Real-time Facial Animation of a Reality Avatar based on Japanese Vowels in a Speech Audio Stream

Ryoto KATO[1], Yusuke KIKUCHI[2], Vibol YEM[2], Yasushi IKEI[3]

kato-ryoto@ed.tmu.ac.jp, {kikuchi,yem}@vr.sd.tmu.ac.jp, ikei@vr.u-tokyo.ac.jp
[1]Faculty of Systems Design, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo, 191-0065 Japan
[2]Graduate School of Systems Design, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo, 191-0065 Japan
[3]Graduate School, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, 113-8656 Japan

Keywords: Real-time, Speech facial expression, Vowel driven, Realistic avatar

ABSTRACT

A realistic facial avatar representing the user is crucial for social activities in a VR (Virtual Reality) space. We developed a real-time method to create natural talk-animation of a realistic facial avatar that is built by photogrammetry. The motion around the mouth follows the magnitude of Japanese five vowel components extracted from a speech audio stream. The result of a user study indicated that our method improved facial expression during speech as compared to the popular method of audio to facial animation (Oculus Lipsync).

1 Introduction

A realistic avatar of a user that appears in a VR space driven by AI (Artificial Intelligence) technology has gained popularity recently as the VR spaces are used in various fields. A human like avatar enhances the quality of social activities including commercial services provided in the virtual space. For this purpose, the avatar needs to respond in real-time to display reliable and realistic facial expression synchronous to the speech to behave as a human agent. The motion of facial skin during speech is highly intricate and humans are extremely sensitive to facial expression that projects mental state now. Although the facial expressions are classified from muscular elements to emotions [1], automated real-time presentation of them from a speech text to the face of an avatar is not established. The implementation ranges from voice with lips synchrony to expressiveness for subtle emotional change.

Edwards et al. [2] has developed comprehensive face animation system (JALI) that creates an expressive talking avatar from speech audio with its transcript. They proposed combination of jaw rigging and the blendshape (shape interpolation) procedure based on English phonemes and corresponding visemes (visible phonemes) [3]. It took a considerable processing time (around a few seconds) not applicable to real-time conversation. Zhou et al. [4] proposed a method to expand JALI system using deep learning, which enabled near real-time facial animation. Although these research could create convincing facial animation for English speech with phonemes and visemes, they are not necessarily applied to Japanese talker animation. It is because Japanese phonemes are different and the structure of phoneme arrangement in a word or a sentence varied from English. Japanese speech is a series of a clearly articulated phoneme that ends with a vowel, which is a salient nature. While the English pronunciation is dominated by many consonants rather than vowels.

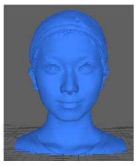
Fig. 1: 3D scanned head model of bare face

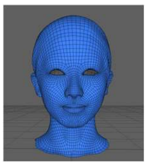
Fig. 2 Shaped polygonal head model

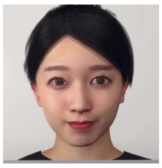
Fig. 3 Avatar in HMD vision

We present a new method to create a realistic facial expression in real-time that uses visemes of Japanese vowels only, which is not realized until now in our best knowledge. This method uses a small amount of user data that enables to apply many faces as compared to JALI system that requires a large control structure and sophisticated facial rigs. In this paper, we present a real-time method to synthesize a natural facial expression of a realistic avatar built by photogrammetry that works on Japanese text.

2 Modeling of VR avatars

Although avatars generally cover the entire human body, this paper discusses the head, which is particularly important in dialogue.

The avatar in this study consists of a 3D model and its deformable facial expression synthesis and speech synthesis systems. We use the blendshape method to synthesize the time course of facial expressions. In the blendshape method, changes in facial expressions are generated by interpolating multiple 3D meshes of the shape of the face [5]. In this study, we use the expression deformation elements of ARFaceAnchor. BlendShape-Location (abbreviated as AARkit), which is published by Apple as a Swift structure of the ARkit. The AARkit defines 52 types of expression deformation elements. The combination of these elements generates various facial expressions on a 3D model of a face [6].

