Accurate Synchronization of Gesture and Speech for Conversational Agents using Motion Graphs

Jianfeng Xu, Yuki Nagai, Shinya Takayama and Shigeyuki Sakazawa
Media and HTML5 Application Laboratory, KDDI R&D Laboratories, Inc., Fujimino-shi, Japan

Keywords: Conversational Agents, Multimodal Synchronization, Gesture, Motion Graphs, Dynamic Programming.

Abstract: Multimodal representation of conversational agents requires accurate synchronization of gesture and speech. For this purpose, we investigate the important issues in synchronization as a practical guideline for our algorithm design through a precedent case study and propose a two-step synchronization approach. Our case study reveals that two issues (i.e. duration and timing) play an important role in the manual synchronizing of gesture with speech. Considering the synchronization problem as a motion synthesis problem instead of a behavior scheduling problem used in the conventional methods, we use a motion graph technique with constraints on gesture structure for coarse synchronization in a first step and refine this further by shifting and scaling the motion in a second step. This approach can successfully synchronize gesture and speech with respect to both duration and timing. We have confirmed that our system makes the creation of attractive content easier than manual creation of equal quality. In addition, subjective evaluation has demonstrated that the proposed approach achieves more accurate synchronization and higher motion quality than the state-of-the-art method.

1 INTRODUCTION

Synchronization of gesture and speech is essential in conversational agents (Cassell et al., 2001; Nishida, 2007), and animators spend an enormous amount of effort using either intuition or motion capture to achieve it. Although synchronization description schemes (e.g. the Behavior Markup Language, BML (Kopp et al., 2006)) have been proposed and widely used in the academic field (Marsella et al., 2013), it remains a challenge to produce such synchronization automatically. In this paper, we propose an effective synchronization solution using a different philosophy than the conventional methods. The basic idea is to consider the synchronization problem as a motion synthesis problem instead of a behavior scheduling problem where the gesture motions are re-scheduled in the timeline according to the speech (Cassell et al., 2001; Neff et al., 2008; Marsella et al., 2013). The greatest benefit is that we can significantly improve both the synchronization accuracy and motion quality simultaneously.

Psychological research has shown that gesture and speech have a very complex relationship (McNeill, 1985). Although they are believed to share a common thought source, the hypothesized *growth point* (McNeill, 2005), the relationship between gesture and speech is many-to-many. For example, to emphasize a word in an utterance, one may use a beat gesture, a nod, or an eyebrow. On the other hand, a nod may mean confirmation rather than emphasis. Furthermore, many other factors affect the relationship between gesture and speech such as personality, gender, culture, conversational context, etc. (Neff et al., 2008; Marsella et al., 2013). For example, Japanese talk to each other with nodding, but a nod means attentiveness rather than agreement. In addition, human perception is highly sensitive to the synchronization of speech and gesture. Although the temporal tolerance is basically dependent on the content and human subject, it is believed that high accuracy (e.g. 150ms) is required by most of human subjects (Miller and D'Esposito, 2005). For example, a level of the phoneme is perceptible for most of people to watch a speaker's mouth movements (McGurk and MacDonald, 1976).

In the field of conversational agents (Cassell et al., 2001) and human-robot interaction (Ng-Thow-Hing et al., 2010), the synchronization of gesture and speech is based on a common practice that synchronizes the gesture stroke (see the definition in Section 3) with the accented syllable of the accompanying speech (Neff et al., 2008). Based on this, the latest system (Marsella et al., 2013) uses an offset/scaling

Xu J., Nagai Y., Takayama S. and Sakazawa S..
Accurate Synchronization of Gesture and Speech for Conversational Agents using Motion Graphs.
DOI: 10.5220/0004748400050014
In *Proceedings of the 6th International Conference on Agents and Artificial Intelligence* (ICAART-2014), pages 5-14
ISBN: 978-989-758-016-1

technique for synchronization of gesture and speech.

Due to the absence of practical guidelines on automating synchronization, in this paper, we investigate the important issues in the manual synchronization of gesture with speech. Two similar but not identical scripts are prepared. We examine the differences among manually created animations of the two scripts and discover practical guidelines for our algorithm. As a result, the above case study reveals that two issues (i.e. duration and timing) play an important role in the manual synchronizing of gesture with speech.

To automatically produce accurate synchronization of gesture and speech, in essence, we consider the synchronization problem to be a motion synthesis problem with certain constraints. At the same time, we observe that many gestures are cyclic or use similar poses, which results in the adoption of the motion graph technique (Kovar et al., 2002; Arikan and Forsyth, 2002; Lee et al., 2002). The motion graph technique is reported to be a powerful tool for synthesizing natural motion from an original motion with constraints such as motion duration (Kovar et al., 2002; Arikan and Forsyth, 2002; Lee et al., 2002). In addition, it is well known that gestural motion has a special temporal structure, which is important in the synchronization of gesture and speech (Neff et al., 2008). Our experimental results show that the proposed algorithm works well in our scenarios.

In this paper, our technical contributions are summarized as follows.

