Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation

Xian Liu[1], Qianyi Wu[2], Hang Zhou[1], Yinghao Xu[1], Rui Qian[1], Xinyi Lin[3],
Xiaowei Zhou[3], Wayne Wu[4], Bo Dai[5], Bolei Zhou[1]
[1]The Chinese University of Hong Kong [2]Monash University [3]Zhejiang University
[4]SenseTime Research [5]S-Lab, Nanyang Technological University
{alvinliu@ie, zhouhang@link, xy119@ie, qr021@ie, bzhou@ie}.cuhk.edu.hk, qianyi.wu@monash.edu,
{shinylin, xwzhou}@zju.edu.cn, wuwenyan@sensetime.com, bo.dai@ntu.edu.sg

Abstract

*Generating speech-consistent body and gesture movements is a long-standing problem in virtual avatar creation. Previous studies often synthesize pose movement in a holistic manner, where poses of all joints are generated simultaneously. Such a straightforward pipeline fails to generate fine-grained co-speech gestures. One observation is that the hierarchical semantics in speech and the hierarchical structures of human gestures can be naturally described into multiple granularities and associated together. To fully utilize the rich connections between speech audio and human gestures, we propose a novel framework named **Hierarchical Audio-to-Gesture (HA2G)** for co-speech gesture generation. In HA2G, a Hierarchical Audio Learner extracts audio representations across semantic granularities. A Hierarchical Pose Inferer subsequently renders the entire human pose gradually in a hierarchical manner. To enhance the quality of synthesized gestures, we develop a contrastive learning strategy based on audio-text alignment for better audio representations. Extensive experiments and human evaluation demonstrate that the proposed method renders realistic co-speech gestures and outperforms previous methods in a clear margin. Project page: https://alvinliu0.github.io/projects/HA2G.*

1. Introduction

When communicating with other people, we spontaneously make co-speech gestures to help convey our thoughts. Such non-verbal behaviors supplement speech information, making the content clearer and more understandable to listeners [11, 48, 62]. Psycho-linguistic studies also suggest that virtual avatars with plausible speech gestures are more intimate and trustworthy [60]. Therefore, actuating embodied AI agents such as social robots and digital humans with expressive body movements and gestures is of great importance to facilitating human machine interaction [55, 56]. To this end, researchers have explored the task of co-speech gesture synthesis [1, 2, 8, 19, 25–27, 41, 53, 68, 69], which aims at generating a sequence of human gestures given the speech audio and transcripts as input.

Traditionally, the task is tackled through building one-to-one correspondences between speech and unit gesture pairs [12, 13, 32, 47]. Such pipelines require huge human efforts, making them inapplicable to general scenarios of unseen speech. Recent studies leverage deep learning to solve this problem by training a neural network to map a compact representation of audio [1, 25, 26, 41, 53] and text [3, 8, 35, 68, 69] to holistic human pose sequence. However, such a straightforward approach fails to capture the micro-scale motions and cross-modal information, *e.g.*, the subtle finger movements and the rich meanings contained in speech audio. The problem of how to learn the fine-grained cross-modal association remains unsolved.

In order to fully exploit the rich multi-modal semantics, we identify two important observations from a human gesture study [48]: 1) Different types of co-speech gestures are related to distinct levels of audio information. For example, the metaphorical gestures are strongly associated with the high-level speech semantics (*e.g.*, when depicting a ravine, one would moving two outstretched hands apart and saying "gap"), while the low-level audio features of beat and volume lead to the rhythmic gestures. 2) The dynamic patterns of different human body parts in co-speech gestures are not the same, such as the flexible fingers and relatively still upper arms. Thus it is improper to generate the upper body pose as a whole like previous studies [1–3, 25, 26, 41, 53, 68].

Inspired by the discussions above, we develop the **Hierarchical Audio-to-Gesture (HA2G)** pipeline, which generates diverse co-speech gestures. Our key insight is to build *hierarchical cross-modal associations across multiple lev-*

els between tri-modal information and generate gestures in a *coarse-to-fine manner*. Specifically, two modules are devised, namely the *Hierarchical Audio Learner*, and the *Hierarchical Pose Inferer*. In the *Hierarchical Audio Learner*, we argue that features extracted from different levels of the audio backbone capture different meanings. Additionally, text information can further strengthen the audio embedding through contrastive learning for more discriminative representations. Afterwards, based on the hypothesis that different levels of audio information contribute to different body joint movements, we associate the multi-level audio features with the hierarchical structure of human body in the *Hierarchical Pose Inferer*. In particular, the association is achieved in correlation with *speaking styles* encoded from speaker appearances. The hierarchy of human upper limb is predicted in a coarse-to-fine manner from shoulders to fingers like a tree structure by cascading multiple bi-directional GRU generators. In addition, we propose a novel physical regularization to enhance the realness of generated poses. Experiments demonstrate that our method synthesizes realistic and smooth co-speech gestures.

To summarize, our main contributions are three-fold: **(1)** We propose the *Hierarchical Audio Learner* to extract hierarchical audio features and render discriminative representations through contrastive learning. **(2)** We propose the *Hierarchical Pose Inferer* to learn associations between multi-level features and human body parts. Human poses are thus generated in a cascaded manner. **(3)** Extensive experiments show that **HA2G** can generate fine-grained co-speech gestures, which outperform state-of-the-art methods on both objective evaluations and subjective human studies.

2. Related Work

Human-Centered Audio-Visual Learning. In recent years, human-centered audio-visual learning has been extensively studied [21–23,44,57–59,66,71,71,72,75]. While some works utilize audio-visual correspondence to solve the problems like music-to-dance [33, 39, 42], and talking face generation [14, 15, 36, 45, 51, 73, 74, 76], the modeling between speech and gesture remains largely unexplored. The difficulty of speech-based gesture generation lies in constructing the correspondence between speech and human gesture, which is more complicated and implicit than music-to-dance or talking face generation.

Human Motion Synthesis. Synthesizing human motions has been of important interest in both computer vision and graphics, where spatial-temporal coherence of pose sequence is used to generate realistic motions [7, 67, 77]. Earlier methods employ statistical models such as kernel-based probability distribution [9, 10, 20, 52] to synthesize human motions. Still, they fail to handle motion details, and the complicated training procedures essentially limit model capacity. Recently, the ability of deep models to generate human motions has been proven on different network architectures, where CNN-based [31, 67], RNN-based [4, 24, 61] and GAN-based [6, 29] methods have been explored. These methods are purely visual-based with the input of history motions, while our work focuses on identifying the strong correlations between speech and gestures in conversational settings to achieve speech-driven motion synthesis.

