生活支援ロボットのタスク遂行のための雑音環境における話者位置推定を用いた命令文理解システムの開発

Development of a Command Sentence Understanding System Using Speaker Localization in Noisy Environments for Task Execution by Life Support Robots

三江啓貴 †, 崔龍雲 †
Hirotaka Mie†and Yongwoon Choi†
† 創価大学大学院　理工学研究科
†Graduate School of Science and Engineering, Soka Univ.

要旨

本研究では，家庭環境において生活支援ロボットが話者位置を推定し，命令文理解を行うシステムの開発を目的とする．従来の Attention 型 Seq2Seq を用いた命令文理解手法は，入力として理想的な命令文を必要とするため，生活音が混在した雑音環境である家庭環境には対応できない．そこで，本システムではマイクアレイと RGB-D センサで推定した話者位置を用いた音源分離を行うことで周囲の雑音を低減し，雑音環境での命令文理解を行う．本システムの話者位置推定と音源分離における評価をする実験では，ロボットと話者の距離が 2[m] 以内の範囲で話者位置を推定し，命令文を 89.09[%] の精度で理解した．

1. はじめに

家庭環境における生活支援ロボットは,「Bring-me task」をはじめとした荷物の運搬や片付けなどの人の作業を，代替して行うことが期待されている [1][2]．このような場面では，ロボットが音声情報から命令の意図を理解することで，人に指示された作業を行うことができる．そのため，ロボットは自然言語の様々な表現の命令から，タスク遂行に必要な情報を一意に理解する必要がある．

音声情報から命令を理解する手法の 1 つとして，ロボットのタスク遂行における Attention 型 Seq2Seq を用いた命令文理解 [3] がある．この手法は単語間の意味関係を表す特徴を Encorder で抽出し，その特徴から命令に必要な情報を Decorder で単語の配列として出力することで，多様な表現の命令文を理解することができる．しかし，入力となる命令文には理想的な文章が想定されており，生活音や話者以外の音声などが混在する雑音環境に対応できないことが課題となる．また，命令文からのみでは「Bring-me task」に必要な情報である話者位置を特定できない．

そこで本研究では，雑音環境に対応した命令文理解を行うために，話者方位の音を強調することで雑音を低減した命令文理解システムを開発する．本システムでは，ロボットに搭載した 2 台のマイクアレイによる音源定位と RGB-D センサを用いた人物検出の統合により，話者位置の推定を行う．その後，ロボットが話者の正面に移動し再度発話を促す．さらに，ビームフォーミング法 [4] を用いて話者方位に対して音源分離することで，話者からの入力を強調した音声を取得し，音声認識と命令文理解を行う．

本稿では，2 台のマイクアレイと RGB-D センサを用いた話者位置推定と，それを用いた命令文理解システムについて述べる．そして，雑音環境における本システムの話者位置推定と命令文理解の精度を評価する実験を行い，本システムの有用性について評価する．

2. マイクアレイを用いた命令文理解

図 1 に，本システムの音声入力からタスク遂行までの流れを示す．まず，図 1 の STEP1 で 2 台のマイクアレイと RGB-D センサを用いて話者位置推定を行う．STEP2 では，ロボットが話者の正面に移動した後，話者方位の入力音声を音源分離することで話者からの音声を強調する．STEP3 では，STEP2 で得られた音声に対して音声認識をすることで命令文を取得する．STEP4 では，命令文に対して Attention 型 Seq2Seq を用いて命令文理解を行う．

2.1. 音源座標と人物検出による話者位置推定

STEP1 では，2 台のマイクアレイを用いて計測した音源座標と，RGB-D センサで検出した人物座標を比較することで話者位置を推定する．まず，ロボットに搭載した 2 台のマイクアレイで TDOA[5] による音源方位の計測を行い，各計測値から三角測量を用いて音源座標を特定する．次に，SSD(Single Shot Multibox Detector)[6] を用いた人物検出結果と，RGB-D センサから取得した 3 次元点群座標を，統合

