5W-03

マスク越しフレーズ認識に向けた ミリ波レーダによる口の形状変化の検出手法

山川 凌太郎 [1]　飯塚 達哉 [1]　石毛 真修 [1]　笹谷 拓也 [1]　高木 健 [1]　川原 圭博 [1]
東京大学 [1]

1 はじめに

新型コロナウイルスの感染拡大の影響によりマスクの着用が普及しているが，音声の減衰や口の形が見えなくなることにより，発話内容が伝わりにくくなる．そこで本論文では，マスクを透過するミリ波帯のレーダを用いて口の動きを読み取り，機械学習により発話内容を推定する手法を提案する．ミリ波レーダの受信信号から検出できる口付近の速度の変化を画像形式の入力データとして用い，それを特徴量として機械学習によるフレーズの分類を行う．本手法は，デバイスを顔に接触させる手法 [1] と異なり非接触で発話内容が検知できるため，接触による不快感を伴わなず，離れた相手の発話内容を検知することもできる．また口の動きのみを利用するため，声が出せない環境におけるコミュニケーションも支援できる．本稿では顔の位置を固定し，マスク着用時と非着用時のそれぞれの場合において，CNN を用いたフレーズの分類を行った．そして，マスクを着用してレーダから 1 m 離れた状況で，4 つのフレーズを 91 % 以上の精度で分類できることを示した．

2 提案手法:ミリ波 FMCW レーダを用いた口の形状変化の検出

ミリ波 FMCW レーダは，チャープ信号を介して離れた対象までの距離やその細かい動きを検出できる．そのため，指先などを用いた細かなジェスチャーの認識にしばしば用いられる [2]．また近年，一部の市販スマートホンに FMCW レーダが搭載され，低コストで利用できる技術になりつつある．本手法ではミリ波 FMCW レーダがマスクを透過することを利用し，口の形状変化を受信信号のドップラー速度成分から検出することを目指す．

図 1 にミリ波レーダの構成および信号処理のプロセスを示す．発話中のユーザの口にレーダを当てると，送信波

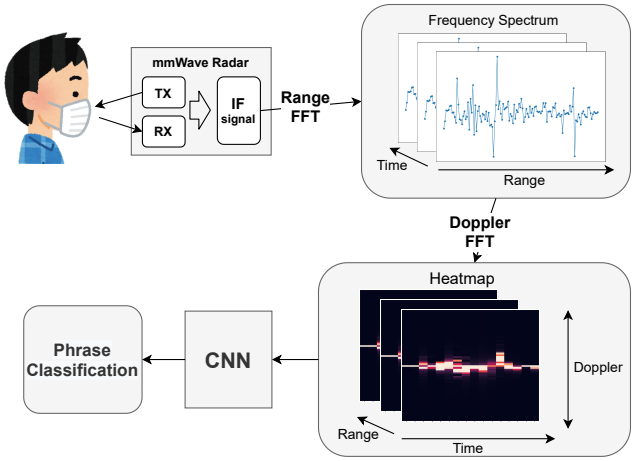


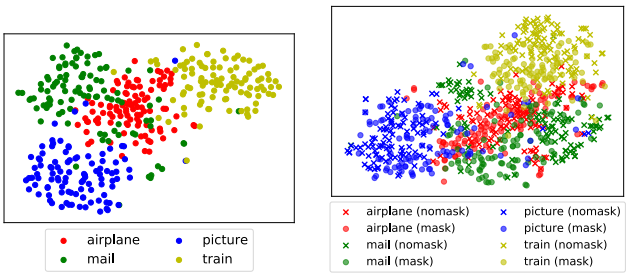
図 1: 提案手法のシステム構成および信号処理の流れ

と受信波の周波数差と位相差に対応する IF 信号が得られる．チャープ信号の特性により，この IF 信号の周波数差が対象（顔）までの距離に対応するため，高速フーリエ変換（FFT）により距離を求められ，この操作を Range-FFT と呼ぶ．しかしミリ波 FMCW レーダの距離分解能は使用できる周波数帯域幅と反比例の関係にあり，電波法などの制約から実用上数 cm 程度の解像度しか得られない．そのため，Range-FFT の情報のみから口の形状を読み取ることは難しい．そこで，Range-FFT で得られたスペクトルの位相の変化に対象の速度成分が含まれることに着目し，Range-FFT の後にもう一度 FFT を行うことにより，口の動きをドップラー速度として検出する．この操作を Doppler-FFT と呼ぶ．この 2 回の FFT を行うことで，口の動きを（時間，距離，速度）の 3 次元テンソルで表すことができる．そしてこれらを CNN に入力することでフレーズの分類を行う．

