伝達関数に着目した
本人発話と録音再生音の判別方法の検討

令和２年度卒業論文
指導教員
喜安千弥教授

長崎大学工学部工学科情報工学コース
喜安研究室
学生番号　35317017
白石　朱理

令和３年２月12日

目 次

第1章　序論

1.1　研究背景

現代では，簡単に音場を模擬することが可能になった．よって，リプレイ攻撃のような問題が音声認識でも起こりうるかもしれない．
リプレイ攻撃とは，ユーザがログインする時にネットワークを流れるデータを盗聴してコピーし，コピーしたデータを認証サーバへ送ることでシステムに不正にログインしようとする行為のことである [1].

また，音場の模擬は人間による聴取においては問題にならないが，真正性の担保が必要とされるシステムにおいては大きな問題となる可能性がある [2]．一旦パッケージ化された信号についてその真正性を担保するには，真正の配信者が介在し，恣意的に電子透かしを埋め込んだり，フィンガープリントを記憶するなどによりそれらを照合することで解決できるが，パッケージ化される前の直接録音されたものである真正性を担保することにはならない．

事実，音声による個人認証・照合システム (ASV : Automatic speaker verification) では，このような録音再生音（リプレイ音声）やディープフェイク音声などによる詐欺がシステムの判断を誤らせる障害となることが取りざたされている．[3]
音声以外でも例えば，工場の異常音を模擬した音が作成され，再生された場合，労働者に誤った判断をさせてしまう．

このような直接録音・現実音場の模擬加工の検知 (Anti-spoofing) は，音声その他のバイオメトリクス認証においては生体検知 (Liveness detection / PAD : presentation attack detection) として考えられ, ASV-PAD についてはその検知性能を競う ASVspoof challenge [4]〜 [6] が開催されている．

1.2 研究目的

本研究では，本人認証・照合システム (以下 ASV) に入力された音声が本人の音声か録音した音声なのか判別することを目標とした．図 (1.1) に本人音声 (bonafide)，図 (1.3) に録音音声 (spoof) が ASV に入力される状況を示す．

まず本人音声は本人発話の場所 (音源) を S とおくと ASV に入力されるまでに室内伝達関数 H がかかり，ASV 位置で観測される音声データは $H \cdot S$ という畳み込みで与えられる．録音した音声の再生音はまず ASV の設置位置とは別のところで録音されたものを bonafide の音源と同じ位置 S で音声を再生することなので，まず攻撃者の録音装置に音声が入力されるまで別の室内伝達関数 P がかかり，$P \cdot S$ という畳み込みが与えられる．最終的な ASV 位置で観測される音声データは $H \cdot P \cdot S$ という畳み込みの式で表される．

図 (1.2)，図 (1.4) に ASV に音声データが入力されるまでの状況を簡単に説明する．赤枠部が室内伝達関数を表しており，図 (1.4) から，録音した音声の方が室内伝達関数が多重化していることが分かる．このように室内伝達関数が多重に重なることで伝達関数が複雑化し，周波数スペクトルを考えた時に極や零点が増えるものと考えられる．

したがって，音声データから室内伝達関数 (bonafide の場合の H，spoof の場合の $H \cdot S$) だけを分離することで録音音響特性を抽出し，そのフィルタ構造の複雑化を評価する．具体的には分離した音声データをケプストラム分析することで，本人の音声か録音した音声か判別できるのではないかと考えた．

また本研究の最終的な目標として，ケプストラム分析をした本人の音声と録音した音声二つのデータを比較して判別するのではなく，一つのデータを見ただけでどちらの音声なのかを判別できるような規則性を見つけることが目的である．

