Rhythmic Gesticulator: Rhythm-Aware Co-Speech Gesture Synthesis with Hierarchical Neural Embeddings

TENGLONG AO, Peking University, China
QINGZHE GAO, Shandong University and Peking University, China
YUKE LOU, Peking University, China
BAOQUAN CHEN, SIST & KLMP (MOE), Peking University, China
LIBIN LIU*, SIST & KLMP (MOE), Peking University, China

Automatic synthesis of realistic co-speech gestures is an increasingly important yet challenging task in artificial embodied agent creation. Previous systems mainly focus on generating gestures in an end-to-end manner, which leads to difficulties in mining the clear rhythm and semantics due to the complex yet subtle harmony between speech and gestures. We present a novel co-speech gesture synthesis method that achieves convincing results both on the rhythm and semantics. For the rhythm, our system contains a robust rhythm-based segmentation pipeline to ensure the temporal coherence between the vocalization and gestures explicitly. For the gesture semantics, we devise a mechanism to effectively disentangle both low- and high-level neural embeddings of speech and motion based on linguistic theory. The high-level embedding corresponds to semantics, while the low-level embedding relates to subtle variations. Lastly, we build correspondence between the

Fig. 1. Gesture results automatically synthesized by our system for a beat-rich TED talk clip. The red words represent beats, and the red arrows indicate the movements of corresponding beat gestures.

hierarchical embeddings of the speech and the motion, resulting in rhythm- and semantics-aware gesture synthesis. Evaluations with existing objective metrics, a newly proposed rhythmic metric, and human feedback show that our method outperforms state-of-the-art systems by a clear margin.

CCS Concepts: • Computing methodologies → Animation; Natural language processing; Neural networks.

Additional Key Words and Phrases: non-verbal behavior, co-speech gesture synthesis, character animation, neural generative model, multi-modality, virtual agents

ACM Reference Format:
Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic Gesticulator: Rhythm-Aware Co-Speech Gesture Synthesis with Hierarchical Neural Embeddings. ACM Trans. Graph. 41, 6, Article 209 (December 2022), 19 pages. https://doi.org/10.1145/3550454.3555435

1 INTRODUCTION

Gesturing is an important part of speaking. It adds emphasis and clarity to a speech and conveys essential non-verbal information that makes the speech lively and persuasive [Burgoon et al. 1990]. There are rich demands for high-quality 3D gesture animation in many industries, such as games, films, and digital humans. However, the difficulties in reproducing the complex yet subtle harmony between vocalization and body movement make synthesizing natural-looking co-speech gestures remain a long-standing and challenging task.
Gestures are grouped into six categories by linguists [Ekman and Friesen 1969; McNeill 1992]—adaptors, emblems, deictics, icons, metaphors, and beats. Among them, the beat gestures are rhythmic movements that bear no apparent relation to speech semantics [Kipp 2004] but serve meta-narrative functions [McNeill 1992] that are crucial to rhythmic harmony between speech and gestures. Generating realistic beat gestures requires modeling the relation between the gestural beats and the verbal stresses. However, it has been observed that these two modalities are not synchronized in a strict rhythmic sense [McClace 1994], making it difficult to learn their temporal connection directly from data using an end-to-end method [Bhattacharya et al. 2021a; Kucherenko et al. 2020; Yoon et al. 2020].

Gestures are associated with different levels of speech information [McNeill 1992]. For example, an emblem gesture such as thumbs-up usually accompanies high-level semantics like good or great, while a beat gesture commonly comes with low-level acoustic emphasis. Many previous studies use only the features extracted at the last layer of an audio encoder to synthesize gestures [Alexanderson et al. 2020; Bhattacharya et al. 2021a; Kucherenko et al. 2020; Qian et al. 2021; Yoon et al. 2020]. This setup, however, may in effect encourage the encoder to mix the speech information from multiple levels into the same feature, causing ambiguity and increasing the difficulty in mining clear rhythmic and semantic cues.

In this paper, we focus on generating co-speech upper-body gestures that can accompany a broad range of speech content—from a single sentence to a public speech, aiming at achieving convincing results both on the rhythm and semantics. Our first observation is that gesture can be considered as a special form of dancing under changing beats. We develop a rhythm-based canonicalization and generation framework to deal with the challenge of generating synchronized gestures to the speech, which segments the speech into short clips at audio beats, normalizes these clips into canonical blocks of the same length, generates gestures for every block, and aligns the generated motion to the rhythm of the speech. This framework, which is partially inspired by recent research in dance generation [Aristidou et al. 2022], provides the gesture model with an explicit hint of the rhythm, allowing the model to learn the pattern of gestural beats within a rhythmic block efficiently. Both the quantitative evaluation with a novel rhythmic metric and the qualitative evaluation with user studies show that the gestures generated by this pipeline exhibit natural synchronization to the speech.

As indicated in linguistics literature [Kipp 2004; Neff et al. 2008; Webb 1996], gestures used in everyday conversation can be broken down into a limited number of semantic units with different motion variations. We assume that these semantic units, usually referred to as lexemes, relate to the high-level features of speech audio, while the motion variations are determined by the low-level audio features. We thus disentangle high- and low-level features from different layers of an audio encoder and learn the mappings between them and the gesture lexemes and the motion variations, respectively. Experiments demonstrate that this mechanism successfully disentangles multi-level features of both the speech and motion and synthesizes semantics-matching and stylized gestures.

In summary, our main contributions in this paper are:

- We present a novel rhythm- and semantics-aware co-speech gesture synthesis system that generates natural-looking gestures. To the best of our knowledge, this is the first neural system that explicitly models both the rhythmic and semantic relations between speech and gestures.
- We develop a robust rhythm-based segmentation pipeline to ensure the temporal coherence between speech and gestures, which we find is crucial to achieving rhythmic gestures.
- We devise an effective mechanism to relate the disentangled multi-level features of both speech and motion, which enables generating gestures with convincing semantics.

2 RELATED WORK

2.1 Data-driven Human Motion Synthesis

Traditional human motion synthesis frameworks often rely on concatenative approaches such as motion graph [Kovar et al. 2002]. Recently, learning-based methods with neural networks have been widely applied to this area to generate high-quality and interactive motions, using models ranging from feed-forward network [Holden et al. 2017; Starke et al. 2022] to dedicated generative models [Henter et al. 2020; Ling et al. 2020]. Dealing with the one-to-many issue where a variety of motions can correspond to the same input or control signal is often a challenge for these learning-based approaches. Previous systems often employ additional conditions, such as contacts [Starke et al. 2020] or phase indices [Holden et al. 2017; Starke et al. 2022], to deal with this problem. Closer to the gesture domain is the speech-driven head motion synthesis, where conditional GANs [Sadoughi and Busso 2018], and conditional VAEs [Greenwood et al. 2017] have been used.

2.1.1 Music-driven Dance Synthesis. Among the general motion synthesis tasks, music-driven dance generation addresses a similar problem to the co-speech gesture synthesis, where the complex temporal relation between two different modalities needs to be modeled accurately. Both motion graph-based methods [Chen et al. 2021; Kim et al. 2006] and learning-based approaches [Li et al. 2021b; Siyao et al. 2022; Valle-Pérez et al. 2021] have been adopted and successfully achieved impressive generation results. To deal with the synchronization between the dance and music, Chen et al. [2021] develop a manually labeled rhythm signature to represent beat patterns and ensures the rhythm signatures of the generated dance match the music. Aristidou et al. [2022] segment the dance into blocks at music onsets, convert each block into a motion motif [Aristidou et al. 2018] that defines a specific cluster of motions, and use the motion motif to guide the synthesis of dance at the block level. Siyao et al. [2022] employ a reinforcement learning scheme to improve the rhythmic performance of the generator using a reward function encouraging beat alignment. Our rhythm-based segmentation and canonicalization framework is partially inspired by [Aristidou et al. 2022]. Similar to [Aristidou et al. 2022], we also segment the gestures into clips at audio beats but learn a high-level representation for each clip via the vector quantization scheme [Oord et al. 2017] instead of the K-means clustering. Moreover, our framework generates gestures in blocks of motion and denormalizes the generated motion blocks to match the rhythm of the speech.
In contrast, Aristidou et al. [2022] synthesize dance sequences in frames conditioned on the corresponding motion motifs.

2.2 Co-speech Gesture Synthesis

The most primitive approach used to generate human non-verbal behaviors is to animate an artificial agent using the retargeted motion capture data. This kind of approach is widely used in commercial systems (e.g., films and games) because of its high-quality motion performance. However, it is not suitable for creating interactive content that cannot be prepared beforehand. Generating co-speech gestures according to an arbitrary input has been a long-standing research topic. Previous studies can be roughly categorized into two groups, i.e., rule-based and data-driven methods.

