



Statistics-based Motion Synthesis for Social Conversations

Yanzhe Yang¹ , Jimei Yang² , Jessica Hodgins¹ 

¹Carnegie Mellon University, USA

²Adobe Research, USA



Figure 1: Conversational gestures that were automatically synthesized from a motion capture database to match a novel audio segment.

Abstract

Plausible conversations among characters are required to generate the ambiance of social settings such as a restaurant, hotel lobby, or cocktail party. In this paper, we propose a motion synthesis technique that can rapidly generate animated motion for characters engaged in two-party conversations. Our system synthesizes gestures and other body motions for dyadic conversations that synchronize with novel input audio clips. Human conversations feature many different forms of coordination and synchronization. For example, speakers use hand gestures to emphasize important points, and listeners often nod in agreement or acknowledgment. To achieve the desired degree of realism, our method first constructs a motion graph that preserves the statistics of a database of recorded conversations performed by a pair of actors. This graph is then used to search for a motion sequence that respects three forms of audio-motion coordination in human conversations: coordination to phonemic clause, listener response, and partner's hesitation pause. We assess the quality of the generated animations through a user study that compares them to the originally recorded motion and evaluate the effects of each type of audio-motion coordination via ablation studies.

CCS Concepts

• **Computing methodologies** → **Motion capture; Motion processing;**

1. Introduction

Interactive and immersive social experiences often include human characters in conversations. These motions should be natural looking for both the speaker and the listener so that they are not distracting. The coordination between the speaker's motion and the primary stresses of their audio have been extensively studied [LTK09; LKTK10; CM11; CM14]. Listener response that is properly synchronized to the speaker's words also plays a key role in creating an engaging rendition of these social settings [KH97; TG67]. In this paper, we propose an efficient motion synthesis technique that generates gestures and body motions in dyadic conversation. The resulting motion respects three forms of audio-motion coordination in human conversations: the coordination to phonemic clause, listener response, and partner's hesitation pause.

The psychological literature has discussed conversation coordination between speakers and listeners as well as between the speaker and their spoken words. In this paper, we focus on three observations from that literature. The first is related to phonemic clause. Phonemic clause is widely accepted as the basic unit in speech encoding because it shows a systematic relationship to body movement. Therefore it has been used to synthesize speaker motions in prior work [BD62; KH97; LTK09]. Phonemic clause is described as a group of words that have one strongly stressed word, also known as the primary stress. The start, peak, and end markers of these clauses provide guidance about how to synthesize the speaker's motions.

The second observation is related to listener response. Dittmann and Leewellyn show that listeners vocally respond to speakers at

the end of a phonemic clause [TG67]. Inspired by their work, we assess whether vocal listener response can be used to synthesize listener behaviors. The vocal listener response are words that seek clarification or brief acknowledgment/reinforcement sounds such as “yes”, “um-hmm”, “yeah”, “I see”. We focused on the latter set of sounds in this work.

The third observation is that when the speaker hesitates or pauses, listeners often nod as a response. Although Dittmann and Leewellyn show that listeners do not always vocally respond to hesitation pauses, we assess whether listeners respond to hesitation pauses using body movements, such as nodding. If the answer to this question is yes, then hesitation pauses can also be used to synthesize plausible listener behaviors. Similar to listener response, we focused on the brief sounds that speakers make while reasoning and thinking, such as “er”, “uh”, “you know”.

In this paper, we describe a motion synthesis technique that automatically creates human motions that respect these three observations for a novel segment of audio from a dyadic conversation. For all three observations, we conducted experiments to verify these hypotheses in our captured dataset and then modeled this synchronization as a search constraint in the motion synthesis process. The key insight behind our approach is that we can leverage the inherent audio-motion coordination in the original database of recordings by matching the talking and listening audio channels of the new audio input with the original audio recording in the database. We factor in the start, peak, and end of the phonemic clause, and start/end of listener response and partner’s hesitation pause.

At the core of our approach is a motion graph that is constructed and then searched with the three observations as constraints. Motion graphs have been used very successfully in the synthesis of locomotion with a sparse set of synthetic transitions. For example, Lee and colleagues describe a system that only allows transitions on contact change states (touchdown and liftoff of the feet) [LCR*02]. Treuille and colleagues blend subsequences that start/end during flight or mid-stance [TLP07]. This sparse set of synthetic transitions creates graphs for locomotion that are rich enough to allow the generation of a variety of motions but small enough to be searched efficiently, usually with an optimization function that approximates energy consumption [KGP02; LCR*02; SH07]. Conversational gestures do not fit easily into this framework because there is no obvious analog to the contact change states of locomotion for the insertion of synthetic transitions. Conversations also are not “optimal” in the sense of minimum use of energy and gestures/body movements have a wide set of stylistic variation because there is no balance requirement to constrain the motion as there is with locomotion.

We address these issues by creating synthetic transitions that mimic the statistics of the natural transitions and adaptively down-sampling the data based on pose change. Given a new audio track, the search finds a motion sequence that has the expected synchronization for natural conversation. The richness of conversation gestures requires a significantly larger motion graph than what was needed for locomotion. To deal with the size of the graph and the lack of an optimality criterion for conversation gestures, we adopt a stochastic greedy search to find a high quality animated sequence.

We demonstrate the efficacy of this approach by conducting a

leave-one-out study where one clip is held out as a test clip, and then the resulting synthesized motion is compared to the original motion for that clip. In a perceptual study, the synthesized motions were judged as more natural 30% of the time. We also evaluate the effects of the two listener-related constraints with ablation studies. The results searched with all constraints are rated as significantly better than the results searched without listener response constraints; these results demonstrate the importance of modeling listener response.

1.1. Contributions

The main contributions of this work are

- An efficient motion synthesis technique that is based on stochastic greedy search. Audio-motion coordination is achieved by three types of constraints: phonemic clause constraints for speaker behaviors, listener response constraints and partner’s hesitation pause constraints for listener behaviors.
- An improved motion graph construction process that is effective for conversational gestures. The process automatically determines the cost thresholds for adding new transition edges so that the resulting graph achieves a good balance between transition smoothness and style diversity.
- A 30-minute conversational database recorded with two actors. The database has been annotated with audio labels including phonemic clause, phonemic clause peaks, hesitation pause, vocal listener response, as well as coarse motion labels such as gesturing and nodding. This database is available at <https://www.cs.cmu.edu/~dyadic-conversation/>.

2. Related Work

In this section, we first briefly introduce the literature related to speaker motions of body, hand, head and fingers, then discuss the related study on synthesizing listener behaviors, and introduce related database and annotation tools, and finally conclude with a brief discussion on motion graph technique.

Speaking motion has been studied extensively and various techniques have been developed to model speech movements using behavior trees, markup languages, and statistical models [CVB01; LM06; KKM*06; NKAS08; XPM14; LM13; TMMK08]. Levine and colleagues map speech motions to audio prosody with Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) [LTK09; LTK10]. Chiu and colleagues model the coordination between audio and motion for speeches with machine learning techniques such as Deep Belief Nets, Deep Conditional Neural Fields, and a combination of CRFs and Gaussian process models [CM11; CM14; CMM15]. Stone et al. and Kopp et al. develop data-driven systems that synthesize motion and speech segments together by reordering the combined captured data [SDO*04; KB12]. Sadoughi and Busso introduce a hybrid rule-based and data-driven system that retrieves gestures from a dataset of motions corresponding to the speech [SB15].

The most similar work to ours for speaker motion is that of Fernandez-Baena et al. which synthesizes gesture for speech based on prosody and gesture strength, and adopts a motion graph technique for motion synthesis [FMA*14]. Similar to their work, we

factor in the prosody features to search the speaker's motions based on phonemic clauses. But while they model motion constraints, we leverage the inherent audio-motion coordination in the original recordings by matching the talking and listening audio channels of the new audio input with the original audio recording in the database. The biggest benefit of search using audio is reducing the requirements for motion labeling. Another key difference between their work and ours is that we synthesize motions for dyadic conversations instead of single-person speech.

