第 26 回日本バーチャルリアリティ学会大会論文集（2021 年 9 月）

ＶＲアバターの表情生成手法に関する研究

Facial expression modeling for a VR avatar

加藤綾斗 [1)]，菊地勇輔 [2)]，ヤェム ヴィボル [2)]，池井　寧 [3)]
Ryoto KATO, Yusuke KIKUCHI, Vibol YEM, and Yasushi IKEI

1) 東京都立大学システムデザイン学部（〒191-0065 東京都日野市旭が丘 6-6，kato-ryoto@ed.tmu.ac.jp）
2) 東京都立大学院システムデザイン研究科（〒191-0065 東京都日野市旭が丘 6-6，{kikuchi,yem}@vr.sd.tmu.ac.jp）
3) 東京大学大学院情報理工学系研究科（〒113-8656 東京都文京区本郷 7-3-1，ikei@vr.u-tokyo.ac.jp）

概要：対話を含む活動を VR 空間で行う場合には，対話者を表わすアバターが有用である．本研究では，フォトグラメトリで構築した頭部モデルを有するアバターが，日本語の発話を行う際に自然な表情を与える手法を開発している．本研究では，テキストから発話・表情変化を与えるシステムを構築して予備的評価を行った結果，提案手法が発話時の表情を改善し得ることが示唆された．

キーワード：表情合成，発話，母音駆動，リアリスティックアバター

1. はじめに

バーチャルリアリティ空間の利用が拡大するとともに，人間の代理をするアバターの重要性が高まっている．実空間の人工物は，人間の身体に適合するように構築され，それを学習した人間にとっては同様の構成を持つ VR 空間の可用性が高い．社会活動では人間がその対象なので，空間と同様に実際の人間に近い，リアリティの高いアバターが円滑な VR 空間利用の基礎の１つとなる．社会的活動では，コミュニケーションが決定的に重要であり，発話時の表情を自然に豊かに表出することが求められる．

本論文では，発話時の表情を音声から生成する手法として，母音成分だけを用いる簡易手法について評価した結果を述べる．

2. アバターモデリング

アバターは一般的には人間の全身を対象としているが，本稿では，対話において特に重要な頭部について論じる．

本研究のアバターは，３D モデルとその変形による表情合成および発話合成系から構成される．表情の時間経過の合成にはブレンドシェイプ法を用いる．ブレンドシェイプ法では，複数の顔形状３D メッシュを補間することで表情の変化を生成する[1]．本研究では，Apple 社が ARkit の Swift 構造体として公開している ARFaceAnchor.BlendShapeLocation（AARkit と略）の表情変形要素を用いる．表情変形要素は，顔の特徴点の相対的な動き（例えば，口を開ける，など）を表わしており，AARkit では 52 種類の表情変形要素が定義されている．これらの組み合わせで顔の３D モデルに多様な表情を生成する[2].

本研究では，アバターの頭部の基本形状として，人間の素顔（中立顔）の 3D スキャンデータを用いた．フォトグラメトリで計測された頭部の約 110 万点の形状データ（図 1）を，人物モデリング用ソフト（R3DS Wrap [3]）で 5284 点の頭部３D メッシュ（図 2）に変換した．この頭部３D メッシュを AARkit の 52 種類の表情変形要素で動かすが，このために変形後の頭部 3D メッシュを 52 個用意した．

中立顔から表情変形後までのメッシュ頂点の移動量としては，iFacialMocap [4]（AARkit を実行し 52 種の表情変形を生成する iOS ソフトウェア）で使用されている能面型のメッシュの頂点の移動量を採用した．このため，表情変形要素を 1 つずつ最大適用した後の 52 個の能面型メッシュを，上記の Wrap により，5284 頂点に変形した．これらと中立顔の頂点との差分を頂点移動量とした．