2.2.1 Rule-based Method. The idea of the rule-based approach is to collect a set of gesture units and design specific rules that map a speech to a sequence of gesture units [Casell et al. 2004; Huang and Mutlu 2012; Kipp 2004; Softbank 2018]. Wagner et al. [2014] have an excellent review of these methods. The results of the rule-based methods are generally highly explainable and controllable. However, the gesture units and rules typically have to be created manually, which can be costly and inefficient for complex systems.

2.2.2 Data-driven Method. Early research in data-driven method learns the rules embedded in data and combines them with predefined animation units to generate new gestures. For example, Kopp et al. [2006]; Levine et al. [2010] use probabilistic models to build correspondence between speech and gestures. Neff et al. [2008] build a statistical model to learn the personal style of each speaker. The model is combined with the input text tagged with the theme, utterance focus, and schema to generate gesture scripts, which are then mapped to a sequence of gestures selected from an animation lexicon. Chin et al. [2015] train a neural classification model to select a proper gesture unit based on the speech input. More recent research has started to take advantage of deep learning and trains end-to-end models using raw gesture data directly, which frees the manual efforts of designing the gesture lexicon and mapping rules. Gestures can be synthesized using deterministic models such as multi-layer perceptron (MLP) [Kucherenko et al. 2020], recurrent neural networks [Bhattacharya et al. 2021a; Hasegawa et al. 2018; Liu et al. 2022; Yoon et al. 2020, 2019], convolutional networks [Habibie et al. 2021], and transformers [Bhattacharya et al. 2021b], or by learning generative models such as normalizing flow [Alexanderson et al. 2020], VAEs [Li et al. 2021a; Xu et al. 2022], and learnable noise codes [Qian et al. 2021]. Our method is also a data-driven framework. We learn the motion generator and the mapping between the speech and gestures from data using a combined network structure of the vector quantized variational autoencoder (VQ-VAE) [Oord et al. 2017] and LSTM. To capture the rhythmic and semantic correspondences between the speech and gestures, we propose a multi-stage architecture that explicitly models the rhythm and semantics in different stages. An earlier system proposed by Kucherenko et al. [2021b] shares a similar high-level architectural design to our framework. However, there are two key differences: (a) our method is essentially an unsupervised learning approach, which learns the gesture lexeme, style code, and the generator directly from the data without detailed annotations; and (b) our system employs an explicit beat-based segmentation scheme which is shown to be effective in ensuring temporal coherence between the speech and the gesture.

2.3 Multi-Modal Data Processing

Co-speech gesture generation is a cross-modal process involving audio, text, motion, and other information related to the speaker and the content of the speech. The representation and alignment of each modality are essential for high-quality results [Baltrušaitis et al. 2019]. Mel-spectrogram and MFCC acoustic features are commonly used as audio features [Alexanderson et al. 2020; Kucherenko et al. 2020; Qian et al. 2021], typically resampled into the same framerate of the motion. For the text features, pre-trained language models like BERT [Devlin et al. 2019; Kucherenko et al. 2020] and FastText [Bojanowski et al. 2017; Yoon et al. 2020] have been used to encode text transcripts into frame-wise latent codes, where paddings, fillers, or empty words are inserted into a sentence to make the world sequence the same length as the motion [Kucherenko et al. 2020; Yoon et al. 2020]. Speaker’s style and emotions can also be encoded by learnable latent codes [Bhattacharya et al. 2021a; Yoon et al. 2020] and are resampled or padded to match the length of the speech. In this work, we employ a pre-trained speech model to extract audio features and fine-tune it using a contrastive learning strategy. We also utilize a BERT-based model to vectorize the text. These multimodal data are then aligned explicitly using the standard approaches discussed above. Notably, a concurrent study [Liu et al. 2022] also extracts audio features using contrastive learning. Their framework considers the learning of the audio features as a part of the training of the gesture generator. Instead, our framework trains the audio encoder in a separate pre-training stage using only the audio data.

2.4 Evaluation of Motion Synthesis Models

Evaluating the generated co-speech gestures is often difficult because the motion quality is a very subjective concept. Previous works have proposed several evaluation criteria. Wolfert et al. [2022] have made a comprehensive review of them. User studies are widely adopted to evaluate different aspects of motion quality, such as human-likeness and speech-gesture matching [Alexanderson et al. 2020; Kucherenko et al. 2020; Yoon et al. 2020], but can be expensive and hard to exclude uncontrolled factors. The absolute difference of joint positions or other motion features, such as velocity and acceleration between a reconstructed motion and the ground truth, is used by several works as an objective metric [Ginosar et al. 2019; Ginosar et al. 2019; Janowski et al. 2017; Yoon et al. 2020]. Speaker’s style and emotions can also be encoded by learnable latent codes [Bhattacharya et al. 2021a; Yoon et al. 2020] and are resampled or padded to match the length of the speech. In this work, we employ a pre-trained speech model to extract audio features and fine-tune it using a contrastive learning strategy. We also utilize a BERT-based model to vectorize the text. These multimodal data are then aligned explicitly using the standard approaches discussed above. Notably, a concurrent study [Liu et al. 2022] also extracts audio features using contrastive learning. Their framework considers the learning of the audio features as a part of the training of the gesture generator. Instead, our framework trains the audio encoder in a separate pre-training stage using only the audio data.

ACM Trans. Graph., Vol. 41, No. 6, Article 209. Publication date: December 2022.
Fig. 2. Our system is composed of three core components: (a) the data module preprocesses a speech, segments it into normalized blocks based on the beats, and extracts speech features from these blocks; (b) the training module learns a gesture lexicon from the normalized motion blocks and trains the generator to synthesize gesture sequences, conditioned on the gesture lexemes, the style codes, as well as the features of previous motion blocks and adjacent speech blocks; and (c) the inference module employs interpreters to transfer the speech features to gesture lexemes and style codes, which are then used by the learned generator to predict future gestures.

3 SYSTEM OVERVIEW

Our goal is to synthesize realistic co-speech upper-body gestures that match a given speech context both temporally and semantically. To achieve this goal, we build a system using neural networks that takes speech audio as input and generates gesture sequences accordingly. Additional speech modalities, such as text and speaker identity, will also be considered by the system when available to enhance semantic coherence and generate stylized gestures.

A gesture motion consists of a sequence of gesture units, which can be further broken down into a number of gesture phases that align with intonational units, such as pitch accents or stressed syllables [Kendon 2004; Loehr 2012]. The action in each of these gesture phases is typically a specific movement such as lifting a hand, holding an arm at a position, or moving both arms down together, which is often referred to as a gesture lexeme by linguists [Kipp 2004; Neff et al. 2008; Webb 1996]. It is also revealed in the literature that there are only a limited number of lexemes used in everyday conversation. These lexemes form a gesture lexicon. A typical speaker may only use a subset of this lexicon and apply slight variations to the motion. We assume such variations cannot be inferred directly from the speech but can be characterized by some latent variables, which we refer to as the gesture style codes. Our system then generates gestures in a hierarchical order. It first determines the sequence of gesture lexemes and style codes and then generates gestural moves based on these motion-related features and other speech modalities.

Our system processes the input speech in a block-wise manner. Considering the temporal and structural synchrony between the gesture and the speech, we leverage a segmentation that aligns with the rhythm of the speech to ensure temporal coherence between the two modalities. Specifically, our system extracts beats from the input speech based on audio onsets and segments the speech into groups, which is typically a specific movement such as lifting a hand, holding an arm at a position, or moving both arms down together, which is often referred to as a gesture lexeme by linguists [Kipp 2004; Neff et al. 2008; Webb 1996]. It is also revealed in the literature that there are only a limited number of lexemes used in everyday conversation. These lexemes form a gesture lexicon. A typical speaker may only use a subset of this lexicon and apply slight variations to the motion. We assume such variations cannot be inferred directly from the speech but can be characterized by some latent variables, which we refer to as the gesture style codes. Our system then generates gestures in a hierarchical order. It first determines the sequence of gesture lexemes and style codes and then generates gestural moves based on these motion-related features and other speech modalities.

Our system processes the input speech in a block-wise manner. Considering the temporal and structural synchrony between the gesture and the speech, we leverage a segmentation that aligns with the rhythm of the speech to ensure temporal coherence between the two modalities. Specifically, our system extracts beats from the input speech based on audio onsets and segments the speech into groups, which is typically a specific movement such as lifting a hand, holding an arm at a position, or moving both arms down together, which is often referred to as a gesture lexeme by linguists [Kipp 2004; Neff et al. 2008; Webb 1996]. It is also revealed in the literature that there are only a limited number of lexemes used in everyday conversation. These lexemes form a gesture lexicon. A typical speaker may only use a subset of this lexicon and apply slight variations to the motion. We assume such variations cannot be inferred directly from the speech but can be characterized by some latent variables, which we refer to as the gesture style codes. Our system then generates gestures in a hierarchical order. It first determines the sequence of gesture lexemes and style codes and then generates gestural moves based on these motion-related features and other speech modalities.