1. With a case study, we have discovered that two issues (i.e. duration and timing) play an important role in the manual synchronizing of gesture with speech, which becomes a practical guideline for our algorithm.
2. We propose a two-step approach based on the motion graph technique (Kovar et al., 2002; Arikan and Forsyth, 2002; Lee et al., 2002) with a temporal structure of gestures that deals with the issues of duration and timing. In the first step, we synthesize a new motion that is coarsely synchronized with the speech. In the second step, we further refine the synchronization by shifting and scaling the synthesized motion.

In addition, we implement our system as an authoring tool, which outputs a synthesized animation with facial expressions and gestures synchronized with the audio signal. To use the authoring tool, we input an audio file that records a speaker's voice and its script with timing tags, and then assign the desired gesture and emotion for each sentence in the script. As a basic unit of generating animation, the authoring tool synthesizes a new motion with facial expressions that is synchronized with the input speech sentence by sentence. As one of the target applications, we create some animations for education by our authoring tool, where we get rather positive feedback from university students in an evaluation experiment.

This paper is organized as follows. Section 2 surveys some techniques related to our approach. As the core of this paper, Section 3 describes the precedent case study and the algorithm for synchronizing gesture and speech in detail. In Section 4, we briefly introduce our authoring tool that can output a rich animation with facial expressions and gestures synchronized with speech. In Section 5, we report our experimental results, including a subjective evaluation that compares our approach to the conventional method (Marsella et al., 2013). In Section 6, we briefly introduce the applications to education using the authoring tool. Finally, we present our conclusions and future work in Section 7.

2 RELATED WORK

In the last two decades, many embodied conversational agents (ECAs) have been developed whose multimodal representation has been shown to be appealing to users. Most ECA systems are composed of three sequentially executed blocks: audio/text understanding, behavior selection, and behavior editing (Cassell et al., 2001). Many techniques are available for audio/text understanding (Marsella et al., 2013; Stone et al., 2004), which provides the needed acoustic and semantic information for behavior selection. Especially, by performing deep analysis of syntactic, semantic and rhetorical structures of the utterance, (Marsella et al., 2013) achieves semantically appropriate behavior, which is their central contribution. For behavior selection, the de-facto method is a rule-based approach (Cassell et al., 2001; Ng-Thow-Hing et al., 2010; Marsella et al., 2013) that maps keywords to behaviors or behavior categories by a large set of predefined rules. For behavior editing, existing systems focus mainly on hand trajectory modification by physical simulation (Neff et al., 2008) or cubic spline interpolation (Ng-Thow-Hing et al., 2010). Unfortunately, there are as yet few techniques for multimodal synchronization, although this is believed to be essential to properly convey the emotional component of communication. Lip synchronization is widely used in ECA systems thanks to TTS (text to speech) techniques (Dutoit, 2001). For synchronization of gesture and speech, the early work (Cassell et al., 2001) aligns the timing of gesture motions with text words, and the latest paper (Marsella et al., 2013) improves

the synchronization level to gesture phases using the offset and scaling approach, in which the timing of the stroke phase in gesture motion is aligned with the speech. However, such an approach will change the quality of motions and even the emotional state of gestures if the scaling factor is too large.

On the other hand, psychological research on multimodal synchronization continues and has provided many valuable insights for ECA systems. The hypothesized *growth point*, proposed by (McNeill, 2005), is a well-known theory to explain the phenomenon of synchronization of gesture and speech. Moreover, based on the fact that the structure of gesture is different from other human motions like dancing, (Neff et al., 2008) pointed out that the stroke phase should be synchronized with the accented syllable of the accompanying speech. However, these discoveries only specify a result without providing the processing needed to produce it automatically (Kopp et al., 2006).

In addition, usage of ECAs has been demonstrated to be potential in many fields besides the original interface agents in human computer interaction such as Rea (Cassell et al., 1999), Greta (Niewiadomski et al., 2009; Huang and Pelachaud, 2012), and RealActor (Čerekovič and Pandžič, 2011). For instance, a virtual human presenter is designed for weather forecasting, and slide presentation in the work of (Noma et al., 2000). It is reported that a navigation agent such as a guide in a theater building (van Luin et al., 2001) or a university campus (Oura et al., 2013) is helpful to visitors. Especially, digital education is attracting much attention from both academic and industrial fields with the rapid development of tablet and mobile devices. Although some systems such as (Beskow et al., 2004) are reported, this paper will further give the evaluation of agent's effectiveness in education.

3 SYNCHRONIZING GESTURE WITH SPEECH

In state-of-the-art systems (Neff et al., 2008; Marsella et al., 2013), the number of gestures in the database is not very large, amounting to just dozens of available gestures, which is comparable to the number used by a TV talk show host (Neff et al., 2008). However, a human being performs each gesture variably according to the context, e.g. synchronizing the gesture with speech. Therefore, gesture variation is rather large in human communication. This indicates that the task of gesture synchronization is in essence to synthesize a new motion from a generic one for a particular portion of speech, which is a motion synthesis problem (Kovar et al., 2002; Arikan and Forsyth, 2002; Lee et al., 2002). As far as we know, this viewpoint is different from the behavior scheduling concept used in conventional methods (Cassell et al., 2001; Neff et al., 2008; Marsella et al., 2013).

