Vol.2019-HCI-183 No.10
Vol.2019-EC-52 No.10
2019/6/10

ユーザの機器操作時の音声情報を用いたとまどい推定手法の検討

各務拓真[†1]　葛岡英明[†2]　原田悦子[†3]　田中伸之輔[†4]

概要：家電製品のような機器を開発する場面においては，ユーザビリティを考慮したインタフェースデザインが重要である．ユーザビリティテストはユーザビリティの評価に有効である一方で，分析に要するコストが問題として指摘されている．本論文では，ユーザビリティの評価を支援することを目的に，ユーザのとまどいを音声情報から推定する手法の検討を行った．実装したシステムでは，ユーザの音声から音響特徴量を抽出し，ユーザのとまどいを機械学習により推定した．抽出された音響特徴量を用いて，サポートベクタマシン（SVM）または，k 近傍法による分類器を構成した．評価実験において，機械学習アルゴリズム，音声データセット，使用する音響特徴量に関してとまどい推定精度の調査を行った．その結果，RBF カーネル SVM を用いて男性のみから構成される音声データに対し MFCC 特徴量で分類を行った時，5 重交差検証において最大で 89.9%の精度が得られた．

Investigation of Confusion Estimation Methods Using Voice Information During Appliance Operation of a User

TAKUMA KAKAMI[†1]　HIDEAKI KUZUOKA[†2]
ETSUKO T. HARADA[†1]　SHINNOSUKE TANAKA[†1]

1. はじめに

家電製品の多機能化・高性能化が進むのに伴って，使いにくい家電製品が多くなっている[1]．家電製品のユーザビリティに問題があると，ユーザが製品の機能を使いこなすことができなくなるだけではなく，ユーザの主観的満足度に大きく影響を与えることが指摘されている[2]．このため，家電製品に限らず人が利用することを想定したシステムの開発において，ユーザビリティを考慮したインタフェースデザインの重要性が指摘されており，ユーザビリティの評価手法として，ユーザビリティテストなどの実験的手法が提案されている．

ユーザビリティテストはユーザにタスクを提示した上で，実際にシステムを使用してもらいデータを収集する手法である．テストの際にはビデオ撮影を行うことで，ユーザが困難を感じている様子や，具体的にどのような操作を行っていたのかをテスト後に観察でき，分析に役立てることができる[3]．しかしビデオデータを用いた分析ではコストが高くなることが指摘されており，1 つのビデオデータを分析するためには少なくともテスト時間の 3 倍以上の時間を費やすと言われている[3]．特に長期間に渡るユーザビリティテストでは，ビデオデータの数も膨大となり，前述の分析コストの問題がより大きくなる．

この問題を解決するために，ユーザのフラストレーションやとまどいを推定することでビデオ分析を支援する手法が提案されている．しかし既存の手法の多くは特定の機器に計測装置を組み込んだり，PC 内のシミュレータを使用するなどの方法を採用しているため，汎用性が低かったり．リアリティが低かったりするという問題がある。

本論文では，ユーザビリティの分析を支援するために，ユーザが機器の操作中に感じるとまどいを発話音声から推定するシステムを実装した．本システムで用いた手法では，ユーザの音声データからパワー，基本周波数，MFCC（Mel-Frequency Cepstral Coefficients：メル周波数ケプストラム係数）などを組み合わせた 20 次元の音響特徴量を抽出し，機械学習を用いてユーザのとまどいを推定する．機械学習に関しては，RBF カーネルと線形カーネルを用いたサポートベクタマシン（SVM），k 近傍法の 3 種類の機械学習アルゴリズムによって構成した分類器を用いてとまどい推定システムを実装し，各分類器のとまどい推定の精度を調査する．そして，音声情報からユーザのとまどいを推定することによる分析作業の効率化への可能性について検討する．

2. 関連研究

2.1 思考発話法

思考発話法は，実験参加者に「課題を達成する間に頭に浮かんだことをすべて語る」よう指示し，その条件下でシステムを使うユーザビリティテストの代表的な手法の一つである[4]．思考発話によって観察者は，参加者のインタフェースを用いた操作とその操作をした理由を知ることが可

