エッジ音声処理と CAN 通信を用いた多数接続可能な発話方向推定システム

市川　直人*，加納　一馬，永田　吉輝，片山　晋，
浦野　健太，米澤　拓郎，河口　信夫（名古屋大学）

CAN based Multi-Node Speech Direction Estimation System with Edge Processing

Naoto Ichikawa, Kazuma Kano, Yoshiteru Nagata, Shin Katayama,
Kenta Urano, Takuro Yonezawa, Nobuo Kawaguchi (Nagoya University)

１．　はじめに

近年、人の位置情報はナビゲーションやセキュリティ、マーケティング、物流管理、災害時の対応など様々な分野で利用されている。中でも倉庫やオフィス、医療施設などでは、作業の効率化や活動状況の分析を目的として、Wi-Fi や Bluetooth ビーコン、距離センサ、カメラなどを用いた屋内測位技術が導入されつつある[1][2]。しかし、位置情報だけではどの向きで、何をしているのかといった詳細な分析は困難である。そこで本研究では、音声から発話の有無・強度・方向の推定をして、人同士がどのようなコミュニケーションをとっているかの分析への利用を目指す。

本研究では赤外線センサとマイクを組み合わせた位置・発話方向推定システムを提案する。システムの概要を図 1 に示す。このシステムでは赤外線センサとマイク、エッジ処理のためのマイコンをパッケージ化したセンサモジュールを天井に複数台設置する。また赤外線センサの温度情報は低解像度で、音声は各センサモジュール内でスペクトル化を行うため、被測定者のプライバシーを確保できる。カメラやレーザースキャナーから画像や点群データを用いる場合に比べ、被測定者への心理的な負荷が抑えられるため、パブリックな施設内でも通路や出入口に制限せず導入できる。

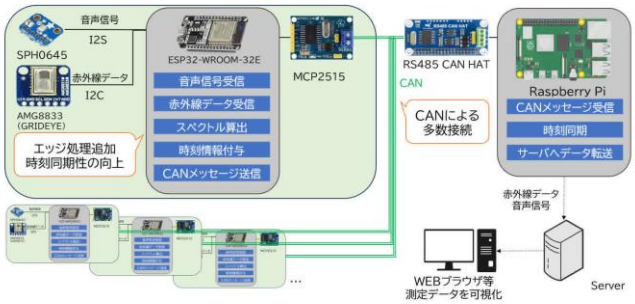

Fig. 1.　System Architecture

２．関連研究

これまでもマイクを用いた位置推定及び発話方向推定の研究は行われてきた。ルールベースの手法ではいくつかの音響的特性に基づいて複数のアプローチが行われてきた。まず始めに、発話者の周囲の音響エネルギー差を用いる手法がある。人が発話を行うと、その指向性に応じて周囲には図 2

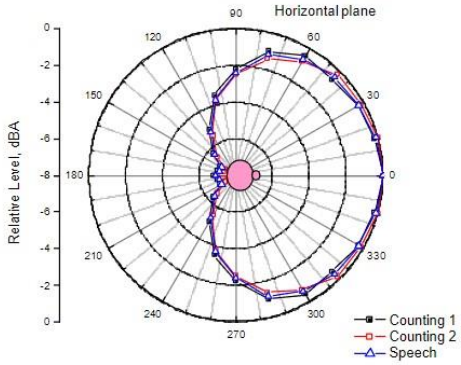

Fig. 2. Repeated test results of relative A-weighted levels in frontal vertical plane of a talker [2]

のような特定の音圧分布が見られる[9]。Levi らは、壁に埋め込んで 449 個、28 クラスタのマイクで音声を測定し、エネルギー放射パターンとのマッチングをすることで発話方向を推定した[3]。他には MUSIC 法やビームフォーミングなど、各マイクに音が伝達するまでの時間差、位相差に注目した手法がある[4]-[6]。

しかし、位相差を活用する手法の多くは一部屋に対し 30 個から 100 個程度のマイクを壁や天井に設置している。また高精度にデータが同期を前提としている。実環境ではこのような設備を導入するには非常にコストがかかり、利用が限られる。

