Vol.2021-HCI-195 No.26
Vol.2021-UBI-72 No.26
2021/12/1

感情表現可能なコミュニケーションツール用 ３Ｄアバタにおける自動翻訳機能の検討

鈴木智也 [1,a)]　田谷昭仁 [2,b)]　戸辺義人 [2,c)]

概要：３Ｄアバタによる動画配信や観光案内は自動翻訳の発展もあり国内外問わずユーザ同士のコミュニケーション方法の一つとして普及しつつある.アバタを利用することで話者の表情や話し方の特徴から調整した感情を表現することが可能となる.このような 3D アバタを提供するツールに自動翻訳機能を付与することで，国外の視聴者や観光客とのコミュニケーションを支援することができる．しかし，自動翻訳に時間を要するため，話者の表情のアバタへの反映と翻訳文章の読み上げに時間差が発生し会話に支障をきたすことが懸念される.この課題を解決するため翻訳された文章が読み上げられるまでの遅延時間を考慮した３D アバタへの感情反映を提案する.話者の発話感情を保存し読み上げ音声やアバタに適応させることで翻訳音声でも適切な感情の円滑な伝達を行う．

キーワード：３Ｄアバタ，感情，コミュニケーション，自動翻訳，機械翻訳

1. はじめに

インターネットが発展することで，ネットショッピングやコールセンター, Web や SNS を利用した音声や映像コンテンツの視聴など，私たちの生活の多くの行動を拡大し変化をもたらせてきた．特に，インターネットを利用したコミュニケーションやオンライン会議は以前から行われてきたが，2020 年度から発生している COVID-19 が大流行していることもあり，国内外の多くのユーザ同士の遠隔のコミュニケーションを容易するため世界的に普及が加速した．

また，自動翻訳・機械翻訳の技術も発展している．各ブラウザでのサイトの自動翻訳に使用するだけでなく，SNS や動画配信などにも使用され国外との対人コミュニケーションにも使用され，容易にコミュニケーションをとれるように言語の壁をなくす働きがみられている．

読み上げ音声ソフトウェアなどの技術も発達し，機械音声とは想像できない実際の人間の声に近い読み上げを行うことができる．対人コミュニケーションに使用することはもちろん駅や各施設でのアナウンス，音楽への利用などで活用されている．このような読み上げ音声でも感情調節が可能であり，状況に適した音声を発話することが可能となる．日本語だけでなく外国語への対応も進んでおり，様々なユーザに適した利用を行うことも可能となっている．

このような多くの形で行われている対人コミュニケーションでは，話者の姿の代わりに３D グラフィックのアバタを利用することへの期待が高まっている．近年では，３D アバタを利用した動画配信や電光掲示板のよる観光案内など，様々な場面での利用が行われている．３D アバタを利用することであらかじめ表情の調整することで感情豊かなコミュニケーションを図ることが可能である．リアルタイムでの利用に関してもカメラやマイクから取得することのできる話者の表情や話し方の特徴から調整した感情を表現する[1][2]ことが可能となる．

しかし，国内外問わないコミュニケーションを行う際にこの３D アバタの利用に関してはいくつかの課題を抱えている．話者の感情を３D アバタへ反映を行う際のリアルタイム性に自動翻訳を組み合わせるときにかかる時間的課題である．自動翻訳には話者がなにを話したかを認識する音声認識に話者の表情をアバタに反映させるまでの遅延，さらに翻訳した文章を読み上げるまでの時間差がリアルタイムに会話することに支障をきたすことが懸念される．

さらに話者の感情伝達にも課題が存在する．現在の多くの３Ｄアバタを利用したコミュニケーションの多くはその感情推定に視覚と音声のどちらかのみを使用している．また，読み上げ音声に関しても感情反映に使用できる感情の種類があまり豊富でないことに加え，リアルタイムでの反映を加味した場合遅延時間を考慮し，棒読みでの利用が多く行われている．

この課題を解決するため，翻訳した話者の発話文章を３D アバタが読み上げるまでの遅延時間を考慮した３D アバタへの感情反映を提案する．話者の発話感情を保存し，読み上げ音声や３D アバタに反映させることで３D アバタを使用した翻訳音声の読み上げでも適切な感情の伝達を行うことが可能となる．本稿では，発話音声を認識した際に発話感情が保存されたと仮定し，３Ｄアバタが読み上げを行うまでにかかる遅延時間を計算し，翻訳発話音声と感情の伝達までにかかるシステム速度を確認する．

本稿では，第２章で関連研究について述べる．第３章では提案システムの設計と機能，実装内容について述べ，第４章で評価実験および考察, 第５章では本稿の結論として，今後の課題と発展について述べる．

