ICT 技術を活用した企業業務補助システムの開発

（音声入力型在庫管理システムの検討）

小 山 洋 太*1
谷 田 幸 宏*2

［要　　旨］

本研究では、フリーの日本語音声認識エンジンを援用し、在庫管理を音声入力で行うシステムの開発を試み、ものづくり現場での音声入力の活用について検討を行った。結果、日本語の認識性については問題がなかったものの、外乱音に反応する課題が認められたため、マイクアレイユニット及びUSB カメラを援用し、外乱音に耐性を持つ試作機の開発に至った。

１　はじめに

新型コロナウイルス感染症の影響により、接触の機会をできうる限り減らしたニューノーマル社会に対応したインフラや機器の需要が高まっている。このような状況では現在、空中ディスプレイを用いたタッチパネル様の制御システム[1)]、及び顔などの画像を用いた認証などが注目されている。他方で、スマートスピーカーなどに代表される音声の入力についても、非接触入力として非常に有効であると考えられるが、ものづくり現場においての展開は少なく、しかもインカムなどの補助具を使う場合が多い。

そこで本研究では、ものづくりの現場で容易に導入できるよう、廉価で簡便に構築することができる音声を使った入力システム構築の可能性を検証するべく、在庫を管理するシステムを念頭に音声入力を擁するシステムの構築を試みた。

２　システム構築について

２．１　音声認識エンジンの選定と性能検証

２．１．１　汎用大語彙連続音声認識エンジン「Julius」について[2-4)]

音声認識エンジンとして何を使用するかは本研究の重要な部分となるが、ここでは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。ここでは「Julius」のバージョンは4.4.2を用いた。

２．１．２　性能の検証（その１）

この「Julius」の認識性能について、検証を行った。まず、「Julius」の「音声認識パッケージ　ディクテーション（自動口述筆記）キット」（以下、「ディクテーションキット」と言う。）を使用した場合での検証を行った。このキットは日本語の自由な発話を解析し文字に変換する標準で準備されているキットであり、ここではバージョンとして「dictation-kit-v4.5」を用いた。また、検証に当たっては、ディクテーションキットに標準で備わっている設定ファイル及び混合ガウスモデル（GMM）の音響モデルを用いた。

＊１　応用技術課　副主査
＊２　企画連携課　技師
（現　企画連携課　主任）

検証対象としたフレーズは「モード 変更 状態 へ 遷移」、「担当 小山」という２つのフレーズを採用し、各 100 回の認識を試みた。また、認証に失敗した場合においては、次の試行において明瞭な発話を心がけることとし、その時の認識の成否（言い直し）についても検証を行った。

表１に結果を示す。結果として、まずどちらのフレーズについても、言い直しを行っても大きな認識率の向上は認められなかった。また、「モード変更状態へ遷移」というフレーズについては 35%ほどの認識率であるものの、「担当小山」というフレーズについては 11%と低いものとなった。

２．１．３　性能の検証（その２）

続いて「Julius」の「記述文法音声認識実行キット」（以下「記述キット」と言う。）を使用した場合での検証を行った。このキットは、あらかじめ認識対象となる単語及びその並び（文法）をユーザーで設定し、その条件下で文字認識を行うものである。今回、設定した単語及び文法の概要を図１に示す。登録した単語としては、担当者・場所・型番の具体、および数字（100 まで）、そして命令としての補助的なキーワードである。また、文法についてはこれらの並びを指定した。

この状況において、登録したものの中から「モード変更状態へ遷移」及び「担当小山」（２．１．２と同様）のフレーズについて、その認識理率を検証した結果が表２である。ここではバージョンとして「grammar-kit-4.4」を用いた。