As described before, most gestures have a temporal structure with multiple consecutive movement phases including a preparation (P) phase, a stroke (S) phase, and a retraction (R) phase (Neff et al., 2008). Only the S phase is essential, as it is the most energetic and meaningful phase of the gesture. In this paper, the parts before and after the S phase are denoted as the P and R phases, respectively. Please see Fig. 1 as a reference.

3.1 Investigation by Case Study

In this section, we look for the guidelines for our synchronization algorithm through a case study. Consider a scenario in which a virtual agent talks with you about your diet when you eat ice cream on two days, which may exceed your preferred calorie consumption. We prepare the following two scripts, with similar content but different lengths and emotional states. Note that both scripts are translated from Japanese. Only Japanese versions are used in the case study.

1. (for Day One) Good morning. Ah–, you must have eaten ice cream last night! *You solemnly promised to go on a diet!* Well, you gotta do some walking today. Since the weather is fine, let's go now.
2. (for Day Two) Good morning. Ah—, you must have eaten ice-cream again last night. That's two days in a row!! *Were you serious when you promised to go on a diet?* Well, now you gotta walk that much farther. The forecast says rain this afternoon, so let's go now.

The speech is recorded by a narrator. A staff member creates the animations manually using the authoring tool MikuMikuDance (no relation to our authoring tool), in which the facial expression and body pose are independently edited in each key-frame after loading a suitable gesture from a motion capture database. First, the staff member manually creates the animation for Script #1 sentence by sentence and uses it directly in Script #2 for the same block of sentences. Then, the staff member manually improves the animation for Script #2. With this processing, we analyze the important issues in manual operation by noting the differences among the animations. Firstly, we observe that our staff member needs to modify the *duration* of gestures to fit with the speech. For example, the punching gesture is used for the italic parts in both

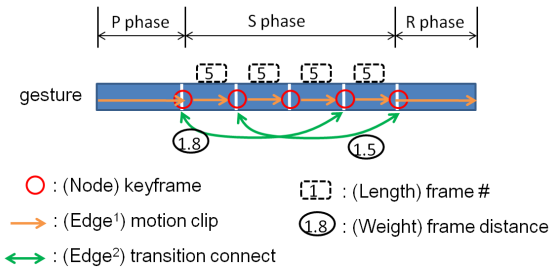

Figure 1: A graph structure with node, edge, weight, and length is constructed in the S phase.

scripts. The cycle number for Script #1 is four while it is changed to six for Script #2. Secondly, we observe that our staff member changes the *timing* of the gesture to fit with the speech. For example, the peak of the hand-lifting motion is arranged to match with the word "let's go". Note that it takes a lot of time for our staff member to synchronize the gestures with the speech.

3.2 Synchronization Algorithm

Discovering the algorithm-friendly guideline in Section 3.1 let us view the synchronization problem as a motion synthesis problem, with both duration and timing issues satisfied automatically as the constraints of our algorithm. Based on the fact that many gestures are cyclic or have similar poses, we adopt the motion graph technique (Kovar et al., 2002; Arikan and Forsyth, 2002; Lee et al., 2002) in this paper upon considering the gesture's structure. Furthermore, we extend the original definition of motion graph by adding the meta data like edge weights and edge lengths, which is specially designed for our task. By using the motion graphs, a new motion can be generated efficiently using dynamic programming (Kovar et al., 2002; Xu et al., 2011), which is applicable to our synchronization task. Then, we refine the animation in an additional step, resulting in a two-step synchronization algorithm.

3.2.1 Coarse Synchronization with Motion Graphs

Motion Graph Construction. Considering the structure of a gesture, we construct a graph structure in and only in the stroke (S) phase for each gesture in the database, as shown in Fig. 1. As many conversational agent systems do (Neff et al., 2008; Marsella et al., 2013), the labeling data for gesture structures are created manually, which is acceptable because the number of gestures is not excessive. Similar to (Kovar et al., 2002; Arikan and Forsyth, 2002; Lee et al., 2002), the motion graphs consist of nodes and edges. In addition, edge weights and edge lengths as defined in this paper are designed to measure the smoothness and duration of motion. All the key-frames $V = \{t^1, t^2, \cdots, t^N\}$ in the S phase are selected as nodes. Two neighboring key-frames t^i and t^{i+1} are connected by a uni-directional edge $e^{i,i+1}$, whose direction is the temporal direction. The edge weight $w^{i,i+1}$ of a uni-directional edge is zero and its edge length $L^{i,i+1}$ is the number of frames between the two nodes t^i and t^{i+1}. Two similar key-frames t^i and t^j are connected by a bi-directional edge $e^{i,j}$, where the similarity or frame distance $d(t^i, t^j)$ is calculated as the weighted difference of joint orientations (Wang and Bodenheimer, 2003) as shown in Eq. (1).

$$d(t^i, t^j) = \sum_{m=1}^{M} w(m) \parallel \log(\mathbf{q}_{j,m}^{-1} \mathbf{q}_{i,m}) \parallel^2 \qquad (1)$$

where M is the joint number, $w(m)$ denotes the joint weight, and $\mathbf{q}_{i,m}$ is the orientation of joint m in the i-th key-frame. The edge weight $w^{i,j}$ of a bi-directional edge is the above frame distance and its edge length $L^{i,j}$ is zero. For smooth transitions, motion blending is performed by the SLERP technique (Shoemake, 1985) for each bi-directional edge. Note that the above construction process can be performed off-line.

