Vowel Priority Lip Matching Scheme and Similarity Evaluation Model Based on Humanoid Robot Ren-Xin

Zheng Liu [a], Xin Kang [a], Shun Nishide [a], Fuji Ren*[a]
[a]School of Information Faculty of Engineering, Tokushima University, Japan

ABSTRACT

At present, the significance of humanoid robots dramatically increased while this kind of robots rarely enters human life because of its immature development. The lip shape of humanoid robots is crucial in the speech process since it makes humanoid robots look like real humans. Many studies show that vowels are the essential elements of pronunciation in all languages in the world. Based on the traditional research of viseme, we increased the priority of the smooth transition of lip between vowels and propose a lip matching scheme based on vowel priority. Additionally, we also designed a similarity evaluation model based on the Manhattan distance by using computer vision lip features, which quantifies the lip shape similarity between 0-1 provides an effective recommendation of evaluation standard. Surprisingly, this model successfully compensates the disadvantages of lip shape similarity evaluation criteria in this field. We applied this lip-matching scheme to Ren-Xin humanoid robot and performed robot teaching experiments as well as a similarity comparison experiment of 20 sentences with two males and two females and the robot. Notably, all the experiments have achieved excellent results.

Keywords: Lip matching; vowel priority; similarity evaluation; humanoid robot; Manhattan distance

1. INTRODUCE

Robots gradually entered people's daily life from the structured environment of the industry since the mid-1980s. These robots can complete work autonomously, complete tasks by cooperation with humans and even complete tasks under the guidance of humans in numerous conditions, such as hospital, office, home and other cluttered and uncontrollable environments. Among them, humanoid robots are the spotlight among many fields since humanoid robots are similar to humans in appearance and structure, and can run, jump, and even carry things like humans. Therefore, we firmly believe humanoid robots will play an essential role in the future human society (Lu H et al. 2018; Ren F and Bao Y 2020).

Atlas is a famous humanoid robot because of its super balance and stability, which was developed by Boston Dynamics (Kuindersma S et al. 2016). Atlas can complete some human tasks and even can perform many skills such as handstand, 360-degree flip as well as rotation. There are also some other humanoid robots, such as the ERICA humanoid robot in the research laboratory invented by Professor Ishiguro of Osaka University, and the Ren-Xin robot modelled by Professor Ren Fuji of Tokushima University. Both of them have human-like rubber skin, and more and more research has been investigated on these robots, enabling them to perform human-like expressions and interact with human beings (Ishi C T et al. 2018; Ren F and Huang Z 2016).

However, the deficiency of intelligence of existing humanoid robots is stifled by the current technology. Hence, a large number of researchers focus on developing the intelligence of robots, so that robots can have perception and cognitive abilities like humans (Hwang J et al. 2017; Herath D C et al. 2017). Speech lip-matching is one of the core technologies of robot expressions and visual speech, and has broad application prospects, while there are only a handful of researches examined on the humanoid robot lip.

The labium superioris and labium inferioris of human beings change with the pronunciation in the process of speaking, which associated with the cooperative interactions between muscle and nerve, such as orbicularis oris muscle and facial nerve. All vocal organs cooperate to produce corresponding sounds. In the process of pronunciation, lip shape will go through three stages repeatedly, forming mouth shape-> keeping mouth shape-> releasing mouth shape. In the first stage, lip is excessively moving towards the state of minimum airflow resistance. The second stage is to keep the mouth shape.

*ren@is.tokushima-u.ac.jp; phone +81886569684; fax +81886566575;

The oral structure is stable, and the airflow resistance is minimum. This stage is usually the vowel keeping stage. In the third stage, the mouth shape is released, in which the vocal cords stop vibration and the lip shape returns to the initial stationary state. If the sound is continuous and uninterrupted, then the third stage will directly become the next round of keeping the mouth shape stage. Through this cyclic vocalization mechanism, we can find that although people's mouth shape is constantly moving and changing during the speech process, there is a rule to follow.

The mainstream languages used by humans all have corresponding characters. Language consists of vocabulary, vocabulary consists of syllables, and syllables consist of vowels and consonants. The vowel is a sound produced by airflow through the oral cavity without resistance during pronunciation. Each vowel has its corresponding mouth shape. The point of this study is how to smoothly convert the vowel into the robot's mouth shape according to the priority of vowels.

Unlike other studies, our study for the first time takes advantage of the priority of vowels in the pronunciation of the language and proposes a robot lip matching scheme. For any language, in the application of lip movement, our method has the characteristics of fast speed, strong versatility and high similarity. At the same time, by using Manhattan distance and computer vision technology, we also propose a method for calculating lip similarity, which makes up for the gap in the calculation of lip similarity in the field of humanoid robots.

The rest of the paper is arranged as follows. The second section of this paper mainly introduces the information of lip shape, humanoid robots and visual tracking by summarizing the related research. In the third section, we introduce our methods, which are vowel priority, lip parameter extraction process, lip similarity evaluation method, and Ren-Xin humanoid robot details. In the fourth section, we introduce a subjective experiment and an objective experiment compared with real persons. In the fifth section, we discuss the performance of our algorithm and the advantages and disadvantages of our lip matching scheme and evaluation model. Last but not least, the sixth section reviews our experiments and introduces our future work.

2. RELATED WORK

Artificial intelligence has attracted considerable attention from many researchers for a long time. As the carrier of artificial intelligence, the research on the appearance of humanoid robots and the research on inner-thinking emotions (Ren F 2009) are hot topics in the research. Fu K et al. (2019) used MobileNet V2 and U-Net to effectively recognize the text contained in the picture, which strengthened the robot's visual ability. In speech emotion recognition, Liu Z et al. (2019) analyzed the feature distribution of different speech lengths, which had an effect on speech emotion recognition and effectively improved the robot's speech hearing ability. Ren F et al. (2015; 2015) strengthened the robot's ability to understand the text by extracting emotional features from the text. For humanoid robots, in order to be more harmoniously integrated into human life, internal thinking and external form need to be considered together.

Human-like pronunciation organs are essential in the humanoid robot. The design of the pronunciation organ structure and system is the basis of robotic speech. Nishikawa K et al. (2004) simulated the interaction between the human-like pronunciation organs. A human-like speech robot constructed by machine not only can produce the most basic vowels but also can produce tones, rubs and nasal sounds through a new flexible mechanism. Based on some formed robot hardware, there are also some researches on relevant lip control systems to make robot speech more realistic (Luo R C et al. 2011; Oh K G 2010), combining the existing robot hardware to enable the robot to make different mouth shapes flexibly. Similarly, there are also some control system design schemes for a specific language (Hara F et al. 1997; Herath D C 2017). These basic structural system designs have promoted the development of humanoid talking robots.