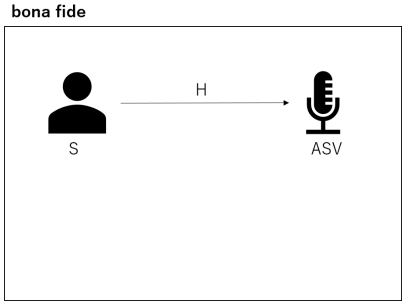

図 1.1 bonafide の図

図 1.2 bonafide の室内伝達関数の図

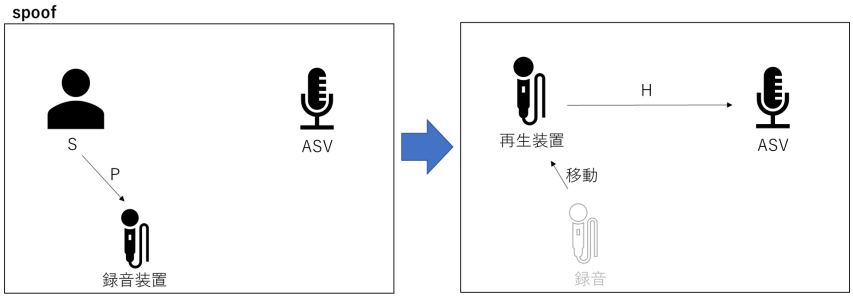

図 1.3 spoof の図

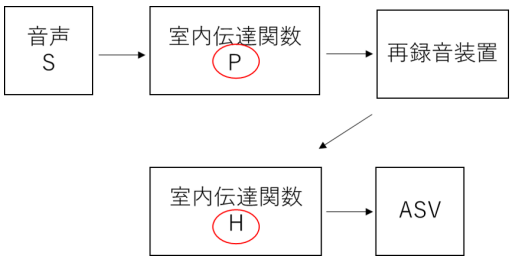

図 1.4 spoof の室内伝達関数の図

第2章　使用した音声データと
ケプストラム分析について

本章では，本研究で使用する音声データの条件と，この研究に関連する既存の研究などについて説明する．

2.1　使用した音声データ

この研究で使用した音声データは 2019 年度の ASVspoof 大会で使用されたものである．ASVspoof 大会は ASVspoof Challenge など音声の生体検知システムの検出精度を競う大会であり，このような音声の生体検知のついての研究は盛んに行われている．

生体検知とは，(図 2.1) のように，Apple 社の Siri [7] や google 社の google アシスタントなどのシステムに使われている．[8]

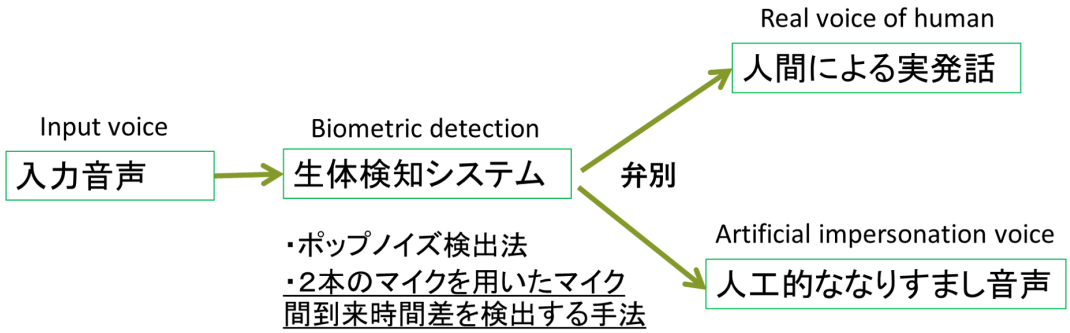

図 2.1 音声の生体検知システムの仕組み

2.2 ASV の録音条件

使用した ASV 音声データの録音条件は，部屋サイズが 2 ～ 5m 残響時間が 50 ～ 200ms，本人発話の場所と ASV までの距離を 10 ～ 50m とする．N0.69，No.70，No.74 の 3 人分の音声データを使用した．それぞれの声の特徴を図 (2.2) に示す．また，ASV はサンプリング周波数 16kHz，量子化ビット数 16bit で音声を取り込み．

No.	69	70	74
性別	低い声の女性	低い声の男性	高い声の女性

図 2.2 声の特徴

2.3 攻撃者の録音再生装置の品質

本人発話の場所と攻撃者が録音している場所までの距離を 10～50(m) とし，攻撃者の録音装置の質を perfect，high，low の 3 種類とする．図 (2.3)，図 (2.4) は，録音装置の質とその状況を簡単に図表化・グラフ化したものである．

	録音再生周波数帯域(OB kHz)	OBの下限(minF Hz)
perfect	infinity	0
high	> 10	< 600
low	< 10	> 600

