

第9回

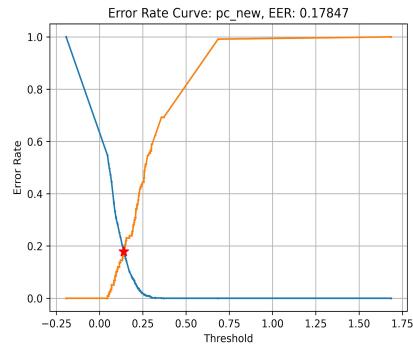
前回は各人が異なるフィルターを音声に提供してエラー率を確認しました。

- 各人で区切りの周波数を統一する
 - 電話の周波数を参考にします。
 - 50, 300, 3400, 7000
 - <https://xtech.nikkei.com/it/atcl/column/14/228621/100100019/>
- 検証結果にはヒストグラムを載せる
 - 載せるヒストグラムは同じ音声のもの
- 音声について
 - 男6女3
 - 計49発話
 - 男(6+7+3+6+6+6) + 女(6+6+3)
 - 1176対
 - 同一話者(117)+異話者(1059)

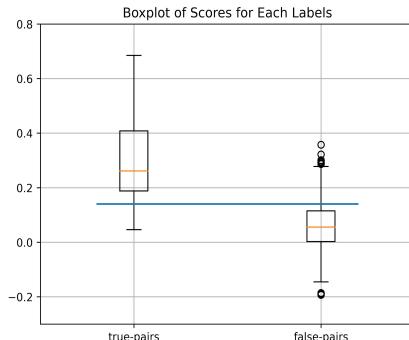
フィルタ無しの場合

PC

EER: 0.1785

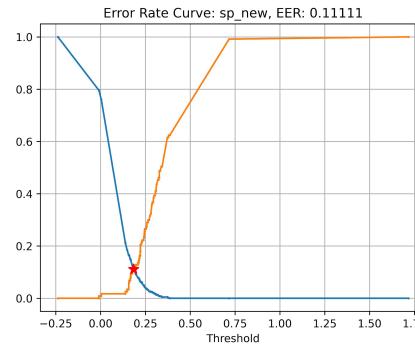


FAR
FRR

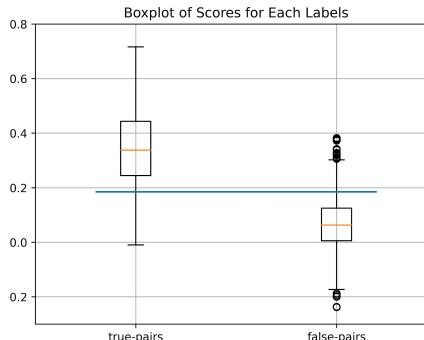


SP

EER: 0.1111



FA
FR



ローパスフィルタ

適用するフィルタは次の通り。

- ~50Hz
- ~300Hz
- ~3400Hz
- ~7000Hz

EER

	フィルタなし	~50Hz	~300Hz	~3400Hz	~7000Hz
PC	0.179	0.350	0.241	0.171	0.180
SP	0.111	0.331	0.188	0.137	0.112

ローバスフィルタ(PC)

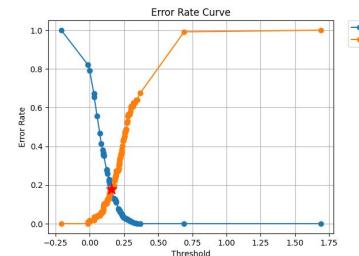
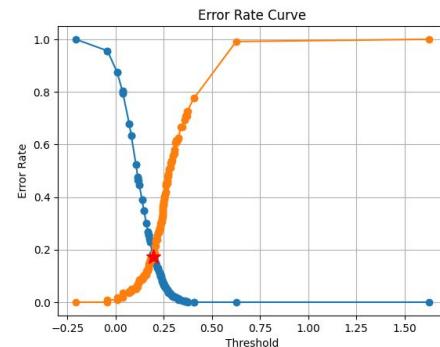
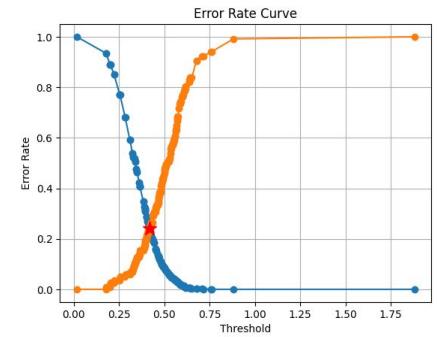
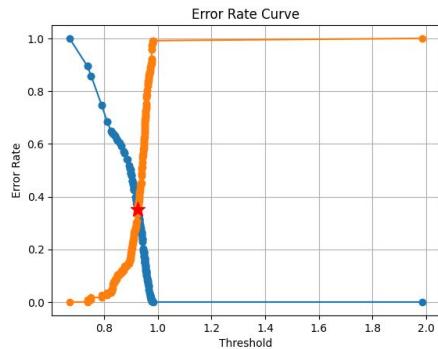
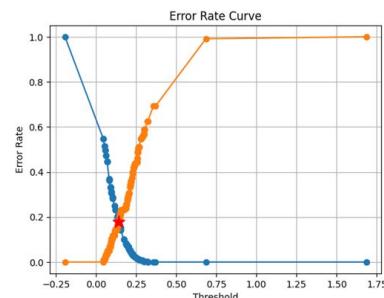
フィルタなし

~50Hz

~300Hz

~3400Hz

~7000Hz

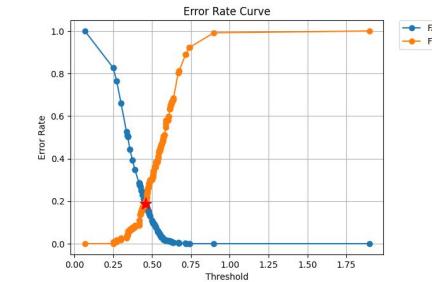
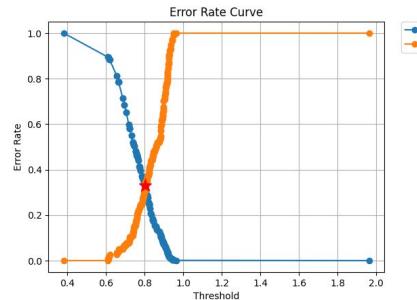
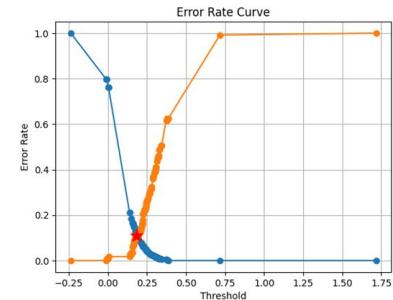


ローパスフィルタ(SP)

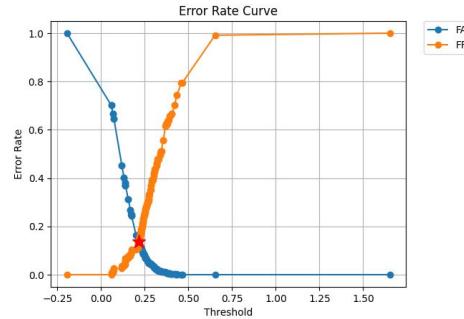
フィルタなし

~50Hz

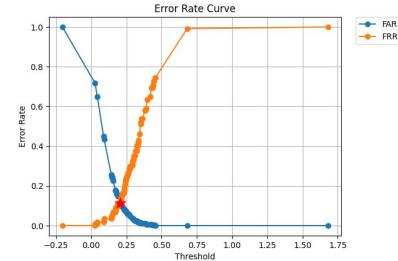
~300Hz



~3400Hz



~7000Hz



ローパスフィルタ(PC)