In recent years, more and more lip-shaped related researches aiming at language characteristics have emerged. Research on lip matching in Chinese was first carried out 20 years ago. Yan Jie (1998) defined the pronunciation of Chinese Pinyin into six basic mouth shapes and proposed a text-driven lip synthesis method. Zeng Hongxin et al. (2013) combined with the pronunciation characteristics of Chinese speech, analyzed the mechanism of matching speech and mouth shape from two aspects of geometric shape matching and time matching. Fan Xinxin (2017) proposed a mouth shape animation synchronization algorithm, which can generate mouth shape animation synchronized with the input speech signal according to the input speech signal, and combine speech and mouth shape. In addition to the Chinese language, there are also many lip-shaped related studies in other languages. Tsuyoshi et al. (2015) coded Japanese

pronunciation patterns and obtained good results in subjective evaluation of the patterns generated according to the codes. There are still other language characteristics such as English, Korean (Morishima S et al. 1991; Kim T H 2008).

The language features of the above lip-shaped studies have several problems. First, the languages studied above are relatively single and not universal. Second, many of them are not combined with robots, and the results are not convincing enough in an application. Of course, the above studies also have a common feature, most of which contain the most basic constituent element of language: vowels. The vowel is an essential component of all languages. There are also many related studies on lip shape based on vowels such as (Saitoh T et al. 2010; Sulistijono I A 2014), which all show that vowels have strong expansibility in lip shape expression.

Visual tracking is quite popular in the current field of visual research. It refers to the detection, extraction, recognition and tracking of moving objects in video sequences for the next step of processing and analysis. From the traditional manual extraction of features in images, until now, neural network feature extraction (Li P et al. 2018), the range of its applications is getting wider and wider. The detection technology of face key points in the video is one of the hot spots in the field of visual tracking. Cootes et al. (2000) proposed the ASM model, which uses parametric sampling shapes to form an object shape model, and uses the PCA method to establish control points that describe the shape. Based on the ASM model, (2001) added the texture information of the objects, and proposed an upgraded AAM model. Sun Yi et al. (2013) first time used deep learning methods to locate key points on the face. The emergence of deep learning has outstanding practical significance for the development of visual tracking models (Li X et al. 2018; Long T 2019; Dai K et al. 2020).

In this paper, by summarizing the previous research, we explore the lip changes under the interaction of different vowels and propose a lip matching scheme for Chinese and Japanese based on vowel priority, which is different from the traditional single vowel viseme study and is universal than the previous research. To measure the effectiveness of our scheme, we also use computer vision technology to construct an evaluation model to evaluate the similarity between lips in the two pronunciation processes.

3. METHOD

In the first part, combining Chinese and Japanese, we propose the priority of vowel transition in pronunciation and define the process of temporal change. The second part introduces how to extract the mouth size of vowel pronunciation through pronunciation video. The third part introduces the evaluation method we put forward to measure lip similarity. In the fourth part, we combine the parameters of a humanoid robot and apply the method to the robot.

Vowel Pronunciation Priority

According to the Chinese Pinyin Scheme, Pinyin consists of 21 initials and 35 finals. The pronunciation of every word in Chinese can be expressed by initial consonants combined with finals. The vowel contains several vowel letters, and the initial consonant determines how the pronunciation organ will block the vowel expressed by the vowel (Binyong Y and Felley M 1990). Therefore, the initial consonant determines the form of the vowel, and the vowel determines the size and duration of the mouth shape.

Japanese is similar to Chinese and has its vowel, which corresponds to Roman characters a, i, u, e, o (Keating P A et al. 1984). Japanese is mainly composed of Chinese characters and pseudonyms. There are two mainstream Roman forms of expression of Japanese, which are Hepburn and Kunrei. In our work, we use the Kunrei form for writing.

Take the Chinese language as an example. During the pronunciation process, the vowel part especially has priority. The vowel with higher priority lasts longer, the feature of which is most obvious. According to previous studies, vowel priority in pinyin is a->e->o-> u(v)->i from high to low. As shown in Tables 1 and 2, we combined the vowel rules of Chinese Pinyin and Japanese and divided both of them into four types according to vowel priority: single sound, front sound, middle sound and rear sound.

The mouth shape of different vowel priority types is different, as shown in Table 3. The single sound contains a vowel, the stress is on only one vowel, the mouth shape transits from 1/3 of the natural state to the vowel, and 2/3 of the time stays above the vowel. The front sound contains two vowels. The stress stays on the first vowel, the mouth shape transits from the natural state 1/4 of the time to the first stressed vowel, after 1/2 of the time stays, 1/4 of the time transits to the second vowel. The middle sound contains three vowels. The stress is on the middle vowel, the mouth shape transits from

the natural state 1/5 time to the first vowel, 1/5 time to the second vowel, after 2/5 time stay, the last 1/5 time transits to the third vowel. The rear sound is similar to the front sound and contains two vowels. The stress stays on the second vowel. The mouth shape transits from the natural state to the first stressed vowel in 1/4 of time, then transits to the second vowel in 1/4 of the time and stays on the second vowel in 1/2 of the time.

Table 1. Types of Chinese Vowel Pronunciation.

Priority Types	Vowel part of Chinese characters
single sound	a , o , e(r) , i , u , v , an , en , ang , eng, ong , in , ing, vn
front sound	ai , ei , ao , ou
middle sound	iao , iou , uai , uei
rear sound	ia , ie , ua , uo , ve , ian , iang , uan , uen , uang , ueng , van , iong

Priority Types	Vowel part of Japanese characters
single sound	a , i , u , e , o
rear sound	ya , yu , yo , yan , yun , yon

Table 2. Types of Japanese Vowel Pronunciation

Table 3.Different Types of Vowel Changes

Vowel priority types	Process of mouth shape
single sound	1/3 S->maxV , 2/3 maxV
front sound	1/4 S->maxV1 , 1/2 maxV1 , 1/4 maxV1->maxV2
middle sound	1/5 S->maxV1 , 1/5 maxV1->maxV2 , 2/5 maxV2 , 1/5 maxV2->maxV3
rear sound	1/4 S->maxV1, 1/4 maxV1->maxV2 , 1/2maxV2

Also, by combining with the above classification, we have observed the proportion of different vowel types distribution commonly used in daily Chinese and Japanese. In Chinese, we collected common dialogue sentences appearing in Weibo, and in Japanese, we collected sentences from Japanese social networking site Himabu. After statistics, as shown in Figure 1, in Chinese, the proportion of single sound, front sound, middle sound and rear sound is 59%, 14%, 8% and 19% respectively. In Japanese, the proportion of single sound and rear sound is 94% and 6%.