Audio/Text-Driven Motion Generation. Early works on speech-driven motion generation are mostly rule-based methods [12, 47], where a predefined set of unit gestures and motion connecting rules are designed manually. With the development of deep learning, data-driven approaches have demonstrated superior performance. Some works map speech text information to co-speech gestures [3, 8, 35, 69]. Yoon *et al.* [69] resort to RNN to map from utterance text to upper body gestures. Some methods use speech audio signals to drive gestures [2, 19, 25–27, 41, 53]. For example, Ginosar *et al.* [25] collect a 2D speaker-specific gesture dataset and train the model with an adversarial loss. To make gestures more expressive, Habibie *et al.* [26] lift the 2D pose to 3D and generate facial expressions simultaneously. However, all of their methods learn a model for each speaker individually, which makes it hard to transfer to general scenes and limit speaker styles to a tiny number. Besides, either audio- or text-driven motion generation methods fail to consider messages from both modalities, which motivates recent methods to jointly tackle multi-modal information [1, 37, 68]. Specifically, Yoon *et al.* [68] propose to encode the trimodal feature embeddings of text, audio, speaker identity and concatenate them together to pass a decoder. But they fail to fully make use of multi-level features. Further, the dynamic patterns of different human body parts are diverse when people talk, *e.g.*, the range and frequency of co-speech finger and arm movement are not the same, which makes it unreasonable to learn holistic human pose directly. In this work, we propose to extract hierarchical audio features with a contrastive learning strategy to excavate cross-modal messages at multiple granularities and learn co-speech gestures in a coarse-to-fine manner.

3. Our Approach

We present **Hierarchical Audio-to-Gesture (HA2G)** that generates a target person's co-speech gestures given speech audio. The generated poses are conditioned on speaker identity and initial poses. Following Yoon *et al.* [68], text information can be provided additionally. The whole pipeline is illustrated in Fig. 1. In this section, we first formulate the problem in Sec. 3.1, and then elaborate the *Hierarchical Audio Learner* which extracts hierarchical audio features in Sec. 3.2. Sec. 3.3 introduces the *Hierarchical Pose Inferer* to perform multi-level feature blending and co-speech gesture synthesis. Finally, training objectives for gesture generation are described in Sec. 3.4.

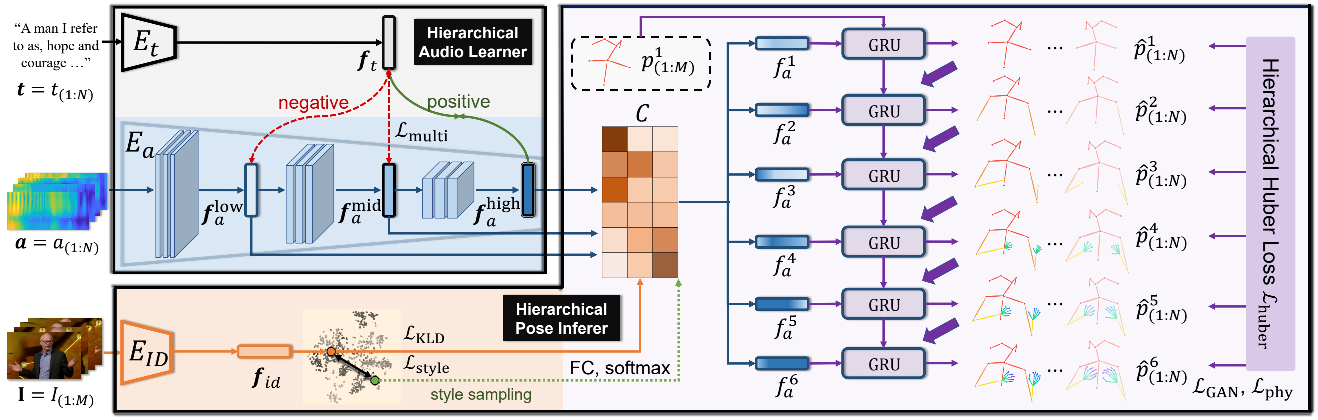

Figure 1. **Illustration of the Hierarchical Audio-to-Gesture (HA2G).** In Hierarchical Audio Learner, E_a encodes speech audio $\mathbf{a}$ into multi-level audio features $\boldsymbol{f}_a^{\text{low}}$, $\boldsymbol{f}_a^{\text{mid}}$ and $\boldsymbol{f}_a^{\text{high}}$ (blue). The speech transcript $\mathbf{t}$ is encoded by E_t into text features $\boldsymbol{f}_t$ (grey). Then a contrastive learning strategy is used to enforce the discriminative audio feature extraction by attracting text feature and high-level audio feature (green) while repelling from low/mid-level features (red). In Hierarchical Pose Inferer, the reference frames $\mathbf{I}$ are encoded by E_{ID} to represent speaker's identity $\boldsymbol{f}_{id}$ (orange), which is then transformed to style coordinator C for multi-level feature blending ($\boldsymbol{f}_a^1$, ..., $\boldsymbol{f}_a^6$). Finally the co-speech gestures $\hat{\boldsymbol{p}}^6_{(1:N)}$ are generated by cascaded bi-GRU based on initial poses $\boldsymbol{p}^1_{(1:M)}$ in a coarse-to-fine manner (purple).