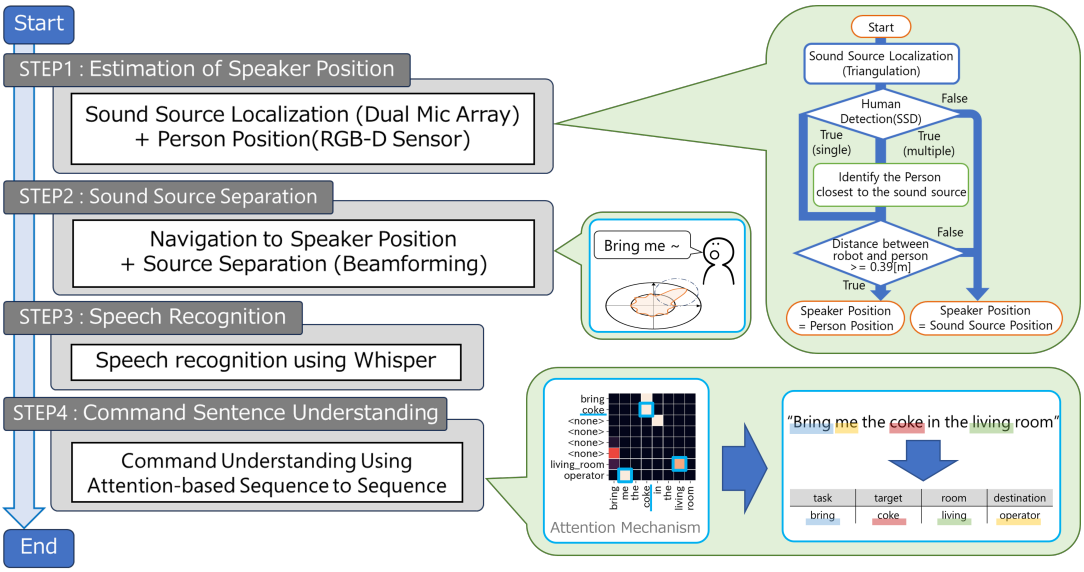


図 1: 話者位置推定を用いた命令文理解システムのフロー図

することで人物座標を取得する．そして，音源座標と最も近い人物座標を統合することで話者位置を推定する．

2.2. 話者方位に対する音源分離

STEP2 では，2.1 節で推定した話者位置の正面にロボットが移動し，話者方位に対して音源分離を行う．本システムでは，命令文に指差しなどの指示を含む場合に対応するために，話者がカメラの画角内に収まる話者からの距離 1.5[m] の地点をロボットの移動先とする．音源分離の手法として，話者方位からの音を強調するためにマイクアレイを用いたビームフォーミング法を用いる．移動後に話者方位となるロボット正面の入力を増幅し，他の方位の入力を低減することで，入力音声から話者による音声のみを強調する．また，音の大きさが閾値以下の部分を雑音として抑制することで，音源分離処理によって発生した音の歪みによる影響を低減する．これにより，話者による発話を強調した音声が得られる．

2.3. 音声認識と命令文理解

STEP3 では 2.2 節で音源分離した音声に対して，音声認識モデルの Whisper[7] を用いて音声認識を行う．さらに，音声認識で取得した命令文に対して，STEP4 で Attention 型 Seq2Seq を用いた命令文理解を行う．それにより，タスクに必要な情報となる単語の配列を出力する．例えば,「Bring me the coke in the living room」の入力には「task:bring，target:coke，room:living，destination:operator」のように順番に意味を持つ単語の配列を出力する．この際,「Bring-me task」などの運搬先が話者である場合には，2.1 節で推定した話者位置を出力することで，タスク遂行に必要な話者位置の情報を取得する．