Our system processes the input speech in a block-wise manner. Considering the temporal and structural synchrony between the gesture and the speech, we leverage a segmentation that aligns with the rhythm of the speech to ensure temporal coherence between the two modalities. Specifically, our system extracts beats from the input speech based on audio onsets and segments the speech into groups, which is typically a specific movement such as lifting a hand, holding an arm at a position, or moving both arms down together, which is often referred to as a gesture lexeme by linguists [Kipp 2004; Neff et al. 2008; Webb 1996]. It is also revealed in the literature that there are only a limited number of lexemes used in everyday conversation. These lexemes form a gesture lexicon. A typical speaker may only use a subset of this lexicon and apply slight variations to the motion. We assume such variations cannot be inferred directly from the speech but can be characterized by some latent variables, which we refer to as the gesture style codes. Our system then generates gestures in a hierarchical order. It first determines the sequence of gesture lexemes and style codes and then generates gestural moves based on these motion-related features and other speech modalities.
We use \( m \) to denote a gesture pose at frame \( m \). The generated gesture motions are then denormalized to match the length of the input speech.

As illustrated in Figure 2, our system consists of three core components: (a) the data module preprocesses a speech, segments it into normalized blocks based on the beats, and extracts speech features from these blocks; (b) the training module learns a gesture lexicon from the normalized motion blocks and trains the generator to synthesize gesture sequences, conditioned on the gesture lexemes, the style codes, as well as the features of previous motion blocks and adjacent speech blocks; and (c) the inference module employs interpreters to transfer the speech features to gesture lexemes and style codes, which are then used by the learned generator to predict future gestures.

We train our system on a speech-gesture dataset with accompanying text and speaker identity (ID). The gesture lexicon is constructed using unsupervised learning based on a vector quantized autoencoder (VQ-VAE) [Oord et al. 2017]. The generator is trained as an autoregressive encoder-decoder network, where we use an LSTM-based decoder combined with a vector quantized encoder to generate gesture motions. We train two separate interpreters to translate speech features into gesture lexemes and style codes, respectively. These interpreters can work with only the audio features and can be retrained to accept other speech modalities. In the following sections, we will provide details about these components and how they are trained in our system.

4 DATA PREPARATION
The data module of our system preprocesses an input speech, segments it into uniform blocks based on speech rhythms, and extracts features that will be used to generate co-speech gestures. In this section, we first introduce the representations of different speech modalities and gesture motion and then describe details of the data preprocessing.

4.1 Representation of Speech Modalities
4.1.1 Motion Representation. We focus on upper-body gestures in this work. Our system employs a character model consisting of 16 upper-body joints, including a rotational root, as shown in Figure 3. A gesture pose is then represented as a list of joint rotations, parameterized using the exponential map, in the hierarchical order. We use \( m_k \in \mathbb{R}^{6m} \) to represent the gesture pose at frame \( k \), and a clip of gestures is represented collectively as \( M = \{m_1, \ldots, m_K\} \).

4.1.2 Text Representation. Text transcription is an important speech modality that provides high-level linguistic information in a compact format. It is typically given as a word sequence, where the number of words per unit time can vary depending on the speed of the speech. Following [Kucherenko et al. 2020], we align the words to the speech and convert the text into frame-level features to overcome this issue, which is done using an off-the-shelf text-speech alignment tool combined with a pre-trained language model.

Text-speech alignment is a standard technique in the field of speech synthesis. In our system, we employ Montreal Forced Aligner (MFA) [McAuliffe et al. 2017] for this task, which pinpoints the beginning and end frames of every word in the speech. MFA also identifies silences and represents them as empty words. Since a speaker typically stops gesticulating in a long silence [Graziano and Gullberg 2018], our system records those silences and uses them during training to reproduce such behaviors, as will be detailed later.

We then pass the text and the empty words into BERT [Devlin et al. 2019], a popular and powerful pre-trained language model, to extract a high-level representation of the text. BERT computes an encoding vector for each word in an input sentence, which is then repeated and used for all the frames that the word occupies. We represent these word vectors collectively as \( T = \{t_1, \ldots, t_K\} \) for a speech clip of \( K \) frames, where each \( t \in \mathbb{R}^{65} \).

4.1.3 Audio Representation. Many recent studies use deep encoders to extract audio features from raw audio signals or audio spectrograms, where only the features extracted at the last layer of the encoder are used to generate gestures [Alexanderson et al. 2020; Kucherenko et al. 2020; Li et al. 2021a; Qian et al. 2021; Yoon et al. 2020]. Such a configuration potentially encourages the encoder to
mix information from multiple levels into the same feature, which can be difficult to disentangle in the downstream generation tasks. In our system, we propose to decouple the multi-level audio features in the encoder and use them in different scenarios. We assume the high-level features correspond to the speech semantics that determines the gesture lexemes, while the low-level features relate to the other audio information and can be used to control the gesture styles. As shown in Figure 4, we employ a pre-trained speech model, vq-wav2vec [Baevski et al. 2020], to extract audio features from raw audio signals and fine-tune it using a contrastive learning strategy.

The encoder of vq-wav2vec has $L = 8$ convolutional layers. When taking a block of audio signals of $K$ frames, $A$, as input, this encoder produces a representation $a_k$ for each frame $k$ of the audio. In this computation, the outputs of every layer can be considered as a set of multi-level features $\{a^l_k\}$, $l = 1, \ldots, L$, and notably, $a^L_k = a_k$. We then encourage the highest-level feature $a^L_k$ to match the speech content and push apart the features of the lower levels $\{a^l_k\}$, $l < L$, to capture crucial content-irrelevant information. Specifically, we utilize the contrastive loss

$$L_{\text{cont}} = -\log \frac{\exp(\text{sim}(\tilde{t}_k, a^L_k)/\tau)}{\sum_{l=1}^{L} \sum_{k=1}^{K} \exp(\text{sim}(\tilde{t}_k, a^l_k)/\tau)}.$$

(1)

where the text feature $t_k$ is extracted from the speech transcription, the $\text{sim}(\cdot, \cdot)$ function computes the cosine similarity between two vectors as

$$\text{sim}(\tilde{t}, \hat{a}) = \frac{\tilde{t} \cdot \hat{a}}{||\tilde{t}|| ||\hat{a}||},$$

(2)

and $\tau$ is the temperature hyperparameter. All the feature vectors are projected into the same vector space using learnable linear projections $\tilde{t} = F_t(t)$ and $\hat{a}^l_k = F^l_a(a^l_k)$, $l = 1, \ldots, L$, respectively. Notably, we consider the highest-level audio feature of the current frame as the positive example and audio features of the other levels and the other frames as the negative examples in this contrastive learning process.

This contrastive learning strategy is partially inspired by the H2G2 model proposed by Liu et al. [2022]. However, unlike their approach, which considers contrastive learning as a part of the training of the gesture generator, we train the audio encoder in a separate pre-training stage using only the speech data. After the training, the features extracted at the second and the last layers of the encoder, represented by $a^{\text{low}} \in \mathbb{R}^{d_a}$ and $a^{\text{high}} \in \mathbb{R}^{d_a}$, respectively, are then used in different training and inference stages in the downstream generation task. They can be represented collectively as $A^{\text{low}}$ and $A^{\text{high}}$ for a speech clip. Although the gesture motions are not considered here, we find that the results of this encoder still demonstrate correlations between the high-level audio features and the gestures. We will discuss these results later in Section 7.4.

4.1.4 Identity Representation. Similar to previous studies [Bhattacharya et al. 2021a; Yoon et al. 2020], our system can leverage the speaker identity (ID) to help distinguish different gesture styles and achieve stylized gesture generation. We represent each speaker as a one-hot vector $I \in \{0, 1\}^N_I$, where $N_I$ is the number of speakers in a dataset.

![Fig. 5. Distribution of time intervals between consecutive audio onsets in two open-source speech-gesture datasets.](image-url)
4.2.2 Normalization. Our system then segments the speech into short clips at every beat. These clips are then time-scaled into uniform blocks of length $D_M$. The speech segments are segmented and time-scaled as well in this process. For the motion, $M$, and text representation, $T$, of a clip, we resample the corresponding features to match the new length. The audio is processed with additional care, where we use the TSM (Time-Scale Modification) algorithm to change the duration of the audio while preserving the pitch. The audio features $A^{low}$ and $A^{high}$ are then recomputed for the time-scaled audio blocks. The speaker ID $I$ is a constant of the whole speech, which will not be changed during the normalization.

5 GESTURE GENERATION

The generator module is the core component of our system. It synthesizes realistic gesture motions according to a sequence of gesture lexemes, the corresponding style codes, and the low-level features of the audio. In this section, we first introduce how we construct the gesture lexicon and then describe the design and training of the gesture generator.