3.1. Problem Formulation

Large amounts of speaking videos with clear co-speech gestures are used for training. Given a video with N frames $\mathbf{V} = \{I_1, \ldots, I_N\}$, the skeletal poses of the upper body can be denoted as $\mathbf{p} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N \mid \boldsymbol{p}_i = [\boldsymbol{d}_{i,1}, \boldsymbol{d}_{i,2}, \ldots, \boldsymbol{d}_{i,J-1}]\}$. Each $\boldsymbol{p}_i$ is represented as the concatenation of unit direction vectors $\boldsymbol{d}_{i,j}$ between J joints. The goal of our model G is to use the video's accompanying speech audio sequence $\mathbf{a} = \{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N\}$ to recover $\mathbf{p}$ according to target's identity representation $\boldsymbol{f}_{id}$ and initial poses $\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_M\}$. Following the setting of Yoon *et al.* [68], the text transcripts $\mathbf{t} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_N\}$ are also provided for training. With encoder E_a for audio information extraction, the overall objective can be written as:

$$\arg\min_{G, E_a} ||\mathbf{p} - G(E_a(\mathbf{a})|\boldsymbol{f}_{id}, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_M)||. \quad (1)$$

3.2. Hierarchical Audio Learner

Hierarchical Audio Feature Extraction. In most previous studies [2, 25, 26, 41, 53, 68], only high-level audio features are extracted to guide the synthesis of desired movements. However, it has been discussed that different semantics in audios contribute to different granularities in the movements of human poses [46,48], which has been mostly ignored in previous works. We identify that such multi-level audio information could be inferred from the hierarchy of an audio encoder E_a to improve the generation resolution. Notably, the rich semantics of hierarchical feature maps embedded at different layers of a deep neural network have been explored in other deep learning tasks [43, 54, 63]. Therefore, the output deep feature $\boldsymbol{f}_a^{\text{high}}$ of E_a, the feature $\boldsymbol{f}_a^{\text{mid}}$ encoded in the middle of the audio encoder and the feature $\boldsymbol{f}_a^{\text{low}}$ encoded in the shallow of E_a are specifically leveraged. We expect $\boldsymbol{f}_a^{\text{low}}, \boldsymbol{f}_a^{\text{mid}}, \boldsymbol{f}_a^{\text{high}}$ to represent the low, middle and high level audio features respectively, as shown in blue block of Fig. 1. These hierarchical features are used for inferring poses in Sec. 3.3.

Contrastive Learning Strategy. Though we expect the audio features can be learned automatically given the property of the encoder, additional text can further enforce the embedding of our desired information. Transcripts, which represent high-level linguistic information, can be directly recognized by Automatic Speech Recognition (ASR) models [30, 49, 70] from speech. Thus we propose to learn the association between provided transcripts and audios in a simple yet effective manner with contrastive learning. Our strategy is to leverage the natural synchronization between text and audio. While the high-level audio features should reflect the temporally-aligned transcripts, text can in turn encourage mid- and low-level audio features to capture crucial speech content-irrelevant information such as tone and cadence.

Specifically, we denote the feature extracted by the text encoder from transcript $\mathbf{t}$ as $\boldsymbol{f}_t = E_t(\mathbf{t})$. In our contrastive learning formulation, the high-level audio features aligned to the transcript serve as positive examples, which are denoted as $\boldsymbol{f}_{a+}^{\text{high}}$. Then we design two types of negative samples: (1) Firstly, high-level features extracted at other time steps, or from other clips are selected as negative samples to enforce the high-level audio feature capture correct semantic information from the aligned text; (2) Secondly, the low/mid-level audio features are expected to be discriminative to reflect other audio information rather than high-level semantics. Therefore, we enforce them all to repel the text feature. With the similarity function defined as $sim(f_1, f_2) = \frac{\boldsymbol{f}_1 \cdot \boldsymbol{f}_2}{|\boldsymbol{f}_1||\boldsymbol{f}_2|}$, we can compute the final multi-level contrastive loss as:

$$\mathcal{L}_{\text{multi}} = -\log \frac{\exp(sim(\boldsymbol{f}_t, \boldsymbol{f}_{a+}^{\text{high}})/\tau)}{\sum_{i=1}^{K} \sum_{l \in L} \exp(sim(\boldsymbol{f}_t, \boldsymbol{f}_{a(i)}^{l})/\tau)}, \quad (2)$$

where $L = \{\text{low}, \text{mid}, \text{high}\}$, and $\boldsymbol{f}_{a(i)}^{\text{low}}, \boldsymbol{f}_{a(i)}^{\text{mid}}, \boldsymbol{f}_{a(i)}^{\text{high}}$ denote the i-th sample of low/mid/high-level audio feature, respectively. K is the number of samples and τ is the temperature parameter that controls the concentration of distribution.

3.3. Hierarchical Pose Inferer

As discussed in Sec. 1, different levels of audio features contribute to different hierarchies of human poses. Thus we propose to hierarchically infer gestures for more delicate audio-based control. To this end, we detach the joints from human body ends (fingers) to the main structure (spine) in H stages as illustrated in Fig. 1 (right). However, two questions still remain: 1) How to associate multiple levels of audios with different levels of joints; 2) How to supervise coarse-to-fine generation process.

Multi-Level Feature Blending with Style Coordinator. Our solution to the first question is to learn automatic feature blending schemes for different levels depending on a person-related style coordinator. As human gestures corresponding to the same speech are diverse across persons, the idea of learning person-specific styles has been adopted in various audio-driven animation tasks [2, 68]. In this work, the style coordinator should be responsible for finding the suitable ratio among hierarchical audio features that contributes to each level of motion hierarchy.

Different from [68] that uses one-hot labels to represent identities, we leverage a more general form by learning from the appearances of reference frames. The encoder E_{ID} is used to extract identity feature from a few frames, $\boldsymbol{f}_{id} = E_{\text{ID}}(I_1, \ldots, I_M)$. Then through a linear layer and softmax function, $\boldsymbol{f}_{id}$ is transformed into the style coordinator $C \in \mathbb{R}^{3 \times H}$, where $\sum_{i=1}^{3} C[i, h] = 1$. In this way, we can associate multi-level audio features with hierarchical body parts by linear blending:

$$\boldsymbol{f}_a^h = C[1, h] \cdot \boldsymbol{f}_a^{\text{low}} + C[2, h] \cdot \boldsymbol{f}_a^{\text{mid}} + C[3, h] \cdot \boldsymbol{f}_a^{\text{high}}, \quad (3)$$

where $\boldsymbol{f}_a^h$ denotes the blended audio feature for the h-th motion hierarchy. The procedure is illustrated in the middle of Fig. 1. To further facilitate style sampling at the inference stage, the Kullback–Leibler (KL) divergence loss $\mathcal{L}_{\text{KLD}}$ between the feature space of $\boldsymbol{f}_{id}$ and $\mathcal{N}(0, \text{I})$ is adopted to assume Gaussian style embedding distribution.