1 青山学院大学大学院理工学研究科理工学専攻
Graduate School of Science and Engineering, Aoyama Gakuin University
2 青山学院大学理工学部情報テクノロジー学科
Department of Integrated Information Technology, Aoyama Gakuin University
a) tomoya21@rcl-aoyama.jp
b) taya@it.aoyama.ac.jp
c) y.tobe@rcl-aoyama.jp

Vol.2021-HCI-195 No.26
Vol.2021-UBI-72 No.26
2021/12/1

2. 関連研究

本章では，本研究の関連研究について述べる．本研究に関連した「会話における感情推定」,「自動翻訳システム」,「対話システム」の３分野に分けて述べる．

2.1 会話における感情推定

ストレス分析やハラスメント防止など，感情のコントロールにつながる感情推定の重要性は高まっている．対人コミュニケーションを行うにあたり，状況に応じた感情推定は多く研究されているが，人間の感情への配慮を考慮した感情推定利用はあまり行われていない．インターネットを通し円滑なコミュニケーションを行うために，感情推定による考慮は重要となる．心理学において基本的な感情が特定されており，いくつかの表現モデルが提案されている[3,4]．Ekman [3]はカテゴリー知覚に関する多くの研究を行なっており，喜び，悲しみ，怒り，恐怖，嫌悪，驚きという６つの基本的感情を提案した．Russell [4]は感情を表す言葉を被験者に提示しその反応評価から”valence”，”arousal”の２つの次元を持った円環的構造を持つことを明らかにした．その後，感情のモデリングが研究者の注目を集め[5]，継続的に改良されてきた[6,7].

このような感情の定義を元に各生体情報からテキスト情報まで様々な情報から人間の感情を推定し，さまざまな用途に感情を反映できるモデルが研究されている．音声においても生の音声信号からリアルタイムで音声駆動の３Ｄフェイシャルアニメーションを実現する深層フレームワークがある[2]．DNN（Deep Neural Network）や CNN（Convolutional Neural Network）を利用し性別やアクセント，言語の異なる話者でも反映を可能としている．表情に関しても Microsoft が提供する API Face API [8]は高度な顔認識機能から顔検出や顔の特徴点抽出から分析，検出を行うことのできる機能から感情を検出することが可能となっている．テキストでは，Transformer [9]の登場から多くのテキストからの感情推定を行う研究が増え，Varsha らは入力テキストからよる細かい感情を推定できる BERT, ELECTRA ベースの知識埋め込み型 Attention による文章の感情検出を行なっている[10]．このような研究も各種会話における情報の一部を取り出し感情推定し利用を行なっているが，人間の感情はただ一つではなく，多くの感情が混ざって出現するため，正確な感情推定とはならない．

2.2 自動翻訳システム

国内外問わない対話を容易に行うために自動翻訳は重要である．Google が提供する機械翻訳[11]の技術だけでなく，DeepL 翻訳[12]やみらい翻訳[13]などの多くの機械翻訳は，他言語を容易に翻訳し理解することのできるツールとして多く発展している．研究においてもこれらの技術を活用しシステムに組み込みコミュニケーションを円滑に行うツールを作成している．Roseline らは英語から他の言語へ，またその逆に翻訳をすることができる Android ベースの言語翻訳アプリケーションを作成している[14]．IBM のリアルタイム翻訳 API の自然言語処理と Java によって作成され，ユーザはアンドロイドにテキストを入力し，言語名のついたボタンをクリックすることでテキストを変換するアプリケーションである．Santosa らは，BPPT が保有している自然言語処理技術のエンジニアリングの能力と成果のよる英語-バハサ・インドネシア語の音声翻訳システムを使用し，Android ベースの携帯電話での会話において，言語ごとに携帯端末を分けて使用する音声翻訳システムのアーキテクチャを提案した[15].

