天井設置型の複数マイクと話者の位置情報を用いた発話方向推定手法

市川　直人[1]　戸出　悠太[3]　片山　晋[1]　浦野　健太[1]　米澤　拓郎[1]　河口　信夫[2,1]

概要：屋内における人の位置と顔向きを推定するためにカメラや距離センサを用いた手法が研究されている．しかし, これらの技術は利用者へのプライバシーに関する心理的負荷, 障害物・人による死角, 高コスト化などの問題を抱えている．そこで本研究では, 赤外線センサを用いた位置情報の取得を前提として, 天井に設置した複数マイクの音声を用いた発話方向推定手法を提案する．まず発話方向による音声信号の音響パワーの違いに注目して, フーリエ変換を用いて特定周波数帯における音響パワー比を算出した．さらに一定時間幅で平均と分散を取り, 8 次元ベクトルとして SVM(Support Vector Machine) に与え, 8 方位にラベリングした発話方向の推定を行った．実際に, $86.4m^2$ の屋内で約 56 分の複数の被験者の音声データを収集し, 1.2 秒の音声信号に対して最大 84%で発話方向推定ができることを確認した．さらに人による学習データ収集の負担を軽減し, 一定環境でのデータ収集を可能とするため, 声帯模型を用いた擬似発話音声データの収集を行い, 検証を行った．

Speech Direction Estimation Method using Multiple Microphones Installed on the Ceiling and Location Information of Speaker

Naoto Ichikawa[1]　Yuta Toide[3]　Shin Katayama[1]　Kenta Urano[1]　Takuro Yonezawa[1]　Nobuo Kawaguchi[2,1]

1. はじめに

人の位置情報は, ナビゲーション, 店舗分析, 感染症対策, セキュリティ管理など多様な分野で利用されている．特に屋内における粒度の細かい位置情報は, スマートホーム・スマートシティ, 倉庫・オフィスの作業最適化システム, 介護・医療施設の見守りサービスへの応用が期待されている．しかしこのようなアプリケーションでは, 人がどのような行動をとっているのか認識するため, 人の顔向きや動き, 温度や湿度, 明るさなどの環境など付随した多くの情報を統合的に利用するシステムが必要となる．特に, 人の向きの推定にはカメラや距離センサを用いた研究がされているが, センサと人, 障害物などの配置によっては死角が生じるため精度が低下することがある．さらにカメラは人の行動や外見, 距離センサは周囲の物体や人々との距離を計測するため, 利用者にプライバシーに関する心理的な負担を与えてしまう．またセンサのコストが非常に高いことも商業利用の際にネックとなる．そこで, 本研究では天井設置型の複数のマイクと人の位置情報を活用した発話方向の推定手法を提案する．ここで位置情報の取得方法は, 本研究室で提案した, 赤外線センサを用いた深層学習での人の位置推定手法の利用を想定している [1]．マイクはカメラ・距離センサより利用者への心理的負荷が小さく, コストも比較的安価である．天井に設置することで, 死角を減らし, 広い空間の情報を観測できる．さらに照明や空調など設備やデバイスの利用のための音声認識, 咳やタイピング音, 機械の駆動音などの詳細な行動状態の推定への拡張性がある [2]．マイクを用いた位置推定・発話方向推定の既存研究には, 音声信号の強度分布, または位相差を活用するものがある．しかし既存研究では, 高精度に同期された多数のマイクを壁や天井に埋め込む場合が多く, 大規模に導入を行わなければならない．本研究では, 発話方向によって生じる音声

[1] 名古屋大学大学院　工学研究科
Graduate School of Engineering, Nagoya University
[2] 名古屋大学　未来社会創造機構
Institutes of Innovation for Future Society, Nagoya University
[3] ヤフー株式会社
Yahoo Japan Corporation