Coarse-to-Fine Pose Generation. We follow the human body dynamic rules to design a H-level (H = 6) body hierarchy (Fig. 1 right). At each level, the generation is affected by both the inferred pose from the previous level and the current level's audio feature rendered by the style coordinator. Such an idea is also similar to previous coarse-to-fine network designs [50].

In particular, we leverage the bi-directional GRU as motion decoder since the recurrent structure effectively captures spatial-temporal dependency in human motion as proved in [40, 64]. With the hierarchical audio feature of the h-th level $\boldsymbol{f}_a^h = \{\boldsymbol{f}_{a(1)}^h, \ldots, \boldsymbol{f}_{a(N)}^h\}$, the h-th level co-speech gesture $\hat{\mathbf{p}}^h = \{\hat{\boldsymbol{p}}_1^h, ..., \hat{\boldsymbol{p}}_N^h\}$ is generated by:

$$\hat{\boldsymbol{p}}_i^h = [\mathbf{h}_i; \hat{\boldsymbol{p}}_i^{h-1}; \boldsymbol{f}_{a(i)}^h] * W^h + \boldsymbol{b}^h, \mathbf{h}_i = \text{GRU}(\mathbf{h}_{i-1}, \hat{\boldsymbol{p}}_{i-1}^h), \quad (4)$$

where $\mathbf{h}_i$ is the i-th hidden state, $[\cdot; \cdot]$ is the concatenation operation and $*$ is the matrix multiplication. $W^h \in \mathbb{R}^{(d_s + d_p^{h-1} + d_a) \times d_p^h}$ and $\boldsymbol{b}^h \in \mathbb{R}^{d_p^h}$ are parameters where d_s, d_a and d_p^h are the dimensions of hidden state, audio feature and the h-th level pose $\hat{\mathbf{p}}^h$, respectively. Note that the poses of the first M frames serve as initial poses and are denoted as $\hat{\mathbf{p}}^0 = \{\boldsymbol{p}_1^0, ..., \boldsymbol{p}_M^0, 0, ..., 0\}$. In this way, fine-grained correspondences between audio sequence and co-speech gestures are jointly built in a coarse-to-fine manner. The last layer's output $\hat{\mathbf{p}}^H$ from the hierarchy is our desired result. This procedure is depicted in the right part of Fig. 1.

3.4. Training Objectives for Gesture Generation

Reconstruction Huber Loss. The generation process is constrained via a hierarchical Huber loss [34] by measuring the distances between generated samples $\hat{\boldsymbol{p}}_i^h$ and ground truth $\boldsymbol{p}_i^h$:

$$\mathcal{L}_{\text{huber}} = \mathbb{E}\left[\frac{1}{HN} \sum_{h=1}^{H} \sum_{i=1}^{N} \text{HuberLoss}(\boldsymbol{p}_i^h, \hat{\boldsymbol{p}}_i^h)\right], \quad (5)$$

where H is the number of motion hierarchy and N is the length of gesture sequence. We feed the blended audio feature to cascaded bi-GRU as generator G and leverage an adversarial loss for preserving realism following [25, 68]:

$$\begin{aligned} \mathcal{L}_{\text{GAN}} = &\min_{G} \max_{D} \mathbb{E}_{\mathbf{p}} [\log D(\mathbf{p})] \\ &+ \mathbb{E}_{\mathbf{a}} [\log(1 - D(G(E_a(\mathbf{a}) | \boldsymbol{f}_{id}, \boldsymbol{p}_{1:M})))] . \end{aligned} \quad (6)$$

Style Diverging Loss. To further avoid posterior collapse on speaker identity $\boldsymbol{f}_{id}$, we guide the generator to synthesize different poses with diverse style input following [68]. Assuming that $\hat{\mathbf{p}}(\boldsymbol{f}_{id})$ is the predicted pose depending on identity feature $\boldsymbol{f}_{id}$, we have:

$$\mathcal{L}_{\text{style}} = -\mathbb{E}\left[\min\left(\frac{\text{HuberLoss}(\hat{\mathbf{p}}(\boldsymbol{f}_{id(1)}), \hat{\mathbf{p}}(\boldsymbol{f}_{id(2)}))}{\|\boldsymbol{f}_{id(1)} - \boldsymbol{f}_{id(2)}\|_1}, \epsilon\right)\right], \quad (7)$$

where $\boldsymbol{f}_{id(1)}$, $\boldsymbol{f}_{id(2)}$ are two different speaker identities and ϵ is the numerical clipping parameter.

Physical Constraint. Previous methods on co-speech gesture generation mostly fail to consider human physical constraint, which leads to unnatural poses and incoherent results. Therefore, we propose to add restrictions on the included angle between bones to ensure reasonable human poses. Concretely, the pose is represented as directional vectors, thus the angle between consecutive bone vectors must obey physical rules. We specifically calculate the mean and variance of each angle within TED-Expressive dataset, and expect our generated ones to fall within such a Gaussian distribution. The loss function for the physics constraint is the log-likelihood function:

$$\mathcal{L}_{\text{phy}} = -\sum_{j=1}^{J-1} \log \mathcal{N}(\theta_j; \mu_j, \sigma_j^2) \tag{8}$$

where θ_j is the j-th bone angle value, μ_j and σ_j^2 are the mean and variance of the j-th angle, respectively.

The overall learning objective for the whole framework is as follows:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GAN}} + \lambda_h \mathcal{L}_{\text{huber}} + \lambda_p \mathcal{L}_{\text{phy}} \\ + \lambda_s \mathcal{L}_{\text{style}} + \lambda_k \mathcal{L}_{\text{KLD}} + \lambda_c \mathcal{L}_{\text{multi}},\end{aligned} \tag{9}$$

where the $\lambda_h, \lambda_p, \lambda_s, \lambda_k, \lambda_c$ are weight coefficients. At the training stage, the hierarchical audio encoder E_a, text encoder E_t, speaker identity encoder E_{ID} and hierarchical pose decoder are trained with back-propagation from the above overall loss function.

4. Experiments

At the inference stage, we use speech audio as guidance while text is not needed. We further extract initial poses and speaker identity from a few reference images. If the reference image is unavailable, we can sample initial poses from dataset and sample speaker identity from normal distribution to generate co-speech gestures since we constrain identity space with $\mathcal{L}_{\text{KLD}}$. In this way, we can generate diverse gestures with multiple styles by sampling style vectors.

