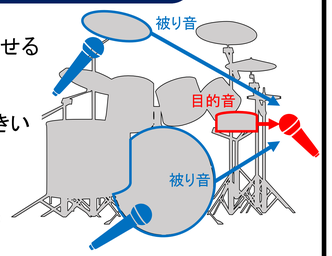
1-R-34

音量比特徴量の重み付きクラスタリングによるドラムセット録音時の被り音抑圧

☆鈴木慶，北村大地（香川高専）

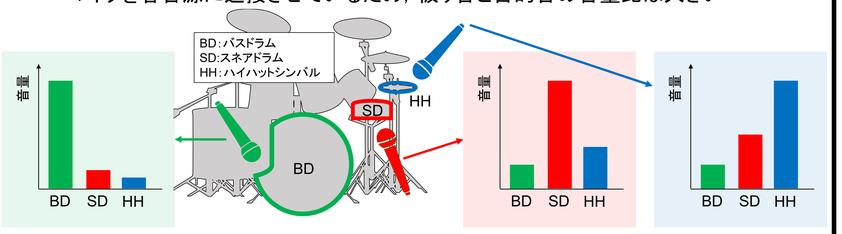
1. 研究背景

- **ドラムセットのマイキング**
  - ドラムセットの録音時にはマイクを各音源に近接させる
  - 目的音源の音（目的音）の観測と
    目的音源以外の音（被り音）の抑圧が目的
  - マイクを各音源に近接させるだけでは被り音が大きい
    - ミキシングの音質低下につながる



➡ 目的音は維持したまま，さらに被り音を
抑圧できる信号処理技術が求められている

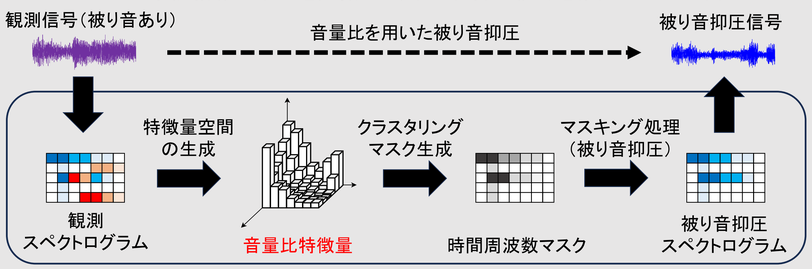
- **DSS（drum source separation）**
  - 目的に近い研究として，ドラムセットの演奏音から各音源に分離するDSSが存在
  - DSSは深層ニューラルネットワーク（deep neural network: DNN）を利用
    - モノラル・ステレオ信号の大規模なデータセットを使用[A. I. Mezza+, 2024]
- **被り音抑圧を目的としたDNN**
  - マイクを各音源に近接させているため，被り音と目的音の音量比は大きい



  - この特徴を表現できる多チャネル信号をDNNのデータセットとして扱う方が効果的
    しかし，大規模な多チャネル信号データセットはほとんど存在せず作成することも困難
- **小規模な多チャネル信号のDNNへの活用方法**
  - 多チャネル信号のデータセットが小規模でもDNNの事前情報としてなら活用可能
  - 例えば，音量比を活用して大まかに被り音を抑圧した信号を，DNNの事前情報として与える

本研究の目的

- **多チャネル信号がもつ音量比を活用し，大まかな被り音抑圧を行う**
  - 音量比特徴量の生成，クラスタリング，マスキング処理を経て被り音抑圧を行う
  - DNNへの事前情報の活用を見据えた，被り音抑圧信号を推定できるか確認



2. データセット作成

- **ドラムセットの多チャネル信号データセット作成**
  - ドラムセットのマイキングを想定し，目的音源にマイクを近接して配置
  - 9種類の音源，3種類の演奏パターンで演奏
  - 全音源を演奏した信号と各音源を個別に演奏した信号を収録
  - 本研究の被り音抑圧実験に使用，今後は公開予定

