5W-08

音声認識によるコミュニケーションツール用 3D モデルの表情反映

鈴木智也† 田谷昭仁† 戸辺義人†
青山学院大学理工学部情報テクノロジー学科†

1. はじめに

2020 年から蔓延し現在も拡大している COVID-19 は，接触を避けるためインターネットを介したコミュニケーション方法の種類を増加させた．その中で近年の VR（Virtual Reality）技術の発達により 3D モデルを用いたコミュニケーションに期待が高まっている．現在，3D モデルの表情を作る技術の多くがフェイストラッキングを採用している．しかし，カメラの存在が不可欠であるこの技術において，話者の行動はカメラの撮影範囲内を限定される．この課題を解決するため，本研究では多くのデバイスに搭載されているマイクを使用した音声認識型の 3D モデルの表情反映を提案する．本研究は 3D 技術を活かしたコミュニケーションツールの作成を目的としており，本稿では設計，実装，実験，評価について述べる．

2. 関連研究

3D モデルの顔アニメーションが音声入力でリアルタイムかつ低遅延駆動する 3D モデル表情反映法[1]がある．DNN や CNN を利用し，性別やアクセント，言語の異なる話者でも反映可能である．しかし，この研究では 1 つの 3D モデルに対し機械学習を行い汎用化させているため，他の 3D モデルを使用するためには一から学習する必要があるという課題がある．また，敵対的生成ネットワークを活用し，生音声入力から直接現実の話者の顔画像を高精度に生成する研究[2]がある．しかし，3D モデルを使用したコミュニケーションにおいて表情反映の正確性は必須ではなく，感情をうまく相手に伝達することが要求される．音声から直接表情を生成するのでなく，感情という中間層を推定することで，異なる 3D モデルへの適応が可能となることや，感情のアニメーションエフェクトを追加することができ，応用の幅が広がるメリットがある．

3D Character Face Representation Based on Voice-based Emotion Recognition
†Tomoya SUZUKI, Akihito TAYA, Yoshito TOBE
/Aoyama Gakuin University

3. 提案システムの設計

3.1 システム概要

提案システムはマイク入力された音声信号から感情の推定を行い，感情に対応した 3D モデルの表情へ反映するコミュニケーションツールである．マイクから入力された音声信号を解析し，音声信号の特徴量を算出し，感情の推定に使用する．感情の推定には 3.4 で説明する機械学習法を使用する．推定した感情に応じて，予め用意しておいた 3D モデルの感情表現を選択し，表情として反映する．

3.2 推定する感情の決定

使用する感情として Eckman[3]が定義した「怒り」「喜び」「嫌悪」「驚き」「悲しみ」「恐怖」の 6 種類に，会話の途切れる瞬間や感情を出さない場合を考慮して「平静」を追加し，計 7 種類を採用した．

3.3 音声信号の特徴量算出

感情の推定には音声からリアルタイムで感情推定する際に有効とされる特徴量である音量の二乗平均平方根（RMS：Root of Mean Square），メル周波数ケプストラム係数（MFCC：Mel-Frequency Cepstrum Coefficients），RMS，MFCC の一次差分，そして次に示す代表値を採用している[4]．使用する代表値は最大値，最小値，最大値と最小値の範囲，平均値，標準偏差値，線形近似した際の勾配とオフセットである．

システムの流れとして生音声データをサンプルレート Fs で処理し，一定間隔で動作するフレーム毎に算出に使用する直前までの音声信号データ L 個を使用し，特徴量を抽出する．表情変化のアニメーションにある程度の時間がかかることと，高頻度な表情変化に違和感があること

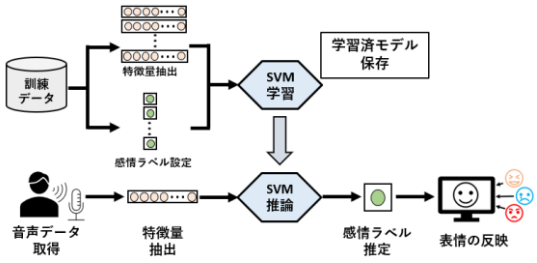

図 1 感情推定までの流れ

から，感情の推定と表情への反映は s 秒おきに行う．s 秒間に算出動作するフレーム数は f とする．

3.4 感情推定の判定法

本研究では，感情推定に SVM（Support Vector Machine）を使用しており，全体的な処理の流れを図 1 に示す．

学習方法として，感情ラベル付き音声データ[5]から 2119 個の特徴量を抽出し，SVM を学習させた．提案システムではこの学習済モデルを用いて抽出した特徴量から感情推論を行う．