図 2.3 録音再生装置の質

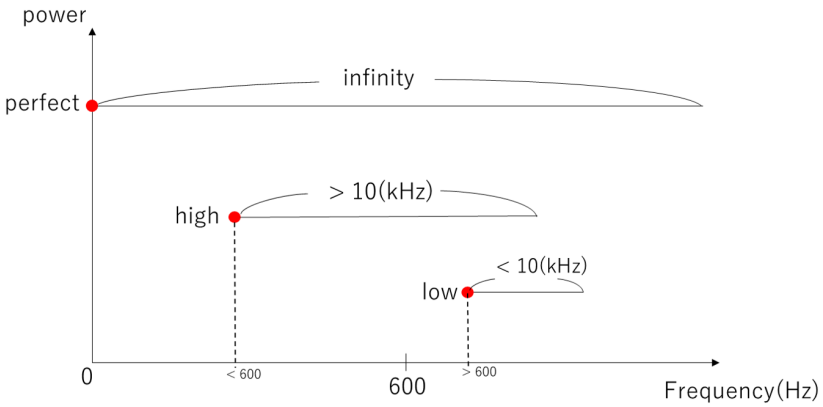

図 2.4 各録音周波数帯域

2.4　関連研究

関連研究では，矢口ら [10] は，複数チャネル間の相互相関関数を用いた話者照合のためのなりすましを検出している．ここでは入力された音声が人間による実発話かスピーカによる再生音かを判別する方法を 2 本のマイクを用いたマイク間到来時間差を検出する手法で研究が行われている．

実発話の場合，無発話区間は音を発していないためマイク間到来時間差を用いて音源定位を行うと音源位置が不安定になるが，スピーカ再生の場合，無発話区間であっても収録時の背景雑音やスピーカの電磁ノイズが発生するため音源定位されやすい傾向がある．また実発話の場合，音素毎に口内の音源位置が若干異なることから，検出される発音位置の変動幅は極めて微小であり，口とマイク間の距離や背景雑音によっては検出が難しく，また精度が発話内容に依存しやすいという問題点がある．よってマイク間到来時間差に着目しつつ，さらに発話内容にも依存しない手法として，実発話を検出するのではなくスピーカ再生を検出する手法で無発話区間におけるマイク間の相互相関値を用いてなりすまし検出を行っている．

この研究は検出性能は高いが，マイクを複数必要としている点，無発話区間を利用している点が問題点と考えられる．本研究では，この点を解決することを目指し，1 つのマイクのみで発話区間を用いたなりすまし検出を試みた．

2.5　ケプストラム分析について

ケプストラムは，時間波形のパワースペクトルの対数のフーリエ変換であり，複数の信号が畳み込まれた信号を分離可能という特徴がある．音声波形から切り出した時間波形を離散フーリエ変換 (DFT) し，対数パワースペクトルを時間波形だと思って逆フーリエ変換して得られる．今回は，攻撃者の録音再生品質の違う 3 種類と本人音声の合わせて 4 種類の音声データのケプストラムの 1 秒ごとのピーク数を見ることで判別を行った．[9]

第3章　録音音響特性の抽出

3.1　録音環境の分離

ASVspoof で対象としているスピーカ再生音声は，本人発話の録音音声のスピーカ再生を再録音したものであり．再録音された信号は再生信号（本人発話録音）と再生再録音時の背景音との合成および，再生再録音時の室内伝達関数との畳み込みと考えられる．

残響特性に着目すると，残響特性は音源に対するフィルタ（畳み込み処理）であり，観測信号のケプストラム上では加算の関係になる．観測した信号を $y(t)$ とすると，$y(t)$ は式 (3.1) のように音源信号 $s(t)$ と残響特性を $h(t)$ の畳み込み演算（積和演算）でモデル化でき，Fourier 変換ドメインのスペクトルでは，式 (3.2) のようにアダマール積となる．この両辺に対して対数をとったものは加算の関係となり，その Fourier 変換（すなわちケプストラム）でも式 (3.3) のように加算となる．

$$y(t) = s(t) * h(t) \tag{3.1}$$

$$Y(\omega) = S(\omega) \cdot H(\omega) \tag{3.2}$$

$$\mathcal{Y}(q) = \mathcal{S}(q) + \mathcal{H}(q) \tag{3.3}$$

ここで $Y(\omega)$, $S(\omega)$, $H(\omega)$ は角周波数 ω における観測信号，音源信号，残響特性それぞれのスペクトル，$\mathcal{Y}(q)$, $\mathcal{S}(q)$, $\mathcal{H}(q)$ は quefrency:q におけるそれぞれのケプストルとする．

