Vol.2020-ICS-197 No.2
2020/2/17

特殊詐欺音声を対象とした韻律的特徴量の考察

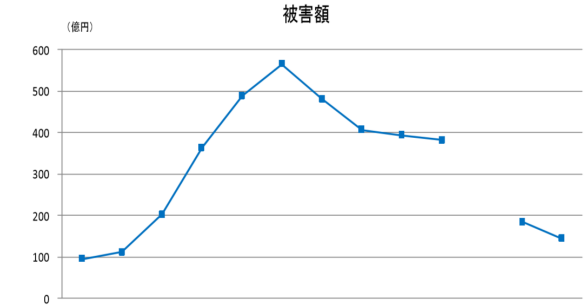
徳島 大河[†1,a)]　清 雄一[1,b)]　田原 康之[1,c)]　大須賀 昭彦[1,d)]

概要：特殊詐欺は，平成 30 年の被害額が 382 億円に及ぶなど，依然として深刻な社会問題である．注意喚起等はなされているが，被害は収まっていない．このような現状の原因に一つには，特殊詐欺の内容の高度化，複雑化があると言える．親族を装う「オレオレ詐欺」だけではなく，役所の職員や警察官などを装う詐欺も被害が確認されている．このような特殊詐欺の被害を少なくするためには，特殊詐欺の検出を行うことが有効であると考えられる．特殊詐欺の検出には，犯人の発話に含まれるキーワードからの検出や，被害者側のストレス状態を検出するといった手法がある．そこで本研究では犯人の声，所謂犯人の発話の韻律的特徴量に着目し，詐欺を働く際の音声を韻律的特徴量を用いて考察を行った．一般人の詐欺を働く発話と通常の会話をするときの発話を収集し，韻律的特徴量を抽出して SVM 分類を行った．収集した音声で作った分類器で，一般人の音声を詐欺を働いているのかどうかで分類することはできたが，実際の犯人の音声の分類は良い結果とならなかった．それらの分類結果をもとに，一般人が詐欺を働く際の声と実際の犯人の声の違いも含め考察を行った．

1. はじめに

1.1 背景

特殊詐欺とは，“犯人が電話やハガキ（封書）等で親族や公共機関の職員等を名乗って被害者を信じ込ませ、現金やキャッシュカードをだまし取ったり、医療費の還付金が受け取れるなどと言って ATM を操作させ、犯人の口座に送金させる犯罪”のことである [1]．特殊詐欺は，平成 26 年に被害額がピークを迎え，現在でも 400 億円近い被害が出ている [2]（**図 1**）．このような状況を改善するため，警察らによる注意喚起等がなされ，図 1 のように被害額は減少の様子が見られるが，特殊詐欺の件数自体は減少していないのが現状である（**図 2**）．警察庁によって公開されている特殊詐欺の手口別認知件数の推移によると，詐欺の内容は複雑化し，種類も増加しているのが分かる．この詐欺の内容の複雑化や種類の増加が詐欺被害の減らない原因の一つであると考えられる．詐欺の内容，シナリオは洗練され，見破るのが難しくなっているのが現状である．実際にテレビ番組の企画で，巧妙な手口の詐欺を見破ることができるのか，というコーナーがあるほど昨今の特殊詐欺はシナリオが高度なものになっている [3]．



区分＼年次	H21	H22	H23	H24	H25	H26	H27	H28	H29	H30	H30 上半期	R1 上半期
被害額	95.8	112.5	204.0	364.4	489.5	565.5	482.0	407.7	394.7	382.9	185.9	146.1

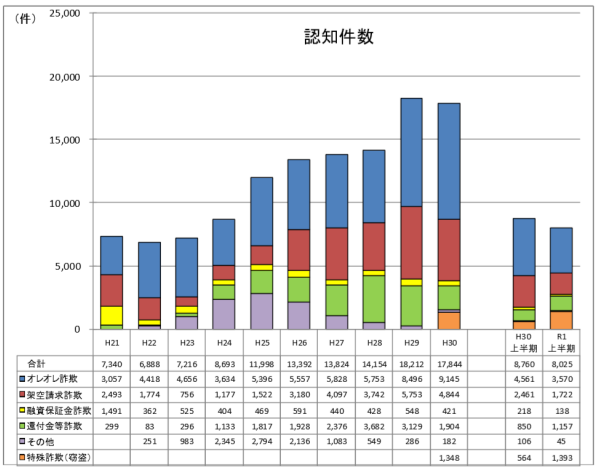
図 1　詐欺被害額の推移 [*1]

1.2 目的

このように高度化，複雑化する特殊詐欺の被害を抑えるには，電話を受ける人が注意するだけでは困難である．電話を受ける人だけではなく，冷静に判断を下す第三者が存在することで詐欺の被害は抑えることができると考えられる．そこで，詐欺かどうかを機械に判別させることを目指すこととした．先述の通り特殊詐欺は内容が高度化，複雑化しているため，内容面を見るだけでは判別するのが困難