しかし，このような研究おいて Android などの手持ち型のデバイスを使用し，テキストや音声に対して翻訳のみを行い，コミュニケーションを円滑に運ぶことでは調査する間とその結果を相手に伝えるまでにラグが生じるだけでなく，そのラグにより感情を正確に伝達することが困難となる可能性がある．

2.3 対話システム

対話システムは，電話が普及されてからは特に多くのユーザに必要不可欠なものである．スマートフォンの普及により，対話システムの形は音声だけでなく，表情を伺いながらの対話を行うことができる．また，テキストベースの対話システムも多く発展しており，SNS としてその形を示している．実際に，人同士が対話を行うだけでなく３Ｄアバタやロボットを使用したシステムも多く発展している．駅や公共施設においても電光掲示板に３Ｄアバタによる案内を設置し，テキストによる対話形式で情報の案内を聞くことが可能である．ロボットにおいてもヒト型ロボットは多くが人間とのコミュニケーションを豊かにしている．多くの研究でもチャットボットやロボットによるコミュニケーションを行う対話システムが存在する．Wan らは顔の表情によるインタラクションを備えた新しいチャットロボットのモデルを開発した[16]．従来の音声アシスタントとは異なり，特定のロールプレイを用いてユーザと対話する．Sakatani らは対人関係を構築するための新しいテレコミュニケーションシステムである，AMC（Avator-Mediated Communication）システムの設計と実装を行った[17]．AMC システムとは，人間のユーザが会話相手と直接対話することなく，ソフトウェア・エージェントを介して会話を楽しむことができるシステムのことである．このシステムを使用し，ユーザの会話体験や対人関係の構築にどのように寄与するかを検証している．

Vol.2021-HCI-195 No.26
Vol.2021-UBI-72 No.26
2021/12/1

このような研究において，実際にアバタを使用することやチャットを行うにあたり感情を考慮していない課題を持つ．３Ｄアバタを使用することで調整された感情を提示することができ，感情を考慮することでより円滑なコミュニケーションを行うことが可能となる．

3. 提案システムの設計と機能

本章では，提案システムの詳細な実装アルゴリズムとその機能について述べる．

3.1 提案システムの特徴

対話システムは多種類にわたって存在する．音声翻訳からテキスト翻訳，画像を利用したチャットボットにアバタへの表情反映などさまざまである．本研究が対話システムとして他のシステムとどの程度相違があり，どの状況において適切に使用するものであるか比較する．提案システムと既存研究との機能的比較を行った結果を表１に示す．

本研究はマイクから音声信号を取得するためテキストベースの発話文章だけでなく音声の特徴も使用することが可能である．既存研究で多く存在するチャットボットと違い，テキストでの対話ではなく音声による会話を行うことで，意図された感情を表現することも可能となる．また，３Ｄアバタを使用することで反映させる感情を調整させることが可能となる．

また，感情推定，反映を行うシステムでは音声またテキスト，表情と単一の感情を使用する．人間が表す感情表現は単一にとどまるものではないため複数の感情を考慮する必要がある．

3.2 システム概要

提案システムでは，軸となるシステム「音声認識システム」，「感情推定システム」，「会話文翻訳システム」，「３Ｄアバタ読み上げシステム」の４つで構成する．マイクで入力された音声を既存の音声認識 API を用いて話者の発話文章を認識，その発話文章を翻訳した文を３Ｄアバタと読み上げソフトを組み合わせ読み上げるコミュニケーションツールである．感情はマイク入力時に入力された感情を保存し，その感情を３Ｄ アバタへと遅延させ反映させることで感情伝達を行う．図１に全体的な処理の流れを示す．

音声認識システムでは，既存の音声認識 API を使用する．認識された発話文章はファイルに保存され翻訳に使用される．

感情推定システムでは，音声や表情，発話文章から感情を推定する．推定結果は読み上げシステムに入力され，読み上げ時の感情表現に使用される．ただし，本稿では発話文章から３Ｄ アバタによる読み上げに要する時間を評価するため，実験では感情推定は行わず，あらかじめ固定値として設定したものを使用した．

会話文翻訳システムでは，音声認識システムで認識し保存された発話文章を翻訳し，翻訳した文章を読み上げシステムに渡し起動させる．翻訳に使用するシステムは既存の翻訳システムを使用する．

表１　既存研究との比較

	提案システム	VFep[1]	Midoriko Chatbot[16]	AMCシステム[17]	Teroらの感情反映システム[2]
テキスト対話型	×	×	○	×	×
音声対話型	○	○	×	○	○
リアルタイム性	△	○	○	○	○
表情感情伝達	△	×	×	×	×
音声感情伝達	△	○	×	×	○
テキスト感情伝達	△	×	○	×	×
３Dアバタの利用	○	○	○	×	○
自動翻訳	○	×	×	○	×
感情を考慮した翻訳文読み上げ	○	×	×	×	×
感情の調整	○	×	×	×	×