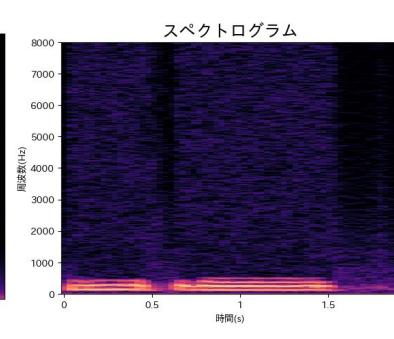
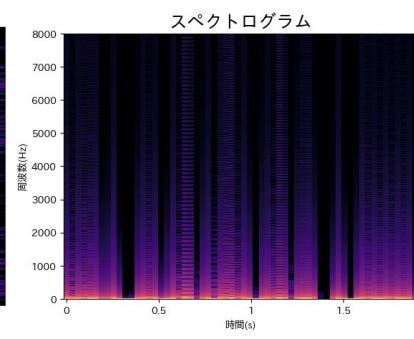
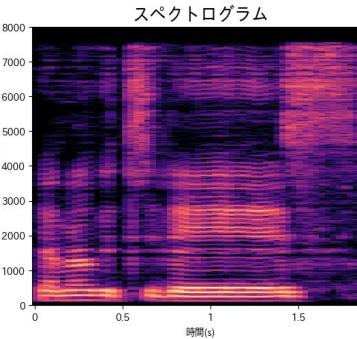
フィルタなし



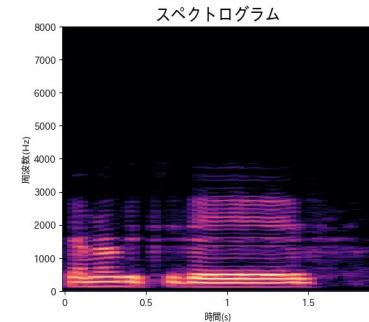
~50Hz



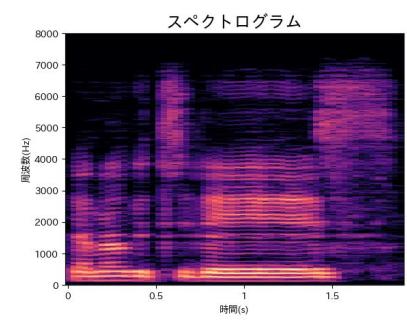
~300Hz



~3400Hz

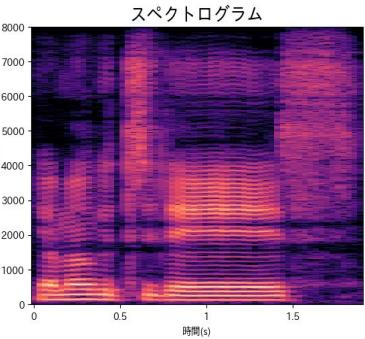


~7000Hz

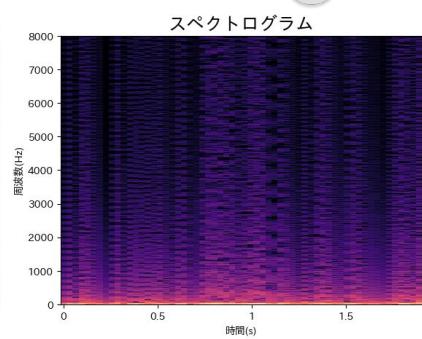


ローパスフィルタ(SP)

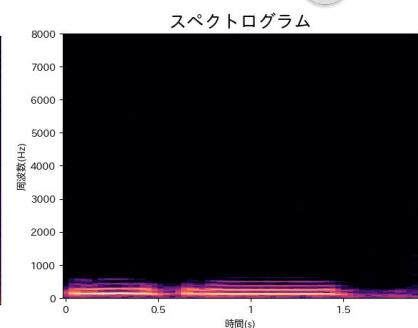
フィルタなし



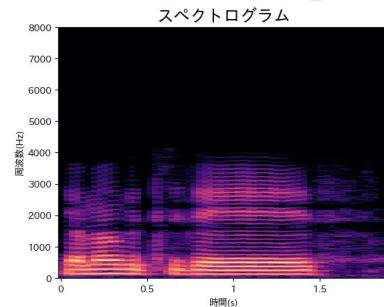
~50Hz



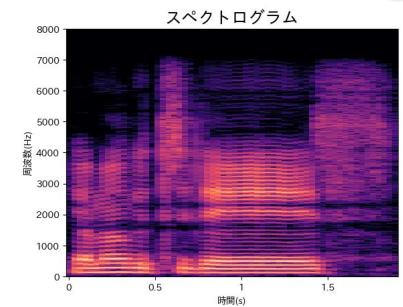
~300Hz



~3400Hz



~7000Hz



バンドパスフィルタ

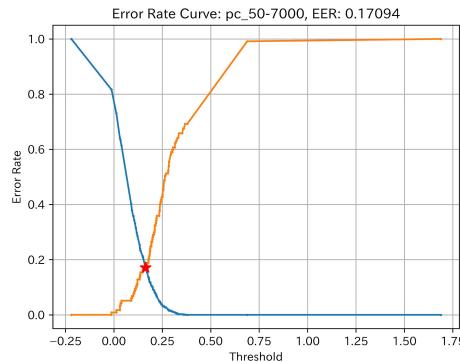
適用するフィルタは次の通り.

- 50~7000Hz
 - VoLTE規格
- 300~3400Hz
 - 標準規格
- 50~3400Hz
- 300~7000Hz

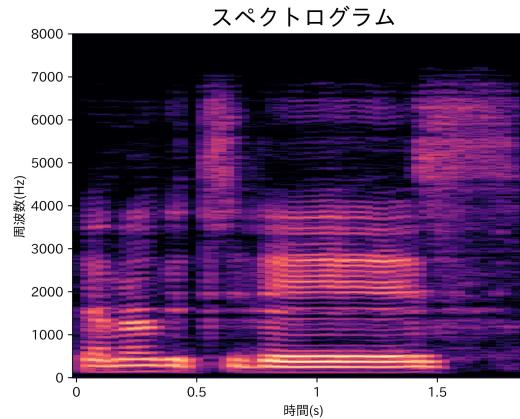
	フィルタなし	50~7000Hz	300~3400Hz	50~3400Hz	300~7000Hz
PC	0.1785	0.1709	0.2308	0.1880	0.1880
SP	0.1111	0.1180	0.1795	0.1379	0.1197

