{l}}}$ represent the trainable parameters.

For the pose representation, we consider a pose graph of the upper body with $J-1$ bones represented with line vectors $u_{1:T_s}$ (Fig. 3b). We consider bones connected to each other or connected through a third bone to be adjacent. We use a set of spatial-temporal graph convolutions to leverage the localized motions of these bones and obtain embeddings $\tilde{u} \in \mathbb{R}^{T_s \times D_u}$ of feature dimension D_u as

$$\tilde{u} = \text{STGCN}_u \left(u_{1:T_s}; \theta_{\text{STGCN}_u} \right), \tag{9}$$

where θ_{STGCN_u} represents the trainable parameters. Similar to the face landmarks, we also consider a hierarchically pooled representation of the bones $v \in \mathbb{R}^{T_s \times L_j \times n_j D_u}$, where $L_j = 3$ are the three anatomical components, the torso and the two arms, represented as single nodes each consisting of n_j nodes from the pose graph. In the pose anatomy graph, we consider the two arms to be adjacent to the torso but not to each other, as they can move independently. We apply a second set of spatial-temporal graph convolutions on the collated features v to obtain the embeddings $\tilde{v} \in \mathbb{R}^{T_s \times L_j \times D_v}$ as

$$\tilde{v} = \text{STGCN}_v \left(v; \theta_{\text{STGCN}_v} \right) \tag{10}$$

where θ_{STGCN_v} represents the trainable parameters. To subsequently obtain the latent space embeddings $\hat{v} \in \mathbb{R}^{T \times D_{\tilde{v}}}$, we apply separate spatial and temporal convolutions on the flattened graph-convolved features $\tilde{v} \in \mathbb{R}^{T_s \times L_j D_v}$, as

$$\hat{v} = \text{ConvT}_{\tilde{v}} \left(\text{ConvS}_{\tilde{v}} \left(\tilde{v}; \theta_{\text{ConvS}_{\tilde{v}}} \right); \theta_{\text{ConvT}_{\tilde{v}}} \right), \tag{11}$$

where $\theta_{\text{ConvS}_{\tilde{v}}}$ and $\theta_{\text{ConvT}_{\tilde{v}}}$ represent the trainable params.

3.2.3 Synthesizing Synchronous Motions

Our synchronous synthesis relies on learning the multimodal distributions of the individual modalities of audio, text, speaker ID, face expressions, and pose expressions given their individual distributions. To this end, we append all the latent space embeddings — $\hat{a}$ for the audio, $\hat{w}$ for the text, $\hat{k}$ for the random speaker representation, repeated over all the T time steps, $\hat{l}$ for the seed landmarks and $\hat{v}$ for the seed poses — into a vector $\hat{e} \in \mathbb{R}^{T \times H}$ representing a multimodal embedding space of all the inputs. Here, $H = D_a + D_w + D_k + D_{\tilde{l}} + D_{\tilde{v}}$ denotes the latent space dimension. On training, our network learns the correlations between the different inputs in this multimodal embedding space. To synthesize our face landmark motions $f_{1:T} \in \mathbb{R}^{T \times L \times 3}$, we apply separate spatial and temporal convolutions on the multimodal embeddings $\hat{e}$ to capture

Figure 4. **Qualitative results.** Snapshots from two of our synthesized samples showing the text transcript of the speech and the corresponding face and pose expressions (row 1). We also zoom in on the eyebrow (row 2) and lip (row 3) expressions for better visualization. We observe a smile, raised eyebrows, and stretched arms (left) for the word ‘excited’, and frowns on the eyebrows and lips (right) for the words ‘very sorry’.

localized dependencies between the feature values followed by fully-connected layers capturing all the dependencies between the feature values (Fig. 3c), as

$$f_{1:T} = \mathrm{FC}_{f\hat{e}}\left(\mathrm{ConvS}_{f\hat{e}}\left(\mathrm{ConvT}_{f\hat{e}}\left(\hat{e};\theta_{\mathrm{ConvT}_{f\hat{e}}}\right);\theta_{\mathrm{ConvS}_{f\hat{e}}}\right);\theta_{\mathrm{FC}_{f\hat{e}}}\right), \quad (12)$$

where $\theta_{\mathrm{ConvT}_{f\hat{e}}}$, $\theta_{\mathrm{ConvS}_{f\hat{e}}}$, and $\theta_{\mathrm{FC}_{f\hat{e}}}$ represent the trainable parameters. The output $f_{1:T}$ from the fully-connected layers has shape $T \times 3L$, which we reshape into $T \times L \times 3$ to get our desired 3D face landmark sequences.

