An Evaluation of the Value of Lip and Jaw Motion in Virtual Reality Avatars

Ubiquitous Computing (Ubicomp) Laboratory
School of Interactive Computing

Ruixuan Sun
Spring 2019

Faculty Member 1

_____

Signed

Faculty Member 2

_____

Signed

Acknowledgement

Thanks to our mentors, Richard Li and Gregory Abowd
Who give us guidance and encouragement during our Undergraduate Research
Thanks to our family and dearest friends
Who always give us love and support in our life

Contents

1. Abstract
2. Introduction
3. Related Work
    3.1 Facial Sensing in VR
        3.1.1 Audio-Generated Lip Motion
        3.1.2 Facial Landmark Tracking
    3.2 Animating Avatars in VR
    3.3 Evaluating Communication in VR
4. User Study Design
    4.1 Purpose
    4.2 Condition
        4.2.1 Landmark Tracking VR
        4.2.2 Audio Driven VR
    4.3 Study One: Language Learning
    4.4 Study Two: Casual Conversation
5. Results
    5.1 Study One Results
    5.2 Study Two Results
6. Analysis
    6.1 Study One Analysis
        6.1.1 Inter-rater Agreement
        6.1.2 Significance Testing
        6.1.3 Subjective Results Analysis
    6.2 Study Two Analysis
7. Discussion
8. Conclusion & Future Work
9. References
10. Appendix

1. ABSTRACT

Virtual reality (VR) is known as a simulated 3D immersive technology. Currently, the emphasis has been mainly placed on tracking the upper face, like eye tracking and eyebrow imitation, where the lower face containing the largest emotions of the face are usually hard for commercial VR headset to capture due to hardware limit. In this study, we explore the role of lip and jaw motions when VR is used as a communication medium by comparing the effectiveness of camera-based facial landmark tracking against audio-driven lip movements.

2. INTRODUCTION

Virtual reality (VR) has emerged as an interactive computer-generated experience taking place within a simulated environment, incorporating auditory and three-dimensional (3D) visual feedback among other types of sensations, including haptics. As a platform, VR has the potential to enable many applications, with additional immersiveness as a benefit over 2D interfaces. One such application might be VR video calling, in which users can see and interact with a 3D representation of their call partner in a virtual environment. The general steps to realize a VR video call can be divided into two phases: facial tracking and 3D avatar rendering. For facial tracking, much of the related work has focused on areas accessible from the headset, such as the eyes, and there already exists many robust models for the upper-half of the face (Parris, J., et al. 2011). In terms of rendering 3D avatars, some example works have also been done, like the re-morphing of the human characters with real-time data stream input, which was achieved with a dynamic 3D model by Feng and his team in 2015. Another example is the integration of the 3D character with a mobile application that could be deployed on a VR device, which was also made possible with the SmartBody SDK (Marsella, S., et al. 2013).

However, two gaps in VR communication are usually neglected. The first is the underdevelopment of the lower-face motion, majorly including the lip and jaw movements, that could also contribute a lot for facial expression and information delivering. Another is the lack of using a relatively low-cost VR for daily human-to-human interaction, like having casual chat. Hence, we decide to fill the research gap by assessing the extent to which the precision of lip and jaw movements can enhance the quality of VR communication. To address the underdevelopment of lip-jaw centered lower face recognition in a real-time model, our research would focus on lip tracking with the improved model modified from the dlib machine learning library (King 2009). The VR rendering step would make use of SmartBody SDK as well as a redesigned Google cardboard prototype, which is only of 1/20 price of a normal Oculus or Vive device and can make the VR technology more accessible to everyone. For enhancing the user experience and network transferring

stabilization during our experimental studies, we utilized Faceware live server and client (Faceware Tech. 2019) implemented in VR mode in Unity. Finally, we compared our results with those from traditional audio-generated Lipsync technology that is currently widely used for virtual avatars in the game and filming industry. The whole experiments would be divided into two phases, including a specific language learning task as well as a general daily chat evaluation.

3. RELATED WORK

3.1 Facial Sensing in VR

3.1.1 Audio-Generated Lip Motion

Generating lip motion and corresponding facial animations can be based on audio input (Karras, et al 2017). Most VR headsets do not have cameras to capture lower face movements. On the other hand, while mobile VR systems like Google Cardboard that there are smartphones inside do have an extra back camera, it is pointed in the wrong direction. Thus, utilizing the speech signal to generate the corresponding lip and jaw movements could improve the quality of audio-visual communications by leveraging the more pervasive and non-directional microphone. This general technology is called lip synchronization (Lipsync), based on the observation that the shape of the mouth over a short interval of time can be correlated with the basic shape of the spectrum of the speech over that same interval. The spectrum is obtained from a Fast Fourier Transform (FFT) and treated like a discrete probability density function. One kind of statistical measurement called moments, the specific quantitative measure of the shape of a function, are used to describe the shape of the FFT (McAllister, David F., 1997). The Lipsync technique has already been adopted in commercial games such as the VR social platform VRChat (VRChatNet, 2019).