バンドパスフィルタ 50~7000Hz PC

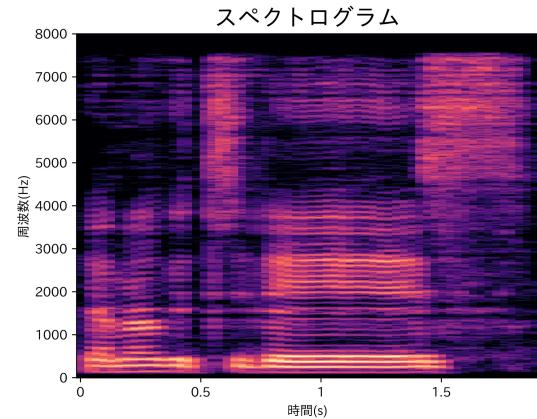
PC, EER: 0.1709



フィルタ適用後



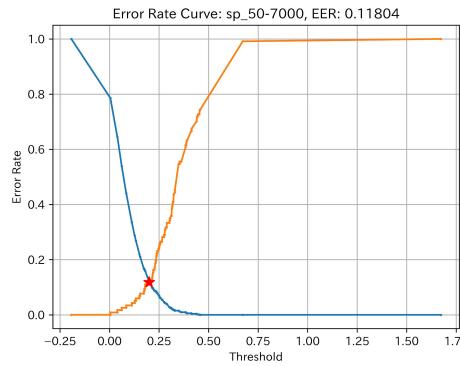
vanilla



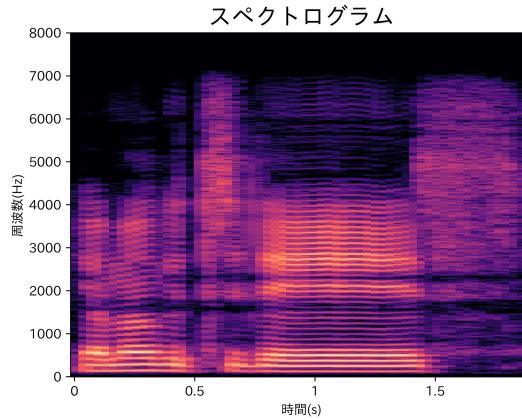
フィルタなし, PC, EER: 0.178

バンドパスフィルタ 50~7000Hz SP

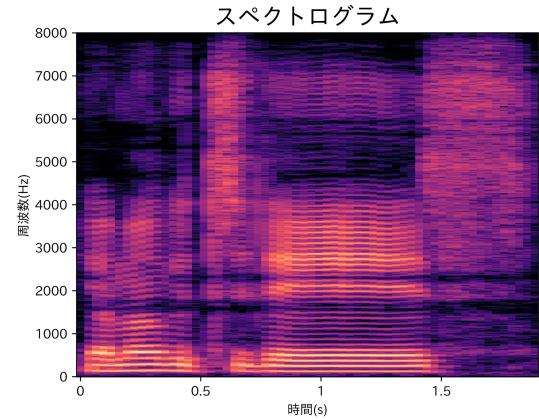
SP, EER: 0.1180



フィルタ適用後



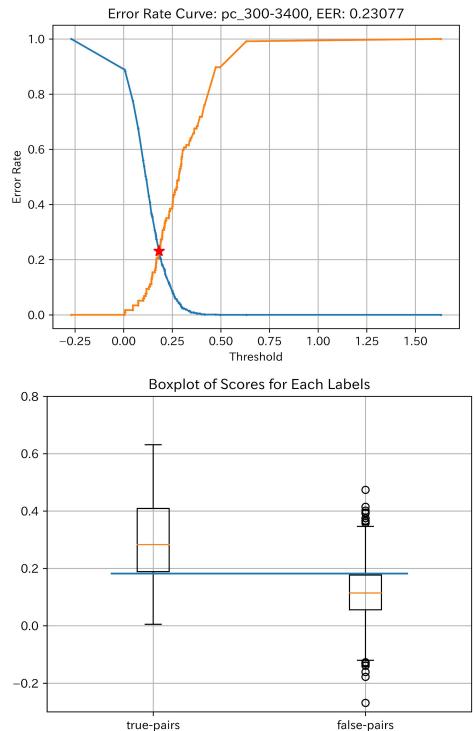
vanilla



フィルタなし, SP, EER: 0.111

バンドパスフィルタ 300~3400Hz PC

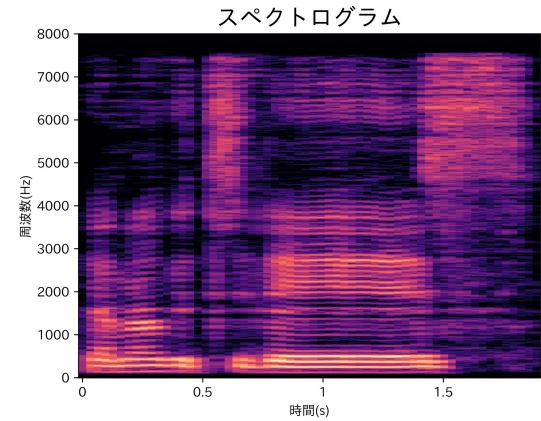
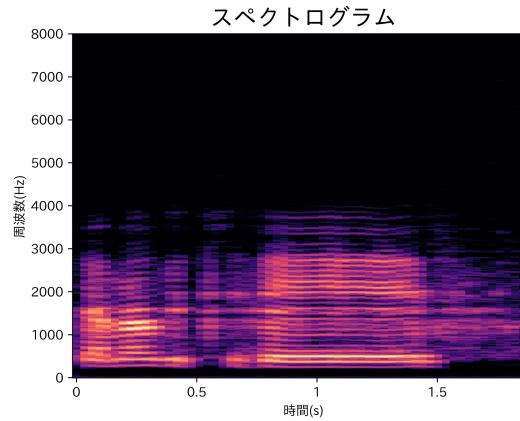
PC, EER: 0.2308



フィルタ適用後



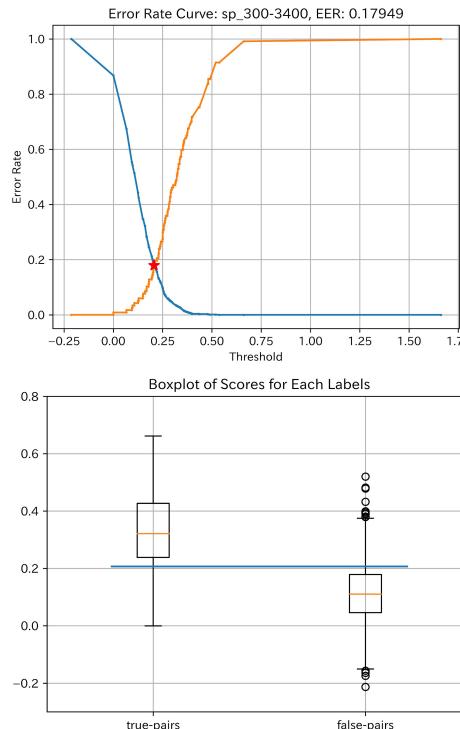
vanilla



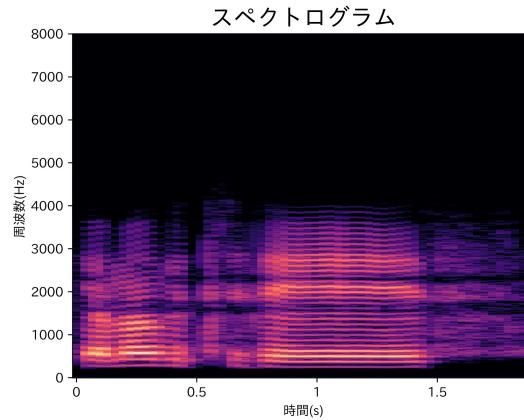
フィルタなし, PC, EER: 0.178

バンドパスフィルタ 300~3400Hz SP

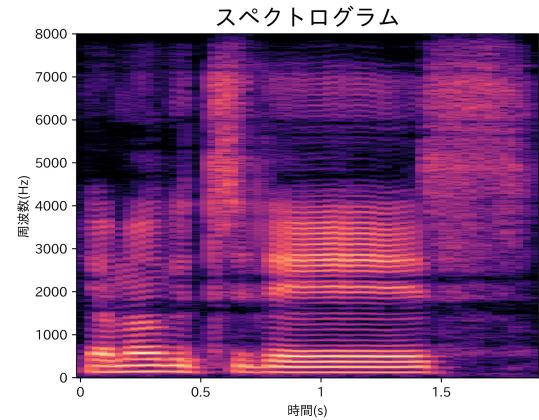
SP, EER: 0.1795



フィルタ適用後



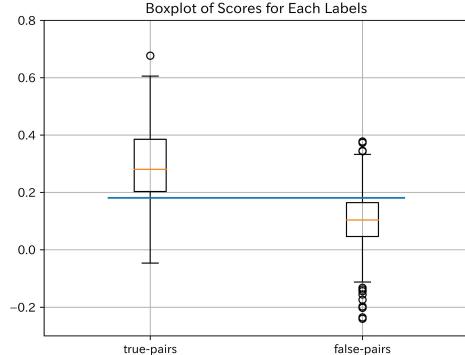
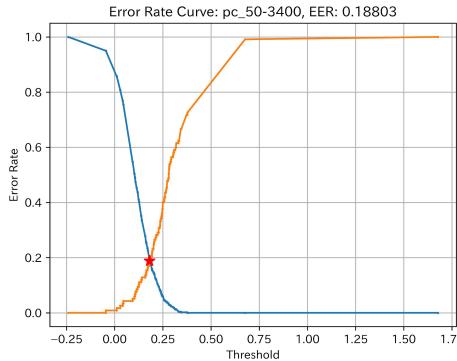
vanilla