また，評価実験の比較対象として，Oculus 社の発声時表情生成ソフトウェア Oculus LipSync [5]の 14 種の表情変形

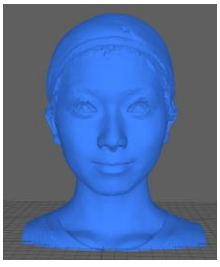

図 1: 素顔の 3D スキャン頭部モデル

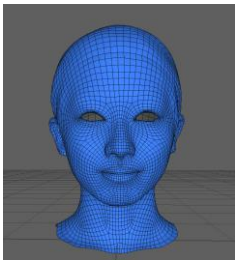

図 2: 整形したポリゴンの頭部モデル

要素も用意した．この表情変形要素は，MPEG-4 Face and Body Animation (ISO14496) [6]に基づき， 音素に対応した表情生成を可能とする．これによって，頭部 3D メッシュを変形させた．

3. アバター表情生成手法

日本語を発話する際の表情を近似的に生成する簡易手法として，発話音声の母音成分を抽出して，母音発話時の（計測済みの）表情を提示する「母音駆動混合形状法」を構築して用いた．

3.1 5 母音強度による表情の決定

母音駆動混合形状法の処理手順を図 3 に示す．用意した日本語の発話テキストを，音声合成ソフトウェア（VoiceRoid2 [7]）によって音声 data file に変換する．（ライブ対話の場合は streaming で利用する．）さらに，評価実験用に，実験協力者（女性，19 歳）の発話も音声 data file に変換した．この音声データを streaming によって，音素認識系（Oculus LipSync [5]）に入力し，日本語の 5 母音に対応した口形素（AA, ih, ou, E, oh）成分の強度 ($x_j | 0<x_j<1.0$, j=1,2,3,4,5)として取り出す．この 5 母音の成分のそれぞれに対応した顔形状 Y_i(表情変形 52 要素のベクトル, AARkit)の値に変換し，この値に基づいて，顔の 3D 形状メッシュの位置を決定して描画する．

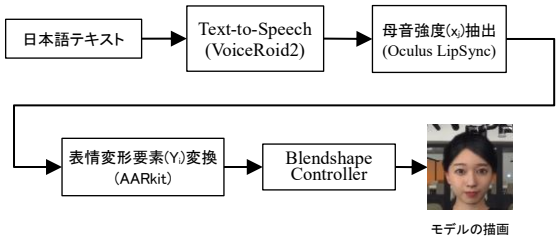

図 3: 母音駆動混合形状法の発話表情生成フロー

3.2 表情変形要素ベクトルの決定

5 母音を発声したときの顔の表情変形要素ベクトルは，5 人の実験参加者に発声させてデータを取得した．発声時に標準的表情で発音した場合と，はっきりと表情の特徴が表れるように強調した発声をした場合の 2 通りを計測した．5 母音を順に発声した際の顔の形から表情変形要素を得るために， iFacialMocap [4]を用いた．5 母音を各実験参加者が 3 回ずつ発声し，発声開始から 0.1 秒間の平均[1]で 52 個の係数ベクトルを求めた．表情変形要素のインデックスを i={1,2,,,,52}, 母音のインデックスを j={1,2,3,4,5},とするとき，この係数ベクトルは，52x5 の行列 Z ($z_{i,j} | 0<z_{ij}<1.0$) で表すことができる．

合成する表情は，各母音の強度に対応した表情変形要素の線形和で表すと近似し，表情変形要素のベクトル Y_i は式(1)で決定する．ここで G は調整ゲインである．

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

これを用いて，日本語を連続発話した際の近似的な表情を生成した．

4. 表情品質評価実験

4.1 目的と実験参加者

5 母音の発声および文章の読み上げをした際のアバター表情を母音駆動混合形状法で生成した場合と Oculus LipSync で生成した場合で比較評価した．評価項目は，表情の自然さ，発話時の口の形状である．実験参加者は，同大学の学生・教員計 8 名（平均 24.8 歳）である．