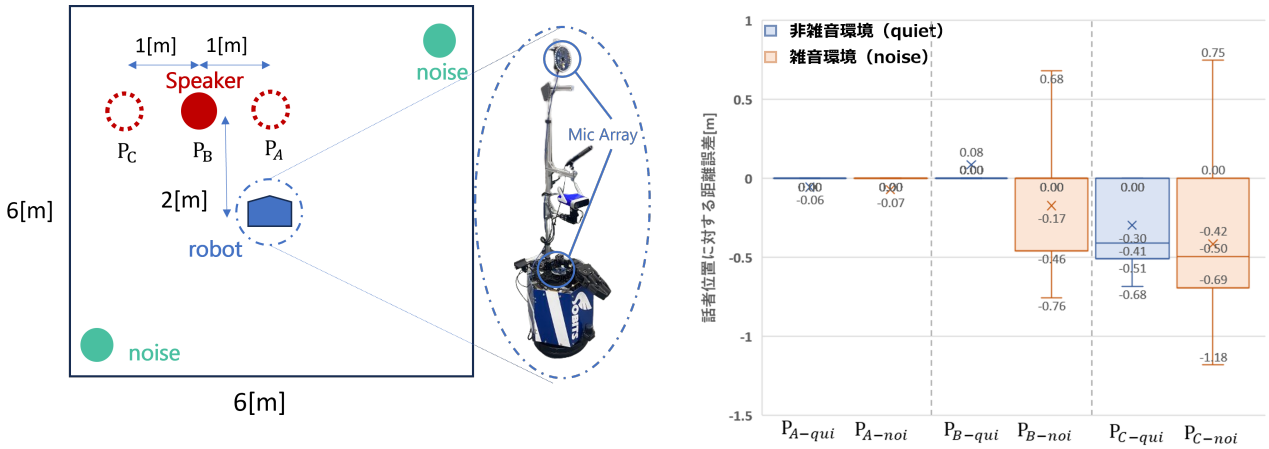
3. 雑音環境におけるシステムの話者位置推定と音源分離の評価実験

本実験の目的は，雑音環境における本システムの図 1 の STEP1，2 の精度を評価することにより，本システムによる命令文理解を評価することである．本実験では，ロボット位置で計測した際に 55[db] となる人の会話音声を，実験環境内に配置したスピーカから出力することで雑音環境とする．話者として 12 人の被験者で実験を行い，被験者の中で発話の平均音圧が 55[db] 以下であった 1 名は評価データから除いた．

3.1. 雑音環境における話者位置推定の精度評価実験

本実験は，本システムの STEP1 の話者位置推定の精度を評価することを目的とする．本実験では人を識別するために，パーソナルスペースである 0.78[m] の半分となる 0.39[m] を基準 [8] に話者位置推定を評価する．

話者位置推定の評価実験に用いる環境を図 2(a) に示す．話者位置推定の評価では本システムの STEP1



(a) 話者位置推定の精度を評価する環境　　(b) 雑音環境と非雑音環境における話者位置推定の精度

図 2: 話者位置推定の精度評価実験

を用いて，図 2(a) に示した 3 地点（P_A，P_B，P_C）でロボットに対して発話する話者位置を推定する．巻き尺で計測した話者位置を真値として，推定した話者位置との誤差を評価する．また，雑音環境による推定精度への影響を示すために，会話音声を出力しない非雑音環境における推定精度との比較を行う．

図 2(b) に雑音環境と非雑音環境の，各地点における話者位置推定の結果を示す．話者位置が P_A の場合には，非雑音環境では平均 0.05[m]，雑音環境では平均 0.07[m] の誤差で話者位置を推定した．P_B の場合の話者位置推定では，非雑音環境では平均 0.08[m] の誤差であり，雑音環境では平均 0.17[m]，最大 0.76[m] の誤差が生じた．また，P_C の場合の話者位置推定では，非雑音環境では平均 0.30[m]，最大 0.68[m] の誤差であり，雑音環境では平均 0.42[m]，最大 1.18[m] の誤差が生じた．両環境において，ロボットと話者の距離が大きくなるほど誤差が大きくなることから，天井などの反響により，推定した音源方位の精度が低下したと考えられる．P_B の結果から，非雑音環境に比べて雑音環境の方が，反響の影響が大きくなり，最大誤差と分散が大きくなることが示唆される．雑音環境の P_A における話者位置推定では，パーソナルスペースの半分の距離である 0.39[m] 以内の誤差であった．このことから，話者とロボットの距離が，P_A とロボット間の距離である 2[m] 以内の環境では，話者位置推定が有用であると考えられる．

3.2. 雑音環境における音源分離を用いた命令文理解の精度評価実験

本実験では，STEP2 以降にあたる音源分離を用いた命令文理解の精度評価を目的とする．命令文理解の精度評価には，SuperGLUE[9] において人が命令文を理解する精度である 89.80[%] を基準とする．

命令文理解の精度を評価する環境を図 3 に示す．雑音環境における本システムの有用性を示すために，図 1 の STEP2 を適用する場合と非適用の場合における命令文理解の精度を比較する．図 3 に示した 3 地点（P_A，P_B，P_C）で話者が発話する命令文に対して，音声認識と命令文理解を行う．本システムを非適用の場合の評価にはロボットの移動を含まないため，ロボットの初期位置（P_S）における命令文理解の精度を用いる．本システムを適用する場合の評価には，STEP2 の処理でロボットの移動先となる話者から 1.5[m] の地点（P_G）における命令文理解の精度を用いる．命令文理解のデータセットには RC2023 の GPSR[10] で使用された命令文生成器で生成した文を用い，発話文章には同様に生成したテスト用の命令文を用いる．話者は各地点で異なる 5 つの命令文を発話するものとし，各命令文ごとに評価を行う．