3.1.2 Facial Landmark Tracking

Traditionally, facial expression tracking and recognition generally involves tracking 68 well-defined landmarks to locate and reflect changes (King 2009). The computer needs several very significant points on one object in order to recognize its class. For the human face, it regards the eyes, nose, mouth as well as the corresponding distance between each of them to identify the object as a human face (Parris, et al 2011). For example, a commercial product, Faceware Live Server (Faceware Tech. 2019), utilizes such a methodology to detect the human face as a whole and labels each part of the whole object as different numbered facial landmark features. Beyond simple landmark recognition, several studies have explored eye tracking and other upper face modeling with their focus mainly on testing and improving different algorithms (Parri, et al 2011), while others in more recent times have proposed a more flexible model-free analysis of 3D facial expression recognition which does not depend on feature extraction (Savran and Sankur 2017).

More specific research on human facial features, like lip recognition and contour extraction, have also studied for a long time. Chen, Tiddeman, and Zhao (2008) presented a newly optimized lip contouring algorithm which can accelerate the machine learning process by addressing usually neglected areas on lip images. Particularly, they add an image gradient term to detect image edge at low contrast areas, thus allowing a more accurate and detailed contouring output.

3.2 Animating Avatars in VR

One of the primary applications for developing such sensing technologies is to eventually animate virtual avatars in a convincing manner. For example, the integration of the 3D character with a mobile application that could be deployed on VR device is made possible with the SmartBody SDK (Marsella, S., et al. 2013). Unity3D is also known to be a useful tool to convert 3D animation to VR in real time. Unity3D has a relatively mature VR plugin that can generate the VR environment through a 3D game scene (Unity 2019), and package the scene to an iOS or Android mobile application which can be run using a Google Cardboard. Therefore, Unity3D would be a convenient tool for us to animate the lip and jaw movements of our avatars.

3.3 Evaluating Communication in VR

We hypothesize that lip and jaw motion can play a significant role in VR communication. To evaluate aspects of VR communication, we reviewed methods that previous researchers have used to study VR/ mixed-reality (MR) in the field of human-computer interaction. For example, to measure how children could learn from MR games (Yannier, 2015), researchers designed a 2x2 experiment in which children played a game in two different controls using two different types of displays. They recorded objective measurements such as the percentage of correctly completed tasks, as well as the subjective measurement of which version of the game children liked the most. Since we hypothesize that lip and jaw motions are significant in information delivering during VR communication, we adapt the idea from this study and designed a language learning task for evaluating the importance of lip and jaw motion.

In addition to the learning experience study, we also reviewed Garau's study (Garau, Maia, et al, 2001) that evaluated the importance of eye gaze in avatars representing people engaged in conversation. According to Garau, a categorized post-experiment questionnaire can provide a quantitative measurement for the quality of communication. The article shows the significance of eye gaze of avatar during conversation. Since we hypothesize that lip and jaw motions are crucial contributors to a high quality conversation, we adapted this experiment for the context of lip and jaw movement.

In general, to evaluate the VR communication system we created, we decided to evaluate our Faceware and Lipsync VR modes in two perspectives: the effectiveness of a purposeful task, and the quality of a casual conversation. Thus, we designed two experimental phases, with the first one including a specific language learning task, and the second one exploring the general quality of communication. Each of the studies would focus on the comparison of our two VR systems, with both objective and subjective measurements.

