信号処理における
機械学習的クラスタリング手法の開発

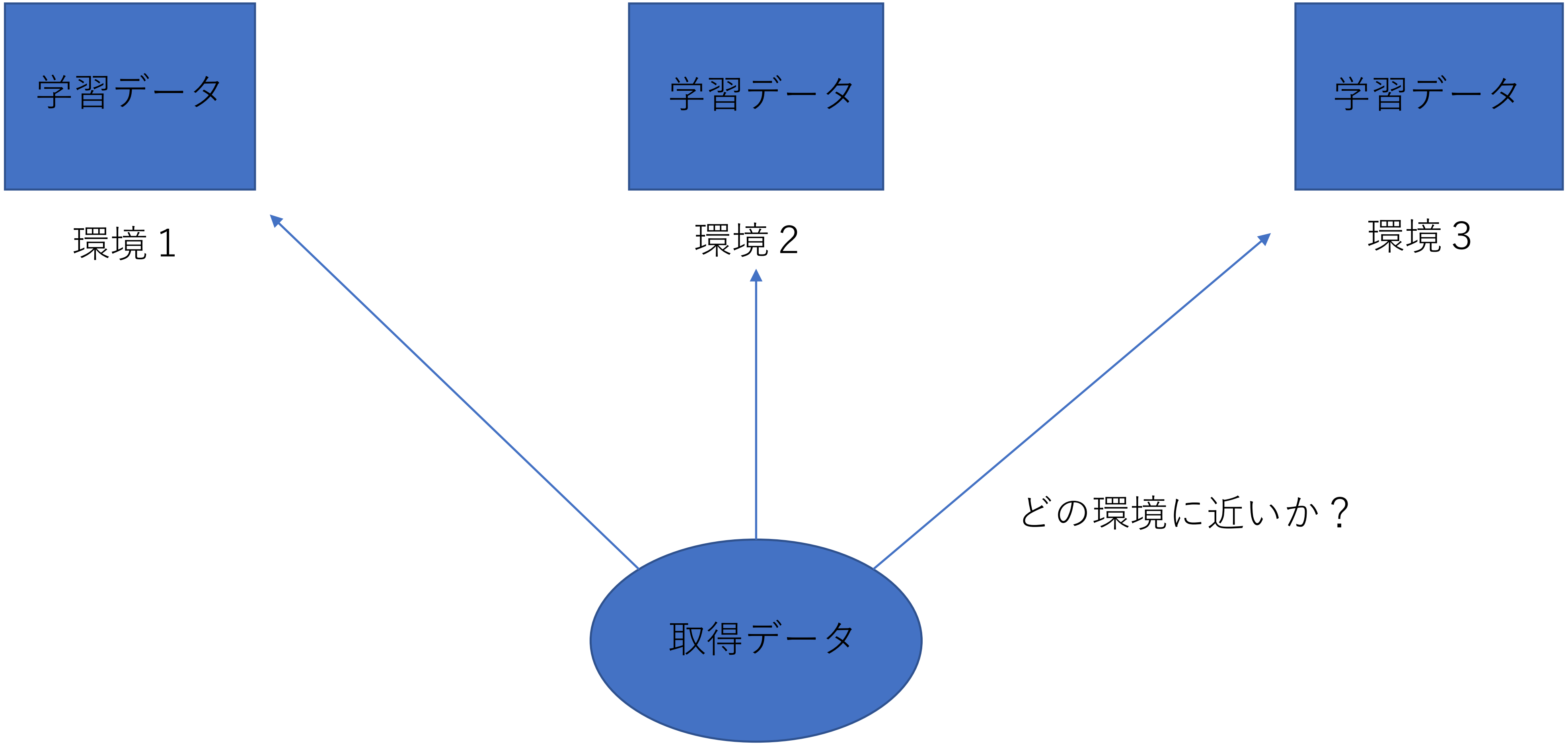
富山県立大学工学部　奥原研究室　麻生到

はじめに

センサや計測器から取得できる信号は、環境の変化の影響で常に同じような信号が取得できるとは限らない

環境の変化に対応できる処理が必要となる

適用手法



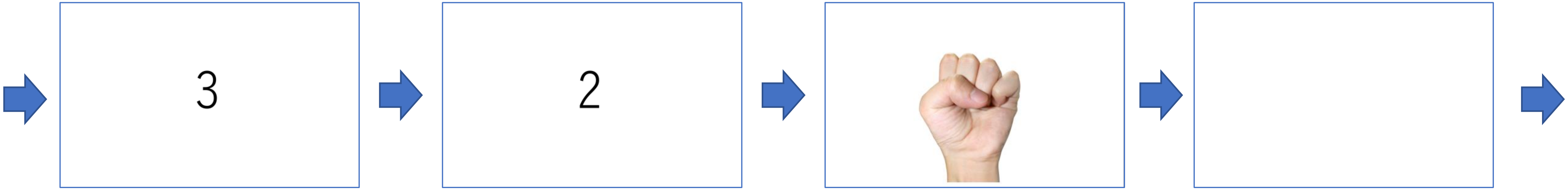
テストデータだったら、どの環境の学習データを使うか？

本研究の目的

環境の変化に対応できる信号処理システムの作成

脳波データを用いてじゃんけんで使用者に勝つシステムの作成

じゃんけんシステムの流れ



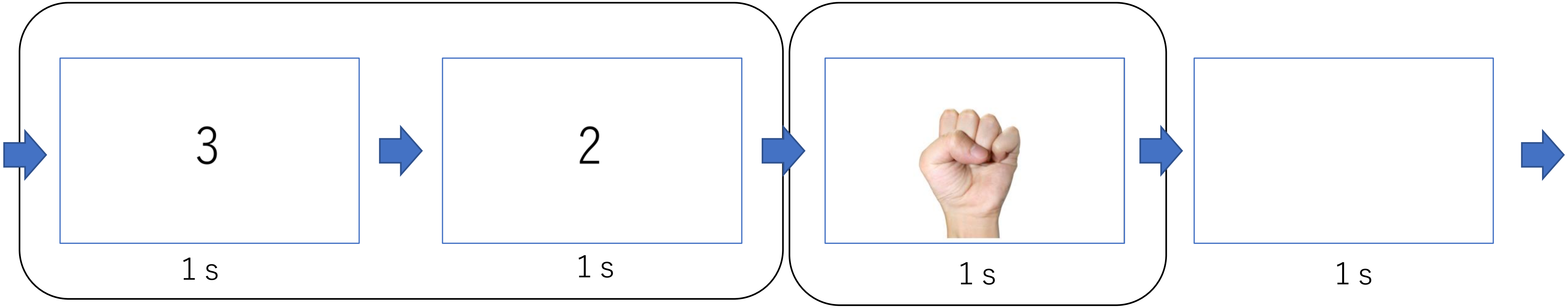
1 s　1 s　1 s　1 s

待機

ランダムでグー・チョキ・パーの
いずれかを表示

前のじゃんけんスライドに
勝つ手を出す

スライド別の処理



３と２スライドから環境の変化の対応を行う

じゃんけんスライド時の
１秒間のデータから判別

クラスタリング

ベイズ理論を用いる

事後分布 ＝ 尤度関数 ＋ 事前分布

・尤度関数
あるパラメータが与えられた時に、データがどのような分布をしているか

・事前分布
どんなパラメータが与えられやすいか（パラメータの散らばり具合）

データの事後分布のパラメータがどのような値になるかを推定

➡ 事後分布を使うことで、未知のデータが予測可能

パラメータを推定することが目的

本研究では、高次元ベクトルが従う確率分布を学習

目的：脳波データから３クラス分類を行いたい（グー・チョキ・パー）

前処理

脳波データを3-50Hzのパワースペクトルに周波数解析

⇒ 分類に用いるデータはパワースペクトル(94次元)