†1 筑波大学
University of Tsukuba
†2 東京大学
University of Tokyo

Vol.2019-HCI-183 No.10
Vol.2019-EC-52 No.10
2019/6/10

能となる[3][5]. 本研究で実装するとまどい推定システムでは，思考発話法を用いたユーザビリティテストを対象とし，参加者の発話を音声データとして収集し，とまどいを推定することで分析を支援する．

2.2 ユーザビリティの評価を支援する研究

ユーザビリティの評価を支援する研究はこれまでにも行われてきた．Reeder ら[5]はキーボードとマウスを用いたシステムにおいて，取得したユーザの操作イベントから hesitation を検出しユーザビリティの評価を支援できることを示した．Qi ら[6]はマウスに圧力センサを取り付け，機械学習を用いることでユーザがシステム操作中に感じた frustration を圧力データから検出できることを示した．Reeder らと Qi らはユーザの心的負荷の状態を推測することがインタフェースの欠陥を検出する際に有効であることを示している[5][6]．また，ユーザの心的負荷と生理反応に関する研究はこれまでに多くされており，心拍数，発汗，眼球運動などがユーザの心的負荷を推定する生理指標として有効であることが報告されている[7][8]．しかし，長期的ユーザビリティテストでは，ユーザにシステムを預けて使用してもらう状況が考えられる．この場合，操作の度に人体にセンサを取り付けることは適切ではない．したがって，これらに代わる方法で心的負荷を推測する必要がある．この問題に対し，本研究では，ユーザビリティテストのビデオデータからユーザの音声情報を抽出し心的負荷を推測する．

2.3 音声感情認識についての研究

音声対話システムなど音声に関連したインタフェースやインタラクションを研究する上で，ユーザの発話に含まれる感情を認識する技術は重要となっている．このような背景から音声情報から話者の感情を推定する研究は古くから行われている．音声から抽出される音響特徴量には，音声の大きさに関するもの（平均パワー，ラウドネス，ゼロ交差数など）や音声のスペクトルに関するもの（MFCC，スペクトル形状に関するものなど），発声に関するもの（基本周波数，ジッタ，シマーなど）がある．音声感情認識ではこれらを組み合わせたものが主に用いられている[9]．

得られた音響特徴量を用いて音声を「喜び」「怒り」といったカテゴリで表現するが，この場合特徴量ベクトルが所属するカテゴリを推定するパターン認識問題となり，サポートベクタマシン（SVM）やニューラルネットワーク，隠れマルコフモデル（HMM）などが統計的分類器として主に用いられている[10]．しかし，音響特徴量は話者の性別や言語などに大きく影響され，また，どの音響特徴量が感情とどの程度強い関連性があるか，感情推定に有効であるかは明らかではない[9][10]．

3. ユーザビリティテストのビデオ分析

本研究に先行してみんなの使いやすさラボ[11]にて実施されたユーザビリティテストのビデオ分析を行い，その後，テスト参加者のとまどい音声と通常時音声からなるデータセットを構成した．はじめに，ユーザビリティテストの概要を以下に示す．

3.1 ユーザビリティテスト概要

実験参加者 高齢者 12 名（男性 6 名，女性 6 名：年齢平均=72.42，SD=4.76），若年者 6 名（男性 3 名，女性 3 名：年齢平均 ＝39.33，SD=5.43）の参加を得た．

実験対象 レンジ，オーブン，過熱水蒸気調理等の機能を，手動，自動で行えるオーブンレンジを用いた（HITACHI MRO-TW1,MRO-VW1，図 1）．

実験課題 手動，自動調理機能を使用する調理課題を行った．また，課題は，個別に遂行する課題と，高齢者 1 名と若年者 1 名の参加者が協同で遂行する課題を行った．調理課題の一例を表 1 に示す．