4.2 刺激

発話時のアバターの表情生成は，前章で述べた母音駆動混合形状法において，標準の発話の場合と強調発話の場合，および Oculus LipSync による場合の 3 水準とした．母音駆動混合形状法の 2 水準（標準，強調）は，Z 行列の値が異なる．Oculus LipSync による表情は，それに付属した 14 の顔形状（Viseme）を用いて生成した．この場合は，図 4 で表される．

3 種類のアバター生成では，Oculus LipSync の音素抽出を利用しているが，変更可能なパラメータはすべてデフォルト値を使用した．

アバターの発話内容は，実験協力者の発話による日本語 5 母音（あ，い，う，え，お）(約 5 秒間の音声, 図 5) と，VoiceRoid2 の合成音声による文章（約 20 秒間の音声）の 2 水準とした．単独母音の合成音声は，人間の発声と異なり，正しい音素抽出が出来なかったことから，本研究では入力音声として使用しなかった．

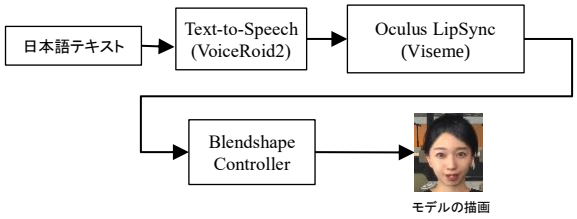

図 4: Oculus LipSync の発話表情生成フロー

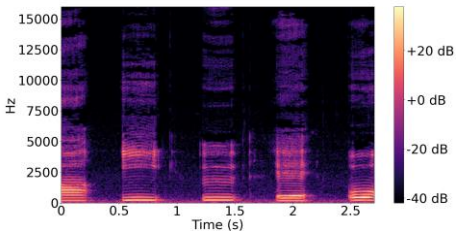

図 5: 日本語 5 母音（あ，い，う，え，お）の音声波形

[1] 発声時の顔変形（口形素）は発声の開始と一致する[8]

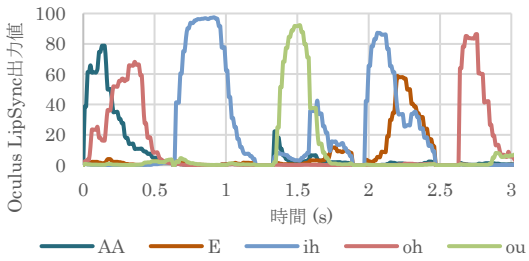

図 6: 日本語 5 母音発声（女性実験協力者の声）の Oculus LipSync の音素解析結果

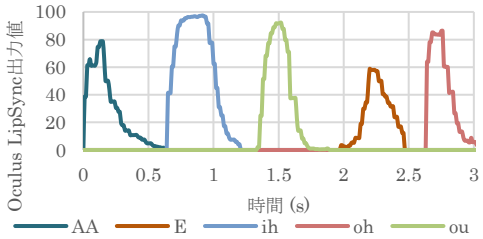

図 7: 日本語 5 母音発声（女性実験協力者の声）の音素解析結果から該当母音成分を抽出した結果

図 5 の波形を持つ実験協力者の 5 母音（あ，い，う，え，お）発声は，Oculus LipSync では図 6 の成分出力となる．1 つの母音発音に対して，このように複数の母音の音素が出力されると，表情が混合し不明確な表情となる．そこで，この実験では，5 母音の発話表情生成の場合，図 7 示すように該当音素のみ値を持たせ，それ以外を 0 とした X 行列を入力として発話表情生成を行った．刺激として提示した Oculus LipSync，母音駆動標準，母音駆動強調の 3 水準の各母音（あ、い、う、え、お）に対する発話表情をそれぞれ図 8，図 9，図 10 に示す．