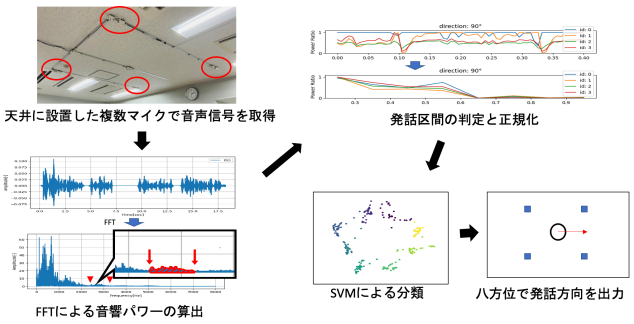

図 1　発話方向推定の流れ

信号の音響パワーの違いを明らかにするため, 音声信号をフーリエ変換後, 特定の周波数領域で積算し, 各チャンネルの信号の強度比を求めた. さらに一定の時間幅 (以降パワーフレームと呼ぶ) の平均と分散を 8 次元ベクトルとして SVM に入力し, 8 方位の発話方向を推定するモデルを構築した. システムの概要を図 1 に示す. 実際に $86.4m^2$ の屋内において, 男性 2 名女性 1 名が約 4 秒の音素バランス文を発話する音声を収集した. 本研究で提案するモデルでは 1.2 秒の発話に対して, 特定の発話者に対しては最大 84%, 発話者を問わない場合は最大 57 %で 8 方位の方向推定が可能であった.

2.　関連研究

2.1　人間の発話音声信号特性

人の発話方向による音声の違いで, 最も大きな特徴は音響パワーの方向性の違いである. 2002 年に Chu と Warnock が無響室で男女各 20 名の英語とフランス語の発話の音声を測定した結果では, 160 ～ 8000Hz の 1/3 オクターブバンドで A 特性音圧レベルを測定したとき, 前後の差が最も大きくなり, A 特性音圧レベルの差はおおよそ 8dB となった. また性別や言語によってスペクトルが異なっても, 方向性の違いはほとんど変化しないことが示された [3]. また音源定位の分野では, 音の到達方向を推定するため, 複数のマイクへの音の到達時間差を用いることが多い. しかし, 発話方向による等距離の地点への音の到達時間差の違いは, 100 ナノセカンドのオーダーとなり, 非常に小さい [4]. 従って非常に高精度に同期したのマイクが必要となる.

2.2　ルールベースの発話方向推定

Avram Levi と Harvey Fellow らは, 音響パワーの違いに基づいて単一の話者の発話方向を推定する手法を提案した [4]. $4m \times 6m$ の部屋内の壁側面に設置したマイクを用いて発話で生じる音響分布を算出した. 空間的ローパスフィルタリングとビームフォーミングを用いることで, 男女各 2 名の英語の発話に対して 22.5° の誤差を許容して, 80 % 以上の推定が可能なことを示した. しかし 449 個と非常に多くのマイクを必要としており, 非常に大規模なシステムであった. 一方で Alessio らは, 音の位相差, つまり干渉を測定することで単一の話者の発話方向を推定する手法を提案した [5]. 各 4 個の分散型マイクアレイを 5 つ用いて, クロスパワー位相分析から得られる干渉場, GCF(Global Cofherence Field) を算出した. これは 2 次元空間上でアクティブな音源が存在する可能性を示す関数である. Alessio らは発話者の位置の円周上の各点に対して GCF のスコアを角度による重みをつけて足し合わせ, その値が最大となる方位を発話方向として推定を行った. しかし, 20 ～ 60 秒の非常に長い音声入力に対して推定を行ったが, 最大で 50° と非常に大きな誤差があった.

2.3　機械学習に基づく発話方向推定

Jackie らは, ほとんどのスマートスピーカーに搭載されている内蔵マイクアレイを活用し, ユーザーの空間位置と頭の向きを音声だけで推測する新しいインタラクション手法を提案した [6]. 16 個のマイクを持つ単一マイクアレイを用いて, 低遅延で発話者の正確な位置と向きを推定するために LSTM(Long Short-Term Memory) アーキテクチャと畳み込みニューラルネットワーク, CNN(Convolutional Neural Network) を使用した機械学習モデルを設計した. 特定の環境で訓練データと検証データのユーザが同じ場合, 位置の平均誤差は 0.31m, 発話方向誤差は 34.3° で推定が可能だった. しかし異なるユーザの場合は位置平均誤差は 0.33m, 発話方向誤差は 40.0°, 異なる環境かつ異なるユーザの場合は位置平均誤差 0.57m, 発話方向誤差は 57.0° であった. 0.1 秒以下の非常に短い音声信号セグメントを入力とした影響はあるが, 十分な汎用性を持つとは言い難い.