Search the Best Path. Given a sentence of script with timing tags and its speech, our system will show a list of gesture candidates that match the category of the text when the creator clicks it in the timeline. For example, "good morning" is a word in the greeting category, where gestures like bowing, hand waving, and light nodding are listed. This rule-based technique is popular for behavior selection in conversational agent systems (Cassell et al., 2001; Marsella et al., 2013). Then the creator will select the best choice interactively. However, the original gesture motion in the database cannot always be a good match for the speech. In this section, our task is to generate a new motion for the given speech.

As (Kovar et al., 2002) pointed out, any path in the motion graph is a new motion, and we search for the best one that best satisfies all the following conditions: (1) as smooth as possible, (2) the one with a length nearest to the desired duration L_{tg}, (3) good connections with the P and R phases. Dynamic programming provides an efficient algorithm for motion graphs to search for the best path (Xu et al., 2011). Basically, edge weight is used in the cost function for Condition (1), i.e. $cost(e^{i,j}) = w^{i,j}$ where $w^{i,j}$ denotes the edge weight from the i-th key-frame to the j-th key-frame, which may be a uni-directional edge or a bi-directional edge. For Condition (2), we check the cumulative length for the desired duration L_{tg} as Eq. (4) shows. For Condition (3), we set the initial node as

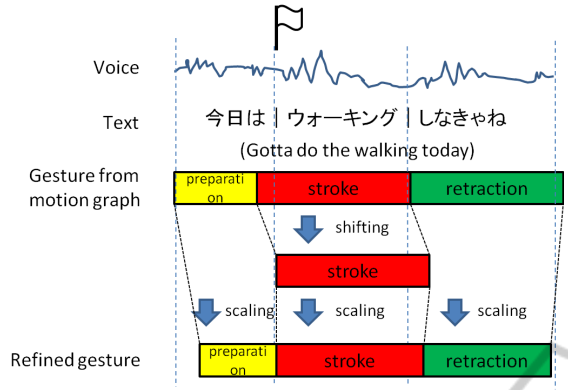

Figure 2: Refine the synchronization by shifting and scaling the gesture motion, where the black flag shows the desired timing in the speech.

the first key-frame t^1 in the S phase as Eq. (2) shows, which makes a natural connection to the P phase. In addition, we require the last node in the path is the last key-frame t^N in the S phase for good connection to the R phase. Finally, we select the best path with minimal cost that can satisfy all the three conditions. This makes the best path a new stroke phase with high quality and the desired duration.

$$P(t^v,1) = \begin{cases} 0 & if \quad t^v = t^1 \\ \infty & others \end{cases} \tag{2}$$

$$P(t^v,k) = \min_{t^i \in V}\{P(t^i,k-1)+cost(e^{i,v}))\} \tag{3}$$

$$P^* = \min_{L(P(t^N,k))\geq L_{tg}}\{P(t^N,k)\} \tag{4}$$

where $P(t^v,k)$ denotes the cost of the best path with k nodes and the last node of t^v, P^* denotes the best path for the speech, $L(P(t^N,k))$ denotes the cumulative length of the best path $P(t^N,k)$.

3.2.2 Fine Synchronization with Shifting and Scaling

In this section, we further improve the accuracy of synchronization with a shifting and scaling operation. As shown in Fig. 2, first the stroke (S) phase is shifted to the desired timing in the speech. Then, the S phase is scaled to match the desired duration as nearly as possible. In order to keep the motion natural and evocative of the desired emotional state, the scaling factor is limited to the range from 0.9 to 1.1. The same scaling factor will be used in the preparation (P) and retraction (R) phases to keep the motion consistency.

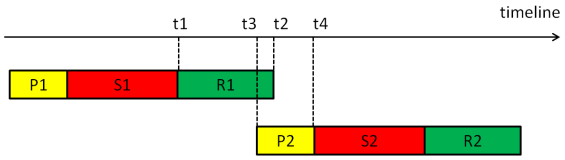

Figure 3: Total scheduling of gestures in the timeline.

Table 1: Total scheduling to avoid conflicts.