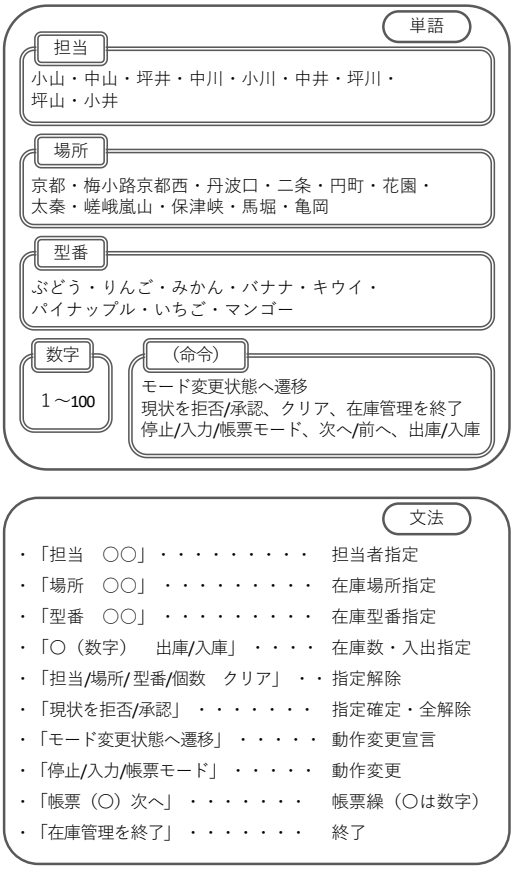

図１　登録した単語と文法

表２は「記述キット」による結果である。結果に示すとおり「モード変更状態へ遷移」というフレーズについては 100%の認識率となった。他方で「担当小山」については５%ほどの誤認識が発生したが、言い直しを行った場合は、100%の認識率となった。

以上より、在庫管理システムを念頭に置いた特定のキーワード認識によって目的が達成されるシステムについては、記述キットを利用することで十分使用に耐えうる認識率となることが分かった。

表１　ディクテーションキットによる認識率

フレーズ	１回目	２回目（言い直し）
「モード変更状態へ遷移」	35.0%	33.8%
「担当 小山」	11.0%	11.2%

表２　記述キットによる認識率

フレーズ	１回目	２回目（言い直し）
「モード変更状態へ遷移」	100%	N/A
「担当 小山」	95.0%	100%

２．２　外乱音に対する反応と対応

２．２．１　外乱音に対する反応

２．１で述べたとおり、記述キットを使用することで目途とするシステムを構成するための音声認識は認識率としては問題ないことが判明したが、他方で、この検証の過程において、外乱音（発話者以外の発話や音）に反応し音声認識が実施されてしまう現象が散発した（以下、便宜上「過認識」と言う）。

この過認識については、「Julius」の設定やモデル等の変更で対応できる可能性もあるが、ここでは外部的な補助装置によって取り除けないかを検討する。

２．２．２　過認識対策その１（音声方向）

過認識の対策として、最も単純なものは音声の到来方向を検知することが考えられる。このためには、複数のマイクを設置し、その音声の位相差（距離）から到来方向を演算することが考えられる。このためには、短時間フーリエ変換やフィルタ処理などデジタル音声処理の実装が必要となってくるが、ここでは市販品のマイクアレイユニットによりこの課題の解決を試みた。

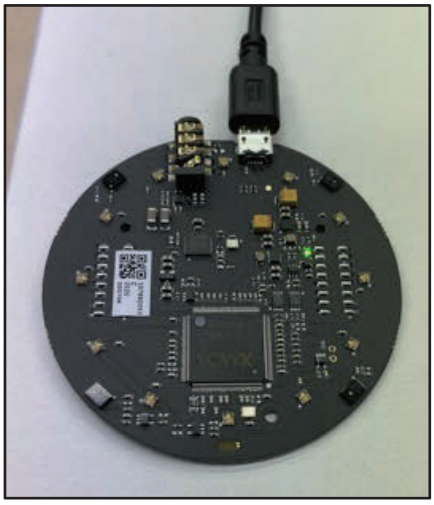

図２　スマートスピーカー基板

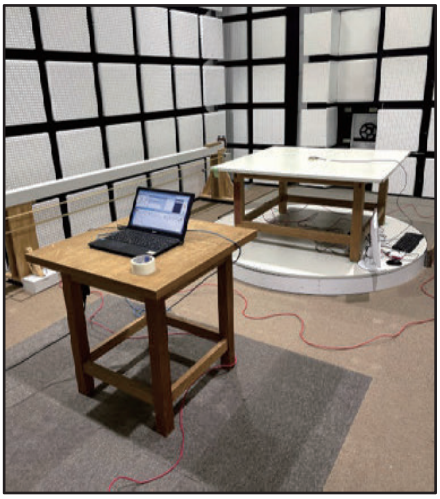

図３　確認セットアップ

本研究では、マイクアレイユニットとしてSeeedTechnology 社の「ReSpeaker Mic Array v2.0」[5]（図２）を採用した。比較的廉価な製品であり、Python を用いた制御についてのチュートリアルもあるため導入についての障壁が低く、音声到来方向の検知のみならず、発話の有無やゲイン調整等、様々な調整が行える。