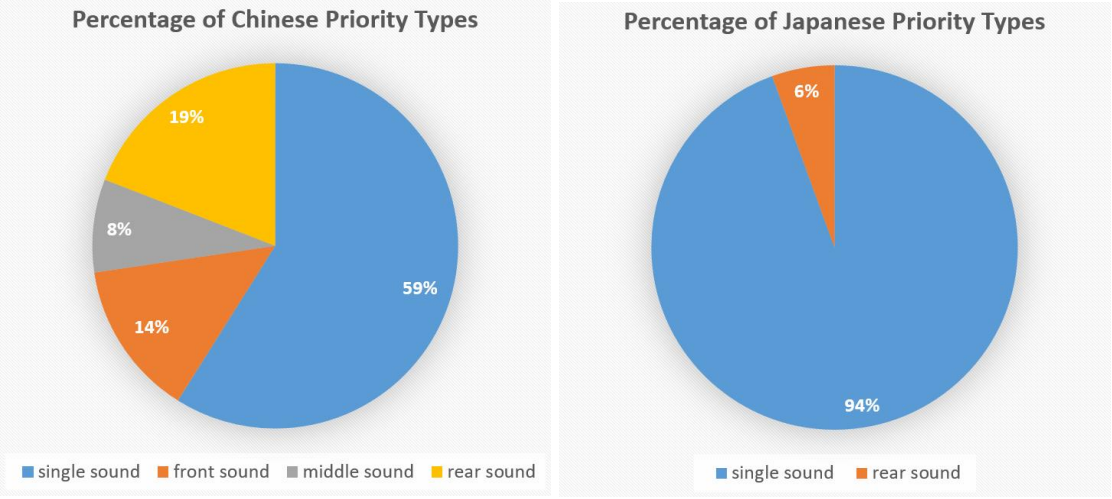


Figure 1. The proportion of Vowel Priority Types in Daily Use of Chinese and Japanese

Mouth Size of Vowels

When different vowels are combined, there is a transition of mouth shape in the process of pronunciation. The upper and lower length of the mouth shape, which is the smallest in the natural state, will increase with the opening pronunciation. In the stage of keeping the mouth shape, the upper and lower length of the mouth shape is the largest. The left and right width of the mouth shape is the largest in the natural state and will decrease with the opening pronunciation. In the stage of keeping the mouth shape, the left and right width of the mouth shape is the smallest.

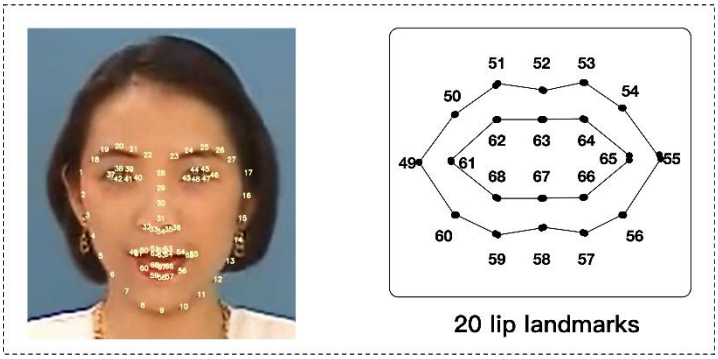


Figure 2. Lip-related feature landmarks

To obtain the maximum length and minimum width when pronouncing different vowel patterns, we extract 68 feature points of the face in the video frame through the existing vowel video, of which 20 feature points are related to the vowel patterns, as shown in Figure 2. The relative position coordinates (Xi, Yi) of each feature point Pi, through the following formulas 1 and 2, calculate the lengths and widths of different frame mouth shapes and finally obtain the upper and lower maximum lengths and the left and right minimum widths of each vowel.

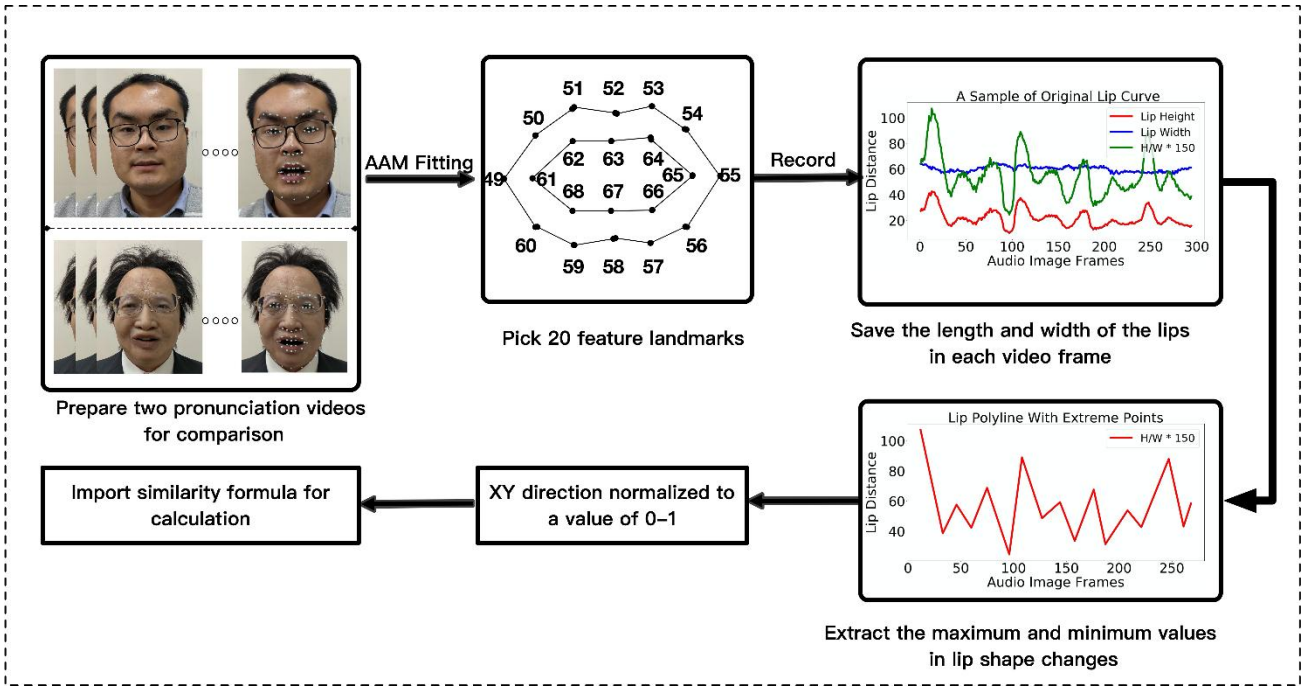
$$H = (Y_{51}+ Y_{52}+ Y_{53}+ Y_{62}+ Y_{63}+ Y_{64}-(Y_{57}+ Y_{58}+ Y_{59}+ Y_{66}+ Y_{67}+ Y_{68}))/6 \quad (1)$$

$$W = (X_{49}+X_{50}+X_{60}+X_{61}-(X_{54}+X_{55}+X_{56}+X_{65}))/4 \quad (2)$$

Through calculation and statistics, we normalized several vowel mouth shapes between 0 and 1, as shown in Table 4, which provides a mathematical basis for the subsequent experimental application of manipulating the lip shape of the humanoid robot.