Conditions	Operations
(1) No conflict	No change
(2) With slight conflict	Scale P2 & R1 phases
(2) With serious conflict	Remove P2 & R1 phases

3.2.3 Total Scheduling

In most cases, our two-step algorithm will handle the duration and timing issues well. However, due to the preparation and retraction phases, neighboring gestures may conflict with each other as shown in Fig. 3. We define a rule to avoid such conflicts as shown in Table 1, where the conditions are (1) $t2 \leq t3$ (which means there is no conflict between R1 and P2), (2) $t2 > t3 \& \frac{(t2-t1)+(t4-t3)}{t4-t1} \leq TH$ (which means there is a conflict between R1 and P2 but the conflict is not serious), and (3) $t2 > t3 \& \frac{(t2-t1)+(t4-t3)}{t4-t1} > TH$ (which means there is a conflict between R1 and P2 and the conflict is too serious to use the scaling operation), respectively. Note that $t1$, $t2$, $t3$, and $t4$ are marked in Fig. 3. TH is a threshold.

4 OVERVIEW OF AUTHORING TOOL

In this section, we will briefly introduce our system, an authoring tool whose interface is shown in Fig. 4. The purpose of the authoring tool is to provide a good balance between the creator's flexibility and his/her efficiency. For example, with our authoring tool, a staff member can create attractive animation for lectures even if he/she has little knowledge of animation creation. To re-use a large amount of CGM resources in characters and motions on the Internet, we use the MikuMikuDance format (http://www.geocities.jp/higuchuu4/index_e.htm) in our system, which is very successful in Japan and East Asia[1]. Using some free software, the data in

[1]In this paper, the so called MikuMikuDance has three meanings. First, it may mean the authoring tool used in Sect. 3.1. Second, it may mean the specifications for mesh model, motion, and other data in animation. Third, it may mean a player to show the animation.

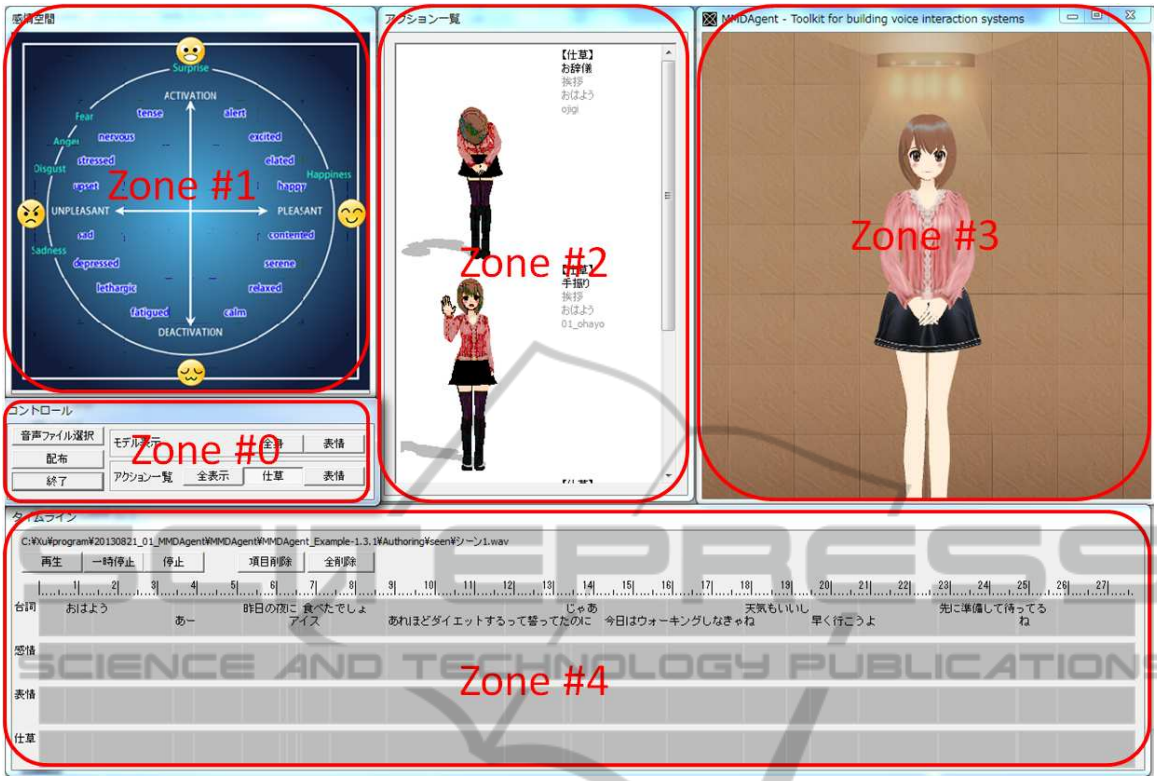

Figure 4: User interface of our authoring tool.

MikuMikuDance format can be transferred from/to other formats like Maya and Blender.

There are five zones in our authoring tool. The top left window (Zone #1) in Fig. 4 is an affective space panel based on the circumplex model (Russell, 1980), where the creator can select a point and see the resulting facial expression in the top right window (preview window, Zone #3) in real time. The middle left window (Zone #0) is the control panel, where the creator can input the audio file and switch the display modes for the preview window and the candidate list window (Zone #2, top middle window in Fig. 4). The bottom window (Zone #4) shows the timeline for text, affect, facial expression, and gesture, respectively, which can be saved as a project file or loaded from a project file.

