第6回：ピッチ・エネルギー・ゼロ交差などの前処理

音声処理と認識

担当：小野原彩香｜北星学園大学｜3年後期

Zoom録画を後日公開します。課題はメール/Google Driveでも提出可。

本日の到達目標

1. ZCR・RMS・F0の定義と計算方法を説明できる
2. 特徴量による区間推定と、その限界を説明できる
3. 波形と3特徴量（ZCR・RMS・F0）を同時可視化できる
4. 有声性と音声活動検出（VAD）の違いを説明できる

> 特徴量は判断の手がかり。性別・年齢・言語をF0だけで確定しない。

フレーム分割と窓掛けの復習

多くの特徴量を**短いフレームごと**に計算する。窓長と窓関数は目的に応じて選ぶ。

パラメータ	例	意味
フレーム長	25 ms × 16 kHz ＝ 400サンプル	時間長＝サンプル数÷f_s
ホップ	10 ms × 16 kHz ＝ 160サンプル	25 ms窓なら60%の重なり
スペクトル分析の窓	Hann窓など	スペクトル漏れを軽減する

本日の実装：22,050 Hz、フレーム2,048（92.9 ms）、ホップ512（23.2 ms）。低いF0の周期を複数含める設定。

RMS・ZCRは波形フレームから直接計算。窓を長くすると短い変化は追いにくくなる。

ゼロ交差率（ZCR）の定義

フレーム内の符号変化の回数を、サンプル数Nで割る。

$$\mathrm{ZCR}[m] = \frac{1}{2N} \sum_{n=1}^{N-1} |s(x_m[n]) - s(x_m[n-1])|$$

s(x)＝+1（$x \geq 0$）、−1（$x < 0$）。0は正側として数える。

傾向	解釈
高い	符号の反転が頻繁。摩擦音や広帯域雑音などで見られる
低い	符号の反転が少ない。有声であることの証明ではない
完全なゼロ信号	この規約ではZCR＝0

ZCRは無次元。交差回数/秒 ≈ ZCR × f_s。サンプリング周波数や雑音条件に依存する。

```
zcr = librosa.feature.zero_crossing_rate(
    y, frame_length=2048, hop_length=512, zero_pos=True)[0]
```

RMS：振幅の大きさとエネルギー

RMS（Root Mean Square：二乗平均平方根）

$$\mathrm{RMS}[m] = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x_m[n]^2}$$

RMSは**振幅**の指標。二乗和Eとの関係は E＝N × RMS²。

基準振幅に対するdB

$$L[m] = 20\log_{10}\left(\frac{\max(\mathrm{RMS}[m], \epsilon)}{A_{\mathrm{ref}}}\right)$$

```
rms = librosa.feature.rms(y=y, frame_length=2048,
                          hop_length=512)[0]
rms_db = librosa.amplitude_to_db(rms, ref=1.0, top_db=None)
```

基準振幅A_{ref}＝1。dB SPLではない。小さいRMSにも弱い発話が含まれ得る。

ZCR + RMS による区間候補の分類

候補	RMS	ZCR
低振幅	低	どちらでも
有声らしい	閾値以上	低
無声らしい	閾値以上	高

```
def rule_based_candidates(zcr, rms_db,
                          zcr_threshold=0.1, rms_threshold=-30):
    labels = []
    for z, r in zip(zcr, rms_db):
        if r < rms_threshold:
            labels.append("low_level")
        elif z < zcr_threshold:
            labels.append("voiced_like")
        else:
            labels.append("unvoiced_like")
    return labels
```

閾値は演習用の設定例。正解ラベルと比較する。**無声・非発話・無音は別の概念**で、囁き声も発話である。

ピッチと基本周波数F0

ピッチは知覚される音の高さ。通常の有声音では**F0**が主な手がかりになる。

$$F_0 = \frac{1}{T_0}$$

T_0は秒単位の周期。例えば100 Hzなら10 ms、200 Hzなら5 ms。

	基本周波数F0	フォルマントF1・F2など
物理量	音源の繰り返しの周波数	声道の共鳴周波数
変化する要因	声帯の状態・イントネーション等	舌・唇の形、声道長等
手がかり	韻律・話者の声域など	母音・声道形状など