3.　提案手法

本節では学習に用いた音声データ, 発話方向データの収集方法とデータの前処理, 学習モデルに関する説明を行う.

3.1　データの収集方法

マイクは無指向性のデジタルマイク SPH0645LM4H を使用した. サンプリング周波数は 16000Hz, 量子化ビット数は 16bit, 感度は 1kHz の音声に対して 94dBSPL, 出力形式は I2S である. I2S とは, Inter-IC Sound の略であり, ESP32 において 2 つのデジタルオーディオデバイス間でオーディオデータを転送するための同期シリアル通信プロトコルである. データを転送するため, M5ATOM Lite を用いて図 2 のようなマイクモジュールを作成して天井に設置した. 本研究では, 図 3 のように 4 つのマイクモジュールを 2m 間隔で天井に設置して利用する. 取得データ量はマイク一つあたり 25.6Kbps で, 送信頻度が非常に高くなるため, Wi-Fi 上で高速に短いメッセージを大量に送受信可能な MQTT(Message Queueing Telemetry Transport) を用いた. システムの構成図を図 4 に示す. ここで電源のラ

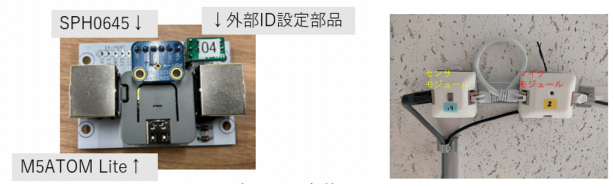

図 2　マイクモジュールと設置図

図 3　マイクモジュール配置図

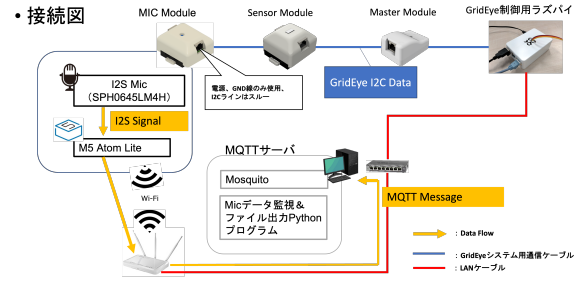

図 4　システム構成図

インは既存の赤外線グリッドセンサのシステムに組み込み, データの転送ルートは別となっている.

3.2　前処理

3.2.1　音響パワーの算出

まずはじめに, 図 5 のように得られた複数チャンネルの音声信号に対してオーバーラップ処理を行う. 下記の実験ではウィンドウ幅を 0.5 秒ごと, スライド幅を 0.05 秒として信号を切り出した. 次に図 6 のように各区間信号に対して高速フーリエ変換を行い, 周波数領域における信号を取得する. そして図 7 のように人の声の周波数領域にあたる, 2400 ～ 3200Hz の部分を足し合わせることで, ノイズを除いて図 8 の通り音響パワーを求められる. この周波数帯は事前実験の推定結果に基づいて決定した. これを各マイクごとに時間方向に値を並べると, 図 8 のマイク毎の音響パワーの推移が得られる.

3.2.2　発話区間の判定

発話区間部分のみを用いるため, 音響パワーが一定の閾値を下回る時間帯のデータを除いた. 図のように実際の値を参照して閾値は 0.05 と定めた.