We similarly synthesize the line vectors $u_{1:T} \in \mathbb{R}^{T\times(J-1)\times 3}$ using separate spatial and temporal convolutions on the multimodal embeddings $\hat{e}$, followed by fully-connected layers (Fig. 3d), as

$$u_{1:T} = \mathrm{FC}_{u\hat{e}}\left(\mathrm{ConvS}_{u\hat{e}}\left(\mathrm{ConvT}_{u\hat{e}}\left(\hat{e};\theta_{\mathrm{ConvT}_{u\hat{e}}}\right);\theta_{\mathrm{ConvS}_{u\hat{e}}}\right);\theta_{\mathrm{FC}_{u\hat{e}}}\right), \quad (13)$$

where $\theta_{\mathrm{ConvT}_{u\hat{e}}}$, $\theta_{\mathrm{ConvS}_{u\hat{e}}}$, and $\theta_{\mathrm{FC}_{u\hat{e}}}$ represent the trainable parameters. Given the synthesized face and pose motions, we use our discriminator to determine how well their affective expressions match that of the corresponding ground-truths in the training data. We obtain our ground-truths as the 3D face landmarks and the 3D pose sequences computed from the full training video data.

3.2.4 Determining Plausibility Using Discriminator

Our discriminator takes in the synchronously synthesized face motions $f_{1:T}$ and pose motions $u_{1:T}$, and encodes them using encoders with the same architecture as our generator (Sec. 3.2.2), with only the number of input time steps being T instead of T_s. This gives us the corresponding latent space embeddings $\hat{l}$ and $\hat{v}$. Similar to our generator, we concatenate these embeddings into a multimodal embedding vector $\hat{e} \in \mathbb{R}^{T\times\left(D_{\hat{l}}+D_{\hat{v}}\right)}$. But different from our generator, we pass these multimodal embeddings through a fully-connected classifier network $\mathrm{FC}_{\mathrm{disc}}$ to obtain class probabilities $c_{\mathrm{disc}} \in [0, 1]$ per sample, as

$$c_{\mathrm{disc}} = \mathrm{FC}_{\mathrm{disc}}\left(\hat{e};\theta_{\mathrm{FC}_{\mathrm{disc}}}\right), \quad (14)$$

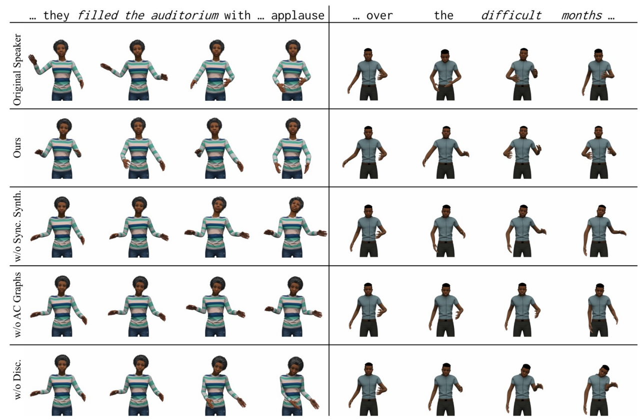


Figure 5. **Qualitative comparisons.** For the same input speech, represented by the text transcript at the top, we compare the visual quality of our synthesized character motions with the original speaker motions and three of our ablated versions: one without synchronous face and pose synthesis, one without our anatomical component (AC) graphs for faces and poses, and one without our discriminator. We observe that our synthesized motions are visually the closest to the original speaker motions compared to the ablated versions. We elaborate on their visual qualities in Sec. 5.4.

where $\theta_{\mathrm{FC}_{\mathrm{disc}}}$ represents the trainable parameters. Our discriminator learns to perform unweighted binary classification between the synthesized face and pose motions and the ground-truths in terms of their synchronous affective expressions. Our generator, on the other hand, learns to synthesize samples that our discriminator cannot distinguish from the ground-truth based on those affective expressions. We provide all our training, testing, and rendering details in the appendix (Secs. C and D).

