Presentation Training System
Based On Imitating Past Famous Speech

Author : SHI Yuhua

Student Number : 6612160065-0

Supervisors :
Prof. Haruo Noma
Assoc.Prof. Roberto Lopez-Gulliver
Lecturer Kohei Matsumura

Ritsumeikan University

Graduate School of Information Science and Engineering

Abstract

Conference presentation is a non-trivial task since it is affected by both presentation content and nonverbal behaviors. To improve presentation ability, we propose a training method based on imitating the famous past speech. Trainees imitate past famous speech by watching recorded videos, then the proposed system matches they trainees' and the orator's motion data and gives real-time visual feedback on how well they imitate the orators' speech. We employed the OpenPose Library to extract orators' motion data from past famous speech 2D video. During training, the system captures trainee's motion in real-time using the Kinect, then calculate the cosine similarity of adjacent limbs to get the similarity of the trainees' and orator's motion data.

In this paper, we evaluate the effectiveness of our motion data matching algorithm and as well as the system effectiveness. We conducted two experiments to verify the effectiveness of our algorithm and our proposed system. The first experiment results show that the proposed algorithm is able to match the trainee's motion and the orator's motion in the past famous speech. In our second experiment, we made a A/B test to evaluate the effectiveness of the proposed system. According to the results, we can find that the trainee perform better after training using our proposed system. Notably, the trainees make more gestures and more suitable pauses during the presentation.

CONTENTS

List of Figures

List of Tables

1. Introduction

The presentation is the art of persuasion, It plays a significant role in our society and has the tremendous impact on the success of everyone [1]. The presentation commonly used to communicate the presenter's ideas to the listener. However, giving a presentation successfully like John F. Kennedy or Steven Jobs is not a simple thing. A great preparation not only contains verbal style but also needs various nonverbal behaviors. On the one hand, the content of preparation must be clear, vivid and appropriate[2]. On the other hand, the significant component of a presentation lies upon nonverbal cues which have the power to change the meaning assigned to spoken words[1].

Nonverbal behaviors of public speakers are expressed via several channels such as voice, gesture and facial expression. They have been proven to have a more significant influence than verbal cues. According to Argyle nonverbal messages are thirteen to fourteen times more powerful than verbal ones[3]. Likewise, Arcy showed that the audience receives more than half of information from nonverbal behaviors[4]. The study of Seiler[1]shows the fact that most people unconsciously believe more in nonverbal behaviors than verbal cues.

Unfortunately, it's hard to practice to express the effective nonverbal behaviors because they are mostly expressed subconsciously. To achieve the learning results, trainees must be provided with appropriate feedbacks from human advisors, but it is costly and not always available. According to Seiler [1], imitation is a kind of effective method to help the trainee to refine their nonverbal behaviors. Imitating those past famous speakers' gesture, sound or enunciation to learn what they're doing right, and then try to make it your own as Pablo Picasso said *" Good Artists Copy. Great Artists Steal "*. It is a great learning experience and can stretch trainees' abilities.

In parallel, the role of nonverbal behaviors in computing is becoming increasingly recognized by the development of the emerging fields, such as social signal processing and affective computing. Such as [5–9], those Deep Learning library can extract user's body key point with high accuracy. Therefore, computers have been equipped with the abilities to decode the complexity of humans nonverbal channels.

Many papers discussed some approaches toward the automatic recognition of nonverbal from trainees. However these approaches all defined some exact rules in advance to give a score of

the presentation, but few have focused on imitating past famous speech to refine their nonverbal behaviors.

In this paper, we propose a presentation training system that allows the trainee to imitate past famous speech to improve their nonverbal behaviors. Nonverbal behaviors include many aspects, and we choose to analyze the gesture of orators as the nonverbal behavior in this paper. In advance, we employed OpenPose library[6] to extract orators' motion data from past famous speech 2D video. While training, the system capture the trainee's motion in real-time using Microsoft Kinect. We choose the cosine similarity of adjacent limbs as the feature to calculate the score that shows the similarity of the trainees' motion and the motion of extracted orators. The trainees will also wear an HMD that show a virtual hall and some virtual audience. The audience will perform some actions according to the score.

Figure 1.1 : System introduction

The rest of this paper is organized as follows. First, we will introduce some existing presentation training method and proposed solution in chapter 2. Then we will introduce the preparation of OpenPose, Kinect and the Evaluation of presentation in the chapter3. Chapter 4 describes our proposed system in detail. In chapter 5, we will evaluate the system and discuss the results. The last chapter is for conclusions and future works.

2. Related Work

Presentation skills give us the power to change the world. Great presenters instill trust, engage our minds and hearts, deliver ideas and information, and inspire and captivate us.

In this paper, we propose a presentation training system that allows trainees to imitate past famous speech to improve their nonverbal behaviors. Many papers discussed some approaches toward the automatic recognition of nonverbal from trainees. However these approaches all defined some exact rules in advance to give a score of the presentation, but few have focused on imitating past famous speech to refine their nonverbal behaviors. Some researches analyze some visual and vocal channels of trainees, thus can provide information about their presentations. For example, the system in Pfister was originated from a vocal emotion detection module [10]. It was similar to the approach of Silverstein [11], which was built solely on vocal cues, by analyzing the voice such as pitch or tempo. The Hincks's system measured the changes in vocal pitch, and then give visual feedback to promote pitch variation by relying on the importance of pitch variance in oral presentations [12].

On the other hand, some include nonverbal behaviors in the analysis. Gao introduced the method based only on visual information [13]. In contrast, Kurihara added face position and orientation as the approximation of eye contact, together with pitch, speaking rate, utterance and filled pauses [14]. Hoque proposed an automated conversation coach to help trainees improve their interview skills by recognizing their motion and facial expression [14].

We will introduce the detail about Kurihara's research [14] in section 2.1 and Nguyen's research [15] in section 2.2.

2.1 Presentation Sensei

In Kurihara's paper, they present a presentation training system that observes a presentation rehearsal and provides recommendations for improving the delivery of the presentation, such as to speak more slowly and to keep eye contact with the audience (Figure 2.1). Kurihara intentionally focuses on the basic behavior patterns because high-level semantics is very difficult to analyze by computers in contrast to basic behavior patterns. Kurihara's goal is to help people improve their base-line presentation skills by reducing inappropriate behavior such as using many fillers

(e.g., “ er ”) or continuously looking down at the script.

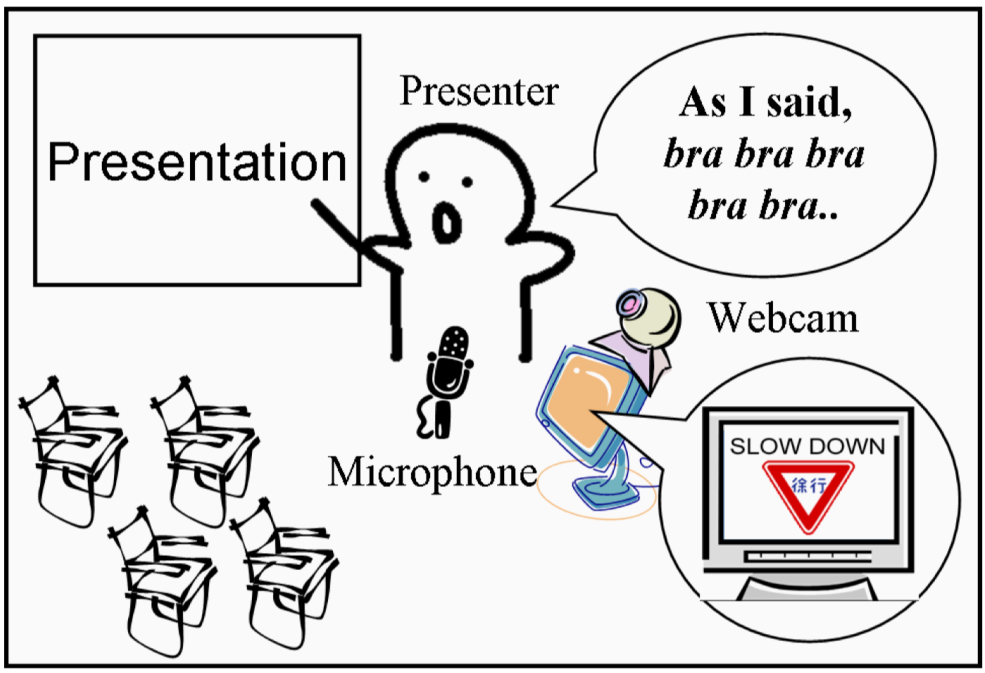

Figure 2.1：Presentation sensei system [14]

In this work, Kurihara focus on the following five aspects of presentation delivery.

- The speaking rate should not be too fast.
- The speech should not be monotonous.
- The speech should not contain too many fillers.
- The speaker should look at the audience and avoid continuously looking down at a script or a screen.
- The speaker should finish the presentation within a certain time limit.

Kurihara selected these because they are emphasized in the existing literature, and they can be detected to some extent using current speech processing and image processing technologies. The Presentation Sensei system visualizes the analysis result in real time communicating with a presentation tool. It can give the presenter both instant “ online ” feedback and post “ offline ” feedback for improvements. The online feedback function shows the analysis result in real-time. When the system detects some inappropriate behavior, it alerts the presenter by showing a visual

signal (Figure 2.2). The offline feedback function shows the visual summaries of the indices for the presenter ' s self-examinations (Figure 2.3).

Figure 2.2：Online feedback. (Left) Real time monitor. (Trafic signals) Visual Alerts.[14]

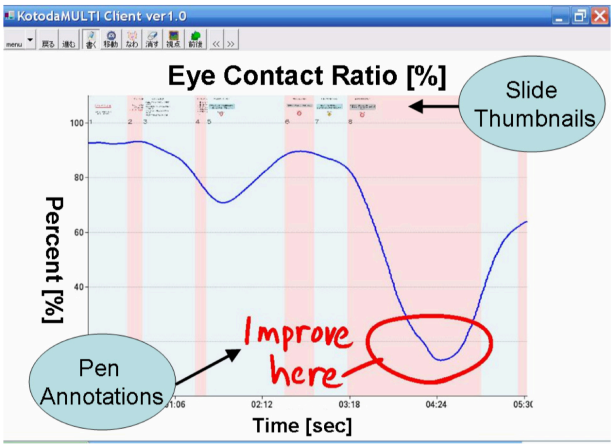

Figure 2.3：Offline feedback. An example of the generated charts[14].

It is often difficult to make speech processing and image processing work robustly in adverse environments. One advantage of Kurihara's target application domain, personal presentation rehearsal, is that we can relatively freely configure the environment. It is realistic to rehearse in a silent room with no visual obstacles, where these technologies perform the best. This feature makes the Presentation Sensei system a highly practical application even with imperfect recognition technologies [14].

Kurihara's system is unique in that, while general multimodal systems help the user to control computers, it tries to help computers guide humans. This way of using a computer is relatively new. Heer et al. [16] investigated the design guidelines for this sort of systems and also introduced an experimental media capture system that acts as a film director.

System Configuration

The Presentation Sensei system consists of several modules connected by a network (Figure 2.4). The audio analysis module continuously analyses the input signal from a microphone and provides the integration module with the results of the utterance duration detection, pitch (F0) detection, and filled pause detection. The speech recognition module also continuously provides the integration module with the mora-based speech recognition results. The image processing module continuously analyzes the input from a webcam and provides the integration module with the result of the face position/orientation detection. The integration module integrates all the provided information and gives feedback to the presenter using various monitors. These modules can be distributed over a local network for the load sharing purpose, and they communicate with each other via RVCP protocol [17]. The system can also be connected to a third party presentation tool to achieve synchronization. Kurihara currently connects their system to an in-house presentation program to receive timing information and thumbnail images of slides.

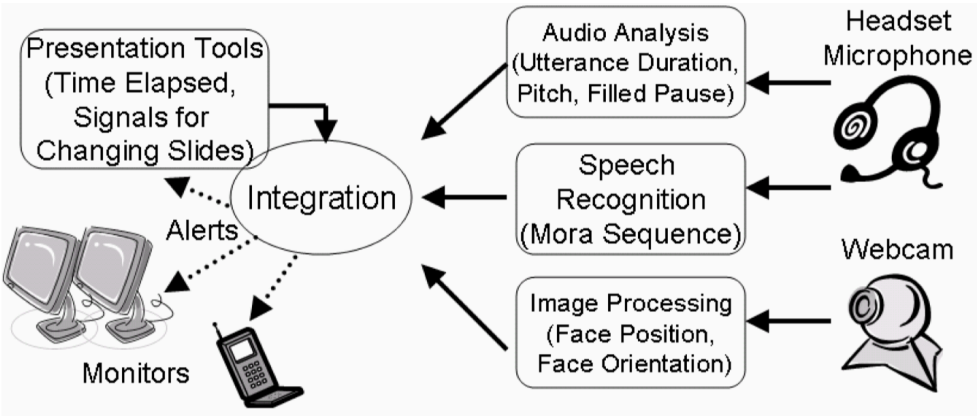

Figure 2.4 : System configuration of presentation sensei [14]

2.2 Intelligent Presentation Skills Trainer

Nguyen's research developed a tutoring system for public speaking, which assesses presentations based solely on the visual behaviors of presenters. Firstly, an empirical study was performed to investigate on the nonverbal cues that impact a presentation, serving as the ground truth. Next, a Microsoft Kinect was implemented for capturing skeletal representations of the presenters' body as input data for the analysis. The recognition process can detect if the behaviors appeared

in real-time. Multi-class support vector machine was used to classify the quality of presentations into a four-degree scale with the recognition rate of 73.9% on a training/test database that includes 76 presentations. For the feedback, the system allows presenters to review their presentation, together with the analysis results. They also developed a simulated conference room as the real-time feedback mechanism.

Automatic Feedback System

In order to support presenters with an effective solution that can help them self-practice even at home, Nguyen aimed to implement the system with the following functions: (1)Automatic analyzes presenters performance; (2)Provides immediate feedback during the presentation; (3)Provides overall analysis about the whole presentation; (4)Lets users review their performance together with the analyzed results, thus allows them to keep track of their practicing progress. To achieve these purposes, they set up a Microsoft Kinect to extract body's skeletal representation as input for the analysis task. In parallel, a regular camera or webcam is positioned to simulate the audience's point of view. The automatic analysis, as well as recording, is processed in real-time using a regular PC. The result is visualized on the PC or an external screen/projector (Figure 2.5). Users also have the chance to review their presentations, together with in-depth analysis of nonverbal cues in the end.

