音声を用いたレトロフィットIoTの開発

小山洋太*1
鴨井督*2
水上莉沙*3

［要　旨］

本研究では、情報ポートを持たない機器のIoTの手法として、音声認識と画像認識の２つの技術を融合したシステムを２例作成し、検討を進めた。結果、両者技術がお互いの短所を補完することで片方の技術だけの場合よりも安定したシステムが構築できることが認められた。しかしながら、音声と画像のどちらからの情報入力を採用するかについては、システムの状況に応じて、検討することが重要であることが明らかとなった。

１　はじめに

近年、ChatGPTなど、目覚ましく発展するAI技術であるが、そこではデータの存在がかなめであり、IoT技術に代表されるような、あらゆるデータを電子化し蓄積することがますます重要となってきている。しかしながら、例えば当センターの測定機器などにおいても、十分な通信ポートを持たない、更には完全に機械的な装置であり、電子的な要素を持たない装置など、アナログ的な機器も多く、このような機器のIoT化は「レトロフィットIoT」と言われ、重要な課題となっている。これまで筆者は、先行する研究にて、比較的利用しやすいリソースを用いた方法で、製造現場における音声の活用[1)]、および画像認識の活用[2)]について検討を進め、人の音声やジェスチャーによりある程度の情報入力が可能であることを明らかにしてきたが、これら技術を用いて、人が介在することにより、レトロフィットIoTが実現できることが十分期待される。しかしながら、音声認識においては、外乱音に対する過認識が課題としてあった一方、画像認識については、細かい判定を行う難しさから、入力できる情報数の制限が見込まれるところであった。そこで本研究では、人が介在したレトロフィットIoTの可能性を確認するべく、これら２つの技術を融合し、各々の短所を補完しあう形で、入出力を持たない機器に対してでも活用可能なシステムの検討を進めた。

２　システム構築について

２．１　音声認識と画像認識

２．１．１　汎用大語彙連続音声認識エンジン「Julius」について[3~5)]

音声認識エンジンとして、本研究では先行する研究と同様、無料利用可能な汎用大語彙連続音声認識エンジン「Julius」を使用することとした。同エンジンは、名古屋工業大学のチームにより現在、開発・管理が進められているエンジンであり、①Linux / Windows / MacOSなど、対応するプラットフォームの幅が広い。②エンジン起動後は音声データを入力するだけで解析結果が出力される。など、汎用性高く簡便に解析することが可能となっている。この度も、文法や単語を事前に指定

＊１　応用技術課　副主査
＊２　応用技術課　主任研究員
＊３　応用技術課　技師

する「記述文法音声認識実行キット」を使用した。

２．１．２　画像認識モジュール「MediaPipe」[6]について

画像認識を行うモジュールについて、こちらも本研究では先行する研究と同様、Alphabet 社（Google）の「MediaPipe」を使用することとした。MediaPipe においては、物体の検出や追跡などの様々な機械学習ベースのソリューションが提供されており、特に人の特徴点（ランドマーク）を取得するソリューションが豊かである。

本研究では、「顔のランドマーク取得（FaceMesh)」、「手のランドマーク取得（Hands)」、そして「体のランドマーク取得（Pose)」の３つを適宜用いた。

２．２　作成したシステムについて

２．２．１　薬品管理音声システム

この度作成したシステムは２種あり、１つが現在、当センターで活用している「薬品管理システム」[7]（図１）の音声補助システムである。

同システムは、薬品使用量の管理をするため、天秤の計量結果を RS-232C により、可搬な端末機である「Raspberry Pi」へデータ登録をする仕組みとなっているが、通信ポートのないアナログ的な天秤でもデータ登録が可能なシステムとして音声とジェスチャーによる登録システムを検討した。なお、端末機は本来、上述のとおり「Raspberry Pi」であるが、ここではデスクトップ型PC（CPU：Intel Core i5@3.20GHz、メモリ：4GB）に、「Ubuntu」（バージョン：20.04）をインストールし、作成した。

２．２．２　測定機データ取得補助システム

もう１つのシステムは、当センター保有の試験装置であるベクトルネットワークアナライザを用いた自由空間法といわれるミリ波性能評価のデータ取得の補助システムである。自由空間法のセットアップを図２に示す。画面中央にある金属製の台が試料をセットする箇所であり、奥に見える測定器左側にある PC がデータ取得用の PC である。

試験中は、試料のセットとデータの取得のそれぞれの作業のため、測定者はこの２つの間を往来することになるが、この度、試料をセットする位置から音声とジェスチャーで PC に指示を与えるシステムを検討した。なお、こちらでは Windows 機（OS：Windows10、CPU：Intel Core i5@2.30GHz、メモリ：8GB）での開発を行った。