採用した基板の簡単な性能評価として、図３に示すセットアップにて到来方向の角度について確認した。図３のセットアップは、比較的音響反射が少ないと考えられる当センターの電波暗室内において、回転テーブルの中心にマイクアレイユニットを設置し、そこからおよそ３m離れた位置に発話者の立ち位置を定め、回転テーブルを１度刻みで回転させた。この回転時の各角度にて「モード変更状態へ遷移」というフレーズを複数回発話し、その際に取得された角度の最頻値と実角度との誤差を図４に示す。図４のとおり、最大５度の角度の誤差が発生しているが、おおまかな到来方向については十分認識できていることが確認でき、このことから、およそ数十度くらいの幅で認識範囲を定めれば、システムに正対した発話者の音声を認識できるものと考えられる。

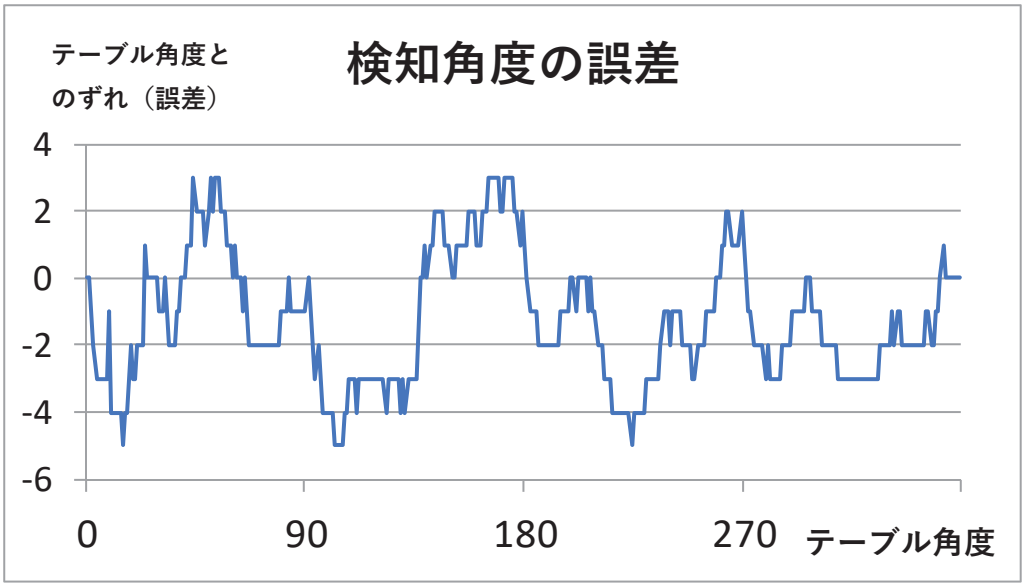

図４　各角度における誤差（ずれ）

２．２．３　過認識対策その２（顔認識）

過認識へのもう１つの対策としては、所定の位置に発話者の存在を検査する方法が考えられる。

そこで、今回は汎用的なUSBカメラとIntel社が開発・公開している画像向けライブラリであるOpenCVを援用した顔認識を採用した。ここでUSBカメラの画角の映像から、顔と思われる部分をOpenCVにあらかじめ準備されているHaar-like特徴分類器[6]「haarcascade_frontalface_alt2」により認識を行うこととした。実際の認識状況のキャプチャ画像を図５に示す。