フィルタなし, SP, EER: 0.111

バンドパスフィルタ 50~3400Hz PC

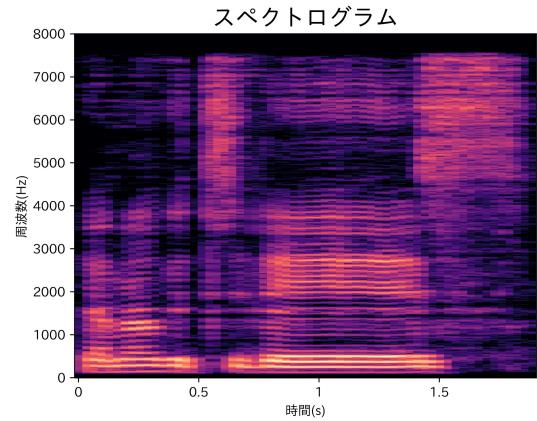
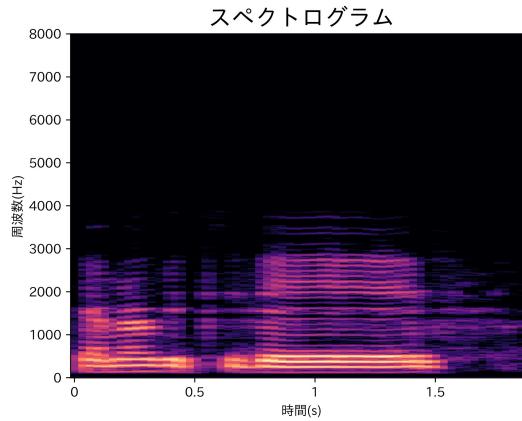
PC, EER: 0.1880



フィルタ適用後



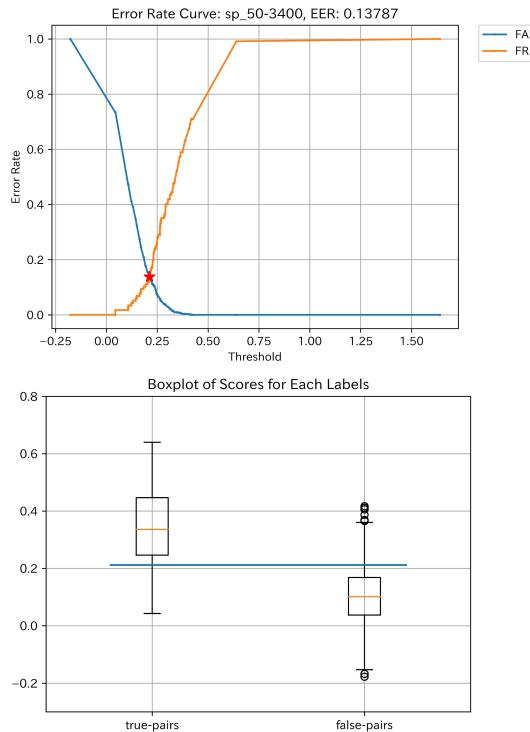
vanilla



フィルタなし, PC, EER: 0.178

バンドパスフィルタ 50~3400Hz SP

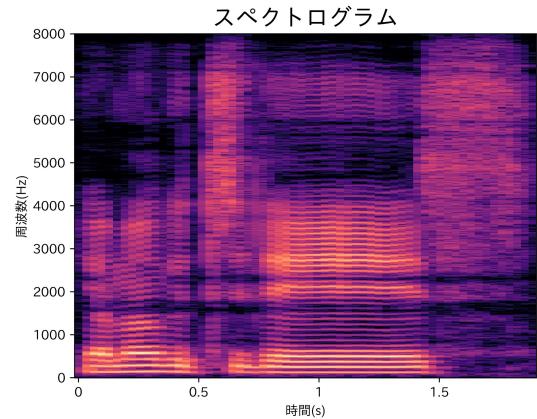
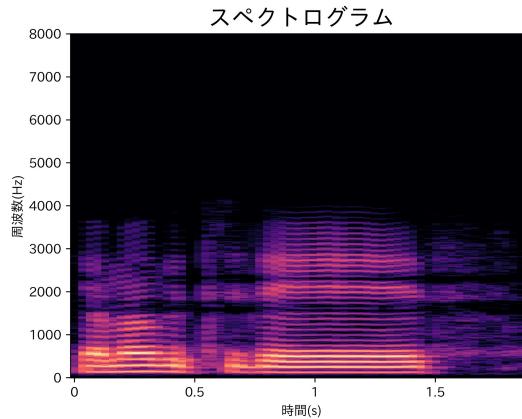
SP, EER: 0.1379