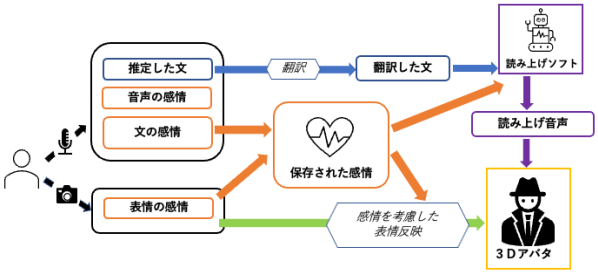


図１　提案システムの流れ

３Ｄアバタ読み上げシステムでは，会話文翻訳システムで起動させた読み上げシステムによって読み上げ音声を作成する．読み上げシステム作成のためのシステムに既存の読み上げシステムを使用する．作成された読み上げ音声を使用し３D モデルへの読み上げ動作を行わせる．この時，あらかじめ設定した感情を読み上げ音声と３D アバタの表情へと反映させる．

3.3 実装

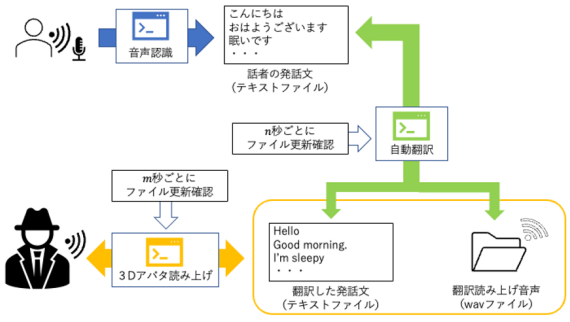
各システムを並列して動作させる必要があるためそれぞれのシステムを独立したプログラムとして実装する．各システムは３D アバタを使用するため Unity による実装を組み合わせるために C#を元に実装を行う．外部システムとして音声認識システムに搭載する API に応じて Python によるシステムも実装を行う．

3.3.1 音声認識システム

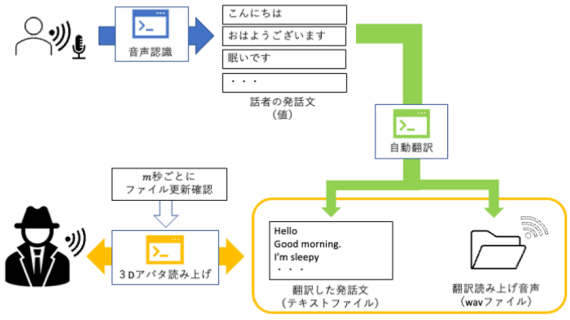
本システムでは，マイクから入力された発話文を認識するシステムを起動する．本研究では各種既存の音声認識 API である Google Cloud Speech-to-Text [18]と Unity の DictationRecognizer [19]の２つを使用し比較した．２つの API の特性からシステム構成をそれぞれに適した形で導入した．図 2 にそれぞれのシステムの流れを示す．

Google Cloud Speech-to-Text を使用した場合，提案システムにおいて外部プログラムとして稼働させるため，マイク入力された発話文を推測，推測した文章を確定させ，その時点でテキストファイルに一文ごとに改行して保存する．

Vol.2021-HCI-195 No.26
Vol.2021-UBI-72 No.26
2021/12/1



i) Google Cloud Speech-to-Text



ii) DictationRecognizer

図 2　各提案システムの機能

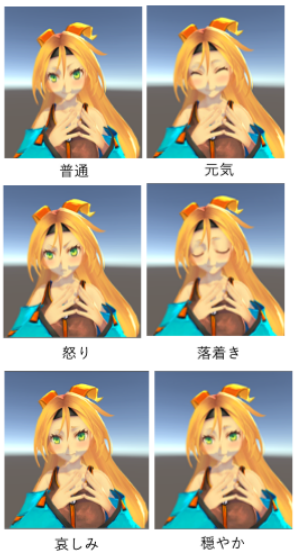


図 3　各感情における 3D アバタの表情

無言の時間が一定時間を超えると，発話文章が終わったと判定し，文章を確定する．

DictationRecognizer は，Unity の機能であり Windows の Cortana [20]のもとに稼働するため，３D 読み上げシステムへの相性からテキストファイルとして保存させず翻訳システムに直接文章を入力する．こちらも同様に，発話文章は読み終わりを判定するため，一定時間発話を行わないことで文章を確定する．

3.3.2 感情推定システム

提案システムでは音声信号，発話文章，話者の表情から感情推定を行い，その結果を統合したものを 3D アバタ読み上げシステムに入力する．推定には，[21]などで提案されている手法が利用できる．ただし，本稿では発話から読み上げまでの遅延時間を評価するため，実験においては感情推定システムの実装は行わず，理想的な推定ができたものとして評価した．