実験手続き 参加者は実験について説明を受け，実験者による実演を含む発話思考法の説明・練習を行った後に，ユーザビリティテストの課題を行った．

図 1 ユーザビリティテスト対象のオーブンレンジ

Vol.2019-HCI-183 No.10
Vol.2019-EC-52 No.10
2019/6/10

表 1　実験手続きの一例

1.　実験説明,発話思考法説明・練習
2.　個別ユーザビリティテスト
課題 1-1 食品あたため（手動あたため課題）
課題 1-2 温めた食品を開封，保存容器へ移す
課題 2　牛乳あたため（自動調理機能課題①）
課題 3　冷凍鶏肉解凍（自動調理機能課題②）
課題 4　鶏のハーブ焼き調理（自動調理機能課題③）
3.　協同ユーザビリティテスト（協同問題解決課題）
課題 5　メニュー選択 （味噌蒸し，鶏の唐揚げ，筑前煮）
課題 6　選択したメニューの調理

3.2 ビデオデータの分析

前節のユーザビリティテストで得られたビデオデータから参加者の発話・行動を書き起こした．書き起こしでは，「発話」,「行動」,「人工物の状態」,「エラー」項目に対し，時系列に沿って要素を書き入れた表を作成した．書き起こしデータに基づき,参加者が操作にとまどっている箇所と，そうでない箇所に分類し，それぞれの音声サンプルをビデオデータから切り出した．これら音声サンプルの切り出しには，オーディオ編集ソフトの Audacity（Windows 版）を使用した．表 2 に各参加者から得られた音声サンプルの概要を示す.

表 2 実験参加者と得られた音声データ

参加者 ID	性別	とまどい 音声(sec)	正常 音声(sec)
AO01	男性	62	140
AO02	男性	67	119
AO03	男性	33	64
AO04	女性	77	124
AO05	女性	62	98
AO06	女性	135	107
AY01	男性	73	60
AY02	男性	43	110
AY03	男性	22	18
AY04	女性	41	93
AY05	女性	16	52
AY06	女性	3	57
TO01	男性	82	129
TO02	男性	64	135
TO03	男性	66	87
TO04	女性	18	69
TO05	女性	93	66
TO06	女性	80	115
合計		1046	1652

4.　実装システムの概要

本研究で実装したシステムの概要を図 2 に示す．本システムはユーザビリティテストで得られたテスト参加者の発話音声を入力として，教師あり学習アルゴリズムよってテスト音声を 2 つのカテゴリ（とまどい/正常）に分類し出力するものである．本システムは，オープンソースソフトウェア openSMILE [13]を用いた音響特徴量の抽出・処理部と機械学習プログラムから構成される．教師データとして入力される音声データには，事前にとまどいの有無がラベリングされているものとする．