フィルタ適用後



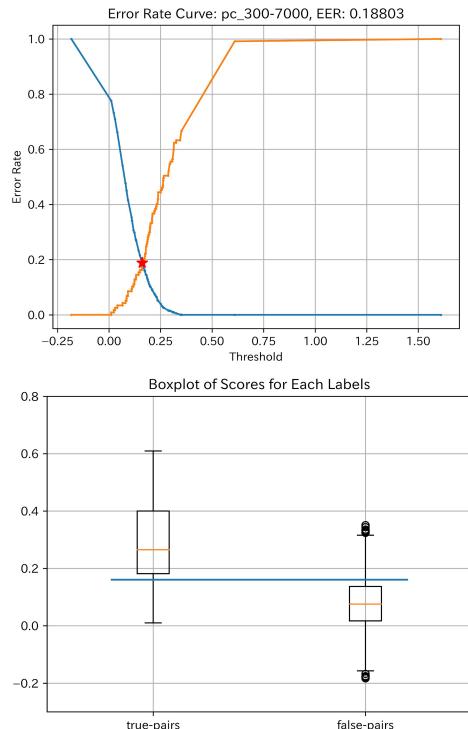
vanilla



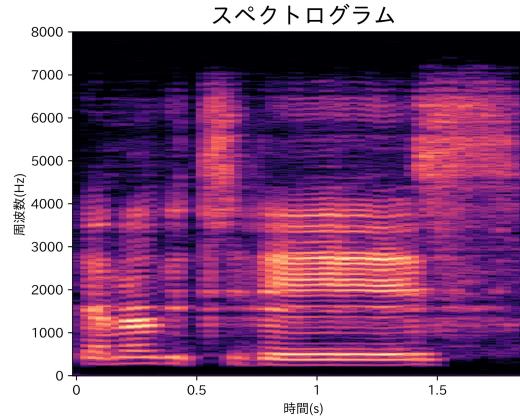
フィルタなし, SP, EER: 0.111

バンドパスフィルタ 300~7000Hz PC

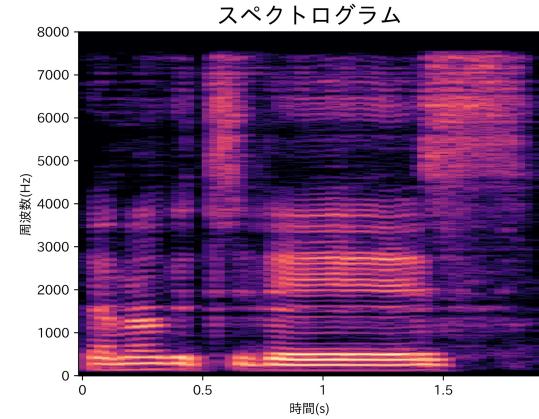
PC, EER: 0.1880



フィルタ適用後



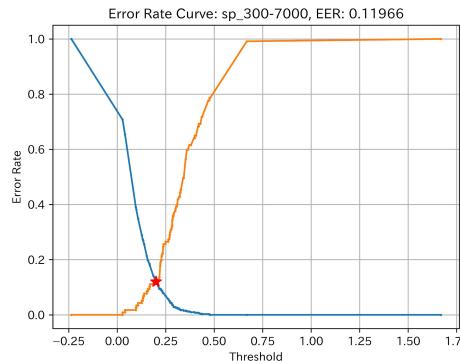
vanilla



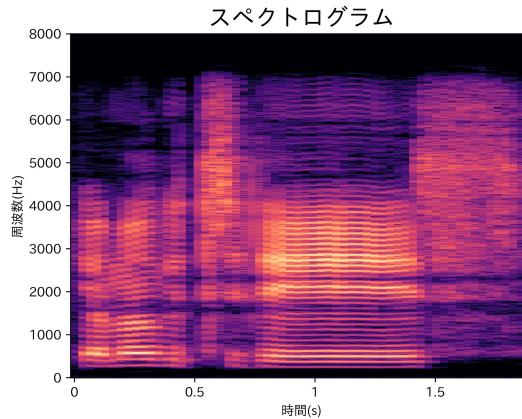
フィルタなし, PC, EER: 0.178

バンドパスフィルタ 300~7000Hz SP

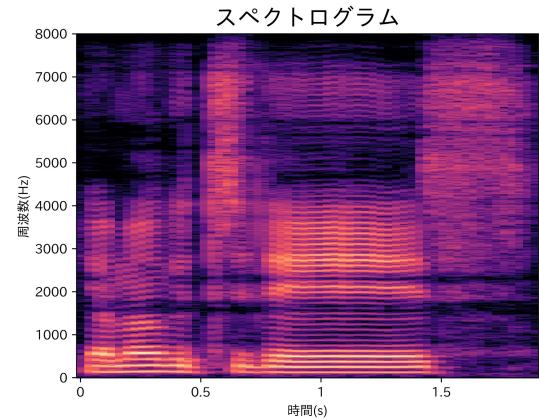
SP, EER: 0.1197



フィルタ適用後



vanilla



フィルタなし, SP, EER: 0.111

ハイパスフィルタ

適用するフィルタは次の通り.

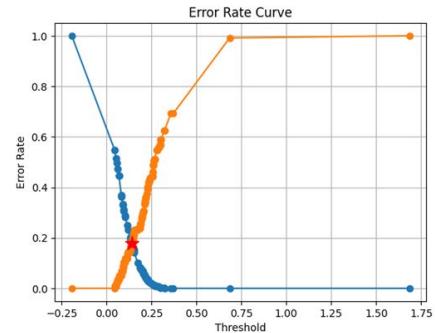
- 50Hz~
- 300Hz~
- 3400Hz~
- 7000Hz~

EER

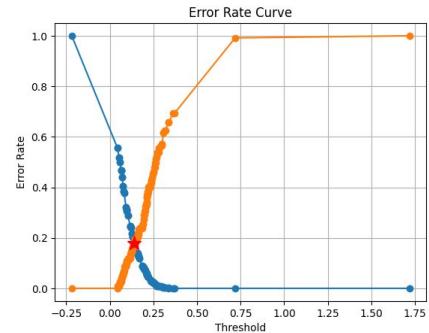
	フィルタなし	50Hz~	300Hz~	3400Hz~	7000Hz~
PC	0.1785	0.1795	0.1795	0.368	0.506
SP	0.111	0.111	0.114	0.316	0.485

ハイパスフィルタ(PC)

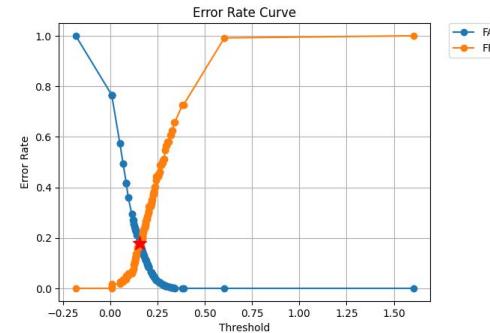
フィルタなし



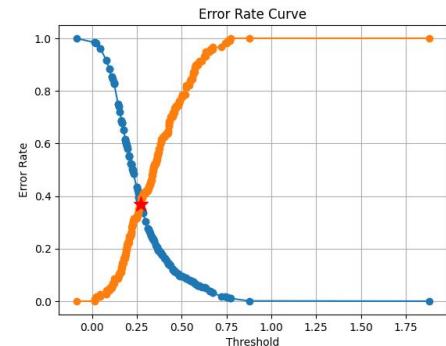
50Hz~



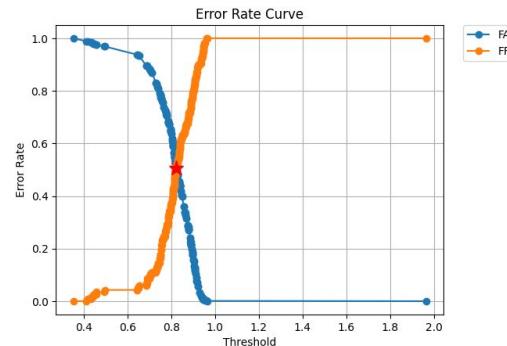
300Hz~



3400Hz~

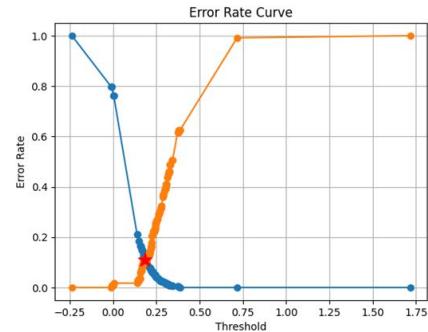


7000Hz~

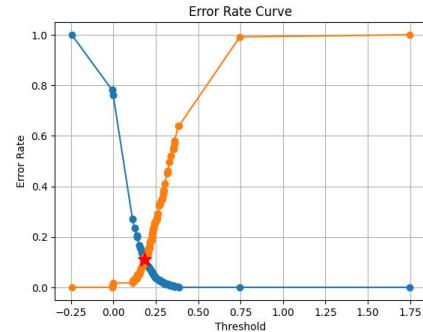


ハイパスフィルタ(SP)

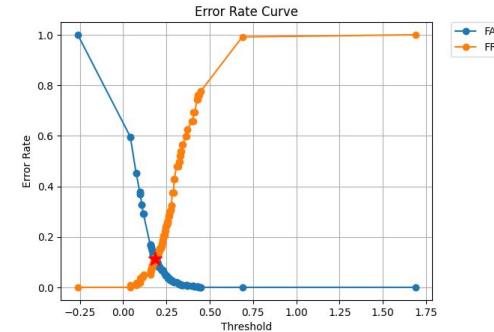
フィルタなし



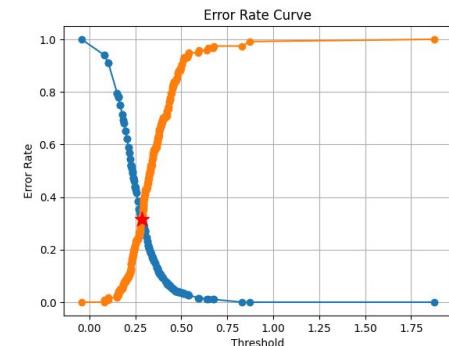
50Hz～



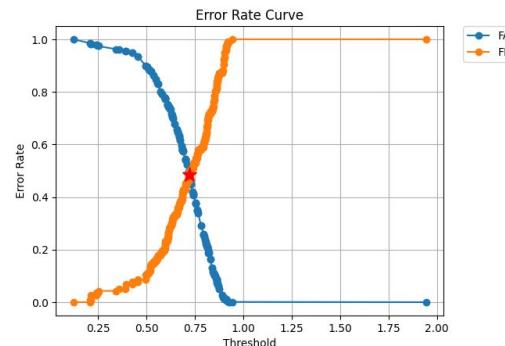
300Hz～



3400Hz～



7000Hz～



ハイパスフィルタ(PC)