図 8: Oculus LipSync による各母音に対する発話表情（左から，あ，い，う，え，お）

図 9: 母音駆動混合形状法による各母音に対する標準発話表情（左から，あ，い，う，え，お）

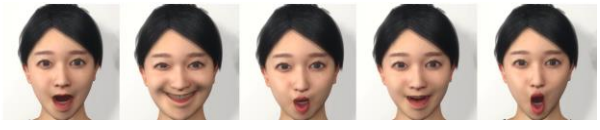

図 10: 母音駆動混合形状法による各母音に対する強調発話表情（左から，あ，い，う，え，お）

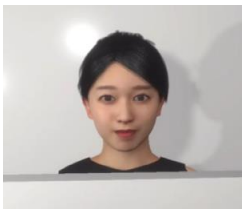

図 11: HMD 視野におけるアバター

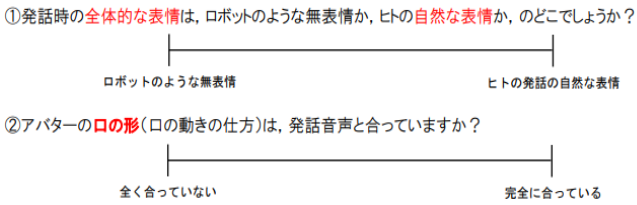

図 12: 発話表情に対する評価項目と尺度

文章の発話時表情評価の際には，Oculus LipSync の 5 母音成分出力を使用した．

4.3 手順

参加者は，着座し Head Mounted Display (VivePro) を装着して， VR 空間の正面で発話するアバター（図 11）の表情を評価した．着座位置での頭部運動は許可した．

最初に母音発声を評価した．3 水準を 2 回ずつ見ることを 1 セットとし，それを 2 セット行い，各提示刺激に対して表情の自然さ（図 12①）とアバターの発話時の口の形の正確さ（図 12②）を VAS（Visual Analogue Scale）で評価させた．次に，文章について，同様の評価項目を用いて評価させた．3 水準を 1 回ずつ見ることを 1 セットとし，それを 2 セット行った．母音の場合，文章の場合，どちらも 3 水準の提示順はランダムとして，順序効果を排除した．

5. 結果および考察

5.1 5 母音発話表情

5 母音発話時の表情の自然さと口の形の評価結果を，図 13，図 14 に示す．VAS の左端を 0，右端を 100 としている．分散分析（一元配置）の結果，表情の自然さは $p=0.0221$ で 5%水準で有意差が認められた．多重比較の結果，Oculus LipSync と母音駆動標準の間のみ $p=0.0214$ で 5%水準で有意差が認められた．また，口の形の正確さの分散分析（一元配置）では，$p=0.0068$ で 1%水準で有意差が認められた．同項目について多重比較を行った結果，Oculus LipSync と母音駆動標準の間のみ $p=0.0067$ で 1%水準で有意差が認められた．

これらより，提案した母音駆動混合形状法の標準の発話条件の表情の自然さ，口の形の評価が高いことが示された．これは，5 母音発音時の顔計測に基づいていること，Oculus と異なり，口以外の部分の変化も用いられているためであると考えられる．また，音素解析結果として，入力音声に該当する母音のみを採用したことも表情の明確さに寄与したと考えられる．