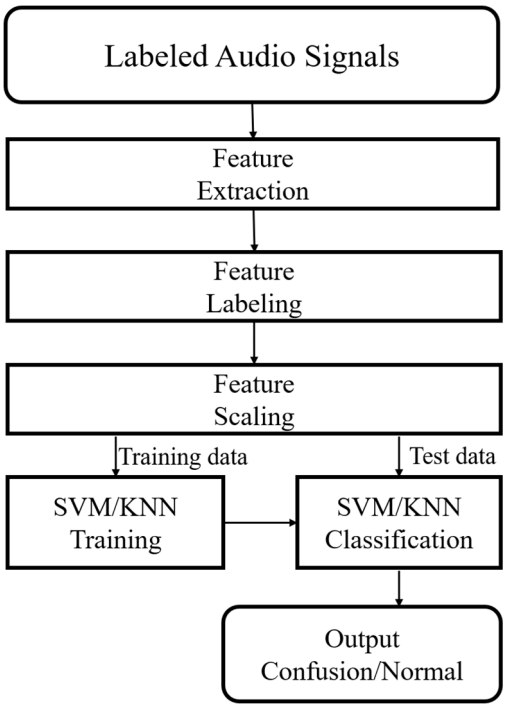

図 2 システムの構成

4.1　特徴量の抽出

音響特徴量の抽出にはオープンソースソフトウェア openSMILE[13]を用いた．音声データは初めに窓関数によってフレーム幅 300 ms，100 ms ずつずらしたフレーム列に分割され，その後，各音声フレームに対し音響特徴量が計算される．本研究で用いた音響特徴量を表 3 に示した．こうして抽出された特徴量は 12 次元のメル周波数ケプストラム係数（MFCC），振幅の RMS 値，基本周波数などと，MFCC を除いた音響特徴量の一次微分から構成される 20 次元のベクトルデータである．

表 3 openSMILE で取得した特徴量

特徴量	説明
RMSenergy	振幅の二乗平均平方根値
MFCC	メル周波数ケプストラム係数（12 次元）
F0	基本周波数

Vol.2019-HCI-183 No.10
Vol.2019-EC-52 No.10
2019/6/10

voiceProb	その時点での音が声である確率（自己相関関数と基本周波数から計算される）
pcm_zcr	ゼロ交差率
（および MFCC 以外の各特徴量の一次微分）	

(1) 発声・音高に関する特徴量

発話音声の音高成分は声帯の緊張と振動に依存することから感情に関する情報を含んでいると考えられている[14][15]．実装システムでは基本周波数（F0）と，自己相関関数と基本周波数から計算される Voicing Probability，およびそれらの一次微分を合わせた 4 次元値を発声に関する特徴量として用いる．

(2) 音声の大きさに関する特徴量

音声の大きさもまた感情に関する情報を含むとされている韻律的特徴量である[14]. 実装システムでは，振幅の二乗平均平方根（RMS）値とゼロ交差率，それらの一次微分を合わせた 4 次元値を音声の大きさに関する音響特徴量として用いる．

(3) 音声スペクトルに関する特徴量

MFCC（Mel-Frequency Cepstral Coefficients : メル周波数ケプストラム係数）は音声認識の研究で広く用いられている特徴量である．人間の音高に対する聴覚特性は，低周波数帯域では細かく，高周波数域では粗い周波数分解能を有していることが知られている[16]．メルケプストラムはこのような聴覚特性を表す非線形周波数軸上で定義されたケプストラムであり，通常のケプストラムの半分程度の次数で音声スペクトルを表現することができる[17]. 図 3 に MFCC 特徴量の抽出過程の概要を示した．音声信号から MFCC を計算する過程は以下の通りである．

1. プリエンファシスフィルタによって高周波数成分を強調する．
2. 窓関数を適用した後に，フーリエ変換し振幅スペクトルを求める．
3. 振幅スペクトルにメルフィルタバンクを適用する．
4. 離散コサイン変換をし，得られたメルケプストラムの低次成分が MFCC となる．

なお，実装システムではメルケプストラムより 12 次元の係数を取り出し，音声スペクトルに関する特徴量として用いる．

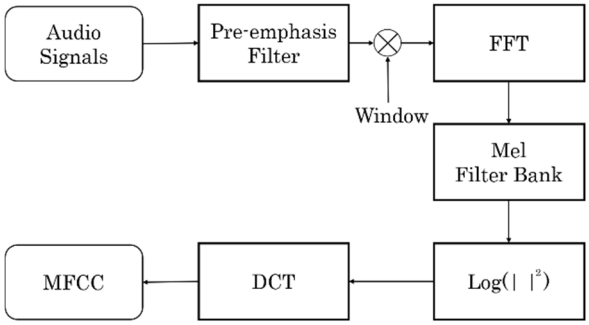

図 3 MFCC の抽出

以上のように得られた教師データの特徴量ベクトルに対し，抽出元となった音声データに基づいてとまどいの有無がラベリングされる．また，教師特徴量データ，テスト特徴量データの各特徴量に対して式(１)の標準化を施す．

$$x' = \frac{x - \mu}{\sigma} \quad (1)$$

ここで，μ は教師データにおける対象の特徴量 x の平均値，σ は教師データにおける対象の特徴量 x の標準偏差を表す．

4.2 機械学習アルゴリズム

本研究のとまどい推定システムで用いた機械学習アルゴリズムについて説明する．本研究では精度比較のため，サポートベクタマシン（以下 SVM）と k 近傍法を用いて分類器を構成した．SVM を用いた分類器では，カーネル関数に放射基底関数（RBF）カーネルと線形カーネルの 2 種類を用いた分類器を構成した．