4. TED Gesture+Face Dataset

We present our TED Gesture+Face Dataset that we use to train and test our network. We elaborate on collecting and processing our dataset for training and testing.

Dataset Collection. The TED Gesture Dataset [50] consists of videos of TED talk speakers together with text transcripts of their speeches, and their 3D body poses extracted in a global frame of reference. The topics range from personal and professional experiences to discourses on educational topics and instructional and motivational storytelling. The speakers themselves come from a wide variety of social, cultural, and economic backgrounds, and are diverse in age, gender, and physical abilities.

Dataset Processing. The 3D poses in the original TED Gesture Dataset [50] are view-normalized to face front and center at all time steps. We compute similarly view-normalized 3D face landmarks of the speakers (Sec. A.1) .

Table 1. **Quantitative evaluations.** Comparison with existing co-speech gesture synthesis methods and our ablated versions (Sec. 5.1) on the metrics MALE (in mm), MAJE (in mm), MAcE for landmarks (MAcE-LM) (in mm/s^2), MAcE for poses (MAcE-P) (in mm/s^2), FLD, and FGD (Sec. 5.2). Lower values are better, bold indicates **best**, and underline indicates second-best.

Method	MALE	MAJE	MAcE-LM	MAcE-P	FLD	FGD
Seq2Seq [50]	–	45.62	–	6.33	–	6.62
S2G-IS [18]	–	45.11	–	7.22	–	6.73
JEM [1]	–	48.56	–	4.31	–	5.88
GTC [51]	–	27.30	–	3.20	–	4.49
Speech2AffectiveGestures [9]	–	24.49	–	2.93	–	3.54
SpeechGestureMatching [21]	–	21.10	–	2.75	–	2.64
Ours w/o Face Synthesis	–	28.32	–	3.89	–	4.01
Ours w/o Pose Synthesis	11.76	–	9.38	–	22.65	–
Ours w/o Vel.+Acc. Losses	26.33	24.41	21.69	7.58	27.54	7.72
Ours w/o Discriminator	14.62	27.40	13.44	11.60	31.93	8.79
Ours w/o Face AC Graph	13.05	25.97	14.24	2.74	25.61	2.25
Ours w/o Pose AC Graph	11.84	25.46	8.12	13.88	19.23	6.94
Ours w/o Synchronous Synthesis	10.72	25.03	7.83	3.22	18.03	3.92
Ours	**9.00**	**18.36**	**6.34**	**2.52**	**15.02**	**1.79**

Similar to the original TED dataset, we divide the 3D pose and face landmark sequences into equally-sized chunks of size $T = 34$ time steps at a rate of 15 fps. Additionally, to reduce the jitter in the predicted 3D face landmarks and pose joints from each video, we sample a set of "anchor" frames at a rate of 5 fps and perform bicubic interpolation to compute the face landmark and pose joint values in the remaining frames. We use the first 4 time steps of pose and face landmarks as our seed values (Sec. 3.2), and predict the next 30 time steps. The processed dataset consists of 200,038 training samples, 26,903 validation samples, and 26,245 test samples, following a split of 80%-10%-10%.

5. Experiments and Results

We run quantitative experiments using ablated versions of our method as baselines. We note that Habibie et al. [20] retrain their network separately for individual speakers belonging to the same profession (talk show hosts), making it unsuitable for our generalized paradigm consisting of less than 50 samples each of multiple, diverse speakers. Yi et al. [49] use VQ with transformers to synthesize faces and gestures, but are limited to the same set of fixed speakers. We also conducted a web-based user study to evaluate the qualitative performance of our method.

