

音声センシングと機械学習を活用した 人流・属性推定システムの検討

長谷川 悠貴[†] 山本 寛[†]

[†] 立命館大学 情報理工学部 情報コミュニケーション学科

1. はじめに

近年、商業・防災分野では人流観測技術の需要が高まっている。しかし、カメラにより撮影された映像を解析する既存技術はプライバシーの問題から敬遠される傾向にある。また、人が所有するスマートフォンなどの無線機器から発信されるビーコンを観測して人流を推定する既存技術では、人の属性(年齢・性別)までは推定できず、マーケティングなどに応用することは難しい[1]。そこで本研究では、マイクにより録音した音声データからプライバシーに関する情報を除去し、その結果を機械学習により解析し、リアルタイムに人の属性・数を推定するシステムを提案する。

2. 音声による人の属性・数推定システム

提案システムの全体像を図 1 に示す。センシング機器は全指向性のマイクを備えた小型ボード PC (Raspberry Pi) であり、周辺の音の録音とメル周波数ケプストラム係数(MFCC)による特徴量抽出を行う。この特徴量(12次元ベクトル)は解析サーバへ送信され、機械学習技術を用いてセンシング機器周辺の人の属性・数の推定が行われる。

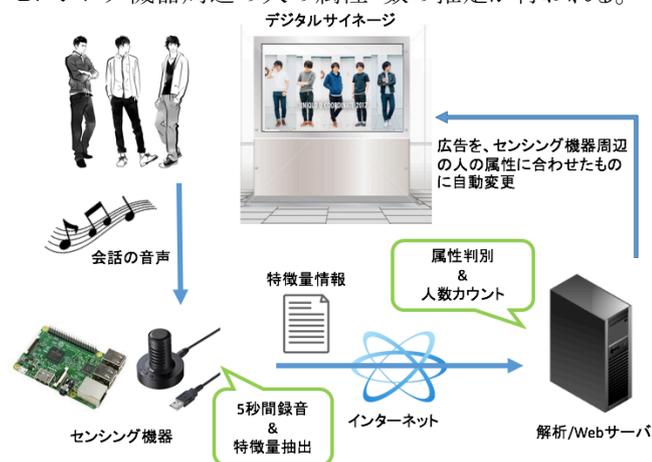


図1. 提案システムの全体像

3. 機械学習とコサイン類似度を用いた推定手法

本提案システムでは、様々な属性(例:20代男性、20代女性、40代男性)の人に対する特徴量を記録した訓練データを用意し、入力された特徴量から対応する属性を推定できるSVM(Support Vector Machine)の学習モデルを構築する。また、センシング機器は音声を0.5秒ごとに録音して音声ファイルを生成し、解析サーバは10個の音声ファイルを一度に解析しており、各音声ファイルに録音されている声に対応する人の属性を推定する。

加えて、センシング機器周辺の人数を推定するには、各音声ファイルに対応する属性を推定するだけでなく、同一人物の声が録音されている音声ファイル群を特定する必要がある。そこで本提案システムでは、声道の形状といった人に固有の特徴が含まれているMFCCによる特徴量を音声ファイル間で比較する。具体的には、音声ファイル間で12次元ベクトルである特徴量のコサイン類似度を算出し、この値が事前に設定された閾値(例:0.85)を超えている場合には、それらの音声ファイルに記録されている音声は同一人物が発したものと判断する。

4. 実証実験

室内に設置したセンシング機器の前で、表1に示す4つのグループに会話を行ってもらい、その際に録音した音声ファイル(10個×8セット)を用いて、提案システムの有効性を評価する実験を行う。表1に示す通り、属性推定では、多くの音声ファイルが「40代男性」と誤って判定された「20代男性」を除いては60%以上の推定精度となり、人数推定では、人数が少ない場合には高い精度で推定できていることが分かる。

表1. 属性推定と人数推定の性能評価

	20代男性1人	20代女性1人	20代男性3人 20代女性2人	20代男性4人
属性推定	43.8%	61.0%	68.8%	40.0%
人数推定	100.0%	90.0%	0%	66.7%

5. まとめと今後の予定

本研究では、音声データから抽出できる特徴量を機械学習により解析することで、プライバシーに配慮しつつ人数・属性を推定するシステムを提案した。今後は、機械学習の訓練データを増やして、属性推定の精度向上を図るとともに、人数が多い場合でも正確な人数推定を行うための新たなアルゴリズムの検討や録音時間の調節を行う。

本研究では、SVMの学習モデルを構築するために、国立情報学研究所音声資源コンソーシアムから提供されている「東北大-松下 単語音声データベース(TMW)」、「CIAIR 子供の声データベース(CIAIR-VCV)」、「鶴岡調査音声データベース 91-92(Tsuruoka91-92)」、および「千葉大学 3人会話コーパス(Chiba3Party)」の音声コーパスを利用しました。ここに感謝の意を表します。

参考文献

[1] 望月祐洋, 他, “Wi-Fi パケット人流解析システムの実環境への適用”, DICOMO2014 論文集, pp.1249-1257, 2014.