[1] 電気通信大学
The University of Electro-Communications
[†1] 現在，電気通信大学
Presently with The University of Electro-Communications
[a)] tokushima.taiga@ohsuga.lab.uec.ac.jp
[b)] seiuny@uec.ac.jp
[c)] tahara@uec.ac.jp
[d)] ohsuga@uec.ac.jp

[*1] 出典) 警察庁　広報資料「令和元年上半期における特殊詐欺認知・検挙状況等について」 p-2
[*2] 出典) 警察庁　広報資料「令和元年上半期における特殊詐欺認知・検挙状況等について」 p-3



	H21	H22	H23	H24	H25	H26	H27	H28	H29	H30	H30 上半期	R1 上半期
合計	7,340	6,888	7,216	8,693	11,998	13,392	13,824	14,154	18,212	17,844	8,760	8,025
オレオレ詐欺	3,057	4,418	4,656	3,634	5,396	5,557	5,828	5,753	8,496	9,145	4,561	3,570
架空請求詐欺	2,493	1,774	756	1,177	1,522	3,180	4,097	3,742	5,753	4,844	2,461	1,722
融資保証金詐欺	1,491	362	525	404	469	591	440	428	548	421	218	138
還付金等詐欺	299	83	296	1,133	1,817	1,928	2,376	3,682	3,129	1,904	850	1,157
その他		251	983	2,345	2,794	2,136	1,083	549	286	182	106	45
特殊詐欺（窃盗）										1,348	564	1,393

図 2 詐欺手口別認知件数の推移 *2

であることが予想される．そこで本研究では内容のみから詐欺かどうかを判別するのではなく，犯人の声からも詐欺検出を行うことを目指し，その前段階として犯人の発話を韻律的特徴量を用いて考察することを目的とした．

2. 既存研究・既存事例

詐欺検出の既存研究では，内容面に注目し詐欺に関連するキーワードや文脈などで検出を行っているものがある．2019 年 8 月にも NTT グループが特殊詐欺のリアルタイム検出の実証実験について発表を行った [4]．NTT グループのものは，犯人の音声をテキスト化し，言語分析によって不審な表現があるかどうかを判別する．音声から発話内容を抽出し，内容に注目することで詐欺判別を行っている．

また，松尾らの研究 [5] では，内容だけではなく韻律的特徴量も用いて詐欺判別を行っていた．松尾らの研究では，犯人側の音声から詐欺関連のキーワードを検出しその数を用いて判別を行う．更に，キーワードの検出数に加えて，被害者側の音声から被害者側のストレス状態を検出し，ストレスを感じているかどうかを加えて詐欺検出を行っている．韻律的特徴量を使用して詐欺判別を行ってはいるが，使用したのは被害者の音声の韻律的特徴量のみで，犯人側の音声の韻律的特徴は使用していない．松尾らの研究では，ストレス状態の検出はかなりの精度で行えていた．

詐欺判別に犯人の発話の韻律的特徴量を使用している既存研究は確認できず，いずれも犯人側の発話内容を主体に詐欺判別を行っていた．よって本研究では，犯人の韻律的特徴量に注目し，詐欺音声を韻律的特徴量を用いて考察を試みる．

3. アプローチ

犯人の発話の韻律的特徴量に着目し，詐欺を行う際の音声について韻律的特徴量を用いて考察を行う．過去に詐欺を行ったことのない一般人による，詐欺を試みる時と通常時の声の二種類で分類器を作成し，実際の犯人の音声を分類することで考察する．

表 1 使用した韻律的特徴量

特徴量名	Δ 成分	素性値 (12 種)
RMS energy	Δ RMS energy	max. min, range
MFCC 1-12	Δ MFCC 1-12	maxPos, minPos, amean
ZCR	Δ ZCR	linregc1. linregc2, linregerrQ
voice Prob	Δ voice Prob	stddev
F0	Δ F0	skewness, kurtosis

3.1 用いる犯人音声について

考察対象として分類実験等に用いる犯人音声は，各都道府県警察が web ページ上に公開しているものを使用した．

3.2 分類器作成のために用いた音声について

犯人音声を分類対象とした分類器の作成のため，学習データとして一般人の発話を収集した．収集した発話は，通常の会話をしている平常時の発話と，詐欺を試みている際の発話の二種類である．

3.3 分類について

収録した二種類の音声を用いて分類器を作成し，実際の犯人の音声を対象として分類実験を行った．犯人の音声が，収録した二種類の音声のうち詐欺を試みている時の発話に近いのか通常時の発話に近いのかを分類し，その結果から犯人の音声を韻律的特徴量を用いて考察する．

3.4 韻律的特徴量

韻律的特徴量とは，パワーやピッチなど，音声に関連する特徴量のことである．本研究では，INTERSPEECH 2009 Emotion Challenge[6] で使用された特徴量を韻律的特徴量として用いた．この特徴量は，音声からの感情認識に関する研究で多く使用されている特徴量であり，本研究では詐欺判別のために用いた．特徴量の一覧は**表 1** の通り．