5.1 Construction of Gesture Lexicon

As revealed in several pieces of literature in linguistics [Kipp 2004; Neff et al. 2008; Webb 1996], only a limited number of lexemes are used in everyday conversation. We assume that each lexeme corresponds to a specific motion category. Our goal is then to extract those motion categories from a large gesture dataset. To achieve this goal, we employ the vector quantized variational autoencoder (VQ-VAE) model [Oord et al. 2017] to learn a categorical representation of the motion and construct the gesture lexicon.

VQ-VAE has been widely used to learn categorical spaces in many successful temporal models [Baevski et al. 2020; Dhariwal et al. 2020; Ramesh et al. 2021; Yan et al. 2021]. Similar to a regular autoencoder, a VQ-VAE also has an encoder-decoder structure but quantizes the latent space using a discrete codebook. The codebook consists of a list of vectors and their associated indices. The output of the encoder network is compared to every vector in the codebook, where the vector that is the closest in Euclidean distance is considered to be the codebook, or the lexicon, and $s$ is a codebook vector, or a lexeme. The first term of Equation (3) penalizes the reconstruction error,

$$L_{\text{lexicon}} = \|M - D(s)\|_2^2$$

$$+ w_a \|E(M) - sg(s)\|_2^2 + w_{\beta}\|sg(E(M)) - s\|_2^2.$$  (3)

where $s = \text{arg \, min}_{s' \in S} \|s' - E(M)\|_2$,  (4)

$M$ is a normalized motion block, $E$ and $D$ represent the encoder and decoder, respectively, $sg$ stands for the stop gradient operator that prevents the gradient from backpropagating through it, $S$ represents the codebook, or the lexicon, and $s$ is a codebook vector, or a lexeme. The while the other two terms pull together the latent representation of motion $M$ and its corresponding lexeme. Notably, since $S$ is discrete, the arg min operator in Equation (4) does not generate a gradient. The gradient of the reconstruction error with respect to the latent code is passed unaltered to the encoder during the backward pass as suggested in [Oord et al. 2017].

We train the VQ-VAE on each dataset in a separate pre-training stage. The encoder is a multi-layer network consisting of four 1-D convolutional layers followed by a fully connected layer, which encodes a motion block into a vector $s \in \mathbb{R}^d_s$. The decoder is a mirror of the encoder structurally. The size of the lexicon is a hyperparameter, which is chosen empirically based on the size and complexity of the dataset. Figure 6 shows the t-SNE visualization of the training results on two speech-gesture datasets, along with sample gestures of several lexemes. Once learned, the gesture lexicon and the gesture lexeme of every motion block are fixed and used by the generator and interpreters in both the training and inference stages.

5.2 Architecture of Generator

As illustrated in Figure 2, the generator module of our system is an autoregressive encoder-decoder network, where a new motion block is conditioned on not only the input speech but also the preceding block, the gesture lexeme, and the style code. More specifically, the generation of a motion block $M_i$ can be formulated as

$$M_i = G(M_{i-1}, A_i, s_i, z_i, P),$$  (5)

where $M_{i-1}$ represents the features extracted from the preceding motion block, $A_i$ stands for the representation of the input audio, $s_i$ and $z_i$ are the gesture lexeme and style code of the new block, respectively. Note that we use an asterisk (*) to indicate a generated quantity. All these motion and feature blocks have the same length of $K$ frames, where $s_i$ and $z_i$ are repeated and stacked into the corresponding blocks, represented as $S_i$ and $Z_i$, respectively.

We extract the motion feature $M_{i-1} \in \mathbb{R}^{K \times d_M}$ as

$$\tilde{M}_{i-1} = E_M(M_{i-1}),$$  (6)

where the encoder $E_M$ is a 1-D convolutional network with three layers.

The audio feature $A_{i} \in \mathbb{R}^{K \times d_A}$ is computed using three consecutive audio blocks

$$\tilde{A}_i = E_A(A_{i-1}^{low}, A_{i-1}^{mid}, A_{i-1}^{high})$$  (7)

Fig. 6. t-SNE visualization of gesture lexicons. Each color stands for a gesture lexeme. (a) lexicon learned on the Trinity Gesture dataset with 50 lexemes. (b) lexicon learned on the TED Gesture dataset with 100 lexemes.
to allow the generator to prepare for the future gestures. Notably, the original duration of an audio block is characterized by the onset interval \([D_{\text{on}}, D_{\text{off}}]\), which is typically \([0.2s, 0.5s]\) in our experiments. Thus the temporal window of this encoder is roughly \([0.6s, 1.5s]\). Each \(A_{\text{low}}^{\text{low}}\) is the low-level feature pre-computed from the raw speech audio. We assume that \(A_{\text{low}}\) already captures necessary information and use a simple network consisting of one fully connected layer as the encoder \(E_A\).

In the training stage, the gesture lexeme \(s_i\) of each motion block is determined during the construction of the gesture lexicon, while the style code \(z_i\) is a learnable variable that will be trained along with the generator. In the inference stage, both \(s_i\) and \(z_i\) are provided by the interpreters, as will be discussed below.

In addition to these features, we include a positional encoding block, \(P \in \mathbb{R}^{K \times dp}\), to let the generator know the progress of the generation in a motion block, which is a standard component of transformers [Vaswani et al. 2017] and many sequence generation tasks [Harvey et al. 2020]. For our normalized blocks with \(K\) frames, we compute \(P = \{p_1, \ldots, p_K\}\) as

\[
P_{2k} = \sin \left( \frac{K}{2\beta k / dp} \right), \quad P_{2k+1} = \cos \left( \frac{K}{2\beta k / dp} \right),
\]

where \(dp\) is the dimension of the encoding and \(\beta = 10,000\) is a constant controlling the rate of change in frequencies along the embedding dimensions.

The generator \(G\) consists of an MLP-based encoder followed by an LSTM-based decoding network. Inspired by the successful systems in generating sequential output [Oord et al. 2017; Richard et al. 2021], we quantized the latent space of the encoder into \(C\) different configurations. We use \(H = 64\) and \(C = 128\) to ensure a large enough categorical space. In addition, we leverage Gumbel-softmax [Jang et al. 2017] to convert a latent code into a codebook vector, which can be viewed as a differentiable sampling operator for the discrete codebook search.

### 5.3 Training

We train the generator \(G\), the encoders \(E_M\) and \(E_A\), and the learnable style codes \(\{z_i\}\) by minimizing a combination of loss terms:

\[
L_{\text{rec}} = w_{\text{rec}} L_{\text{rec}} + w_{\text{perc}} L_{\text{perc}} + w_{\text{lexeme}} L_{\text{lexeme}} + w_{\text{z}} L_{\text{z}}.
\]

The reconstruction loss

\[
L_{\text{rec}} = ||M_i - M'_i||_2^2
\]

is simply the MSE loss between the generated motion block and the ground truth. We additionally include a perceptual loss to ensure the similarity between the generated motion and the ground truth in the feature level as well, which is defined as

\[
L_{\text{perc}} = ||E(M_i) - E(M'_i)||_2^2,
\]

where \(E\) is the motion encoder pre-trained in Section 5.1.

We assume that the gesture lexeme determines the type of gesture motion, and the other speech modalities only affect the motion variations. To enforce this assumption, we develop a new perceptual loss, namely the lexeme loss. We first generate a number of new motion blocks using the current gesture lexeme but random sets of other features,

\[
M'_i = G(M_{i-1}, A_j, s_j, z_j, P),
\]

where \(M_{i-1}, A_j, \) and \(z_j\) correspond to a random speech block \(j\). Then the lexeme loss is defined as

\[
L_{\text{lexeme}} = \frac{1}{N_j} \sum_{j \in \mathcal{J}} ||s_i - E(M'_i)||_2^2,
\]

where \(J\) is a random subset of all the motion blocks in the training dataset. The size of \(J, N_j\) is chosen based on the size of the dataset.

Lastly, we regularize the learning of the style code by applying a KL-divergence loss

\[
L_{\text{z}} = D_{KL}(N(\mu_z, \sigma_z^2) || N(0, 1)),
\]

where \(\mu_z\) and \(\sigma_z^2\) are the mean and variance vectors of the style codes in a mini-batch, respectively.

### 6 CO-SPEECH GESTURE INFERENCE

When given a speech as input, our system segments it into normalized feature blocks \(\{A_{\text{low}}^{\text{low}}, A_{\text{high}}^{\text{high}}, T_i, I\}\) and then generates motion blocks \(\{M'_i\}\) recursively, where

\[
M'_i = G(E_M(M_{i-1}^{\text{low}}), E_A(A_{\text{low}}^{\text{low}}, A_{\text{low}}^{\text{high}}, T_i, I), s_i, z_i, P),
\]

and \(E_M, E_A\) are the components of the learned gesture generator. The generated motion blocks are then denormalized to their original length in the input speech, producing a realistic co-speech gesture animation. Note that we again use the asterisk (\(*\)) to indicate a computed quantity that is not provided directly in the speech.