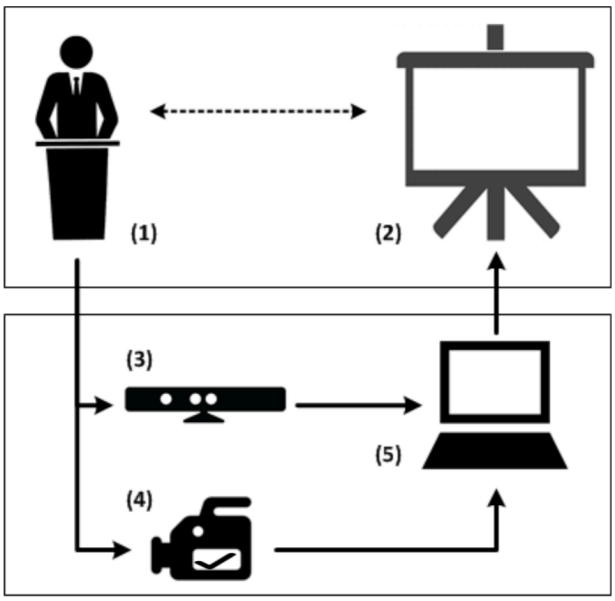

Figure 2.5 : Setup of the Nguyen's system [15]

The Two Methods of Giving Feedbacks

Nguyen's system provides two ways of delivering feedback to the audience [15]. The first one shows users their recorded presentation, the appearances of each behavior and results on the four nonverbal aspects, plus the overall result. In parallel, with the purpose to give presenters the helpful feedbacks, also aim to provide them the experience as presenting for the real audience, Nguyen developed a virtual conference room as one method to deliver feedbacks. The environment was built using the Unity3D engine, simulates the classroom that we collected data for observation. Avatars can perform several animations that may bring either a positive or negative feeling for presenters. These animation clips are sorted based on the increase of negative feeling: (1) Nodding; (2) Sitting still; (3) Sleeping; (4) Yawning.

Figure 2.6 : The simulated conference room [15]

3. Preparation Work for Proposed System

In this chapter, we will introduce some preparation work for the proposed system. At first, we employed the OpenPose library [6] to extract the orator's joint data from past speech 2D video, and we will introduce the OpenPose library in section 1. Then we set up a Microsoft Kinect for Windows Version device [18] to extract the trainee's joint data in training, and we will introduce the Kinect camera in section 2. To evaluate the effectiveness of the proposed system, we need to know how to evaluate a presentation, and we will introduce some evaluation points of a presentation.

3.1 Joint Data Extraction Method (2D)

We employed the OpenPose library to extract the orator's joint data from 2D speech video [6]. OpenPose is a library for real-time multi-person keypoint (Figure 3.1) detection and multi-threading written in C++ using OpenCV and Caffe [19]. OpenPose can detect the human body, hand and facial key points on single images. Also, the system computational performance on body keypoint estimation is invariant to the number of detected people in the image.

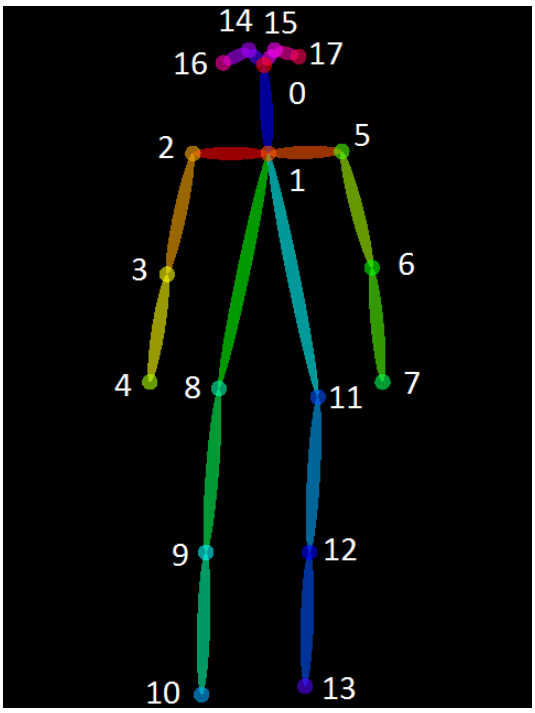

Figure 3.1 : Keypoints detected by OpenPose [6]

Convolutional Pose Machine (CPM)

Convolutional Pose Machine (CPM) use Convolutional Neural Networks (CNNs) to detect the human joint data from 2D image or video. CPMs consist of a sequence of convolutional networks that repeatedly produce 2D belief maps for the location of each part (Figure 3.2). At each stage, image features and belief maps produced by the previous stage are used as input [20].

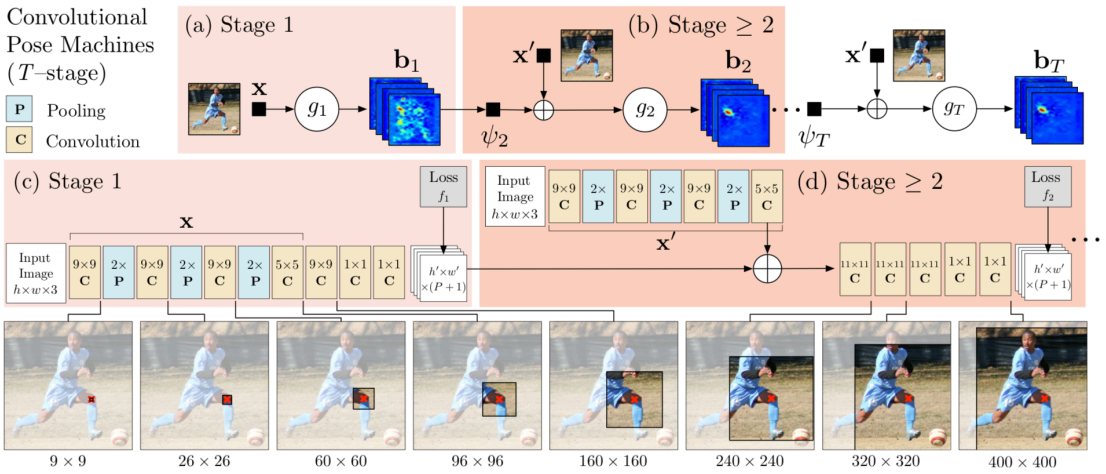

Figure 3.2 : Architecture of Convolutional Pose Machines (CPMs) [20]

The belief maps provide the subsequent stage an expressive non-parametric encoding of the spatial uncertainty of location for each part, allowing the CPM to learn rich image-dependent spatial models of the relationships between parts. The overall proposed multi-stage architecture is fully differentiable and therefore can be trained in an end-to-end fashion using back propagation [20].

At a particular stage in the CPM, the spatial context of part beliefs provides strong disambiguating cues to a subsequent stage. As a result, each stage of a CPM produces belief maps with increasingly refined estimates for the locations of each part.

To capture long-range interactions between parts, the design of the network in each stage of our sequential prediction framework is motivated by the goal of achieving a large receptive field on both the image and the belief maps [20].

OpenPose

Based on CPM architecture, OpenPose is an efficient method for multi-person pose estimation what uses a non-parametric representation of association scores via Part Affinity Fields (PAFs), a

set of 2D vectors field that encodes the location and orientation of limbs over the image domain.

The part affinity is a 2D vector field for each limb for each pixel in the area belonging to a particular limb, which encodes the direction that points from one part of the limb to the other. Each type of limb has a corresponding affinity field jointing its two associated body parts. A greedy parsing algorithm is sufficient to produce high-quality parses of body poses, which maintains efficiency even as the number of people in the image increases [6].

Figure 3.3 illustrates the overall pipeline of OpenPose library.

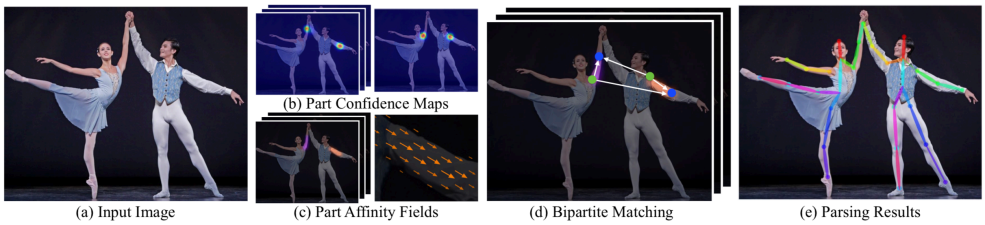

Figure 3.3 : Overall pipeline of OpenPose [6]

- OpenPose takes, as input, a color image of size of $w \times h$ (Figure 3.3a) and produces, as output, the 2D locations of anatomical key-points for each person in the image (Figure 3.3e) [6].

- First, a feed-forward network simultaneously predicts a set of 2D confidence maps of body part locations (Figure 3.3b) and a set of 2D vector fields of part affinities, which encode the degree of association between parts (Figure 3.3c).

- Finally, the confidence maps and the affinity fields are parsed by greedy inference (Figure 3.3d) to output the 2D key points for all people in the image.

The architecture of OpenPose, shown in Figure 3.4, simultaneously predicts detection confidence maps and affinity fields that encode part-to-part association. The network is split into two branches: the top branch, shown in beige, predicts the confidence maps, and the bottom branch, shown in blue, predicts the affinity fields. Each branch is an iterative prediction architecture, Following the typical structure of a CMP, which refines the predictions over successive stages, with intermediate supervision at each stage [6].

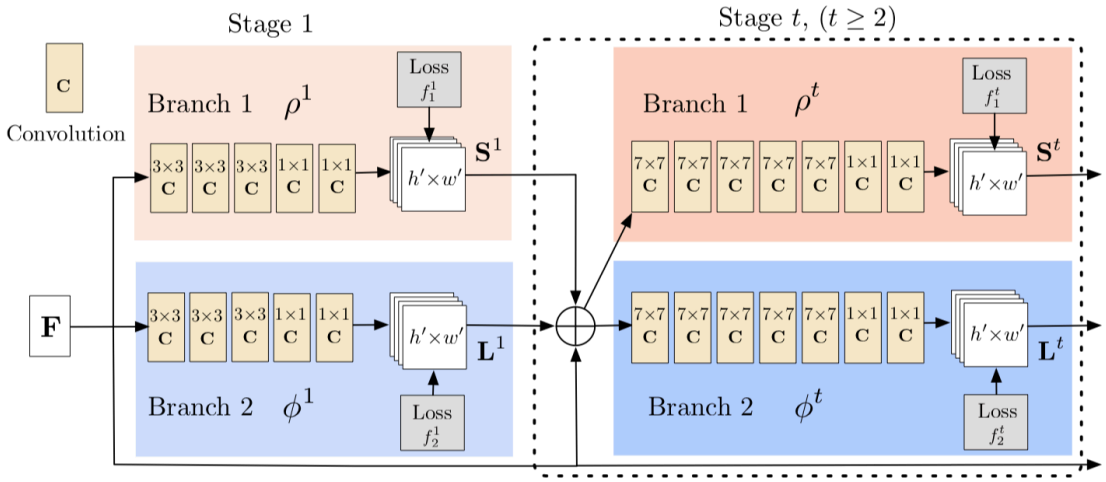

Figure 3.4 : Architecture of the two-branch multi-stage CNN in OpenPose [6]

In our proposed system, we employed OpenPose library to extract the orator's joint data from past famous 2D speech video (Figure 3.5). We use those joint data to do template matching to get the score.

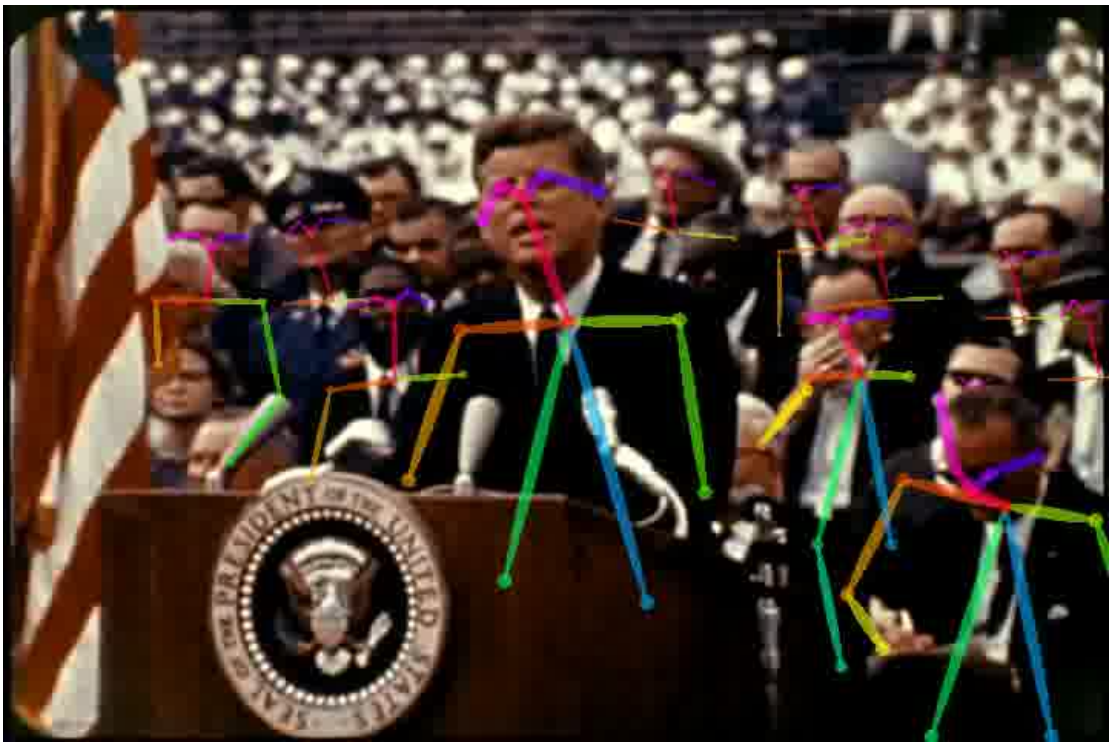

Figure 3.5 : Joint detected by OpenPose (Example)

3.2 Joint Data Extraction Method (3D)

Recent advances in 3D depth cameras such as Microsoft Kinect sensors have created many opportunities for multimedia computing. The Kinect sensor lets the computer directly sense the third dimension (depth) of the players and the environment. It also understands when users talk, knows who they are when they walk up to it and can interpret their movements and translate them into a format that developers can use to build new experiences [21].

Kinect V2 Sensor

The Kinect V2 sensor incorporates several advanced sensing hardware. Most notably, it contains a depth sensor, a color camera, and a microphone array that provide full-body 3D motion capture, facial recognition, and voice recognition capabilities. Figure 3.6 shows the arrangement of the infrared (IR) camera, the color camera, and the IR illuminator.