特徴量の内容は，エネルギーの二乗平均平方根，12 次元の MFCC（メル周波数ケプストラム係数），ゼロ交差率，Voice probability（全パワーにおける調波成分の割合），基本ピッチ周波数の計 16 次元とその Δ 成分，その 32 次元に対して 12 種類の素性値（最大値，最小値，最大値と最小値の差，最大値を出力した場所，最小値を出力した場所，算術平均，線形近似の勾配度，線形近似のオフセット，線形近似の二乗誤差，標準偏差，歪度，尖度）をとった計 $32 \times 12 = 384$ 次元である．

韻律的特徴量の抽出には，openSMILE[7] を使用した．openSMILE はミュンヘン工科大学が開発した「音声認識」等の研究によく用いられる音声特徴量抽出のオープンソフトウェアである [8][9][10]．

表 2　学習データと分類対象データの組み合わせ

学習データ	分類対象データ
一般人 20 人分の模倣音声と通常音声（各 1000 ファイルずつ）	一般人 10 人分の模倣音声と通常音声（各 500 ファイルずつ）
一般人 20 人分の模倣音声と通常音声（各 1000 ファイルずつ）	実際の犯人 30 人分の詐欺音声（810 ファイル）
一般人 30 人分の模倣音声と通常音声（各 1500 ファイルずつ）	実際の犯人 30 人分の詐欺音声（810 ファイル）
演技経験者 5 人の模倣音声と通常音声（各 250 ファイルずつ）	実際の犯人 30 人分の詐欺音声（810 ファイル）
演技未経験者 5 人の模倣音声と通常音声（各 250 ファイルずつ）	実際の犯人 30 人分の詐欺音声（810 ファイル）
演技未経験者 5 人の模倣音声と通常音声（各 250 ファイルずつ）	演技未経験者 25 人の模倣音声と通常音声（各 1250 ファイルずつ）
演技未経験者 25 人の模倣音声と通常音声（各 1250 ファイルずつ）	演技経験者 5 人の模倣音声と通常音声（各 250 ファイルずつ）
演技未経験者 25 人の模倣音声と通常音声（各 1250 ファイルずつ）	実際の犯人 30 人分の詐欺音声（810 ファイル）

3.5　分類器の作成

分類器は機械学習オープンソフトウェアである Weka の Support Vector Machine（SVM）を用いて作成した．入力を発話の韻律的特徴量として，出力するクラスを学習に用いた「模倣音声」か「通常音声」の二択とした．学習データ及びテストデータは先述の音声である．

4.　実験

アプローチに記した内容について具体的に記載する．

4.1　音声データの用意

4.1.1　犯人音声データ

犯人音声データは青森県警察，茨城県警察，京都府警察，警視庁，静岡県警察，千葉県警察，長野県警察，兵庫県警察，北海道警察（50 音順）が Web ページ上で公開している音声を収集した．収集した音声の中には著しくノイズの激しいものもあり，分類実験への使用が困難なものは除去した．Savita Sondhi らの研究 [12] を参考に，収集した音声をフリーソフトウェアである Audacity を用いて編集し，ノイズ除去等の処理を施した．Kun Han らの研究 [11] を参考に，音声は 16000Hz をサンプリング周波数とし，ビットレートは 512kbps に揃えた．結果，収集し実際に実験に使用したファイルは，男性の犯人 30 人分，810 個になった．後述の一般人の発話内容に合わせて，音声ファイルは犯人が敬語で話しているもののみで揃えた．いずれのファイルも，3 秒から 20 秒ほどの長さとなっている．これらのファイルを以降「犯人音声」と呼称する．

4.1.2　一般人発話音声データ

一般人の発話に関しては，実際に著者と会話をする形で収録を行った．一般人の発話は二種類を収録している．一つ目は詐欺を試みる際の音声，二つ目は通常の会話をしている際の音声である．しかし，一般人に詐欺を試みてもらうのは困難であるため，本研究では用意した詐欺を模倣したスクリプトをもとに会話することで詐欺を模倣した発話（以降「模倣音声」と呼称）を収録した．詐欺模倣スクリプトは，実際の詐欺を参考に，クレジットカード番号を聞き出そうと店員のふりをする犯人との会話を作成した．この詐欺模倣スクリプトは，事前に発話者に目を通しておくよう依頼した．詐欺を模倣した発話の収録後，雑談をすることで通常時の発話（以降「通常音声」と呼称）を収録した．

以上の模倣音声と通常音声を，20 代から 60 代までの男性 30 人から 50 発話ずつ収録した．男性のみのデータを収集したのは，入手できた犯人音声データがすべて男性の犯人のものだったためである．男性 30 人のうち，5 人は舞台俳優など過去 1 年以内に演技経験のある人物の発話を収録し，その他 25 人は過去 1 年以内に演技経験がない，もしくは演技経験の全くない人物の発話を収録した．収録した発話は，犯人音声と同様にフリーソフトウェアである Audacity を用いて編集し，ノイズ除去等の処理を施した．犯人音声同様サンプリング周波数は 16000Hz，ビットレートは 512kbps に全て揃えた．ファイル数は模倣音声，通常音声ともに発話者一人当たり 50 個ずつ収集した．よって実験に使用したファイル数は模倣音声，通常音声それぞれ 1500 個ずつである．いずれのファイルも 3 秒から 20 秒ほどの長さとなっている．