4.1. Datasets and Annotation[1]

TED Gesture. TED Gesture dataset [68,69] is a large-scale English-language dataset for speech-driven motion synthesis, which contains 1,766 TED videos of different narrators covering various topics. The extracted 3D human skeletons, aligned English transcripts and speech audio are all available. Following [68], we resample human poses with 15 FPS and sample the consecutive 34 frames with stride of 10 frames as input segments. We finally get 252,109 segments with length of 106.1h. In this dataset, human pose $\boldsymbol{p}$ is represented by direction vectors of 10 upper body joints.

TED-Expressive. The pose annotations of TED Gesture limit to 10 upper body keypoints without expressive co-speech finger movements. Hence, to harvest more detailed pose annotation as training data, we use the state-of-art 3D pose estimator ExPose [16] to extract 3D human skeleton as pseudo ground truth. In particular, we first annotate the 3D coordinates of 43 keypoints, including 13 upper body joints and 30 finger joints. Then we convert 3D coordinates into 42 unit direction vectors following [68] to represent each bone for eliminating the influence of various bone lengths in training data. In this way, our 3D representation is invariant to root joint motion and body shape. At the inference stage, the mean bone length over dataset is multiplied to the predicted bone vectors for visualized results.

4.2. Experimental Settings

Baselines. We compare our method with : (1) **Attention Seq2Seq** [69] which generates gestures from speech text by attention mechanism; (2) **Speech2Gesture** [25] that takes the whole-length audio spectrogram as input and generates motion sequence with an encoder-decoder architecture and adversarial training scheme; (3) **Joint Embedding** [3], a representative method that maps the text and motion to the same embedding space and creates motion from description text; (4) **Trimodal** [68], the state-of-art method that considers the trimodal context of text, audio and speaker identity to learn co-speech gestures. Note that some recent works [41,53] lack open-source codes so far, thus we do not compare with them. All works are trained on the TED Gesture and TED-Expressive datasets for the same number of epochs with hyper-parameters optimized by grid search for best evaluation results. We also show the evaluation directly on the pseudo **Ground Truth** annotated in the dataset.

Implementation Details.[1] Following the settings of [68], we set $N = 34$ and $M = 4$, so that the data are segmented into 34-frame sequences and the first 4 frames serve as reference frames. The number of joint J is 10 for TED Gesture dataset and 43 for TED-Expressive dataset as mentioned in Sec. 4.1. The audio encoder backbone is a ResNetSE34 [17] and the structure of text encoder E_t is borrowed from [5]. The reference video frames are resized into 224×224, then passed into the speaker identity encoder E_{ID} with visual backbone of ResNet-18 [28] to extract speaker identity. The raw audios are converted to mel-spectrograms with FFT window size 1024, hop length 512. The word sequence is inserted with padding tokens to align with gestures. For each frame, 16 padded words and 0.25s mel-spectrogram with the target frame time-step in the middle are sampled as condition. The pose decoder is a cascaded 4-layer bi-directional GRU with a hidden size d_s of 300 for each level of pose hierarchy. Empirically, we set $\tau = 0.07$, $\epsilon = 1000$, $d_a = 32$, $\lambda_h = 200$, $\lambda_p = 0.1$, $\lambda_s = 0.05$, $\lambda_k = 0.1$, $\lambda_c = 0.1$. The models are trained using Adam Optimizer with the learning rate of $1e-4$ on 1 GTX 1080Ti GPU.

[1] Please refer to Supplementary Material for more details.

	TED Gesture [68, 69]			TED-Expressive		
Methods	FGD ↓	BC ↑	Diversity ↑	FGD ↓	BC ↑	Diversity ↑
Ground Truth	0	0.795	110.821	0	0.723	175.231
Attention Seq2Seq [69]	18.154	0.186	92.176	54.920	0.155	122.693
Speech2Gesture [25]	19.254	0.764	98.095	54.650	0.714	142.489
Joint Embedding [3]	22.083	0.177	91.223	64.555	0.131	120.627
Trimodal [68]	3.729	0.688	102.539	12.613	0.592	154.088
HA2G (Ours)	**3.072**	**0.769**	**108.086**	**5.306**	**0.715**	**173.899**

Table 1. **The quantitative results on TED Gesture [68, 69] and TED-Expressive.** We compare the proposed Hierarchical Audio-to-Gesture (**HA2G**) against recent SOTA methods [3, 25, 68, 69] and ground truth under three metrics. For FGD the lower the better, and the higher the better for other metrics. Note that the FGD results of [3, 25, 68, 69] on TED Gesture are reported from [68].

4.3. Quantitative Evaluation

Evaluation Metrics. We take the evaluation metrics that have been previously used in the co-speech gesture generation and music2dance for quantitative analysis.

Fréchet Gesture Distance (FGD) is used in [68] to measure how close the distribution of generated gesture is to the real one. Note that for the evaluation on TED Gesture dataset, we use the feature extractor provided in [68] for fair comparison. For the TED-Expressive dataset, we similarly train an auto-encoder on the TED-Expressive dataset and take the encoder part for feature extraction. FGD is calculated as the fréchet distance between the latent representations of real gesture and generated gesture.

Beat Consistency Score (BC) is a metric for motion-audio beat correlation as proposed in [39, 42]. However, since the kinematic velocities vary from different joints, we propose to use the change of included angle between bones to track motion beats. Concretely, we calculate the mean absolute angle change (MAAC) of angle θ_j in adjacent frames by:

$$\mathrm{MAAC}(\theta_j) = \frac{\sum_{s=1}^{S} \sum_{t=1}^{T-1} \|\theta_{j,s,t+1} - \theta_{j,s,t}\|_1}{S * (T-1)}, \quad (10)$$

where S is the total number of clips over dataset, T is the number of frames for a clip and $\theta_{j,s,t}$ is included angle between the j-th and the (j+1)-th bone of the s-th clip at timestep t. In this way, the angle change rate of frame t for the s-th clip is $\frac{1}{J-1}\sum_{j=1}^{J-1}(\|\theta_{j,s,t+1} - \theta_{j,s,t}\|_1 / \mathrm{MAAC}(\theta_j))$. Then we extract the local optima whose first-order difference is higher than a threshold[1] to get kinematic beats. We follow [39] to detect audio beat by onset strength [18] and compute the average distance between every audio beat and its nearest motion beat as Beat Consistency Score:

$$\mathrm{BC} = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\min_{\forall t_j^x \in B^x} \|t_i^x - t_j^y\|^2}{2\sigma^2}\right), \quad (11)$$

where $B^x = \{t_i^x\}$ are the kinematic beats, $B^y = \{t_j^y\}$ are the audio beats and σ is a parameter to normalize sequences that is empirically set to 0.1 for experiments.