The creation procedure of a new project is as follows. First, an audio file and its text file with timing tags are loaded. The text will be displayed in the timeline. When a creator clicks a sentence in the timeline, its audio will be played and some candidate gestures will be listed in Zone #2. The creator can select the affect in Zone #1 and the best gesture in Zone #2 to fit the audio and text. The system will automatically synthesize a facial expression and gesture motion synchronized with the speech in real time, which will be displayed in Zone #3. After repeating this step for all desired sentences, the creator can watch the entire animation by clicking the “play” button in Zone #4. For those parts without any arrangement from the creator, the system will automatically deal with it. In addition, blinking and lip synchronization are automatically embedded in the animation. Finally, the creator can edit the animation at any time and release it. Note that facial expression is generated based on the Facial Action Coding System (FACS) technique (Ekman et al., 2002), where totally 18 action units are defined in the case of our system.

5 EXPERIMENTAL RESULTS AND DISCUSSIONS

In our experience, it takes much less time to create animations with our authoring tool than by manual creation. Although the creation time is dependent on the experience of creator (which infers that the detailed time data have little meaning), it only takes about 10% of time for our staff member to create the content used in our experiments by our authoring tool. Moreover, in our opinion, their quality is almost the same. However, evaluation of the proposed method is challenging. Objective evaluation of synchronization is difficult if not impossible because it is difficult to define

Table 2: Scripts of Twitter comments on a news item about a company's new software, which are used in subjective evaluation.

Comment	Script (Translated from Japanese)
#1	No plan to release it, although they have a Japanese version!
#2	No plan to release a Japanese version! (The company) cannot seem to get motivated.

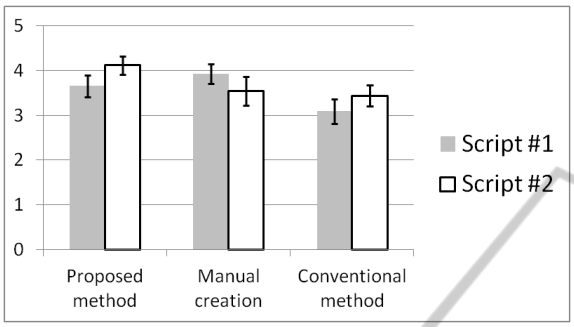

Figure 5: Mean opinion scores for Q1 in diet scenario. The standard errors are shown in black lines (±1 SE).

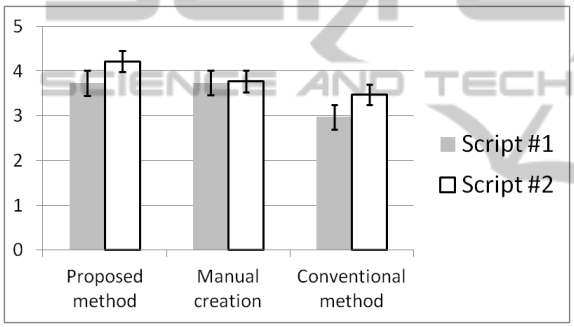

Figure 6: Mean opinion scores for Q2 in diet scenario. The standard errors are shown in black lines (±1 SE).

the ground truth of synchronization. Currently, subjective evaluation is the only option.

Because our target user is the general consumer, the participants are non-expert volunteers. In our experiments, eleven participants are asked to evaluate two kinds of content including the diet scenario (see details in Section 3.1) and a news comment scenario (see details in Table 2). The animation of the diet scenario lasts about 30 seconds, the news comment scenario about 5 seconds. The 11 participants, 5 male and 6 female, range in age from their 30s to their 50s, most with little experience or knowledge of animation creation.

For the diet scenario, three animations are shown that come (in random order) from manual creation, the proposed method, and the conventional method where the only difference from the proposed method is that the synchronization algorithm comes from (Marsella et al., 2013). Two questions (Q1: How good is the animation quality? Q2: How good is the synchronization of gesture and speech?) are evaluated using the following rating scheme. 5: Excellent; 4: Good; 3: Fair; 2: Poor; 1: Bad. The mean opinion scores (MOS) for Q1 and Q2 are listed in Fig. 5 and Fig. 6 respectively, where our method performs better in all cases than the conventional method, and in most cases is better than manual creation. Especially in Script #2, our method performs much better than manual creation, receiving a MOS of more than 4 (Good) in both Q1 and Q2. This is because manual creation simply repeats the motion cycles while the best path in the proposed method provides more variation. In the animation from the conventional method, we observe that the gesture's speed is changed so much that it becomes unnatural, and synchronization is not clear due to an unsharp phase boundary, which explains why the conventional method performs worst for both Q1 and Q2. T-test results in Table 3 consistently confirm that a significant difference exists in both Q1 and Q2 for Script #2 between the proposed method and the conventional method at the 5% significance level and p values for Script #1 between the proposed method and the conventional method are rather low.