取得したパワースペクトルはガウス分布に従うと仮定

ガウス分布のパラメータ
・平均ベクトル
・分散共分散行列

つまり、３クラスそれぞれのガウス分布のパラメータを求め、確率分布を推定する

クラスタリング

訓練データx_nについて

$$\bar{x}_k = \frac{1}{N_k}\sum_{n=1}^{N} x_n$$

$$\boldsymbol{S}_k = \frac{1}{N_k}\sum_{n=1}^{N}(x_n - \bar{x}_n)(x_n - \bar{x}_n)^T$$

あるクラスkにおける

N_k は各クラスの要素数であり，$\bar{x}_k$ は各クラスの訓練データの平均ベクトルと $\boldsymbol{S}_k$ は各クラスの分散共分散行列である．

事後分布は以下のように表される

$$q(\mu_k, \Lambda_k) = N(\mu_k | m_k, (\beta_k \Lambda_k)^{-1}) W(\Lambda_k | W_k, \nu_k)$$

パラメータは以下のように定義した

$$\beta_k = \beta_0 + N_k$$

$$m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \bar{x}_k)$$

$$W_0 = \alpha E$$

$$W_k^{-1} = W_0^{-1} + N_k \boldsymbol{S}_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(x_n - \bar{x}_n)(x_n - \bar{x}_n)^T$$

$$\nu_k = \nu_0 + N_k$$

$$\ln \widetilde{\Lambda_k} = \sum_{i=1}^{D} \psi\left(\frac{\nu_k + 1 - i}{2}\right) + D \ln 2 + \ln |W_k|$$

D：次元数　E：単位行列

クラスタリング

y_n：テストデータ

各クラスの訓練データで求めたパラメータβ_k、ν_k、m_k、$\ln\widetilde{\Lambda_k}$ を以下の近似式に代入

$$\ln r_{nk} \simeq \frac{1}{2}\ln\widetilde{\Lambda_k} - \{-\frac{D}{2\beta_k} - \frac{\nu_k}{2}(y_n - m_k)^T W_k (y_n - m_k)\}$$

$$\ln r_n = \{\ln r_{n1}, \ln r_{n2}, \cdots \ln r_{nM}\} \quad (12)$$

負担率 r_n の集合の中の最大値 c とし，各要素で引く

$$\ln r_n - c = \{\ln r_{n1} - c, \ln r_{n2} - c, \cdots \ln r_{nM} - c\} \quad (13)$$

そして，集合の要素を自然対数 e のべき乗をとる

$$\mathrm{e}^{(\ln r_n - c)} = \{\mathrm{e}^{(\ln r_{n1} - c)}, \mathrm{e}^{(\ln r_{n2} - c)}, \cdots \mathrm{e}^{(\ln r_{nM} - c)}\} \quad (14)$$

集合 (14) の総和 S_r を求め，各要素を総和 S_r で割ることで採択確率の集合 p を求める．

$$S_r = \sum_{k=1}^{M} \mathrm{e}^{(\ln r_{nk} - c)} \quad (15)$$

$$p = \{\frac{\mathrm{e}^{(\ln r_{n1} - c)}}{S_r}, \frac{\mathrm{e}^{(\ln r_{n2} - c)}}{S_r}, \cdots \frac{\mathrm{e}^{(\ln r_{nM} - c)}}{S_r}\} \quad (16)$$