Diversity evaluates the variations among generated gestures corresponding to various inputs [38]. Similarly, we use the same feature extractor in measuring FGD to map synthesized gestures into latent feature vectors and calculate the average feature distance for evaluation. Concretely, we randomly sample 60 speech audios from the test set to generate co-speech gestures and compute the average feature distance between 500 random combinated pairs.

Evaluation Results. The results are shown in Table 1. We can see that our **HA2G** framework outperforms existing methods on both datasets. Since our method establishes motion hierarchy and generates gestures in a coarse-to-fine manner, we can learn the diverse motion pattern of different human body parts and perform the best on FGD metric. Note that the improvement of FGD is smaller on TED Gesture dataset compared to TED-Expressive. This is due to the absence of finger information in TED Gesture dataset, which makes the motion hierarchy lower and the improvement brought by our hierarchical framework less significant. We can find that both Speech2Gesture [25] and ours synthesize synchronous gestures to speech with high values on BC. But they tend to create unnatural poses and hence perform fair on FGD. In terms of Diversity, the discriminative feature extraction at multiple granularities enables us to excavate fine-grained audio-pose associations, thus capturing diverse speaking styles compared to baseline methods.

4.4. Qualitative Evaluation

Subjective evaluation is crucial for judging the quality of results in generation tasks. Here we show the key frames comparison of our method against ground truth and SOTA baselines (as listed in Sec. 4.2) in Fig. 2. For two cases, both Attention Seq2Seq [69] and Joint Embedding [3] generate slow and invariant motions that are misaligned to speech as demonstrated in red rectangles of Fig. 2. While Trimodal [68] generates diverse gestures, the rigid motion pattern makes them mismatch to audio beats. For example, they stiffly move hands up and down with asynchronous beats to speech audio (see the red rectangle on the

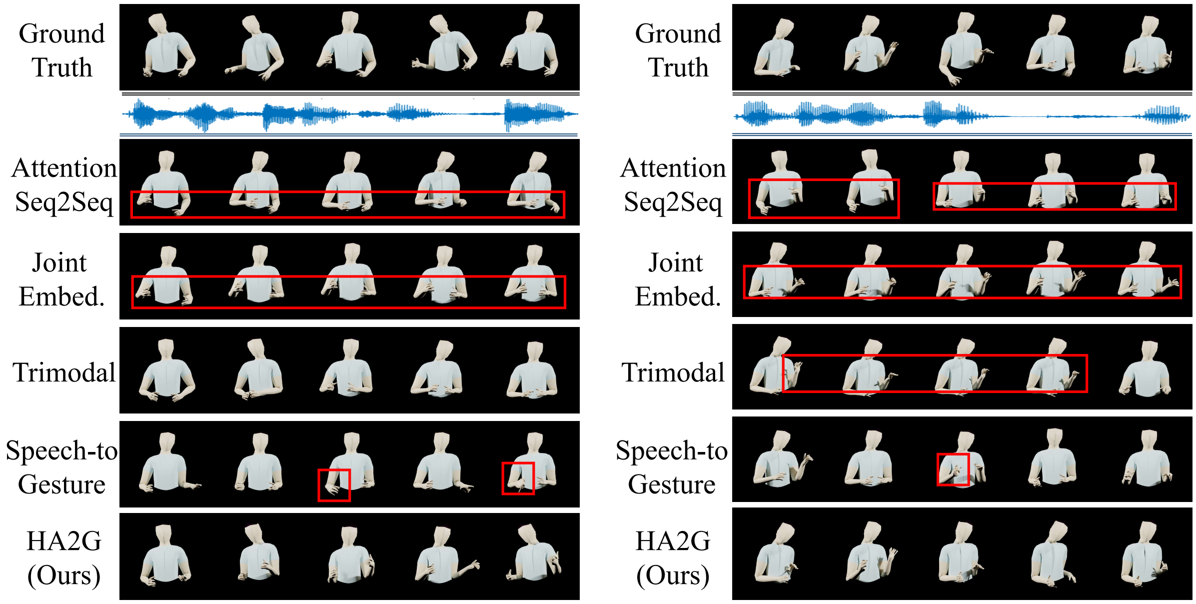

Figure 2. **The visualized results in two example clips.** We show the key frames of the generated motions from ground truth and baseline methods [3, 25, 68, 69]. Please **zoom in for better visualization**. More high-resolution results can be found in the demo video.

Methods	GT	Seq2Seq [69]	Joint. [3]	Tri. [68]	S2G. [25]	**HA2G (Ours)**
Naturalness	4.16	1.36	1.52	3.66	2.88	**4.13**
Smoothness	3.97	**4.48**	4.32	3.87	2.23	3.92
Synchrony	4.28	1.24	1.18	3.21	3.89	**4.06**

Table 2. **User study results on motion naturalness, smoothness and synchrony.** The rating is on a scale of 1-5, with the larger the better.

right). Both our method and Speech2Gesture [25] create synchronous motions, but they synthesize unnatural poses, *e.g.*, the twisted hands in both cases as highlighted in Fig. 2. The hierarchical cross-modal association against single-level design also leads to more diverse results than [25].

User Study.[2] We conduct a user study on motion naturalness, smoothness and the generated co-speech gestures' synchrony to speech. In particular, we randomly sample 20 speech clips from test set of TED-Expressive to generate results for ground truth (tracked) annotations, baselines and our method. The study involves 24 participants. We adopt the widely-used Mean Opinion Scores (MOS) rating protocol, which requires the participants to rate three aspects of generated motions: (1) *Naturalness*; (2) *Smoothness*; (3) *Synchrony between speech and generated gestures*. The rating is based on a scale of 1 to 5, with 5 being the most plausible and 1 being the least plausible.