3.2.3　正規化と平滑化

マイク毎の値を学習モデルに入力するために, 各時点のパワー値の最大値で全チャンネルの値を割って 0 ～ 1 になるように正規化を行った. さらに時間変化に対してロバスト性を持つようにするため, 再度 1 秒以下の時間区間における各マイク毎の平均と分散の値をとった. この時間区間のことを今後パワーフレームと呼ぶ. 中心地点の実際の音声信号と各マイク毎のパワー値の比の平均の変化を図 9 に示す. ここまでの処理により, 複数チャンネルの音声信号はパワー値の比の平均と分散の 8 次元ベクトルに変換される. これに対し, 正解の水平発話方向を 8 方位でラベリングしてデータセットとした.

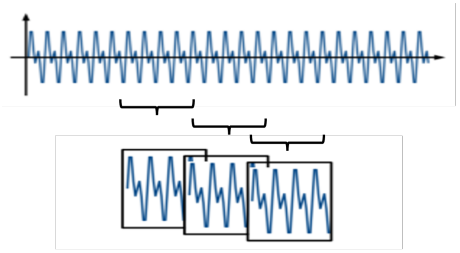

図 5　オーバーラップ処理

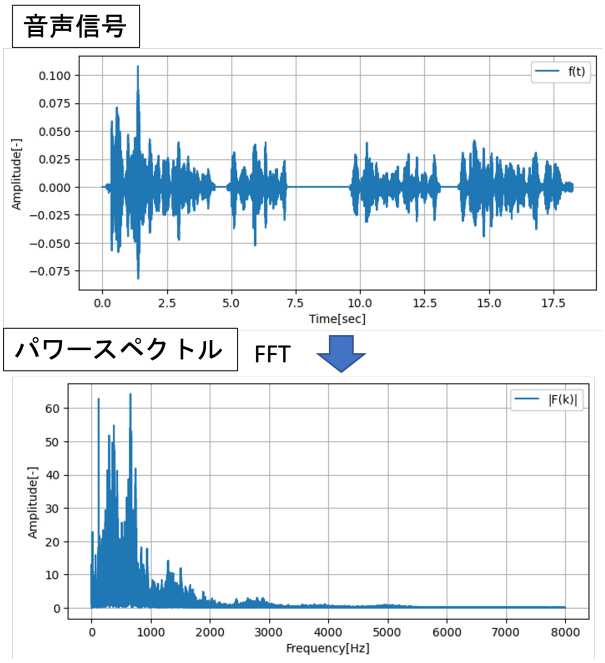

図 6　音声信号の高速フーリエ変換 (FFT)

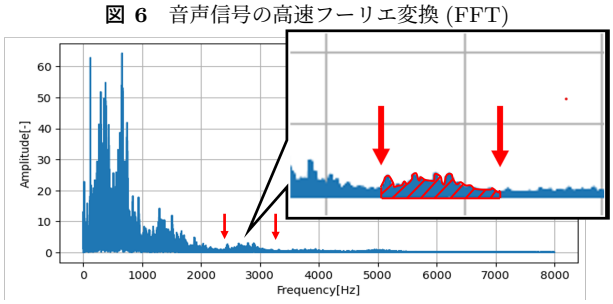

図 7　周波数領域積分範囲

3.3　SVM を用いた発話方向推定

同じ室内であっても場所によって発話者からマイクまでの音響伝達関数が変化するため, 本研究では 0.5 ～ 1m の粒度で発話地点別にモデルを作成した. 図 9 を見てわかる