Table 3: p values in T-test for diet scenario. PM: Proposed method vs. Manual creation. PC: Proposed method vs. Conventional method. Red fonts: $p < 0.05$. Blue fonts: $p < 0.1$.

		PM	PC
Script #1	Q1	0.4131	0.1443
	Q2	0.9817	0.0690
Script #2	Q1	0.1469	0.0422
	Q2	0.2050	0.0356

Because the scores are rather different from different participants, we analyze the rank rating of three methods for the same content. The tally of assigned No. 1 ranks for Q1 and Q2 are shown in Fig. 7 and Fig. 8. As you can see, 9 of 11 participants rank our method No. 1 for Script #2 for Q1 and Q2, which is a much better evaluation than other methods. Note that because more than one method may get No. 1 rank rating, the total number of No. 1 ranking is a little more than 11 as shown in Fig. 7 and Fig. 8.

For the news comment scenario, we ask the participants to choose the better animation: the one produced by our proposed method or one by the conventional method (presented in random order) in terms of the two criteria above. For Comments #1 and #2, 8 of 11 participants select the animation produced using our proposed method. Five participants select our

Table 4: Statistics for different gesture styles.

gesture style	no gesture	regular gestures	highlighted gestures
average number of gestures	0.0	9.7	15.3
gesture example	-	point a finger up	beat on the slide

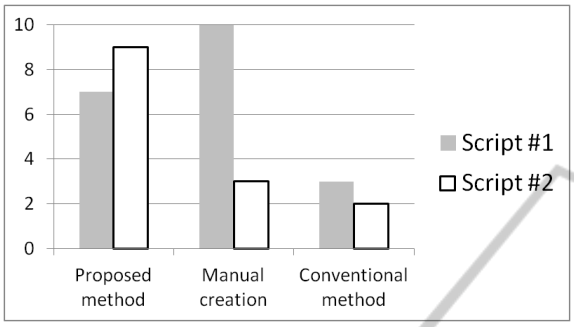

Figure 7: Rank No.1 ratings for Q1 in diet scenario.

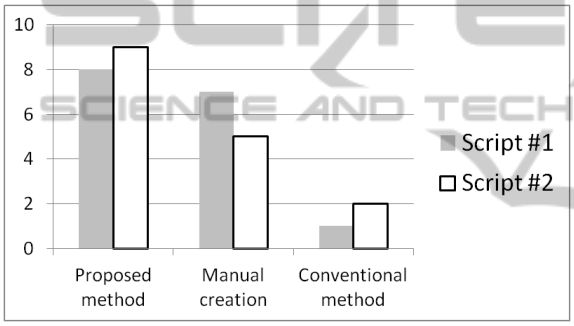

Figure 8: Rank No.1 ratings for Q2 in diet scenario.

proposed method in both comments, while none selects the conventional method in both comments. This suggests that the proposed method is very promising.

6 APPLICATIONS TO EDUCATION

Using our authoring tool, our staff creates several animations for digital education, where the ECA acts as a lecturer as shown in Fig. 9. To avoid the possible bias from different levels of students' knowledge, short animations in four different categories are created and evaluated by each student including information technology, history, chemistry, and geography. In each content, there is a short slide show with an ECA that lasts about one minute and a test with three questions including a question about figures. By selecting different facial expressions and gestures, three different styles of animations are evaluated, which include no expression, moderate expression, and intensive expression or no gesture, regular gestures, and highlighted gestures. As Table 4 shows, more gestures and

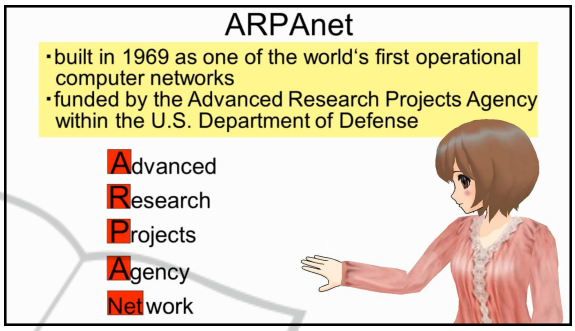

Figure 9: Screen capture of an education animation. Note that the text in the slide is translated from Japanese. Only Japanese versions are used in the evaluation.

stronger gestures are used in the style of highlighted gestures to emphasize the key points in the education content.