4. USER STUDY DESIGN

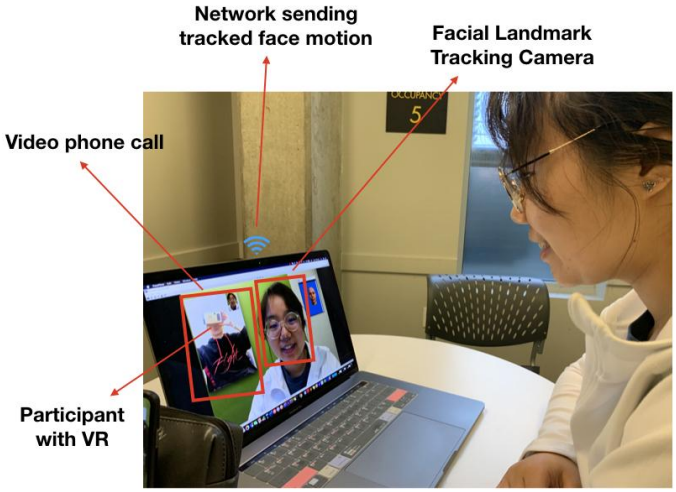
4.1 Purpose

We hypothesize that lip and jaw motions are critical to communication. To evaluate this hypothesis in the context of VR, we conduct two studies to compare the Faceware and Lipsync systems as communication methods.

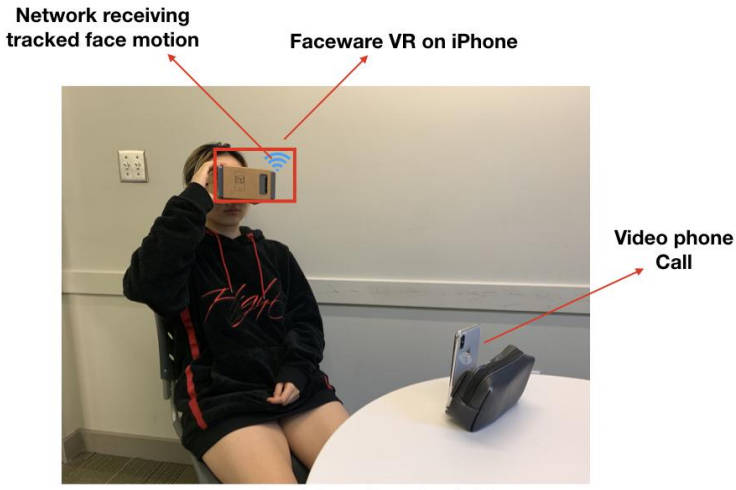
4.2 Conditions

4.2.1 Landmark Tracking VR

For the facial landmark tracking system, we used the library of Faceware Tech (Faceware). Faceware is a commercial product that enables using the webcam on a laptop for tracking facial features and movements. The information is then sent to a client device. In our case, the laptop's webcam tracks the facial landmarks of the speaker's face and uses this information to manipulate a 3D avatar in the headset of the speaker's partner. To build the VR application, we imported the client side into Unity3D, and built an iOS mobile application to display the avatar in VR. Since we used Faceware Tech software for the landmark tracking, we will refer to this system as Faceware for the rest of the paper.



Faceware - Researcher(Input) side

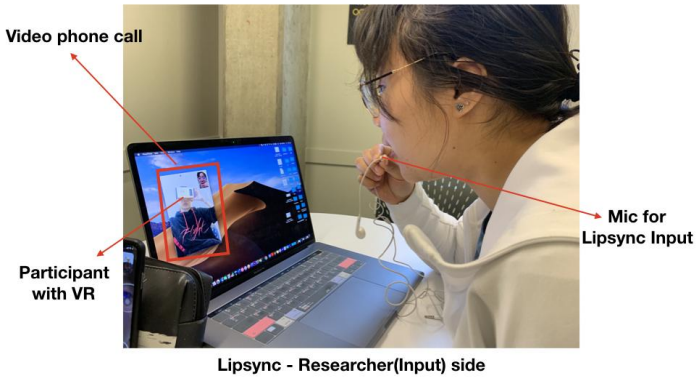


Faceware - Participant(Receiver) side

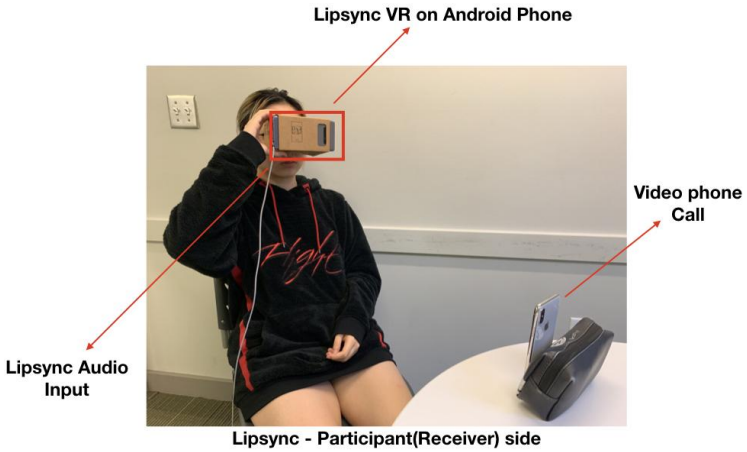
4.2.2 Audio Driven VR

Since audio driven VR is commonly used in currently existing applications, like the VR

Chat room, we used the Oculus lip sync API that can be called from Unity3D asset store as a baseline for comparison and used it to animate the same avatar used in Faceware. Then we built an Android mobile application that would take in a real time audio stream and translate the audio into mouth motion. The application then displayed the avatar with moving mouth in VR. Since we used Oculus lip sync plugin for Unity, we will refer to this system as Lipsync for rest of the paper.