Previous researchers also studied head and finger motions for speakers. Ben-Youssef et al. build HMMs for head motions [BSB13]. Mu et al. model head movement patterns with Classification and Regression Trees and create motions based on lexical and prosodic features [MTCY10]. Jörg et al. introduce a data-driven method to automatically add finger motions to body motions for characters giving a speech [JHS12]. Wheatland et al. automate hand motions by Principal Component Analysis [WWS*15]. Ding et al. present systems that synthesize head motions and eyebrow movements for a single speaker with a Hidden Markov model and Bi-Directional Long Short Term Memory (BLSTM) [DPA13; DZX15]. Barbulescu et al. develop a model for speaker's head motions with different attitudes from audio prosody and capture a dataset of head motions from actors who performed as if another person is standing in front of them [BRB17]. Jin et al. present a deep learning model that generates head and eye motion in three-party conversations [JDZD19].

Psychology studies have shown the importance of interaction in listener behaviors. Knapp and Hall discuss the literature on interaction synchrony [KH97] which is the coordination of the speech and movement between two or more people in a conversation. For example, listeners always nod at the end of a phrase where a speaker makes a point. McDonnell et al. report that people can detect inter-character synchronization between characters using body motion alone and human sensitivity to such synchronization is not affected by the number of speakers or topics [MEDO09]. Ennis et al. obtain similar results on human sensitivity for inter-character synchronization in three-character conversations and report no effect for whether characters are female or male [EMO10].

Our system not only models speaker behaviors, but also listener behaviors. Early work on listener behaviors employed hand-crafted rules or a hand-designed markup language. Cassell et al. introduce a rule-based system for multiple agents by generating and analyzing dialogue [CPB*94]. Gratch and colleagues develop a behavior mapping that correlates detected speaker events to listener motion behaviors in human-to-virtual interactions [GOL*06; GWG*07]. Marsella et al. present a framework using a hybrid of lexical and prosody information. The listener in their system performs head movements that mirror the speaker with some delay [MXL*13]. Instead of synthesizing motions based on linguistic and speech components, Jan and Traum synthesize conversational motions using behavior trees [JT05]. Sun et al. determine conversation events based on agent attributes and spatio-temporal context and create motions using behavior trees [SSH*12]. In recent work, Greenwood et al. model speaker and listener head motions using BLSTM [GLM17]. Ahuja et al. propose a model based on temporal convolutional network that predicts upper-body motions

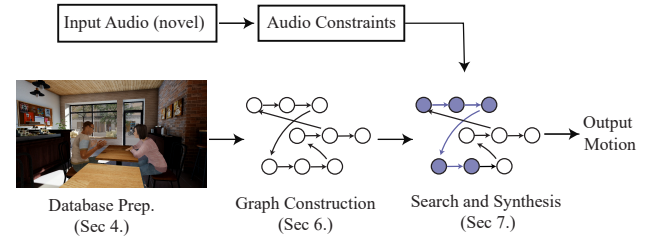


Figure 2: An overview of our motion synthesis system. The system takes in captured audio and motion sequences of dyadic conversations as input, and extracts motion features and audio features to create a motion database. From this database, the system constructs a motion graph that can be searched to find motion clips that match a novel input audio clip.

from an audio clip and the partner's motions in dyadic conversations [AMMS19]. In contrast to their work, our system does not require partner's motions as input and is able to generate motions for both characters. In addition to dyadic conversations, Joo et al. predicts body motions from the partner's motions using Convolutional Autoencoder in three-party conversations [JSCS19].

We captured our own dataset for dyadic conversations. The audio is annotated using the IBM Speech to Text service [IBM18] and then the transcription and audio segmentation is manually corrected. The motion is annotated with an automatic algorithm to the level of gesturing and nodding, which is then used in the database analysis and hypothesis validation. Others have developed annotation tools for more detailed annotation and specifically for head motion. Martell developed an annotation scheme named FORM that annotates gestures based on kinematics [Mar02]. Kipp and colleagues introduce a scheme for hand/arm gesture phases using the ANVIL video annotation tool [Kip01; KNA07]. Vandeventer et al. contribute a dataset of head motions reconstructed from images from different views [VARM15].

Our work is a data-driven system that synthesizes full-body motions for both a speaker and a listener based on an enhanced motion graph technique. Motion graphs are not new to animation [KGP02; LCR*02; AFO03; MC12; HLL16] but have not been used extensively for conversational gestures. Safonova and Hodgins introduced an interpolated motion graph that allows a sequence of constraints to be matched more accurately [SH07]. McCann and Pollard use very short motion fragments (0.1s) for an online controller for locomotion [MP07].

3. System Overview

Our system synthesizes motions from the audio recordings of a conversation between two characters. Figure 2 outlines our approach. First, we capture the motion and audio of a series of dyadic conversations to create a database (Section 4). The database of each character is further annotated with audio and motion features such as emphasis and prosody. Second, our system builds motion graphs by adding synthetic transition edges that preserve the statistics of

the original frames (Section 6). Finally, our system synthesizes motions from the motion graphs with a stochastic greedy search algorithm that achieves reasonable time and memory efficiency (Section 7).

4. Database Preparation

We prepared an annotated database with synchronized audio and motion data for dyadic conversations on different topics. The database of each character is annotated with the start/peak/end timings of audio and prosody features for what the character hears and says, and annotated with motion timings of strokes. In the following subsections, we discuss the motion capture and database annotation in detail.

4.1. Captured Database

We recorded 30 minutes of different types of dyadic conversations in a motion capture lab. The motions were captured at 120 Hz and downsampled to 30 Hz in our dataset. Our performers were trained as actors at the undergraduate level and were instructed to have free-ranging conversations. The recordings were done by a male and a female actor. They were given only high-level guidance about the conversation scenario, such as “a meeting with a friend that you have not seen since high school” or “a job interview”, and they were told which of five emotions to use in each conversations (happy, sad, angry, excited, nervous). For each scenario, we recorded two to three conversations featuring different levels of emotional intensities, resulting in 17 conversations total. For simplicity of cleaning and processing the data, we minimize contact by setting no objects on the table and asking the actors not to intertwine their fingers or touch their hairs.

The motion of the actors was captured with a Vicon motion capture system. In all conversation scenarios, the actors were sitting at a table and facing each other. They wore 62 markers, including four markers on the back of each hand. Finger motions were manually animated by an artist after the capture due to the limitations of our motion capture system. In addition, we recorded reference video of the performance and synchronized it with the captured motion data. Two character models were created using Adobe Fuse CC, and then rigged by an artist to skeletons from the motion capture system to avoid errors introduced by retargeting.

Our database consists of captured motion frames and the matching audio. We use a Y-up and right-handed coordinate system. Each motion frame is represented by $\{q, y^r, \Delta x^r, \Delta z^r, \Delta q^{yaw}\}$, where q is the joint rotation relative to the parent joint and the root orientation along the X and Z axes, y^r is the root position along the vertical axis (the Y axis), Δx^r and Δz^r are the root linear velocities on the ground plane (the XZ plane), and Δq^{yaw} is the root angular velocity along the vertical axis (the Y axis).

4.2. Audio Segmentation

The recorded audio is segmented into and labeled with one of four labels: phonemic clause, vocal listener response, partner hesitation pause, and idle. Phonemic clause is defined as a group of words that has a strongly stressed word. The average number of words

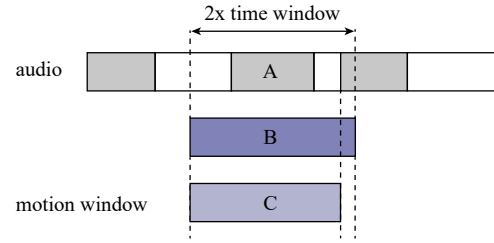


Figure 3: The audio segment provides a window in which we can look for a corresponding movement segment. A is an audio segment (speech). B is an expanded time window that is twice as long as the audio segment. The candidate time window for the matched movement segment is C, which has been trimmed to remove any overlap with neighboring audio or motion segments.

in a phonemic clause is 4.29 ± 2.16 for the female character and 3.86 ± 1.36 for the male in our database. Listener responses are brief sounds such as “yes”, “um-hmm”, and “I see”. Partner hesitation pauses are the brief sounds that a subject’s partner makes while reasoning and thinking, such as “er”, “uh” and “you know”. The idle label indicates a segment that does not belong to one of the other labels.