ここで rcunwrap を行う．rcunwrap を行うことで位相アンラッピング処理後に直接位相をキャンセル処理することができる．

次にケプストラム分析を行う．ケプストラム分析ではオールパスケプストラム分析を使用する．音声をケプストラム分析するとケフレンシーの低い部分にスペクトラムの包絡成分．高い部分にスペクトルの微細構造が表れる．低ケフレンシー (スペクトラムの包絡成分) には音源 (最小位相ケプストラム) とフィルター成分である室内伝達関数が含まれる．オールパスケプストラムと低ケフレンシーのリフタリングを行うことで，低ケフレンシー内の最小位相ケプストラムと高ケフレンシーの微細構造を省くことができ，室内伝達関数のみを取り出すことができる．

ここからはオールパスケプストラム分析について説明する．まずケプストラムを 2 種類作る．
実数ケプストラム $\mathcal{Y}_{\mathcal{A}}(q)$ と複素数ケプストラム $\mathcal{Y}(q)$ は以下のように表すことができる．

$$\begin{cases} \text{①実数ケプストラム} & \mathcal{Y}_{\mathcal{A}}(q) = \mathrm{F}^{-1}\left[\log|Y(\omega)|\right] \\ \text{②複素ケプストラム} & \mathcal{Y}(q) = \mathrm{F}^{-1}\left[\log|Y(\omega)| + j\angle Y(\omega)\right] \end{cases} \tag{3.4}$$

①について最小位相化する．

$$W = \begin{cases} 0 & (q < 0) \\ 1 & (q = 0) \\ 2 & (q > 0) \end{cases} \tag{3.5}$$

とすると，
最小位相ケプストラム C_{min} は，

$$C_{min} = W \cdot \mathcal{Y}_{\mathcal{A}}(q) \tag{3.6}$$

と表せる．
よってオールパスケプストラム $C_{Allpass}$ は，

$$C_{Allpass} = \mathcal{Y}(q) - C_{min} \tag{3.7}$$

と表せる．
以上より室内伝達関数のみのケプストラム分析を行うことができる．

3.2 提案手法について

手順は以下の通りである．

1. 観測信号 $y(t)$ の DFT を行い，パワースペクトル $Y(\omega)$ を求める．
2. $Y(\omega)$ の fftshift を行う．
3. fftshift した $Y(\omega)$ の対数をとる．
4. rcunwrap を行う．
5. オールパスケプストラム上で低ケフレンシーのリフタリングを行う．
6. 1 秒ごとのピーク数を数える．

今回設定したウインドウサイズとフレーム長は 4096 である．リフタリングはウインドウサイズを 2 で割った長さとした．ピーク数を数える order は 5 である．

第4章　提案手法の評価と考察

第3章で提案した手法から評価音声を用いて実験を行った．この章では，図(1.1)，図(1.3)における録音環境における音声データの検出性能の評価を行う．

4.1　音声データの評価

図(4.1)，図(4.3)，図(4.5)にASV音声の時間波形及び，本人発話，再録音(品質3段階)のそれぞれで推定された室内伝達関数のケプストラムピーク数の時間変化を示す．No.69(低い声の女性)は全体的にピーク数変化が大きい．bonafideのピーク数がspoofの3種類のピーク数よりも低い方にあることが分かった．No.70(低い声の男性)も全体的にピーク数変化が大きいが，bonafideのピーク数が他のspoofのピーク数と比べて一定であることが分かった．No.74(高い声の女性)はspoofはピーク数変化が大きいが，bonafideのピーク数変化は小さい．ピーク数もほとんどが180〜195の中に収まっていることが分かった．

また，図(4.1)，図(4.3)，図(4.5)のピーク数の平均値を図(4.2)，図(4.4)，図(4.6)に示す．No.69(低い声の女性)は，bonafideのピーク数はspoof-lowよりも低くなった．しかし、spoof-perfectとspoof-highとは同じぐらいだった．No.70(低い声の男性)は，bonafideのピーク数はどのspoofよりも低くなった．No.74(高い声の女性)は，bonafideのピーク数はspoof-highよりも低くなった．しかし、spoof-perfectとspoof-lowとは同じぐらいだった．