5.1. Baselines

We use seven ablated versions of our method as baselines. The first two ablations correspondingly remove the entire face (Figs. 3a, 3c) and pose components (Figs. 3b, 3d) from our network, making our network learn only talking head and only co-speech gesture syntheses. The third ablation removes the velocity and acceleration losses from our reconstruction loss (Eqn. C.2) , leading to jittery motions. The fourth ablation removes the discriminator and its associated losses (Eqn. C.4) from our training pipeline, leading to unstable motions without appreciable expressions. The fifth and the sixth ablations correspondingly remove the "higher-level" anatomical component (AC) graphs of the faces (Eqn. 7) and the poses (Eqn. 10), leading to reduced movements. The final ablation trains the face and the pose expressions separately, learning marginal embeddings for the two modalities based on the speech but not attending to their mutual synchronization. This ablation is a direct evaluation of the co-speech motions when combining separately synthesized face and pose expressions. For completeness, we also compare with co-speech gesture synthesis methods that only synthesize body poses. We evaluate all the methods on our TED Gesture+Face Dataset.

5.2. Evaluation Metrics

Inspired by prior work [51], we evaluate using four *reconstruction errors* and two *plausibility errors* (PEs). Our reconstruction errors include the mean absolute landmark error (MALE) for the faces, the mean absolute joint error (MAJE) for the poses, and their respective mean acceleration errors (MAcEs). MALE and MAJE indicate the overall fidelity of the synthesized samples w.r.t. the corresponding ground-truths, and the MAcEs indicate whether or not the synthesized landmarks and poses have regressed to their mean absolute positions. To report these metrics, we multiply our ground-truth and synthesized samples by a constant scaling factor such that they all lie inside a bounding box of diagonal length 1 m. For our PE, we use the Fréchet Gesture Distance (FGD) designed by [51] to indicate the perceived plausibility of the synthesized poses. To similarly indicate the perceived plausibility of the synthesized face landmarks, we also design the Fréchet Landmark Distance (FLD). We train an autoencoder network to reconstruct the full set of face landmarks at all time steps for all the samples in the training set of our TED Gesture+Face Dataset. To compute FLD, we then obtain the Fréchet Inception Distance [22] between the encoded features of the ground-truth and the synthesized samples.

5.3. Quantitative Evaluations

We show our quantitative evaluations in Table 1.

Comparison with Co-Speech Gesture Synthesis. Since co-speech gesture synthesis methods do not synthesize face expressions, we leave those numbers blank. For these methods, we have taken the numbers reported by Bhattacharya et al. [9]. For the method of SpeechGestureMatching [21], we retrain their method on the TED Gesture Dataset to report the numbers. However, we were unable to perform similar comparative evaluations with co-speech face synthesis methods as existing methods synthesize dense landmarks [23] or blendshape-like features [13], which cannot be mapped one-to-one with our sparser face landmarks.

Comparison with Ablated Versions Removing either the face or the gesture components of our network leads to poorer values across the board compared to using both of them. Without the velocity and acceleration losses, the motions are jittery, and the MAcE losses are higher, especially MAcE for the face landmarks. Without the discriminator, the synthesized samples suffer from mode collapse and often produce implausible motions, leading to higher values across the board. Without the AC graphs, there are fewer movements in the synthesized and the reconstruction errors are higher. When synthesizing face and pose expressions separately and not synchronizing them, we observe some mismatches in when the expressions from either modality appear and how intense they are. This indicates that synchronous synthesis of facial expressions and body gestures leads to more accurate and plausible movements for both the modalities, including a 30% improvement on MALE and a 21% improvement on MAJE, compared to trivially combining synthesized outputs of the individual modalities.

5.4. Qualitative Comparisons

We visualize some of our synthesized samples in Fig. 4 and provide more results in our supplementary video. We observe the synchronization between the face and the pose expressions for two contrasting emotions. We also visually compare with the original speaker motions rendered using their face landmarks and the poses extracted from the videos, and three of our ablated versions in Fig. 5. The original speaker motions provide an "upper bound" of our performance. The three ablated versions we compare with are: one without the synchronous synthesis, one without our face and pose AC graphs, and one without our discriminator. The ablated versions without either the face or the pose synthesis, without the velocity and acceleration losses, and without our discriminator are visually inferior in obvious ways, therefore we leave them out. Without either face or pose synthesis, that modality remains static while the other one moves. Without the velocity and the acceleration losses, the overall motions regress to the mean pose. Without our discriminator, our generator often fails to understand plausible movement patterns, leading to unnatural limb and body shapes. Of these, we only keep the ablations without our discriminator as our "lower bound" baseline because, unlike the other two, this ablation has visible movements in both the face and the pose modalities.