図１　薬品管理システム

図２　自由空間法のセットアップ

２．２．３　その他

システムの開発は python3 を使用して行った。また、音声の取得には SeeedTechnology 社の「ReSpeaker Mic Array v2.0)」を使用した。その他、測定器データ取得補助システムについては、測定器からデータを取得する既存の VBA プログラムと連動させるため、「xlwings」モジュールを活用した。

３　最終構築システム

３．１　薬品管理音声システム

図３はこの度作成した薬品管理音声システムのユーザインターフェース(UI)画面である。画面左上にカメラによる撮像状況を映しており、顔の有無を判定している。画面中央には、薬品管理システム上で登録を必要とする、測定の「場所」、「天秤」の管理番号、「利用者」の指名、薬品の使用・棚卸など測定の種別を示す「所作」、薬品「瓶」の管理番号、そして天秤で計測した「重量」の項目欄となっている。

各項目は、図４に示す親指と人差し指だけを伸ばしたジェスチャーで繰ることができ、人差し指の向きを変えることで、繰る方向を変更することができる。また、選択されている項目は赤くハイライトされる。音声の入力は図５に示す親指と小指だけを伸ばしたジェスチャーで認識し、そのジェスチャーが確認されると、右側に「聞き取り中・・・」という音声認識を開始した旨を知らせる文言が現れる。

図３　薬品管理音声システムのUI

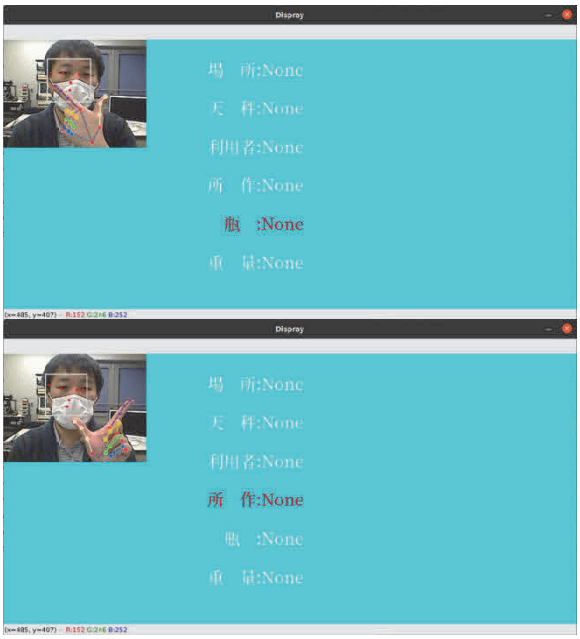

図４　項目移動のジェスチャー

図５　音声認識のジェスチャー

その後、音声が発話され、発話が終わったことを認識し、「Julius」による音声認識が実行される。なお、その際の辞書・文法は項目ごとに準備し、選択されている項目に合わせて、読み込まれる仕組みとしている。また、重量の入力に際しては、ここでは単数字読み上げ（例えば 323.3g の場合「さん・にー・さん・てん・さん」と読み上げる。）にて実装している。

必要事項を入力した後、データの登録を行うが、ここではそのジェスチャーとして、データの登録

図６　データ登録のジェスチャー

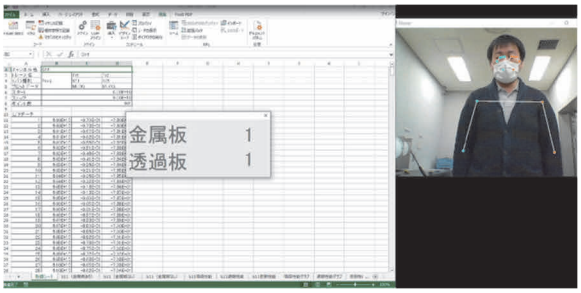

図７　測定器データ取得補助システムのUI

という重要性の高い行為であるため、両手を使った図６のジェスチャー（片方の手の甲に、もう片方の人差し指を突き立てる。）を採用し、これを５秒間留めることで実装した。

以上が、この度作成した薬品管理音声システムである。ジェスチャーと連動させたため、音声の過認識の課題には対応できているところであるが、実際の試行においては、単数字読み上げで認識が難しい場面があったものの、注意しながら再度読み上げた場合は認識が可能であった。また、先行する研究[2]では、すべての入力を音声で実施したが、比較して、ジェスチャー認識を援用しすぎるとわずらわしさ（例えば、項目の選択）が目立つ点もある。このため、修正の利く入力等については音声で実施し、音声認識のトリガやデータの登録など不可逆的なところについて、ジェスチャーを援用することが望ましいと思われる。