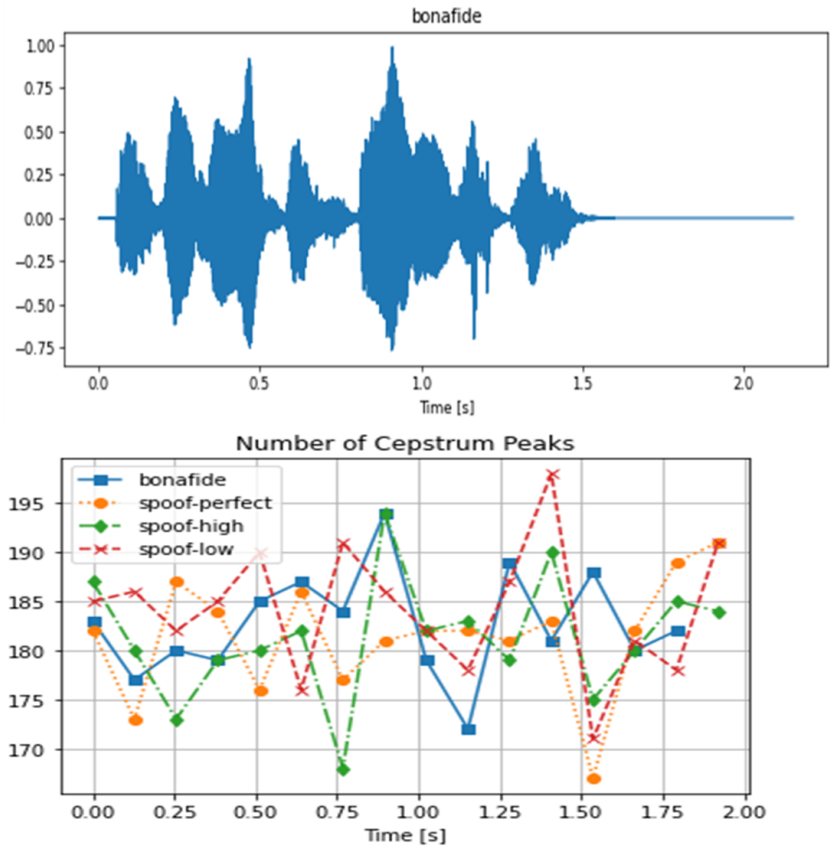

図 4.1 No.69(低い声の女性) の音声の波形とケプストラムピーク数

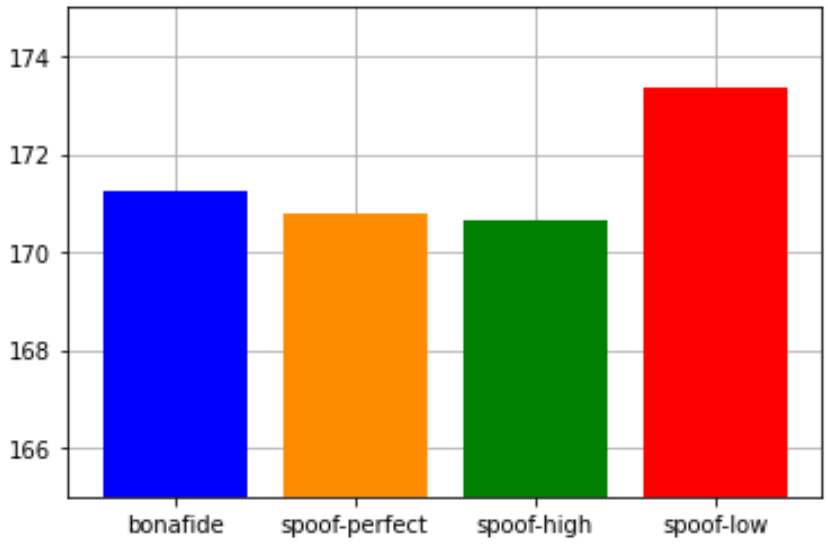

図 4.2 No.69(低い声の女性) のピーク数の平均値

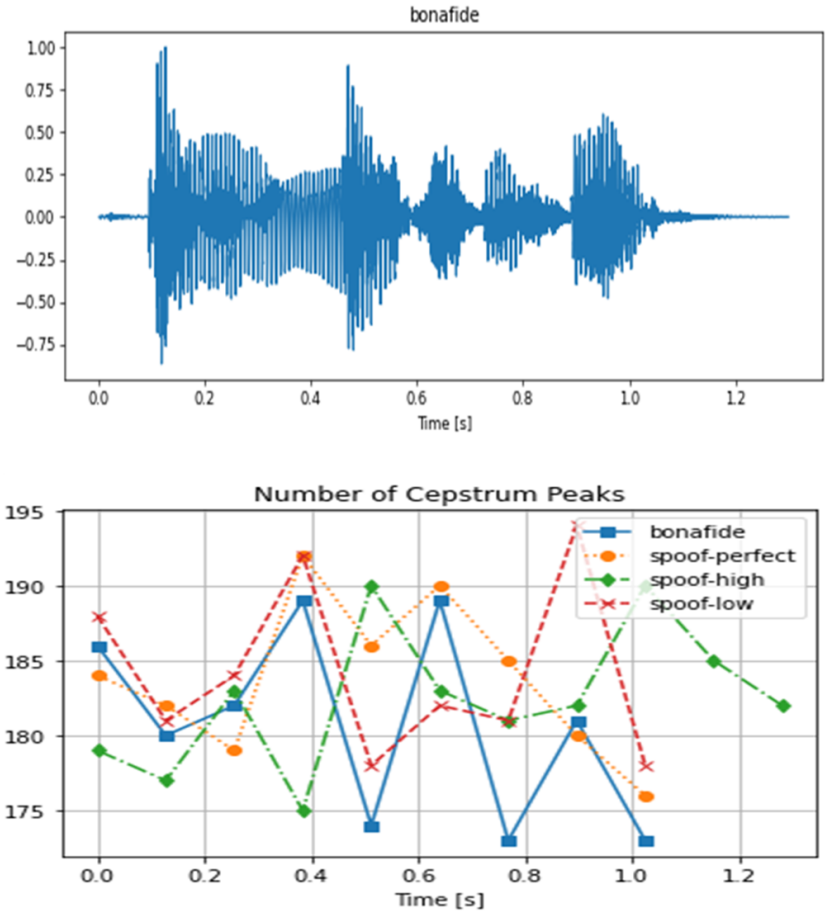

図 4.3 No.70(低い声の男性) の音声の波形とケプストラムピーク数

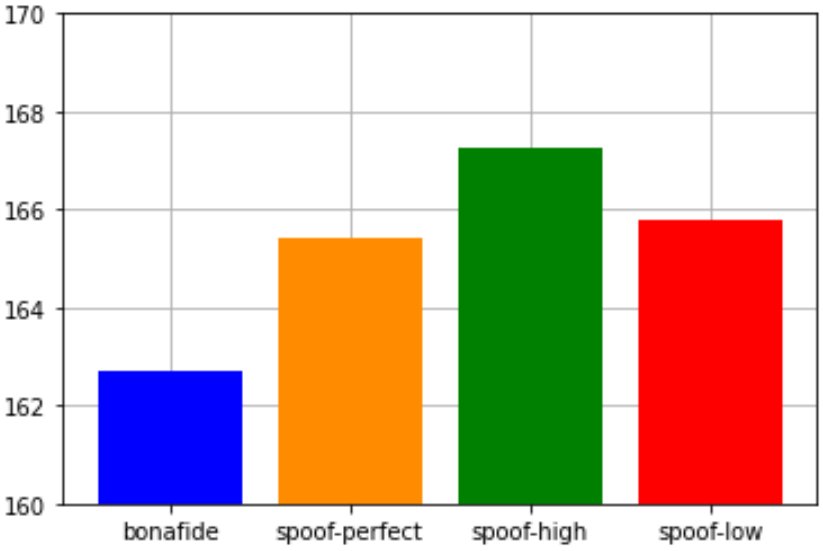

図 4.4 No.70(低い声の男性) のピーク数の平均値

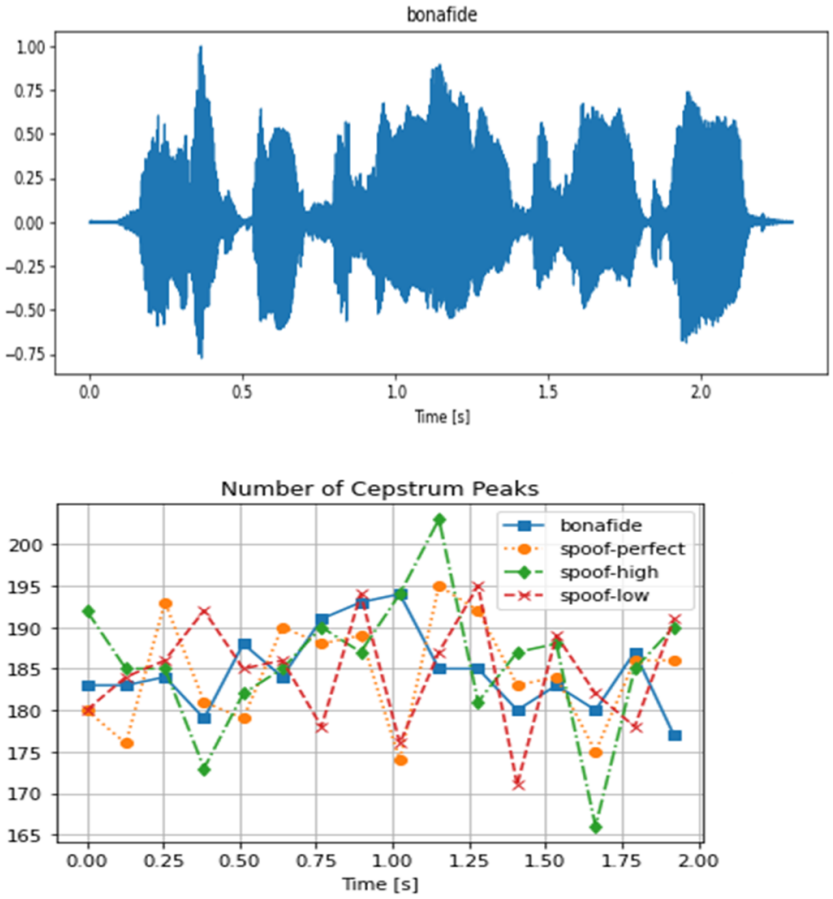

図 4.5 No.74(高い声の女性) の音声の波形とケプストラムピーク数

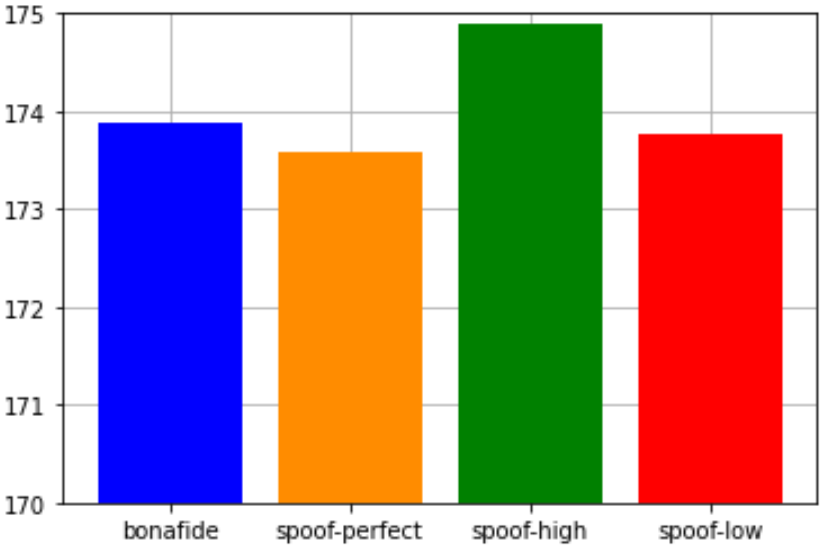

図 4.6 No.74(高い声の女性) のピーク数の平均値

4.2 考察

実験の結果，No.69(低い声の女性) の 1.1 秒，No.70(低い声の男性) の 0.5 秒，0.8 秒，1.0 秒，No.74(高い声の女性) の 1.2 秒では bonafide が最もピーク数が少なく，録音品質が劣化するに従いピーク数が増えている．また，最も伝達関数の多重が少ない bonafide のピーク数が一番小さく，最も残響の大きい spoof が一番大きくなるのではないかと考えたが結果からは陽に観測できなかった．