4. 実装および評価実験

4.1 実装

本研究では 3D モデルに UNITY-CHAN![6]を使用した．また，オープンプラグインであるリップシンク[7]を採用したことで会話をしている臨場感を与えた．

実装には Unity を使用しており，3.3 に示したパラメータの実装を行った．音声サンプルレート Fs は 44100 Hz に設定し，特徴量の算出に使用したデータ個数 L は 4096，動作フレーム数 f を 50 回/s とした．最後に，表情変化の周期 s を 1 秒と設定した．

4.2 評価実験

実験はすでにシステム操作経験のある 3 人と未経験の 7 人の計 10 人が行った．実験はマイクに外音が入力されるのを抑えるため静かな空間で行った．実験機器はピンマイクが USB 接続された PC を使用した．会話を想定しているため二人一組で実験を行い，一人が実験を行っている間もう一人は動作状況を聞き手として観察してもらった．被験者には首元にピンマイクを装着してもらい，発話を行った．

始める前に予備実験として感情指定のあるセリフ 21 個を発話してもらった．評価実験は全体で二種類行った．実験 1 では感情指定のないセリフ 20 個を話者が想像する感情を含め発話し，3D モデルの表情変化を観察した．実験 2 では，感情変化の頻度を実験 1 で使用した 1 秒の他に 0.5 秒，1.5 秒，2 秒の計 4 種類を観察しどの表情変化が適切かを見てもらった．動作確認後，簡単なアンケートに回答してもらい評価値とした．アンケート項目として「自分の声と表情が対応していると思うか」や「友達との会話で使用したいと思うか」などを評価値 1−5 の 5 段階で回答してもらった．実験環境を図 2 に示す．

4.3 評価結果

話し手の立場として声と表情の対応を見た際，評価値 3，4 の合計が 8 人と半数以上がどちらでもないまた多少一致したと回答した．聞き手としては評価値 3，4 の合計が 7 人と話し手とあまり変わらない回答となった．また，どのような関係の方と使用したいかについては評価平均値 3.7 と友達と使用したいという意見が多かった．初対面の方々とは使用したいかについては評価値 1，2 の合計 5 人，4，5 の合計 4 人と別れる形となった．使用者からは低い声の方は「怒り」が出やすいという意見や感情によって表情出現頻度が変わったほうが良いという意見があった．総評として，音声による表情の変化がよく出ており会話のアクセントとなっていることから被験者が使用したいと感じるものとなった．

図 2　評価実験

5. 結論

本研究では，音声認識型の 3D モデルの表情反映システムについての提案を行った．提案システムではリアルタイムで表情豊かな 3D モデルのコミュニケーションを可能としている．現在のところ，コミュニケーションに使用する 3D モデルは「UNITY-CHAN!」のみだが，将来的にどのような 3D モデルでも簡単に利用することが可能となる．今後の展望として，より複雑な表情の生成や他の 3D モデルへの応用，複数デバイスでの通話での利用を検討している．

参考文献

1) Tero, K., Timo, A., Samuli, L., Antti, H., and Jaakko, L.: Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion, *ACM Trans. Graph*, Vol.36, No.4, Article 94(2017).
2) Amanda, D., Francisco, R., Miquel, T., and et al.: WAV2PIX: Speech-Conditioned Face Generation Using Generative Adversarial Networks, *ICASSP 2019*, pp.8633-8637(2019)
3) Paul, E.: An Argument for Basic Emotions, *Cognition and Emotion*, Vol 6, No.3-4, pp.169-200(1992).
4) Tang, B.N.， 目良和也，黒沢義明，竹澤寿幸：音声に含まれる感情を考慮した自然言語対話システム，HAI シンポジウム 2014，pp.87-91（2014）
5) 音声資源コンソーシアム：慶應義塾大学 研究用感情音声データベース(Keio-ESD)，音声資源コンソーシアム(オンライン)，入手先〈http://research.nii.ac.jp/src/Keio-ESD.html〉(参照 2020-07-14)
6) Unity Technologies Japan：UNITY-CHAN!，Unity Technologies Japan(オンライン)，入手先〈https://unity-chan.com/contents/guideline/〉(参照 2020-06-12)
7) Oculus：Oculus LipSync Unity，Oculus(オンライン)，入手先〈https://developer.oculus.com/downloads/package/oculus-lipsync-unity/〉(参照 2020-06-22)