Audio segmentation and labeling is implemented with a tool chain that includes the IBM Speech to Text service, our GUI tools, word detection, and OpenSmile. First, we transcribe the input audio to text, speaker id, and timestamps using IBM Speech to Text service [IBM18]. In their output, brief hesitation pauses such as “uh” and “hmm” are usually transcribed as “%hesitation”. We manually correct the output of the automatic process, i.e. transcription and segmentation, using a GUI tool. Second, we defined a corpus for hesitation pause and listener response, and each segmentation is labeled automatically by detecting whether the words fall into a corpus or category. Hesitation pauses are short sounds or a short phrase with one or two words that indicates the speaker is thinking, such as “huh”, “uh”, “umm”, “well”, or “%hesitation” as transcribed in IBM Speech to Text. Examples of listener response are “ah ha”, “wow”, “um hum”, “absolutely”. The word corpus is a part of the database and is available at <https://www.cs.cmu.edu/~dyadic-conversation/>. Finally, we use OpenSmile [EWGS13] to extract prosody features from the input audio, and identify the peak indicators by finding the local maxima of the fundamental frequency (F_0) in each audio segment. After audio processing, our system has knowledge about timings and duration of the phonemic clause, listener response, and hesitation pause for each actor in the motion database.

5. Statistics of Audio-Motion Co-occurrence

Our approach focuses on three hypotheses: 1. For the phonemic clause, speakers often use hand gestures simultaneously for emphasis; 2. In the vocal listener response, listeners usually nod at the same time. 3. When another character is hesitating or pausing, listeners nod to respond. The question is how often do these forms of coordination occur in our database and will they provide useful guidance when synthesizing conversation motions? To answer

	Hypothesis	Matched Motion	Total Audio Segments	Rate
Female	PC - LH	319	515	62%
	PC - RH	336	515	65%
	PC - LH/RH	387	515	75%
	LR - Nod	70	121	58%
	PH - Nod	21	69	30 %
Male	PC - LH	314	614	51%
	PC - RH	309	614	50%
	PC - LH/RH	382	614	62%
	LR - Nod	81	152	53%
	PH - Nod	14	40	35 %

Table 1: Audio and motion co-occurrence. Each row lists the aggregated data of audio motion co-occurrence for each hypothesis and each actor. “Hypothesis” defines the type of audio and motions, “matched motion” is the number of audio-motion co-occurrences, and “rate” is the occurrence rate of the matched motion among all audio segments. In this table, PC: phonemic clause; LH: left hand gesture; RH: right hand gesture; LR: listener response. PH: partner hesitation;

these questions, we measure these three forms of audio-motion co-occurrence.

To analyze whether motion and audio co-occur, we create an automatic motion segmentation and labeling process using the known audio segmentation as a candidate time window to look for matching body movements. Figure 3 illustrates this process. Second, motion segmentation is calculated based on relative joint speed curves. Hand speeds are calculated in a local coordinate system based at the shoulder, and head speeds are measured in a local coordinate system based at the neck and then projected on the y axis (vertical). Annotated audio and motion are shown in the supplementary video.

Let v denote joint speed, which is computed from the joint rotations q using forward kinematics and derivatives [Par12]. A segment is labeled “idle” when the maximum joint speed is smaller than an activation threshold. In our implementation, the activation threshold θ_S is the minimum joint speed plus a small tolerance proportional to speed range:

$$\theta_S = \min_S(v) + \alpha_S \times (\max_S(v) - \min_S(v)) \quad (1)$$

where S is the set of frames in the entire clip. In our experiment, we set $\alpha_S = 0.1$. The activation thresholds are determined adaptively for each clip because motion clips that have different emotional content contain different amounts of motion.

When a gesture stroke is detected, timings are detected with crossing threshold events similar to Levine and colleagues [LTK09]. A segment starts when joint speed v is greater than a preset threshold θ_W for the first time in the candidate time window, and ends on the last frame within the candidate time window where the velocity is greater than the threshold. In our implementation, θ_W is computed adaptively based on the speed curve in each candidate window W :

$$\theta_W = \min_W(v) + \alpha_W \times (\max_W(v) - \min_W(v)) \quad (2)$$

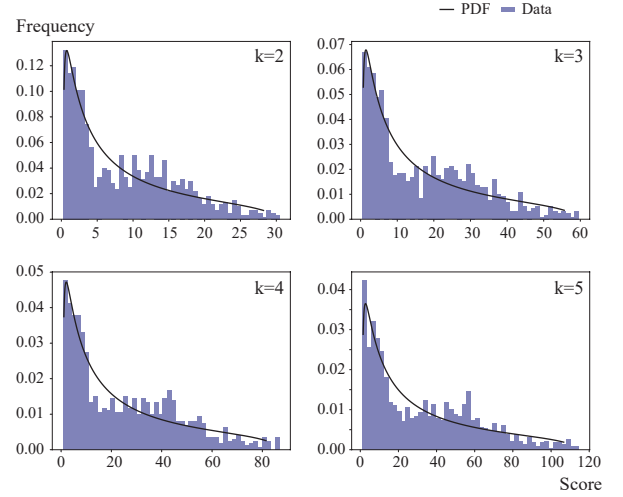


Figure 4: Statistics of transition cost $c_{i, i+k}$ for $k = 2, 3, 4, 5$ (Equation 3). The purple bar represents the histogram of transition cost. The black line represents the fitted Johnson SB distribution.

where α_W is 0.1 in our experiment.

Let PC denote a phonemic clause, LH denote a left hand gesture, RH denote a right hand gesture, LR denote a listener response, and PH denote a partner hesitation pause. The frequency of co-occurrence is shown in Table 1. In our experiment, we observed that the phonemic clause has a strong co-occurrence with a left or right hand gesture. The occurrence rate of a left or right hand gesture around a phonemic clause is 75% for the female character and 62% for the male character. Vocal listener response also has strong co-occurrence with visual listener response such as nodding: 58% for the female and 53% for the male. Partner hesitation pause also shows co-occurrence with nodding, though this effect is less strong compared to the other two. The occurrence rate of a nodding motion near a partner’s hesitation pause is 30% for the female and 35% for the male.

The well-known theory that hand gestures have strong co-occurrence with phonemic clause is supported by this statistical analysis of our database. In addition, we also observed a strong co-occurrence between vocal listener response and nodding, as well as a weaker but still significant co-occurrence between partner’s hesitation pause and nodding. Given this analysis, we developed search constraints that maintain these statistics in the generated motion. We designed start/peak/end constraints for phonemic clauses which is similar to previous work, and proposed listener response constraints and hesitation pause constraints. As we focus on short sounds or a short phrase with one or two words for listener responses and hesitation pauses, only start/end timings are considered for these two constraints.

6. Motion Graph Construction with Distance Distribution

In order to generate natural conversational motions, our system needs to construct a motion graph that is rich enough to allow

the generation of a variety of motions but small enough to allow efficient search. Synthetic transitions added to the graphs should appear as natural as the captured motions. In this section, we first introduce the standard pipeline for motion graph construction and then discuss why creating a motion graph for conversation is more difficult than for locomotion. Finally, we illustrate a statistics-guided approach to address these challenges.

A motion graph MG is composed of states and transitions. Each state is a motion frame in the database as defined in Section 4.1 and its associated audio signal (the fundamental frequency $F0$ in our implementation). A transition in a motion graph G indicates that a character can naturally move from one state to another in two consecutive frames. The graph construction process initially adds all original transitions between consecutive frames in the captured motions to the motion graph, and proceeds to find and add *synthetic* transitions between non-consecutive motion frames to the graph.

The key to constructing a motion graph that yields novel and natural-looking motions is to identify new transitions that are indistinguishable from those in the original captured motion, at least after blending. To evaluate whether a new transition from frame i to frame j yields a plausible motion, Lee and colleagues estimate the probability of transitioning based on a measure of frame similarity [LCR*02]. Similarly, our system calculates a the *transition cost* $c_{i,j}$ between the two frames. A lower cost indicates higher likelihood of a natural transition. The transition cost from frame i to frame j is defined as the weighted sum of distances between the pose in frame i and frame $j - 1$. We assume that any pairs of consecutive frames in the original dataset are natural transitions. Under this definition, the transition cost from frame i to frame $i + 1$ is 0, which is consistent with this assumption. Let p and \dot{p} denote joint positions and velocities relative to root joint, which is computed from the joint rotations q using forward kinematics and derivatives [Par12]. The transition cost $c_{i,j}$ is

$$c_{i,j} = d(p_i, p_{j-1}) + w_v d(\dot{p}_i, \dot{p}_{j-1}) \quad (3)$$

where w_v scales the velocity term to match the range of the position term. The first term $d(p_i, p_{j-1})$ measures the joint position difference between frame i and frame $j - 1$:

$$d(p_i, p_{j-1}) = \sum_{b=0}^{N_b} w_b \|(p_i[b] - p_{j-1}[b])\| \quad (4)$$

where w_b are weights for each joint and N_b is the number of all joints. We set the weight to be zero for finger joints and 1 for other joints in our experiments. The second term $d(\dot{p}_i, \dot{p}_{j-1})$ measures the joint velocity distance in the same formula as the first term.