4.2　韻律的特徴量の抽出

韻律的特徴量の抽出には先述の openSMILE を用いた．INTERSPEECH 2009 Emotion Challenge で使用された 384 次元の韻律的特徴量に，模倣音声か通常音声かを示すクラスを付与した 385 次元の特徴量を抽出した．

4.3　分類器の作成

分類器は weka の SVM で作成した．カーネル関数は多項式カーネルを使用した．多項式カーネルは一般的に次の式で表される．

$$k(x_i, x_j) = (x_i x_j + r)^d$$

分類器の作成には，模倣音声と通常音声を学習データとして用いた．学習データおよび分類対象データの組み合わせは**表 2** の通り．

まずは収録した一般人の音声 20 人で学習し分類器を作成し，残り 10 人の音声を分類した．次に，その分類器を用いて実際の犯人 30 人分の音声を分類した．次に，新た

に一般人の音声 30 人分で分類器を作成し，実際の犯人 30 人分の音声を分類した．そして今までの分類結果を踏まえて，発話者の演技経験の有無によって別々の分類器を作成した．演技経験のある人は 5 人だけだったため，演技経験者 5 人のみで分類器を作成し，犯人 30 人分の音声を分類した．演技経験のない人は 25 人いたが，25 人の音声で分類器を作成すると演技経験者 5 人の分類器との結果が著しく異なってしまうため，学習データの人数は 5 人分の音声に揃えて分類器を作成した．25 人中 5 人で分類器を作成するため，本研究では 25 人中 5 人をランダムに選択し分類器を作成するのを 10 回繰り返し，学習データの異なる分類器 10 個で犯人音声を分類，結果を検証した．演技経験者と犯人の音声，演技未経験者と犯人の音声の違いを確認するために，演技未経験者 25 人の音声で学習した分類器を新たに作成した．演技経験者 5 人のみで作成した分類器で演技未経験者のと犯人の音声の違いを，演技未経験者 25 人のみで作成した分類器で演技経験者と犯人の音声の違いを比較した．

4.4 評価方法

本研究の目的は韻律的特徴量を用いた考察のため，分類率，学習に用いたデータ，特徴量の重み等を総合的に検討した．分類結果から分類器の性能を評価し，その分類器が重視している特徴量などから詐欺に関連する韻律的特徴量などについて考察した．

5. 実験結果

分類率及び重み（係数）が大きな特徴量 10 個を示す．

5.1 一般人 20 人分で学習，10 人分を分類した結果

学習データは一般人音声 20 人分（模倣音声ファイル：1000 個，通常音声ファイル：1000 個）である．分類対象データは一般人音声 10 人分（模倣音声ファイル 500 個，通常音声ファイル：500 個）である．このうち，演技経験者は学習データに 3 人，分類対象データに 2 人含まれている．

分類結果の混合行列は**表 3** の通り．模倣音声が模倣音声，通常音声が通常音声というように正しく分類されたのは 1000 個のファイル中 821 個で，82.1%であった．

5.2 一般人 20 人分で学習，犯人 30 人分を分類した結果

学習データは一般人音声 20 人分（模倣音声ファイル：1000 個，通常音声ファイル：1000 個）である．分類対象データは実際の犯人の音声 30 人分（ファイル数：810 個）である．このうち演技経験者は学習データに 3 人含まれている．

分類結果は**表 4** の通り．多くの詐欺音声が通常音声として分類された．よって詐欺音声は，この分類器の学習した通常音声に近いということが言える．

表 3 分類結果
学習：一般人 20 人分 (模倣，通常各 1000 ファイル)
分類対象：一般人 10 人分（模倣，通常各 500 ファイル)

		分類されたクラス	
		模倣音声	通常音声
実際のクラス	模倣音声	444	56
	通常音声	123	377

表 4 分類結果
学習：一般人 20 人分（模倣，通常各 1000 ファイル）
分類対象：実際の犯人 30 人分（810 ファイル）

	数	割合 (%)
模倣音声	23	2.8395
通常音声	787	97.1605

表 5 分類結果
学習：一般人 30 人分（模倣，通常各 1500 ファイル）
分類対象：実際の犯人 30 人分（810 ファイル）

	数	割合 (%)
模倣音声	24	2.963
通常音声	786	97.037

5.3 一般人 30 人分で学習，犯人 30 人を分類した結果

収録した一般人音声全てを用いて分類器を作成し，犯人音声の分類を行った．学習データは一般人音声 30 人分（模倣音声ファイル：1500 個，通常音声ファイル：1500 個）である．分類対象データは実際の犯人の音声 30 人分（ファイル数：810 個）である．