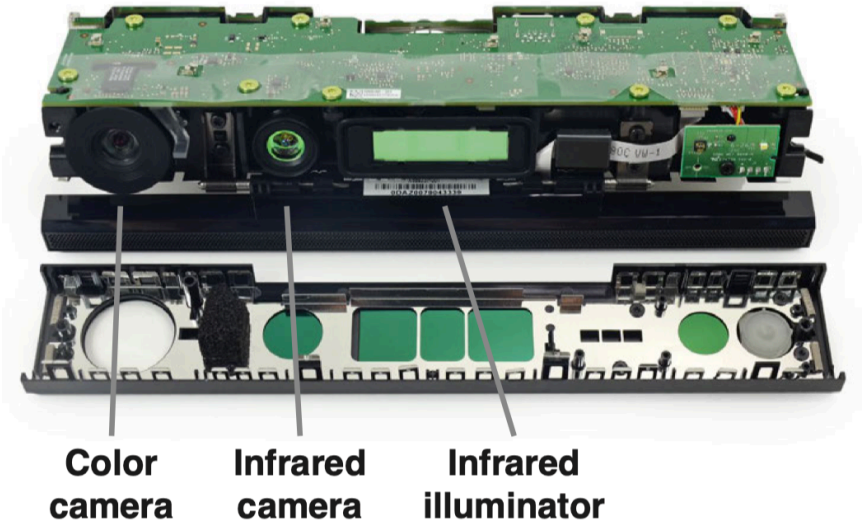

Figure 3.6 : Microsoft Kinect V2 sensor [22]

The Kinect V2 depth sensor is based on the time-of-flight measurement principle. An infrared strobe light (see Figure 3.6) illuminates the scene, the light is reflected by obstacles, and the time of flight for each pixel is registered by the infrared camera. Internally, wave modulation and phase detection are used to estimate the distance to obstacles (indirect ToF) citeFankhauser2015a. Details on the depth measurement method of the Kinect V2 are given in Sell's paper [23].

Kinect Skeletal Tracking

In skeletal tracking, a human body is represented by some joints representing body parts such as head, neck, shoulders, and arms (see Figure 3.7). Each joint is represented by its 3D coordinates.

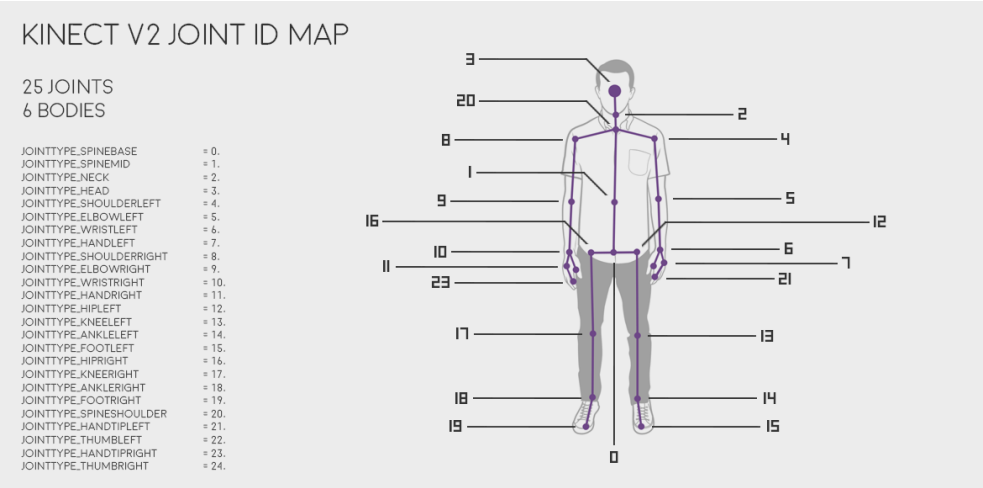

Figure 3.7 : Microsoft Kinect V2 joint id map [1]

Figure 3.8 illustrates the whole pipeline of Kinect skeletal tracking. The first step is to perform per-pixel, body-part classification. The second step is to hypothesize the body joints by finding a global centroid of probability mass through the mean shift. The final stage is to map hypothesized joints to the skeletal joints and fit a skeleton by considering both temporal continuity and prior knowledge from skeletal train data [21].

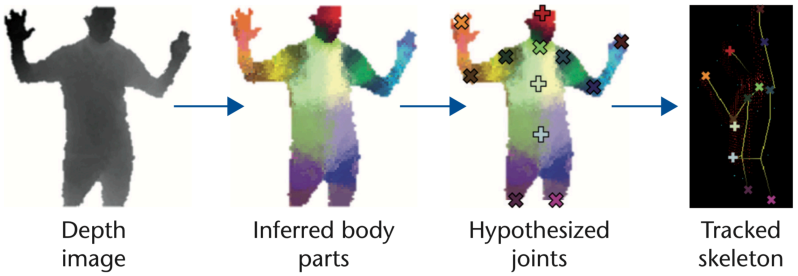

Figure 3.8 : The Kinect skeletal tracking pipeline [21]

In our proposed system, we set up a Kinect V2 to extract the joint data of trainee in real-time. We only employ the 2D coordinates to fit the joint data extracted by OpenPose from the past speech.

[1] https://vvvv.org/documentation/kinect

3.3 Evaluation of Presentation

To evaluate the presentation skill of trainees, we need to know what kind of behaviors will have the impact on the presentation. Nguyen's research performs an observation to analysis the behaviors of presenters [15]. They collected data from a training class about public speaking skills for postgraduate students. They ask the learners to give short presentations (about one minute) in front of the audience, which includes about ten other learners and one or two coaches. The presenters can freely choose the content of the presentations. In fact, all presenters chose to talk about their research, in the ways that it can be understood by all of the audience that might come from the different fields. After each presentation, the audience gave feedbacks and suggestions on how the presentation should be improved, regarding nonverbal expressions. They set up a regular camera to record the presentations. They also set up a Microsoft Kinect to capture the whole body movement for their further signal processing, as well as behavioral studies. They stored the data from Kinect as the *.ONI files using the OpenNI SDK. They removed the unsatisfied videos (e.g. presenters moved out of the camera range) and finally collected 39 presentations of 11 presenters (four females and seven males).

In their research, they use regular videos for behavioral analysis. This task was done through the collaboration with an expert in public speaking. The role of the expert was to review the recorded videos and then specifying the nonverbal cues that affected the performance of the speakers, together with the duration that they appeared. Thus, for each video, a set of behaviors was created. They collected the nonverbal cues and then annotated their appearance using the commercial software Noldus Observer XT [24]. Behaviors were categorized into either *State event* if their duration is necessary to be studied, or *Point event* otherwise. The software provided them with the statistical analysis on the appearance of these behaviors, including the number of presentations that contain the behaviors, the rate that they appeared (point events) and the percentage of time that they accounted for (Table 3.1) [15].

Table 3.1 : **The list of observed nonverbal cues** [15]

#	Behaviors	Event Type (S/P)	No.	Rate of occurrences (times/minute)			Percentage during observation of the occurrences(%)		
				M	SD	Range	M	SD	Range
Postural behaviors									
1	(-) Shoulder too tight	S	19				60.94	23.80	12.67 - 98.50
2	(-) Legs closed	S	12				73.02	36.44	5.15 - 100
3	(-) Legs too stretch	S	3				61.42	11.33	19.18 - 100
4	(-) Weight in on foot	S	20				65.42	28.69	5.20 - 100
5	(-) Chin too high	S	14				64.94	23.80	12.67 - 98.50
6	(-) Hands in pockets	S	3				11.85	4.89	12.76 - 92.60
7	(+) Lean forward	S	19				32.50	28.66	3.70 - 82.78
8	(-) Lean backward	S	17				62.80	28.07	12.73 - 96.20
Vocal behaviors									
9	(-) Speak too fast	S	19				45.88	36.55	7.32 - 100
10	(-) Start too fast	P	18						
11	(-) Energy decreases at the end	P	23	2.88	1.77	0.53 - 6.31			
12	(+) Vocal emphasis	P	33	5.51	4.51	0.59 - 17.50			
13	(+) Suitable pause	P	33	4.63	3.16	0.53 - 12.50			
14	(-) Unsuitable pause	P	20	1.73	1.14	0.53 - 5.19			
15	(-) Monotone	S	20				92.49	13.08	56.29 - 100
16	(-) Fillers	P	34	5.17	4.22	1.44 - 19.03			
17	(-) Stuttering	P	12	1.72	0.83	0.53 - 3.42			
Behaviors of eye contact									
18	(-) Make eye contact	S	39				93.81	8.24	75.00 - 100
19	(-) Contact avoidance	S	28				9.98	8.47	1.12 - 25.00
19.1	(-) Look up to ceiling	S	14				4.23	2.95	1.12 - 9.61
19.2	(-) Look down to floor	S	19				7.67	4.67	2.84 - 14.17
19.3	(-) Look at hands	S	11				10.24	3.15	4.40 - 13.15
Behaviors related to facial expression									
20	(+) Facial mimicry	S	30				39.31	25.97	4.50 - 91.81
21	(-) Smile	S	22				13.62	11.54	3.54 - 41.08
22	(-) Flat face	S	8				80.61	24.16	40.41 - 100
Behaviors related to whole body movement									
23	(-) Too much movement	S	11				42.21	25.97	4.50 - 91.81
24	(-) Too little movement	S	23				50.62	29.21	10.05 - 100
25	(-) Step backward	P	31	1.83	1.27	0.36 - 4.36			
26	(+) Step forward	P	34	2.06	1.04	0.59 - 4.61			
Behaviors related to hand gesture									
Amount of hand gesture									
27	Hand gesture occur	P	38	16.83	7.15	0.93 - 28.42			
28	(-) Too little gestures	S	20				69.55	34.64	17.21 - 100
29	(-) Too much gestures	S	10				61.49	31.82	27.34 - 96.10
Quality of hand gestures									
30	(-) Bounded gestures	P	30	6.75	5.33	1.00 - 19.77			
31	(+) Relaxed gestures	P	29	7.41	4.95	1.15 - 15.79			
32	(-) Casual gestures	P	10	5.16	3.14	1.56 - 10.28			
33	(-) Uncompleted gesture	P	27	3.23	2.78	0.93 - 10.27			
34	(+) Gestural emphasis	P	20	4.43	4.05	0.36 - 11.99			
35	(-) Repeated gestures	P	31	6.57	2.49	1.09 - 12.31			

The observed behaviors can be separated based on the nonverbal channels that they were generated: (1) Posture (the static configuration of body), (2) Voice (concerning the paralinguistic characteristics), (3) Eye contact, (4) Facial Expression, (5) Globe body movement, (6) Hand gesture. This method of categorization is similar to the literature of public speaking skills [2]. From Nguyen's observation, as well as advices from the expert, they find the following aspects are the most important:

- *Eye Contact* : Similar to social interaction, maintaining good eye contact is the first thing the presenters must keep in mind. It initiates and strengthens the connection between them and the audience (#18, 19 in Table 3.1). It might have the first and foremost influence on the performance of a presentation, as well as regular communications [24].

- *Amount of energy* : This aspect concerns the dynamic characteristics of a presentation, thus can reflect the internal state of the presenters. It has the impact on most behaviors that they have found (except posture as the static channel). For example, the amount of whole body movement (#23, 24), the amount of hand gesture (#28, 29), vocal behaviors (partly via tempo, emphases) and most features of hand gesture.

- *Variety* : The presentations with strong variations significantly increase the attention of the audience [15]. Lacking variation results in monotone (#15), flat face (#22), and hand gesture repeated (#35). In fact, variety can be separated as one single measurement to analyze a presentation. It takes the role as rhythm in music. Even a beautiful piece of music, without changes in rhythm, will steadily lose the attention of the audience.

To evaluate the effectiveness of our proposed system, we need to select some import cues from Nguyen's research. We made a previous experiment, in which we let the evaluator score the trainee's presentation with all 35 cues. However, We find it's too hard for the evaluator to score by all those cues. So we select 20 most occurred cues, and we make a evaluate sheet, like table 3.2 to evaluate the presentation skills of the trainees before and after the training. We will explain the details about our experiment in chapter 5.

Table 3.2：Nonverbal cues for evaluation

#	Behaviors	Event Type (S/P)
Postural behaviors		
1	(-) Hands in pockets	S
2	(+) Lean forward	S
3	(-) Lean backward	S
Vocal behaviors		
4	(-) Speak too fast	S
5	(+) Vocal emphasis	P
6	(+) Suitable pause	P
7	(-) Unsuitable pause	P
Behaviors of eye contact		
8	(-) Make eye contact	S
9	(-) Contact avoidance	S
10	(-) Look up to ceiling	S
11	(-) Look down to floor	S
Behaviors related to facial expression		
12	(-) Smile	S
13	(-) Flat face	S
Behaviors related to whole body movement		
14	(-) Too much movement	S
15	(-) Too little movement	S
16	(-) Step backward	P
17	(+) Step forward	P
Behaviors related to hand gesture		
18	(+) Hand gesture occur	P
19	(-) Too little gestures	S
20	(-) Too much gestures	S

4. Proposed Presentation Training System

In this paper, we propose a presentation training system that allows trainees to imitate past famous speech to improve their nonverbal behaviors. Figure 4.1 shows the overview of our proposed system. Firstly, we employed OpenPose library [6] to extract orators' behaviors as motion data from past famous speech 2D video. Next, the system captures skeletal representations of the trainees' body in real-time using Microsoft Kinect. Then we calculate the cosine similarity of adjacent limbs as features to get the similarity of the trainees' motion and the motion of extracted famous orators. After that, We normalize the similarity and get a score between 0 and 100. Finally, our proposed system will give the trainees some feedback according to their score. We have two kinds of feedback. One shows trainees the number of their score. The other is a kind of visual feedback that contains a virtual hall and some virtual audiences. The audiences will show different actions according to the trainees' score in real-time.

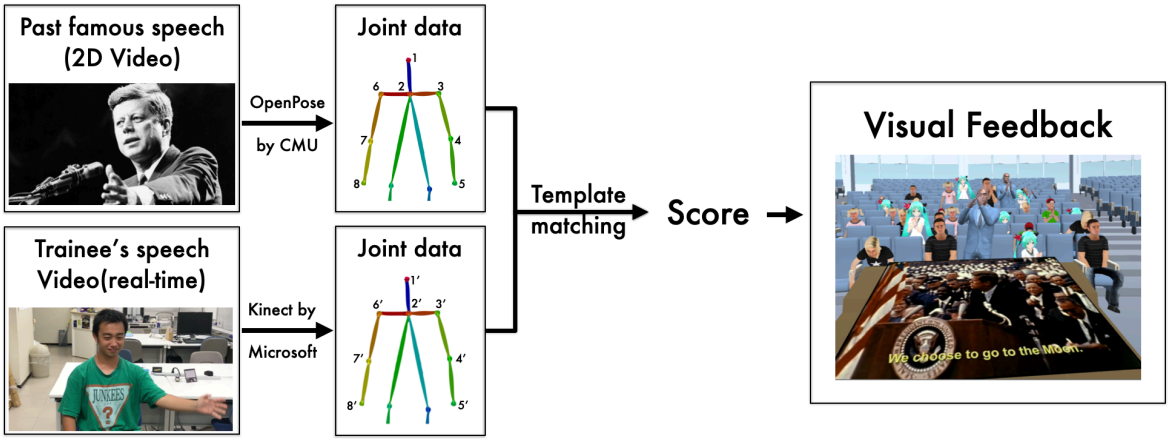

Figure 4.1 : Overview of proposed system

4.1 Extract Pose Data from Past Speech Video

We employed the OpenPose library to extract the orator's joint data from 2D speech video [6]. OpenPose is a library for real-time multi-person keypoint (Figure 3.1) detection and multi-threading written in C++ using OpenCV and Caffe [19]. OpenPose can detect the human body, hand and facial key-points on single images. The detail about how OpenPose works are introduced in chapter 3.