採択確率の集合 p の中で最も大きい値をもつクラスにテストデータを分類する．

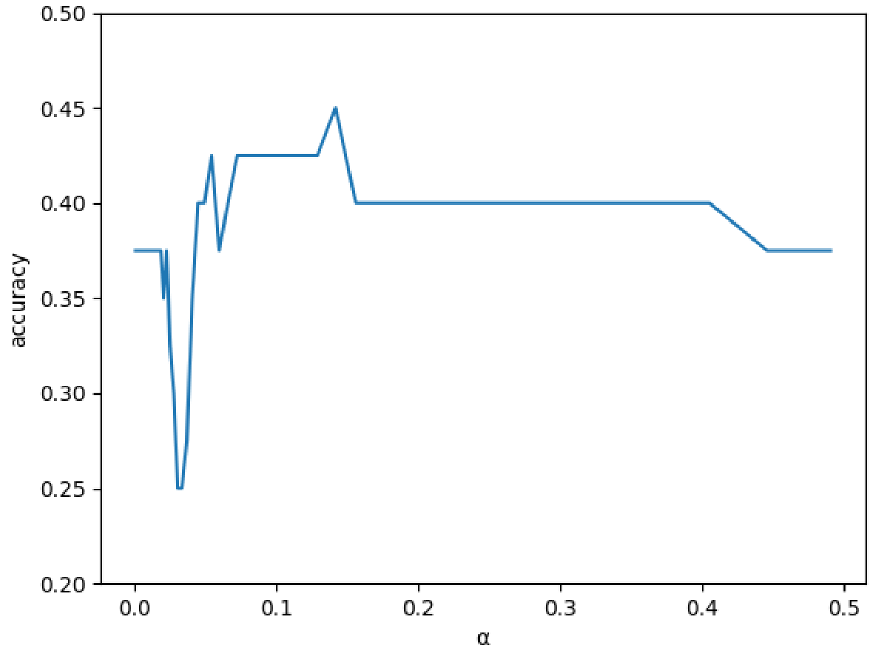
本研究では、３クラス分類を行うので集合ｐの要素は３つとなる

評価実験

1日分の環境変動なしの訓練データ60セットとテストデータ40セットを用いて分類を行った

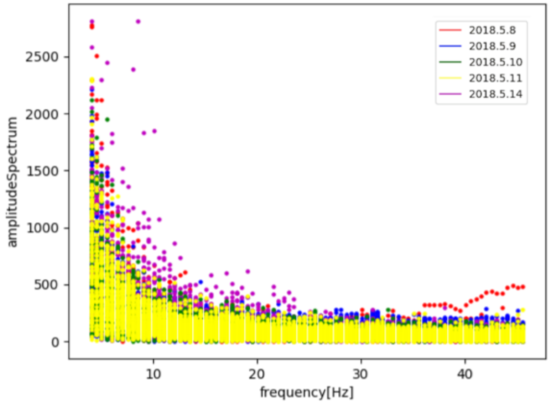
$$W_0 = \alpha E$$

パラメータαの値と精度の関係を以下のグラフに示す

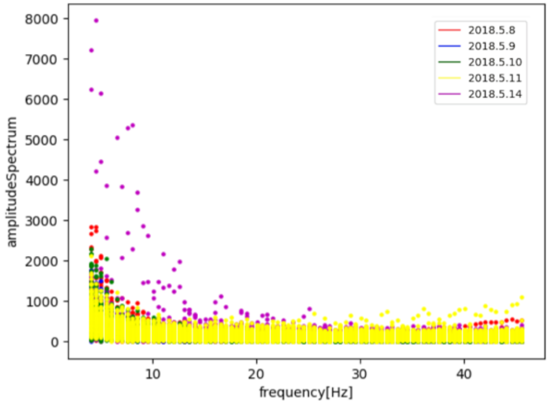


α＝0.142のときに 0.45と最も精度が高かった

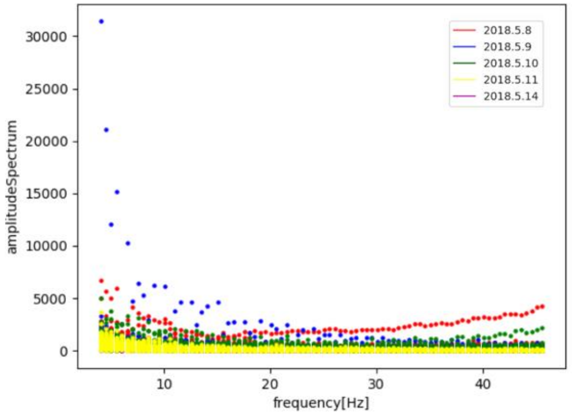