本来，室内伝達関数は録音状況が変わらない限り変化しないはずなので，音声データ全時間区間のピーク数の平均値を見てみる．図 (4.2)，図 (4.4)，図 (4.6) の結果，bonafide が一番値が小さく，spoof の品質が劣化するにつれてだんだん大きくなる形を期待したが，どの No. も期待の形にはならなかった．以上より，使用音声データによって振幅やピーク数の規則性がないことから，うまく低ケフレンシーから室内伝達関数が分離できていないのではないかと考えられる．

第5章　結論

本研究では，ASVspoof の 2019 年度の音声データを使用し，その音声が本人の音声か録音した音声かの判別方法を検討した．録音環境の分離では，オールパスケプストラム上でリフタリングを行うことで低ケフレンシー内の室内伝達関数のみを取りだし，分析を行った．しかし，本人の音声と録音した音声をケプストラム分析した 1 秒ごとのピーク数を示した図 (4.1)，図 (4.3)，図 (4.5) を見てみても本人音声と録音した音声を判別する違いが見つからなかった．しかし図 (4.3) を見ると期待した形ではないものの，bonafide のピーク数がどの spoof の品質よりも低かった．他の図も，bonafide は全ての spoof の品質ではないものの，どれか 1 つの品質よりは低くなった．しかし，時間波形の振幅との関連性を見つけることができなかったので，本研究の最終的な目標であった，本人の音声と録音した音声二つのデータを比較して判別するのではなく，1 つのデータを見ただけでどちらの音声なのかを判別しできるようにすることができなかった．