第5回：音響特徴量（MFCC）の原理とPython実装

音声処理と認識

担当：小野原彩香｜北星学園大学｜3年後期

この回は動画回です。Google Formsでテストを受けてください。

本回の到達目標

1. メル尺度が人間の聴覚特性を反映する理由を説明できる
2. MFCCの計算ステップを順に説明できる
3. Pythonで MFCC を計算・可視化できる
4. デルタ特徴量の意味を説明できる

周波数の差と、音の高さの尺度

同じHzの差でも、周波数比や音程は異なる。

物理的差	比率・音程としての違い
100 Hz → 200 Hz（差100 Hz）	周波数が2倍、1オクターブ
5000 Hz → 5100 Hz（差100 Hz）	周波数が1.02倍、1オクターブより小さい変化

高域の100 Hz差が聞き取れないとは限らない。
弁別できるかは、音圧・長さ・周波数・聴取者などによる。

> メル尺度は**音の高さの知覚**に基づく尺度。
> オクターブを等間隔に置く尺度や、周波数差の弁別閾、蝸牛のフィルタ幅とは区別する。

メル尺度：HTKで使われる近似式

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

逆変換（メル → Hz）：

$$f = 700\left(10^{m/2595} - 1\right)$$

周波数 (Hz)	メル (Mel)
0	0
700	≈ 781
1000	≈ 1000
7000	≈ 2702

700 Hzに折れ目はない。fが700より十分小さいと線形に近く、十分大きいと対数に近づく。
librosaの既定はSlaney方式。本回は**htk=True**で式に合わせる。

メルフィルタバンク

メル尺度で **等間隔に並んだ三角形フィルタ群** でスペクトルをまとめる

> パワースペクトル（本実習：FFT長512 → 257ビン）
> ↓
> 三角フィルタをかける（例：40個）
> ↓
> 各フィルタ内のパワーを重み付きで合計
> ↓
> 40次元のメルエネルギーベクトル

k番目のフィルタ出力：

$$S[k] = \sum_n X[n] \cdot H_k[n]$$

- 低周波：フィルタが**狭い**（細かく分析）
- 高周波：フィルタが**広い**（まとめて分析）

X[n]はパワースペクトル、H_k[n]はフィルタの重み。
本回はフィルタの面積正規化を使わず、`norm=None`とする。

ログメルスペクトログラム

対数の役割：メルパワーの広い数値範囲を圧縮する。
聴覚に着想を得た処理だが、ラウドネスそのものではない。

$$\mathrm{LogMel}[k] = \log(\max(S[k], \epsilon))$$

- S[k]：メルフィルタバンクの出力
- ε：log(0)を避ける正の下限（例：10^{-10}）
- 実習では自然対数の代わりに`power_to_db`を使用。底・基準・下限・切り詰め条件で数値は変わる

> **Whisperは対数メルスペクトログラム、wav2vec 2.0は波形**を入力する。MFCCとは区別する。

出典：power_to_db・Whisper・wav2vec 2.0

DCT（離散コサイン変換）の役割

$$C[n] = \alpha_n \sum_{k=0}^{K-1} L[k] \cos\left(\frac{\pi n(k+0.5)}{K}\right)$$

L[k]：対数メル値、K：帯域数。正規直交DCT-IIでは
$\alpha_0 = 1/\sqrt{K}$、$\alpha_n = \sqrt{(2/K)}$（$n > 0$）。コードの`norm="ortho"`に対応。

- **全係数を保てば可逆。**低次係数の選択で次元を減らす
- 相関を低減しやすいが、無相関性・統計的独立性は保証しない
- **C[0]は対数メルパワーの平均に比例。**波形の総エネルギーとは別の量

係数	表す変化
低次	メル軸上の緩やかなスペクトル包絡
高次	より細かな変化。音韻・話者・雑音などの情報を含みうる

各係数とF1・F2に一対一の対応はない。`n_mfcc=13`は**C[0]〜C[12]**の13係数。

MFCCの計算ステップ全体

入力：音声波形 $x[n]$
↓
① プリエンファシス（採用する場合）$x'[n] = x[n] - \alpha\, x[n-1]$
↓
② フレーム分割＋Hann窓：25 ms窓、10 msホップ（重複15 ms）
↓
③ 必要に応じてゼロパディング、FFT → パワースペクトル
↓
④ メルフィルタでパワーを重み付き集約
↓
⑤ 正の下限を設け、対数へ（実習では基準1のdB値）
↓
⑥ 正規直交DCT-II → 残す係数を選択
↓
出力：MFCC（例：C[0]〜C[12]）× フレーム数