All the variables in Equation (15) are known except the gesture lexeme \(s^*\) and style code \(z^*\). As shown in Figure 2, our system learns two interpreters to compute them: the lexeme interpreter \(P_\delta\) translates high-level speech features into the gesture lexemes \(s^*\), and the style interpreter \(P_\gamma\) predicts the style code \(z^*\) according to the low-level speech features.

#### 6.1 Lexeme Interpreter

As illustrated in Figure 7, the lexeme interpreter is formulated as

\[
s_i^* = P_\delta(s_{i-1}^*, A_{\text{low}}^{\text{high}}, \hat{T}_i, I),
\]

which is conditioned on the gesture lexeme of the last motion block \(s_{i-1}^*\) and the high-level features of the current speech block. Like the

Fig. 7. Architecture of the lexeme interpreter.
generator, the high-level audio features $\mathbf{A}^H_i \in \mathbb{R}^{K \times d_A}$ are computed using three consecutive audio blocks

$$\mathbf{A}^H_i = E^\text{lex}_A(\mathbf{A}^{\text{high}}_{i-1}, \mathbf{A}^{\text{high}}_i, \mathbf{A}^{\text{high}}_{i+1}),$$

(17)

where each $\mathbf{A}^{\text{high}}$ contains the high-level representation of the input speech audio, and the encoder $E^\text{lex}_A$ is a single-layer fully connected network. The text feature $\tilde{T}_i \in \mathbb{R}^{K \times d_T}$ is also extracted from the text representation of the speech block as

$$\tilde{T}_i = E^\text{lex}_T(T_i),$$

(18)

where $E^\text{lex}_T$ is again a single-layer network. Lastly, the one-hot representation of the speaker ID $I$, is repeated $K$ times and converted into a feature block.

Those feature blocks can then be concatenated together and fed to an LSTM-based decoder to predict the next gesture lexeme $s^*_i$. However, considering that the lexemes are selected from the discrete gesture lexicon, we can convert this regression problem into a classification problem. Specifically, instead of directly evaluating Equation (16), we can let $P_s$ predict the probability that $s^*_i$ is a specific lexeme in the gesture lexicon, then the lexeme with the maximum likelihood will be considered as the result.

6.2 Style Interpreter

The style interpreter shares a similar structure with the lexeme interpreter. It computes $z^*_i$ as

$$z^*_i = P_s(\tilde{z}_i s^*_i, E^\text{style}_A(T_i, E^\text{low}_A(\mathbf{A}^{\text{low}}_{i-1}, \mathbf{A}^{\text{low}}_i, \mathbf{A}^{\text{low}}_{i+1}))),$$

(19)

which is conditioned on the last style code and the new gesture lexeme computed by the lexeme interpreter. The low-level audio representation $\mathbf{A}^{\text{low}}$ is used in the style interpreter.

6.3 Audio-Only Inference

Both the two interpreters can be reformulated to take only the speech audio as input, where the features related to the text representation $T$ and optionally the speaker ID $I$, will be removed from Equation (16) and (19).

In practice, these audio-only interpreters allow cross-language gesture generation, where the speech audio in another language can be taken as input to synthesize realistic gestures without further training. For example, we can utilize a pre-trained model on an English dataset to generate gestures that accompany a Chinese speech. We will show related experiments in Section 7.2.

6.4 Training

During the training of the generator, we have computed the gesture lexeme $s_i$ and the style code $z_i$ of every motion block in the training dataset. We then train the two interpreters using these results as the ground truth. We minimize the standard categorical cross-entropy loss to train the lexeme interpreter, while the MSE loss is used for the style interpreter.

6.4.1 Silent Period Hint. A speaker typically stops gesticulating during a silent pause [Graziano and Gullberg 2018]. Such behaviors are often crucial to the naturalness of a co-speech gesture animation. However, we find that it is often difficult for a gesture generator to deal with silent periods well, even in recent successful systems such as [Alexanderson et al. 2020; Kucherenko et al. 2020]. The speech-gesture datasets may lack necessary motion, and some specific generator models, such as LSTM, may exhibit generative inertia that makes it difficult to become stationary in time.

To solve this problem, we develop a new approach, which we refer to as the silent period hint, to encourage the lexeme interpreter to compute a specific silent lexeme that corresponds to a silent gesture when encountering a silent period. We check all the lexemes in the lexicon and label a number of stationary ones as the silent lexemes. Notably, the silent lexemes can be automatically labeled by finding such a lexeme corresponding to an empty text word. Then, when a training audio block is in a silent period, which can be detected by the data module of our system, we will force the lexeme interpreter to output the silent lexeme that is the nearest to the current lexeme in the latent space. Moreover, a silent data augmentation is applied when training the generator. We find data blocks that contain empty words and randomly insert 0 ~ 10 consecutive silent blocks after them. The silent block above includes four different features: (a) the audio feature is the environmental noise; (b) the style code is set to zero; (c) the gesture lexeme is the silent lexeme that is the nearest to the previous lexeme in the latent space; and (d) the motion is a stationary pose that is the same as the last frame of the previous motion block. In total, the amount of the inserted silent blocks accounts for 5% of the whole training set.

7 RESULTS

In this section, we first introduce the setup of our system and then evaluate its performance, followed by quantitative and qualitative comparisons with other recent systems. Lastly, we do the ablation study to make an overall analysis of our system.

7.1 System Setup

7.1.1 Datasets. Three speech-gesture datasets are used in this paper: the Trinity dataset [Ferstl and McDonnell 2018], the TED dataset [Yoon et al. 2019], and a Chinese dataset collected for this work.

Trinity Gesture dataset is a large database of speech and gestures jointly collected by Ferstl and McDonnell [2018]. This dataset consists of 242 minutes of motion capture and audio of one male actor talking on different topics. The actor’s motion was captured with a 20-camera Vicon system and solved onto a skeleton with 69 joints.
In this paper, we use the official release version of The GENEA Challenge 2020 Dataset [Kucherenko et al. 2021a], where 221 minutes are used as training data, and the remaining 21 minutes are kept for testing.

TED Gesture dataset [Yoon et al. 2019] is a 3D upper-body gesture dataset collected using 3D pose estimation from English TED videos. This dataset includes 3D upper-body gestures of the speakers, the aligned speech audios, and text transcripts. In total, there are 253,186 data samples, 80% of which are training samples, 10% belong to the validation set, and the remaining 10% are test samples. The duration of each data sample is 34 frames at a rate of 15 fps, so the total length of this dataset is about 97 hours. Notably, we adapt our model to take 3D joint positions instead of rotations for the TED dataset. The generated gestures are also represented in 3D joint positions, which are then converted into joint rotations for visualization.

Additionally, we collected a 4-hours (80% are the training data and 20% are used for testing) Chinese Gesture dataset using the Noitom Perception Neuron Pro system. This dataset contains 3D full-body gestures of five speakers, aligned speech audios, and Chinese text transcripts. The text transcripts were recognized by Alibaba Cloud Automatic Speech Recognition (ASR) service. The skeleton of this dataset is retargeted to be consistent with the Trinity Gesture dataset. To ensure semantic richness, speakers are instructed to cover a diverse set of topics, such as cooking, fiction, philosophy of life, and academic reporting. Figure 8a illustrates the distribution of onset intervals in our dataset, and Figure 8b shows the visualization of the learned gesture lexicon on our dataset.

7.1.2 System Settings. All the motion data are downsampled to 20, 20, and 15 frames per second on the Trinity, Chinese, and TED datasets, respectively. The range of onset intervals $[D_m, D_M]$ is $[0.2s, 0.5s]$ for both the Trinity and TED datasets, but $[0.3s, 0.6s]$ for the Chinese dataset. The length of each normalized block $K = [D_M \times fps]$. The generator synthesizes 4 seconds of gestures at a time. The dimensions of $d_1, d_2, d_3, d_6, d_7, d_8, d_9$ and $d_p$ are 768, 128, 128, 128, 192, 32, and 32 respectively. The size of the gesture lexicon $N_i$ is 50 for both the Trinity and Chinese datasets but 100 for the TED dataset. We train our framework using the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ and a learning rate of 0.0003. The loss weights $w_{re}, w_{s}, w_{rec}, w_{perc}, w_{lexeme}$, and $w_2$ are set as 1.0, 1.0, 1.0, 0.5, 0.2, and 1.0, respectively. At runtime, we use a Gaussian filter with a kernel size of 5 to smooth the denormalized gesture sequence, where $K = 5$ is chosen to generate the results presented in this paper.

We train separate gesture generators on the Trinity, TED, and Chinese datasets. The cross-language capability of a generator can be further enhanced by pre-training the audio encoder (Section 4.1.3) using datasets in different languages. We have tried pre-training the audio encoder using both an English dataset (such as the Trinity or TED datasets) and our Chinese dataset and using it to train the generator on the Chinese dataset only. The gesture results can be found in the supplementary video.