分類結果は**表 5** の通り．大半の詐欺音声が通常音声として分類された．

5.4 演技経験者 5 人分で学習，犯人 30 人分を分類した結果

多人数で学習した分類器よりも精度は欠くが，演技経験者のみで学習した分類器で犯人音声を分類した．学習データは一般人演技経験者音声 5 人分（模倣音声ファイル：250 個，通常音声ファイル：250 個）である．分類対象データは実際の犯人の音声 30 人分（ファイル数：810 個）である．

分類結果は**表 6** の通り．学習データが少ない分精度は落ちるが，多くの犯人音声がこの分類器が学習した模倣音声に近いという結果となった．

5.5 演技未経験者 5 人分で学習，犯人 30 人分を分類した結果

学習データ数による差をなくすため，演技未経験者においても 5 人分で学習し，分類を行う．演技未経験者については 25 人いたため，5 人をランダムに選択して作成した分類器を 10 個用意し，犯人音声を分類した．つまり，用意した 10 個の分類器の学習データはそれぞれすべて異なるものとなっている．よって結果は，10 個の分類器による分

表 6 分類結果
学習：演技経験者 5 人分（模倣，通常各 250 ファイル）
分類対象：実際の犯人 30 人分（810 ファイル）

	数	割合 (%)
模倣音声	666	82.2222
通常音声	144	17.7778

表 7 10 回の合計の分類結果
学習：演技未経験者 5 人分（模倣，通常各 250 ファイル）
分類対象：実際の犯人 30 人分（810 ファイル）

	数	割合 (%)
模倣音声	1997	24.6543
通常音声	6103	75.3457

表 8 分類結果
学習：演技経験者 5 人分（模倣，通常各 250 ファイル）
分類対象：演技未経験者 25 人分（模倣，通常各 1250 ファイル）

		分類されたクラス	
		模倣音声	通常音声
実際のクラス	模倣音声	1096	154
	通常音声	665	585

類結果の合算を示す．一つ一つの分類器の学習データは一般人演技未経験者音声 5 人分（模倣音声ファイル：250 個，通常音声ファイル：250 個）である．分類対象データは実際の犯人の音声 30 人分（ファイル数：810 個）である．

分類結果は**表 7** の通り．810 個のファイルの分類 10 回分の結果の合計のため，分類したファイル数の合計は 8100 個となっている．多くの犯人音声がこれらの分類器の学習した通常音声に近いという結果となった．

5.6 演技経験者 5 人分で学習，未経験者 25 人分を分類した結果

演技経験者と演技未経験者の分類器には大きな性能差があるのを確認したため，次は演技未経験者と実際の犯人の音声を比較する．収録した音声と収集した犯人音声とでは収録環境等が大きく違うため，直接分類を行い比較を行っても良い結果が得られない．そこで，演技経験者 5 人の音声で学習した分類器によって演技未経験者と実際の犯人の音声を分類し，その結果を比較することで音声の比較を行う．学習データは一般人演技経験者音声 5 人分（模倣音声ファイル：250 個，通常音声ファイル：250 個）である．分類対象データは演技未経験者音声 25 人分（模倣音声ファイル：1250 個，通常音声ファイル：1250 個）である．

分類結果は**表 8** の通り．

5.7 演技未経験者 25 人分で学習，演技経験者 5 人分を分類した結果

演技経験者と実際の犯人の音声を比較するために，演技未経験者で学習した分類器で演技経験者と実際の犯人の音声を分類した．学習データが多い方が精度が向上することが見込まれるため，演技未経験者 25 人の音声全てを学習データとした分類器を新たに作成した．学習データは演技未経験者音声 25 人分（模倣音声ファイル：1250 個，通常音声ファイル：1250 個）である．分類対象データは一般人演技経験者音声 5 人分（模倣音声ファイル：250 個，通常音声ファイル：250 個）である．

表 9 分類結果
学習：演技未経験者 25 人分（模倣，通常各 1250 ファイル）
分類対象：演技経験者 5 人（模倣，通常各 250 ファイル）

		分類されたクラス	
		模倣音声	通常音声
実際のクラス	模倣音声	174	76
	通常音声	42	208

表 10 分類結果
学習：演技未経験者 25 人分（模倣，通常各 1250 ファイル）
分類対象：実際の犯人 30 人分（810 ファイル）

	数	割合 (%)
模倣音声	13	1.6049
通常音声	797	98.3951

分類結果は**表 9** の通り．

5.8 演技未経験者 25 人分で学習，犯人 30 人分を分類した結果

演技経験者の音声と犯人音声を比較するために，同一の分類器で犯人 30 人分の音声を分類する．学習データは演技未経験者音声 25 人分（模倣音声ファイル：1250 個，通常音声ファイル：1250 個）である．分類対象データは犯人音声 30 人分（ファイル数：810 個）である．

分類結果は**表 10** の通り．大半の犯人の音声が，分類器の学習した通常音声に近いという結果になった．

5.9 各分類器の特徴量の重み

本実験において作成した分類器は，一般人 20 人の音声で学習したもの，一般人 30 人の音声で学習したもの，演技経験者 5 人で学習したもの，演技未経験者 5 人で学習したもの，演技未経験者 25 人で学習したものの 5 種類である．それぞれの分類器において，重みの大きな特徴量の上位 10 個を示す．