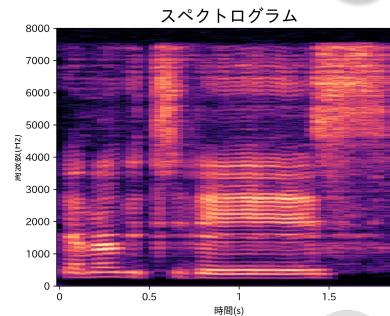
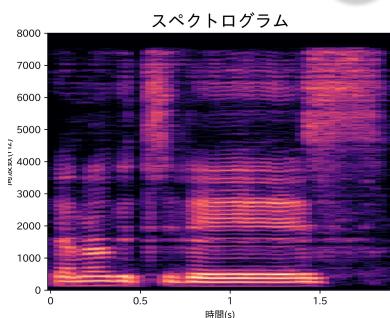
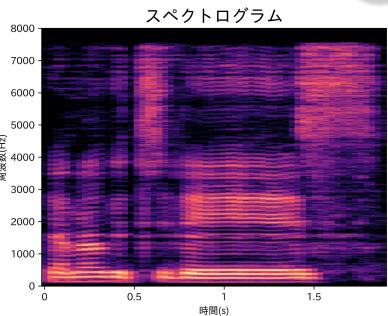
フィルタなし



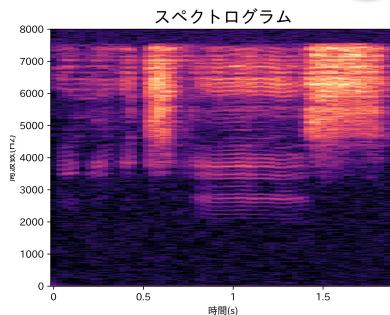
50Hz～



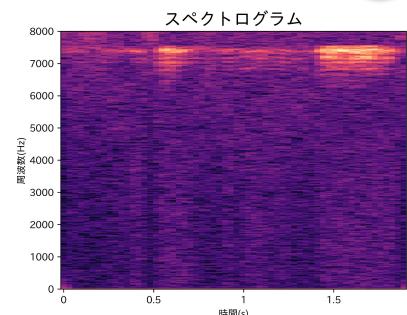
300Hz～



3400Hz～



7000Hz～

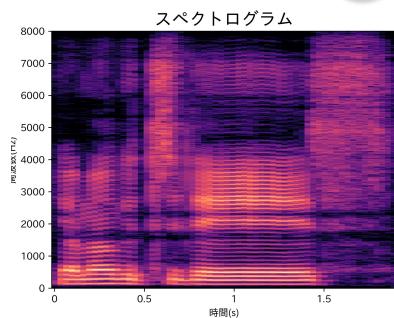
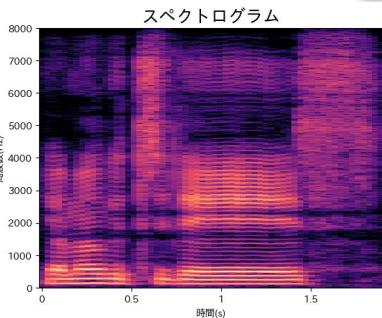


ハイパスフィルタ(SP)

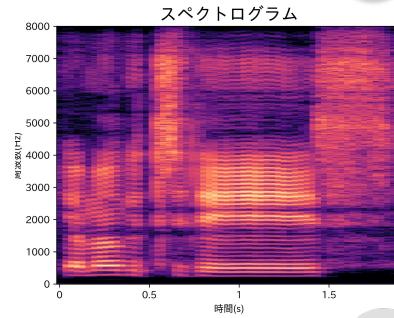
フィルタなし



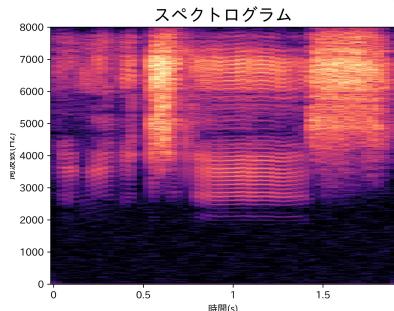
50Hz～



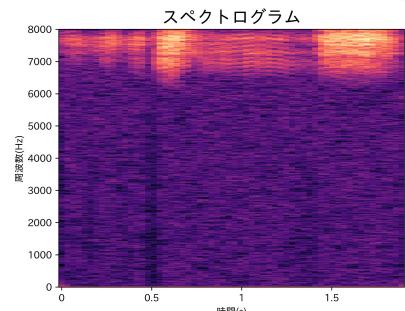
300Hz～



3400Hz～



7000Hz～



班が分かれてからのまとめ

第7回

パソコンで録音したものとスマホで録音したものを用意して、照合精度の違いを比較する

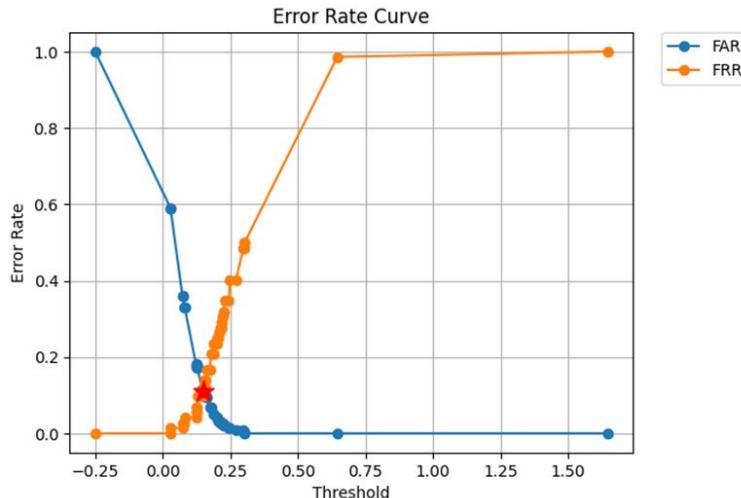
男性3人、女性3人の計6人で検証

(同じ人のペアが72、違う人のペアが393の計465ペア)

第7回

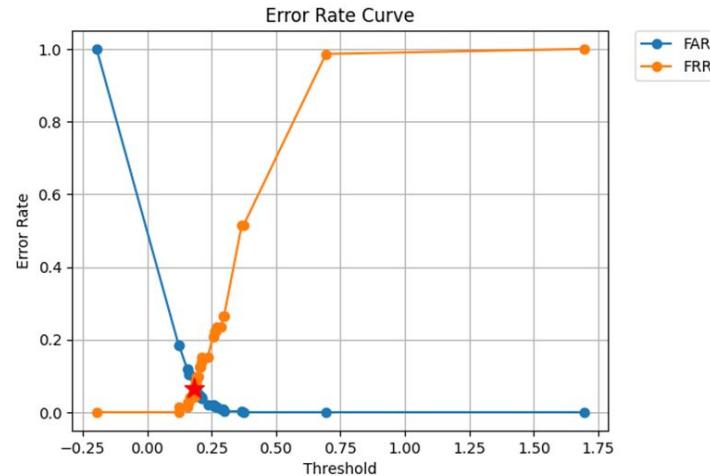
それぞれのError Rate Curve

PC



0.111

スマホ



0.066

第7回

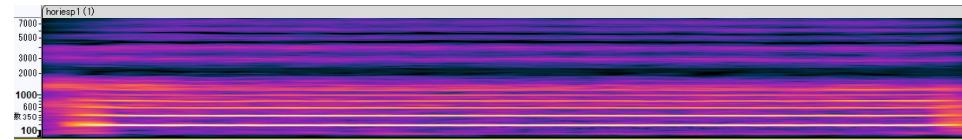
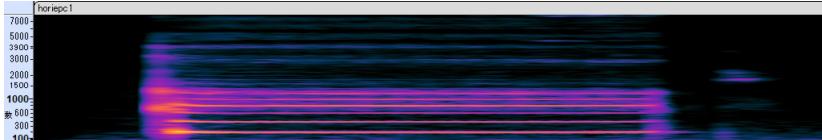
この結果から、PCで録音した音声よりスマホで録音した音声の方がEERが低いため、照合精度が高いと分かった。