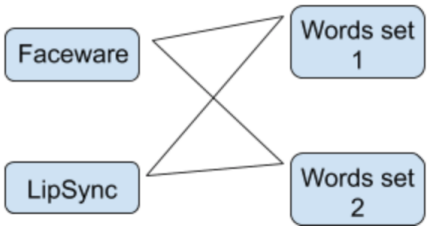
Lipsync - Researcher(Input) side



Lipsync - Participant(Receiver) side

4.3 Study One: Language Learning

In order to assess if Faceware helps the effectiveness of completing a task more than Lipsync for the same user, we designed a language learning task in which participants would benefit from having visual feedback, since seeing lip movement during training can significantly help second language learners (Hirata 2010). To ensure similar familiarity level with the second language for all participants, we chose participants who self-reported as having never studied Chinese before.



2x2 experiment

With 8 native English speaking college students (5 male, 3 female; aged between 20 to 23) who have never studied Chinese before, we compared two lip motion generating techniques in the context of learning Chinese. It is a 2x2 experiment where we consider two conditions for our study: Faceware and LipSync, and two different sets of monosyllabic words. The reason for choosing two sets of words is to mitigate the learning effect. Since we need to switch whether participants use Faceware or LipSync first, we do not want our learner to study the same set of words that they have learned in the previous VR mode. To counterbalance the bias of the order of words and different VR modes for each

participant, we utilized two Latin squares, and multiplied them to decide which combinations of VR modes and words sets each participant would use. Hence, we could largely balance the learning effect of exposing to a new language and avoid potentially biased familiarity.

Person/Order	1	2
Person 1	Faceware	Lipsync
Person 2	Lipsync	Faceware

2x2 Latin squares for VR modes

Person/Order	1	2
Person 1	Word Set 1	Word Set 2
Person 2	Word Set 2	Word Set 1

2x2 Latin squares for word sets

Conditions	1st	2nd
Person 1	Faceware x Animals	Lipsync x Colors
Person 2	Lipsync x Colors	Faceware x Animals
Person 3	Faceware x Colors	Lipsync x Animals
Person 4	Lipsync x Animals	Faceware x Colros

The Experimental Latin Square combined with two 2x2 ones above

During the learning session, all of the participants' pronunciation attempts were recorded. The pronunciation of each single word was then clipped out. After the learning session, two Chinese native speakers with the same accent and geographical backgrounds raters listened to the word clips in randomized order and gave binary rating scores for "accurate" or "inaccurate" based on a set of criteria attached in the appendix.

To measure the effectiveness of Faceware and Lipsync in language learning, we evaluated participants using two metrics. The objective measurement is the native speaker's binary rating scores for each participant's pronunciation based on three language components that will be used for further statistical analysis, while the subjective measurement is based on users' own preferences and their experience about overall immersiveness for the two VR systems.

Objective Measurement	Subjective Measurement
Native Speaker's binary rating scores for each participant's pronunciation based on three language components	Users' own preference for two VR modes and overall immersivene ss

Objective and Subjective Measurements

4.4 Study Two: Casual Conversation

The second phase of user study involved a more general chatting task. Instead of a

language learning task between researchers and participants, a free talk between four pair of users was conducted, including one pair of female-female, one pair of male-male, and two pairs of female-male participants. To encourage natural and organic conversation, none of the researchers were in the rooms with the participants, allowing them to freely converse about anything.

Participants were recruited in pairs, and they were separated into two different study rooms. One participant has his or her face tracked by Faceware camera or voice captured by Lipsync mic, while the other participant wore a VR headset. The order of Faceware and Lipsync was determined through a Latin square similar to the first study, and which participant wore the VR was randomly assigned.

Participants first filled out a pre-survey about their perceived closeness between with their partner, and if they had ever experienced VR before. Then a ten-minute free chat without a predetermined topic was conducted between them for each VR mode (either Faceware or Lipsync), followed by filling out a 15-question questionnaire by each participant after finishing each VR chat. One post survey question about VR mode preferences would be asked at the end of two chats.