The results are shown in Table 2. Since both Attention Seq2Seq [69] and JointEmbedding [3] generate slow and near-stationary results, they score reasonably low on naturalness and synchrony, and trivially perform well on smoothness, which is even better than ground truth due to the motion jitter in ExPose annotation. Although Speech2Gesture [25] performs well on synchrony, unnatural poses lead to fair results on naturalness and smoothness. Moreover, as our hierarchical design can capture fine-grained associations between multi-level features and diverse body parts, we score better than Trimodal [68] on all three aspects, with comparable results against ground truth. Note that to measure the disagreement on scoring among the participants, we also calculate the Fleiss's-Kappa[3] statistic on 24 participants' ratings over all methods. The Fleiss-Kappa value is 0.837, which is comparatively high and can be interpreted as "almost perfect agreement".

4.5. Ablation Study

In this section, we present ablation studies on two key modules proposed in our framework. We report the results implemented on the TED-Expressive dataset.

Hierarchical Audio Learner. To show the effect of multi-level audio feature in generating co-speech gesture, we conduct experiments on our model (1) $\boldsymbol{f}_a^{\text{low}}$ only, which means we only use low-level feature from hierarchical audio encoder, *i.e.*, the weight for low-level is set as 1 and weights for mid/high level features are set as 0 in Eq. 3; (2) $\boldsymbol{f}_a^{\text{mid}}$ only; (3) $\boldsymbol{f}_a^{\text{high}}$ only; (4) w/o $\boldsymbol{f}_{a-}^{\text{high}}$, which means we do not involve high level audio negative samples mentioned in

[2] Please refer to Supple. for more details about user study.

[3] https://en.wikipedia.org/wiki/Fleiss%27_kappa

Methods	FGD ↓	BC ↑	Diversity ↑
f_a^{low} only	6.588	0.704	171.482
f_a^{mid} only	7.212	0.682	168.223
f_a^{high} only	7.421	0.661	165.741
HA2G w/o f_{a-}^{high}	7.982	0.652	163.649
HA2G w/o $f_{a-}^{\mathrm{low,mid}}$	6.998	0.701	169.021
HA2G w/o text	9.228	0.619	158.236
HA2G-ASR	5.319	**0.716**	173.058
HA2G Full	**5.306**	0.715	**173.899**

Table 3. **Ablation study results of Hierarchical Audio Learner.**

Methods	FGD ↓	BC ↑	Diversity ↑
Holistic	11.989	0.594	156.079
w/o hand hierarchy	10.832	0.606	158.823
w/o body hierarchy	5.882	0.709	173.066
Same audio f_a^h	6.801	0.701	170.085
w/o $\mathcal{L}_{\mathrm{phy}}$	5.907	0.708	172.651
HA2G Full	**5.306**	**0.715**	**173.899**

Table 4. **Ablation study results of Hierarchical Pose Inferer.**

Sec. 3.2 for contrastive learning; (5) w/o $f_{a-}^{\mathrm{low,mid}}$, which states the situation without cross-level negative samples; (6) w/o text, in this setting the input of speech text is not used, so we do not use the contrastive loss $\mathcal{L}_{\mathrm{multi}}$ for audio-text alignment and discriminative audio feature extraction. The results are shown in Table 3, which indicates the efficacy of Hierarchical Audio Learner. Concretely, the only use of single-level audio feature fails to excavate information at multiple granularities, thus leading to degradation in performance. Besides, the contrastive learning strategy further improves performance since it achieves discriminative audio feature extraction with the self-supervision of audio-text alignment. More importantly, we find that our method **without** text outperforms Yoon *et al.* [68] **with** the input of text. This demonstrates that the hierarchical design and coarse-to-fine generation manner can synthesize gestures of higher quality despite lack of text, enabling our method to handle general scenarios where video transcripts are unavailable.

Another ablation study relates to the Hierarchical Audio Learner is why we adopt contrastive learning strategy for discriminative feature extraction. We take inspiration from the fact that ASR models can semantically align text and audios, thus multi-level semantic information can be extracted from audio itself. However, the amount of data provided in the dataset is insufficient for training an expert ASR model, which leads to our choice of hierarchical contrastive design. For the ablation experiment, we use a well-trained ASR model [65] as the audio encoder and generate co-speech gestures without contrastive strategy. The low, middle and high level features are also extracted from the backbone in a similar way as our method. We denote this variant of HA2G as HA2G-ASR. The comparisons on the TED-Expressive dataset are shown in the Table 3. We can notice that the prior knowledge of pretrained ASR network prevents outlier predictions, which achieves competitive results compared to ours. This illustrates that using different levels of ASR features will benefit gesture generation. Note that the pretrained ASR network is trained on **a large amount of additional data**, while HA2G is trained with just a multi-level contrastive loss **without involving other pretrained networks and additional data**.

Hierarchical Pose Inferer. The experiments of Hierarchical Pose Inferer on our model contain: (1) Holistic, which means we do not use pose hierarchy and directly generate whole-body pose like previous methods [3, 25, 68, 69]; (2) w/o hand hierarchy, where the hand poses are generated holistically while body hierarchy remains; (3) w/o body hierarchy, where body poses are generated holistically while hand hierarchy remains; (4) Same audio f_a^h, which means we pass identical hierarchical audio features to each level of motion hierarchy, *i.e.*, all columns of style coordinator C are same in Eq. 3; (5) w/o $\mathcal{L}_{\mathrm{phy}}$. Table 4 shows the results, which verify that Hierarchical Pose Inferer improves the performance. The pose hierarchy and distinct audio feature of each level enable the model to grasp fine-grained audio-pose associations of different body parts, making generated pose more vivid. The physical regularization $\mathcal{L}_{\mathrm{phy}}$ enhances FGD with more realistic human poses. Note that w/o body hierarchy outperforms w/o hand hierarchy. This is reasonable since the hand motion is more subtle, so hierarchical architecture's impact on hand is more significant.

5. Discussion

Conclusion. In this paper, we propose a novel framework Hierarchical Audio-to-Gesture (**HA2G**) for co-speech gesture generation. We introduce Hierarchical Audio Learner with a contrastive learning strategy that extracts discriminative audio representations across semantic granularities. Then we propose Hierarchical Pose Inferer with a physical regularization to render the entire human pose gradually in a hierarchical manner. Extensive experiments demonstrate the superior performance of our proposed approach on co-speech gesture generation with high fidelity.