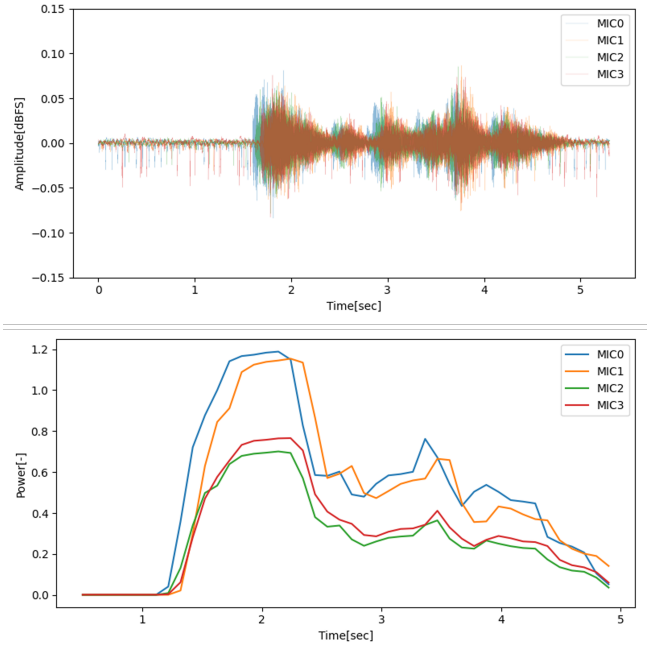

図 8　音声信号と音響パワー

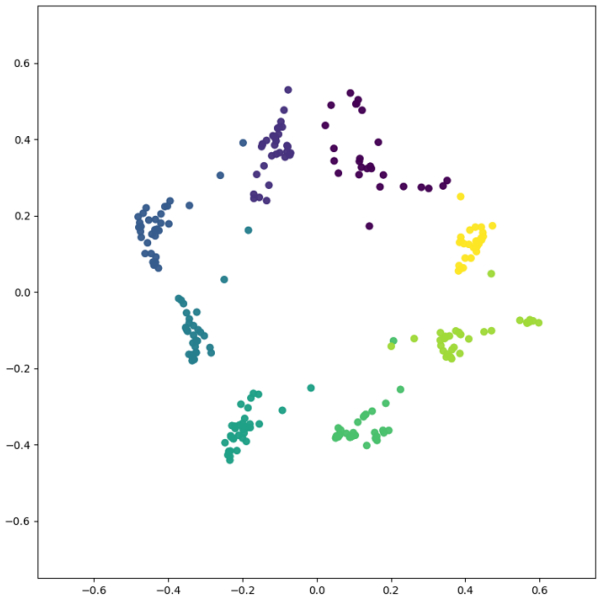

図 10　音響パワー比の特徴量

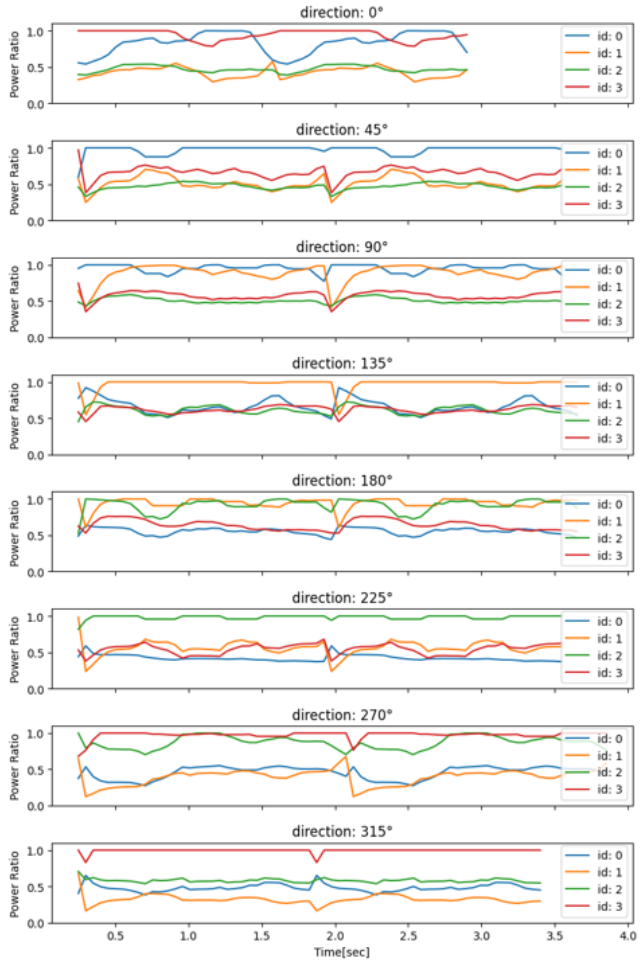

図 9　音響パワー比

ように, 発話方向によって各マイクの音響パワー比に特徴が見られる. 発話方向は本来連続的な値をとるが, ここでは 45° ごとに離散化している. したがって, 各マイクの音響パワーの比を 8 次元空間上の点のベクトルとすれば, 8 方位に対応した空間に分類できるはずである. SVM ではこの 8 次元空間の境界面を求め, 未知の入力に対して分類を行うことができる. 実際に後述の実験で測定されたデータのベクトルを, 次元削減後, 方向別に色分けして 2 次元空間上にプロットすると図 10 のようになった. 各方向ごとにデータが偏在していることが確認できる.

SVM の入力を音響パワー比の 8 次元ベクトル, 出力を 8 方位の発話方向として教師あり学習を行った. データセットはハイパーパラメータはグリッドサーチにより探索した. 場所によって音の反響が異なり, 発話者からマイクまでの伝達関数が変化するため, まずは部屋内における発話地点別にモデルを作成した.

4. 評価実験

上記の推定精度を評価するため, データ収集実験を行った.

4.1 設定

実験に用いた部屋は幅 14.4m, 横 6.0m, 高さ 2.7m であった. 図 11 のように, マイク配置は部屋中心上の 2m 四方, 天井付近高さ 2.53m のアルミフレーム上に設置した. 男性 2 名, 女性 1 名を対象として, 図 12 に示す 14 地点で静止した状態で約 4 秒の音素バランス文を 8 方位で発話した. 1 人あたり, 約 560 秒の音声データを収集した. 図 13 に実験の様子を示す.