３．２　測定器データ取得補助システム

図７はこの度作成した「測定器データ取得補助システム」のUIである。ここではエクセルと連動させており、エクセルには取得したデータ列の表示と、測定モード（本測定には金属板を使用したモードと穴の開いた透過板を使用したモードが存在する。）のポップアップが表示されている。また、画面右側にはカメラ像が移っており、顔と体の認識が行われている。

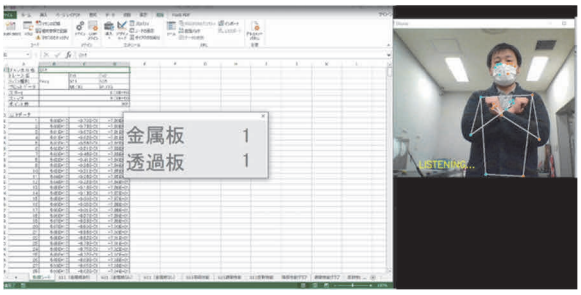

図８　音声認識のジェスチャー

このシステムでは、前述の「薬品管理音声システム」とは異なり、ジェスチャーは音声認識のトリガとしてだけ援用している。そのトリガは図８に示すように、カメラに正対したうえ、腕をクロスするジェスチャーを採用している。

図８のジェスチャーを認識すると、カメラ像の左下に「LISTENING...」との表示が現れ、音声の認識・Juliusでの解析と処理が流れていく。

図９はシステム起動中のサンプル取り換え時の状況であるが、ここでは腕がクロスされず、また上半身が作る四角（左右の肩と腰の位置）内に収まっていないため、前述の認識状態とは判断されず、音声認識は行われない。また、追加での過認識防護策として、図８でポップアップの文字が灰色になっているところ、図９のポップアップでは一部の文字が黒くなっている。これは測定のモードを表しているが、同時に、システムの起動を示しており、すべて灰色の際はシステム起動の言葉「音声認識を開始」というワード以外に反応しな

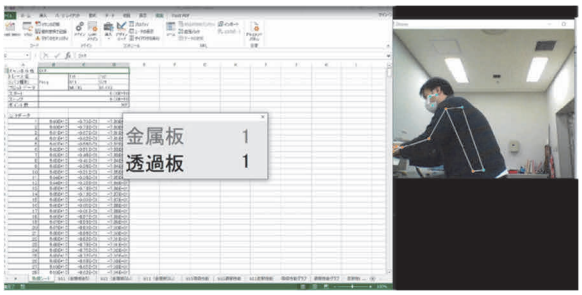

図９　サンプル取り換え時の状況

いようにしている。これにより、例えば測定間の打ち合わせなど、測定を実施していない状態でシステムをアイドル状態にすることが可能である。

このシステムでは前述の「薬品管理音声システム」と比較して、音声認識のトリガとしてのみジェスチャーを使用した。他方で、このシステムでは測定モードが２つだけであるため、ジェスチャーによる切り替えの方が望ましいとも考えられる。

４　まとめ

本研究では、通信ポートを持たない機器の IoT 化というレトロフィット IoT 背景から、音声と画像を用いた入出力方法を、２つのシステム例を作成し検討した。結果、機器の通信ポートを使用せず、音声による情報入力の実現とその制御としての画像認識（ジェスチャー）の組み合わせが一定効果的であることが確認できた。

しかしながら、音声と画像のどちらに役割を配分するかは重要なポイントであり、選択肢の多さやその所作の重要性から、どちらにどの役割を持たせるかは状況に応じて検討する必要があると考えられる。また、そのシステムの置かれる状況（音の状況や人の往来）などによっても、どちらからの情報を主とするかについても検討する点が多いと考えられる。

（参考文献）

１）小山洋太，他：京都府中小企業技術センター技報，No.49, p1（2021）

２）小山洋太，他：京都府中小企業技術センター技報，No.50, p21（2022）

３）A. Lee and T. Kawahara: Julius v4.5 (2019) https://doi.org/10.5281/zenodo.2530395

４）A. Lee, T. Kawahara and K. Shikano. ″Julius ―― An Open Source Real-Time Large Vocabulary Recognition Engine″. In Proc. EUROSPEECH, pp. 1691―1694, 2001.

５）A. Lee and T. Kawahara. ″Recent Development of Open-Source Speech Recognition Engine Julius″ Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2009.

６）https://google.github.io/mediapipe/

７）水上 莉沙：クリエイティブ京都M&T No. 173, p22（2022）