3.3.3 会話文翻訳システム

本システムでは，音声認識システムにて認識した発話文章を翻訳する．翻訳には Google Cloud Translation[22]を使用して行う．認識した発話文章は 3.2.1 にて API の利用で保存された発話文章を使用して翻訳を行う．

Google Cloud Speech-to-Text を使用した場合，認識した発話文章はテキストファイルとして保存し，各行に認識した文章があるため各行に対して翻訳を行う．発話した文章は n 秒ごとにテキストファイルの更新を確認し，更新を確認後更新された各文章に対して翻訳を行う．提案システムでは 1 秒で設定する．更新の確認するため更新のタイミングが合わない場合，翻訳までに遅延時間がかかる可能性がある．

DictationRecognizer を使用した場合，発話が終了した段階で翻訳システムに入力される．そのため更新に要する時間がかからないが，発話を連続で行うことが困難となる．

翻訳には Google Cloud Translation を使用するため翻訳したい文章を送信し，json ファイルとして結果を受信し，json ファイルから翻訳された文章を取り出す．翻訳した文章も発話文章ごとに各行で分け，テキストファイルに保存する．

3.3.4 ３Ｄアバタ読み上げシステム

本システムでは，翻訳した文章を読み上げシステムに渡し起動させる．この時，あらかじめ保存した感情を追加で設定する．読み上げには CeVIO AI 弦巻マキ[23]を使用するため外部で API として起動させ実行を行う必要がある．CeVIO AI 弦巻マキでは日本語，英語両言語が対応しているため日本から英語と英語から日本語の両方向のからの翻訳と読み上げが可能である．本稿では日本語から英語への翻訳のみを行う．読み上げる文章は 3.2.3 で翻訳した文章を保存したテキストファイルから各行ごとに読み込み読み上げ音声ファイルを作成し wav ファイルとして保存する．その際３D アバタの読み上げ動作システムはバックグラウンドで実行させる．保存されている翻訳発話文章のテキストファイルからまだ読まれていない文章を m 秒ごとに探索し，３D アバタの読み上げを実行させる．この時，読み上げ音声で設定されていた感情の表情を３D アバタに設定する．それぞれの表情に関する感情は使用する読み上げシステムに CeVIO AI 弦巻マキでプリセットとして用意されている「元気」「怒り」「落着き」「哀しみ」「穏やか」の５感情を使用する．また，会話の途切れる瞬間や感情を出さ

Vol.2021-HCI-195 No.26
Vol.2021-UBI-72 No.26
2021/12/1

ない場合を考慮して「普通」を追加し，計６感情を使用する．定義した感情と３D アバタの表情を同期させ反映させる．本研究ではユニティ・テクノロジーズ・ジャパン合同会社が提供する３D アバタにユニティちゃん[24]を使用する．顔の各パーツが個別にモデル化されているため，よりリアルな表情を表現することができる．また，提案システムではオープンライブラリであるリップシンク[25]を採用し，実際に会話をしているかのような臨場感を演出する．図 3 は，ユニティちゃんの表情から，各感情に対応する顔の表情を示したものである．

3.4 ３D アバタの読みあげタイミング

３D アバタの読み上げタイミングとして「外部プログラム方式」「内部機能方式-逐次処理」「内部機能方式-一括処理」の 3 種類を実装した．それぞれのタイムチャートとシステムの遅延時間を図 4 に示す．遅延時間の詳細は次章で述べる．

3.4.1 外部プログラム方式

本手法では音声認識システムに Google Cloud Speech-to-Text を使用する．Google Cloud Speech-to-Text は Python による外部システムのため全体的なシステムを分けて動作させるシステム構成とした．そのため，その他提案システムである会話文翻訳システムや３Dアバタ読み上げシステムも並列し処理を行うことが可能であるため処理速度の効率化を図る．