Target Joint

To decide which joint should be detected, we watched about 20 past famous speech, and we fond that most nonverbal behaviors showed in those past famous speeches are the motion of the upper half of the body. In parallel, we also fond that there is always a podium behind the orator (see Figure 4.2). According to those reasons, we chose eight kinds of joints (1 - 8 in Figure 4.3) of the body that includes head, neck, shoulders, elbows, and wrists to determine a motion pattern.

Figure 4.2: An example of past famous speech
(John F.Kennedy's Inaugural Adress in January 20, 1961)

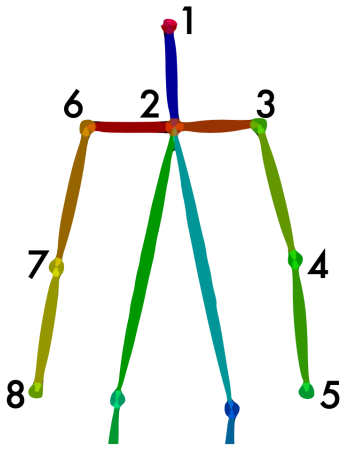

Figure 4.3 : Extracted Joint

Extract Joint Data From Speech Video

To train the presentation skill by imitating speech video, we need an appropriate video clip that should contain some gestures, proper vocal behaviors, and not too long. We watched some speech video and chose the famous *We choose to go to the Moon* as our target video (Figure 4.4).

We choose to go to the Moon is the famous tagline of a speech about the effort to reach the Moon delivered by President John F. Kennedy to a large crowd gathered attic Rice Stadium on September 12, 1962. The speech was intended to persuade the American people to support the Apollo program, the national effort to land a man on the Moon. In his speech, Kennedy characterized space as a new frontier, invoking the pioneer spirit that dominated American folklore. He infused the speech with a sense of urgency and destiny, and emphasized the freedom enjoyed by Americans to choose their destiny rather than have it chosen for them [25]. In this speech, he used gestures to stress his words and use some suitable pause to make his speech more impressive.

Figure 4.4：Speech video：*We choose to go the Moon* (September 12,1962 in Rice Stadium)

We cut the speech video into a short clip, which is 55 seconds, about 1654 frames. We download the source code of the OpenPose library from Github and compile it on a Linux desktop. Then we use the OpenPose library to extract joint data from target clip, and we will get a set of the processed image like figure 4.5 and a set of JSON format data that contains the body joint of the orator like figure 4.6.

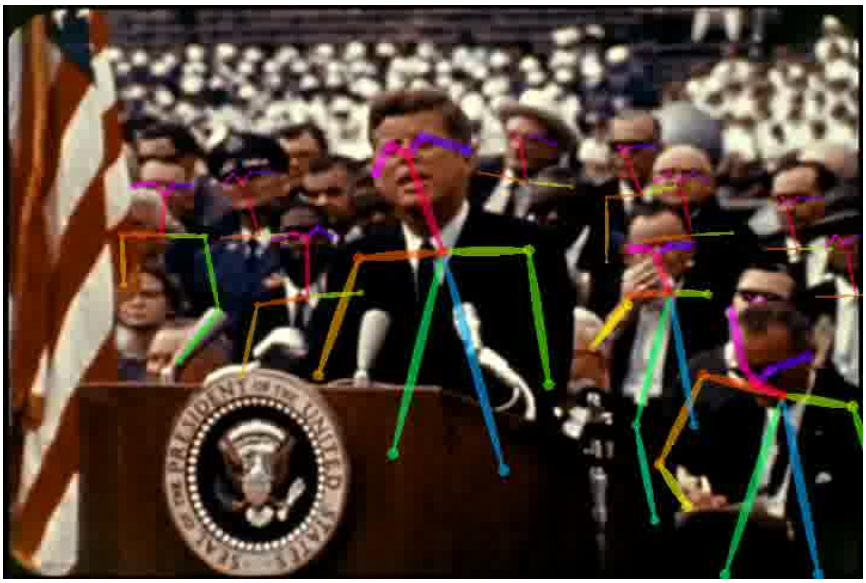

Figure 4.5：An example of processed image

The Figure 4.6 shows the extracted data that contains the 18 kinds of joint. Each line contains the X coordinate, Y coordinate of each joint. The joint is ordered as table 4.1


```
{
  "version": "0",
  "pose_keypoints[]": [
    "382.856","171.406",
    "433.733","213.148",
    "355.439","202.733",
    "318.873","265.39",
    "0","0"
    "505.518","221.006",
    "535.485","346.286",
    "446","368",
  ]
}
```


Figure 4.6 : JSON format data example

#	Joint
1	Head
2	Neck
3	Right shoulder
4	Right wrist
5	Right hand
6	Left shoulder
7	Left wrist
8	Left hand

Table 4.1 : Joint order

Joint data editor

However, OpenPose library can't detect some joint (see Figure 4.7) or detect wrong joint (see Figure 4.8) sometimes due to the quality of speech clip and the effect of the podium.

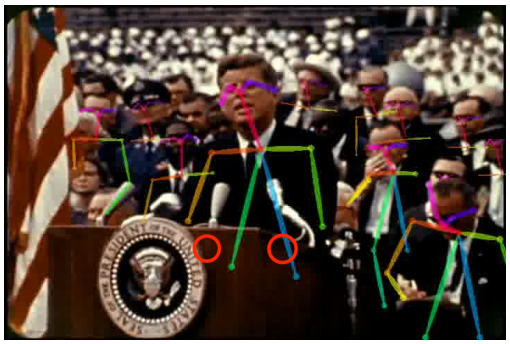

Figure 4.7 : An example of undetected joint

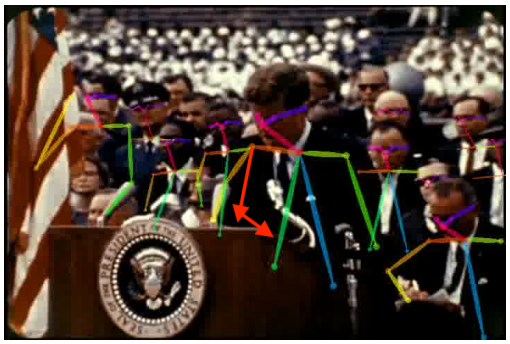

Figure 4.8: An example of wrongly detected joint

In the figure 4.7, we can found that the OpenPose library can't detect the orator's hand (red circle) because of the orator is stand behind the podium. In the figure 4.8, we can find that the orator's right arm is detected wrongly due to the effect of the other people in the video, and the arm should be the position of red arrows. The selected video has 1654 frames, and we found that about half of all frames are not detected correctly.

To make the teacher data more accurate, we developed a joint data editor that can edit each joint data of each frame (Figure 4.9). This editor is wright by HTML and JavaScript and runs in the browser, and we can edit the joint data easily by it. At first, we need to click the joint name

(such as RWrist in Figure 4.9) which is undetected or detected wrongly. Then, we need to click the selected joint's right position in the picture. Finally, we need to click the next button (⟶ in Figure 4.9), the joint data will be saved, and we can edit the next frame.

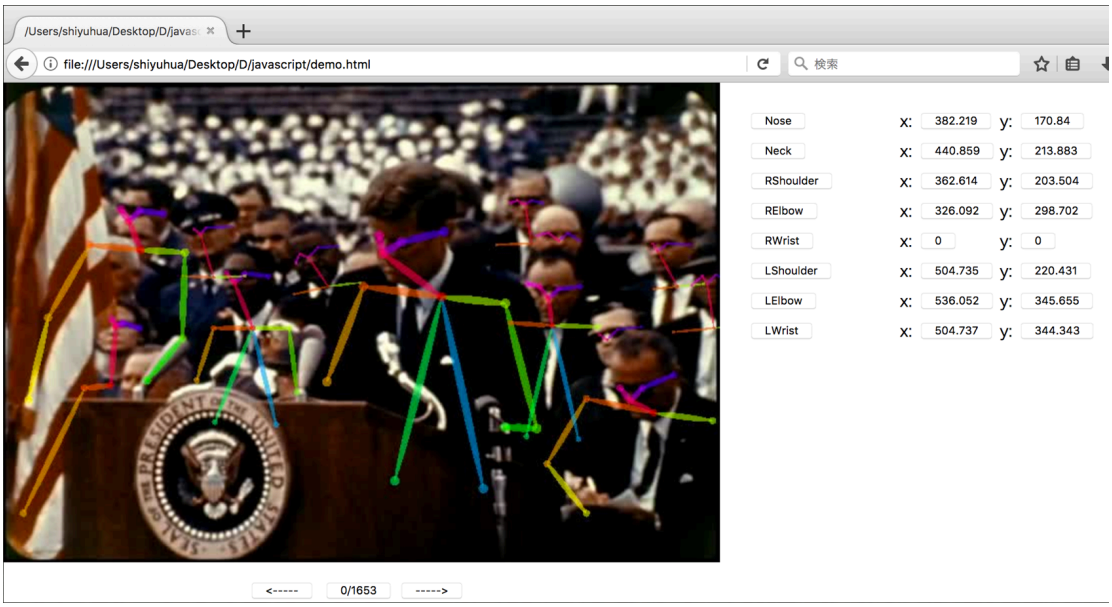

Figure 4.9：Joint data editor

After edited all the frame, we will get a set of joint data. Then we will use those data as teacher data to do a template matching which will be introduced in section 4.3.

4.2 Extract Joint Data from Trainees

To evaluate the trainee's speech and give trainee feedback in real-time. We also need to extract the trainee's joint data in real-time. Although the OpenPose library provides a method for real-time human key points extraction, it can only process about ten frames per second, and it isn't enough for real-time motion matching.

To get trainees' joint data in real-time, we set up a Microsoft Kinect and employ Kinect for Windows SDK to extract joint data. The detail about Kinect was introduced in chapter 3. The Kinect can extract 25 kinds of joints information, and we only use the same 8 kinds of joints data (Fig 4.3) as OpenPose to decrease computational time. Although the Kinect camera can extract X, Y, Z coordinates of the joint, we only use X and Y coordinates because the OpenPose can only extract X and Y coordinates from the 2D video.

4.3 Evaluate Motion of Trainee

To evaluate how well does the trainee imitate past famous speech, we need to calculate the similarity of the trainee's motion and orator's motion in past speech. We employ a template matching algorithm to calculate the similarity. The joint data which extracted from past famous speech is called template data, and the joint data which extracted from the trainee's real-time video is called real-time data. The OpenPose library can extract 18 kinds of the key joint from a human and Kinect can extract 25 kinds of the key joint. However, we only use 8 kinds of key joint (see table 4.1). The correspondence of template data and real-time data is showed in Figure 4.10.

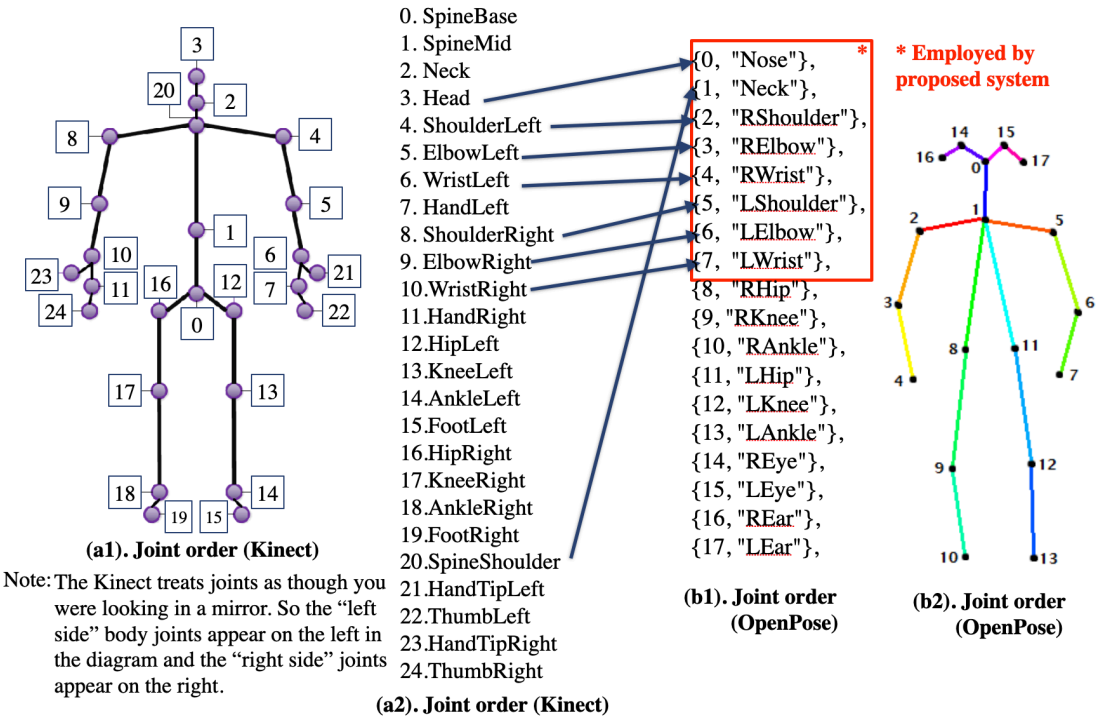

Figure 4.10：Joint correspondence map

Template Matching Alogrithm

The traditional Euclidean distance based method allows calculating the Euclidean distance of past famous speech data and the real-time data to estimate the similarity. However, the Euclidean distance can't reflect the similarity reliably, and it's very sensitive to the position of the camera and noise.

To estimate the similarity between the orator's motion in the past famous speech and the motion of trainees reliably, we calculate the cosine similarity of two adjacent limbs. The Figure 4.11 and 4.12 shows the pipeline of the template matching algorithm.