しかし、それぞれの音声の長さや発話のタイミングが異なるため、あまり正しく検証できていないかもしれない。そこで、タイミングや長さを揃えて再検証を行う。

また、PCとスマホで周波数による成分に違いがあったため、フィルターを適用して、ある周波数だけを取り出して再検証を行う。

PC

スマホ



第8回

1. 発話開始のタイミングとファイルの長さを揃え、再検証する
2. 音声にフィルタを適用して、照合スコアの変化を検証する

第8回

1. 発話開始のタイミングとファイルの長さを揃え、再検証する

第7回時点での発話ファイルについて

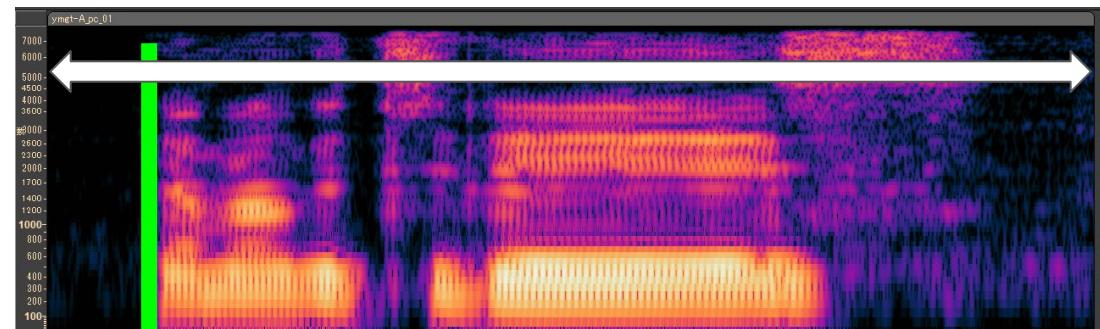
- 発話者が発話した音声を、パソコンとスマホで同時に録音した
- 収録した時点では、録音機器以外の条件は同一
- 録音環境毎に、発話数は31
 - 男3人女3人、男16発話(6+7+3)、女15発話(6+6+3)
 - 照合発話対は465(同一話者62対、異話者393対)
- 一度に複数の発話を録音し、発話ごとに切り出しを行った

切り出しに伴い、同一発話を記録した異デバイスの音声対では

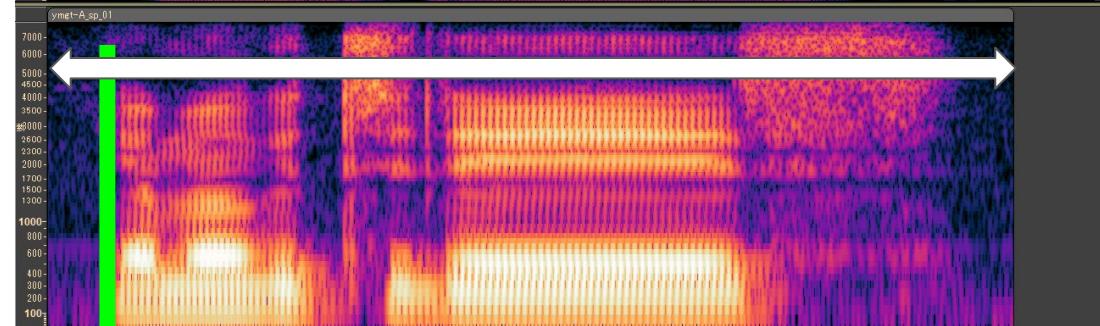
- 発話開始のタイミングが異なる
- 音声自体の長さが異なる

第8回

1. 発話開始のタイミングとファイルの長さを揃え、再検証する
 - 発話開始のタイミングが異なる
 - 音声自体の長さが異なる



上がパソコンで録音した音声
下がスマートフォンのもの



発話開始タイミングと音声の長さ
が異なっている。

第8回

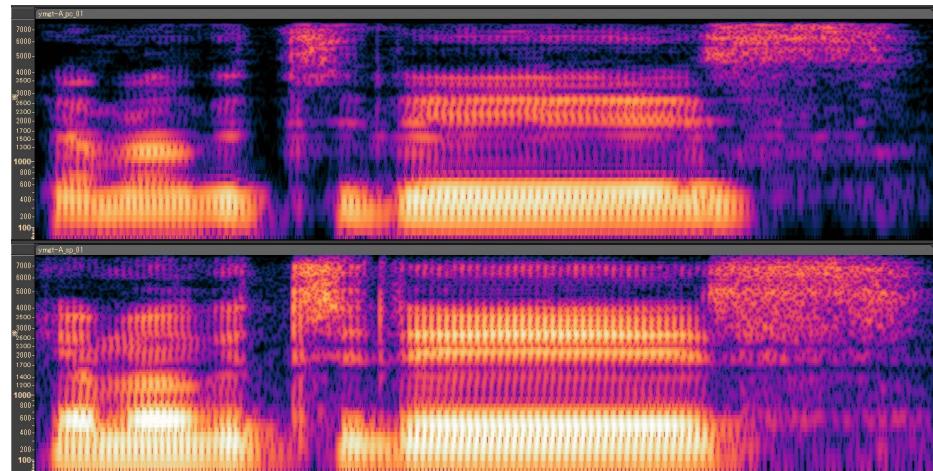
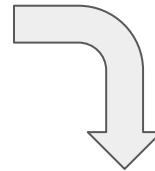
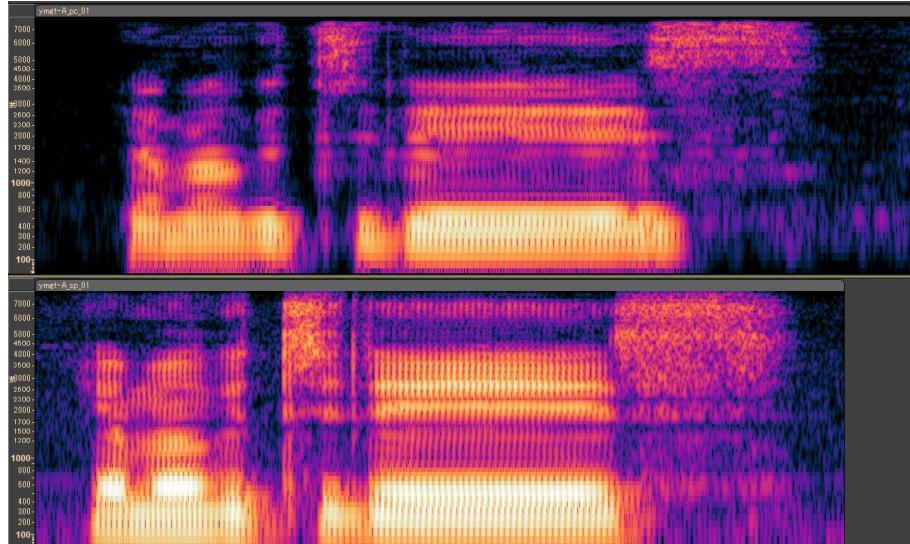
1. 発話開始のタイミングとファイルの長さを揃え、再検証する

音声が少し変わると話者埋め込みは大きく変化してしまう[1].