命令文理解の結果を表 1 に示す．表 1 から，本システムを非適用の場合では，P_A，P_B，P_C でそれぞれ 54.54[%]，61.81[%]，49.09[%] の精度で命令文を理解するという結果が得られた．音声認識の処理において，発話音声が雑音と区別されずに音声認識ができない場合や，単語の誤認識があったことから，命令文理解の精度に雑音による影響があったと考えられる．それに対して，雑音環境において本システムを適用した場合では STEP1 の話者方位に対する音源分離を用いたことで，89.09[%] の精度で命令文を理解できるという結果が得られた．STEP2 の処理により発話部分が強調されたことで，雑音による影響

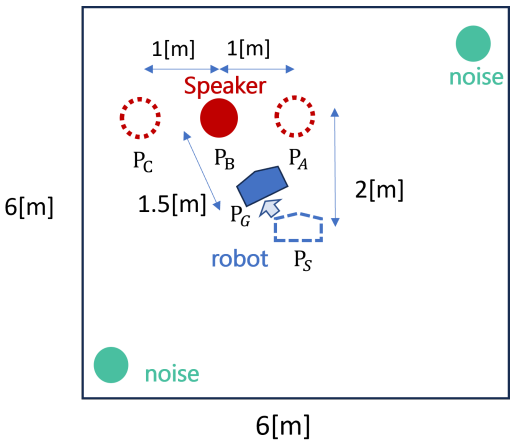


図 3: 雑音環境における命令文理解の実験環境

表 1: 雑音環境における命令文理解の精度

	P_A[%]	P_B[%]	P_C[%]
本システム非適用	54.54	61.81	49.09
本システム適用	89.09		

を低減することができ，全ての発話において発話音声と雑音を区別して音声認識ができていた．さらに，非適用の場合と比べて雑音による単語の誤認識が減ったことで，命令文理解の精度が向上した．本システムによる命令文理解の精度は，SuperGLUE で人の理解精度とされる 89.80[%] には達しないが，非適用の場合と比べて 27.28[%] 以上向上したことから，雑音環境における命令文理解に有用であると言える．

4. まとめ

本研究では，話者方位に音源分離した音声を用いた命令文理解システムの開発を行い，話者位置推定と音源分離における評価を行った．実験では，雑音環境と非雑音環境で話者位置推定の精度を評価し，本システムを適用した場合と非適用の場合における命令文理解の精度を評価した．実験結果では話者とロボットの距離が 2[m] 以内の場合，話者位置推定が可能であることを示したが，距離が大きくなるほど天井などの反響によって位置推定の精度が低下することがわかった．命令文理解では，雑音環境において本システムを適用することで 89.09[%] の精度で命令文を理解でき，適用しない場合に比べて 27.28[%] 以上向上したという結果が得られた．このことから，本システムは雑音環境における命令文理解に有用であることを示した．今後は，点群による音源方位の距離を用いて，話者位置推定における音源座標を補正する手法を検討する．

参考文献

[1] 伊藤 魁一, 諸橋 一穂, 三浦 純,“Bring-me task におけるユーザ指示のあいまいさ解消のための質問生成”, ロボティクス・メカトロニクス講演会講演概要集, 1P1-D11, 2020.
[2] NEDO,“生活支援ロボット実用化プロジェクト”, 2011, Ref.:2023-01-04.
[3] 鶴江 匠, 崔 龍雲,“生活支援ロボットのタスク遂行のための未知語を含む命令文理解手法の開発”, 創価大学大学院 修士論文，2022.
[4] I. Andras, P. Dolinsky, L. Michaeli and J. Saliga,“Beamforming with small diameter microphone array”, Proc. 28th Int. Conf. Radioelektronika, pp. 1-5, Feb. 2018.
[5] C.Junu Jahana, et al.,“Direction Of Arrival Estimation using Microphone Array”, 2021 Fourth International Conference on Microelectronics, Signals & Systems, Kollam, India, pp.1-6, 2021.
[6] W.Liu, et al. ,“SSD: Single shot multibox detector”, European conference on computer vision, pp.21-37，2016.
[7] A. Radford, et al.,“Robust speech recognition via large-scale weak supervision”, Tech. Rep., OpenAI, 2022.
[8] 山口 千晶, 山 祐嗣,“現実世界状況法によるパーソナル・スペースの測定”, 対人社会心理学研究, 16 pp.1-8，2016.
[9] A. Wang, et al.,“SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”, NIPS, 2019.
[10] J. Hart, et al.,“RoboCup@home 2023:Rules and Regulations”, 2023, Ref.:2023-10-23.