一般人 20 人の音声で学習した分類器における重みの大きな特徴量は**表 11** の通り．

一般人 30 人の音声で学習した分類器における重みの大きな特徴量は**表 12** の通り．

演技経験者 5 人の音声で学習した分類器における重みの大きな特徴量は**表 13** の通り．

演技未経験者 5 人による分類器は 10 個作成した．具体例として一つの分類器における重みの大きな特徴量は**表 14**

表 11　重みの大きい特徴量（一般人 20 人分で学習した分類器）

特徴量の重み（係数）	特徴量名
-2.6136	pcm_fftMag_mfcc_sma[4]_skewness
2.068	pcm_fftMag_mfcc_sma[5]_skewness
-1.9778	voiceProb_sma_amean
1.9562	pcm_fftMag_mfcc_sma[4]_amean
1.9222	pcm_RMSenergy_sma_de_kurtosis
1.8227	pcm_fftMag_mfcc_sma_de[5]_linregc1
1.8196	pcm_fftMag_mfcc_sma_de[1]_skewness
1.6364	pcm_fftMag_mfcc_sma[3]_linregc1
1.6036	pcm_zcr_sma_skewness
1.4227	pcm_fftMag_mfcc_sma[2]_skewness

表 12　重みの大きい特徴量 (一般人 30 人分で学習した分類器）

特徴量の重み（係数）	特徴量名
-2.9759	pcm_fftMag_mfcc_sma[4]_skewness
-2.6753	voiceProb_sma_amean
2.4469	pcm_fftMag_mfcc_sma[5]_skewness
-2.1069	pcm_fftMag_mfcc_sma[3]_linregc1
2.0752	pcm_fftMag_mfcc_sma[4]_amean
1.995	pcm_fftMag_mfcc_sma_de[5]_linregc1
-1.9759	pcm_zcr_sma_de_linregc2
1.8636	pcm_fftMag_mfcc_sma[1]_kurtosis
1.7521	pcm_fftMag_mfcc_sma[10]_linregerrQ
1.63	pcm_fftMag_mfcc_sma[3]_linregc2

表 13　重みの大きい特徴量
（演技経験者 5 人分で学習した分類器）

特徴量の重み（係数）	特徴量名
1.1312	pcm_fftMag_mfcc_sma[8]_amean
-1.0715	pcm_fftMag_mfcc_sma_de[2]_kurtosis
0.935	pcm_fftMag_mfcc_sma_de[1]_skewness
-0.9303	pcm_fftMag_mfcc_sma[1]_srddev
0.9284	pcm_fftMag_mfcc_sma[6]_max
0.8055	pcm_fftMag_mfcc_sma[7]_skewness
-0.7865	pcm_fftMag_mfcc_sma_de[1]_linregerrQ
0.768	pcm_fftMag_mfcc_sma_de[8]_linregc2
0.7396	pcm_fftMag_mfcc_sma_de[1]_min
0.7303	pcm_fftMag_mfcc_sma[6]_amean

表 14　重みの大きい特徴量
（演技未経験者 5 人分で学習した分類器）

特徴量の重み（係数）	特徴量名
1.2318	pcm_fftMag_mfcc_sma[8]_amean
1.207	pcm_fftMag_mfcc_sma[8]_linregc1
1.1616	pcm_fftMag_mfcc_sma[3]_amean
-1.0686	pcm_fftMag_mfcc_sma[1]_minPos
-1.0442	pcm_fftMag_mfcc_sma[12]_skewness
1.0392	pcm_fftMag_mfcc_sma[8]_skewness
-1.0194	pcm_fftMag_mfcc_sma[5]_minPos
1.015	pcm_fftMag_mfcc_sma[9]_skewness
-0.9664	pcm_fftMag_mfcc_sma[4]_skewness
0.8771	pcm_fftMag_mfcc_sma[4]_linregc1

表 15　重みの大きい特徴量
（演技未経験者 25 人分で学習した分類器）

特徴量の重み（係数）	特徴量名
-2.8649	pcm_fftMag_mfcc_sma[4]_skewness
2.5477	pcm_fftMag_mfcc_sma[5]_skewness
-2.3939	voiceProb_sma_amean
-2.2224	pcm_zcr_sma_de_linregc2
-2.1818	pcm_fftMag_mfcc_sma_de[3]_linregc1
2.0253	pcm_fftMag_mfcc_sma[10]_linregerrQ
-1.8908	voiceProb_sma_linregc1
1.8803	voiceProb_sma_skewness
-1.854	pcm_fftMag_mfcc_sma[5]_min
1.8501	pcm_fftMag_mfcc_sma[9]_stddev