In this study, we used 3D scan data of a human bare face (neutral face) as the basic shape of the avatar's head. Approximately 1.1 million points of head shape data measured using photogrammetry (Fig. 1) were converted to 5284 points of head 3D mesh (Fig. 2) using a software program used for human modeling (R3DS Wrap [7]). The head 3D meshes were deformed with 52 different facial expression deformers in the AARkit. Fifty-two deformed head 3D meshes were prepared for this purpose.

As the amount of mesh vertex movement from the neutral face to the deformed face, we added the amount of mesh vertex movement to the Noh mask mesh used in iFacialMocap [8] (iOS software that executes AARkit and generates 52 types of facial deformations). For this reason, the 52 active meshes were deformed to 5284 vertices using the aforementioned Wrap after applying the maximum number of expression deformation elements one by one. The difference between these and the vertices of the neutral face was used as the vertex displacement.

We also prepared 14 kinds of expression deforming elements of Oculus Lipsync [9], a software program used for generating facial expressions during speech, as comparison objects for the experimental evaluation. These elements are based on MPEG-4 Face and Body Animation (ISO14496) [10] and can generate facial expressions corresponding to phonemes in speech audio stream. The 3D mesh of the head is deformed by these elements.

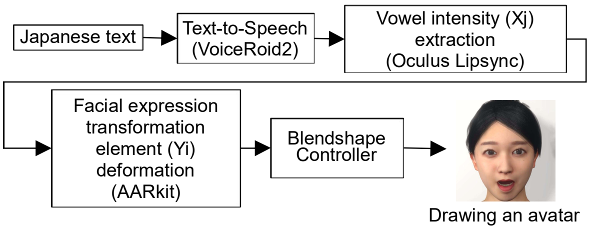
Japanese text
Text-to-Speech (VoiceRoid2)
Vowel intensity (Xj) extraction (Oculus Lipsync)
Facial expression transformation element (Yi) deformation (AARkit)
Blendshape Controller
Drawing an avatar

Fig. 4: Speech facial expression generation flow of VDFE

3 Avatar expression generation method

As a simple method of reasonably good facial expressions during Japanese speech, we constructed and used vowel-driven facial expression (VDFE), which extracts vowel components of speech and measured facial expressions during vowel speech.

3.1 Determination of facial expression by the five-vowel intensity

Fig. 4 shows the process of VDFE. The prepared Japanese speech text is converted into a speech data file using speech synthesis software (VoiceRoid2 [11]). In the case of live dialogue, streaming is used. In addition, the speech of the collaborator (female, 19 years old) was also converted into a speech data file for the experimental evaluation. These audio data are input to a phoneme recognition system (Oculus Lipsync [9]) by streaming and extracted as the intensity (Xj | 0 < Xj < 1.0, j = 1, 2, 3, 4, 5) of the visual phoneme (viseme: AA, ih, ou, E, and oh) components corresponding to the five Japanese vowels. After each of these five vowel components is converted to a value of the face shape Yi (vector of 52 elements of expression deformation, AARkit) on the basis of this value, the position of the 3D shape mesh is determined and drawn.

3.2 Determination of facial expression deformation element vectors

The facial expression deformation element vectors of the five vowels were obtained by having five research participants vocalize them. We used iFacialMocap [8] to obtain the facial deformations from the shape of the face when the five vowels were uttered. Each participant uttered three times for each vowel and 52 coefficient vectors were obtained by averaging the first 0.1 second of utterance [13]. When the indexes of the expression components are i = {1, 2, …, 52} and the indexes of the vowels are j = {1, 2, 3, 4, 5}, the coefficient vector can be represented as a 52 × 5 matrix Z (zi, j | 0 < zij < 1.0).