- **各音源の音量比**
  - BD・SDマイクに対する被り音は小音量
  - HHマイクに対するSDの被り音は
    目的音よりも大音量で観測

➡ 音量比を用いたHHマイクの
被り音抑圧は困難であると予想

目的音を0 dBとした，被り音の音量比（dB）

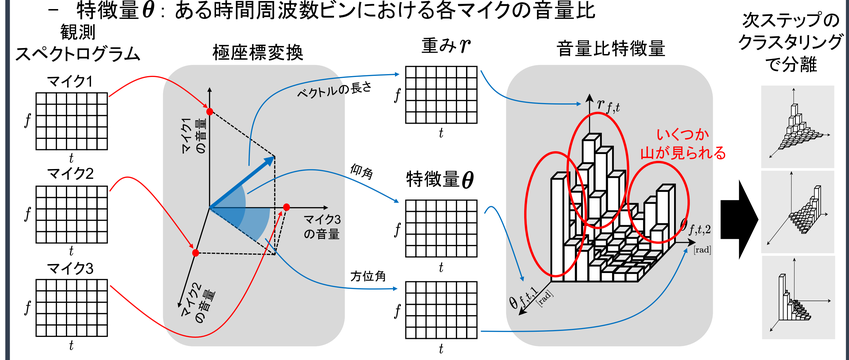
マイク ＼ ソース	KD	SD	HH
KD	0	-13.41	-42.43
SD	-25.98	0	-23.41
HH	-6.62	9.24	0

3.教師無し手法による被り音抑圧

- 音量比を活用し，以下の3ステップで大まかな被り音抑圧を推定

――― *Step1*: 音量比特徴量への変換 ―――

- 各信号の音量比を極座標で表現
  - 重み r：ある時間周波数ビンにおける全音源の音量の平均
  - 特徴量 $\boldsymbol{\theta}$：ある時間周波数ビンにおける各マイクの音量比



――― *Step2*: 重み付きクラスタリング ―――

- **重み付きk平均法**
  - 重みを考慮した更新式が収束するまでセントロイド μ_kとデータの属するクラスタを推定

$$\mu_k = \frac{\sum_{(f,t)\in C_k} r_{f,t}\theta_{f,t}}{\sum_{(f,t)\in C_k} r_{f,t}}$$

重みが大きいほど影響力が高い

k ：クラスタのインデクス
C_k ：クラスタ k に属する f,t の集合

  - 得られたセントロイドと各データ点から，
    分散共分散行列Σ_kを導出し，各クラスタを形成するガウス分布を推定

$$\boldsymbol{\Sigma}_k = \frac{\sum_{(f,t)\in C_k} r_{f,t}(\boldsymbol{\theta}_{f,t}-\boldsymbol{\mu}_k)(\boldsymbol{\theta}_{f,t}-\boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{(f,t)\in C_k} r_{f,t}}$$

各クラスタを形成するガウス分布 $\mathcal{N}_k(\boldsymbol{\theta}|\mu_k,\Sigma_k)$

- **重み付き混合ガウス分布モデル（Gaussian mixture model: GMM）**
  - 期待値ステップ（E-step）と最大化ステップ（M-step）を
    収束するまで繰り返し，各クラスタを形成するガウス分布を推定
  - E-step

$$\gamma_{f,t,k} = \frac{r_{f,t}\,\pi_k \mathcal{N}(\boldsymbol{\theta}_{f,t}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{k'}\pi_{k'}\mathcal{N}(\boldsymbol{\theta}_{f,t}|\boldsymbol{\mu}_{k'},\boldsymbol{\Sigma}_{k'})}$$