7.2 Evaluation

Figure 9 shows the gesture synthesis results for the speech excerpts from the test set of the Trinity dataset. Our system generates different types of realistic gestures. The character makes a metaphoric gesture when saying fine and an iconic gesture for defend. There are beat gestures for words like thing and selling, and a deictic gesture appears when the character says me.

We also did a cross-language synthesis experiment to test the robustness of our system. We use the pre-trained model trained on the Trinity dataset (an English dataset) to generate gestures for a Chinese speech clip. Since different languages do not share the same word embedding, we generate gestures by taking only the speech audio as input (Section 6.3). As illustrated in Figure 13b, when encountering a different language, our system still generates beat-matched and stylized gestures, reflecting the robustness of our system. We also trained our system on the Chinese dataset and then did another cross-language experiment by generating gestures for an English speech excerpt. Please refer to the supplementary video for the visualization results.

7.2.1 Style Editing. Inspired by the idea of Alexenderson et al. [2020], we augment our system to achieve motion editing by adding a feature block $C \in \mathbb{R}^{K \times d_c}$ of the gesture motion as an extra input of the generator and the lexeme interpreter (Figure 7). The computations in Equation (5) and (16) are then reformulated as

$$M_i^c = G(M_{i-1}, \hat{A}_i, s_i, z_i, P, C_i),$$

$$s_i^c = P_c(s_{i-1}^c, \hat{A}_i^H, \bar{T}_i, I, C_i),$$

which allows the network to learn the relationship between a desired motion feature and a gesture motion. During inference phase, we can easily edit the motion style feature of the generated gestures by modifying the feature block $C$ as needed.

Similar to [Alexenderson et al. 2020], we have experimented with three different style features using our system: the height of the right-hand (hand height), the average speed of both hands (hand speed), and the average distance from the hand positions to the up-axis through the root joint of the character (hand radius). We train a separate generator for each style feature. The training data is computed using the reference motions and averaged within a four-second sliding window, forming $K$-frame feature blocks.

We have synthesized three animations for each of the motion styles. Each animation has a constant desired low, mid, or high feature value, as shown in the first three columns of Figure 10. The last column of Figure 10 shows the accuracy of the generated motion features. These results indicate that all the editing signals could efficiently affect the generated gestures.

7.3 Comparisons

In this section, we compare our system with several state-of-the-art systems to demonstrate the advances made by our system. We first briefly introduce these systems and then our quantitative and qualitative comparisons. We also propose a simple but effective objective metric (PMB) to evaluate the rhythm performance for co-speech gesture synthesis.
Fig. 9. Qualitative results on the gestures synthesized by our method for four sample speech excerpts from the Trinity Gesture dataset [Ferstl and McDonnell 2018]. The character makes a metaphoric gesture when saying fine and an iconic gesture for defend. There are beat gestures for the words like thing and selling, and a deictic gesture appears when the character says me. The drawing of this figure is inspired by [Yoon et al. 2020].

Fig. 10. Results of style editing for the right-hand height (the first row), the hand speed (the second row), and the hand radius (the third row). The graphs on the right show the editing input (flat line) and the corresponding values of the output motions. The box plots show the statistics of the output.
Table 1. Comparison of our system to SG [Alexanderson et al. 2020], Ges [Kucherenko et al. 2020], GTC [Yoon et al. 2020], and S2AG [Bhattacharya et al. 2021a] on the TED and Trinity datasets. The system without beat segmentation (w/o BC) uses a fixed interval of $D_M$ for segmentation, which is $0.5s \sim 0.6s$ depending on which dataset is used. The system without gesture lexeme (w/o SC) uses the gesture lexicon and lexeme interpreter modules. The generator is retrained to predict future gestures based on only the previous motion, the audio, and the style code. Similarly, the system without gesture style code (w/o ZC) excludes the style code and the style interpreter modules. Only the motion, the audio, and the lexeme are used by the generator. Ours (audio only) denotes the audio-only inference.

| Dataset | System          | MAJE (mm) | MAD (mm/s²) | FGD | PMB (%) |
|---------|-----------------|-----------|-------------|-----|---------|
| Trinity | Real Gesture    | 0.0       | 0.0         | -   | 95.74   |
|         | SG              | 97.29     | 4.26        | 36.98 | 54.54  |
|         | Ges             | 82.41     | 3.62        | 31.04 | 71.0   |
|         | S2AG            | 54.93     | 1.49        | 20.36 | 79.53  |
|         | Ours (w/o BC)   | 59.11     | 1.89        | 16.13 | 78.18  |
|         | Ours (w/o SC)   | 70.10     | 2.51        | 29.75 | 85.74  |
|         | Ours (w/o ZC)   | 52.85     | 1.35        | 12.53 | 91.36  |
|         | Ours (audio only)| 57.99   | 1.83        | 15.79 | 87.35  |
|         | Ours            | 49.53     | 0.97        | 10.78 | 91.36  |
| TED     | Real Gesture    | 0.0       | 0.0         | -   | 93.10   |
|         | GTC             | 26.95     | 3.03        | 3.73  | 71.72  |
|         | S2AG            | 24.49     | 2.93        | 3.54  | 75.57  |
|         | Ours (w/o BC)   | 27.10     | 3.11        | 3.88  | 67.88  |
|         | Ours (w/o SC)   | 30.07     | 3.53        | 5.22  | 83.10  |
|         | Ours (w/o ZC)   | 21.33     | 2.61        | 2.47  | 88.67  |
|         | Ours (audio only)| 27.28   | 3.17        | 3.96  | 81.33  |
|         | Ours            | 18.13     | 2.29        | 2.04  | 89.52  |

Fig. 11. PMB results of continuously adjusting the threshold $\delta$ on the Trinity and TED datasets.

7.3.1 Systems of Comparison. We choose four recent successful 3D co-speech gesture synthesis systems for comparison. On the Trinity dataset, we compare with the systems of Style Gesture (SG) [Alexanderson et al. 2020] and Gesticulator (Ges) [Kucherenko et al. 2020]. SG generates gestures based on only the speech audio and uses a normalizing flow model. Ges leverages the audio and text of the speech to generate semantically consistent gestures. On the TED dataset, we compare the systems of Gestures from Trimodal Context (GTC) [Yoon et al. 2020] and Speech to Affective Gestures (S2AG) [Bhattacharya et al. 2021a]. GTC uses speech audio, text transcript, and speaker identity to generate gestures. Based on the three modalities used in GTC, S2AG adds another new modality of affective expressions from the seed poses into their model.

Because we have no access to the official pretrained model of SG, we strictly follow the official configuration and run the codes offered by authors to train a model. For other systems, we use the pretrained models provided by the authors. For a fair comparison, we use the same skeleton and motion frame rate as the baselines.

7.3.2 Quantitative Evaluation. We first adopt three commonly used evaluation metrics (MAJE, MAD, and FGD) to compare these systems quantitatively. MAJE measures the mean of the absolute errors between the generated joint positions and the ground truth over all the time steps and joints, which indicates how closely the generated joint positions follow the ground truth. MAD measures the mean of the $L_2$ norm differences between the generated joint accelerations and the ground truth over all the time steps and joints, which indicates how closely the generated joint movements match. Fréchet Gesture Distance (FGD) was proposed by Yoon et al. [2020], which measures the difference between the distributions of the latent features of the generated gestures and ground truth, where the latent features are extracted using an auto-encoder trained on the Human 3.6M dataset [Ionescu et al. 2013]. FGD could assess the perceived plausibility of the synthesized gestures.

Calculating the matching rate between audio and motion beats is a standard method of evaluating rhythm performance, which has been widely used in music-driven dance synthesis [Chen et al. 2021; Li et al. 2021b]. However, we cannot simply apply this method to the gesture generation task because the onset of beat gesture usually precedes the corresponding speech beat by a small amount of time [Pouw and Dixon 2019]. Thus, we need a distance threshold.
Ours Ours Ours
(b) Mean ratings for the Chinese speech clips.

The sequence of motion beats $B^m = \{b^m_1, \ldots, b^m_{N^m}\}$, where $N^m$ is the number of motion beats, is identified using the algorithm proposed by Ho et al. [2013] based on the local minima of joint deceleration. The audio beats $B^a = \{b^a_1, \ldots, b^a_{N^a}\}$ are the onset sequence, where $N^a$ is the number of onsets. $j^a_{i-1}$ represents the index of the audio beats that the last motion beat matches. $\delta$ denotes the indicator function. The threshold $\delta$ is set to 0.2s by default, but we can adjust $\delta$ continuously and observe the changes of the PMB values, which provides a more comprehensive picture of the rhythmic performance.