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{52} \end{bmatrix} = \mathrm{G} \begin{bmatrix} z_{1,\ 1} & \cdots & z_{1,\ 5} \\ \vdots & \ddots & \vdots \\ z_{52,\ 1} & \cdots & z_{52,\ 5} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix} \qquad \textbf{(1)}$$

The expression to be synthesized is approximated as a linear sum of the expression transformation elements corresponding to the intensity of each vowel. A vector of expression transformation elements Yi is determined using Equation (1), where G is the adjustment gain.

Using this method, we generated appropriate speech expressions during continuous speech in Japanese.

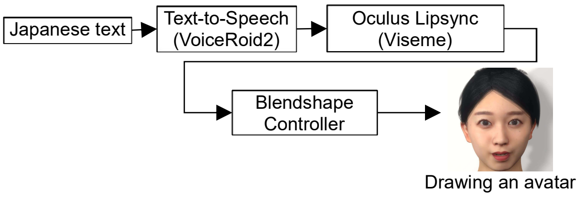

Fig. 5: Speech facial expression generation flow of Oculus Lipsync

4 Experimental evaluation

4.1 Objective and experiment participants

We compared and evaluated the avatar's facial expressions during the utterance of five vowels and the speech of sentences generated by both VDFE and Oculus Lipsync. The evaluation items were the naturalness of facial expressions and the accuracy of shape of the mouth during speech. A total of eight students and faculty members at the university participated in the experiment (average age: 24.8 years).

4.2 Stimulus

The two levels of the VDFE (standard and emphasized) have different values for the Z-matrix. The facial expressions by Oculus Lipsync for comparision were generated using 14 face shapes (viseme) included with this Software (Fig. 5). Three types of avatars were generated using phoneme extraction of Oculus Lipsync; however, all changeable parameters were used as default values.

The speech content of the avatar consisted of two levels: five Japanese vowels (International Phonetic Alphabet (IPA[12]): a, i, ɯ, e, and o) uttered by the collaborator (approx. 5 s of speech) and a sentence (approx. 20 s of speech) synthesized using VoiceRoid2. The synthetic voice of a single vowel was not used as an input voice in this study because it differed from human speech, and correct phoneme extraction was not possible. The five vowels (IPA: a, i, ɯ, e, and o) of the volunteer for this research with the waveform shown in Fig. 6 are output as the components shown in Fig. 5 in Oculus Lipsync. In this experiment, we used an X-matrix with only the corresponding phonemes having values as shown in Fig. 7 and the other phonemes being 0.

Figs. 8, 9, and 10 show the speech expressions for the three levels of vowels (IPA: a, i, ɯ, e, and o) presented as stimuli: Oculus Lipsync, vowel-driven standard, and vowel-driven stress.

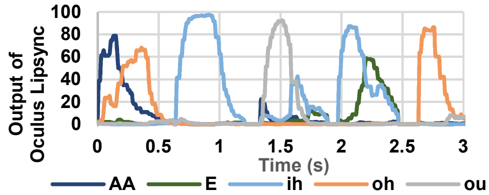

Fig. 6: Phoneme analysis results of Oculus Lipsync on Japanese five-vowel vocalizations by a female collaborator

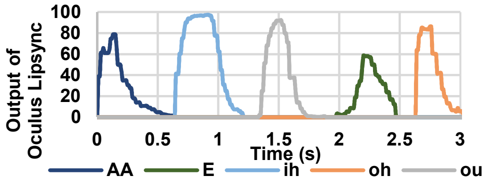

Fig. 7: Phoneme analysis results of Oculus Lipsync on Japanese five-vowel vocalizations by a female collaborator

4.3 Procedure

The participant sat on a chair wearing a head-mounted display (Vive Pro) and evaluated the facial expressions of the avatar as it spoke five vowels and a 20-s sentence in the VR space (Fig. 3); head movement in the seated position was allowed. We presented each participant of the three stimuli (standard, emphasized, and Oculus) two times, and we randomized the order of presentation to eliminate order effects. If a participant asked, we presented the stimulus as many times as needed until the participant finished the evaluation.