プリエンファシスは必須ではない。librosaは自動では行わず、本実習も省略。

Python実装：librosaでMFCC計算

```
import librosa
import numpy as np

# 朗読音声。16 kHz、窓400点＝25 ms、ホップ160点＝10 ms
y, sr = librosa.load(librosa.example("libri1"), sr=16000, mono=True)
S_mel = librosa.feature.melspectrogram(
    y=y, sr=sr, n_fft=512, win_length=400, hop_length=160,
    n_mels=40, htk=True, norm=None)
# 同じログメル値をDCTへ。フィルタのnormとDCTのnormは別。
S_db = librosa.power_to_db(S_mel, ref=1.0)
mfcc = librosa.feature.mfcc(
    S=S_db, n_mfcc=13, dct_type=2, norm="ortho")
print(mfcc.shape)  # (13, フレーム数)、C[0]〜C[12]
```

FFT長512と窓長400は異なる。librosaの既定Slaney方式と本回のHTK設定を区別する。

`mfcc(y=...)` で同じ条件にする場合、フィルタ側は `mel_norm=None` を指定。

デルタ（Δ）特徴量

1フレームのMFCCに、周辺フレームから求めた**時間変化を明示的に追加**する方法。

$$\Delta C_t[n] = \frac{\sum_{k=1}^{K} k\,(C_{t+k}[n] - C_{t-k}[n])}{2\sum_{k=1}^{K} k^2}$$

例：K = 2なら前後2フレーム（width = 5）。上式は1フレームあたりの傾き。
秒あたりにするならホップ秒数で割る。二階微分ならその二乗で割る。

13次元のMFCC＋13次元のΔ＋13次元のΔΔ = 39次元

```
# フレーム番号に対する微分の近似。端は最近傍値で延長。
mfcc_d = librosa.feature.delta(mfcc, width=5, order=1, mode="nearest")
mfcc_dd = librosa.feature.delta(mfcc, width=5, order=2, mode="nearest")
mfcc_39 = np.concatenate([mfcc, mfcc_d, mfcc_dd], axis=0)
```

librosaはSavitzky–Golay法を使用。`order=2` は二階微分の直接推定で、一次を2回適用する方法と常に同じではない。MFCC系列を時系列モデルに渡す方式もある。

MFCCの次元数と実習の設定

n_mfcc	n_mels	設定の例
13	26〜40	従来の音声認識で使われる構成
20	40	C[0]〜C[19]を残す
40	80〜128	より多くの係数を残す。精度は実験で判断

Whisperは80または128帯域の**対数メル**、wav2vec 2.0は**波形**。80次元MFCCを使うという意味ではない。

パラメータ	本実習	意味
sr	16000 Hz	サンプリング周波数
win_length	400点	25 msの窓
n_fft	512点	FFT長。窓長とは別に設定
hop_length	160点	フレーム中心の間隔10 ms

Google Forms テスト（全5問）

締め切り：翌週末 23:59

問1：メル尺度について正しい説明はどれか
A. 全帯域を線形スケールで表したもの
B. 高音域を拡張したもの
C. 低音域を細かく、高音域を粗く扱うもの ✓
D. 基本周波数F0を検出するための尺度

問2：MFCCの正しい計算順序はどれか
A. FFT → メルフィルタ → 対数 → DCT ✓
B. DCT → FFT → メルフィルタ → 対数
C. メルフィルタ → FFT → 対数 → DCT
D. FFT → 対数 → メルフィルタ → DCT

まとめ

概念	ポイント
メル尺度	音の高さの心理尺度。弁別閾や蝸牛のフィルタ幅とは別
メルフィルタバンク	三角形フィルタをメル尺度で等間隔に配置
ログメルスペクトログラム	メルエネルギーの対数。深層学習で多用
DCT	メル帯域方向の変換。全係数なら可逆、係数の選択で次元削減
MFCC	対数メル値をDCTで変換。13など、残す係数数は設計で選択
Δ・ΔΔ特徴量	時間変化を明示的に追加。13＋13＋13なら39次元

次回（第6回・オンライン同期）：ピッチ・エネルギー・ゼロ交差などの前処理手法と音声活動検出（VAD）

Google Forms テスト

第5回の理解度テストを受けてください

- 全5問・各1点
- 締め切り：**翌週末 23:59**
- Google Formsのリンクはメールまたは授業ページで確認してください

お疲れさまでした！