環境変動について



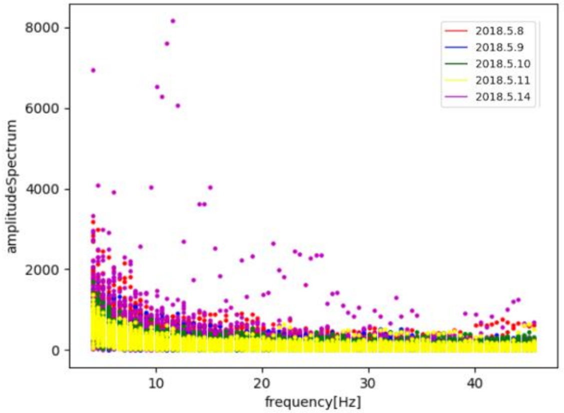
CH1



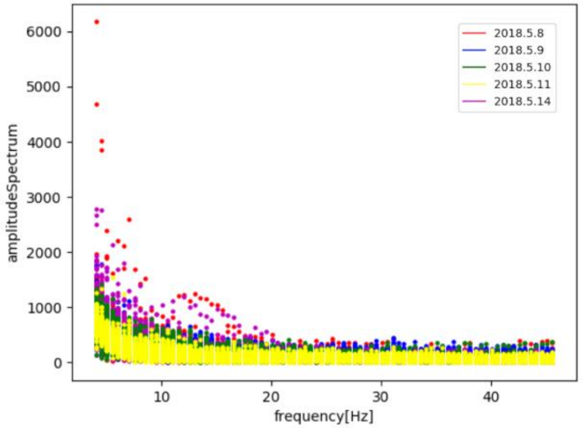
CH2



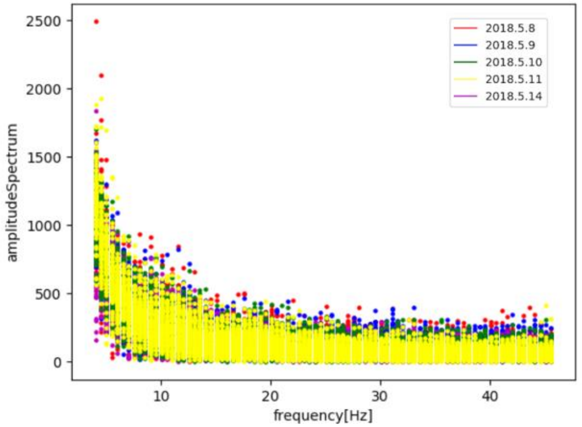
CH 3



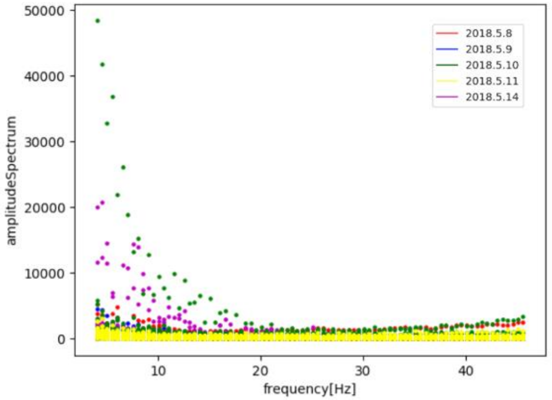
CH 4



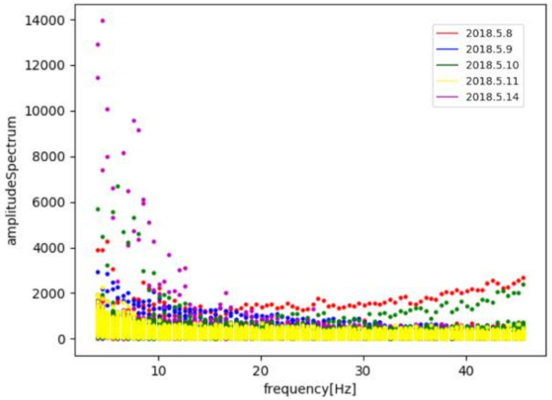
CH 5



CH 6



CH 7



CH 8

今後の課題

・環境変動を考慮した仕組みをつくる
（正規分布のパラメータである平均と分散を用いる。指標はKLダイバージェンス）