図 11　マイクの配置

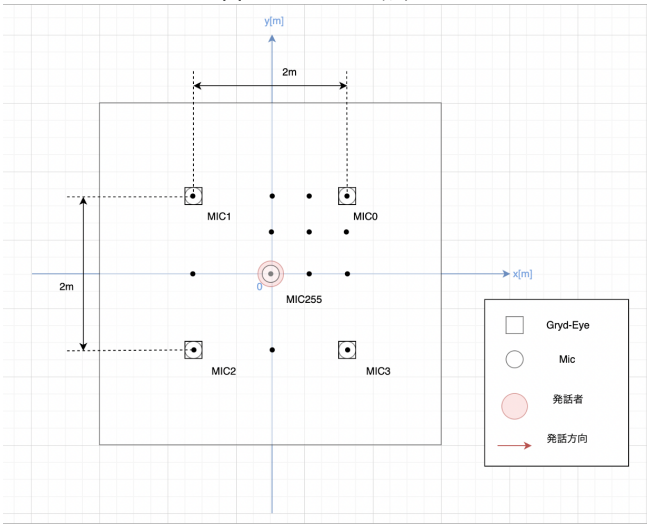

図 12　マイク・発話者の位置

図 13　発話音声データ収集実験の様子

4.2　結果

複数チャンネルの音声信号を 2 分割して, それぞれ上記の前処理を行うことで学習データ,　検証データに分割して推定精度を求めた. ハイパーパラメータは各モデルごとにグリッドサーチによって最適化を行った. 以下に実際にとった値の幅を示す.

$$C = 1 \sim 10000, decision\_function\_shape = ovr,$$

$$gammma = 0.1 \sim 10, kernel = rbf$$

またフーリエ変換のフレームサイズは 1.0 秒で, スライド幅は 0.1 秒とした. パワーフレームのサイズはフーリエ変換後の各値に対して 3, 実質的には 1.0 秒のフレームを 0.1 秒ごとに 3 つずらした 1.2 秒の音声信号入力で推定を行った. まず各地点で各発話者ごとにモデルを分けて学習を行ったところ, 図 14 のようになった. 中心地点で 82%, $(0.5, 0.5)$ の地点で最大 84%であった. 中心から遠のくほど精度は低下し, 特に計測した範囲の端, マイクの直下の点で最小 24%となった. 次に各地点でモデルを分け, 全発話者のデータを用いて学習を行ったところ図 15 のようになった. 中心地点で 47%, $(1.0, 0)$ の地点で最大 57%であった. 外側に行くにつれて精度は低下し, 同様の傾向があった. 中心地点における推定発話方向と正解方向は図 refpredict のような対応となった. 隣接した角度に推定することが予想されたが, 離れた特定の角度に集中的に誤っていることもある. 例えば, 図 refpredict の $(correct\_angle = 90°, estimate\_angle = 180°)$ のセルなどである.