Pruning techniques are often employed in previous work to keep the motion graph concise and the search process efficient. Lee and colleagues allowed transitions only on contact change states, i.e., when one or both feet touch or leave the ground [LCR*02]. Safonova and Hodgins allowed transitions inside the contact phase but only the optimal segments were kept [SH07]. Similarly, we allowed transitions only on key states. A key state is either a state where audio status changes, or represents a keyframe. Keyframes are sampled adaptively according to joint position change, with more samples when a joint is moving and fewer samples when it is not. In our implementation, 22.25% of the states are key states.

	Before SCC		After SCC	
k	#states	#edges	#states	# edges
2	50,568	130,397	3,899	13,182
3	50,568	408,234	40,673	376,300
4	50,568	1,823,041	47,854	1,810,036

Table 2: Number of states and edges before and after removing states not in the largest strongly connected component (SCC) during the graph construction process. In our experiments, we observed $k = 3$ yields the best results.

Furthermore, given the transition probability between all pairs of frames, Lee et al. propose that transitions with a transition probability less than a chosen threshold should not be added to the graph [LCR*02]. In other words, those transitions with a transition cost higher than a chosen threshold should not be added to the graph. This threshold of transition costs represents a trade-off between transition quality and graph connectivity. If the threshold is low, the transitions added to the motion graphs will look more natural, but the number of transitions that can be added is reduced. On the other hand, if the threshold is too high, the process will identify many synthetic transitions but some of them are bad transitions which will affect motion quality.

Manually setting a threshold is not too difficult for a motion graph consisting of locomotion because the motion is cyclic and therefore provide many opportunities for good transitions. Setting a reliable threshold is harder for conversational motions because of the significant variance in the gestures and postures that occur during conversation. Usually this threshold is set through multiple iterations of trial and error. Our method adaptively sets the thresholds using statistics collected from captured database.

We fit the histogram of transition costs to different analytic models and observed that the Johnson SB distribution fits our data best (lowest residual sum of squares between predicted and empirical values). Figure 4 illustrates the distribution of $c_{i,i+k}$ which indicates the cost of transitioning from frame i to frame $i + k$ and $k \in \{2, 3, 4, 5\}$. Most $c_{i,i+k}$ values are distributed near 0 because of resting motions.

A smaller k creates synthetic transitions that are of higher quality. Specifically, setting $k = 1$ results in all synthetic transitions having the same quality as the original transitions. In practice, this causes too few synthetic transitions being added to the graph. Fortunately, by adopting motion blending techniques [Par12], lower quality transitions can still be used without creating unnatural motions. In general, more transitions in the graph means more candidates during searching, which leads to more diverse results. In our experiments, we found that $k = 3$ yields the best balance between transition quality and connectivity. Table 2 summarizes the number of states and edges in the graphs with different k values. We observed that when $k = 2$, less than 10% of the frames remain after removing states that are not in the largest strongly connected component [Tar72]. This number grows to 80% when $k = 3$. Further increasing k results in motion graphs with stronger connectivity but also bad transitions that affect the quality of the resulting motion.

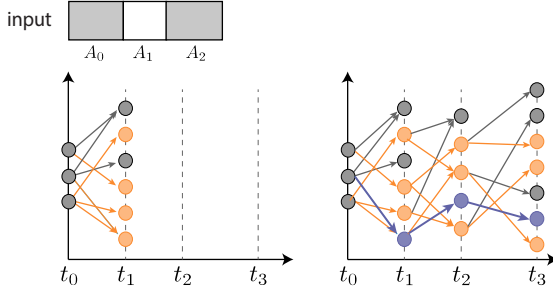


Figure 5: Illustration of the search process. In this example, $n_{start} = 3, n_{save} = 4$. At each step, n_{save} low-cost actions (orange) are saved. States that result from taking these actions are the start states for the next step. In the final step, the best complete path (purple) is selected.

In our implementation, we select a 95% confidence level for the transition cost $c_{i, i+k}$ as a threshold and $k = 3$.

After constructing the motion graph, motions can be synthesized by finding a path on the graph. A sequence of poses that are stored on the path is a motion clip for a character. The quality of the constructed graphs are evaluated in the perceptual study.

7. Motion Synthesis Algorithm

Synthesizing motions using a motion graph is equivalent to searching for a path on the graph that matches a set of constraints while minimizing an optimization criteria. Each node on the resulting path is a frame (a state in the graph) that can be played in order to produce an animation. In this section, we first introduce the framework of synthesizing motion clips using stochastic greedy search. Next, we explain the key component in this search algorithm that finds matching motion segments given an input audio segment. Finally, we illustrate how to augment results with facial animation.

7.1. Stochastic Greedy Search

Finding a path through a locomotion graph is typically performed with an optimal search algorithm such as the Anytime A* algorithm [LGT04; SH07] or branch-and-bound algorithms [KGP02]. However, conversational motions do not have a global optimization criteria such as energy consumption. Many possible gestures and postures may be appropriate for a given speech segment. For this reason, our system employs a stochastic greedy search algorithm to find a plausible solution in a best-effort attempt.

Given a novel input audio clip, our system first analyzes and segments the audio using the same technique as used for the original database (Section 4.2). Next, our system synthesizes a motion sequence by matching the audio segments in order. Figure 5 illustrates the search process. Let $[A_0, A_1, A_2, \dots]$ denote the input audio segments. At start time t_0 , the system picks a random set of graph states as the start states. Starting from this set of start states (indicated by gray dots at t_0), the algorithm finds multiple motion segments (all lines between t_0 and t_1) from the motion graph that

matches the input audio segment A_0 . We explain how to find matching motion segments for a specific audio segment A_i in Section 7.2. Each candidate motion segment has a penalty cost that sums all transition costs along the motion path. Note that transition costs are calculated using Equation 3 as described in Section 6. Among all motion segments that satisfy search constraints (the ends of these segments are indicated by the dots at t_1), only n_{save} low-cost candidate segments (orange) are sampled and preserved. In our implementation, we set $n_{save} = 200$. Segments that involve only original transitions will receive a lower cost than those involving many synthetic transitions, and thus are more likely to be preserved for the next round.

In the next round, the end states of the preserved candidate segments (orange dots at t_1) are used as new start states. The algorithm repeats these steps until all input audio segments are processed, and the path with the best score is returned as the solution (purple line in Figure 5, right). The full-body motions are realized by playing the poses p_i that are stored at the states along the path in time order. Root translations and rotations are computed by integrating root velocity $\Delta x^r, \Delta z_i^r$ and angular velocity Δq^{yaw} plus an initial root translation and rotation given chair locations.

To smooth transitions, our system blends the neighboring frames of the synthetic transition using a cosine function, and joint angles are interpolated using the Slerp technique during blending [Sho85].

7.2. Finding Matching Motion Segments

One key component in our search algorithm is to find matching motion segments given an input audio segment. With motion graphs, synthesizing motion segments that matches an audio input is to find paths that satisfy audio constraints.

In classical motion graph techniques, user constraints are commonly expressed as spatial constraints, such as an approximate path on the ground for the character to follow, or a goal location, or an obstacle to avoid or leap over. In our application, user constraints are temporal. Our insight is to find a path where the timings of the new input audio matches the original audio associated with this motion path (refer to *associated audio* below). This insight allows us to model constraints purely on audio signals, and eliminates the step of inferring motion constraints from audio signals and then searching for a matching motion. The new motion should implicitly model the three synchronization behaviors described above while avoiding the need to explicitly construct a model of those synchronization behaviors.