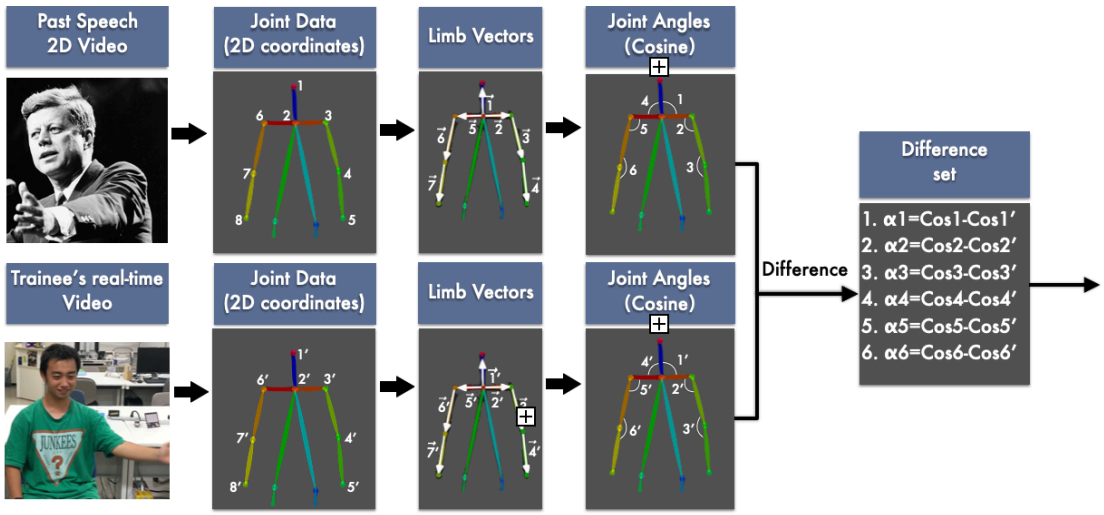

Figure 4.11 : The pipeline of template matching algorithm (part 1)

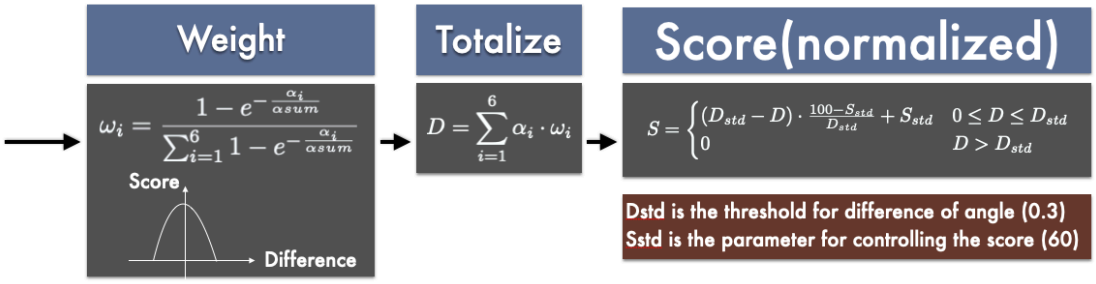

Figure 4.12 : The pipeline of template matching algorithm (part 2)

Suppose that two adjacent joints' coordinate are (X_1,Y_1) and (X_2,Y_2), that limb vector will be:

$$\boldsymbol{n} = (X_1, Y_1) - (X_2, Y_2)\, F \tag{4.1}$$

In accordance with the formula.4.1, we can get a assemble $\boldsymbol{P} = (\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots \boldsymbol{n}_7)$that include seven limbs' vectors (Figure 4.13).

Then calculate the cosine similarity of each two adjacent limb vectors A and B (eg. A for $\vec{1}$, B for $\vec{2}$ in Fig. 4.13) with the formula:

$$\boldsymbol{cos}\ \boldsymbol{\alpha}_i = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{4.2}$$

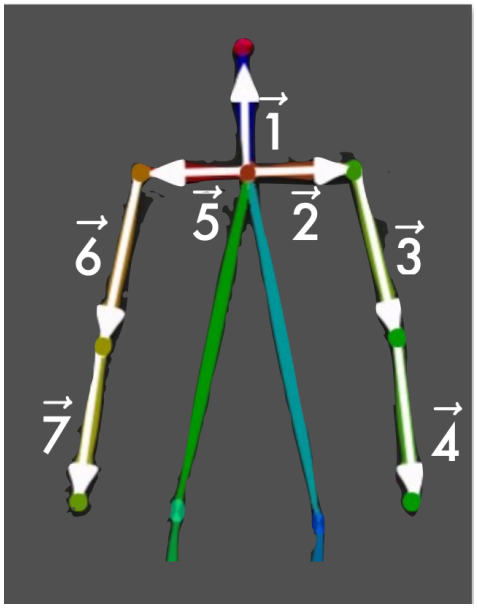

Figure 4.13 : The vector of each limb

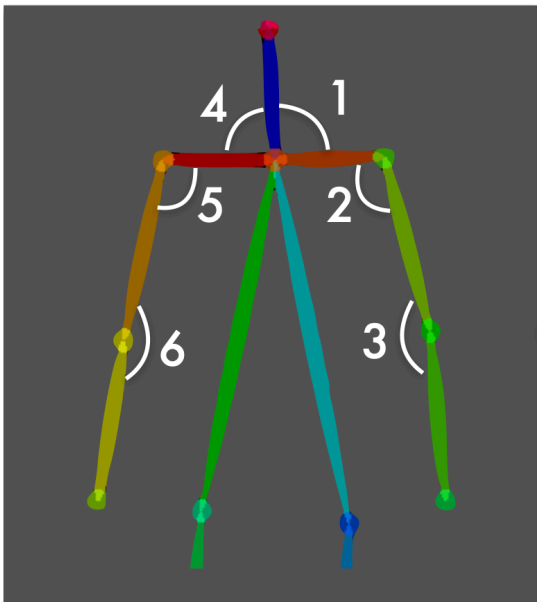

Figure 4.14 : The angle of each two adjacent limbs.

where A_i and B_i are components of vector A and B respectively.

Because each assemble $\boldsymbol{P}$ has 7 vectors, so we get a assemble $\boldsymbol{\theta}_{template} = \{\boldsymbol{\alpha}_{t1},\boldsymbol{\alpha}_{t2},...\boldsymbol{\alpha}_{t6}\}$ that include six cosine similarity of each adjacent limbs (Figure 4.14) , which determine a motion template. In the same way, we can get a assemble $\boldsymbol{\theta}_{real-time} = \{\boldsymbol{\alpha}_{r1},\boldsymbol{\alpha}_{r2},...\boldsymbol{\alpha}_{r6}\}$ with the joint data extracted by the Kinect that determine the real-time motion of trainee's body.

Then calculate the difference of template motion and real-time motion by:

$$\boldsymbol{\theta} = \{(\boldsymbol{\alpha}_{r1} - \boldsymbol{\alpha}_{t1}), (\boldsymbol{\alpha}_{r2} - \boldsymbol{\alpha}_{t2})...(\boldsymbol{\alpha}_{r6} - \boldsymbol{\alpha}_{t6})\} \tag{4.3}$$

And the sum of cosine similarity difference will be:

$$\alpha_{sum} = \sum_{i=1}^{6} \alpha_i \tag{4.4}$$

The α_i represents the cosine similarity of different body parts, and the various parts' contribution for motion matching are diverse. For example, we wouldn't move our neck in a presentation, so the angle 1 (Figure 4.14) is always about 90°, and it shouldn't affect obviously. In parallel, other bodies joint like hand or arm are always used to make some gesture in a presentation, so we should let it affect the score. Therefore we have to give each α_i a weight to compensate for the contribution's difference by:

$$\omega_i = \frac{1 - e^{-\frac{\alpha_i}{\alpha sum}}}{\sum_{i=1}^{6}(1 - e^{-\frac{\alpha_i}{\alpha sum}})} \tag{4.5}$$

According to the formula 4.5, the large difference will have a high weight, and a light weight will be given to the insignificant difference (see the Weight in Figure 4.12). Then we get a assemble $\boldsymbol{W} = \{\omega_1, \omega_2 ... \omega_6\}$, and the sum of $\boldsymbol{\omega_i}$ will be:

$$\sum_{i=1}^{6} \boldsymbol{\omega_i} = \mathbf{1} \tag{4.6}$$

For the difference of cosine similarity, the greater it is, the greater weight will be given. According to the difference of the adjacent limbs' angle and the weight, we can get a totalize as :

$$D = \sum_{i=1}^{6} \alpha_i \cdot \omega_i \tag{4.7}$$

The totalize calculated by the formula.4.7 can show the difference between real-time motion and template motion uniquely. And we normalize the totalize to a score between 0 and 100 by:

$$S = \begin{cases} (D_{std} - D) \cdot \frac{100 - S_{std}}{D_{std}} + S_{std} & 0 \leq D \leq D_{std} \\ 0 & D > D_{std} \end{cases} \tag{4.8}$$

The range of matching degree score is 0-100. The greater the score $\boldsymbol{S}$ is, the real-time motion more similar to the template motion. In the formula 4.8, $\boldsymbol{D_{std}}$ is predetermined threshold for difference of angle, and the matching will be stringent if $\boldsymbol{D_{std}}$ is diminished. To avoid the negative number of score, the score will be set as 0 when $\boldsymbol{D < 0}$. $\boldsymbol{S_{std}}$ is a predetermined parameter for controlling the score in the appropriate range. By several experiments, we set $\boldsymbol{D_{std} = 60}$, $\boldsymbol{S_{std} = 60}$ for best results.

4.4 System UI and Feedback for speech

In training, we allow trainee imitate the past famous speech and the system calculate the similarity of trainee's motion and motion of orator in past famous speech. Then the trainee will get feedback about how well do they imitate.

In our prototype system, we allow trainee watch their skeleton model, past speech video and score in real-time like Figure 4.15. The score will update every second, and the trainee can adjust their motion to get a high score. The more similar the trainee imitating, the higher the score will be.

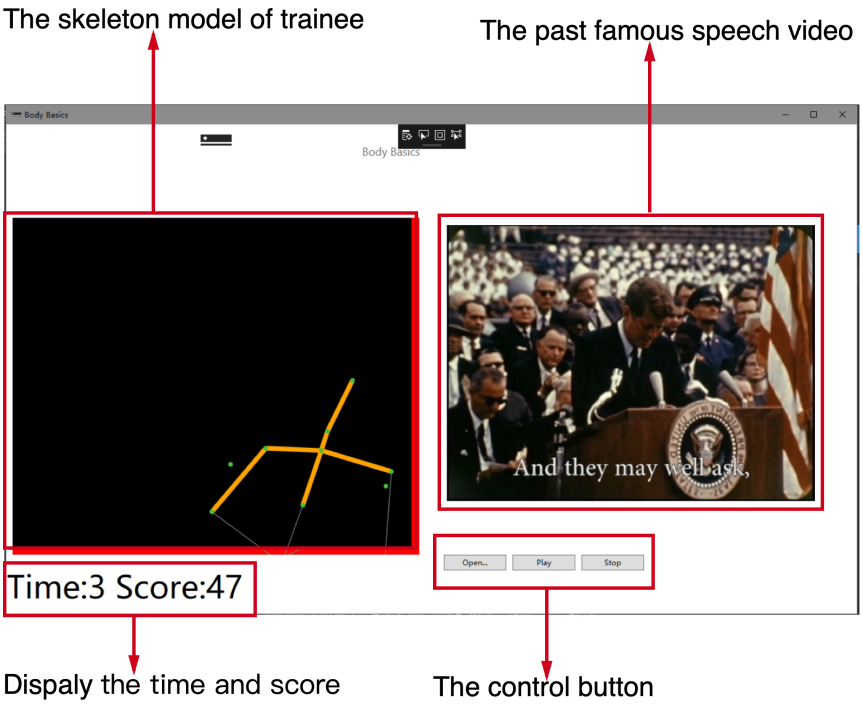

Figure 4.15：Prototype system UI

We let some trainees try our prototype system and did a short interview. After the interview, we found that the trainee always looks at both the score and their skeleton model in training, so they can't focus on the speech. To solve this problem, we improved our system by connecting our system with a real-time visual feedback system like Figure 4.16.

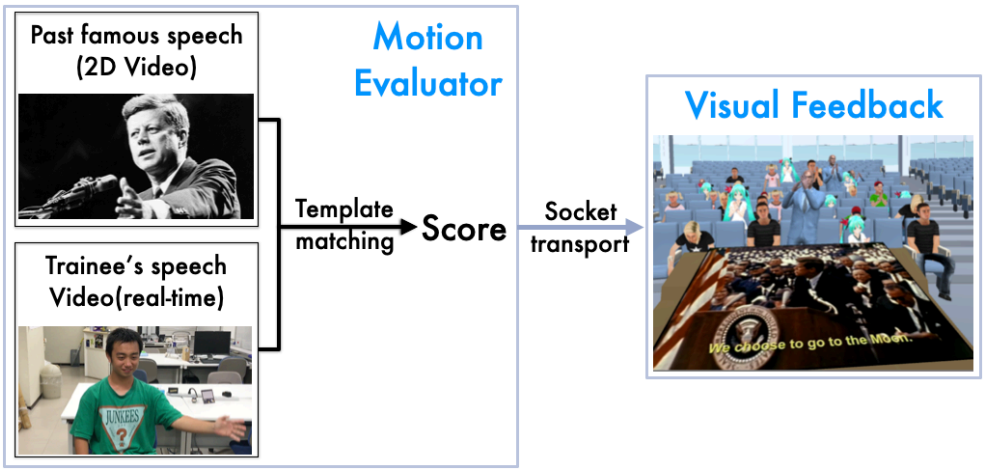

Figure 4.16：System structure

Visual Feedback System [26]

The visual feedback system is an independent system build by Unity (see Figure 4.17). When the trainee wear an Oculus Rift (a kind of HMD), they can see a virtual hall and some virtual audiences. In the front of the virtual hall, there is a podium, and the past speech video will play over it. There are about 30 audiences, and when the trainee start imitating, the audiences will do different actions according to the score.

Figure 4.17：Visual feedback [26]

5. Experiment

5.1 Evaluation of the Algorithm Effectiveness

To verify the effectiveness of the matching algorithm, we make a simple experiment.

Method

We let an orator record a simple video instead of past famous speech which includes some common nonverbal behaviors. The video has 12 seconds (360 frames) that include 5 kinds of nonverbal behaviors such as wave hand. We recruited three subjects and introduced the system briefly, then showed them the video once. Then let them imitate the video one by one for three times and record the score. For reference, we also let the orator imitate the same video, and the orator can imitate exactly and get a high score.