今後の課題として，ケプストラム分析を行った音声データの規則性を見つけるための対策を行うことにより，利用可能なものにすることがあげられる．そのために音声から室内伝達関数を正しく分離するためにアルゴリズムを改善する必要があると考えた．

謝辞

本研究を進めるにあたり，直接のご指導と多くの助言を頂いた長崎大学情報データ科学部薗田光太郎助教に深く感謝致します．また，適切な指導と有益な助言を頂いた喜安千弥教授に深く感謝致します．喜安研究室の皆様，ならびにその他関係者各位に心より感謝致します．

参考文献

[1] http://www.chuu-information.com/security/gyou_ra_wa_5.html

[2] 薗田光太郎，「音響信号の非現実的ミックスの検出」， 信学技報，vol.119， no.396，EMM2019-95， pp.7-10， 2020.

[3] https://japan.zdnet.com/article/35142255/

[4] Z. Wu,J. Yamagishi,T. Kinnunen,C. Hanilci,M. Sahidullah,A. Sizov,N. Evans,M. Todisco, and H. Delgado,"Asvspoof: the automatic speaker verification spoofing and countermeasures challenge,"IEEE Journal on Selected Topics in SignalProcessing, vol.11, no.4, pp.588-604, 2017.

[5] Z. Wu,T. Kinnunen,N. Evans,J. Yamagishi,C. Hanilci,M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge,"Interspeech 2015, pp.2037-2041, Sept. 2015.

[6] T. Kinnunen,M. Sahidullah,H. Delgado,M. Todisco,N. Evans,J. Yamagishi, and k.A.Lee,"The asvspoof 2017 challenge: Assessing the limits of replay spoofing attach detection," Interspeech 2017, pp.2-6, 2017.

[7] Apple,Siri,https://www.apple.com/jp/siri/

[8] 黒田 康弘， 薗田 光太郎， 喜安 千弥， 「音のパンニング加工の検出に関する一検討」， 信学技報，vol.118， no.494， EMM2018-96， pp.25-28， 2019.

[9] https://www.slideshare.net/ShinnosukeTakamichi/lpc-49065650.html

[10] 矢口凌也， 塩田さやか， 小野順貴， 貴家仁志， 「複数チャネル間の相互相関関数を用いた話者照合のためのなりすまし検出」， 日本音響学会講演論文集，2018 年