> 演習では50〜500 Hzを探索する。人間の声の普遍的な範囲や属性の判定基準ではない。

周期性を推定する：自己相関・YIN・pYIN

有限長の自己相関は、重なり部分の積を足す：

$$R_x[\tau] = \sum_{n=0}^{N-1-\tau} x[n]x[n+\tau], \quad 0 \le \tau < N$$

τはサンプル単位。周期候補τ_0から F0 ≈ f_s/τ_0。

- **YIN**：遅延した波形との差を二乗和にし、累積平均で正規化する
- 閾値以下の最初の局所最小を選び、放物線補間する
- `librosa.yin()` はF0候補を返す。fmaxは無声フラグではない
- **pYIN**：F0に加えて有声フラグ・有声確率を推定する

```
f0, voiced_flag, voiced_prob = librosa.pyin(
    y, fmin=50, fmax=500, sr=sr,
    frame_length=2048, hop_length=512)
f0_voiced = np.where(voiced_flag, f0, np.nan)
```

pYIN公式：NaNは無声と推定した区間。非発話の判定とは異なる。

Python実習：波形と3特徴量の複合可視化

```
import librosa, librosa.display
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec

# LibriSpeechの英語発話サンプル（最初の10秒）
y, sr = librosa.load(librosa.example("libri1"),
                     sr=22050, duration=10.0)
hop_length, frame_length = 512, 2048
zcr = librosa.feature.zero_crossing_rate(
    y, frame_length=frame_length, hop_length=hop_length)[0]
rms = librosa.feature.rms(
    y=y, frame_length=frame_length, hop_length=hop_length)[0]
rms_db = librosa.amplitude_to_db(rms, ref=1.0, top_db=None)
f0, voiced_flag, voiced_prob = librosa.pyin(
    y, fmin=50, fmax=500, sr=sr,
    frame_length=frame_length, hop_length=hop_length)
f0_voiced = np.where(voiced_flag, f0, np.nan)
times = librosa.frames_to_time(
    np.arange(len(zcr)), sr=sr, hop_length=hop_length)
```

音声活動検出（VAD）の基礎

VAD：発話区間と非発話区間を推定する。声帯振動の有無を判定する課題とは異なる。

アプローチ	特徴・例
単純な規則	エネルギー等の閾値。本日の分類は原理を試す簡略例
統計的	周波数帯の特徴をGMMで比較するWebRTC VADなど
深層学習	学習した区間判別モデル。Silero VAD、pyannoteなど

- 後段の処理・保存・送信する区間の制御に利用できる
- 非発話への不要な認識出力を減らせる場合がある
- 発話を取り逃すと文字起こしも欠落する

VADだけで、データ収集や保存が自動的に制御されるわけではない。

WebRTC VADの入力条件とAPI：16-bit PCM、対応サンプリング周波数、10/20/30 msフレームを確認。

演習：特徴量から候補と限界を考える

22,050 Hzでの模式例。dBの基準振幅は1、NaNはpYINの無声推定。

ケース	ZCR	RMS	F0
A	低、時々上昇	高低が交互	130〜180 Hz
B	高い	−50 dB以下	ほとんど未検出
C	低い区間と高い区間	−10〜−15 dB	200〜280 HzとNaN

1. 考えられる音・状況を複数挙げる
2. 候補を区別するために必要な追加情報を考える

> F0だけで性別・年齢・言語を決めない。未検出でも発話がないとは限らない。

チャットまたは共有ドキュメントで回答。

まとめ

特徴量	定義・推定	意味
ZCR	符号変化の回数÷サンプル数	符号反転の頻度
RMS	二乗平均平方根	録音信号の振幅
F0	周期の逆数。pYIN等で推定	有声区間の周期性

区間候補の分類・有声性推定・VADは目的が異なる。
モデルや閾値の出力を正解ラベルと照合する。

```
librosa.feature.zero_crossing_rate()  # ZCR
librosa.feature.rms()                 # RMS
librosa.pyin()                        # F0・有声フラグ・有声確率
librosa.amplitude_to_db()             # 基準振幅に対するdB
```

次回（第7回・動画回）：HMM・GMMと自己教師あり表現学習。

予習：「状態」「遷移確率」「出力確率（密度）」を調べる。