The participants rated the naturalness of the facial expressions and the accuracy of mouth shape. For the former, a visual analog scale (VAS) is used and the scale defined as the left end was "robot-like and expressionless" and the right end was "natural expression of human speech.". For the latter, a scale described the left end was "completely mismatch" and the right end was "completely match". The same items were used to evaluate both the five-vowel and the sentence speech, with each of the three levels being viewed once for two sets. In both cases, the order of presentation was randomized to eliminate order effects.

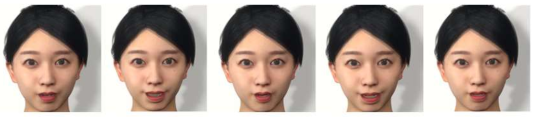

Fig. 8: Speech facial expression for each vowel by Oculus Lipsync (from left to right: a, i, ɯ, e, and o)

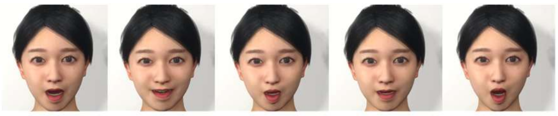

Fig. 9: Speech facial expression for each vowel by standard VDFE (from left to right: a, i, ɯ, e, and o)

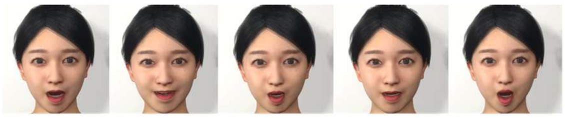

Fig. 10: Speech facial expression for each vowel by emphasized VDFE (from left to right: a, i, ɯ, e, and o)

5 Results and discussion

5.1 Five-vowel speech facial expressions

The results of the evaluation of the naturalness of facial expression and accuracy of mouth shape during the utterance of the five vowels are shown in Figs. 11 and 12, with the left end of the VAS set at 0 and the right end at 100. The results of the one-way analysis of variance showed that the naturalness of facial expressions was significantly different at the 5% level, p = 0.0221. As a result of multiple comparisons on this result, there was a significant difference between Oculus Lipsync and the vowel-driven standard (p = 0.0214) only at the 5% level. The one-way analysis of variance for the accuracy of mouth shape showed a significant difference at the 1% level with p = 0.0068. Multiple comparisons of the same items revealed a significant difference between Oculus Lipsync and the vowel-driven standard only at p = 0.0067 at the 1% level.

These results indicate that the proposed VDFE is highly evaluated for the naturalness of facial expressions and the accuracy of shape of the mouth in the standard speech condition. The VDFE is based on facial measurements during the pronunciation of five vowels and, unlike Oculus Lipsync. In addition, this method also uses changes in areas other than the mouth. The fact that only the vowels corresponding to the input speech were used as the phoneme analysis results may have