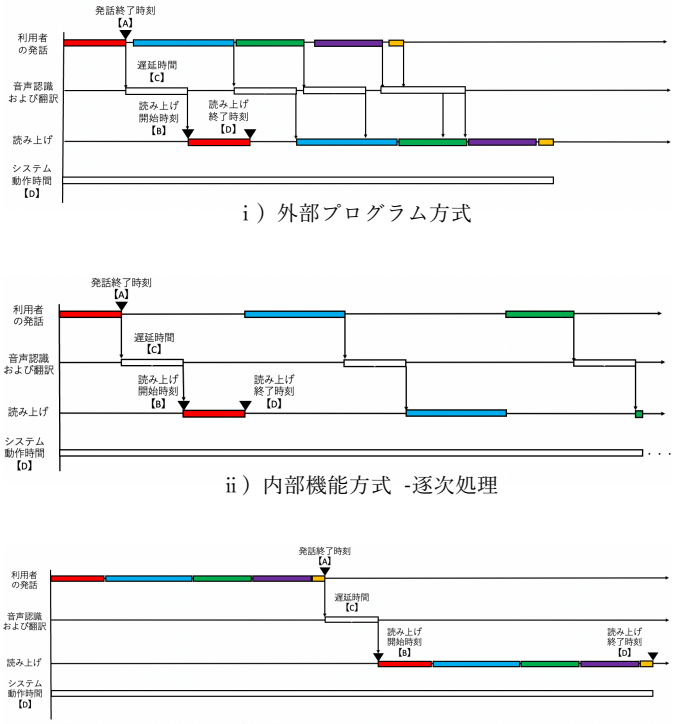
また，翻訳を行う際にテキストファイル更新を確認して行っているため，音声認識のための発話をその都度終了させる必要がないため全体的な発話を連続して行うことが可能となる．

3.4.2 内部機能方式-逐次処理

本手法では音声認識システムに DictationRecognizer を使用する．DictationRecognizer は Unity に搭載されている機能の一つであるため Unity のできるだけ単一で処理を行うことで Unity 自体の処理機能への負担を減らす必要があることから，読み上げ音声作成のためのシステム以外のすべてを Unity 上で動作させている．また，DictationRecognizer では各文に対して発話を行なった際に翻訳システムを起動させるため，各発話が終了し度音声認識を完了させた段階で待機し，その後３Dアバタの読み上げ終了を待つ必要がある．そのため逐次的な処理によるシステム動作を行う必要があり，発話を連続では行えない．

3.4.3 内部機能方式-一括処理

本手法では音声認識システムに DictationRecognizer を使用する．3.4.2 の同様のシステム動作ではあるが，音声認識までの流れを変えることで 3.4.2 がもつ発話終了後，読み上げが終了するまで待機する必要がある課題を解決させる．DictationRecognizer は発話終了を無言としてとらえその時点までの発話内容を発話文章としてシステムに渡している．そのため，発話したい内容を無言で行わない状態で連続して発話を行い，一つの複数の発話文章を一つの文章として認識させる．これにより，逐次的な処理を行う必要がなく一括で行うことで処理までの時間を短縮することが可能となる．



ｉ）外部プログラム方式

ii）内部機能方式 -逐次処理

iii）内部機能方式 -一括処理

図 4　提案手法の流れ

4. 評価実験および考察

本章では，提案システムの評価実験と考察について述べる．

4.1 実験概要

本実験では，実際にマイクに向かって発話を行いシステムの速度計測を行った．提案システムにおいて，マイクの性能は重要な要素となる．周囲の環境にノイズが多い場合，認識精度に支障をきたす可能性があるため集音性能やノイズキャンセル性能などが重要である．本研究では，これらの機能を持つマイクとして Yeti を使用して行う．単一指向性を持つため外音が多い場所でも話者の発話音声の集音を可能とする．

3.4 に示した提案システムをもとに，あらかじめ定めた５つの発話文章について発話を各発話文章に対して５回ずつ行い，各一文のみを発話した場合の読み上げまでの反映速度，５つの発話文章すべてに対して発話を行い，すべての読み上げを行った場合の２つで計測を行った．表 2 に発話を行った文章と実際に翻訳された結果の文章，また 3D アバタの読み上げに使用された翻訳読み上げ音声の再生時間を示す．

各システムの更新時間は$n = 1$, $m = 1$と設定する.

4.2 各システムの読み込み速度

表 3 に 3.4 にて記した提案手法から Google Cloud Speech-to-Text と Unity の DictationRecognizer の２つの API を搭載した提案システムによるシステムの各手法について速度計測を行った結果を示す．各数値は各発話文章に対して各文の発話を５回ずつ行いそのシステム速度のそれぞれの平均値である.遅延時間は図４に C で示した箇所であり，実際に読み上げが完了した時刻（図 4 の D）から表 2 の読み上げ音声の再生時間を引いた値を発話から読み上げ開始までにかかる時間（図４の B）とし，そこから発話文章の発話が終了した時刻（図４の A）を引いた値である.

表 4 に，各手法に対して発話開始から全ての発話文章を発話し，３D アバタの読み上げが各発話文章に対する翻訳文章全てに対して完了するまでに時刻（図 4 の E）を，システムの動作時間として結果を示す.

この時, 3.4.3 の手法では発話文は１～５まで一つの発話文として発話されているため，実際に読み上げが完了した時刻である D とシステムの動作時間である E は同じ計測結果を持つ.