得られた教師データを用いて汎化能力の高い分類器を構成するためには，ハイパーパラメータが適切に決定されている必要がある．線形カーネルを用いた SVM では正則化の強さ C，RBF カーネルを用いた SVM の場合，これに加えて分類境界の複雑さγを決定する必要がある．また，k 近傍法を用いた分類器の場合は参照近傍数 K を決定する．これらの決定には，5 分割交差検証によって，とまどい推定への精度が最も高くなるパラメータを用いる. このとき，C の値は 2.0～5.6 の 5 種類，γの値は 0.03～1 の 4 種類，K の値は 2^2～2^9 の 8 種類から決定した．

4.3 とまどい推定

事前に得られている他ユーザのユーザビリティテストのビデオデータに対しては，人手によって音声データの切り出しととまどいの有無のラベリングが行われたものとし，これを教師データとして用いる．新規に得られたユーザの音声データに対してとまどい推定を行う際には, SVM および k 近傍法による 2 クラス（とまどい/正常）分類を行う．機械学習アルゴリズムの実装には Pedregosa らによって提供される python 用の機械学習ライブラリの Scikit-learn[13] を用いた．

Vol.2019-HCI-183 No.10
Vol.2019-EC-52 No.10
2019/6/10

5. 評価実験

本研究では，ユーザのとまどいを音声情報から推定することに対する有効性の検証を目的に，構成したとまどい推定モデルの精度について調査した．3 章で構成した音声データセットを入力とし，とまどい音声からは合計 10462 個（男性：5164 個，女性：5297 個），正常音声からは 16519 個（男性：8654 個，女性：7864 個）の特徴量ベクトルが抽出された．

5.1 音声データの評価

実装システムの分類モデルの精度を評価するための指標として accuracy と F 値（Recall と Precision の調和平均）を用いる．これらの評価値は 5 分割交差検証を行って算出される．S 分割交差検証は図 4 に示した通り，得られた特徴量データを S 分割し，1 つの分割をモデルの評価セットとし，残りの分割を教師データとして分類器の構成に用いる．これを S 回繰り返し評価値の平均を取ることでモデルの評価をする手法である．

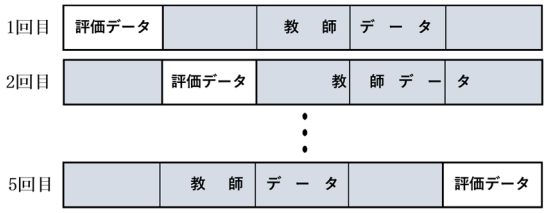

図 4　5 分割交差検証

とまどい推定モデルの評価実験は，発話者の性別・使用する特徴量・分類器がとまどい推定精度に与える影響を調査することを目的に，複数の条件に分けて実施した．発話者の性別に関しては，男性のみの音声データ，女性のみの音声データ，男女混合の音声データの 3 条件に対して行った．また，とまどい推定に用いる音響特徴量に関しては，MFCC のみ，音声の大きさ・発声に関する特徴量（Pitch + Energy），その両方を組み合わせた場合（MFCC + Pitch + Energy）の 3 条件に分け，各条件に対して評価値を算出した．この他，前述のように，線形カーネル，RBF カーネルの SVM と k 近傍法を用いた各分類器に対して評価値を算出した．したがって合計 27 条件で評価を行った．