の通り．この分類器単独の結果は，810 個の犯人音声のうち 197 個（24.321%）が模倣音声，613 個（75.679%）が通常音声と分類された．

演技未経験者 25 人全員分で学習した分類器における重みの大きな特徴量は**表 15** の通り．

6. 考察

6.1 収録音声の考察及び犯人音声との比較

まず，表 3 の結果より，収録した一般人の発話では，詐欺模倣の場合と通常の場合で声が異なっていることが分かった．しかし，表 4 の通り，実際の犯人の発話は一般人の詐欺模倣ではなく，一般人の通常発話に近いという結果となった．同一の分類器によってここまで分類結果に差が生じるということは，犯人の発話は明確に通常発話に近いと言える．この差は単純に，犯人と一般人の技量の差から生じていると考えられる．今回の収録では，一般人による詐欺模倣は事前に目を通すだけで練習してもらうことを依頼しなかった．そのため，詐欺模倣では不自然なトーンでの発話になってしまったと考えられる．今回の結果を見ると，犯人の発話は非常に自然であり，一般人の普通の発話と遜色がなかったのだと考えられる．次に，表 5 の結果を見ると，こちらも同様に犯人音声のほとんどが一般人の通常発話に近いという結果になった．ここで特徴量に注目すると，表 11 と表 12 より，重みの大きな特徴量の上位 10 個のうち 6 個の特徴量が共通している．つまり，これらの特徴量が一般人の音声，実際の犯人の音声の分類どちらにおいても重要な特徴量であったと考えられる．これらの結果から言えるのは，一般人が詐欺を試みる際は声が変化するが，実際の犯人の音声とは大きく異なるということである．実際の犯人は，一般人が普通に会話している声と同様に，自然な声で詐欺を行っていると考えられる．

6.2 一般人の演技経験の有無による違いの考察

次に，一般人の演技経験の有無による分類結果の相違について考察する．まず，表 6 と表 7 を比較すると，分類

結果は大きく異なっていることが分かる．これらの結果から，犯人の音声は演技経験者の模倣発話に近いと言える．演技経験者（本研究では過去 1 年以内に舞台等で演技をした人）は，普段の自分ではない存在を演じることに長けている．そのため，普段とは違う「詐欺をする」という発話も違和感なくできたのだと考えられる．しかし，この分類器は学習データが少ないため，精度に不安が残る．実際に表 13，表 14 にて分類器における重みの大きな特徴量を確認すると，学習データの多い分類器に比べて係数の絶対値が小さい．更に，上位 10 個にある特徴量も共通のものがかなり少なく，また学習データが多い分類器では共通して確認された特徴量が上位に来ていないため，学習の不完全さが伺える．だが，演技経験の有無によって分類結果が変わっていたのは確かであるため，演技経験者をより集めて学習させると，また新たな発見があると考えられる．

実際の犯人の発話は，演技経験者の模倣発話に近いという結果が確認できたため，次は実際の犯人と演技経験者の発話を比較する．犯人音声と演技経験者の模倣音声が近いのならば，同一の分類器で分類を行った結果も同様になるのではないかと考えていたが，実際は大きく異なる結果が得られた．表 9 と表 10 を見比べると，実際の犯人の発話は大半が通常音声として分類されているが，演技経験者の模倣音声で通常音声として分類されたのは 3 割程度であった．つまり，この結果だけを見ると演技経験者の模倣発話の韻律的特徴量は犯人のものと大きく異なるということが分かる．表 15 を確認すると，表 11，表 12 に見られる特徴量が確認でき，精度は十分であると考えられる．

次に，演技経験者の発話で学習した分類器で検討する．表 8 を見ると，演技未経験者の模倣音声の大半が模倣音声として分類されている．しかし，通常音声に関しては通常音声だと正しく分類されたのは約 46%と，分類率は芳しくなかった．学習データが少ないため，学習が不十分ではなく結果が 50%に近くなったと考えられる．やはり学習データが 5 人だけだと，分類器としての信頼性は不安でが残る．だが，この分類器で模倣音声は高い精度で模倣音声と分類された．また，先述の通り犯人音声は演技経験者の模倣音声に近いという結果が出ている．これらのことから，演技経験者の模倣音声は分類器作成において非常に有用なデータであると考えられる．

演技経験者の模倣音声が有用であると考えられるが，演技未経験者 25 人で学習した分類器で分類した結果は実際の犯人の分類結果と大きく異なっていた．これは，詐欺模倣音声はスクリプトをもとに，収録した一般人が同じセリフを発言をしていたことが原因であると考えられる．表 3 と表 8 は，どちらも分類対象は収録した音声だが，どちらの結果も模倣音声が正しく分類された割合が通常音声が正しく分類された割合よりも大きかった．つまり，学習データと同じ発言は，分類が容易なのだと考えられる．本研究では，学習データに同一セリフが含まれていることを憂慮し，同一セリフが学習データに存在しすぎない分類器も作成したが，その分類器で犯人音声の分類を行ったところ，大半の犯人音声が通常音声として分類されるという結果は変わらなかった．学習データに同一セリフがあることは分類対象が犯人音声の場合はほとんど影響が見られなかったが，分類対象に同一セリフがある場合は結果に影響を及ぼすのではないかと考えられる．表 9 を見ると，分類対象を収録音声としている結果の中で，この結果だけは模倣音声が正しく分類された割合が通常音声が正しく分類された割合よりも小さいことが分かる．このことから，演技経験者の模倣音声の声自体は演技未経験者の通常音声のものと近かったが，同一セリフを発言している影響で模倣音声と分類されたと考えられる．つまり，演技経験者の模倣音声は犯人の音声と近いのだが，学習データと同一セリフを発言しているため表 9 と表 10 のような結果の相違が生じたのだと考えられる．同一セリフを分類していると考えると，表 8 の結果で，模倣音声が模倣音声として分類される割合が異様に高いのも説明がつく．