Table 1 summarizes the performance of all the systems on the two English datasets, Trinity and TED. For the MAJE, MAD, and FGD metrics, our system achieves the lowest values on both datasets. The PMB values drop rapidly without the gesture lexeme component (w/o SC), indicating that beat segmentation is crucial to gesture quality. Figure 11 shows the PMB results with different thresholds. It can be seen that even the ground-truth motion does not match the speech beats precisely, while the behavior of our method is closer to the ground truth compared with the baselines.

7.3.3 User Study. We further conduct user studies to assess the performance of our system qualitatively, where SG [Alexanderson et al. 2020] and Ges [Kucherenko et al. 2020] are used as the baseline methods. We generate 14 test video clips, each consisting of the results synthesized by our methods, the baselines, and the ground truth in random order. Among the 14 clips, nine are English clips generated from the test set of the Trinity dataset, and five are Chinese clips generated from the test set of our Chinese dataset. Notably, all the clips are synthesized using the models trained on the English Trinity dataset. The duration of each clip is around 30s.

We have recruited 30 volunteer participants to participate in our study, of which 19 are male and 11 are female. 19 participants are 18 – 22 years of age, 10 are between 23 – 30, and one is above 30. When participants watch each video clip, they will be asked to answer three questions and rate the videos from 1 to 5, with 1 being the worst, 3 being average, and 5 being the best. The three questions are: (1) human likeness (neglecting the speech audio), (2) speech-to-gesture content matching, and (3) speech-to-gesture beat matching.

The results of these user studies are shown in Figure 12, our system receives higher scores than the other systems and is closer to the real gestures (GT). A one-way ANOVA reveals main effects of human likeness, content matching, and beat matching, and a post-hoc Tukey multiple comparison test identifies a significant difference ($p < 0.005$) between our system and all the other methods. As also illustrated in Figure 13a, the end-to-end systems SG and Ges are less sensitive to rhythm than ours, and the resulting motions lack diversity, which affects their performance in the user studies.

The statistical results of the cross-language test (Figure 12b) demonstrate the better robustness of our system. When dealing with a completely different language (Figure 13b), the gestures generated by SG are more rigid and do not match the beats correctly. In contrast, our model (audio-only) is still able to perceive beats...
We conduct a variety of ablation studies to analyze the performance we can obtain their corresponding motion latent codes. As illustrated in Figure 14a, these latent codes (gray dots) only appear in specific gesture lexemes (orange, purple, and red), and it can be seen that the sample gestures of these latent codes indeed convey the semantics of the cluster \( C_0 \). The same observation is not true for the low-level audio features. If we also cluster all low-level audio features and pick a representative cluster \( C_0^{\text{low}} \), Figure 14b shows that the corresponding motion latent codes (various color dots) appear nearly uniformly in most of the gesture lexemes. The experiments above confirm the correlation between the high-level audio features and the gesture lexemes, as well as the semantic disentanglement between high-level and low-level audio features.

### 7.4.2 Gesture Style Code
The low-level audio feature \( a^{\text{low}} \) contains semantic-irrelevant information, e.g., pitch and volume. Presumably, it should affect the motion variations within a gesture lexeme. In our system, we introduce the learnable gesture style code \( z \) combined with the low-level audio feature to jointly determine the motion variations, which can be considered a fine-grained style. As illustrated in Figure 15b, the low-level audio feature increases the variety of the generated gestures but cannot fully decide the variations.

Fig. 15. t-SNE visualization of real gestures (GT) versus generated gestures in the motion latent space. We randomly choose two gesture lexemes for visualization. (a) results of our system without low-level audio feature and gesture style code. (b) with low-level audio feature but without gesture style code. (c) with low-level audio feature and gesture style code. See Section 7.4.1 for details.
Table 2. Comparison of style interpreters w/ and w/o low-level audio features and gesture style code. See Section 7.4.1 for details.

| Dataset | System     | Variance ↑ | FGD ↓   | PMB (%) ↑ |
|---------|------------|------------|---------|-----------|
| Real Gesture | w/o $\alpha_{\text{low}, z}$ | 0.09 | 17.99  | 89.35    |
|          | w/ $\alpha_{\text{low}} + w/o z$ | 0.21 | 12.53  | 91.36    |
|          | w/ $\alpha_{\text{low}, z}^\text{random}$ | 0.30 | 11.96  | 91.28    |
|          | w/ $\alpha_{\text{low}, z}$ | 0.37 | 10.78  | 91.36    |
| Trinity | w/o text feature | 2.72 | -      | 93.10    |
|          | w/ text feature | 0.89 | 3.71   | 84.57    |
| TED     | w/o text feature | 1.87 | 2.47   | 88.67    |
|          | w/ text feature | 2.49 | 2.13   | 88.89    |
|          | w/ $\alpha_{\text{low}, z}$ | 2.45 | 2.04   | 89.52    |
|          | w/ $\alpha_{\text{low}, z}^\text{random}$ | 0.21 | 12.53  | 91.36    |

Table 3. Comparison of style interpreters w/ and w/o text features.

| Dataset | System     | FGD ↓   | PMB (%) ↑ |
|---------|------------|---------|-----------|
| Trinity | w/o text feature | 10.91  | 91.36    |
|          | w/ text feature | 10.78  | 91.36    |
| TED     | w/o text feature | 2.09   | 89.22    |
|          | w/ text feature | 2.04   | 89.52    |

Table 4. Effects of the range of onset intervals.

| Dataset | Range of Onset Intervals | FGD ↓   | PMB (%) ↑ |
|---------|--------------------------|---------|-----------|
| TED     | 0.5-1.0s                 | 25.45   | 73.87     |
| Trinity | 0.2-1.0s                 | 19.16   | 90.75     |
|          | 0.2-0.5s                 | 10.78   | 91.36     |
| TED     | 0.5-1.0s                 | 3.61    | 65.34     |
|          | 0.2-1.0s                 | 2.55    | 89.10     |
|          | 0.2-0.5s                 | 2.04    | 89.52     |

The introduction of the gesture style code (Figure 15c), the distribution of generated gestures becomes closer to the ground-truth distribution.

Table 2 compares several settings in terms of motion variance, FGD, and PMB. We measure the Euclidean distance from the motion latent code of a motion to the corresponding gesture lexeme and compute the variance of these distances corresponding to the same lexeme. The motion variance is then defined as the average of such variances of every lexeme. $\zeta$ denotes the output of the style interpreter, while $\zeta^{\text{random}}$ is a random style code sampled from the normal distribution in the latent space. Consistent with Figure 15, Table 2 also shows that combining the low-level audio feature and styles achieves more significant motion variance and lower FGD values. Besides, the FGD values in Table 2 also indicate that the style codes computed by the interpreter generate gestures that are more perceptually acceptable than that created using random style codes.

We have also evaluated the importance of the text features in interpreting the gesture style codes (see Equation 19). The result of Table 3 shows that interpreting the style codes with text features does improve the FGD value. However, the improvement is marginal. Considering the inference efficiency, we can interpret the style codes conditional on only the low-level audio features.

7.4.3 Range of Onset Intervals. Choosing a proper range of onset intervals is crucial to achieving quality gestures. Intuitively, the lower bound of this range affects the model’s sensitivity to beat. If the lower bound is too high (such as 0.5s), the interval between the identified beats becomes large, causing the model to respond sluggishly to beats. The upper bound of the onset intervals regularizes the variance of the duration of the speech blocks, which should not be too large either. If the length of the speech blocks differs too much, the side effect of the normalization becomes visible, causing unnatural transitions between the generated gesture blocks.

To verify the above hypothesis, we compare the performance of three different interval ranges: 0.2-0.5s, 0.2-1.0s, and 0.5-1.0s, where 0.2-0.5s is our default setting (Section 4.2.1). As shown in Table 4, our default range of onset intervals achieves the best FGD and PMB values on both the Trinity and TED Gesture datasets. PMB drops significantly when the minimum distance between onsets is high (0.5-1.0s), which indicates that the model is insensitive to the rhythm. When the variance is too large (0.2-1.0s), PMB is barely affected, but FGD drops a lot, showing that the generated motions exhibit a lower quality. This result is consistent with our hypothesis.

Moreover, to explicitly show the effect of the minimum onset interval, we visualize the synchronization between the motion and audio beats under different interval ranges in Figure 16. Similar to Aristidou et al. [2022], we calculate the motion beats based on the motion deceleration [Davis and Agrawala 2018]. As shown in Figure 16, the motion beats extracted under the interval range of 0.2-0.5s (denoted by the green stars) synchronize with the audio beats (dashed red lines). In contrast, the motion beats extracted under the interval range of 0.5-1.0s (denoted by the orange stars) do not match well with the audio beats, which proves that a high minimum interval (0.5s) will cause a low beat sensitivity of the model.

