# 音声認識によるコミュニケーションツール用バーチャルモデルの表情反映

鈴木智也[1]　田谷昭仁[1]　戸辺義人[1]

**概要：**本研究ではマイク入力によって取得した音声信号を解析し，指定した音声信号の特徴量によって感情を推定し3D モデルに表情を反映させるコミュニケーションツールを作成することを目的としている．予め決定した閾値に沿って感情を推定し表情を反映させる実験を行ったが，正確性において課題が残ったため今後機械学習の導入を検討していく．

## 1. はじめに

近年のバーチャル技術の普及により様々な製品やシステムが生み出されている．今後，さらなる普及が見込まれるこの分野において新しい繋がり方の形としてコミュニケーションへの応用が期待される．本稿では，コミュニケーションに重要な要素の一つ 3D モデルの表情に着目する．現在，3D モデルの表情を作る技術として多くがフェイストラッキングを採用している．しかし，カメラやゴーグルの存在が不可欠であるこの技術において，話者の行動範囲がカメラの撮影範囲内に限定される，ゴーグルを長時間着用し続けるため心身に負荷がかかるという課題が存在する．この課題を解決するため，本稿では様々な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．提案手法ではマイクから取得した音声信号をリアルタイムで解析して感情の推定を行い，3D モデルの表情へ感情を反映する．本稿では，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，設計，実装について述べる．

## 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レイテンシで駆動するシステムの研究がある(1)．この研究では，音声では説明できない顔の駆動を表現しており，よりリアリティのある表情の生成を可能としている．また，別の研究では敵対的生成ネットワークを活用し，生の音声入力から直接話者の顔画像を高精度に生成するものがある（2）．この研究では，3D モデルの生成ではなく現実の話者の顔をピクセルレベルで再現している．

どちらの研究も音声から正確にモデルの表情を再現させることを目的としている．しかし，本研究の目的であるコミュニケーションツールとして正確性は必須でなく，感情を相手にうまく伝えることが要求される．音声から直接表情を再現するのでなく，感情を推定しておくことで，異なる 3D モデルへの適応が可能となることや，感情表現のアニメーションエフェクトの追加を行えるなどといった，応用の幅が広がるメリットがある．

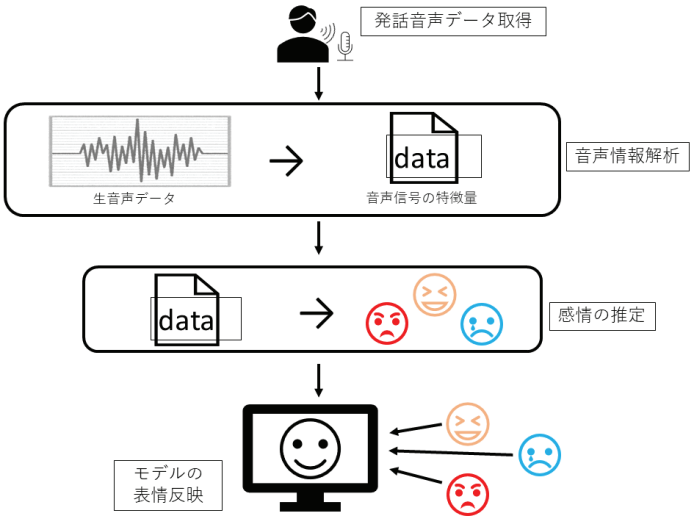


図 1 システムの流れ

## 3. 音声認識型 3D モデルの表情反映システム

### 3.1 システムの概要

提案システムはマイクから入力された音声を使用し感情の推定を行い，その結果から使用する 3D モデルの表情を反映するコミュニケーションツールである．システムの概要を図 1 に示す．

マイクに入力された音声信号を解析し数値的に出力，そのデータをもとに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 で説明する判定式を使用する．推定した感情に応じて，あらかじめ用意しておいた 3D モデルの感情表現を選択し，表情として反映する．