以上の通り，演技経験者の模倣音声は実際の犯人の音声と近いと考えられる．

7. おわりに

7.1 まとめ

本研究では，詐欺音声の検出に役立てるため，一般人が詐欺を試みる時と通常の時の音声を収録，実際の犯人の音声を分類することで特殊詐欺音声について韻律的特徴量を用いて考察を行った．結果は直接的に有用な特徴量を発見すること，集めた音声での実際の犯人音声の分類はできなかった．しかし，一般人が詐欺を試みる際は韻律的特徴量に変化が生じること，発話者の演技経験の有無によって詐欺を試みる際の韻律的特徴量には違いがあることを確かめられた．また，演技経験者が詐欺を試みる際の韻律的特徴量は，実際の犯人の音声と比較的近いのではないかということを考察することができた．

7.2 今後の課題

今後の課題としては，演技経験者の模倣音声は本当に実際の犯人の音声と近いのかということを検証しなければならない．今回は演技経験者の人数自体が少なく，十分な検証はできなかった．人数が増えれば，また新たに発見できることがあると考えられる．今回の研究では一つのスクリプトで模倣音声を収録したため同一セリフの存在の影響を受けてしまったが，今後音声を集める際は，スクリプトを一人一人違うものにするなど集める音声も改良したい．また，本研究では一般人を対象に収録を行ったため，詐欺を試みる時の声の変化が，詐欺を試みているからなのか，営業電話を模しているからなのかが分からなかった．今後

は，実際に電話で営業をしている方のみを対象として，普段の営業時の音声と詐欺を試みる時の音声を収録しデータベースを作成したい．

今回は特徴量として INTERSPEECH 2009, Emotion Challenge で用いられた特徴量を使用したが，今回のようなデータ数では特徴量としては数が多すぎたかもしれない．特徴量選択で真に必要な特徴量のみを分類に使用するなど，分類器自体にも改良を加えたい．

以上の点を考慮し，新たに音声を集めていけば，特殊詐欺を声から判別できるような分類器を作成できるのではないかと期待している．

謝辞 本研究は JSPS 科研費 JP17H04705, JP18H03229, JP18H03340, JP18K19835, JP19H04113, JP19K12107 の助成を受けたものです．

参考文献

[1] 警察庁：特殊詐欺とは，入手先 ⟨https://www.keishicho.metro.tokyo.jp/kurashi/tokushu/furikome/furikome.html⟩（2020/01/10 参照）．
[2] 警察庁：特殊詐欺認知・検挙状況等について，入手先 ⟨https://www.npa.go.jp/publications/statistics/sousa/sagi.html⟩（2020/01/10 参照）．
[3] 日本テレビ：OA まとめ 最新「ダマされたフリ作戦」詐欺に注意！，入手先 ⟨https://www.ntv.co.jp/choumon/articles/83o187f9dtmw-kuprzn.html⟩（2020/01/10 参照）．
[4] ITmedia NEWS：「特殊詐欺の電話」AI で見破る NTT グループが実証実験 犯人が使う表現を学習，入手先 ⟨https://www.itmedia.co.jp/news/articles/1905/09/news-116.html⟩ (2020/01/10 参照)
[5] 松尾直司，早川昭二，原田将治，武田一哉，降旗喜和男，”音声からのストレス状態検出システムの開発”，マルチメディア, 分散, 協調とモバイル (DICOMO2014) シンポジウム (2014)
[6] B. Schuller, S. Steidl, and A. Batliner, ”The interspeech 2009 emotion challenge” Proc. INTERSPEECH 2009, pp.312-315
[7] 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. 5).
[8] 松井辰哉，萩原将文，”感情推定と知識獲得機能を有する対話システムの構築”，日本感性工学会論文誌 Vol.16 No.1 pp.35-42（2017）
[9] Tang Na Nhat，目良和也，黒澤義明，竹澤寿幸，”音声に含まれる感情を考慮した自然言語対話システム”，HAI シンポジウム（2014）
[10] Amanda Chow, John Louie, ”Detecting Lies via Speech Patterns”
[11] Kun Han, Dong Yu, Ivan Tashev, ”Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine” INTERSPEECH 2014
[12] Savita Sondhi, Ritu Vijay, Munna Khan, Ashok K. Salhan, ”Voice Analysis for Detection of Deception” 11th International Conference on knowledge, Information and Creativity Support Systems (KICSS) (2016)