- 発話の切り出しによって生じた音声の長さと発話開始タイミングの違いが話者埋め込みの類似度に影響している可能性
 - 検証したい条件以外の差異が結果に影響している可能性がある
- これらの違いを除去し、再度検証する

第8回

1. 発話開始のタイミングとファイルの長さを揃え、再検証する

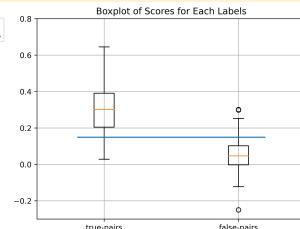
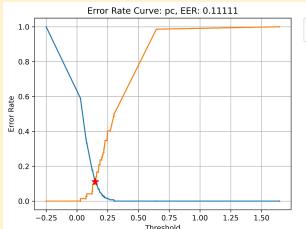


第8回

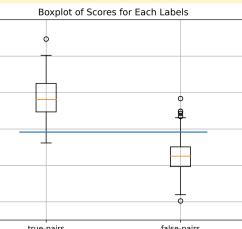
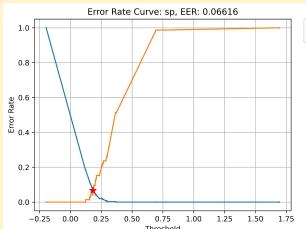
1. 発話開始のタイミングとファイルの長さを揃え、再検証する

第7回(音声が不揃い)

PC: EER=0.111

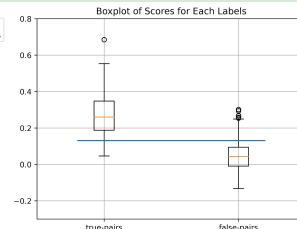
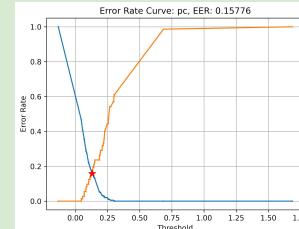


SP: EER=0.066

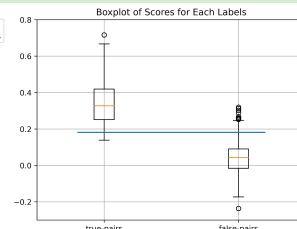
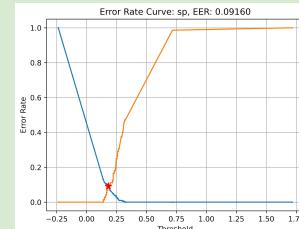


第8回(音声揃え済)

PC: EER=0.158



SP: EER=0.092



- 録音環境毎に、発話数は31

- 男3人女3人、男16発話(6+7+3)、女15発話(6+6+3)
- 照合発話対は465(同一話者62対、異話者393対)

第8回

- 録音環境毎に、発話数は31
 - 男3人女3人、男16発話(6+7+3)、女15発話(6+6+3)
 - 照合発話対は465(同一話者62対、異話者393対)

1. 発話開始のタイミングとファイルの長さを揃え、再検証する

	音声調整前	音声調整後
PC	0.111	0.158
SP	0.066	0.092

発話タイミングと音声ファイルの長さを調整したところ、照合のEERがおよそ3~4ポイント増加した。

録音環境以外の差異がほぼ無くなったため、音声調整後の値が本来の照合精度を表していると考えられる。

第8回

- 録音環境毎に、発話数は31
 - 男3人女3人、男16発話(6+7+3)、女15発話(6+6+3)
 - 照合発話対は465(同一話者62対、異話者393対)

1. 発話開始のタイミングとファイルの長さを揃え、再検証する

	音声調整前	音声調整後
PC	0.111	0.158
SP	0.066	0.092

デバイス間の照合精度を比較すると、スマートフォンの方が精度が良いことが分かる。

デバイスの性能はパソコンの方が高いため、音声に載るノイズが相対的に少なくなり、照合精度が上がると想定していたが、そのようにはならなかつた。

スマートフォンは音声通話が主とした機能の1つだが、パソコンはそうではなく補助的な機能である。このため、録音回路の性能はパソコンの方が高いが、マイクの性能自体はスマートフォンの方が高い可能性が考えられる。

第8回

2. 音声にフィルタを適用して、照合スコアの変化を検証する
 - 用意した音声の全てに同じフィルタを適用して照合を行い、エラー率を比較する。
 - 検証に用いる音声について
 - 適用するフィルタについて
 - 検証結果

第8回

2. 音声にフィルタを適用して、照合スコアの変化を検証する

検証に用いる音声について

- 前項で使用した音声群に男性3人の発話と加えたもの
 - 録音環境毎に49発話
 - 男6人、女3人
 - 男34発話(6+7+3+6+6) 女15発話(6+6+3)
- 各発話について、異なるデバイスにより同時に録音
- 発話開始のタイミングと音声ファイルの長さは、デバイス間でそろえてある

第8回

2. 音声にフィルタを適用して、照合スコアの変化を検証する

適用するフィルタについて

- フィルタの定義と適用はSciPyライブラリを使用する
 - ver: 1.11.4
- Butterworthフィルタを作成する。
 - 次数は5次(バンドパスは10次)
 - ローパス, ハイパス, バンドパス

異なる周波数でフィルタを作成し、音声ファイルに適用、EERを算出する。

適用するフィルタの種類による照合スコアの変化を検証する。

第8回

2. 音声にフィルタを適用して、照合スコアの変化を検証する

検証結果

適用するフィルタ		フィルタなし	ローパス		バンドパス			ハイパス
通過周波数(Hz)		-	~3000	~5000	100~1000	300~3400	3000~6000	500~
EER	PC	0.178	0.171	0.188	0.264	0.230	0.419	0.141
	SP	0.111	0.130	0.178	0.213	0.179	0.393	0.195

- ローパスフィルタとバンドパス(3000~6000Hz)の結果から、高周波の成分は照合にあまり重要でないと考えられる。
- 500Hz以上を通すハイパスフィルタを適用した結果が最も照合精度が高かったため500Hzより下の周波数成分は話者間であまり差がない可能性がある
- フィルタを設計する際に注目する周波数がフィルタごとに異なっており、正確な検証がしづらい
 - 第9回にて、注目する周波数をフィルタ間で統一させ、再度検証する。

第9回

第8回で用いた音声に以下のようなフィルタを適用し, EERを比較する

適用するフィルタは第8回のものと同じもの. ただし、周波数は統一する

- 50, 300, 3400, 7000Hzのハイパスフィルタとローパスフィルタ
- 50~7000Hz(VoLTE規格), 300~3400Hz(標準規格), 50~3400Hz, 300~7000Hzのバンドパスフィルタ

第9回

それぞれのフィルタについて、適用した中で最もEERが小さかったのは以下の周波数のときであった

PC収録の場合

- ハイパスフィルタ: 50Hz~
- ローパスフィルタ: ~3400Hz
- バンドパスフィルタ: 50~7000Hz

第9回

スマホ収録の場合

- ハイパスフィルタ: 50Hz~
- ローパスフィルタ: ~7000Hz
- バンドパスフィルタ: 50~7000Hz

以上の結果より、どちらの機器においても、~50Hzと7000Hz~にエラー率を上げるノイズが多くあると考えられる。

しかし、ローパスフィルタについて、PCとスマホでEERが最小の周波数に違いがったため、実際はどの周波数域にノイズが多くあるのかをもっと周波数を細かく分けて実験する必要があると考えられる。