### 3.2 感情推定のため指標の決定

感情の推定を行うにあたって（3）で使用された感情推定において有効とされている特徴量を参考に，有効度が高く設計を行うことの可能な特徴量を採用した．本研究では，算出する値に感情推定で主に使用される音量の二乗平均平方根値（RMS：Root of mean square），またこの値を使用した音声信号の特徴量を使用し，その値から感情の推定を行っていく．

1 青山学院大学理工学部情報テクノロジー学科

### 3.3 音声信号の特徴量

音声信号の特徴量の算出はフレームごとに行う．フレームレートは $f$ とする．フレームごとにマイクから入力された音声信号をサンプルデータとして取り出す．サンプルデータは算出時から直前の $N$ 個のデータを使用する．フレームごとにサンプルデータの RMS 値を求め，直近の $M$ フレーム内での RMS 値の最大値（max），平均（mean）をそのフレームの特徴量として計算する．

### 3.4 感情推定の判定式

特徴量を利用した感情推定の判定式を図 2 に示す．感情の種類は（4）で扱われた感情の種類から本研究で使用するモデルと組み合わさるもの計 5 種類に分け，「怒り」,「喜び」,「驚き」,「平静」,「悲しみ」で判定を行った．$M$ フレーム終了時に判定式から感情の推定を行い，判定結果から表情を画面に反映させる．

事前に測定した情報から通常話者が会話をする際に測定される平均的な RMS 値は $1 \leqq \text{mean} \leqq 5$ の範囲である．しかし，マイクは音声以外にも様々な音を入力してしまう．パソコンのファンや起動音，エアコンの風の音に掃除機の操作音など多くの外的要因が存在するため，$0 \leqq \text{mean} < 1.5$ の範囲において雑音の範囲として確保する必要がある．そのため，会話を行っている平均的な RMS 値を $1.5 \leqq \text{mean} \leqq 5$ と設定する．また，表情は常に閾値に当てはまるとも限らないことから例外処理を確保することも必要である．よって,「平静」は発話をしていない状況また，その他感情が当てはまらない状況下での表情と仮定し，雑音と沈黙，さらに例外の状態とするための例外処理として設定する．

通常の中で会話をする際，平均的にほぼ同じ音量で行われる感情は「喜び」,「悲しみ」,「驚き」である．しかし，すべての感情が一定の音量で会話をするわけではないので，最大音量の程度で閾値を設定する．すでにこれらの音量は $1.5 \leqq \text{mean} \leqq 5$ に収まると設定されているため，追加の判定材料として max で判定を行う．喜びは通常気持ちが高まっている際に発生する感情であり，感情の高まりが音量と比例するならば音量は大きめであると考えられる．そのため「喜び」は $3 \leqq \text{max} \leqq 5$ と設定する．反対に「悲しみ」は気持ちの落ち込みが発生する際に起こる感情であるため，音量は小さいと考えられる．よって,「悲しみ」は $1.5 \leqq \text{max} \leqq 2.5$ と設定する．「驚き」は例外として瞬間的に max が 5 の範囲を越える場合が存在する．およそ二倍近くの音量が発生すると仮定し，「驚き」では $\text{max} \geqq 10$ が存在する場合と設定する．この中に当てはまらない感情が存在した場合例外処理として「平静」となる．

最後に「怒り」である．この感情は「喜び」を越えた感情の高まりが起こると仮定する．そのため，通常の会話中より大きい音量を出すと考えられる．そのため,「怒り」は $5 < \text{mean}$ とし音量の大きさと感情を最大に近い気持ちの高まりと設定する．