Basically, we want to see the effectiveness of agent by the scores (percentage of correct answers) in the test after watching the animations. Totally 34 participants from a university conduct the evaluation on the animations in four categories, whose styles of gesture and expression are randomly selected. The scores for facial expressions and gestures are shown in Table 5 and Table 6 respectively. The results show that both the highlighted gestures and intensive expressions are effective to obtain better scores. Especially, a high score of 72.7% is obtained in the questions about figures for the highlighted gestures. A T-test is performed between no gesture and highlighted gestures, which gives a significant level of 6.96%. A similar T-test between no expression and intensive expressions gives a significant level of 7.74%. Both are rather near to 5%, which is commonly used as a significant difference. In addition, many participants report that the intensive expressions and highlighted gestures are impressive in the questionnaire.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we have described an authoring tool we have implemented to facilitate the creation of high quality animation for conversational agents, with facial expressions and gestures that are accurately synchronized with speech. In a precedent case study, we

Table 5: Average scores (percentage of correct answers) for different gesture styles.

gesture style	no gesture	regular gestures	highlighted gestures
average scores for all questions	56.7 %	60.7 %	66.7 %
average scores for questions about figures	57.4 %	60.0 %	72.7 %

Table 6: Average scores (percentage of correct answers) for different expression styles.

expression style	no expression	moderate expressions	intensive expressions
average scores for all questions	56.3 %	60.0 %	67.4 %
average scores for questions about figures	55.6 %	66.7 %	67.4 %

have investigated the important issues (i.e. duration and timing) in the manual synchronizing of gesture with speech, which has led us to consider the synchronization problem to be a motion synthesis problem. We have proposed a novel two-step solution using the motion graph technique within the constraints of gesture structure. Subjective evaluation of two scenarios involving talking and news commentary has demonstrated that our method is more effective than the conventional method.

In the future, we plan to improve the generation of facial expressions, where realistic facial dynamics can further improve animation quality. At the same time, we are extending the target applications to new categories such as remote chat and human-robot interaction.

REFERENCES

Arikan, O. and Forsyth, D. (2002). Interactive motion generation from examples. *ACM Transactions on Graphics*, 21(3):483–490.

Beskow, J., Engwall, O., Granstrom, B., and Wik, P. (2004). Design strategies for a virtual language tutor. In *INTERSPEECH-2004*, pages 1693–1696.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., and Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '99, pages 520–527.

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '01, pages 477–486.

Dutoit, T. (2001). *An Introduction to Text-to-Speech Synthesis*. Springer.

Ekman, P., Friesen, W. V., and Hager, J. C. (2002). *Facial Action Coding System: The Manual on CD ROM*. A Human Face, Salt Lake City.

Huang, J. and Pelachaud, C. (2012). Expressive body animation pipeline for virtual agent. In *Intelligent Virtual Agents*, volume 7502 of *Lecture Notes in Computer Science*, pages 355–362.

Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thórisson, K., and Vilhjlmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*, pages 205–217. Springer Berlin Heidelberg.

Kovar, L., Gleicher, M., and Pighin, F. (2002). Motion graphs. *ACM Transactions on Graphics*, 21(3):473–482.

Lee, J., Chai, J., Reitsma, P., Hodgins, J., and Pollard, N. (2002). Interactive control of avatars animated with human motion data. *ACM Transactions on Graphics*, 21(3):491–500.

Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., and Shapiro, A. (2013). Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '13, pages 25–35.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746 – 748.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, 92(3):350–371.

McNeill, D. (2005). *Gesture and Thought*. University of Chicago Press.

Miller, L. M. and D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience*, 25(25):5884–5893.

Neff, M., Kipp, M., Albrecht, I., and Seidel, H.-P. (2008). Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27(1):5:1–5:24.

Ng-Thow-Hing, V., Luo, P., and Okita, S. (2010). Synchronized gesture and speech production for humanoid robots. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 4617–4624.

Niewiadomski, R., Bevacqua, E., Mancini, M., and Pelachaud, C. (2009). Greta: an interactive expressive eca system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '09, pages 1399–1400.

Nishida, T. (2007). *Conversational Informatics: An Engineering Approach*. John Wiley & Sons, Ltd.

Noma, T., Zhao, L., and Badler, N. (2000). Design of a

virtual human presenter. *Computer Graphics and Applications, IEEE*, 20(4):79–85.

Oura, K., Yamamoto, D., Takumi, I., Lee, A., and Tokuda, K. (2013). On-campus, user-participatable, and voice-interactive digital signage. *Journal of The Japanese Society for Artificial Intelligence*, 28(1):60–67.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Shoemake, K. (1985). Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '85, pages 245–254.

Stone, M., DeCarlo, D., Oh, I., Rodriguez, C., Stere, A., Lees, A., and Bregler, C. (2004). Speaking with hands: creating animated conversational characters from recordings of human performance. *ACM Transactions on Graphics*, 23(3):506–513.

van Luin, J., op den Akker, R., and Nijholt, A. (2001). A dialogue agent for navigation support in virtual reality. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '01, pages 117–118, New York, NY, USA. ACM.

Čerekovič, A. and Pandžič, I. (2011). Multimodal behavior realization for embodied conversational agents. *Multimedia Tools and Applications*, 54(1):143–164.

Wang, J. and Bodenheimer, B. (2003). An evaluation of a cost metric for selecting transitions between motion segments. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, SCA '03, pages 232–238, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association.

Xu, J., Takagi, K., and Sakazawa, S. (2011). Motion synthesis for synchronizing with streaming music by segment-based search on metadata motion graphs. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6.