Result and Discussion

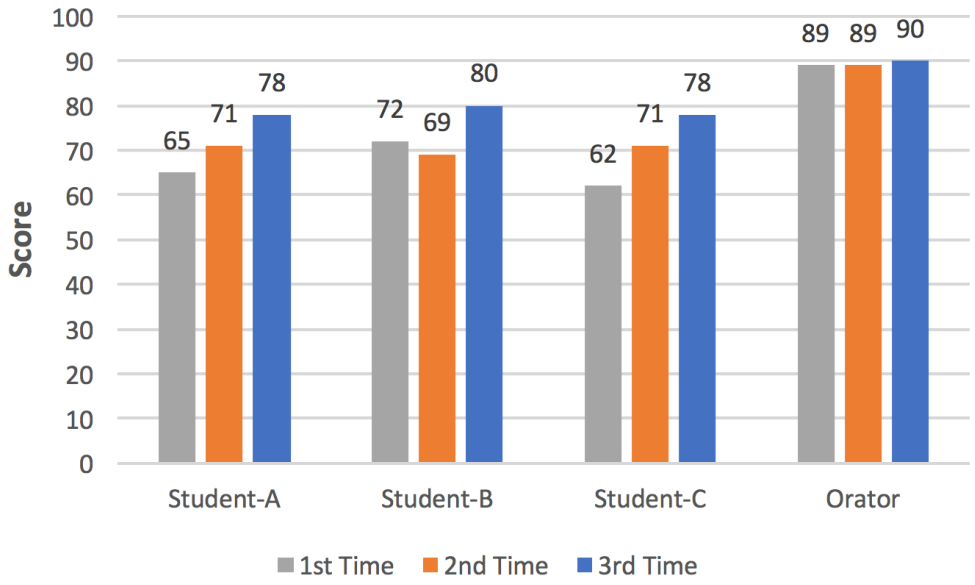

Figure 5.1 : The score of subjects and orator

In the Figure.5.1, we find that the orator gets a high score in all three times that prove the effectiveness of the matching algorithm. The three subjects didn't get a good score for the first time, but they imitate the orator better for the next two times.

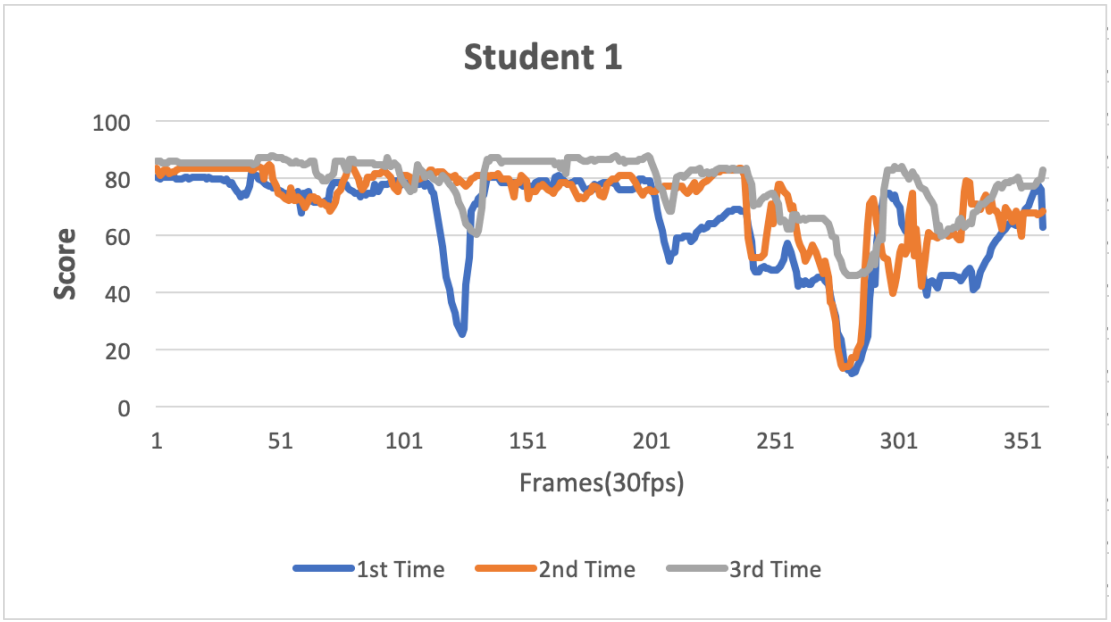

Figure 5.2：The score of student-A

The Figure.5.2 shows the time variable score of student-A. In the Figure .5.2, we can find that the student-A did not imitate the motion well in the frame 121 and frame 281 for the first time. Moreover, after training, he imitates better at the third time and gets a higher score. The orator can imitate his motion well and get a score in all three time. The student can not imitate those motion at first and get a low score. However, when the student imitates three times, they might remember the motion and imitate well, so they get a higher score at the last time. This result shows the effectiveness of the proposed matching method.

5.2 Evaluation of the System Effectiveness

We made a experiment to verify the effectiveness of our proposed system.

Method

We made an A/B test to verify will the trainee's presentation skill be better after training by our proposed system. We recruited 10 subjects from Ritsumeikan University, which include some undergraduate students and some graduate students, and divide them into two groups (Group A and Group B) randomly. We also recruited 3 evaluators from Ritsumeikan University. The Figure 5.3 shows the process of this experiment.

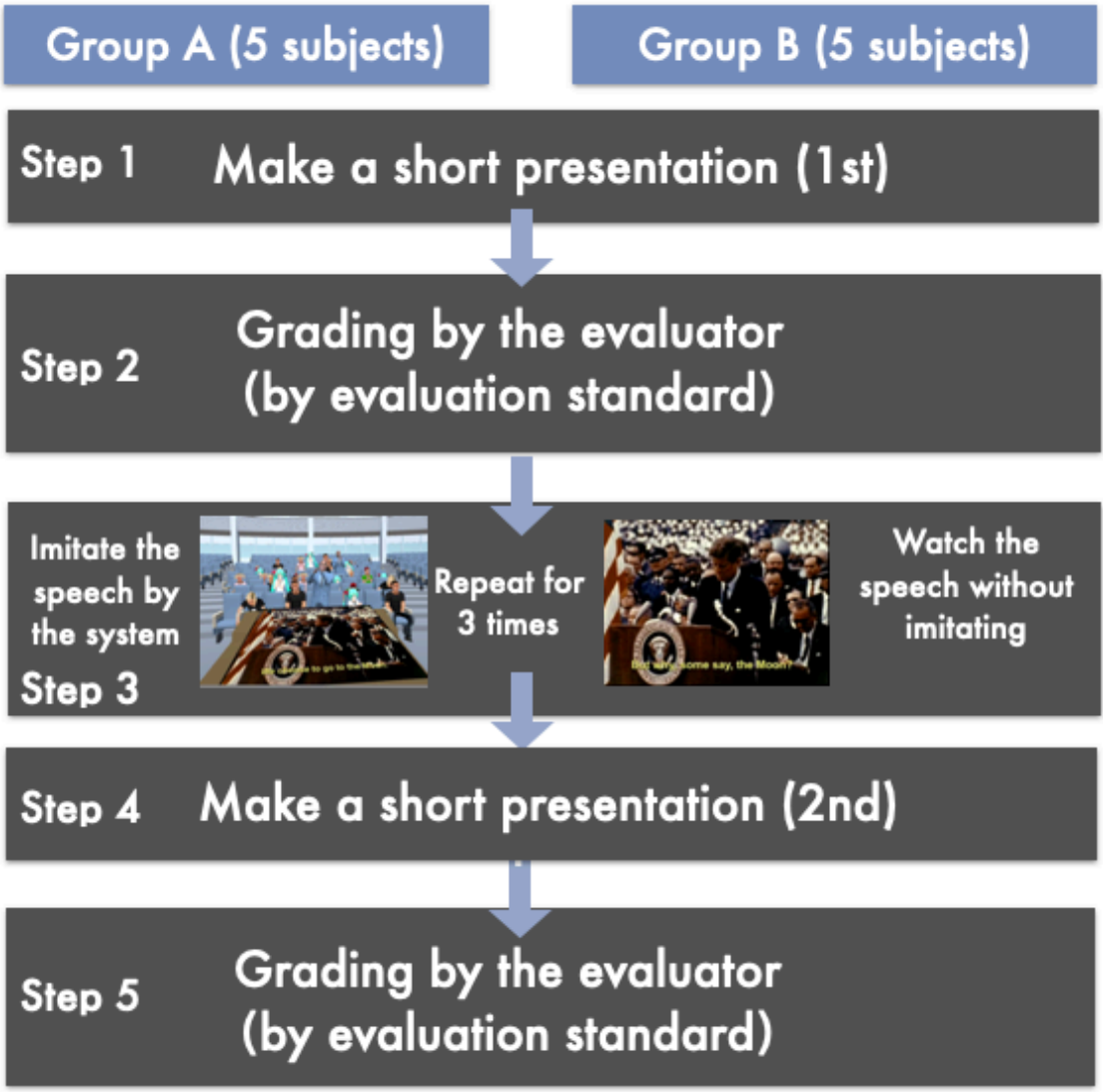

Figure 5.3 : The process of experiment

Step1 At first, we let the subject make a short presentation. We offer the subject a PowerPoint about the student's life in Ritsumeikan University. It has eight pages and includes some images and simple text (see Figure 5.4). We give each subject ten minutes to prepare for the presentation, and then let them make the presentation freely in front of three evaluators and one stuff. We also record video to review the subject's presentation.

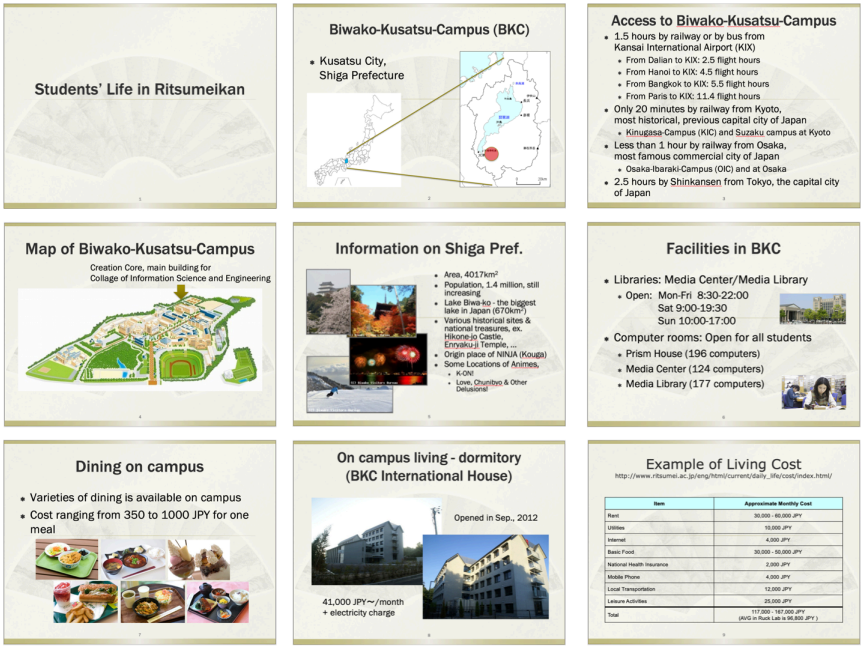

Figure 5.4 : Offered PowerPoint

Step2 At the same time, we let three evaluators score the subject's speech by an evaluation sheet like Figure 5.5. The evaluation standard has been introduced in chapter 3.3 and table 3.2.

Step3 After the first presentation, we let the subjects do some training. We let the subjects of Group-A (A for short) use our proposed system. In training, they were asked to imitate the speech, and they can get the visual feedback from the system. In parallel, we just let the subjects of Group-B (B for short) watch the original speech video. Both subjects of A and B should do the training for three times.

Step4 After training, we let the subjects make a short presentation again. The content of the presentation is the same as the presentation for the first time.

Step5 Finally, the evaluator score the subject's speech again (Figure 5.5).

Result

We collected the evaluation sheet from evaluators and calculated the score of each subject. The Figure 5.5 shows an example of the result sheet.

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets	1					
-1	Lean backward						
1	Lean forward	1					1
Whole body movement							
-1	Too much movement						
-1	Too little movement	1	1	1			
-1	Step backward						
1	Step forward						
Vocal behaviors							
-1	Speak too fast						
-1	Unsuitable pause	2	2	3			2
1	Suitable pause	1			2	1	
1	Vocal emphasis	1					
Behaviors of eye contact							
1	Make eye contact				3	1	1
-1	Contact avoidance	1	1	1			
-1	Look up to ceiling		1				
-1	Look down to floor						1
Facial expression							
1	Smile				1		1
-1	Flat face	1	1	1			
Hand gesture							
1	Hand gesture occur		2		5	5	1
-1	Too little gestures	1		1			
-1	Too much gestures						
Total point of each evaluator		-4	-4	-7	11	7	1
Average Point (Rounding)		-5			6		
DIfference		11					

Figure 5.5：An example of result sheet

In the Figure 5.5, the first row is the factor of each behavior. 1 means it's positive behavior and will plus one point, -1 means it's negative behavior and will minus one point. The three rows behind 1*st* or 2*nd* show the point which is given by evaluator 1, 2, 3. We calculate the average score of three evaluators and the increased score of the score before training and the score after training. The Figure 5.6 and Figure 5.8 is the score of each subject in Group-A and Group-B. The blue bar shows the score of each subject's first presentation, and the green bar shows the score of the subject's score after training. The Figure 5.7 and Figure 5.9 is the increased score of each subjects.