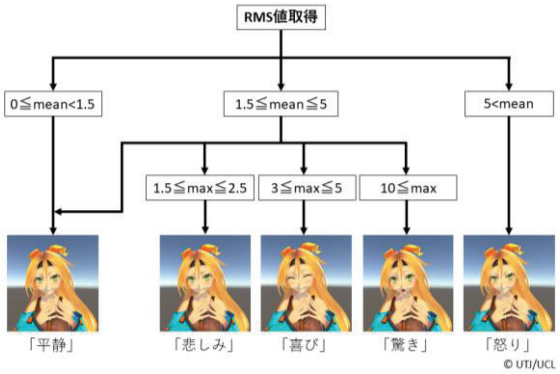


図 2 感情推定の判定の流れ

## 4. 実装と結果

### 4.1 使用 3D モデル「ユニティちゃん」

本稿で使用する 3D モデルはユニティ・テクノロジーズ・ジャパン合同会社が提供する 3D キャラクタ「UNITY-CHAN!」（5）である．このモデルでは顔のモデルがそれぞれのパーツごとに分割して作成されており，よりリアリティのある表情を表現することが可能となっている．また，オープンアセットであるリップシンク（6）を採用し，より会話をしているように感じさせるツールとなるよう実装している．

### 4.2 実験

実装には Unity を使用しており，パソコンに内蔵また接続されているマイクを識別し，マイクから入力される音声信号を数値的に出力することが可能である．このサンプルデータを利用し，3.3 で説明した特徴量を算出し推定を行っていく．

実験を行うにあたり，それぞれの変数を設定していく．フレームレートは定期的な算出を大量に行う必要があるため $f$=50 と設定した．サンプルデータ数は音声処理で多く使用される $N$=4096 個に設定した．最後に周期だが今回表情が不自然に変化しないよう 1 秒で表情の変化を行うことと設定したため $M$=50 と設定した．

### 4.3 実験結果と評価

3.3 で設定した判定式の元、システムを稼働した結果，マイクから入力された話者の音声の大きさによってモデルの表情が変化した．実験から話者が発話することによるモデルの表情は「平静」が多く出現する結果となった．しかし，外的要因に起因する音声は話者の状況下で変化するため，外的な音が大きい場合「喜び」や「怒り」といった表情が多く反映される結果となる．一番反映率が低いものは「悲しみ」という結果となった．1 秒ごとに変化するモデルの表情変化は円滑に変化しているとは言えないが，表情の機敏を表現することはできた．

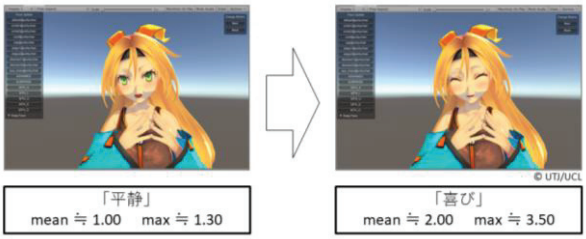


図 3 実行結果

## 5. 結論

本稿では，マイク入力から音声信号を取得しその音声データを解析して感情推定し，その結果をモデルへ表情として反映させるシステムの提案を行った．提案システムでは音声信号の特徴量から感情推定する判定式を導入し，話者の音声に反応し 3D モデルの表情を変化させた．しかし，判定式が簡潔なため話者の感情を正確に反映できていない．表情はコミュニケーションにおいて重要なファクターであり，正確性は維持しなければならない．今後，機械学習の導入とともにさらなる機能改善を行っていく必要がある．

## 参考文献

1) Tero, Karras. Timo, Aila. Samuli, Laine. Antti, Herava. and Jaakko, Lehtinen.: 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, Duarte. Francisco, Roldan. Miquel, Tubau, and et al.: WAV2PIX: Speech-conditioned Face Generation using Generative Adversarial Networks, *ICASSP 2019*, pp.8633-8637(2019)
3) Tang, Ba, Nhat. 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）
4) 田名網那由多，林実：感情音声における音響特徴量の選択，大学コンソーシアム八王子（2018）
5) UNITY-CHAN! http://unity-chan.com/
6) Oculus LipSync Unity
https://developer.oculus.com/downloads/package/oculus-lipsync-unity/