重みが大きいほど影響力が高い

$\gamma_{f,t,k}$ ：k番目のガウス分布から生成された事後確率
π_k ：k番目のガウス分布の寄与率

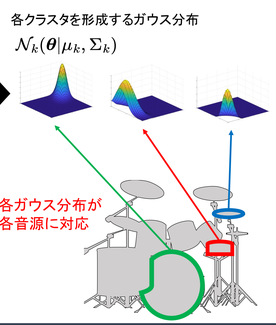
  - M-step

$$\boldsymbol{\mu}_k = \frac{\sum_f\sum_t \gamma_{f,t,k}\boldsymbol{\theta}_{f,t}}{\sum_f\sum_t \gamma_{f,t,k}}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_f\sum_t \gamma_{f,t,k}(\boldsymbol{\theta}_{f,t}-\boldsymbol{\mu}_k)(\boldsymbol{\theta}_{f,t}-\boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_f\sum_t \gamma_{f,t,k}}$$

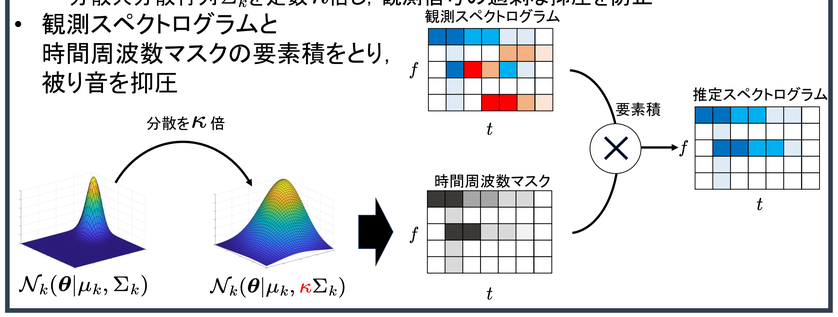
$$\pi_k = \frac{\sum_f\sum_t \gamma_{f,t,k}}{\sum_{k'}\sum_f\sum_t \gamma_{f,t,k'}}$$

各クラスタを形成するガウス分布 $\mathcal{N}_k(\boldsymbol{\theta}|\mu_k,\Sigma_k)$



――― *Step3*: マスク生成と被り音抑圧 ―――

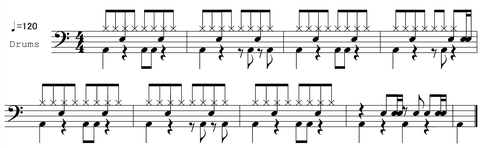
- クラスタリングから得られたガウス分布より時間周波数マスクを生成
  - 分散共分散行列Σ_kを定数 κ倍し，観測信号の過剰な抑圧を防止
- 観測スペクトログラムと
  時間周波数マスクの要素積をとり，
  被り音を抑圧



4. 実験

- **実験条件**
  - 2節で作成したデータセットのうち，BD，SD，及びHHを使用
  - クラスタリングは20回分異なる乱数を使用し，その中から最良のモデルを採用

使用した演奏パターン



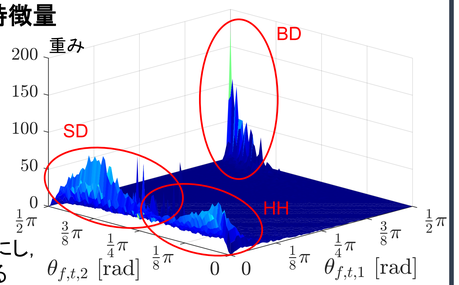
観測信号及びSTFTの設定値

サンプリング周波数	48000 Hz
STFT窓関数	Blackman
STFT窓長	2048
STFTシフト長	1024

- **実録音から生成された音量比特徴量**
  - 山が3つ見られ，
    BD,SD及びHHの音源に対応
  - BDの山は孤立している
  - SDとHHは干渉する部分が存在

HHマイクにSDの被り音が
大きく観測されたことが原因

この干渉はクラスタリングを困難にし，
延いては被り音抑圧を困難にする



- **κ による各クラスタリング手法の被り音抑圧性能**

SDR改善量：全体的な音質の良さ
SIR改善量：被り音に対する目的音の音量比
SAR：信号処理によって生じる人工的な歪みの少なさ