Table 4.The Mouth Size of Different Vowels

Vowels	The ratio of upper and lower length	The ratio of left and right width
a	1	0.97
o	0.68	0.09
e	0.70	0.79
i	0.71	0.94
u(v)	0.35	0
Silence	0	1



Evaluation Model

Figure 3. Evaluation model process.

To verify the effectiveness of our method, the evaluation model is a significant and crucial step. As for the evaluation methods of robot lip similarity, most of them currently adopt subjective evaluation methods to evaluate whether the two lips are similar according to the judgment of human eyes and perception. For example, in the paper (Hyung H J et al. 2016), by recording videos of real people and robots, viewers can evaluate whether they are similar. This method is not objective enough. Therefore, we also use the computer vision method, combining with Manhattan distance, and propose a new evaluation model.

Figure 3 is the whole process of the evaluation method. For the two videos to be compared, we first extract 64 feature points of the face in each picture frame in the video, of which 20 feature points are about lip. Using the calculation method in 3.2, we record the lip shape along with the length and width of the frame. To facilitate the calculation and reduce the error caused by the distance between the cameras, we define the change value of the lip shape length-width ratio. Next, we extract the extreme points in the aspect ratio as the key points for evaluation. These extreme points can be used to reflect the maximum position reached by the lip shape at a specific time point and can effectively retain the time and space data of the lip shape change process. Next, we normalize the XY direction values of all extreme points to values between 0 and 1. Finally, the similarity calculation formula is used.

Manhattan distance is one of the most widely used evaluation models. For two points $X(X_1, X_2)$ $Y(Y_1, Y_2)$, the point X is (X_1, X_2) and the point Y is (Y_1, Y_2), and its Manhattan distance can be expressed as formula 3. In this evaluation model, Manhattan distance means the difference between lip shapes in two videos.

$$MH(X,Y) = |X_1 - Y_1| + |X_2 - Y_2| \tag{3}$$

When m points are used for comparison

$$\mathbf{X}=\{X_1(X_{11}, X_{12}), X_2(X_{21},X_{22}) \ldots X_m(X_{m1},X_{m2})\} \tag{4}$$

$$\mathbf{Y}=\{Y_1(Y_{11}, X_{12}), Y_2(Y_{21},Y_{22}) \ldots Y_m(Y_{m1},Y_{m2})\} \tag{5}$$

We calculate the Manhattan distance of each corresponding point separately, so there are m coordinate differences. MH(X, Y) represents the Manhattan distance sum of two corresponding point sets.

$$MH(\boldsymbol{X},\boldsymbol{Y}) = \sum_{j=1}^{m} \left|X_{j1} - Y_{j1}\right| + \left|X_{j2} - Y_{j2}\right| \tag{6}$$

Since the x and y directions have been normalized by 0-1, the sum of the distances in the x or y directions is between 0-1 for any two points., Finally, as shown in formula 7, and further reduced to formula 8, by calculating the average value of all points, we can get a similarity of 0-1.

$$L(\boldsymbol{X},\boldsymbol{Y}) = 1 - \frac{1}{2m}\sum_{j=1}^{m} \left|X_{j1} - Y_{j1}\right| + \left|X_{j2} - Y_{j2}\right| \tag{7}$$

$$L(\boldsymbol{X},\boldsymbol{Y}) = 1 - \frac{1}{2m}\sum_{j=1}^{m}\sum_{i=1}^{2} \left|X_{ji} - Y_{ji}\right| \tag{8}$$

If the two lips videos are the same, the distance of the extreme points is 0, and the similarity of the lips is 1.

Actorid Ren-Xin Robot

Actorid Ren-Xin robot is a humanoid robot made by Kokoro Company based on Professor Ren. Its appearance, such as clothes and skin are very similar to real people. The servos of Ren-Xi are covered with silicone, which is very similar to human muscles. With some advanced control technologies, Ren-Xi robots can flexibly use facial muscles to make facial movements like human beings. However, the current Ren-Xin robot is not realistic enough in the speech process compared with human beings.

Ren-Xin robot has a total of 12 controllable parameters, 7 for face control, 3 for the head, and 2 for the upper body. The following is a detailed description of each parameter:

Eyebrows up and down:

Parameter range 0-255, 0 means the lowest position of eyebrows, 255 means the highest position.

Facial stretch:

Parameter range 0-255, 0 means that the face is pulled up to the lowest position, 255 means that the face is pulled up to the highest position.

Eyes closed and opening:

The parameter range 0-255, 0 means the closed position of the eye, 255 means the opening to the maximum position.

Eyes moving left and right:

Parameter range 0-255, 0 means the rightmost position of the eye, 255 means the leftmost position of the eye.

Eyes moving up and down:

Parameter range 0-255, 0 means the lowest position below the eye, 255 means the highest position above the eye.

Mouth closed and opening:

Parameter range 0-255, 0 means the mouth closed position, 255 means the maximum open position.

Mouth opening roundly:

Parameter range 0-255, 0 means the normal position of the mouth, 255 means the position of the circular pouting with the largest opening.

Head left tilt:

Parameter range 0-255, 0 means the normal state, 255 means head tilted to the leftmost position.

Head right tilt:

Parameter range 0-255, 0 means the normal state, 255 means head tilted to the rightmost position.

Head rotation:

Parameter range 0-255, 0 means the head is rotated to the rightmost angular position, 255 means the head is rotated to the leftmost angular position.

Shrugging:

Parameter range 0-255, 0 means no shrug, 255 means shrug to the highest position.

Lean forward and backward:

Parameter range 0-255, 0 means the body is tilted to the rearmost position, 255 means the body is tilted to the front-most position.