5.5. User Study

We conducted a user study in two sets to evaluate the visual quality of our synthesized motions in terms of their plausibility and synchronization. We provide an overview of the results here and elaborate on all the details in the appendix (Sec. E). The first set compares between our method and its ablations without the AC graph and the discriminator.

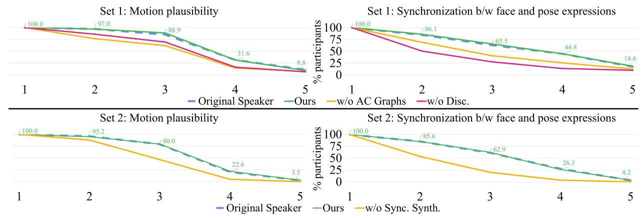


Figure 6. **Cumulative lower-bound of participant responses.** We plot the cumulative lower-bound (LB) percentage of responses across the Likert-scale scores for each type of character motion in each set. A cumulative LB percentage X for a Likert-scale score s denotes $X\%$ of responses had a score of s or higher. We observe that the curve for our synchronously synthesized motions stays at the top, indicating that the participants preferred it over the other motions.

The second set compares between our method and its ablation without synchronous face and pose synthesis. In each set, we collect responses from 90 responses on 5-point Likert scales (1=worst, 5=best) to evaluate two aspects, plausibility and synchronization. We plot the cumulative lower bound of participant responses for each Likert-scale score for each type of motion in each set in Fig. 6. We note that the scores for our synchronously synthesized samples remain close to the original speaker scores and consistently above the other ablated versions, indicating a clear preference.

6. Conclusion, Limitations and Future Work

We have presented a method to synthesize synchronous co-speech face and pose expressions for 3D digital characters. Our method learns to synthesize these expressions from 3D face landmarks and 3D upper-body pose joints computed directly from videos. Our work also has some limitations. We use sparse face landmarks and pose joints to synthesize co-speech face and pose expressions. To synthesize more fine-grained expressions, we plan to extract more detailed face meshes and additional pose joints from videos. Further, given the sparsity of our face and pose representations and the noise associated with extracting them from videos, the quality of our synthesized motions do not match those synthesized from high-end facial scans and motion-capture data. We aim to bridge this gap by building techniques to develop more robust face and pose representations from videos. We also plan to combine our work with lower-body actions such as sitting, standing, and walking to synthesize 3D animated digital humans in a wider variety of scenarios. In terms of its running-time cost, our method uses high-end GPUs to obtain real-time performance. We plan to explore knowledge distillation techniques to reduce our running-time cost and implement our method in real-time on commodity devices such as digital personal assistants.

References

[1] C. Ahuja and L. Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728, 2019. 7

[2] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992. 2

[3] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Trans. Graph.*, 41(6), 2022. 3

[4] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Trans. Graph.*, 2023. 3

[5] Abhishek Banerjee, Uttaran Bhattacharya, and Aniket Bera. Learning unseen emotions from gestures via semantically-conditioned zero-shot perception with adversarial autoencoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):3–10, 2022. 2

[6] T. Baur, I. Damian, P. Gebhard, K. Porayska-Pomsta, and E. André. A job interview simulation: Social cue-based interaction with a virtual character. In *2013 International Conference on Social Computing*, pages 220–227, 2013. 1

[7] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, page 1342–1350. AAAI Press, 2020. 2

[8] Uttaran Bhattacharya, Christian Roncal, Trisha Mittal, Rohan Chandra, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping. In *Computer Vision – ECCV 2020*, pages 145–163, Cham, 2020. Springer International Publishing. 2

[9] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. *Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning*, page 2027–2036. Association for Computing Machinery, New York, NY, USA, 2021. 3, 7

[10] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR)*. IEEE, 2021. 2

[11] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008. 2

[12] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[13] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3, 7

[14] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014. 1

[15] Paul Ekman. Are there basic emotions? 1992. 2

[16] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games*, New York, NY, USA, 2019. Association for Computing Machinery. 3

[17] Panagiotis Giannopoulos, Isidoros Perikos, and Ioannis Hatzilygeroudis. *Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013*, pages 1–16. Springer International Publishing, Cham, 2018. 2