The fundamental variable of interest was the quality of users' communication. Adapted from the evaluation of eye gaze experiment (Garau, Maia, et al, 2001), we can evaluate our participants on four broad indicators:

1. *Face-to-face*: The extent to which the conversation was experienced as being like a real face-to-face conversation.

2. *Involvement*: The extent to which the participants experienced involvement in the conversation.

3. *Co-presence*: The extent of co-presence between the participants - that is, the sense of being with and interacting with another person rather than with a computer interface

4. *Partner Evaluation*: The extent to which the conversational subjects positively evaluated their partner, and the extent to which the conversation was enjoyed.

The responses to these variables were elicited by means of the post-experiment questionnaire after each VR mode chat, with each response being on a 9-point Likert scale, where 1 was anchored to "strongly disagree" and 9 to "strongly agree". The questions were grouped in the second part of the appendix.

5. RESULTS

5.1 Study One Results

Each of the 8 participants listened and repeated each of the 20 words at least three times until they got the correct pronunciation. Hence, we had 488 words clips. Then, the two native speaker raters gave each of these clips an "accurate" or "inaccurate", totaling 976 scores.

Besides the numerical data, we also collected comments from them on the effectiveness and immersiveness of Faceware and Lipsync, and their personal preferences between Faceware and Lipsync.

5.2 Study Two Results

We had eight participants in four pairs, where in each pair one participant was tracked by Faceware or Lipsync, and the other one was watching an animated avatar through a VR headset. All of the participants were given the same survey with 15 questions, which were categorized as Face-to-Face (five questions), Involvement (three questions), Co-presence (two questions), and Partner Evaluation (five questions). Participants would rate these questions from 1 to 9. Overall, we obtained 15 (#questions) x 8 (#participants) x 2 (Faceware or Lipsync) = 240 responses. Then we calculate the average and standard deviation among four survey categories in Faceware and Lipsync.

6. ANALYSIS

6.1 Study One Analysis

6.1.1 Inter-rater Agreement

As outcomes, we first calculate the inter-rater reliability (Spitzer RL, Gibbon M, Williams JBW, 1994) for all the rating scores to make sure those two native speaker raters agree with each other.

$$k = \frac{p_{11} + p_{00} - (p_{1+}p_{+1} + p_{0+}p_{+0})}{1 - (p_{1+}p_{+1} + p_{0+}p_{+0})}.$$

Kappa coefficient (K)

Calculated with the formula above, the overall kappa value for agreement is 0.73. According to the standard by Landis and Koch (1977), the overall agreement of 0.73 indicates a substantial agreement of all participants. Therefore, since these two native speaker rater agree with each other on their ratings, we can compare the ratings further for analyzing their statistical results.

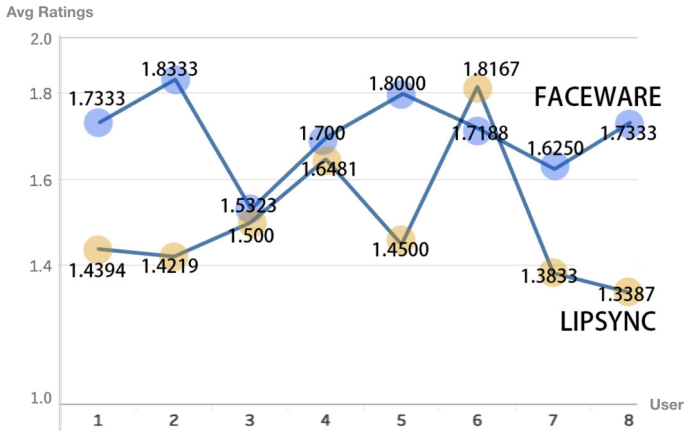
6.1.2 Significance Testing

After that, we calculated the overall average score and standard deviation for Faceware and Lipsync, considering the binary ratings of either 1 ("unacceptable") or 2 ("acceptable"), Faceware scored about 20% higher on average than Lipsync, as well as about 5% lower standard deviation.

Avg./Std. Faceware	Avg./Std. Lipsync
1.7082 ± 0.45507	1.4959 ± 0.50050

Overall Average ± Standard deviation for Faceware and Lipsync

We also calculated the average rating score for each participant, as shown below, 7 out of 8 participants performed better on average using Faceware, among which 5 participants have over 0.2 difference considering the range from 1 to 2. Combined with the overall statistical results above, we can infer that Faceware's greater accuracy is more helpful for pronunciation learning.