7.4.4 Gesture Lexeme. Table 1 has shown that FGD increases significantly without the gesture lexeme, indicating the importance of the gesture lexeme in achieving quality motions. Besides, as demonstrated in Figure 17, the variety of the generated gestures is...
Table 5. Diversity of motions generated w/ and w/o gesture lexemes.

| Dataset    | System                        | Diversity ↑ |
|------------|-------------------------------|-------------|
| Trinity    | Real Gesture w/o gesture lexeme | 1.91        |
| TED        | Real Gesture w/o gesture lexeme | 2.99        |
| TED        | Real Gesture w/ gesture lexeme  | 3.40        |
|            | w/o gesture lexeme            | 4.09        |

Also significantly reduced without the gesture lexeme. To show this conclusion quantitatively, we calculate the entropy of the gesture lexemes to measure the motion diversity as

\[
\text{Diversity} = - \sum_{i=1}^{N_s} p_i \log p_i, \tag{23}
\]

where \(N_s\) is the size of gesture lexicon, \(p_i\) indicates the occurrence frequency of the \(i\)-th lexeme in the generated gestures. As shown in Table 5, our system with the gesture lexeme creates much higher diversity than the system without it, which further testifies the conclusion of Figure 17 and again emphasizes the importance of the gesture lexeme.

7.4.5 Size of Gesture Lexicon. Figure 18 shows the performance of our system under different gesture lexicon sizes. It can be seen that neither too small nor too large lexicons achieve good results, measured as FGD values. On the one hand, a small gesture lexicon forces a diverse range of gesture motions to be merged into the same gesture lexeme, which aggravates the one-to-many mapping issue and causes the generator hard to learn all the motions. On the other hand, an excessively large lexicon forcibly splits many lexemes into sub-lexemes. These sub-lexemes are typically close together, making the gesture lexeme interpretation more challenging and thus negatively affecting the gesture quality. We note that the PMB metric is less affected by the gesture lexicon size, possibly because our beat-based segmentation and normalization mechanism explicitly enforces the gesture rhythm. Based on these experiments, we set the sizes of the gesture lexicon for the Trinity and TED dataset to 50 and 100, respectively, as shown in Figure 6.

7.4.6 Different Interpreters. In our system, the lexeme interpreter determines the sequence of gesture lexemes. Our learning-based interpreter described in Section 6.1 accomplishes this task according to the speech features and the previous lexemes. Besides, inspired by [Aristidou et al. 2022], we further experiment with a statistical interpreter that matches the frequency of each lexeme in the generated gestures with the reference. Specifically, for a speaker \(I\), we calculate the frequency distribution of gesture lexemes, \(f_{I} \in \mathbb{R}^{N_s}\), and a transition matrix, \(L \in \mathbb{R}^{N_s \times N_s}\), describing the frequency of transitions between lexemes using the training data. These quantities can be considered as a global representation of speaker’s gesture style. During inference, we configure the statistical interpreter to ensure that the lexeme distribution of the generated gesture sequence \(f_{I}\) matches the corresponding global distribution \(f^I\). To achieve this goal, at each generation step, we compute a multinomial distribution characterized by \(f_{I+1}^*\) using the transition matrix and the difference between the current and target lexeme frequencies \(f_{I}\) and \(f_{I}\), where

\[
f_{I+1}^* = \text{softmax}(f^I - f_I) \cdot L. \tag{24}
\]

Then the next lexeme is sampled from the multinomial distribution. This statistical interpreter does not consider the input speech but only the statistics of the generated gestures when selecting the lexemes. In practice, the result motions still look acceptable but more random. This can be confirmed by Table 6, where the statistical
Table 7. Effects of the positional encoding block.

| Dataset | System                  | FGD ↓  | PMB (%) ↑ |
|---------|-------------------------|--------|-----------|
| Trinity | w/o positional encoding | 11.15  | 89.98     |
|         | w/ positional encoding  | 10.78  | 91.36     |
| TED     | w/o positional encoding | 2.19   | 88.13     |
|         | w/ positional encoding  | 2.04   | 89.52     |

7.4.7 Positional Encoding. The positional encoding block (Equation 8) informs the generator about the frame-level progress of the synthesis in a motion block, which helps the generator model the temporal structure, especially the rhythm, of the sequence. As shown in Table 7, the positional encoding block can improve the beat-matched rate (higher PMB) while enhancing the perceptual quality of generated movements (lower FGD).

8 CONCLUSION

In this paper, we present a rhythm- and semantics-aware co-speech gesture synthesis system that can generate realistic gestures to accompany a speech. For the rhythm, we utilize a segmentation pipeline that explicitly enforces beat alignment to ensure the temporal coherence between the speech and gestures. For the semantics, we successfully disentangle both low- and high-level neural embeddings of speech and motion based on linguistic theory. Then, we devise two neural interpreters to build correspondence between the hierarchical embeddings of the speech and the motion. To evaluate the rhythmic performance, we propose a new objective metric, PMB, to measure the percentage of matched beats. Our method outperforms state-of-the-art systems both objectively and subjectively, as indicated by the MAJE, MAD, FGD, PMB metrics, and human feedback. The cross-language synthesis experiment demonstrates the robustness of our system for rhythmic perception. In terms of application, We show our system’s flexible and effective style editing ability that allows editing of several directorial styles of the generated gestures without manual annotation of the data. Lastly, we have systematically conducted detailed ablation studies that justify the design choices of our system.

There is still room for improvement in our current research. First, our beat detection algorithm is not perfect. We have assumed that the gesture beats coincide with the verbal stresses, but in practice, it has been observed that gesture beats may not always correspond to stressed syllables [McClave 1994]. How to accurately model the complex gestural rhythm is an exciting topic for further exploration. Second, our system can only capture semantics-related gestures repeatedly appearing in the dataset. Learning semantically meaningful gestures that are sparsely distributed in a dataset and allowing a user to control the gesture corresponding to specific semantics is still challenging. Third, our system hypothesizes that each audio onset should correspond to a beat gesture. However, in reality, humans do not make a beat gesture at every point of verbal emphasis. We believe our framework can be easily augmented by employing another model to predict whether the character should gesture at a specific moment, as suggested by [Speech2Properties2Gestures], and replacing the corresponding lexeme with the silent lexeme. Finally, we only consider the upper body gestures in this work. Generating full-body gestures that include locomotion, facial expressions, finger motions, and the temporal and semantic correspondence among them is a valuable topic for future exploration.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments. This work was supported in part by NSFC Projects of International Cooperation and Exchanges (62161146002).

REFERENCES

Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. Computer Graphics Forum 39, 2 (2020), 487–496.

Andreas Aristidou, Daniel Cohen-Or, Jessica K Hodgins, Yiorgos Chrysanthou, and Ariel Shamir. 2018. Deep motifs and motion signatures. ACM Transactions on Graphics (TOG) 37, 6 (2018), 1–13.

Andreas Aristidou, Anastasios Yiannakidis, Kifer Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. 2022. Rhythm is a Dancer: Music-Driven Motion Synthesis with Global Structure. IEEE Transactions on Visualization and Computer Graphics (2022), 1–1.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020. vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. In International Conference on Learning Representations.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 2 (2019), 423–443.

J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler. 2005. A Tutorial on Onset Detection in Music Signals. IEEE Transactions on Speech and Audio Processing 13, 5 (Sept. 2005), 1035–1047.

Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021a. SpeechAffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning. In Proceedings of the 29th ACM International Conference on Multimedia (Virtual Event, China) (MM ’21). 2021b. Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents. CoRR abs/2101.11101 (2021).

Piotr Bojanowski, Edouard Grave, Armand Jolion, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the association for computational linguistics 5 (2017), 135–146.

Jude K. Burgeon, Thomas Birk, and Michael Pfau. 1990. Nonverbal behaviors, persuasion, and credibility. Human communication research 17, 1 (1990), 140–169.

Justine Cassell, Hannes Hogni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In Life-Like Characters. Springer, 163–185.

Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. 2021. Choreomaster: choreography-oriented music-driven dance synthesis. ACM Transactions on Graphics (TOG) 40, 4 (2021), 1–13.

Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. 2015. Predicting co-verbal gestures: A deep and temporal modeling approach. In International Conference on Intelligent Virtual Agents. Springer, 152–166.

Abe Davis and Maneesh Agrawala. 2018. Visual rhythm and beat. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2532–2535.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Prashulla Dhariwal, Heewoo Joo, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. ArXiv abs/2005.00341 (2020).

Paul Ekman and Wallace Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica 1, 1 (1969), 49–98.

Daniel PW Ellis. 2007. Beat tracking by dynamic programming. Journal of New Music Research 36, 1 (2007), 51–60.
Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2022. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. IEEE Transactions on Human-Machine Systems 52, 3 (2022), 379–389.

Jing Xu, Wei Zhang, Yalong Bai, Qibin Sun, and Tao Mei. 2022. FreeForm Body Motion Generation from Speech. arXiv preprint arXiv:2203.02291 (2022).

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021).

Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) 39, 6 (2020), 1–16.

Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. 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). 4303–4309.