また機械学習を用いた手法もいくつか提案されている。Yang らは VR ヘッドセットを用いて大量のデータを収集し、単一マイクアレイと、CNN+LSTM を用いた学習モデルで最大で位置推定誤差 0.3m, 角度推定誤差 30° の精度を達成した[7]。Ahuja らは単一マイクアレイを用いて複数の音響特徴量を入力としてエクストラツリーで学習を行い、位置が特定された発話者について 65.4%の発話方向推定を実現した[8]。しかし、これらの手法は単一地点での音声データを用いるため、まだ推定精度に問題がある。

本研究では実環境での利用を想定して、比較的安価かつ導入が容易な推定システムを提案する。かつ複数地点で取得した音声スペクトルを用いて、位置が既知の対象について発話方向の推定を行う。データ伝送のためにエッジでデータ量を削減する必要があること、現時点で本システムの時刻同期の

精度がミリ秒オーダーであることを踏まえると、各音声信号の位相差を利用して発話の方向を推定するのは困難である。したがって音響エネルギー分布の特性に着目して推定を行った。

３．実装

センサモジュール内には赤外線アレイセンサ Grid-Eye AMG8833、小型デジタルマイク SPH0645LM4H、エッジデータ処理用のマイコン ESP32、及び CAN コントローラが搭載されている。マイクのサンプリング周波数は 16,000Hz、量子化ビット数は 16bit、感度は 94dBSPL(1kHz)、出力形式は Inter-IC Sound(I2S)である。サイズは縦 75mm、横 60mm、厚さ 40mm で手の平上に乗る程度である。マスタモジュールには RaspberryPi 4 Model B に RS485 CAN HAT を搭載して使用した。

通信方式の検討について示す。センサモジュールは赤外線センサの測定範囲を考慮して高さ 2.7m の天井において 2m 間隔で設置する。一般の会議室、オフィスでの利用を想定しておおよそ 20 台、配線距離は約 50m を目安とした。接続台数、配線距離の他、通信速度、拡張性、動作安定性、電源供給、コストなどの評価から表 1 のようにいくつかの通信方式を比較した。以上より、接続台数及び安定性、電源供給の利便性から車載通信にも利用される Controller Area Network, CAN を採用した。

Table 1　Communication system

通信方式	I2C※1	Wi-Fi	Ethernet	CAN	A2B※2
通信速度[bps]	400K	1.6G	1G	1M	50M
最大接続台数[台]	112	約 50	ポート数依存	約 100	10
接続形式	バス型	スター型	スター型	バス型	バス型
拡張性	乏しい	非常に良い	良い	良い	非常に良い
接続可能距離[m]	約 5	約 100	100	40	40
対ノイズ性	低い	電波環境に依存	高い	高い	高い
電源供給	不可	不可	可(PoE)	同一配線で可	可
コスト	低い	低い	低い	低い	高い
備考	接続数に応じて通信読度低下	伝送ロス有り安全性に懸念有り	分岐にはハブが必要	接続数・距離に応じて通信速度低下	音声伝送に特化

※1. Philips 社が開発したシリアルバス通信方式(Inter-Integrated Circuit)

※2. 車載用オーディオ・バス(Automotive Audio Bus)

４．推定手法

位置推定には我々が以前提案した CNN ベースの推定手法[10]の利用を前提として、ここでは位置情報と音声情報を用いていかに発話の方向を推定するかを扱う。推定の流れを図 3 に示す。