Individual Average for Faceware and Lipsync

McNemar's significance test was used for comparing the binary ratings within a case-control study, with continuity correction applied. We found the two-tailed P value when comparing the scores from Faceware versus LipSync is less than 0.0001, which could indicate extremely statistical significance between the difference of Faceware and Lipsync and reject the null hypothesis. The Chi Square for this dataset equals to 41.123 with 1 degree of freedom. Hence, based on McNemar's definition, there is an association between the risk factor and the disease, meaning that our data are of statistical significance.

6.1.3 Subjective Results Analysis

When asked which mode the participant prefer, 6 out of 8 participants said they liked the Faceware system more. Upon further prompting, they suggested Faceware's detailed lip and jaw movements seemed more realistic. Comparing to Lipsync, Faceware can present more accurate motion of lip and jaw.

P1: "I definitely like Faceware more. It has more details on the lower face, and some of them are much exaggerated than Lipsync motion. Also it is more accurate, while Lipsync only display a set of general motions."

P3: "I prefer faceware. The quality of phone calling was actually not quite good. Since I cannot hear clearly, see how the mouth move exactly is very helpful."

P5: "Definitely the first one(Faceware). Well it is just more detailed and more vivid. Lipsync is….sometimes you can not really tell if there is a big difference between two different pronunciation."

There was also, however, one participant not comfortable with the visual distraction of Faceware:

P7: "Well honestly I prefer Lipsync….you know it is less distracting than the Faceware."

When talking about the major benefits of VR, participants mentioned immersiveness and compared these two systems that which gave them better experience as in a real learning experience.

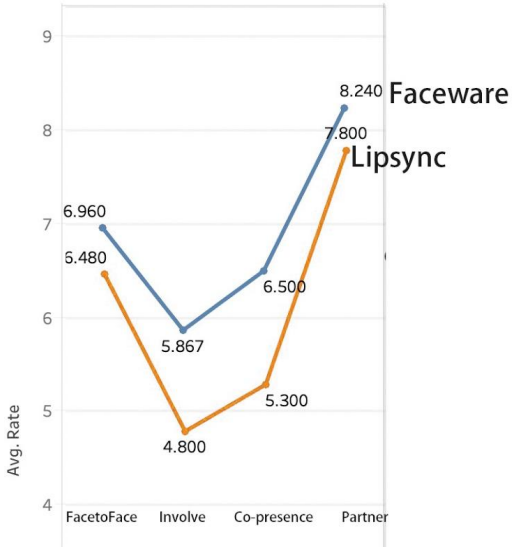
P2: "I would prefer faceware, you know it is really like talking to you face-to-face except sometimes the laggy frame….that is understandable yeah. l love the immersiveness!"

With talking to the avatar through Faceware, some of the participants did not regard him as an AI, which shows the quality of a VR conversation.

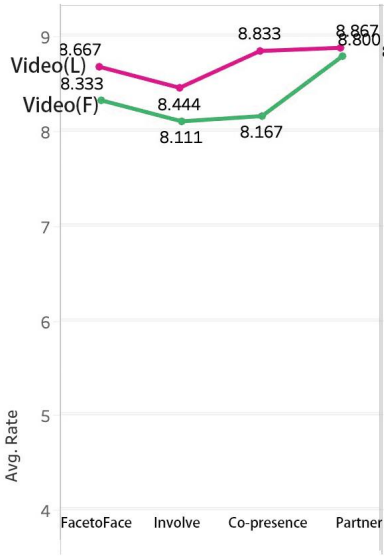
P7: "...But yeah I have to admit Lipsync is more like an AI, it is not real! And Faceware is more immersive, like a real virtual teacher!"

P8: "And Faceware is also good for learning language as a student. I would definitely like to have a virtual teacher like Victor(our avatar's name)!"

6.2 Study Two Analysis



Mean Raw Questionnaire Response for Users wearing two different VR sets



Mean Raw Questionnaire Response for Users having video calls without VR sets

As the results for general chat pairs of shown, Faceware (blue line) is always higher than Lipsync (orange line), but

interestingly, all four observers who did not try VR set would instead give a video recording of Lipsync a higher average score than video Faceware.

A similar trend happened to the mean and standard deviation data of all four different rating averages again, with Faceware having higher average and lower standard deviation than the Lipsync score, but Video Faceware would have lower average and higher standard deviation than the corresponding Video Lipsync mode.