In our system, search constraints consist of four types: phonemic clause, vocal listener response, partner's hesitation pause, and idle. Each constraint restricts different audio types and requires matching different timings between the input audio and the associated audio. Phonemic clause constraints require matching the start/peak/end timings between the input audio and the associated audio. Vocal listener response constraints require matching the start/end timings between the input audio and the associated audio. Partner's hesitation pause constraints require matching the start/end timings in input audio and the associated audio at partner's channel. Idle constraints requires the audio type to be idle but no restriction on timings. In our implementation, we set a tolerance of 0.1s for

matching timings. Based on these definitions, a search constraint could be derived from an input audio segment.

Given an audio segment A_i and a set of start states in the graph, recursive depth-first traversal on motion graphs is performed to find candidate motion paths which satisfy the audio constraint derived from A_i . During the recursive depth-first traversal, our system checks whether the associated audio on the path satisfies audio types and the start/end timings (if the constraint is not idle constraint) at each step. Peak timings will be checked at the last step if the constraint is a phonemic clause constraint.

7.3. Facial Animation

Facial animation is not the focus of this paper. However, to avoid the distraction of rigid, unmoving faces, we augment the motions with facial animations that are driven automatically from the input audio using Motion Builder [Aut19]. Given the intended distant positioning of the background characters in the scene, this facial animation is sufficient to give the faces a sense of life without requiring significant animation effort.

Eyeblinks are inserted by our system automatically according to the audio peaks. Existing studies have revealed that eyeblinks occur near the audio peaks [Sch64; Loe07], and the average frequency of blinks is 26 per minute for conversational motions [BBC*97]. To reproduce realistic eyeblinks, our system inserts an eyeblink at the first audio peak point and at all subsequent audio peaks if at least one second has passed since the last blink. Finally, additional blinks are keyed randomly in any intervals in which no blink has occurred for 5 seconds. The animation of each blink lasts 0.17 seconds, using 0.07s to close the eye and 0.1s to reopen the eye.

8. Experimental Evaluation

In Section 5, we have shown evidence that upheld the three hypotheses on audio-motion co-occurrence between phonemic clause and hand gestures, between vocal listener response and nodding, and between partner hesitation pause and nodding. Accordingly, we have proposed four types of constraints including constraints on phonemic clause (PC), constraints on vocal listener response (LR), constraints on partner hesitation pause (HP), and idle constraints to cover the rest of the frames. In this section, we evaluate the quality of our results using a perceptual study and demonstrate the effects of LR and HP constraints via an ablation study.

Since our technique is only dependent on the start/peak/end timings in prosody features of an input audio clip and its transcript, it generalizes well to conversations from novel subjects. This generalization ability is demonstrated by the results in the supplementary video.

8.1. User Study Setup

To evaluate the quality of synthesized motions and the significance of considering each type of constraint, we designed four pairwise comparisons:

1. comparing the original captured motion to our results with all constraints;
2. comparing synthesized results with all three constraints to result including PC and PH constraints but leaving LR out;
3. comparing synthesized results with all three constraints to results including PC and LR constraints but leaving HP out;
4. comparing synthesized results with all three constraints to results with no constraints, i.e., only the motion quality objectives of the original motion graph.

The first comparison is a quality test of our method. It studies how synthesized motions that maintain all types of correlations compare to original captured ground-truth. The latter three comparisons are ablation studies for the different types of constraints.

During each user study session, a participant is presented with several pairs of side-by-side videos. Each video features the same conversation dialog, with motions sourced from two different techniques showing side-by-side, randomly assigned to the left or right. The total stimuli consisted of 20 (segments) x 4 (pairs) x 10 (repetitions) = 800 clips.

These 20 segments are randomly selected from the test cases in leave-one-out tests. From the 17 captures in our dataset, we used 16 of them to construct a motion graph and left one out. Test cases are obtained by sliding the test audio with a 15-second window and 14-second interval between windows. We use 15 seconds as window size because we hope participants will remain focused while watching the whole clip. With the sliding-window technique, we obtained 116 test cases. Our system can find a solution for 81 of them when searching with all constraints, resulting 69.8% success rate, and can find a solution for 100 of them if removing hesitation pause constraints (86.2% success rate). We used SpireEngine to render the videos for the user study [He20], and implemented the skinning and blendshape computations in the Slang [HFF18] shading language.

Ethical approval was granted for all user studies, and participants were recruited via Amazon MTurk. Each user study session consists 20 rounds of comparisons that are randomly sampled from all stimuli and lasts about 20 minutes. At the end of each pair of videos, the participant is asked which motion they preferred. We adopted the force-choice methodology used by Chang et al. [CYW*16], which forces a choice of the left or the right side. A total of 40 people participated in our experiment, including 25% female and 75% male. 50% of participants were of ages between 25 and 34, and 97.5% of them were under age 55. 80% of the participants have no experience with animation, 17.5% of them have less than 2 years experience with animation, and the rest 2.5% have over 2 years experience.

8.2. Analysis of User Study

Figure 6 visualizes the responses from the user study. The results suggest that the improvement by adding listener response constraints is statistically significant and our results are judged as better than the original motion capture data in 31% of the tests, although the captured motions are still significantly better than our results.

Comparison between ground-truth and our method. A total of 206 questions are asked to compare captured ground-truth motions with motions synthesized using our method respecting all

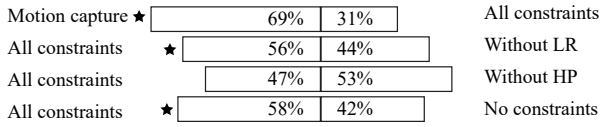


Figure 6: User study that compares original captured motion with our results, and compares the search results that respects all constraints with the search results without LR constraint, without HP constraint, and with no constraints. Results marked with ★ are statistically significant ($p < 0.03$) according to 2-sample test for equality of proportions with continuity correction and two sided alternatives.

three types of correlations. 64 of the responses (31%) favor synthesized motions over the ground-truth. That result is encouraging as people are highly sensitive to motion naturalness. Not surprisingly the original motion capture data is still significantly better than our results ($\chi^2 = 57.563, df = 1, p = 3.272e - 14$). The top comments on why the participant favors the ground-truth over our synthesized motions include that the ground-truth motions are more emotional and more heated, and better matches the sound of hitting the table. This suggests interesting future work to study how emotions and contacts can be leveraged to further improve motion quality.

Ablation study on the impact of each type of correlations. In the ablation tests, 108 out of 193 responses (56%) are in favor of the synthesized results with all three constraints over the results with PC and HP constraints only (leaving listener response constraint out), suggesting that listener response correlation plays a significant role in motion naturalness ($\chi^2 = 5.0155, df = 1, p = 0.02512$).

Meanwhile, 116 out of 200 responses (58%) selected the synthesized results with all three constraints over the search without constraints, as expected. This result suggests that modeling conversation constraints has a positive effect on the generated motions ($\chi^2 = 9.61, df = 1, p = 0.001935$).

Interestingly, we find that 94 out of 201 responses (47%) are in favor of the synthesized results with all three constraints over leaving HP constraint out. This result means that explicitly considering the correlation of partner hesitation pause does not significantly impact on the final motion quality ($\chi^2 = 1.4328, df = 1, p = 0.2313$). This result is likely because the vocal listener response already covers the hesitating motions well, thus explicitly labeling and matching hesitating states are not always necessary.

8.3. Computation Performance

As discussed in Section 6, we adaptively downsample frames and only allow transitions at key states instead of at each frame while constructing the graph. This optimization reduces the number of candidate states for synthetic transitions by 78%, which translates into a 20x speedup over a naive graph construction process that considers all the states as candidates. Our motion graph consists of approximately 50,000 states. Table 3 shows the wall-clock time of our graph construction and search algorithms. The run-time performance is measured on a machine with an Intel i9-9980XE CPU

procedure	avg. tim (s)	std. dev.
graph construction	47.94	3.30
search	1.16	0.56

Table 3: Wall-clock time for search and graph construction in the leave-one-out tests. From the 17 captures in our dataset, we used 16 of them to construct a motion graph and left one out. A total of 116 test cases are obtained by sliding the test audio with 15-second window and 14-second interval. Database in leave-one-out tests has about 50,000 frames at 30fps.

(18 cores at 3GHz) and 64GB memory. The current implementation uses at most 4 cores and 16GB memory. On average, our graph construction takes 50s. The average wall-clock time required to search for a 15 s motion clip for one character is 1.16 s, which potentially allows our system to be used interactively in scenarios where proper scheduling is implemented.