In this experiment, two parameters of the robot are mainly used, closing and opening of the mouth and the pouting parameters of the mouth. According to the following formula 9 and 10, the vowel size ratios extracted in 3.2 are mapped into the numerical values of robot control parameters.

$$RobotH = \frac{255}{max\,(\boldsymbol{H}) - min\,(\boldsymbol{H})} * (H_i - min\,(\boldsymbol{H})) \tag{9}$$

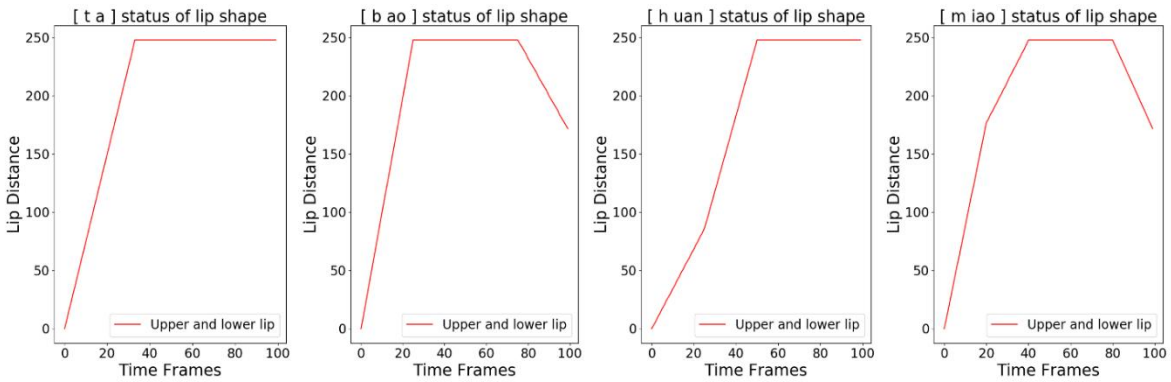
$$RobotW = \frac{255}{max\,(\boldsymbol{W}) - min\,(\boldsymbol{W})} * (max\,(\boldsymbol{W}) - W_i) \tag{10}$$

In the formula, H_i represents the up-down length scale of each vowel, $\boldsymbol{H}$ represents the up-down size scale set of vowels, W_i represents the left-right width scale of each vowel, and $\boldsymbol{W}$ represents the width scale set of vowels. Through calculation, we have obtained the maximum values of different vowel patterns of the adaptive robot as shown in Table 5.

Table 5. Mouth Shape Control of Vowel in Ren-Xin Robot

vowels	upper and lower length	upper and lower length
a	248	7
o	168	229
e	176	53
i	178	14
u(v)	87	250
Silence	0	0

In combination with the four types of vowel pronunciation mentioned in 3.1, we process each word into a control value of 0-255 for controlling the robot according to the duration of pronunciation. As shown in Figure 4, we have respectively listed four types of words: ta, bao, huan, miao, which are single sound, front sound, rear sound, and middle sound. The value in the X direction represents time. The value in the Y direction represents the size and position of the robot lip that need to be moved corresponding to different frames. In this way, the robot can read frame by frame, and each frame



moves its lip towards the corresponding position.

Figure 4. Single sound, front sound, rear sound, middle sound operation value examples for the robot.

4. EXPERIMENTS AND RESULTS

To verify the effect of our proposed scheme on humanoid robots, we conducted two different types of experiments. In the first experiment, we applied the proposed scheme to a robot Japanese teaching project and collected some subjective evaluations through a questionnaire survey. In the second experiment, we constructed 20 Chinese sentences and compared the pronunciation similarity between robots and real people.

Japanese Experiments and Subjective Evaluation

With the advancement of robot technology, robots have been increasingly applied to many fields. Among them, combining robots with education is also a prevalent direction (You Z J et al. 2006; Verner I M et al. 2016), which can combine classroom teaching with humanoid robots, achieve less teacher workload, and enhance students' learning enthusiasm.

Figure 5. Ren-Xin robot classroom teaching experiment.

In this experiment, we used a Ren-Xin robot and combined the proposed vowel priority lip matching scheme to conduct a 20-minute robot teaching experiment. The classroom situation is shown in Figure 5. For the lip matching subjective questionnaire evaluation, we have raised three questions for evaluation. The first is the subjective feeling of the pronunciation of a single word during the robot's pronunciation process. The second is how the subjective feeling of one sentence. Moreover, the third is how to feel about the overall pronunciation. The score is between 1-5, 5 means the best, and 1 means the worst.

Figure 6 shows the subjective evaluation of the questionnaire we collected. The X coordinate represents the corresponding score of 1-5 points, the Y coordinate gives the number of people corresponding to the score, and the three different colors indicate the different questions. In the subjective evaluation, a total of 20 students participated in the evaluation. As a result, we found that our scheme worked well in the pronunciation of single words, and was lacking in pronunciation of a sentence. The overall evaluation tended to be good.

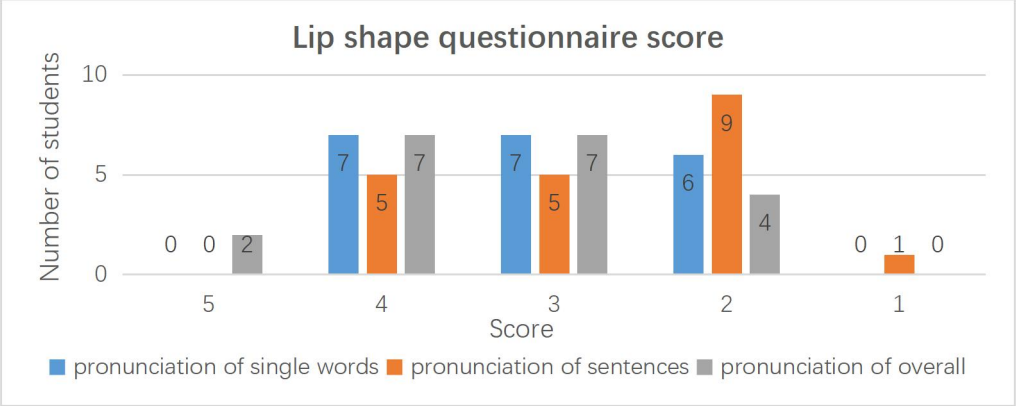


Figure 6. Questionnaire evaluation of humanoid robot teaching experiment.

Chinese Experiments and Model Evaluation

Different from the first experiment, to more objectively evaluate our proposed lip-matching scheme, we combined our new evaluation method and used the program to screen out 20 sentences. As shown in Table 6, the length of each sentence is 6 to 9 words, the type ratio of the single sound, front sound, rear sound and middle sound of Chinese finals is 1: 1: 1: 1.

Then we input 20 sentences into the robot control system one by one and record the pronunciation video of the robot's lip from the front of the robot's face. Similarly, as shown in Figure 7, in addition to the robot R, we selected two males and two females, M1, M2, W1, W2, and recorded 20 sentences of lip reading videos. Finally, the robot and the real person are compared one by one, and a total of 10 experimental comparison results are obtained.