6. 実験結果と考察

構成したとまどい推定システムに対して評価実験を行い，前述の各条件の下で評価値を算出した．最も高い精度が得られたのは，RBF カーネル SVM により特徴量として MFCC のみを用いて男性のみの音声データに対してとまどい推定を行った場合で，Accuracy：89.9%，F 値：0.861 の精度でとまどいを推定することができた．この時の混同行列を表 4 に示す．以下，各条件に基づいて結果を考察する．

表 4　RBF カーネル SVM，MFCC を用いた男性のみの音声データに対するとまどい推定の混同行列

		Confusion Recognition (%)	
		Confuse	Normal
Actual	Confuse	82.9	17.1
	Normal	6.5	93.5

(1) 音響特徴量に関する結果・考察

とまどい推定に使用する音響特徴量による推定精度について考察する．評価実験の結果を表 5 と表 6 に示す．すべての分類器，音声データにおいて Pitch + Energy を特徴量として用いた場合に F 値，accuracy 共に最も低くなった．MFCC のみを用いた場合と，MFCC + Pitch + Energy を用いた場合の評価値の間にはほとんど差は見られなかった．以上の結果より，本研究のとまどい推定モデルでは音声の大きさ・音高に関する音響特徴量は推定精度の向上にはあまり寄与しないと考えられる．

表 5 RBF カーネル SVM を用いた男女混合音声に対するとまどい推定精度

	Accuracy (%)	F 値
MFCC	87.4	0.834
Pitch + Energy	70.8	0.562
MFCC + Pitch + Energy	86.7	0.817

表 6 k 近傍法を用いた男女混合音声に対するとまどい推定精度

	Accuracy (%)	F 値
MFCC	83.1	0.753
Pitch + Energy	69.8	0.576
MFCC + Pitch + Energy	82.7	0.745

(2) 発話者の性別に関する結果と考察

音声データを構成する発話者の性別がとまどい推定モデルの精度に与える影響について考察する．評価実験結果の一部を表 7 と表 8 に示す．男性のみから構成される音声データを用いて，分類器の構成・とまどい推定を行った場合では，他の音声データを用いた場合と比べて，わずかに精度が高くなる傾向があった．

表 7　RBF SVM を用いた Pitch + Energy によるとまどい推定の精度

	Accuracy (%)	F 値
General	70.8	0.562
Male	75.6	0.640
Female	68.0	0.563

表 8　k 近傍法を用いた MFCC によるとまどい推定の精度

	Accuracy (%)	F 値
General	83.1	0.753
Male	85.4	0.780
Female	83.0	0.762

(3) 分類器に関する結果と考察

分類器を構成するために使用した機械学習アルゴリズムを変えた場合の評価値を表 9 と表 10 に示す．多くの条件で，線形カーネル SVM によってとまどい推定した場合に精度の低さが顕著に見られるものの，音響特徴量として Pitch + Energy を用いた場合には，分類器の構成に用いたアルゴリズムによる差は見られなかった．

表 9　男女混合の音声データを対象にした MFCC を特徴量に用いたとまどい推定の精度

	Accuracy (%)	F 値
RBF SVM	87.4	0.834
Linear SVM	70.6	0.614
K-NN	83.1	0.753

表 10　男女混合の音声データを対象にした Pitch + Energy を特徴量に用いたとまどい推定の精度

	Accuracy (%)	F 値
RBF SVM	70.8	0.562
Linear SVM	69.1	0.610
K-NN	69.8	0.576

7. おわりに

本論文では，ユーザビリティの分析支援における音声を用いたとまどい推定の有効性を検討することを目的として，3 種類の音声データセット、3 種類の機械学習手法によってとまどいの推定精度を比較した．音声データの特徴量として，MFCC，基本周波数，RMS 値，ゼロ交差率，Voicing Probability，およびそれらの一次微分を抽出した．これらを用いて, SVM および k 近傍法によってユーザのとまどいを推定するシステムを実装し，分類器の構成に用いる機械学習アルゴリズムや特徴量の組み合わせ，テスト参加者の性別によるとまどい推定精度の評価実験を行った．評価実験の結果，RBF カーネル SVM により MFCC を特徴量として男性のみの音声データに対してとまどい推定を行った場合，最大で 89.9%の Accuracy，0.861 の F 値という高い精度でとまどいを推定することができることが確認された．この結果から発話音声によってユーザのとまどいを推定することに有効である可能性が示された．しかし，実際にユーザビリティテストにおいて，実装手法によるとまどい推定がユーザビリティの評価の効率向上に対して有効であるかについては検討しておらず，今後の課題としたい．