4.3 結果と考察

各発話文章を一つずつ発話した場合，外部プログラム方式と内部機能方式（逐次処理）の提案手法から結果を比較する.

表 3 の結果から，Google Cloud Speech-to-Text を使用している外部プログラム方式では各遅延時間は約 4~5 秒かかり，DictationRecognizer を使用している内部機能方式（逐次処理）では約 2.5~3.5 秒かかる結果となった．このことから DictationRecognizer の使用した場合の提案システムが遅延時間を少なくなるということがわかる．また，発話開始から読み上げ開始までにかかる時間にしても外部プログラム方式と内部機能方式（逐次処理）では約２秒の差がみられる．これらの理由として DictationRecognizer は Google Cloud Speech-to-Text と違い発話文章の更新を行わないだけでなく，もともと Unity の機能であるためシステムの組み合わせとして相性のよいものであり，３D アバタへの反映が行いやすいことがあげられる．しかし，欠点として提案システムに共有の機能である，外部システムである翻訳文読み上げ音声作成のシステムをバックグラウンドで展開し，交互に実行をおこなっているため，一連の実行の流れにより Unity の音声認識システムに負荷がかかり，プログラムが異常終了しやすい点があった．また，DictationRecognizer は Google Cloud Speech-to-Text と比較し活舌が悪いと誤認識をしやすい傾向もわかった.

発話から３D アバタ読み上げ終了までのシステム全体の時刻では, 外部プログラム方式と内部機能方式(逐次処理)の提案手法だけでなく，内部機能方式（一括処理）からも結果を比較する.

表２　各発話文章に対する翻訳結果

No.	発話文	翻訳文	翻訳文の読み上げ時刻(秒)
1	こんにちは	Hello	0.85
2	今日は良い天気ですね	The weather is good today, is not it	2.33
3	すこし肌寒くなりましたね	It's getting a little chilly	1.69
4	明日は雨が降るそうなので傘があるとよいと思います	It's going to rain tomorrow so I hope you have an umbrella	3.83
5	さようなら	good bye	0.88

表３　各発話文章に対する遅延時間

i) 外部プログラム方式

No.	発話終了時刻（図４, A）	読み上げ開始までの時刻（図４, B）	遅延時間（図４, C）	読み上げ完了時刻（図４, D）
1	2.89	5.52	4.67	8.41
2	3.23	7.71	5.38	10.94
3	3.60	5.99	4.30	9.59
4	5.48	7.68	3.85	13.16
5	2.68	5.98	5.10	8.66

ii) 内部機能方式 –逐次処理

No.	発話終了時刻（図４, A）	読み上げ開始までの時刻（図４, B）	遅延時間（図４, C）	読み上げ完了時刻（図４, D）
1	1.60	3.99	3.14	5.59
2	2.42	5.46	3.13	7.88
3	2.76	5.14	3.45	7.90
4	5.08	6.49	2.66	11.57
5	1.50	4.31	3.43	5.81

iii) 内部機能方式 –一括処理

No.	発話終了時刻（図４, A）	読み上げ開始までの時刻（図４, B）	遅延時間（図４, C）	読み上げ完了時刻（図４, D）
1-5	9.55	12.60	3.06	22.18

表４　提案システムの動作時間

外部プログラム方式	内部機能方式 逐次処理	内部機能方式 一括処理
26.88	36.89	22.18

表 4 から，内部機能方式（一括処理）が約 22 秒と全体的なシステム動作としてもっとも動作時刻が短い結果であることがわかる．また，内部機能方式（逐次処理）が３６秒と終了までに一番遅い結果であることもわかる．これは内部機能方式（逐次処理）では一度の発話を行うまでに終了

を待つ必要があるため，発話を被せることが困難な点であると考えられる．また，外部プログラム方式においても更新時間によるラグにより，発話の終了から読み上げまでに多少の誤差を生じてしまうことからであると考えられる．

発話から読み上げまでの開始時間（図４の B）を比較すると,最初の読み上げまでにかかる時間は内部機能方式(一括処理）が最も時間がかかることがわかる．外部プログラム方式と内部機能方式（逐次処理）では読み上げ開始までの時間はあまり差がない結果であることもわかる．

内部機能方式（一括処理）は発話文章を一括で処理させるため無言のタイミングを発生させない方式である．実際に対話を行うにあたり無言状態が必ず生じてしまうため，常に内部機能方式(一括処理)を使用することはできない．そのため，連続かつ単文章ごとの対話においては外部プログラム方式がもっともよい結果であると考える．