なお、今回はほぼ無調整で実施しているが、画角のトリミングや認識される顔の大きさを調整することで、顔を認識する位置（範囲）をある程度限定することも可能と思われる。

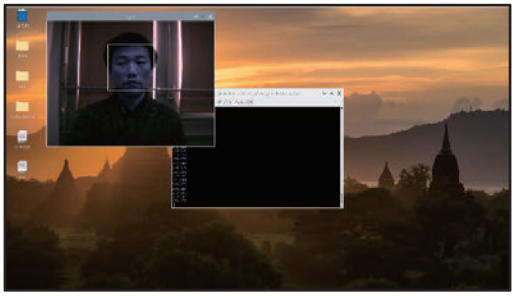

図５　OpenCVによる顔認識の試行

また、必要に応じては、背景にホワイトボードなどを置く、もしくは画角として壁面を捕らえるなど、誤認識を防ぐ方策を施すことも可能である。実際に顔認識を実行した状況を図５に示す。なお、OpenCVのバージョンは4.5.1.48を利用した。

２．３　最終試作版について

以上、過認識への対応を実装し、この度は試作機を開発した。（図６）

試作機のハードウェア的構成は

①本体（ラズベリーパイ）

②USBカメラ

③マイクアレイユニット

となっており、②、③はUSBケーブルにより本体に接続されている。

また、ソフトウェアとしては「Julius」をサーバライクに動作させるためにWebアプリケーションフレームワークの「Cherrypy（ver. 3.8.1）」を利用し、Webサーバへのデータポスト・解析結果返しという形式を採用した。その他、ユーザーインターフェース（UI）はWeb

図６　在庫管理システムの試作機

ブラウザにより実装しており、ブラウザを Python で制御するために selenium(ver. 3.141.0) を援用している。入力された在庫データを管理するデータベースには SQLite を使用している。

図７に、この度の試作機の簡単な動作の流れを示す。

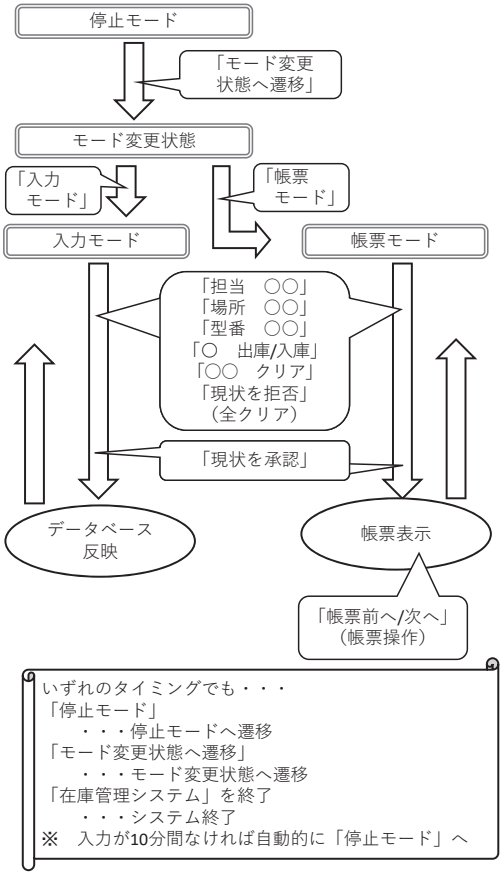

図７　動作の流れ

基本的にシステムは起動中、「停止モード」の状態にあり、「モード変更状態へ遷移」というフレーズ以外には応答しない状況となっている。また、「モード変更状態へ遷移」後はモード入力待ち状態となり、この時は「入力モード」「帳票モード」そして「停止モード」の３つのフレーズにしか反応しない。これにより不用意な外乱音により機器が誤作動することを防いでいるものである。また、同様の対策として、「入力モード」でも「帳票モード」でも、発話者は、型番や場所・担当者などのワードを発話して入力することになるが、データベースへの反映や帳票の出力は「現状を承認」というフレーズが発話されるまで実行されず、何度でも言い直すことができる仕組みとしている。