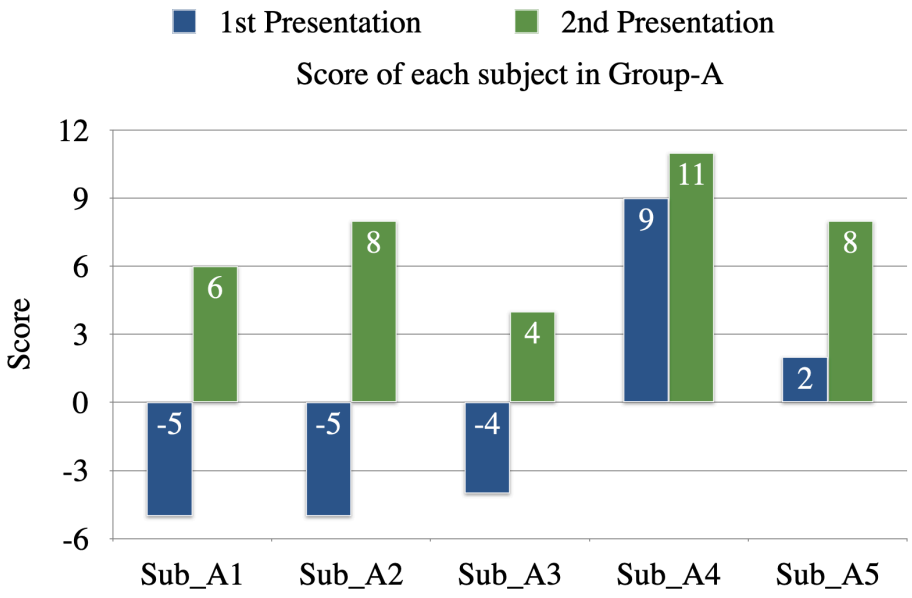

Figure 5.6：Score of each subject in Group-A

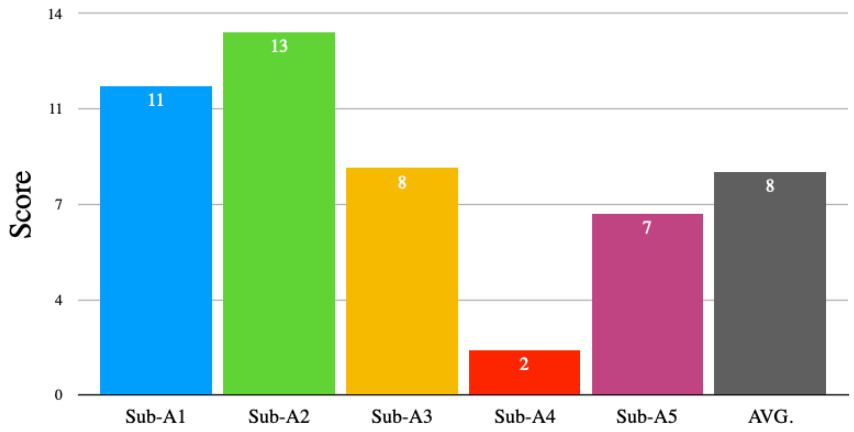

Figure 5.7：Increased score of the subjects in Group-A

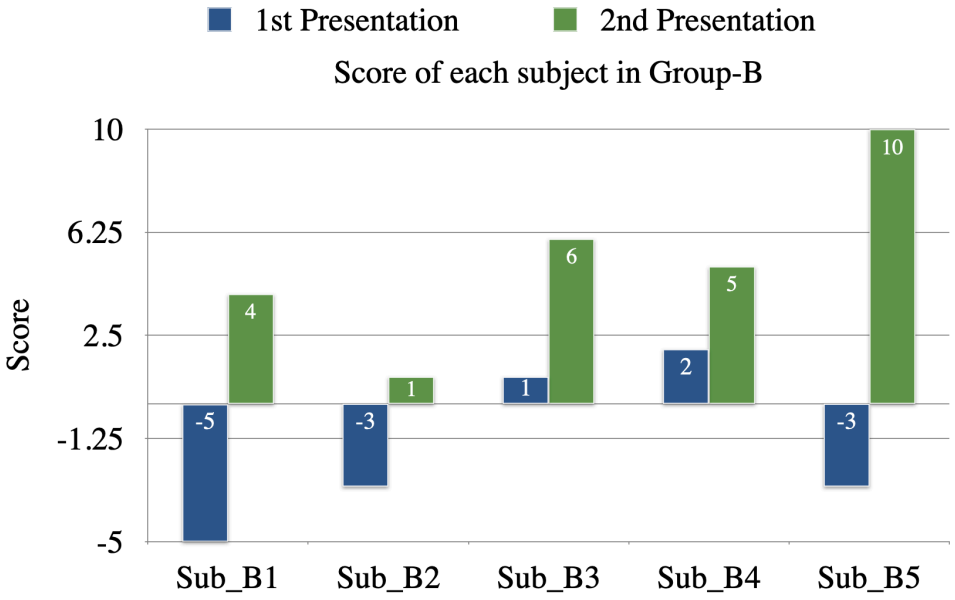

Figure 5.8：Score of each subject in Group-B

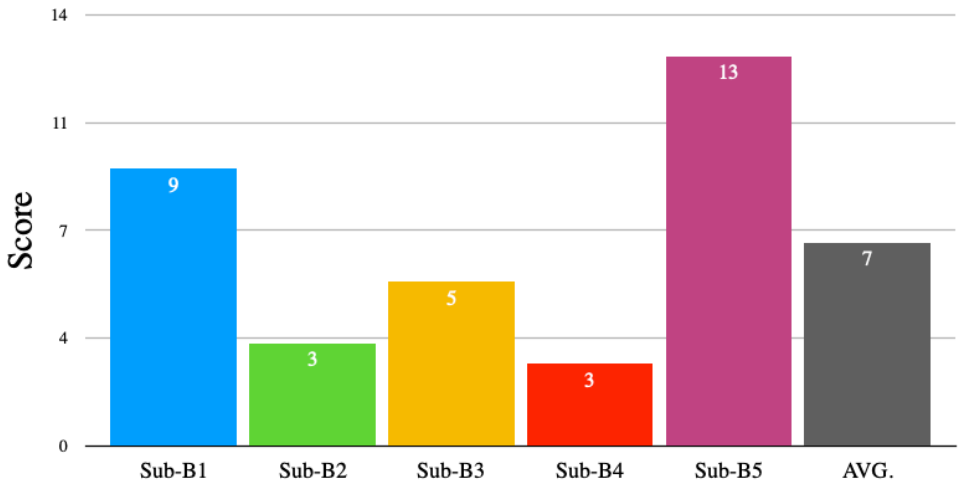

Figure 5.9：Increased score of the subjects in Group-B

The Figure 5.6 shows the score of each subject in Group-A. From the figure, we can see that the subject A1, A2, and A3 didn't give a good presentation at the first time. After the training, we can find that these three subjects perform better than the first time. The subject A4 and A5 give a not lousy presentation at the first time. Notably, the subject A4 get a very high score before the training. After training, the subject A4 and A5 also perform better than the first time but nor very obviously.

The Figure 5.8 shows the score of each subject in Group-B. From the figure, we can find that

the subject B1, B2, and B5 give a bad presentation at the first time. In parallel, the subject B3 and B4 perform not bad for the first time. After the training by watching speech video, all of the subjects perform better but not very obviously except subject B5. The detail about the subject's presentation will be discussed in section 5.2.

According to the Figure 5.7 and Figure 5.9, all of subjects perform better after training by our system or not. We find that the average increased score of Group-A is 8, and the average increased score of Group-B is 7. It proves that the Group-A improved more than the Group-B and the proposed system is effective to train the presentation skill. We did an independent two-sample t-test and the P-value approximately equal to 0.285 (single tail). So we can't reject the null hypothesis. However, the score of subject A4 and B5 are not representative. If we ignore the score of subject A4 and B5, the P-value will be 0.026 (single tail), and we reject the null hypothesis. According to this result, we improved the effectiveness of our proposed system.

Discussion

After the experiment, we made a short interview and analysis some subject's presentation.

From the recorded video, we find that the subject A2 didn't give a great presentation at the first time, he always avoided the eye contact with the evaluator and stay the same gesture (see Figure 5.10). After training, he made more gesture and more eye contact with the evaluator consciously(see Figure 5.11).

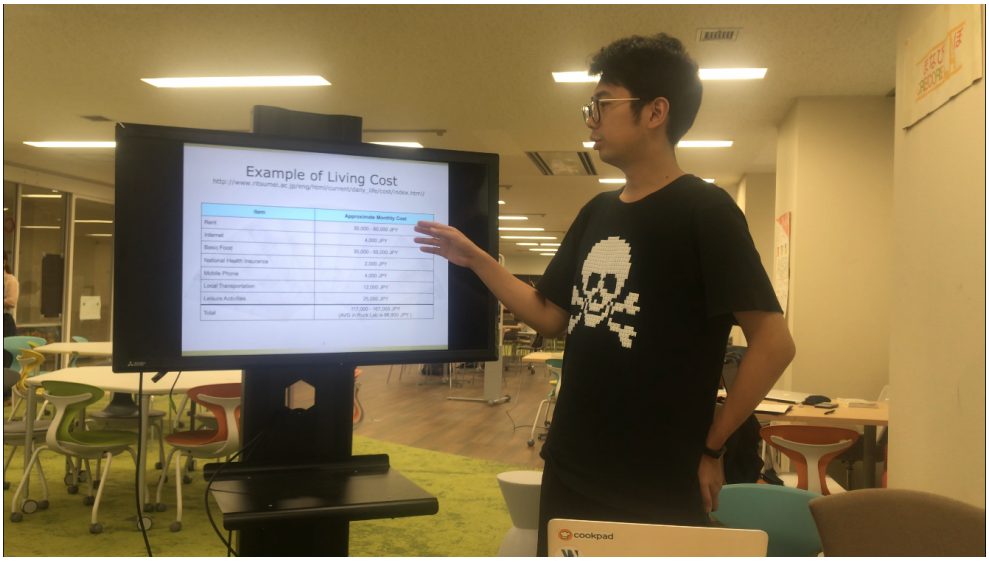

Figure 5.10 : The subject A2 before training

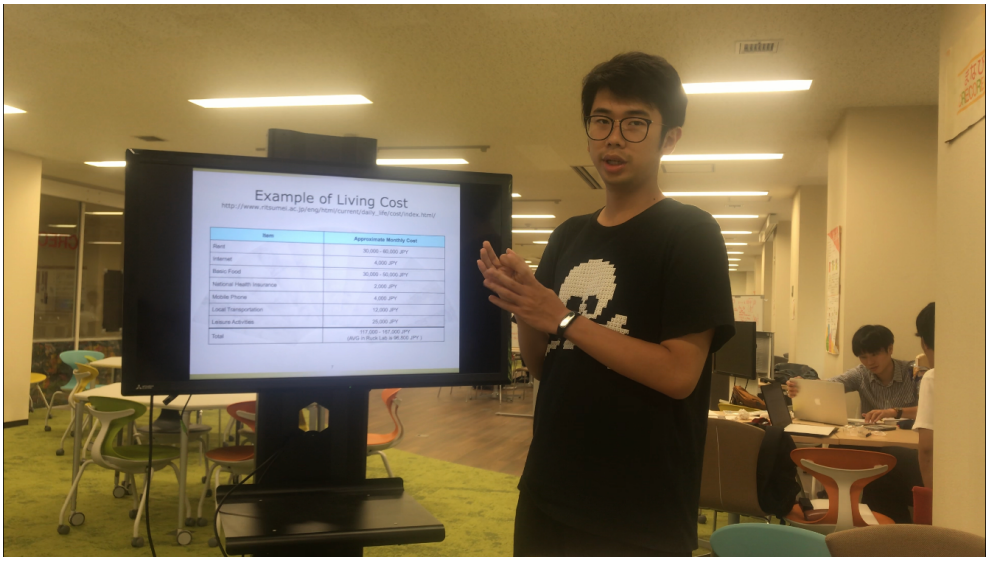

Figure 5.11 : The subject A2 after training

The Figure 5.12 shows a part of the evaluation sheet for subject A2. From the sheet, we also find that all of the three evaluators marked *Contact avoidance* and *Too little gestures* at the first time. After training, the evaluators marked *Make eye contact* and *Hand gesture occur*.

In the interview, the subject A2 tell us that Kennedy's speech is awe-inspiring because not only what he said but also his gesture and vocal behavior. In training, the subject A2 was asked to imitate those behaviors, so he does some gesture and eye contact consciously at the second presentation.

The subject A4 got a very high score in both twice presentations. From the recorded video, he shows his presentation skills for the first time. After training by propose system, he uses more gesture to make his presentation more intelligible. The evaluation sheet also shows that he performs better in the second time.

According to the interview with other subjects, we also know that imitating famous past speech lets the trainees improve their skills in gesture, vocal behavior and eye contact. In parallel, some subjects said the visual feedback make their training more interesting and have effect on overcome the nervous.

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Behaviors of eye contact							
1	Make eye contact				1	1	1
-1	Contact avoidance	1	1	1			
-1	Look up to ceiling						
-1	Look down to floor						
Facial expression							
1	Smile						1
-1	Flat face	1	1	1			
Hand gesture							
1	Hand gesture occur			1	5	6	5
-1	Too little gestures	1	1	1			
-1	Too much gestures						

Figure 5.12：A part of evaluation sheet for subject A2

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Hand gesture							
1	Hand gesture occur	5	6	5	6	7	6
-1	Too little gestures						
-1	Too much gestures					1	

Figure 5.13：A part of evaluation sheet for subject A4

6. Conclusion

In this paper, we propose a presentation training system that allows the trainee to imitate past famous speech to improve their presentation skills.

In advance, we employed OpenPose library[6] to extract orators' motion data from past famous speech 2D video. OpenPose is a library for real-time multi-person key-point detection, which can extract 2D coordinates from 2D image or video. We select the John F.Kennedy's *We choose to go to the Moon* as our target video. In this speech, he used some gestures to stress his words and used some suitable pause to make speech more impressive. We think it's a suitable speech for our proposed system. We select eight kinds of joint as our target joints because most behaviors showed in our target speech are the behaviors of the upper half of the body.

While training, the system capture the trainee's motion in real-time using a Microsoft Kinect camera (Kinect for short). The Kinect can extract 25 kinds of joint's 3D coordinates from a human body. We only employ eight kinds of joint's 2D coordinates to pair the joints which extracted by OpenPose from famous past speech.

Then, we employ a template algorithm to compare the trainee's speech with the famous past speech. We choose the cosine similarity of adjacent limbs as the feature to calculate the score that shows the similarity of the trainees' motion and the motion of extracted orators.

Finally, the system gives the trainee visual feedback to make the training more effective. In training, the trainee wears an HMD (Oculus Rift) that shows a virtual hall and some virtual audience. The audience will perform some actions according to the score in real-time. For example, when the trainee gets a high score, the audience applauds for the trainee.

We made two experiments to verify the effectiveness of our algorithm and system. In the first experiment, we proved that our algorithm could compare the trainee's motion and the orator's motion of famous past speech. In the second experiment, we made an A/B test to verify will imitating past famous speech by our proposed system can improve trainee's presentation skill. The experiment result shows that the trainee performs better after training. From the interview and the recorded video, we find that imitating past famous speech let the trainee use more gestures and eye contact in their presentation. It shows the effectiveness of our proposed system.

In our present system, we only employ the upper body's gestures to judge the trainee's presentation. However, we also consider some other behaviors in the experiment such as whole body

movement, vocal behaviors, and postural behaviors. Those behaviors are also necessary because there are many kinds of style of presentation. Some speech like inaugural speech (A speech given during this ceremony which informs the people of his or her intentions as a leader.) or congress speech needs the speaker stands behind a podium and can't move a lot. In the other recent speech style like TED or conference, the speaker moves in the stage and make more interaction with the audience.

In the future, we would like to consider more behaviors such as eye contact, whole body movement and vocal behaviors in our proposed system. We also need to improve our algorithm and feedback for more effective training.