Table 6. 20 Chinese sentences for testing

No.	Sentences	Sentences_Pinyin
1	我叫任心来自中国。	wo jiao ren xin lai zi zhong guo
2	欢迎来到我的科研室。	huan ying lai dao wo de ke yan shi
3	小白流下了泪水。	xiao bai liu xia le lei shui
4	老刘给买了两栋高楼。	lao liu gei mai le liang dong gao lou
5	我想给你讲个笑话。	wo xiang gei ni jiang ge xiao hua
6	九月里有好多台风。	jiu yue li you hao duo tai feng
7	就聊聊天好不好。	jiu liao liao tian hao bu hao
8	妹妹带着红玫瑰拍照。	mei mei dai zhe hong mei gui pai zhao
9	天冷多穿衣服。	tian leng duo chuan yi fu
10	明天去旅游吧。	ming tian qu lv you ba
11	想你给我一个机会。	xiang ni gei wo yi ge ji hui
12	小牛想回家睡觉。	xiao niu xiang hui jia shui jiao
13	最优秀的老师和学生。	zui you xiu de lao shi he xue sheng
14	小鸟在外面鸣叫。	xiao niao zai wai mian ming jiao
15	小鬼悄悄睡觉。	xiao gui qiao qiao shui jiao
16	外面的月亮又圆又亮。	wai mian de yue liang you yuan you liang
17	谁也不要讲话。	shui ye bu yao jiang hua
18	昨天老姐教我照相。	zuo tian lao jie jiao wo zhao xiang
19	老周每月都要吃烤肉。	lao zhou mei yue dou yao chi kao rou
20	宝贝爱好购物。	bao bei ai hao gou wu

Table 7 shows the results of 10 groups of comparisons. ID means the sequence number of the sentence, and other tags mean the comparison between different objects. The value is the percentage of similarity. From the table, we can see that for the same sentence, the similarity between different persons is different. To confirm the effectiveness of our method, for the robot, we calculated the average of the similarity between robot and four real persons in each sentence, and the most significant similarity in each sentence. For real persons, we counted the smallest similarity between real people. If the best results or average results of the robot are better than the worst results among real persons, then the proposed method is effective.

Figure 8 shows the statistical results. Blue MaxR means the best similarity between the robot and the real persons, orange AvgR means the average similarity between the robot and the real persons, and green MinH means the worst similarity between the real persons. Among them, the number of sentences with MaxR> MinH accounts for 80%; the number of sentences with AvgR> MinH accounts for 85%.

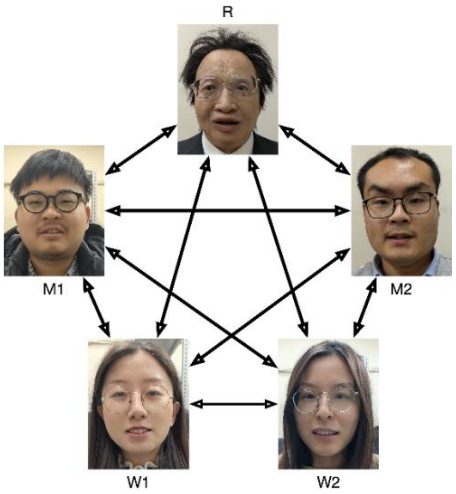


Figure 7. 10 groups of Chinese similarity measurement

Table 7. Robot and real human similarity results for a total of 10 groups

ID	R_M1	R_M2	R_W1	R_W2	M1_M2	M1_W1	M1_W2	M2_W1	M2_W2	W1_W2
1	92.24	87.6	90.48	91.8	85.82	92.87	94.13	84.25	85.85	90.52
2	90.96	89.34	89.83	86.64	90.32	93.53	89.14	90.14	89.98	87.44
3	91.98	90.71	90.6	90.78	92.85	92.27	93.26	92.86	92.45	92.01
4	88.46	90.68	88.76	88.57	93.36	93.01	95.6	91.56	92.76	95.42
5	92.1	91.83	93.42	86.42	94.32	93.25	89.09	95.23	91.43	90.47
6	88.75	91.29	86.27	88.72	90.81	84.6	93.75	90.45	90.36	84.48
7	82.07	87.95	91.03	87.95	88.83	87.46	85.43	90.38	92.55	91.21
8	87.92	88.93	90.17	86.8	92.14	93.39	87	91.55	86.11	92.3
9	91.95	94.06	93.95	84.2	92.56	90.89	87.55	92.36	86.5	86.23
10	86.92	89.16	88.85	86.08	90.16	87.84	94.8	92.48	87.91	85.07
11	90.55	89.2	89.46	87.71	89.64	91.29	88.11	90.94	90.23	89.73
12	90.55	92.56	86.87	88.94	91.5	90.1	91.59	87.93	90.54	89.28
13	90.72	91.09	86.9	85.41	91.3	90.55	89.08	87.78	87.77	89.83
14	91.73	90.72	91.63	89.89	95.33	95.17	88.95	94.36	90.13	90.13

15	91.99	93.55	90.74	92.37	94.11	92.78	90.08	92.98	90.6	89.52
16	85.19	86.92	88.01	84.79	92.59	92.83	88.27	93.89	91.11	88.81
17	92.14	92.97	93.03	91.31	92.9	92.71	92.27	91.58	91.54	90.39
18	91.35	91.35	91.18	87.43	93.99	91.92	92.04	89.85	88.76	91.86
19	88.23	89.22	90.27	87.08	93.48	90.95	89.06	94.04	92.99	89.67
20	86.32	88.42	87.72	82.85	91.27	91.29	90.05	93.24	85.81	84.52