3 実験・評価

口の細かい動きの変化をレーダのドップラー速度成分から識別できるか検証するため，まずマスクを着用していない状態で，顔の位置を固定し，フレーズの認識を行った．実験には Texas Instruments 製のミリ波レーダ評価モジュール（IWR1443）を使用し，周波数帯域は 77~81 GHz に設定した．そして著者（20 代男性）を被験者とし，レーダと顔との

Detecting Method of Mouth Shape Change with mm-Wave Radar for Phrase Recognition through a Mask
Ryotaro Yamakawa[1], Tatsuya Iizuka[1], Matthew Ishige[1], Takuya Sasatani[1], Ken Takaki[1], and Yoshihiro Kawahara[1]
1. The University of Tokyo



(a) マスク非着用　(b) マスク非着用&マスク着用

図 2: 距離 50 cm における t-SNE 散布図

距離が 15 cm，50 cm，100 cm の場合について，各フレーズの口の動きを 100 回ずつ測定した．本稿では初期検討として [3] で用いられた 4 つの英文フレーズ “Turn on Airplane mode.”，“Check for new mail”,“Take a picture”,“When next train departs” を用いた．

まず分類が可能か検討を行うために，t-SNE 散布図を用いて視覚的にデータの分布を確認した．図 2a に距離 50 cm，マスク非着用時のデータの分布を示した．同じフレーズを示す点が固まって分布しており，フレーズごとの口の形状変化の特徴を捉えていることがわかる．

次にマスクを着用した状態で同様の実験を行った．図 2b は，距離 50 cm における，マスク着用時および非着用時のデータの散布図である．マスクの有無にかかわらず，各フレーズが概ね近くに分布しているため，各フレーズ毎に口の動きの変化が特徴づけられていることがわかる．

さらに，各フレーズの 100 個のデータをそれぞれ訓練用データ 85 個と試験用データ 15 個に分けて CNN による分類精度を評価した．図 1 のヒートマップは時間とドップラー速度を軸に持つが，この画像に口の動きの特徴が表れているため，Range 方向を CNN のチャネル成分として入力した．表 1 に各距離における精度をそれぞれまとめた．マスク非着用時の精度は距離が離れるほど低下する一方で，マスク着用時には 50 cm よりも 100 cm の方が精度が高くなっており，マスク着用時に 100 cm 離れていても，91% 以上の精度で 4 つのフレーズを分類できることがわかった．

表 1: マスク非着用，着用，それぞれの条件下での分類精度

距離	15 cm	50 cm	100 cm
マスク非着用	0.9533	0.9517	0.8867
マスク着用	0.9683	0.8767	0.9200

次に，マスク非着用時と着用時のそれぞれのデータを訓練用データと試験用データに分けて CNN で分類精度を比較した．表 2 の条件 A と条件 B にどの割合で分けたかを示している．条件 B ではマスク非着用時のデータの学習でマスク着用時のデータを分類できるかを評価した．

表 3 にそれぞれの分類精度を示した．A の場合は各距離共にマスク着用時とほぼ精度が変わらないが，B の場合は 100 cm の時の精度が一番高くなっている．これより，マスク非着用時のデータでの学習がマスク着用時のデータの識別にも活用できることがわかった．

表 2: データ分割条件

条件 A	訓練	試験	条件 B	訓練	試験
マスク非着用	0.85	0.15	マスク非着用	1.00	0.00
マスク着用	0.85	0.15	マスク着用	0.00	1.00

表 3: マスク非着用，着用のデータを合わせた際の分類精度

距離	15 cm	50 cm	100 cm
条件 A	0.9683	0.8933	0.9184
条件 B	0.7950	0.5355	0.8525

4　おわりに

本稿ではマスク越しにフレーズの認識をするために，ミリ波レーダを用いた口の形状変化の検出手法を提案した．そして 4 つのフレーズを分類する際の精度を評価し，マスク着用時にも高い精度でフレーズの分類ができることを示した．今後は人との対話場面を想定してフレーズを増やすほか，顔の位置が変動する場合の影響の評価や，複数の被験者がいる場合における評価を行う．本研究は JST ERATO 川原万有情報網プロジェクト（JPMJER1501）の助成を受けたものである．

参考文献

[1] Kimura, N., Kono, M. and Rekimoto, J.: SottoVoce: an ultrasound imaging-based silent speech interaction using deep neural networks, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–11 (2019).

[2] Lien, J., Gillian, N., Karagozler, M. E., Amihood, P., Schwesig, C., Olson, E., Raja, H. and Poupyrev, I.: Soli: Ubiquitous gesture sensing with millimeter wave radar, *ACM Transactions on Graphics*, Vol. 35, No. 4, pp. 1–19 (2016).

[3] Fukumoto, M.: Silentvoice: Unnoticeable voice input by ingressive speech, *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, pp. 237–246 (2018).