Acknowledgement

本論文の執筆にあたり、多岐に渡るご指導をしてくださり、私を導いてくださった野間春生教授に深く感謝の言葉を申し上げます。至らぬ私が無事に論文の執筆を終えるができたのは野間春生教授の教えがあったからです。本当にありがとうございます。また、鋭い指摘や温かい助言を下さった Lopez Rober Gulliver 准教授に深く感謝の意を述べさせて頂きます。スイカバーを嬉しく召し上がる姿には、心を癒されました。そして同研究室の松村耕平講師と大井翔講師からは、ゼミナールを中心に一歩引いた視点からの意見をいただきました。ありがとうございました。

研究の日々を互いに切磋琢磨し、励まし合いながら過ごした研究室の皆様に輝かしい未来があるようお祈り申し上げます。自らもお忙しいなか、快く実験への協力を承諾いただきました被験者の皆様への恩は一生涯胸に残します。研究のみならず数々の困難を共に乗り越えてきた同級生の方々が, これからも健やかに過ごし、皆様の力を存分に発揮できるような日々を送られますよう心からお祈りいたしま。

そして何よりも、この 24 年間私を育て、支えてくれた両親に心から感謝いたします。

References

[1] W. J. Seiler and M. L. Beall, "Communication: making connections," 2002.

[2] G. Rodman and R. B. Adler, "Style: Delivery and Language Choices," *The New Public Speaker*, 1996.

[3] M. Argyle, F. Alkema, and R. Gilmour, "The communication of friendly and hostile attitudes by verbal and non-verbal signals," *European Journal of Social Psychology*, vol. 1, no. 3, pp. 385–402, 1971.

[4] EnglishClub, "Communicating with Body Language," *Technically Speaking*, p. 2, 2002.

[5] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded Pyramid Network for Multi-Person Pose Estimation,"

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," in *Cvpr*, vol. 1, p. 7, 2017.

[7] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional Multi-person Pose Estimation," nov 2016.

[8] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards Accurate Multi-person Pose Estimation in the Wild," jan 2017.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," mar 2017.

[10] T. Pfister and P. Robinson, "Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 66–78, 2011.

[11] D. A. Silverstein and Tong Zhang, "System and method of providing evaluation feedback to a speaker while giving a real-time oral presentation," apr 2006.

[12] R. Hincks and J. Edlund, "Promoting increased pitch variation in oral presentations with transient visual feedback," *Language Learning & Technology*, vol. 13, no. 3, pp. 32–50, 2009.

[13] T. Gao, C. Wu, H. A. C. V. W. (ICCV, and undefined 2009, "User-centric speaker report: Ranking-based effectiveness evaluation and feedback," *ieeexplore.ieee.org*.

[14] K. Kurihara, M. Goto, and J. Ogata, "Presentation sensei: a presentation training system using speech and image processing," *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, pp. 358–365, 2007.

[15] A. T. Nguyen, W. Chen, and M. Rauterberg, "Intelligent presentation skills trainer analyses body Movement," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9095, pp. 320–332, 2015.

[16] J. Heer, N. Good, A. Ramirez, M. D. . . . . C. on human . . . , and undefined 2004, "Presiding over accidents: system direction of human action," *dl.acm.org*.

[17] M. Goto, K. Itou, T. Akiba, and S. Hayamizu, "Speech Completion: New Speech Interface with On-demand Completion Assistance," *Proceedings of HCI International*, vol. 1, pp. 198–202, 2001.

[18] J. Shotton, A. Fitzgibbon, A. Blake, A. Kipman, M. Finocchio, B. Moore, and T. Sharp, "Real-Time Human Pose Recognition in Parts from a Single Depth Image," jun 2011.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," jun 2014.

[20] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," jan 2016.

[21] Z. Zhang, "Microsoft Kinect Sensor and Its Effect," *IEEE Multimedia*, vol. 19, pp. 4–10, feb 2012.

[22] P. . Fankhauser, M. . Bloesch, D. . Rodriguez, R. . Kaestner, M. . Hutter, R. Y. Siegwart, P. Fankhauser, M. Bloesch, D. Rodriguez, R. Kaestner, M. Hutter, and R. Siegwart, "Kinect v2 for Mobile Robot Navigation: Evaluation and Modeling," 2015.

[23] J. Sell, O. P. I. Micro, and undefined 2014, "The xbox one system on a chip and kinect sensor," *ieeexplore.ieee.org*.

[24] P. H. Zimmerman, J. E. Bolhuis, A. Willemsen, E. S. Meyer, and L. P. J. J. Noldus, "The Observer XT: A tool for the integration and synchronization of multimodal signals," *Behavior Research Methods*, vol. 41, pp. 731–735, aug 2009.

[25] J. J. R. &. P. Affairs and undefined 2003, "Kennedy's romantic moon and its rhetorical legacy for space exploration," *muse.jhu.edu*.

[26] T. yi Huang, "The Proposal of Virtual Audience for Public Speech Training System," 2018.

Appendix

Evaluation sheet of each subject

Subject A1

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets	1					
-1	Lean backward						
1	Lean forward	1					1
Whole body movement							
-1	Too much movement						
-1	Too little movement	1	1	1			
-1	Step backward						
1	Step forward						
Vocal behaviors							
-1	Speak too fast						
-1	Unsuitable pause	2	2	3			2
1	Suitable pause	1			2	1	
1	Vocal emphasis	1					
Behaviors of eye contact							
1	Make eye contact				3	1	1
-1	Contact avoidance	1	1	1			
-1	Look up to ceiling		1				
-1	Look down to floor						1
Facial expression							
1	Smile				1		1
-1	Flat face	1	1	1			
Hand gesture							
1	Hand gesture occur		2		5	5	1
-1	Too little gestures	1		1			
-1	Too much gestures						
Total point of each evaluator		-4	-4	-7	11	7	1
Average Point (Rounding)		-5			6		
DIfference		11					

Figure A. 1 ：Evaluation sheet of subject A1

Subject A2

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets						
-1	Lean backward	1	1				
1	Lean forward						
Whole body movement							
-1	Too much movement				1		
-1	Too little movement	1	1	1			
-1	Step backward		1			1	
1	Step forward		1		1	1	1
Vocal behaviors							
-1	Speak too fast			1			
-1	Unsuitable pause	1		2	1		
1	Suitable pause		1			1	1
1	Vocal emphasis	1			1	1	1
Behaviors of eye contact							
1	Make eye contact				1	1	1
-1	Contact avoidance	1	1	1			
-1	Look up to ceiling						
-1	Look down to floor						
Facial expression							
1	Smile						1
-1	Flat face	1	1	1			
Hand gesture							
1	Hand gesture occur			1	5	6	5
-1	Too little gestures	1	1	1			
-1	Too much gestures						
Point of each evaluator		-5	-4	-6	6	9	10
Average Point (Rounding)		-5			8		
DIfference		13					

Figure A. 2 ：Evaluation sheet of subject A2

Subject A3

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets						
-1	Lean backward		1				
1	Lean forward			1	1		
Whole body movement							
-1	Too much movement						
-1	Too little movement	1	1				
-1	Step backward		1				
1	Step forward		1	1			1
Vocal behaviors							
-1	Speak too fast			1			
-1	Unsuitable pause	1	1	1	1		
1	Suitable pause					1	1
1	Vocal emphasis	1			1		
Behaviors of eye contact							
1	Make eye contact			1	1	1	1
-1	Contact avoidance	1	1	1			
-1	Look up to ceiling			1			1
-1	Look down to floor						
Facial expression							
1	Smile						1
-1	Flat face	1	1	1		1	
Hand gesture							
1	Hand gesture occur			1	1	5	1
-1	Too little gestures	1	1	1			
-1	Too much gestures						
Point of each evaluator		-4	-6	-2	3	6	4
Average Point (Rounding)		-4			4		
DIfference		8					

Figure A. 3 ：Evaluation sheet of subject A3

Subject A4

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets						
-1	Lean backward						
1	Lean forward	1	1	1	1	1	1
Whole body movement							
-1	Too much movement	1	1	1	1	1	1
-1	Too little movement						
-1	Step backward						
1	Step forward						
Vocal behaviors							
-1	Speak too fast						
-1	Unsuitable pause						
1	Suitable pause	1	1	1	2	2	2
1	Vocal emphasis	1	1	1	1	1	1
Behaviors of eye contact							
1	Make eye contact	1	1	1	1	1	1
-1	Contact avoidance						
-1	Look up to ceiling						
-1	Look down to floor						
Facial expression							
1	Smile	1	1	1	1	1	1
-1	Flat face						
Hand gesture							
1	Hand gesture occur	5	6	5	6	7	6
-1	Too little gestures						
-1	Too much gestures					1	
Point of each evaluator		9	10	9	11	11	11
Average Point (Rounding)		9			11		
DIfference		2					

Figure A. 4 ：Evaluation sheet of subject A4

Subject A5

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets						
-1	Lean backward						
1	Lean forward	1		1	1		1
Whole body movement							
-1	Too much movement	1		1	1	1	1
-1	Too little movement		1				
-1	Step backward						
1	Step forward	1			1		
Vocal behaviors							
-1	Speak too fast						
-1	Unsuitable pause			1	1		1
1	Suitable pause		1		1	1	1
1	Vocal emphasis				1		1
Behaviors of eye contact							
1	Make eye contact				1	1	1
-1	Contact avoidance	1	1	1			
-1	Look up to ceiling						
-1	Look down to floor			1			
Facial expression							
1	Smile	1		1	1	1	1
-1	Flat face		1				
Hand gesture							
1	Hand gesture occur	3	3	4	6	5	7
-1	Too little gestures		1	1			
-1	Too much gestures				1		1
Point of each evaluator		4	0	1	9	7	9
Average Point (Rounding)		2			8		
DIfference		7					

Figure A. 5 ：Evaluation sheet of subject A5

Subject B1

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets	1	1	1			
-1	Lean backward						
1	Lean forward			1			
Whole body movement							
-1	Too much movement		1				
-1	Too little movement	1					
-1	Step backward			1			
1	Step forward						
Vocal behaviors							
-1	Speak too fast			1			
-1	Unsuitable pause	1	1	1		1	
1	Suitable pause				1		1
1	Vocal emphasis	1	1		1	1	1
Behaviors of eye contact							
1	Make eye contact				1	1	1
-1	Contact avoidance	1	1	1			
-1	Look up to ceiling						
-1	Look down to floor	1					
Facial expression							
1	Smile				1	1	1
-1	Flat face						
Hand gesture							
1	Hand gesture occur				1	1	1
-1	Too little gestures	1	1	1			
-1	Too much gestures						
Point of each evaluator		-5	-4	-5	5	3	5
Average Point (Rounding)		-5			4		
DIfference		9					

Figure A. 6 ：Evaluation sheet of subject B1

Subject B2

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets						
-1	Lean backward		1	1		1	
1	Lean forward	1	1			1	
Whole body movement							
-1	Too much movement	1		1			
-1	Too little movement						
-1	Step backward		1	1		1	
1	Step forward			1			
Vocal behaviors							
-1	Speak too fast	1		1	1		1
-1	Unsuitable pause	1		1	1		
1	Suitable pause						
1	Vocal emphasis					1	
Behaviors of eye contact							
1	Make eye contact	1		1	1		1
-1	Contact avoidance		1			1	
-1	Look up to ceiling						
-1	Look down to floor						
Facial expression							
1	Smile						
-1	Flat face	1	1	1	1	1	1
Hand gesture							
1	Hand gesture occur	1	1	2	2	2	3
-1	Too little gestures	1	1	1			
-1	Too much gestures						
Point of each evaluator		-2	-3	-3	0	0	2
Average Point (Rounding)		-3			1		
Difference		3					

Figure A. 7 ：Evaluation sheet of subject B2

Subject B3

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets						
-1	Lean backward		1				
1	Lean forward		1	1		1	1
Whole body movement							
-1	Too much movement						
-1	Too little movement	1	1	1			
-1	Step backward						
1	Step forward						
Vocal behaviors							
-1	Speak too fast		1				
-1	Unsuitable pause	1	1	1			
1	Suitable pause				1	2	1
1	Vocal emphasis			1			
Behaviors of eye contact							
1	Make eye contact	1	1	1	1	1	1
-1	Contact avoidance						
-1	Look up to ceiling						
-1	Look down to floor	1	1	1			
Facial expression							
1	Smile						
-1	Flat face	1		1	1		1
Hand gesture							
1	Hand gesture occur	3	3	3	4	3	4
-1	Too little gestures						
-1	Too much gestures						
Point of each evaluator		0	0	2	5	7	6
Average Point (Rounding)		1			6		
DIfference		5					

Figure A. 8 ：Evaluation sheet of subject B3

Subject B4

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets	1	1	1	1	1	1
-1	Lean backward						
1	Lean forward						
Whole body movement							
-1	Too much movement						
-1	Too little movement	1	1	1	1	1	1
-1	Step backward						
1	Step forward						
Vocal behaviors							
-1	Speak too fast						
-1	Unsuitable pause						
1	Suitable pause	2	1	3	3	2	3
1	Vocal emphasis	1	1	1	1	1	1
Behaviors of eye contact							
1	Make eye contact	1	1	1	1	1	1
-1	Contact avoidance						
-1	Look up to ceiling						
-1	Look down to floor						
Facial expression							
1	Smile	1	1	1	1	1	1
-1	Flat face						
Hand gesture							
1	Hand gesture occur				1	1	1
-1	Too little gestures	1	1	1			
-1	Too much gestures						
Point of each evaluator		2	1	3	5	4	5
Average Point (Rounding)		2			5		
DIfference		3					

Figure A. 9 ：Evaluation sheet of subject B4

Subject B5

		1st			2nd		
Evaluator number		**1**	**2**	**3**	**1**	**2**	**3**
Postural behaviors							
-1	Hands in pockets			1			
-1	Lean backward						
1	Lean forward			1		1	1
Whole body movement							
-1	Too much movement			1			
-1	Too little movement	1					
-1	Step backward		1	1		1	
1	Step forward					1	1
Vocal behaviors							
-1	Speak too fast	1	1	1			
-1	Unsuitable pause						
1	Suitable pause				1	4	1
1	Vocal emphasis	1			1	1	
Behaviors of eye contact							
1	Make eye contact				1	1	1
-1	Contact avoidance	1	1	1			
-1	Look up to ceiling		1	1			
-1	Look down to floor	1					
Facial expression							
1	Smile		1		1	1	1
-1	Flat face			1			
Hand gesture							
1	Hand gesture occur	1	5		4	4	5
-1	Too little gestures			1			
-1	Too much gestures	1					
Point of each evaluator		-3	2	-7	8	12	10
Average Point (Rounding)		-3			10		
DIfference		13					

Figure A. 10 ：Evaluation sheet of subject B5