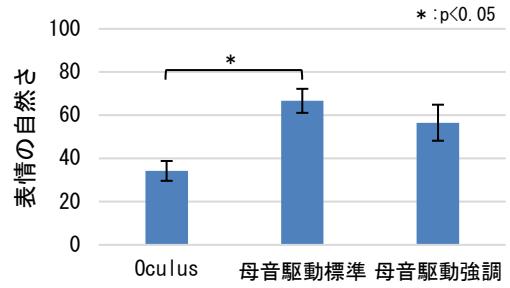

図 13: 5 母音発話時の表情の自然さの評価結果

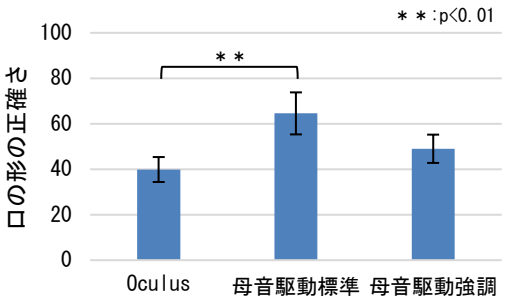

図 14: 5 母音発話時の口の形の評価結果

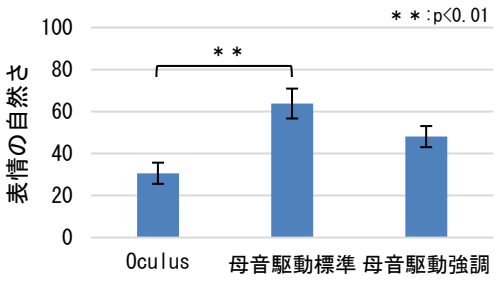

図 15: 文章発話時の表情の自然さの評価結果

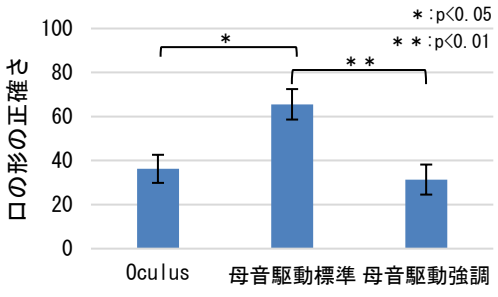

図 16: 文章発話時の口の形の評価結果

5.2 文章発話表情

文章発話時の表情の自然さと口の形の評価結果を, 図 15, 図 16 に示す. 分散分析の結果, それぞれ p=0.0045, p=0.0053 となり，両者とも 1%水準の有意差が認められた．また，両者についてそれぞれ多重比較を行った結果，文章発話時の表情の自然さについては，Oculus LipSync と母音駆動標準間のみ p=0.0037 で 1%水準で有意差が認められた．文章発話時の口の形の評価については，Oculus LipSync と母音駆動標準間で p=0.0223 で 5%水準の有意差が認められ，母音駆動標準と母音駆動強調間で p=0.0079 で 1%水準の有意差が認められた.

これらの結果より，母音駆動混合形状法の標準の発話は，文章の発話においても表情の自然さと口の形の評価が高いことが示された．5 母音発話と異なる点は，Oculus LipSync の出力を加工せずに用いていることであり，連続発話にそれが有効であることが示唆された.

6. おわりに

本研究で提案した手法は，母音発音時だけの顔形状の計測で実現でき，処理が容易であるにも関わらず，子音も含めて 15 種の Viseme を使う Oculus LipSync より，表情の自然さが高く，有効であることが示された.

今後の課題としては，子音の表情を明確にすること，音声と CG のより正確な同期を行うことなどが挙げられる.

謝辞 本研究は，著者（池井）の所属講座および総務省 SCOPE#191603003 の支援を受けて実施された．また，本研究の機会は，東京大学 廣瀬通孝 名誉教授に頂いた．ここに謝意を表する.

参考文献

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

[2] Apple Arkit ARFaceAnchorARFaceAnchor. BlendShape Location:https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation

[3] R3DS Wrap:https://www.russian3dscanner.com/docs/Wrap3/Nodes/Wrapping/Wrapping.html

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

[5] OculusLipSync:https://developer.oculus.com/documentation/unity/audio-ovrlipsync-unity/?locale=ja_JP

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

[7] VoiceRoid2, 結月ゆかり:https://www.ah-soft.com/voiceroid/yukari/

[8] Bailly G.: Learning to speak. Sensori-motor control of speech movements, Speech Communication, Vol. 22, Issues 2–3, pp. 251-267, 1997. doi.org/10.1016/S0167-6393(97)00025-3. (1997)