各モードの UI の画面状態を図８、図９に示す。

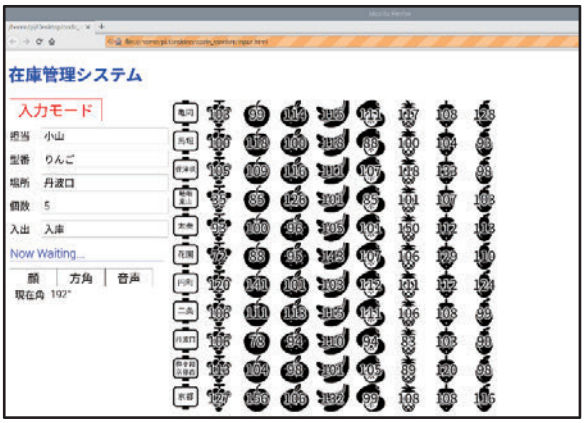

図８　UI（入力モード）

図９　UI（帳票モード）

どちらも画面左側に発話者が入力した情報及び各種過認識対策として導入した検知内容の検知状況（顔、音声、方向）についての表示部が左側にある。「入力モード」では画面右側には場所とその場所の各型式の在庫数を示している。「帳票モード」では一覧表が右側に映り、場所・型番を特定しきらない場合（例えば場所だけ特定）は、対象となるものの総和（特定された場所の各型番の総数）を表示し、両者ともが特定された場合は、その場所・型式の時系列的な入出庫記録が表示される仕組みとしている。なお、表は10行しかないが、「帳票次へ/前へ」というフレーズで次/前データに表示が切り替わる仕組みを実装した。（この際「次へ/前へ」の前に数字を述べるとその数字分、ページが移動する。）

また、いずれのタイミングにおいても、「停止モード」、「モード変更状態へ遷移」というキーワードは認識するようにしており、10 分間の入力がない状態が続けば、自動的に停止モードへ移行するようにしている。

本システムのセットアップ状況を図 10 に示す。UI の表示はプロジェクタにより実施しており、発話者の前にカメラとマイクアレイユニットを準備している。また、発話者の正面にはカメラとマイクアレイユニットを準備し、スピーカーの感知角度は正面に対して左右20度（全方位角40度）

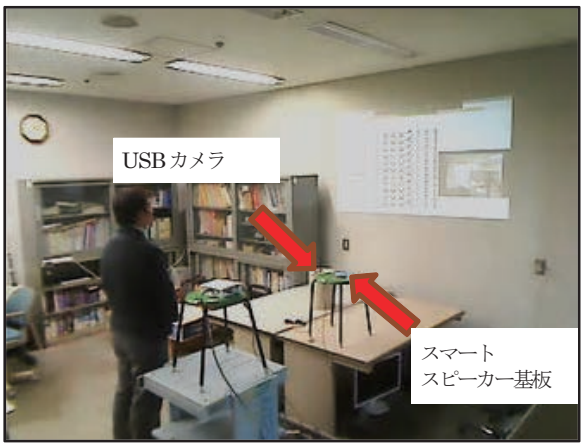

図10　システムのセットアップ状況

としている。構築した場所は当センター研究室内の静かな場所ではあったが、発話認識・顔認識ともに問題なく運用が出来た。

3　まとめ

本研究では、ものづくりの現場で容易に導入できるよう、廉価で簡便に構築することができる音声を使った入力を擁するシステム構築の可能性を検証するべく、在庫を管理するシステムを念頭に音声入力システムの構築を試みた。音声認識部については、無料利用可能な汎用大語彙連続音声認識エンジン「Julius」の利用を検討したところ、記述キットを利用することで、目的とする発話認識をストレスなく認識することを確認した。しかしながら、外乱音に反応する過認識が問題としてあり、この点について、市販のマイクアレイユニットとUSBカメラ像からの顔認識を援用することにより外乱音に対する耐性を高める実装を行い、試作機の作製を完了した。

この度の研究より、特定キーワードを認識させる形で、少なくとも静音下ではなんら問題のない音声により入力システムが簡便に構成できる可能性が認められたため、在庫管理システム以外の展開も大きく期待が持てるものとの結論に至った。

（参考文献）

1）谷口　友一，他：京都府中小企業技術センター技報，No47，p31（2019）

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

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

4）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.

5）ReSpeaker Mic Array v2.0 Wiki https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/

6）カスケード型分類器（Haar Feature-based Cascade Classifier for Object Detec-tion) http://opencv.jp/opencv2.2/c/objdetect_cascade_classification.html