表情反映の観点では，今回固定値として設定していたため表情反映におけるラグは存在しなかった．また，表情変化も用意されている３D アバタの各表情を変化させて行なっているため，スムーズな表情変化を行うことが可能であった．ここで実際に感情推定を行いその感情に応じた表情に反映させるには学習を行う必要があるため，その場合に生じる遅延時間も検証する必要がある．

5. 結論

本稿では，翻訳した話者の発話文章を３D アバタが読み上げるまでの遅延時間を考慮した３D アバタへの感情反映を提案した．提案システムでは，音声認識した発話文章を保存し３Dアバタに読みあげを行うまでにかかる遅延時間を計算し，どの程度感情伝達までにかかる遅延時間を考慮，改善できるかを確認した．現在のところ遅延時間は大きく，かつ使用する音声認識によっても異なる動作が行われている．今後の展望としてシステムだけなくオーディオインターフェースなどのハードウェアによる制御も視野にいれ改善を検討している．

参考文献

[1] Suzuki, T., Taya, A., and Tobe, Y.: VFep: 3D Graphic Face Representation Based on Voice-based Emotion Recognition, 5th Workshop on emotion awareness for pervasive computing beyond traditional approache(2021).
[2] Karras, T., et al.: Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion, ACM Trans. Graph, Vol.36, No.4, Article 94(2017).
[3] Ekman, P.: An Argument for Basic Emotions, Cognition and Emotion,Vol 6, No.3-4, pp.169-200(1992).
[4] Russell, J. A.: A Circumplex Model of Affect, Journal of Personality and Social Psychology, 39, pp.1161-1178(1980).
[5] Pfeifer, R.: Artifical Intelligence Models of Emotion, Cognitive perspectives on emotion and motivation, Springer, pp. 287 - 320 (1988).
[6] Marsella, S., Gratch, J., and Petta, P.: Computational Models of Emotion, Blueprint for Affective Computing, Oxford University Press(2010).
[7] Luis-Felipe, R. and Felix, R.: Development of Computational Models of Emotions for Autonomous Agents: A Review, Cognitive Computation Vol. 6, No.3, pp. 351 – 375(2014).
[8] Microsoft : Face API, Microsoft Azure，入手先〈https://azure.microsoft.com/ja-jp/services/cognitive-services/face/〉
[9] Vaswani, A., et al.: Attention is all you need, Advances in neural information processing systems(2017).
[10] Suresh, V., and Desmond C. O.: Using Knowledge-Embedded Attention to Augment Pre-trained Language Models for Fine-Grained Emotion Recognition, arXiv preprint arXiv:2108.00194 (2021).
[11] Google : Google 翻訳 〈https://translate.google.co.jp/〉
[12] DeepL GmbH : DeepL 翻訳 〈https://www.deepl.com/translator〉
[13] 株式会社みらい翻訳 : みらい翻訳〈https://miraitranslate.com/〉
[14] Ogundokun, R. O., et al.: An android based language translator application, Journal of Physics: Conference Series. Vol. 1767. No. 1. IOP Publishing(2021).
[15] Santosa, A., et al.: The Architecture of Speech-to-Speech Translator for Mobile Conversation, 2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA). IEEE(2019).
[16] Wan, Y., et al.: Midoriko Chatbot: LSTM-Based Emotional 3D Avatar, 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE). IEEE(2019).
[17] Sakatani, Y., et al.: An Avatar-Mediated Communication System for the Construction of Interpersonal Relationships, 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE(2018).
[18] Google : Google Cloud Speech-to-Text, Google Cloud，入手先〈https://cloud.google.com/speech-to-text〉
[19] Unity : DictationRecognizer, Unity(オンライン)，入手先〈https://unity3d.com/jp/legal?_ga=2.182475942.1415518277.1635917999-178299773.1635917999〉
[20] Microsoft : Cortana〈https://support.microsoft.com/en-us/topic/what-is-cortana-953e648d-5668-e017-1341-7f26f7d0f825〉
[21] Mittal, T., et al.: M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 02.(2020).
[22] Google : Google Cloud Translation, Google Cloud，入手先〈https://cloud.google.com/translate〉
[23] 株式会社 AHS : CeVIO AI 弦巻マキ トークボイス，株式会社 AHS，入手先〈https://www.ah-soft.com/commercial/cevio/〉
[24] Unity Technologies Japan:UNITY-CHAN!,Unity Technologies Japan(オンライン) ,入手先〈https://unity-chan.com/contents/guideline/〉
[25] Oculus:Oculus LipSync Unity,Oculus(オンライン) ,入手先〈https://developer.oculus.com/downloads/package/oculus-lipsync-unity/〉