	Face ware	Lipsy nc	Video (F)	Video (L)
Face2 Face N = 5	6.96 ± 2.01	6.48 ± 2.33	8.33 ± 1.05	8.67 ± 0.62
Invol ve N = 3	5.87 ± 1.51	4.80 ± 2.46	8.11 ± 1.27	8.44 ± 0.73
Co-Prese nce N = 2	6.50 ± 1.90	5.30 ± 2.54	8.17 ± 0.98	8.83 ± 0.41
Partn er N = 5	8.24 ± 1.74	7.80 ± 1.96	8.80 ± 0.41	8.87 ± 0.35

Mean ± standard deviations of Count Response Variables.
N = number of questions on which the count is based.

Finally, we calculated the paired t test for Faceware and Lipsync. With a total of 76 pairs of participant responses, the result for the ratings that VR experiencing users give are of statistical significance, with two-tailed p value equals 0.0063 and t value equals 2.8130. Since p value is less than alpha=0.05, the difference is statistically significant.

7. DISCUSSION

Our hypothesis is to evaluate if lip and jaw motion are critical in communication. The language learning task is designed for a purpose-task user scenario. Our expectation for the study is that Faceware would outperform Lipsync, since in this specific task, the exact lip and jaw movement is significant important for a language learner. When dealing with new language pronunciation, the more details one could receive from the lip and jaw motion, expectedly more accurate the learner could speak. Therefore, as Faceware could landmark one's face and track it exactly, it was shown to be a more helpful language learning tool comparing to Lipsync. In the result, both objective result and subjective result match our expectation, that show the fact that users read and learn better through Faceware than through Lipsync, and they also prefer Faceware rather than Lipsync.

Our expectation for the second study was to evaluate the overall impact of Faceware and Lipsync on conversation, with a free

environment designed for two users to have casual talk. Based on our hypothesis, Faceware should capture the lip and jaw motion more accurately and provide participants a more immersive and informative ways to deliver and understand information during their communication. The statistical results from the VR wearers indeed matched our expectation, but the video observers seemed to find the Lipsync to be somehow more helpful when communicating with the VR experiencers. Our interpretation of this interesting fact is the multiple focus requirement for Video observers while using Faceware. Since for Lipsync, the Video observers only need to hold the mic and focus on the video call screen to communicate with their partner. However, when using Faceware, they have to focus on the Faceware camera that capture and update their facial landmark in real time, the animation avatars on the screen that mimic their emotions, as well as the video call screen for their partners' reaction, which combined together could be more distracting than a single focus requirement during a Lipsync talk. So a more robust and user-friendly algorithms can make it easier for people to use lip tracking technologies.

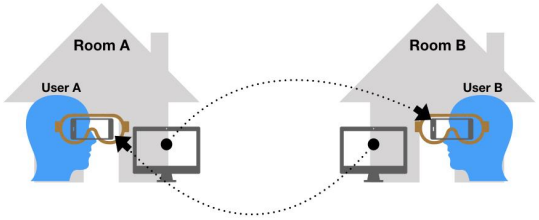
However, though both our studies with participants wearing the VR set show Faceware performs better than Lipsync, we still do not know whether Faceware is better than Lipsync in all the user scenarios. For example, during the language learning user study, we received feedback complaining about Faceware's visual distraction. For example, one of the two users who dislike Faceware commented that "Faceware has too many expressions on the avatar's face, sometimes I cannot tell which are the keys to the pronunciation." Therefore, Lipsync has its own advantage of cleanness, which may also be applied in other fields. Therefore, landmark tracking techniques should also strive to appear as visually clean as Lipsync.

8. CONCLUSION & FUTURE WORK

Our first studies compared facial landmark tracking (Faceware) with audio-driven lip movement (LipSync), the current gold standard used in social VR environments. We perform this comparison in the context of learning a foreign language (Chinese) and show that participants using Faceware obtain significantly better pronunciation ratings than those using the audio-driven model. Moreover, 6 out of 8 participants preferred learning from the Faceware system. Therefore, we have shown that Faceware is more helpful than Lipsync in completing a purposeful task. As for the general communication study, VR-experiencing participants would generally prefer Faceware than Lipsync, while the observer participants could concentrate better while using Lipsync.

In our studies, several steps needs to be improved. First, the word sets chosen by language learning task were not proven

balance. If we can have the chance to further conduct the study, we would consult Chinese linguists to find two equally difficult word sets. Second, when native speakers rated the recordings, they only gave a overall grade. It would be more informative to obtain ratings based on three components of language separately.



Future VR system Sketch

The potential future for this study includes a two way VR system with both users wearing VR. Moreover, the research could be extended by combining this technology with eye tracking and sensing technology of other facial features to create a whole face imitation and therefore enhance the future of social VR.