[18] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 7

[19] David Greenwood, Stephen Laycock, and Iain Matthews. Predicting head pose from speech with a conditional variational autoencoder. ISCA, 2017. 3

[20] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, page 101–108, New York, NY, USA, 2021. Association for Computing Machinery. 2, 3, 7

[21] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 Conference Proceedings*, New York, NY, USA, 2022. Association for Computing Machinery. 3, 7

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 7

[23] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.*, 36 (4), 2017. 2, 3, 7

[24] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. page

242–250, New York, NY, USA, 2020. Association for Computing Machinery. 1

[25] Rudolf Laban and Lisa Ullmann. The mastery of movement. 1971. 2

[26] 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. 3, 4

[27] Jamy Li, René Kizilcec, Jeremy Bailenson, and Wendy Ju. Social robots and virtual agents as lecturers for video instruction. *Computers in Human Behavior*, 55:1222 – 1230, 2016. 1

[28] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11293–11302, 2021. 3

[29] M. Liao, C. Sung, H. Wang, and W. Lin. Virtual classmates: Embodying historical learners' messages as learning companions in a vr classroom through comment mapping. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 163–171, 2019. 1

[30] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*, page 3764–3773, New York, NY, USA, 2022. Association for Computing Machinery. 3

[31] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. In *Advances in Neural Information Processing Systems*, pages 21386–21399. Curran Associates, Inc., 2022.

[32] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10462–10472, 2022. 3

[33] S. Mascarenhas, M. Guimarães, R. Prada, J. Dias, P. A. Santos, K. Star, B. Hirsh, E. Spice, and R. Kommeren. A virtual agent toolkit for serious games developers. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–7, 2018. 1

[34] Albert Mehrabian and James A Russell. *An approach to environmental psychology.* the MIT Press, 1974. 2

[35] Batja Mesquita and Michael Boiger. Emotions in context: A sociodynamic model of emotions. *Emotion Review*, 6(4): 298–302, 2014. 2

[36] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 1359–1367. AAAI Press, 2020. 2

[37] Daniel Neiberg, Kjell Elenius, and Kornel Laskowski. Emotion recognition in spontaneous speech using gmms. In *Ninth international conference on spoken language processing*, 2006. 2

[38] NVIDIA Omniverse. *NVIDIA Omniverse, https://www.nvidia.com/en-us/omniverse/*, 2021. 1

[39] Brian Parkinson, Agneta H Fischer, and Antony SR Manstead. *Emotion in social relations: Cultural, group, and interpersonal processes*. Psychology press, 2005. 2

[40] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11077–11086, 2021. 3

[41] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1173–1182, 2021. 3

[42] N. Sadoughi and C. Busso. Novel realizations of speech-driven head movements with generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6169–6173, 2018. 3

[43] I. Santos, N. Nedjah, and L. de Macedo Mourelle. Sentiment analysis using convolutional neural network with fasttext embeddings. In *2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–5, 2017. 4

[44] Adalberto L Simeone, Marco Speicher, Andreea Molnar, Adriana Wilde, and Florian Daiber. Live: The human role in learning in immersive virtual environments. In *Symposium on Spatial User Interaction*, New York, NY, USA, 2019. Association for Computing Machinery. 1

[45] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1320–1327. International Joint Conferences on Artificial Intelligence Organization, 2022. Main Track. 3

[46] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2):211–223, 2012. 2

[47] Lijuan Wang and Frank K Soong. Hmm trajectory-guided sample selection for photo-realistic talking head. *Multimedia Tools and Applications*, 74(22):9849–9869, 2015. 3

[48] Katie Watson, Samuel S. Sohn, Sasha Schriber, Markus Gross, Carlos Manuel Muniz, and Mubbasir Kapadia. Storyprint: An interactive visualization of stories. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, page 303–311, New York, NY, USA, 2019. Association for Computing Machinery. 1

[49] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black.

Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, 2023. 2, 3, 7

[50] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *Proc. of The International Conference in Robotics and Automation (ICRA)*, 2019. 1, 2, 6, 7

[51] 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, 3, 7

[52] 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. 2, 4

[53] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Trans. Graph.*, 37(4), 2018. 3