9. Conclusion and Discussion

In this paper, we present a system for synthesizing conversational motions that respect both speaker and listener behaviors. The key insight is that the self and interaction synchrony in the original recordings can be reproduced by finding a path on the motion graph where the timing of the recorded audio signatures in the motion graph database matches that of the new input audio in both the speaking and listening channels. We contribute a database of face-to-face dyadic conversations featuring a variety of emotions and scenarios (friends talking, arguing couples, job interviews). We extend the motion graph technique to include conversational constraints based on statistics computed from a motion capture database and add a threshold tuning scheme during graph construction based on the distance distribution in the database.

In designing proper constraints for motion synthesis, we tested three hypotheses about audio-motion co-occurrence for speaker and listener behaviors, and further tested their effects in search by including them as constraints. Our results show the importance of considering vocal listener response as constraints, in addition to the well-accepted constraints on phonemic clause. Furthermore, we developed a stochastic greedy search algorithm to efficiently generate motion sequence that respects these audio constraints.

While these results have shown significant improvement by respecting listener response constraints, they also illustrate some limitations in our method. The biggest limitation is that our method does not model emotions and semantics of a conversation. This leads to overly neutral gestures for emotional moments, e.g. moments of excitement or anger, as compared to the ground-truth motion. Modeling emotion as an audio magnitude constraint that measures excitement, or as a frequency constraint that measures how fast a speaker talks would be an interesting direction for future work. Semantics could be supported by detecting iconic words and using text constraints to search and synthesize motions. While our system can easily be extended with more features and search constraints, we would likely need a significantly larger database to ensure that an acceptable path still exists in the motion graph for the entire input audio when those additional constraints are applied.

Our system focuses on studying the audio-motion co-occurrence on head and hand motions. In addition to these motions, other type of body movements such as changing the posture to move closer to the conversation partner or mirroring the speaker posture as a listener may also be important. Studying how to model these behaviors and how these behaviors affect motion realism would be an interesting direction for future work.

Our method currently does not model contact with the environment. When the input audio contains a collision sound of a character's hand with the table, the synthesized motion should also include a corresponding collision. One way to factor in this kind of environment interaction is to adopt the hierarchical motion editing technique introduced by Lee and colleagues [LS99]. Furthermore, handling the contact with objects and self contact are also an interesting future system extension.

As with all motion graph approaches, synthesized motions maintain individual styles in the dataset. For example, the female character in our dataset has a variety of postures, such as sitting straight or slouching. At the same time, we have a limitation as other motion graph approach that the range of synthesized motions is limited to what is in the database. For example, our system cannot synthesize motions with characters sitting side by side or conversations that involve other emotions beyond those recorded in the database. Extending our system to handle these scenarios would likely require capturing a dramatically larger database. One way to augment data is re-timing motions, such that motion graphs have more flexibility to match an input audio clip by speeding up or slowing down captured motions. Another way is to have a larger database with pose-tracking techniques. With the development of pose-tracking techniques in computer vision, we expect that obtaining a large-scale motion database may soon become significantly easier [CSWS17]. We are interested in scaling up our system to handle databases that are larger by several orders of magnitude. A significantly larger database would likely require a hierarchical approach for graph construction and searching [LCL06]. We hypothesize that such a hierarchy can be built by constructing a set of smaller subgraphs representing particular kinds of motions based perhaps on emotion or on semantics. These subgraphs can then be connected by constructing synthetic transitions between them.

Besides motion graph approaches, we have seen a growing interest in learning gestures using deep neural networks. Kucherenko and colleagues introduce a model that learns Japanese gestures using a denoising variational autoencoder [KHH*19]. Ferstl and colleagues address gesture dynamics by training generative adversarial models with gesture phasing [FNM20]. Alexanderson and colleagues propose learning gestures using normalizing flow with a style-control framework and experiment controls over the hand height, velocity, gesture radius, and gesture symmetry [AHKB20]. Although their models are trained for a single subject, their frameworks are good starting points for designing dyadic conversation models. In addition to dyadic conversation, three-party conversations are also studied in recent work [JDZD19; dCYS*19; JSCS19]. While motion graph approaches are good at maintaining styles and high motion quality in terms of dynamics and smoothness from the captured dataset, learning based methods have the

potential for fast inference after models are trained, and may find more diverse results.

10. Acknowledgement

The authors would like to thank our artists Michelle Ma, Eric Yu, Kevin Carlos, Melanie Danver for their help in modeling the scene, rigging models, and improving the quality of finger motions in our motion database. The authors would like to thank Justin Macey for his assistance in collecting and cleaning the motion capture data. The authors would like to thank the participants for attending the user study. Special thanks to Yong He for his useful discussions.

References

- [AFO03] ARIKAN, OKAN, FORSYTH, DAVID A., and O'BRIEN, JAMES F. "Motion synthesis from annotations". *ACM Trans. Graph.* 22 (2003), 402–408 3.
- [AHKB20] ALEXANDERSON, SIMON, HENTER, GUSTAV EJE, KUCHERENKO, TARAS, and BESKOW, JONAS. "Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows". *Computer Graphics Forum*. Vol. 39. 2. Wiley Online Library. 2020, 487–496 10.
- [AMMS19] AHUJA, CHAITANYA, MA, SHUGAO, MORENCY, LOUIS-PHILIPPE, and SHEIKH, YASER. "To React or not to React: End-to-End Visual Pose Forecasting for Personalized Avatar during Dyadic Conversations". *2019 International Conference on Multimodal Interaction*. 2019, 74–84 3.
- [Aut19] AUTODESK. *Motion Builder Docs*. <http://help.autodesk.com/view/MOBPRO/2019/ENU/>. 2019 8.
- [BBC*97] BENTIVOGLIO, ANNA RITA, BRESSMAN, SUSAN B, CASSETTA, EMANUELE, et al. "Analysis of blink rate patterns in normal subjects". *Movement disorders* 12.6 (1997), 1028–1034 8.
- [BD62] BOOMER, DONALD S and DITTMANN, ALLEN T. "Hesitation pauses and juncture pauses in speech". *Language and speech* 5.4 (1962), 215–220 1.
- [BRB17] BARBULESCU, ADELA, RONFARD, RÉMI, and BAILLY, GÉRARD. "A generative audio-visual prosodic model for virtual actors". *IEEE computer graphics and applications* 37.6 (2017), 40–51 3.
- [BSB13] BEN YOUSSEF, ATEF, SHIMODAIRA, HIROSHI, and BRAUDE, DAVID A. "Articulatory features for speech-driven head motion synthesis". *Proceedings of Interspeech, Lyon, France* (2013) 3.
- [CM11] CHIU, CHUNG-CHENG and MARSELLA, STACY. "How to train your avatar: A data driven approach to gesture generation". *International Workshop on Intelligent Virtual Agents*. Springer. 2011, 127–140 1, 2.
- [CM14] CHIU, CHUNG-CHENG and MARSELLA, STACY. "Gesture Generation with Low-dimensional Embeddings". *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*. AAMAS '14. Paris, France: International Foundation for Autonomous Agents and Multiagent Systems, 2014, 781–788. ISBN: 978-1-4503-2738-1. URL: <http://dl.acm.org/citation.cfm?id=2615731.2615857> 1, 2.
- [CMM15] CHIU, CHUNG-CHENG, MORENCY, LOUIS-PHILIPPE, and MARSELLA, STACY. "Predicting co-verbal gestures: a deep and temporal modeling approach". *International Conference on Intelligent Virtual Agents*. Springer. 2015, 152–166 2.
- [CPB*94] CASSELL, JUSTINE, PELACHAUD, CATHERINE, BADLER, NORMAN, et al. "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents". *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. ACM. 1994, 413–420 3.
- [CSWS17] CAO, ZHE, SIMON, TOMAS, WEI, SHIH-EN, and SHEIKH, YASER. "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". *CVPR*. 2017 10.