9. REFERENCES

[1] Chen, Jingying, Bernard Tiddeman, and Gang Zhao. "Real-time lip contour extraction and tracking using an improved active contour model." *International Symposium on Visual Computing*. Springer, Berlin, Heidelberg, 2008.

[2] Franklin, Rachel. "Facebook spaces: A new way to connect with friends in vr." 2017.

[3] Feng, Andrew, Dan Casas, and Ari Shapiro. "Avatar reshaping and automatic rigging using a deformable model." *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*. ACM, 2015.

[4] "Award-Winning, Gold Standard Facial Motion Capture Solutions." *Faceware Tech*, 18 Apr. 2019, www.facewaretech.com/software/live#realtime-for-unity.

[5] Garau, Maia, et al. "The impact of eye gaze on communication using humanoid avatars." *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2001.

[6] "Getting Started with VR Development." *Unity*, 20 Apr. 2019, unity3d.com/learn/tutorials/topics/xr/getting-started-vr-development.

[7] Gupta, Rajat, Rohan Nawani, and Vishal P. Talreja. "Virtual Reality Content Creation using Unity 3D and Blender." *Int. J. Comput. Appl.* 156.3 (2016): 8-12.

[8] Hirata, Yukari, and Spencer D. Kelly. "Effects of lips and hands on auditory learning of second-language speech sounds." *Journal of Speech, Language, and Hearing Research* (2010).

[9] Karras, Tero, et al. "Audio-Driven Facial Animation by Joint End-to-End Learning of

Pose and Emotion." *ACM Transactions on Graphics* (2017).

[9] King, Davis E. "Dlib-ml: A machine learning toolkit." Journal of Machine Learning Research 10, Jul, 2009

[10] Landis JR, Koch GG, The measurement of observer agreement for categorical data. Biometrics. 1977.

[11] Mann, Henry B. "The construction of orthogonal Latin squares." The Annals of Mathematical Statistics 13.4, 1942.

[12] Marsella, Stacy, et al. "Virtual character performance from speech." *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. ACM, 2013.

[13] McAllister, David F., et al. "Lip synchronization of speech." Audio-Visual Speech Processing: Computational & Cognitive Science Approaches, 1997.

[14] Parris, J., et al. "Face and Eye Detection on Hard Datasets." IEEE Int. Joint Conf. on Biometrics: 10, 2011.

[15] Savran, Arman, and Bülent Sankur. "Non-rigid registration based model-free 3D facial expression recognition." *Computer Vision and Image Understanding* 162, 2017.

[16] Katchhi, Shrutik, and Pritish Sachdeva. "A Review Paper on Oculus Rift." *International Journal of Current Engineering and Technology E-ISSN*, 2014.

[17] Spitzer RL, Gibbon M, Williams JBW. Biometrics Research Department: New York State Psychiatric Institute; Structured Clinical Interview for Axis I DSM-IV Disorders. 1994.

[18] VRChatNet. “VRChat.” *VRChat*, 20 Apr. 2019, www.vrchat.net/.

[19] Yannier, Nesra, Kenneth R. Koedinger, and Scott E. Hudson. "Learning from Mixed-Reality Games: Is Shaking a Tablet as Effective as Physical Observation?." Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 2015.

10. APPENDIX

1. Rating criteria:

When you are rating this pronunciation in Chinese, please also note the following rules besides your native speaker intuition:

1. Please check the consonance is correct

Specifically, please pay attention for the following cases:

z vs. zh, c vs. ch, s vs. sh, l vs. n

2. Please check the vowel is correct

Specifically, please pay attention for the following cases:

ing vs. in, eng vs. en, ong vs. on,

ang vs. an, ü vs. u

3. Please check the tone is correct

2. Questions for Study 2:

Face-to-face:
1. I was able to take control of the conversation when I wanted to.
2. It was easy for me to contribute to the conversation.
3. The conversation seemed highly interactive.
4. There were frequent and inappropriate interruptions.
5. This felt like a natural conversation.

Involvement:
6. I found it easy to understand my partner.
7. I felt completely absorbed in the conversation.
8. I can sense a clear emotion of my partner.

Co-presence:
9. I had a real sense of personal contact with my conversation partner.
10. I was very aware of my conversation partner.

Partner-evaluation:
11. My partner was friendly.
12. My partner did NOT take a personal interest in me.
13. I trusted my partner.
14. I enjoyed talking to my partner.
15. I would be interested in meeting my partner face-to-face.