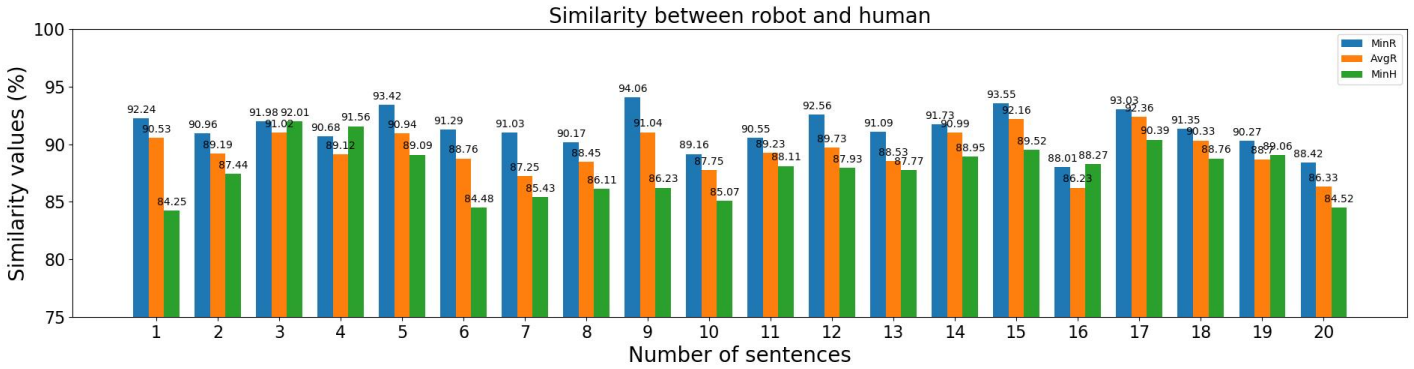


Figure 8. The similarity between robot and human

5. DISSCUSSION

In this paper, the lip shape matching scheme based on vowel priority that we proposed has played a massive role in the application of robots to imitate human lip. The proposed similarity evaluation model makes up for the gap of the humanoid robot lip shape evaluation. In Section 4, the experimental results show that our proposed lip shape matching scheme based on vowel priority, compared with the lip shape of real people, has achieved a high similarity result. In this section, we want to analyze further the performance of our algorithm, the advantages and disadvantages of the lip scheme and the lip similarity evaluation model.

Algorithm Speed

In this section, we focused on testing the performance of our algorithm. Our system environment is Windows 10 Enterprise 64-bit, CPU: Intel (R) Core (TM) i7-6700 3.40GHz, RAM: 16.0GB, Python version: 3.6.3. For Chinese and Japanese, we use Pypinyin and Pykakasi python packages, which will be used for the conversion of Roman characters.

We use the program to read 1,000 Chinese sentences and 1,000 Japanese sentences from the file into the computer's memory. After that, we set the start test time, and then convert each sentence into Roman characters and format it, then generate control codes for robots, and record the end time after all generations are completed.

Table 8. Running speeds of algorithm

Languages	Running time of 1 thousand sentences	Average time of 1 sentence
Chinese	4353 ms	4.3 ms
Japanese	1179 ms	1.2 ms

Table 8 shows the running time of the test procedure. For the generation of the robot control value of 1,000 sentences, Chinese and Japanese consumed 4353ms and 1179ms respectively, and the generation of one sentence also achieved the millisecond level of 4.3ms and 1.2ms. The time consumption can effectively operate the robot in practical applications.

Lip Matching Scheme Analysis

Our method is based on vowels. Vowels are indispensable in all languages, so our scheme is very versatile. Our method defines a set of template rules for the interaction of vowels, so when generating robot lip opcodes, the speed is breakneck. Also, after our two experiments, the results show that our algorithm results are also close to real people.

The shortcomings of the algorithm are limited by the reaction speed of the robot's lip. As shown in Figure 9, the solid line represents the curve of the opcode generated by the algorithm, and the dashed line represents the curve of the actual movement of the lip. When the first word transitions to the second word, for each control frame, the robot's lip will move gently to the position of the control code. However, due to the limitation of the hardware conditions of the robot control system, if the speech speed is too fast when speaking, especially during the lip-shaped switching of b, p, m consonants between words, the robot lip cannot move as fast as the preset control code to the designated location.

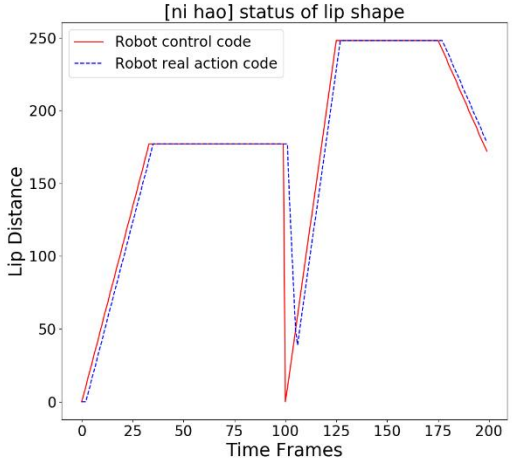


Figure 9. Example of robot control code and real action code

Evaluation Model Analysis

Our evaluation model mainly uses the technology of face key point detection in the field of computer vision. After extracting the position coordinates of all feature points in each picture frame, we will record the lip shape change curve. For each word, the lip will move from the normal position to the specified maximum vowel position. So in order to facilitate calculation without losing essential information, we extracted the extreme points of the change curve. The coordinate positions of the extreme points represent the time of words in the sentence and the maximum position of the lip shape, which effectively retains the information in the process of lip shape change. However, in the process of extracting feature points, redundant feature points will appear, and it is necessary to remove them and then compare them manually.

6. CONCLUSION AND FUTURE WORK

To make the robot's lip behave more like real people, we proposed a lip matching scheme for Chinese and Japanese based on vowel priority. Unlike traditional research about the viseme, we combined the pronunciation features of the language and increased the priority of vowels during the pronunciation process. Besides, we also used computer vision lip feature points and created a Manhattan distance-based evaluation model for measuring the similarity of lip shape.

We conducted two types of experiments. In the questionnaire survey of the robot classroom teaching experiment, a total of 20 students rated the pronunciation of a single word, the pronunciation of the entire sentence, and the overall pronunciation, and finally achieved a good score. In another experiment, we make the robot compare with real people by using 20 sentences and 80% ~ 85% of the sentences, the robot’s lip is close to the real people.

We propose a simple and effective method with the characteristics of fast speed, strong versatility and high similarity. It is conducive to the humanoid robot's lip shape closer to a real person. Our method can achieve an effect similar to a real person for a sentence. However, it cannot express sentences with rich emotions. For the future, we will incorporate some emotional factors and combine the features of speech to make the humanoid robot's lip shape behave more like real humans.

7. ACKNOWLEDGMENT

This research has been partially supported by JSPS KAKENHI Grant Number 19K20345.

REFERENCES

Lu H, Li Y, Chen M et al (2018) Brain intelligence: go beyond artificial intelligence. Mob Netw Appl 23(2):368-375. https://doi.org/10.1007/s11036-017-0932-8

Ren F, Bao Y (2020) A Review on Human-Computer Interaction and Intelligent Robots. Int J Inf Technol Decis Mak 19 (01):5-47. https://doi.org/10.1142/S0219622019300052

Kuindersma S, Deits R, Fallon M et al (2016) Optimization-based locomotion planning, estimation, and control design fo r the atlas humanoid robot. Auton Robots 40(3):429-455. https://doi.org/10.1007/s10514-015-9479-3

Ishi C T, Machiyashiki D, Mikata R et al (2018) A speech-driven hand gesture generation method and evaluation in andr oid robots. IEEE Robot Autom Lett 3(4):3757-3764. https://doi.org/10.1109/LRA.2018.2856281

Ren F, Huang Z (2016) Automatic facial expression learning method based on humanoid robot XIN-REN. IEEE Trans H um Mach Syst 46(6):810-821. https://doi.org/10.1109/THMS.2016.2599495