Limitation. From the dataset perspective, our model is trained on an English-based corpus, which brings inductive bias on language. How to build a versatile model to generate co-speech gesture of diverse languages is a worthy direction for the community to explore.

Acknowledgments This work has been supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Fund, the RIE2020 Industry Alignment Fund–Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. No gestures left behind: Learning relationships between spoken language and freeform gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1884–1895, 2020. 1, 2

[2] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020. 1, 2, 3, 4

[3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 1, 2, 5, 6, 7, 8

[4] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019. 2

[5] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018. 5

[6] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018. 2

[7] Neeraj Battan, Yudhik Agrawal, Sai Soorya Rao, Aman Goel, and Avinash Sharma. Glocalnet: Class-aware long-term human motion synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 879–888, 2021. 2

[8] 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 Virtual Reality and 3D User Interfaces (VR)*, pages 1–10. IEEE, 2021. 1, 2

[9] Richard Bowden. Learning statistical models of human motion. In *IEEE Workshop on Human Modeling, Analysis and Synthesis, CVPR*, volume 2000. Citeseer, 2000. 2

[10] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192, 2000. 2

[11] Justine Cassell, David McNeill, and Karl-Erik McCullough. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & cognition*, 7(1):1–34, 1999. 1

[12] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420, 1994. 1, 2

[13] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *Life-Like Characters*, pages 163–185. Springer, 2004. 1

[14] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. What comprises a good talking-head video generation?: A survey and benchmark. *arXiv preprint arXiv:2005.03201*, 2020. 2

[15] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7832–7841, 2019. 2

[16] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. 5

[17] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. In defence of metric learning for speaker recognition. *arXiv preprint arXiv:2003.11982*, 2020. 5

[18] Daniel Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36:51–60, 03 2007. 6

[19] Ylva Ferstl, Michael Neff, and Rachel McDonnell. Adversarial gesture generation with realistic gesture phasing. *Computers & Graphics*, 89:117–130, 2020. 1, 2

[20] Aphrodite Galata, Neil Johnson, and David Hogg. Learning variable-length markov models of behavior. *Computer Vision and Image Understanding*, 81(3):398–413, 2001. 2

[21] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 2

[22] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7062, 2019. 2

[23] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 2

[24] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017. 2

[25] 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*, pages 3497–3506, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[26] 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. *arXiv preprint arXiv:2102.06837*, 2021. 1, 2, 3

[27] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. Evaluation of speech-to-gesture

generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 79–86, 2018. 1, 2

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[29] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7134–7143, 2019. 2

[30] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012. 3

[31] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 2016. 2

[32] Chien-Ming Huang and Bilge Mutlu. Robot behavior toolkit: generating effective social behaviors for robots. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 25–32. IEEE, 2012. 1

[33] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*, 2020. 2

[34] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992. 4

[35] Carlos T. Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. A speech-driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters*, 3(4):3757–3764, 2018. 1, 2

[36] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[37] 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. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 242–250, New York, NY, USA, 2020. Association for Computing Machinery. 2

[38] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 6

[39] Buyu Li, Yongchi Zhao, and Lu Sheng. Dancenet3d: Music based dance generation with parametric motion transformer. *arXiv preprint arXiv:2103.10206*, 2021. 2, 6

[40] C. Li, Z. Zhang, W. S. Lee, and G. H. Lee. Convolutional sequence to sequence model for human dynamics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 4

[41] 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. *arXiv preprint arXiv:2108.06720*, 2021. 1, 2, 3, 5

[42] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation. *arXiv preprint arXiv:2101.08779*, 2021. 2, 6

[43] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[44] Xian Liu, Rui Qian, Hang Zhou, Di Hu, Weiyao Lin, Ziwei Liu, Bolei Zhou, and Xiaowei Zhou. Visual sound localization in the wild by cross-modal interference erasing. *arXiv preprint arXiv:2202.06406*, 2022. 2

[45] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. *arXiv preprint arXiv:2201.07786*, 2022. 2

[46] Daniel P Loehr. Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory phonology*, 2012. 3

[47] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 25–35, 2013. 1, 2

[48] David McNeill. *Hand and mind*. De Gruyter Mouton, 2011. 1, 3

[49] Ara V Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2002(11):1–15, 2002. 3

[50] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 4

[51] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 2

[52] Katherine Pullen and Christoph Bregler. Animating by multi-level sampling. In *Proceedings Computer Animation 2000*, pages 36–42. IEEE, 2000. 2

[53] Shenhan Qian, Zhi Tu, YiHao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. *arXiv preprint arXiv:2108.08020*, 2021. 1, 2, 3, 5

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image com-*

puting and computer-assisted intervention, pages 234–241. Springer, 2015. 3

[55] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4(2):201–217, 2012. 1

[56] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. A friendly gesture: Investigating the effect of multimodal robot behavior in human-robot interaction. In *2011 Ro-Man*, pages 247–252. IEEE, 2011. 1

[57] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[58] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*. Springer, 2020. 2

[59] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2

[60] Susanne Van Mulken, Elisabeth Andre, and Jochen Müller. The persona effect: how substantial is it? In *People and computers XIII*, pages 53–66. Springer, 1998. 1

[61] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8639–8648, 2018. 2

[62] P. Wagner, Z. Malisz, and S. Kopp. Gesture and speech in interaction: An overview. *Speech Communication*, 57:209–232, 2014. 1

[63] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 3

[64] Mao Wei, Liu Miaomiao, Salzemann Mathieu, and Li Hongdong. Learning trajectory dependencies for human motion prediction. In *ICCV*, 2019. 4

[65] Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, and Pascale Fung. Lightweight and efficient end-to-end speech recognition using low-rank transformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6144–6148. IEEE, 2020. 8

[66] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021. 2

[67] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4394–4402, 2019. 2

[68] 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 (TOG)*, 39(6):1–16, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[69] 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 *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE, 2019. 1, 2, 5, 6, 7, 8

[70] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6984–6988. IEEE, 2020. 3

[71] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1735–1744, 2019. 2

[72] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 2

[73] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 2

[74] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[75] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[76] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 2

[77] Yi Zhou, Jingwan Lu, Connelly Barnes, Jimei Yang, Sitao Xiang, et al. Generative tweening: Long-term inbetweening of 3d human motions. *arXiv preprint arXiv:2005.08891*, 2020. 2