- [CVB01] CASSELL, JUSTINE, VILHJÁLMSSON, HANNES HÖGNI, and BICKMORE, TIMOTHY. "BEAT: The Behavior Expression Animation Toolkit". *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '01. New York, NY, USA: ACM, 2001, 477–486. ISBN: 1-58113-374-X. DOI: [10.1145/383259.383315](https://doi.org/10.1145/383259.383315). URL: <http://doi.acm.org/10.1145/383259.383315>.
- [CYW*16] CHANG, HUIWEN, YU, FISHER, WANG, JUE, et al. "Automatic triage for a photo series". *ACM Transactions on Graphics (TOG)* 35.4 (2016), 1–10 [8](#).
- [dCYS*19] De CONINCK, FERDINAND, YUMAK, ZERRIN, SANDINO, GUNTUR, et al. "Non-Verbal Behavior Generation for Virtual Characters in Group Conversations". *AIVR*. 2019, 41–49 [10](#).
- [DPA13] DING, YU, PELACHAUD, CATHERINE, and ARTIERES, THIERRY. "Modeling multimodal behaviors from speech prosody". *International Workshop on Intelligent Virtual Agents*. Springer. 2013, 217–228 [3](#).
- [DZX15] DING, CHUANG, ZHU, PENGCHENG, and XIE, LEI. "Blstm neural networks for speech driven head motion synthesis". *Sixteenth Annual Conference of the International Speech Communication Association*. 2015 [3](#).
- [EMO10] ENNIS, CATHY, McDONNELL, RACHEL, and O'SULLIVAN, CAROL. "Seeing is Believing: Body Motion Dominates in Multisensory Conversations". *ACM SIGGRAPH 2010 Papers*. SIGGRAPH '10. Los Angeles, California: ACM, 2010, 91:1–91:9. ISBN: 978-1-4503-0210-4. DOI: [10.1145/1833349.1778828](https://doi.org/10.1145/1833349.1778828). URL: <http://doi.acm.org/10.1145/1833349.1778828>.
- [EWGS13] EYBEN, FLORIAN, WENINGER, FELIX, GROSS, FLORIAN, and SCHULLER, BJÖRN. "Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor". *Proceedings of the 21st ACM International Conference on Multimedia*. MM '13. Barcelona, Spain: ACM, 2013, 835–838. ISBN: 978-1-4503-2404-5. DOI: [10.1145/2502081.2502224](https://doi.org/10.1145/2502081.2502224). URL: <http://doi.acm.org/10.1145/2502081.2502224>.
- [FMA*14] FERNÁNDEZ-BAENA, ADISO, MONTAÑO, RAÚL, ANTONIO-JOAN, MARC, et al. "Gesture synthesis adapted to speech emphasis". *Speech communication* 57 (2014), 331–350 [2](#).
- [FNM20] FERSTL, YLVA, NEFF, MICHAEL, and McDONNELL, RACHEL. "Adversarial gesture generation with realistic gesture phasing". *Computers & Graphics* (2020) [10](#).
- [GLM17] GREENWOOD, DAVID, LAYCOCK, STEPHEN, and MATTHEWS, IAIN. "Predicting head pose in dyadic conversation". *International Conference on Intelligent Virtual Agents*. Springer. 2017, 160–169 [3](#).
- [GOL*06] GRATCH, JONATHAN, OKHMATOVSKAIA, ANNA, LAMOTHE, FRANCOIS, et al. "Virtual rapport". *International Workshop on Intelligent Virtual Agents*. Springer. 2006, 14–27 [3](#).
- [GWG*07] GRATCH, JONATHAN, WANG, NING, GERTEN, JILLIAN, et al. "Creating rapport with virtual agents". *International workshop on intelligent virtual agents*. Springer. 2007, 125–138 [3](#).
- [He20] HE, YONG. *Spire Engine*. 2020. URL: <https://github.com/spire-engine/spire-engine> [8](#).
- [HFF18] HE, YONG, FATAHALIAN, KAYVON, and FOLEY, TIM. "Slang: language mechanisms for extensible real-time shading systems". *ACM Transactions on Graphics (TOG)* 37.4 (2018), 1–13 [8](#).
- [HLL16] HYUN, KYUNGLYUL, LEE, KYUNGHO, and LEE, JEHEE. "Motion grammars for character animation". *Computer Graphics Forum*. Vol. 35. 2. Wiley Online Library. 2016, 103–113 [3](#).
- [IBM18] IBM. *IBM Cloud Docs*. <https://console.bluemix.net/docs/services/speech-to-text/index.html>. Accessed: 2018. 2018 [3](#), [4](#).
- [JDZD19] JIN, AOBO, DENG, QIXIN, ZHANG, YUTING, and DENG, ZHIGANG. "A Deep Learning-Based Model for Head and Eye Motion Generation in Three-party Conversations". *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 2.2 (2019), 1–19 [3](#), [10](#).
- [JHS12] JÖRG, SOPHIE, HODGINS, JESSICA, and SAFONOVA, ALLA. "Data-driven Finger Motion Synthesis for Gesturing Characters". *ACM Trans. Graph.* 31.6 (Nov. 2012), 189:1–189:7. ISSN: 0730-0301. DOI: [10.1145/2366145.2366208](https://doi.org/10.1145/2366145.2366208). URL: <http://doi.acm.org/10.1145/2366145.2366208>.
- [JSCS19] JOO, HANBYUL, SIMON, TOMAS, CIKARA, MINA, and SHEIKH, YASER. "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, 10873–10883 [3](#), [10](#).
- [JT05] JAN, DUŠAN and TRAUM, DAVID R. "Dialog Simulation for Background Characters". *Intelligent Virtual Agents*. Ed. by PANAYIOTOPOULOS, THEMIS, GRATCH, JONATHAN, AYLETT, RUTH, et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, 65–74. ISBN: 978-3-540-28739-1 [3](#).
- [KB12] KOPP, STEFAN and BERGMANN, KIRSTEN. "Individualized gesture production in embodied conversational agents". *Human-Computer Interaction: The Agency Perspective*. Springer, 2012, 287–301 [2](#).
- [KGP02] KOVAR, LUCAS, GLEICHER, MICHAEL, and PIGHIN, FRÉDÉRIC. "Motion Graphs". *ACM Trans. Graph.* 21.3 (July 2002), 473–482. ISSN: 0730-0301. DOI: [10.1145/566654.566605](https://doi.org/10.1145/566654.566605). URL: <http://doi.acm.org/10.1145/566654.566605> [2](#), [3](#), [7](#).
- [KH97] KNAPP, MARK L and HALL, JUDITH A. *Nonverbal communication in human interaction*. Harcourt Brace College Publishers, 1997 [1](#), [3](#).
- [KHH*19] KUCHERENKO, TARAS, HASEGAWA, DAI, HENTER, GUSTAV EJE, et al. "Analyzing input and output representations for speech-driven gesture generation". *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 2019, 97–104 [10](#).
- [Kip01] KIPP, MICHAEL. "Anvil-a generic annotation tool for multimodal dialogue". *Seventh European Conference on Speech Communication and Technology*. 2001 [3](#).
- [KKM*06] KOPP, STEFAN, KRENN, BRIGITTE, MARSELLA, STACY, et al. "Towards a common framework for multimodal generation: The behavior markup language". *International workshop on intelligent virtual agents*. Springer. 2006, 205–217 [2](#).
- [KNA07] KIPP, MICHAEL, NEFF, MICHAEL, and ALBRECHT, IRENE. "An annotation scheme for conversational gestures: how to economically capture timing and form". *Language Resources and Evaluation* 41.3-4 (2007), 325–339 [3](#).
- [LCL06] LEE, KANG HOON, CHOI, MYUNG GEOL, and LEE, JEHEE. "Motion patches: building blocks for virtual environments annotated with motion data". *ACM Transactions on Graphics (TOG)*. Vol. 25. 3. ACM. 2006, 898–906 [10](#).
- [LCR*02] LEE, JEHEE, CHAI, JINXIANG, REITSMA, PAUL S. A., et al. "Interactive Control of Avatars Animated with Human Motion Data". *ACM Trans. Graph.* 21.3 (July 2002), 491–500. ISSN: 0730-0301. DOI: [10.1145/566654.566607](https://doi.org/10.1145/566654.566607). URL: <http://doi.acm.org/10.1145/566654.566607> [2](#), [3](#), [6](#).
- [LGT04] LIKHACHEV, MAXIM, GORDON, GEOFFREY J, and THRUN, SEBASTIAN. "ARA*: Anytime A* with provable bounds on sub-optimality". *Advances in neural information processing systems*. 2004, 767–774 [7](#).
- [LKT10] LEVINE, SERGEY, KRÄHENBÜHL, PHILIPP, THRUN, SEBASTIAN, and KOLTUN, VLADLEN. "Gesture controllers". *ACM Transactions on Graphics (TOG)*. Vol. 29. 4. ACM. 2010, 124 [1](#), [2](#).
- [LM06] LEE, JINA and MARSELLA, STACY. "Nonverbal behavior generator for embodied conversational agents". *International Workshop on Intelligent Virtual Agents*. Springer. 2006, 243–255 [2](#).
- [LM13] LHOMMET, MARGAUX and MARSELLA, STACY C. "Gesture with meaning". *International Workshop on Intelligent Virtual Agents*. Springer. 2013, 303–312 [2](#).