謝辞

本研究を行うにあたり，熱心なご指導やご助言を頂きました，葛岡英明教授に深く感謝いたします．また，様々な面でご助言くださり，ご協力いただいた川口一画氏と大槻麻衣助教，原田悦子教授，田中伸之助氏に心より感謝いたします．

参考文献

[1]Jokela, T. When good things happen to bad products: where are the benefits of usability in the consumer appliance market? Interactions, 11, 6, 28-35 (2004).
[2]Marks, J.: The usability problem for home appliances: engineers caused it, engineers can fix it!, ACM SIGSOFT (2005).
[3]ニールセン，篠原監訳・三好訳: ユーザビリティエンジニアリング原論 ユーザのためのインタフェースデザイン，東京電機大学出版局(2002).
[4]海保・原田編: プロトコル分析入門―発話データから何を読むか,光明社 (1993)
[5]樽本: ユーザビリティエンジニアリング―ユーザ調査とユーザビリティ評価実践テクニック，オーム社 (2005).
[6]Reeder, R., Maxion, R.: User Interface Defect Detection by Hesitation Analysis; In Proc. DSN 2006, IEEE, pp61-72 (2006).
[7]Qi, Y., Reynolds, C., and Picard, R.: The Bayes Point Machine for computer-user frustration detection via pressuremouse. In Proc. PUI 2001, ACM Press, pp. 1-5 (2001).
[8]水科晴樹，阪本清美，金子寛彦:課題遂行時の作業負荷により誘発された心理的ストレスとサッカード眼球運動の動特性との関係，電子情報通信学会論文誌 D, J94-D(10), pp.1640-1651 (2011)
[9]梅野克身,浜出絵理子,横井秀輔,堀悦郎,小野武年,西条寿夫: 精神ストレス負荷時の自律神経反応と手掌からの皮膚喪失水分量(TEWL)との相関性，The Autonomic nervous system 43(5), pp.416-423 (2006)
[10]鈴木基之：音声に含まれる感情の認識，日本音響学会誌 71(9), pp. 484-489 (2015)
[11]原田: みんラボ,発進 ： 高齢者のための使いやすさ検証実践センターについて，人間生活工学 13(1), pp.71-74.(2012).
[12]Ayadi, M., Kamel, M., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44, pp. 572-587 (2011)
[13]F. Eyben, M.Wöllmer, B. Schuller, : openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor, ACM Multimedia Conference – MM, pp. 1459-1462 (2010).
[14]Scikit-learn: Machine Learning in Python, Pedregosa et al, JMLR, Vol. 12, pp. 2825-2830 (2011).
[15]Iker, L., Navas, E., Hernaez, I., Sanchez, J.: Emotion Recognition

Vol.2019-HCI-183 No.10
Vol.2019-EC-52 No.10
2019/6/10

using Prosodic Parameters, Interspeech, pp. 433-442.(2005).
[16]Carlos, B., Lee, S., Narayanan, S.: Analysis of Emotionally Salient Aspects of Fudamental Frequency for Emotion Detection, IEEE Transactions on Audio, Speech, and Language Processing 17(4), pp. 582-596(2009)
[17]古井:音声情報処理,森北出版 pp.85-86(1998).
[18]Keiichi, T., Kobayashi, T., Fukuda, T., SAITO, H., IMAI, S.: Spectral Eastimation of Speech Based on Mel-Cepstral Representation, 電子情報通信学会論文誌 A 74(8), pp. 1240-1248(1991).