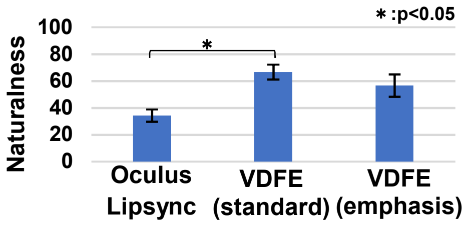

Fig. 11: Naturalness of facial expressions for five-vowel speech

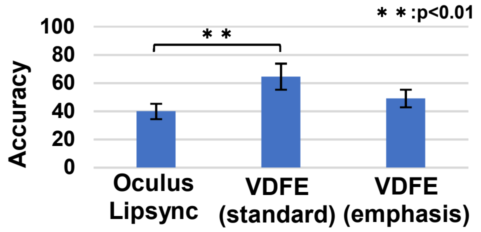

Fig. 12: Accuracy of mouth shape for sentence speech

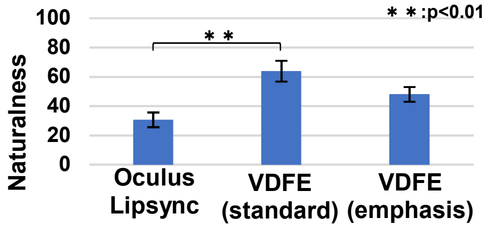

Fig. 13: Naturalness of facial expressions for five -vowel speech

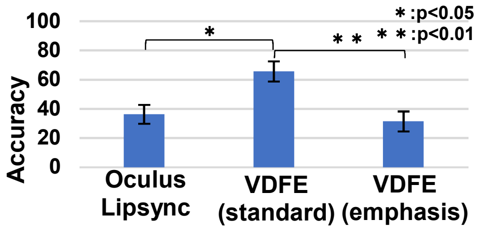

Fig. 14: Accuracy of mouth shape for sentence speech

contributed to the clarity of the facial expressions.

5.2 Sentence speech facial expression

The results of the evaluation of the naturalness of facial expression and the accuracy of mouth shape during sentence speech are shown in Figs. 13 and 14. The results of the analysis of variance showed, p = 0.0045 and p = 0.0053 respectively. There are significant differences at the 1% level for both naturalness and accuracy of mouth shape. As the result of multiple comparisons between the Oculus Lipsync and the VDFE standard (p = 0.0037), there is a significant difference between them at the 1% level. As for the evaluation of mouth shape during speech, there was a significant difference between Oculus Lipsync and vowel-driven standard (p = 0.0223) at the 5% level and between VDFE standard and emphasis (p = 0.0079) at the 1% level.

These results indicate that the standard VDFE speech is highly evaluated for the naturalness of facial expression and mouth shape, even for written speech. These results are also suggesting that VDFE is effective for continuous speech compared to Oculus Lipsync during continuous speech.

6 Conclusions

We proposed an expressive facial animation method that generated a live face motion of a realistic avatar. It is efficient in producing a photorealistic Japanese talker in real-time by using only vowels detected in speech audio stream. This vowel-based speech expression relies on the feature studied in Japanese phonetics.

The future work includes modification of the expression in terms on consonants in addition to synchronization of voice and graphics presentation..

Acknowledgement

This research was supported by SCOPE #191603003. We thank Professor Emeritus Michitaka Hirose at the University of Tokyo for this research opportunity.

References

[1] P. Ekman and W.V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, 1 ed. Consulting Psychologists Press, Palo Alto, California, Aug (1978).

[2] P. Edwards, C. Landreth, E. Fiume, K. Singh, JALI: An Animator-Centric Viseme Model for Expressive Lip Synchronization, SIGGRAPH ’16 Technical Paper,, July 24 - 28, Anaheim, CA (2016).

[3] C. G. Fisher, Confusions Among Visually Perceived Consonants, Journal of Speech and Hearing Research, Vol. 11, Issues 4, pp. 796–804 (1968).

[4] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, VisemeNet: Audio-Driven Animator-Centric Speech Animation, ACM Transactions on Graphics, Vol. 37, No. 4, Article 161, (2018).

[5] J.P. Lewis, K. Anjyo, T. Rhee, M. Zhang, F. Pighin, Z. Deng, Practice and Theory of Blendshape Facial Models, *Eurographics* (2014).

[6] AppleArkit, ARFaceAnchorARFaceAnchor.BlendShapeLocation:https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation

[7] R3DSWrap:https://www.russian3dscanner.com/docs/Wrap3/Nodes/Wrapping/Wrapping.html

[8] iFacialMocap:https://www.ifacialmocap.com/tutorial/unity/

[9] OculusLipsync:https://developer.oculus.com/downloads/package/oculus-lipsync-unity/

[10] Visage Technologies, MPEG-4 Face and Body Animation (MPEG-4 FBA), An overview, pp. 37-40.

[11] VoiceRoid2,Yukari Yuzuki:https://www.ah-soft.com/voiceroid/yukari/

[12] International Phonetic Association, Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge: Cambridge University Press. (1999)

[13] G. Bailly, Learning to speak: Sensori-motor control of speech movements, Speech Communication, Vol. 22, Issues 2–3, pp. 251-267 (1997)