センサモジュール内ではマイクから取得した音声信号に対して短時間フーリエ変換（Short Time Fourier Transform, STFT）

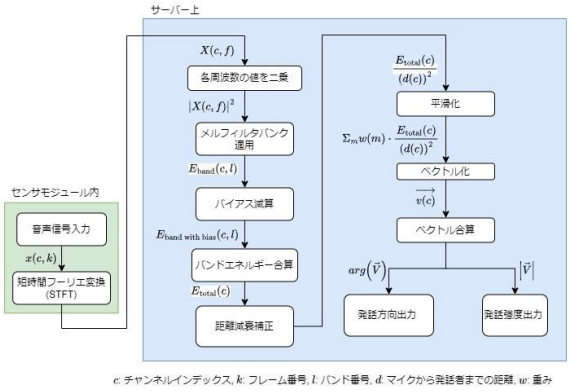

Fig. 3. Flow of speech direction estimation

を行う。サンプリングサイズを 256 として、オーバーラップは行わなかった。窓関数にはハミング窓を用いた。フレーム長は 16ms、周波数分解能は 62.5Hz である。発話時に得られた音声スペクトルの例を図 4 に示す。またセンサモジュール内で、各ウィンドウの始めの時間をタイムスタンプとしてデータに付与した。

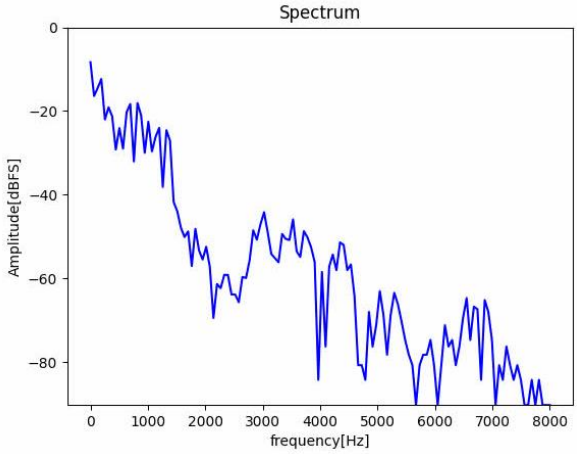

Fig. 4. Spectrum (while speaking)

まず各センサモジュールから送信された振幅スペクトルから、各周波数の値を二乗してパワースペクトルに変換後、メルフィルタバンクを適用する。メルフィルタバンクとは、複数のバンドパフィルタをメル尺度（人間の音高知覚特性を考慮した尺度）に応じて等間隔に配置したものである。周波数をfとしたとき、メル尺度は以下の式で表される。

$$mel(f) = 1127 \cdot \log\left(1 + \frac{f}{700}\right)$$

メルスペクトルからは各バンドのエネルギーの大きさを得られるが、この値は同一の音声に対しても各センサモジュールによって異なる値が得られた。おそらく各マイクや回路などの特性や録音環境下における定常的なノイズなどのバイアス成分が含まれていると考えられる。そこで無音環境下における各バンドのエネルギーとの差分をとり、これらのバイアス成分を除去した。

$$E_{band} = E_{band\ with\ bias} - E_{silence}$$

その後各バンドのエネルギーを合算し、各チャンネルのエネルギーを求めた。ここでcはセンサモジュールのチャンネル番号($c = 0,1,2,3$)、lはバンド番号である。

$$E_{total}(c) = \sum_{l} E_{band}(c, l)$$

次に距離による音声の減衰を補正するため、発話位置―マイク間の距離$d(c)$の二乗で除算した。さらに時間方向への平滑化を行うため、過去のフレームほど重みwが小さくなるように定め、過去 1 秒分の値で加重平均をとった。ウィンドウのインデックスmとしたとき以下のように表される。

$$w_{m-1} = rw_m \ (r = 0.8)$$

$$\sum_{m=0,-1,\ldots} w_m = 1$$

この値は発話地点から各マイク方向へ伝わるエネルギーのベクトル$\overrightarrow{v(c)}$の大きさと考えられる。

$$\left|\overrightarrow{v(c)}\right| = \sum_{m} \left(w(m) \frac{E_{total}(c)}{\left(d(c)\right)^2} \right)$$