Hwang J, Tani J (2017) Seamless integration and coordination of cognitive skills in humanoid robots: A deep learning ap proach. IEEE Trans Cogn Dev Syst 10(2):345-358. https://doi.org/10.1109/TCDS.2017.2714170

Herath D C, Jochum E, Vlachos E (2017) An experimental study of embodied interaction and human perception of social presence for interactive robots in public settings. IEEE Trans Cogn Dev Syst 10(4):1096-1105. https://doi.org/10.1 109/TCDS.2017.2787196

Ren F (2009) Affective information processing and recognizing human emotion. Electron Notes Theor Comput Sci 225: 39-50. https://doi.org/10.1016/j.entcs.2008.12.065

Fu K, Sun L, Kang X et al (2019) Text Detection for Natural Scene based on MobileNet V2 and U-Net. In: 2019 IEEE In ternational Conference on Mechatronics and Automation (ICMA), pp 1560-1564. https://doi.org/10.1109/ICMA.20 19.8816384

Liu Z, Ren F, Kang X (2019) Research on the Effect of Different Speech Segment Lengths on Speech Emotion Recognit ion Based on LSTM. In: Proceedings of 2019 the 9th International Workshop on Computer Science and Engineerin g, pp 491-499. https://doi.org/10.18178/wcse.2019.06.073

Ren F, Kang X, Quan C (2015) Examining accumulated emotional traits in suicide blogs with an emotion topic model. I EEE J Biomed Health Inform 20(5):1384-1396. https://doi.org/10.1109/JBHI.2015.2459683

Ren F, Matsumoto K (2015) Semi-automatic creation of youth slang corpus and its application to affective computing. I EEE Trans Affect Comput 7(2):176-189. https://doi.org/10.1109/TAFFC.2015.2457915

Nishikawa K, Takanobu H, Mochida T et al (2004) Speech production of an advanced talking robot based on human aco ustic theory. In: 2004 IEEE International Conference on Robotics and Automation(ICRA), pp 3213-3219. https://do i.org/10.1109/ROBOT.2004.1308749

Luo R C, Chang S R, Huang C C et al (2011) Human robot interactions using speech synthesis and recognition with lip s ynchronization. In: 2011 IECON 2011-37th Annual Conference of the IEEE Industrial Electronics Society, pp 171-176. https://doi.org/10.1109/IECON.2011.6119307

Oh K G, Jung C Y, Lee Y G et al (2010) Real-time lip synchronization between text-to-speech (TTS) system and robot mouth. In: 19th International Symposium in Robot and Human Interactive Communication, pp 620-625. https://doi. org/10.1109/ROMAN.2010.5598656

Hara F, Endou K, Shirata S (1997) Lip-Configuration control of a Mouth robot for japanese vowels. In: Proceedings 6th IEEE International Workshop on Robot and Human Communication, pp 412-418. https://doi.org/10.1109/ROMAN.1997.647022
Herath D C, Jochum E, Vlachos E (2017) An experimental study of embodied interaction and human perception of social presence for interactive robots in public settings. IEEE Trans Cogn Dev Syst 10(4):1096-1105. https://doi.org/10.1109/TCDS.2017.2787196
Yan J (1998) Research on the viseme of chinese phonetics. Comput Eng Des 19(1):31−34. (in Chinese)
Zeng H, Hu D, Hu Z (2013) Simple analyzing on matching mechanism between Chinese speech and mouth shape. Audio Eng 10:44-48. (in Chinese)
Fan X, Yang X (2017) A speech-driven lip synchronization method. J Donghua Univ (Nat Sci) 4:2. (in Chinese)
Miyazaki T, Nakashima T (2015) Analysis of Mouth Shape Deformation Rate for Generation of Japanese Utterance Images Automatically. In: Software Engineering Research, Management and Applications, pp 75-86. https://doi.org/10.1007/978-3-319-11265-7_6
Morishima S, Harashima H (1991) A media conversion from speech to facial image for intelligent man-machine interface. IEEE Journal on selected areas in communications 9(4):594-600. https://doi.org/10.1109/49.81953
Kim T H (2008) A Study on Korean Lip-Sync for Animation Characters-Based on Lip-Sync Technique in English-Speaking Animations. Cartoon Animat Stud 13:97-114. (in Korean)
Saitoh T, Konishi R (2010) Profile lip reading for vowel and word recognition. In: 2010 20th International Conference on Pattern Recognition, pp 1356-1359. https://doi.org/10.1109/ICPR.2010.335
Sulistijono I A, Baiqunni H H, Darojah Z et al (2014) Vowel recognition system of Lipsynchrobot in lips gesture using neural network. In: 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp 1751-1756. https://doi.org/10.1109/FUZZ-IEEE.2014.6891843
Li P, Wang D, Wang L et al (2018) Deep visual tracking: Review and experimental comparison. Pattern Recognit 76:323-338. https://doi.org/10.1016/j.patcog.2017.11.007
Cootes T, Baldock E R, Graham J (2000) An introduction to active shape models. Image processing and analysis 243657: 223-248.
Cootes T, Edwards G J, Taylor C J (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 23(6):681-685. https://doi.org/10.1109/34.927467
Sun Y, Wang X, Tang X (2013) Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3476-3483. https://doi.org/10.1109/CVPR.2013.446
Li X, Wang T (2018) A long time tracking with BIN-NST and DRN. J Ambient Intell Humaniz Comput 1-7. https://doi.org/10.1007/s12652-018-1025-7
Long T (2019) Research on application of athlete gesture tracking algorithms based on deep learning. J Ambient Intell Humaniz Comput 1-9. https://doi.org/10.1007/s12652-019-01575-w
Dai K, Zhang Y, Wang D et al (2020) High-Performance Long-Term Tracking with Meta-Updater. arXiv preprint arXiv:2004.00305
Binyong Y, Felley M (1990) Chinese romanization: Pronunciation & orthography. Peking
Keating P A, Huffman M K (1984) Vowel variation in Japanese. Phonetica 41(4):191-207. https://doi.org/10.1159/000261726
Hyung H J, Ahn B K, Choi D et al (2016) Evaluation of a Korean Lip-sync system for an android robot. In: 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), pp 78-82. https://doi.org/10.1109/URAI.2016.7734025
You Z J, Shen C Y, Chang C W et al (2006) A robot as a teaching assistant in an English class. In: Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06), pp 87-91. https://doi.org/10.1109/ICALT.2006.1652373
Verner I M, Polishuk A, Krayner N (2016) Science class with RoboThespian: using a robot teacher to make science fun and engage students. IEEE Robot Autom Mag 23(2):74-80. https://doi.org/10.1109/MRA.2016.2515018