5.　考察

本手法にはいくつかの問題点が見られた. まず一つ目は発話者から各マイクへの方向が特定方向に偏るような発話位置で, 推定精度が低下する点である. 図 14 を見ると $(1.0, 0)$ のような “辺”にあたる部分よりも, $(1.0, 1.0)$ のような “角”にあたる部分では著しく精度が低下している. これは発話者が外側に行くほど, 発話者から見て 4 つのマイクの方向が特定の方向に集まることになる. 例えば上記の “辺”の部分では各マイクは発話者から見て 180° の角度をカバーするが, “角”の部分では 90° の角度しかカバーできていない. これはより部屋の端点にマイク位置を変更する, もしくは数を増やすことで発話者の周囲を均等にカバーできるようにすることで改善すると思われる. 二つ目の問題点は図 8 のように, 取得した音声データに約 0.1 秒以下の時刻のズレが生じている点である. 音響パワーの値も時刻成分でズレが生じるため, 推定精度が下がる要因となる. この要因は M5ATOM Lite から送信された音声信号の MQTT メッセージが受信されるまでに, 通信状況によって遅れが生じることだと思われる. 送信する音声情報にタイムスタンプ情報を付与する, もしくはデータ収集時に特定の周期で特定の音を鳴らして前処理に同期処理を追加するなどの改善方法が考えられる.

6.　まとめと展望

本研究では, 位置情報と天井に設置した複数マイクの音声を用いて, 特定周波数帯の音響パワー算出と SVM モデルをによる人の発話方向推定手法を提案した. 実際に複数人の発話のデータを収集し, 1.2 秒の音声信号に対して, 特定の発話者の場合約 80%, 発話者を問わない場合 50%程度