発話をした際、他に音源がなく、反響が発生しない理想的な環境下では図 2 のような音圧分布が生じるはずである。つまり、発話者の前後左右では伝達するエネルギーの大きさに差が生じるはずである。ここで、ベクトル$\overrightarrow{v(c)}$の合計を$\vec{V}$と定義する。

$$\vec{V} = \sum_{c} \overrightarrow{v(c)}$$

周囲にある程度均等にマイクが配置されている場合では、各マイクへ伝わるエネルギーのベクトルの合計が発話の方向に対応すると考えられる。また、発話の音量が大きいほど各マイクへ伝わるエネルギーの差も大きくなるはずである。したがって$\vec{V}$の角度を発話方向、大きさを発話強度として出力した。

４．結果

本研究室内で発話を行い、発話方向推定の結果を WEB ブラウザ上にリアルタイムで表示をした。この際、無音環境下における各バンドのエネルギーは 10 秒程度で測定した。斜め上からの撮影動画と各センサの取得データ、発話方向推定結果を合わせた様子を図 5 に示す。図 5 左の白線の円が発話者の位置を示し、推定された発話の方向を赤の弧の部分で示している。この弧は推定された発話強度に対応しており、値が大きいほどより濃くなる。

発話した際の様子を見たところ、概ね実際の発話に応じて赤の弧が表示された。推定方向は 45 度程度の幅で揺れることがあるが、各マイクの中心地点ではおおよその発話方向を推定できた。ただし、推定結果が特定の方向に全体的に偏る傾向があった。環境にノイズとなる音源があったか、もしくはスペクトル減算だけでは補正しきれなかった各マイクの音響特性が要因だと考えられる。また中心地点以外では精度が低下する傾向が見られた。

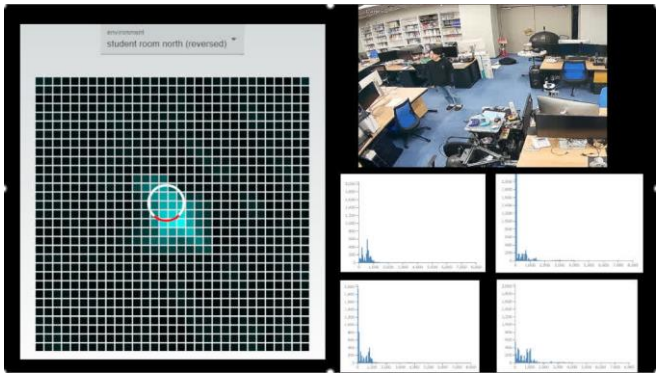

Fig. 5.　Estimation results at the time of speech

５．まとめと展望

本研究では赤外線センサとマイクを用いた発話方向推定システムを提案した。組み込みシステムで低コストかつ導入が容易なセンサモジュールを制作し、複数地点での単一マイクの音声を活用するため、CAN 通信とエッジでの音声スペクトル変換及び時刻付与の仕組みを実現した。今後は、推定精度の評価、発話時の音響エネルギー分布の形状を考慮したアルゴリズムの検討、高周波数帯と低周波数帯のエネルギー比(HLBR)の活用などに取り組む予定である。

謝辞　本研究の一部は科学技術振興機構(JST) CREST 課題番号 JPMJCR22M4、NICT 委託研究(22609)に支援いただいています。また、本研究の一部はパナソニック ホールディングス株式会社の泉様に助言をいただき推進しております。

文　献

[1] Brena　他：Journal of Sensors, 2630413, 2017
[2] Liu　他：IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), pp1067-1080, 2007
[3] Levi　他：IEEE transactions on audio, speech, and language processing Vol.18, No.2, pp.277-285.
[4] 中島　他：IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.676-683, 2009
[5] Brutti　他：Interspeech, Vol5, pp.2337-2340, 2005
[6] 石井　他：日本ロボット学会, Vol.34, No.3, pp.199～204, 2016
[7] Yang　他：CHI Conference on Human Factors in Computing Systems, pp.1-12, 2020
[8] Ahuja　他：Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, pp.1121-1131, 2020
[9] Chu　他：National Research Council of Canada, 2002
[10] 戸出　他:マルチメディア, 分散, 協調とモバイルシンポジウム, pp.909-915, 2021.