- [Loe07] LOEHR, DANIEL. "Aspects of rhythm in gesture and speech". *Gesture* 7 (Jan. 2007), 179–214. DOI: [10.1075/gest.7.2.04loe8](https://doi.org/10.1075/gest.7.2.04loe8).
- [LS99] LEE, JEHEE and SHIN, SUNG YONG. "A Hierarchical Approach to Interactive Motion Editing for Human-like Figures". *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH fffdfffdfffd99. USA: ACM Press/Addison-Wesley Publishing Co., 1999, 39fffdfffdfffd48. ISBN: 0201485605. DOI: [10.1145/311535.311539](https://doi.org/10.1145/311535.311539). URL: <https://doi.org/10.1145/311535.311539>.
- [LTK09] LEVINE, SERGEY, THEOBALT, CHRISTIAN, and KOLTUN, VLADLEN. "Real-time prosody-driven synthesis of body language". *ACM Transactions on Graphics* 28.5 (2009), 1. ISSN: 07300301. DOI: [10.1145/1618452.1618518](https://doi.org/10.1145/1618452.1618518). URL: <http://portal.acm.org/citation.cfm?doid=1618452.1618518> 1, 2, 5.
- [Mar02] MARTELL, CRAIG. "Form: An extensible, kinematically-based gesture annotation scheme". *Seventh International Conference on Spoken Language Processing*. 2002 3.
- [MC12] MIN, JIANYUAN and CHAI, JINXIANG. "Motion Graphs++: A Compact Generative Model for Semantic Motion Analysis and Synthesis". *ACM Trans. Graph.* 31.6 (Nov. 2012), 153:1–153:12. ISSN: 0730-0301. DOI: [10.1145/2366145.2366172](https://doi.org/10.1145/2366145.2366172). URL: <http://doi.acm.org/10.1145/2366145.2366172> 3.
- [MEDO09] McDONNELL, RACHEL, ENNIS, CATHY, DOBBYN, SIMON, and O'SULLIVAN, CAROL. "Talking Bodies: Sensitivity to Desynchronization of Conversations". *ACM Trans. Appl. Percept.* 6.4 (Oct. 2009), 22:1–22:8. ISSN: 1544-3558. DOI: [10.1145/1609967.1609969](https://doi.org/10.1145/1609967.1609969). URL: <http://doi.acm.org/10.1145/1609967.1609969> 3.
- [MP07] McCANN, JAMES and POLLARD, NANCY. "Responsive Characters from Motion Fragments". *ACM Trans. Graph.* 26.3 (July 2007). ISSN: 0730-0301. DOI: [10.1145/1276377.1276385](https://doi.org/10.1145/1276377.1276385). URL: <http://doi.acm.org/10.1145/1276377.1276385> 3.
- [MTCY10] MU, KAIHUI, TAO, JIANHUA, CHE, JIANFENG, and YANG, MINGHAO. "Mood Avatar: Automatic Text-driven Head Motion Synthesis". *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. ICMI-MLMI '10. Beijing, China: ACM, 2010, 37:1–37:4. ISBN: 978-1-4503-0414-6. DOI: [10.1145/1891903.1891951](https://doi.org/10.1145/1891903.1891951). URL: <http://doi.acm.org/10.1145/1891903.1891951> 3.
- [MXL*13] MARSELLA, STACY, XU, YUYU, LHOMMET, MARGAUX, et al. "Virtual character performance from speech". *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. ACM, 2013, 25–35 3.
- [NKAS08] NEFF, MICHAEL, KIPP, MICHAEL, ALBRECHT, IRENE, and SEIDEL, HANS-PETER. "Gesture modeling and animation based on a probabilistic re-creation of speaker style". *ACM Transactions on Graphics (TOG)* 27.1 (2008), 5 2.
- [Par12] PARENT, RICK. *Computer animation: algorithms and techniques*. Newnes, 2012 5, 6.
- [SB15] SADOUGHI, NAJMEH and BUSO, CARLOS. "Retrieving Target Gestures Toward Speech Driven Animation with Meaningful Behaviors". *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ICMI '15. Seattle, Washington, USA: ACM, 2015, 115–122. ISBN: 978-1-4503-3912-4. DOI: [10.1145/2818346.2820750](https://doi.org/10.1145/2818346.2820750). URL: <http://doi.acm.org/10.1145/2818346.2820750> 2.
- [Sch64] SCHEFLEN, ALBERT E. "The Significance of Posture in Communication Systems". *Psychiatry* 27.4 (1964). PMID: 14216879, 316–331. DOI: [10.1080/00332747.1964.11023403](https://doi.org/10.1080/00332747.1964.11023403). eprint: <https://doi.org/10.1080/00332747.1964.11023403>. URL: <https://doi.org/10.1080/00332747.1964.11023403> 8.
- [SDO*04] STONE, MATTHEW, DeCARLO, DOUG, OH, INSUK, et al. "Speaking with hands: Creating animated conversational characters from recordings of human performance". *ACM Transactions on Graphics (TOG)*. Vol. 23. 3. ACM, 2004, 506–513 2.
- [SH07] SAFONOVA, ALLA and HODGINS, JESSICA K. "Construction and Optimal Search of Interpolated Motion Graphs". *ACM Trans. Graph.* 26.3 (July 2007). ISSN: 0730-0301. DOI: [10.1145/1276377.1276510](https://doi.org/10.1145/1276377.1276510). URL: <http://doi.acm.org/10.1145/1276377.1276510> 2, 3, 6, 7.
- [Sho85] SHOEMAKE, KEN. "Animating rotation with quaternion curves". *ACM SIGGRAPH computer graphics*. Vol. 19. 3. ACM, 1985, 245–254 7.
- [SSH*12] SUN, LIBO, SHOULSON, ALEXANDER, HUANG, PENGFEI, et al. "Animating synthetic dyadic conversations with variations based on context and agent attributes". *Computer Animation and Virtual Worlds* 23.1 (2012), 17–32 3.
- [Tar72] TARIAN, ROBERT. "Depth-first search and linear graph algorithms". *SIAM journal on computing* 1.2 (1972), 146–160 6.
- [TG67] T. DITTMANN, ALLEN and G. LLEWELLYN, LYNN. "The Phonic Clause as a Unit of Speech Decoding". *Journal of personality and social psychology* 6 (Aug. 1967), 341–9. DOI: [10.1037/h0024739](https://doi.org/10.1037/h0024739) 1, 2.
- [TLP07] TREUILLE, ADRIEN, LEE, YONGJOON, and POPOVIĆ, ZORAN. "Near-optimal character animation with continuous control". *ACM SIGGRAPH 2007 papers*. 2007, 7–es 2.
- [TMMK08] THIEBAUX, MARCUS, MARSELLA, STACY, MARSHALL, ANDREW N, and KALLMANN, MARCELO. "Smartbody: Behavior realization for embodied conversational agents". *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*. 2008, 151–158 2.
- [VARM15] VANDEVENTER, JASON, AUBREY, ANDREW J, ROSIN, PAUL L, and MARSHALL, A DAVID. "4D Cardiff Conversation Database (4D CCDb): a 4D database of natural, dyadic conversations." *AVSP*. 2015, 157–162 3.
- [WWS*15] WHEATLAND, NKENG, WANG, YINGYING, SONG, HUAGUANG, et al. "State of the Art in Hand and Finger Modeling and Animation". *Computer Graphics Forum* (2015). DOI: [10.1111/cg.12595](https://doi.org/10.1111/cg.12595) 3.
- [XPM14] XU, YUYU, PELACHAUD, CATHERINE, and MARSELLA, STACY. "Compound Gesture Generation: A Model Based on Ideational Units". *Intelligent Virtual Agents*. Ed. by BICKMORE, TIMOTHY, MARSELLA, STACY, and SIDNER, CANDACE. Cham: Springer International Publishing, 2014, 477–491 2.