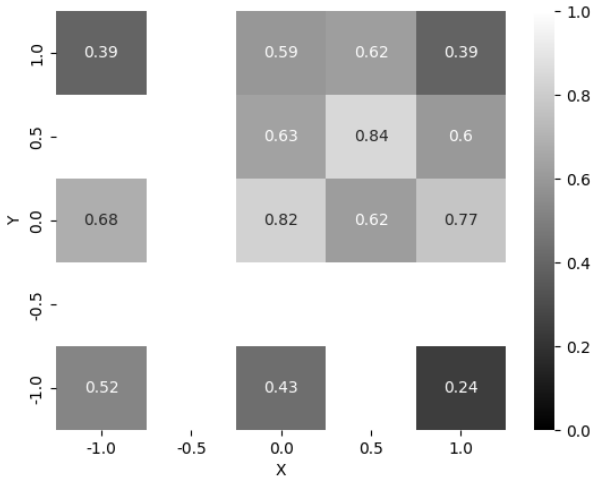

図 14　各発話者別学習モデル平均推定精度

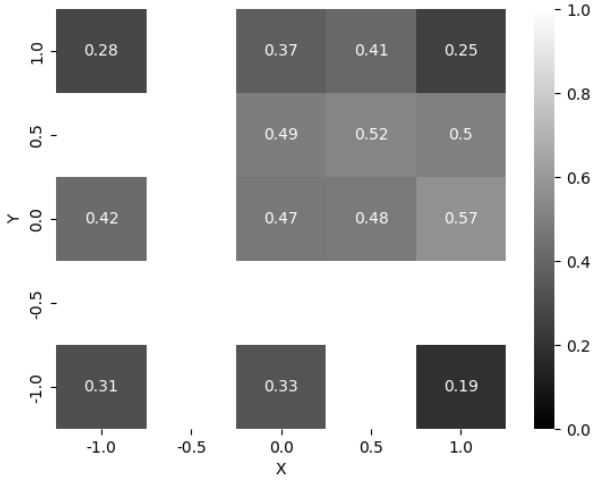

図 15　全発話者学習モデル推定精度

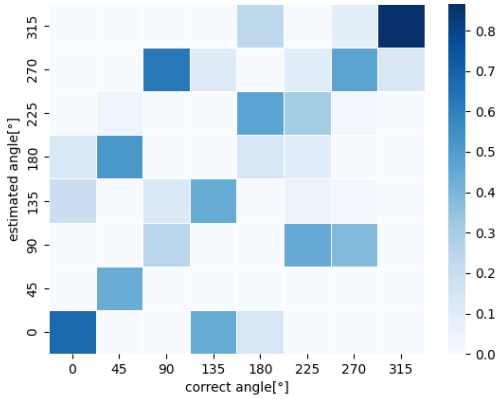

図 16　全発話者学習モデル推定結果

の推定ができることを確認した．本手法は，音声信号間の位相差を利用しないためマイク同士の同期が低精度の音声信号でも利用できる．さらにマイク・部屋の特性をモデルに学習させられるため，データの収集が可能であれば異なる特性のデバイスの音声信号であっても応用できる可能性がある．今後の展望としては，図 17 のようなリアルタイム発話方向推定システム，赤外線グリッドセンサを用いた統

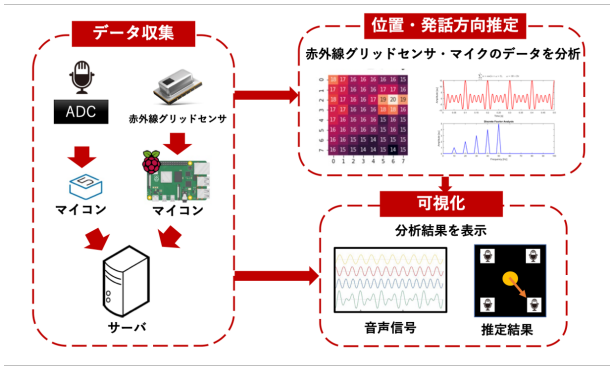

図 17　赤外線グリッドセンサ・マイクを用いた位置・発話方向推定及び行動認識システム

合的な位置・方向システム導入を目指している．

謝辞　本研究の一部は，JST CREST JPMJCR21F2，パナソニックホールディングス株式会社，NICT 委託研究 22609 に支援いただいています．

参考文献

[1] 戸出 悠太, 片山 晋, 浦野 健太, 青木 俊介, 米澤 拓郎, 河口 信夫. 赤外線グリッドセンサを用いた深層学習での人の位置推定手法の検討. マルチメディア, 分散, 協調とモバイル DICOMO2021 シンポジウム.

[2] Gierad Laput, Karan Ahuja, Mayank Goel, Chris Harrison, Carnegie Mellon University. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. 5000 Forbes Ave, Pittsburgh, PA 15213gierad.laput, kahuja, mayank, chris.harrison@cs.cmu.edu.

[3] Chu, W. T.; Warnock, A. C. C. Detailed directivity of sound fields around human talkers. NRC Publications Archive Archives des publications du CNRC. IRC-RR-104, December 2002.

[4] Avram Levi, Student Member, IEEE, and Harvey Silverman, Life Fellow, IEEE. A Robust Method to Extract Talker Azimuth Orientation Using a Large-Aperture Microphone Array. IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 2, FEBRUARY 2010.

[5] Alessio Brutti, Maurizio Omologo, Piergiorgio Svaizer, Istituto Trentino di Cultura(ITC)-irst, Via Sommarive, 18, Povo-Trento, Italy. Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays. INTERSPEECH 2005. brutti/omologo/svaizer@itc.it.

[6] Jackie (Junrui) Yang, Gaurab Banerjee, Vishesh Gupta, Monica S. Lam, James A. Landay. Soundr: Head Position and Orientation Prediction Using a Microphone Array. CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing SystemsApril 2020, Pages 1–12.