1V-03

少量の教師データを用いた自己教師あり音高推定

石江 悠眞† 大野 将樹‡ 獅々堀 正幹‡
徳島大学大学院 創成科学研究科† 徳島大学大学院 社会産業理工学研究部‡

1. はじめに

音高推定（pitch estimation）とは，音声や音楽の音響信号から最も低い周波数である基本周波数を推定するプロセスである．近年，教師あり学習に基づく深層学習モデル[1]が，高精度に音高推定できることが報告されている．しかし，大規模かつ高品質なラベル付き訓練データを人手で作成することは困難である．

SPICE（Self-supervised PItCh Estimation）[2]は，訓練データから教師ラベルを自動生成し，教師あり学習を行う自己教師あり学習（self-supervised learning）に基づく音高推定手法である．SPICE は 2 つの音響信号の音高差を推定する深層学習モデルを訓練することにより，間接的に音高値を推定する．推定された音高値は，周波数とは異なる中間的な値であるため，キャリブレーション処理により周波数に変換する必要がある．

本研究の目的は，少量の教師ラベルを用いて SPICE のキャリブレーション処理を最適化し，音高推定の精度を向上させることである．

2. 自己教師あり音高推定手法 SPICE

SPICE は入力部，エンコーダ，キャリブレーション部から構成される．

2.1 入力部

入力部の概要を図 1 に示す．

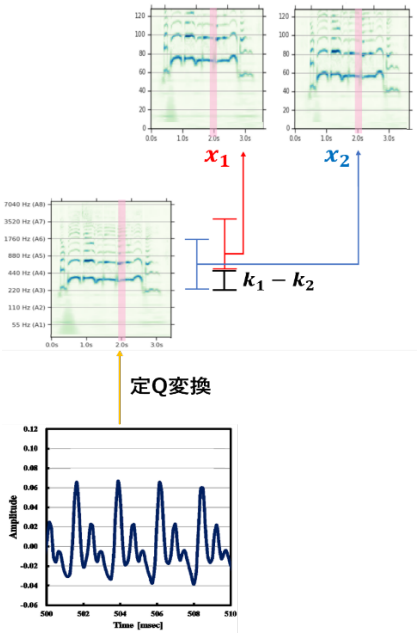

図 1：入力部の概要

入力部には，サンプリング周波数 16kHz，512 サンプルの WAV 形式ファイルを入力する．入力信号を定 Q 変換（Constant-Q Transform）した後，それぞれ異なるシフト幅k_1，k_2でピッチシフトし，2 つの 128 次元ベクトルx_1，x_2を生成する．定 Q 変換の周波数ビン数は 190 とし，k_1，k_2は$[0, 8]$の範囲から一様乱数に従って決定する．

2.2 エンコーダ

エンコーダは 6 層の畳み込みニューラルネットワークで構成されており，x_1，x_2に対する音高の推定値y_1，y_2を出力する．エンコーダの損失関数Lを式(1)に示す．

$$e = |(y_1 - y_2) - \sigma(k_1 - k_2)|$$
$$L = \frac{1}{T}\sum H(e) \quad \cdots(1)$$

σはeを$[0, 1]$の範囲に納めるための係数，Tはフレーム数であり，Hは Huber 損失を表す．式(1)より，エンコーダはピッチシフトした信号x_1，x_2の音高差$k_1 - k_2$を教師ラベル値として訓練される．同時に，推定値yは入力ベクトルxに対応する音高の推定値として学習される．

2.3 キャリブレーション

キャリブレーション部の概要を図 2 に示す．

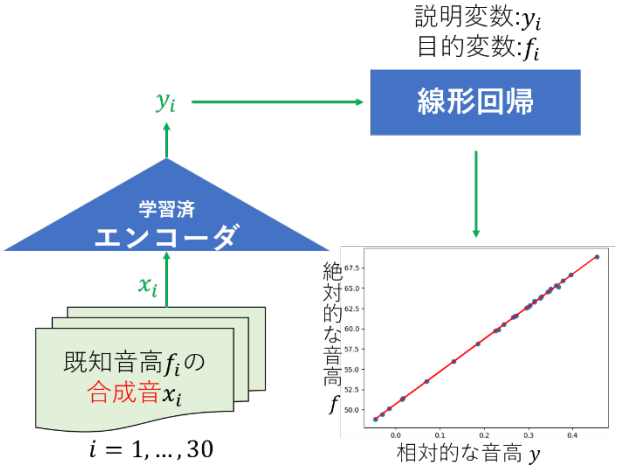

図 2：キャリブレーション部の概要

エンコーダの出力yは音高に対応する中間的な値であり，周波数に変換する必要がある．キャリブレーション部では，音高fが既知の合成音 F を

用い，式(2)に従って線形回帰することで，傾きs, 切片bを得る．

$$f = s \cdot y + b \quad \cdots(2)$$

F は音高が既知の純音 F0, F1, F2 を合成した和音であり，F1 は F0 の 2 倍音，F2 は 3 倍音である．A2 から A4 の音高から一様乱数に従って F0 を決定し，30 個の F を用いて線形回帰モデルを訓練した．

しかし，F は純音から合成した 3 和音であり，音声や楽器音と調波構造が異なるため，最適なキャリブレーションが実現できないという問題がある．

予備実験により，10 回のキャリブレーションを行った場合，音高推定精度は最高 89.3%，最低 86.3%とばらつきがあることがわかった（差分 2.96）．

3. 提案手法

提案手法の概要を図 3 に示す．

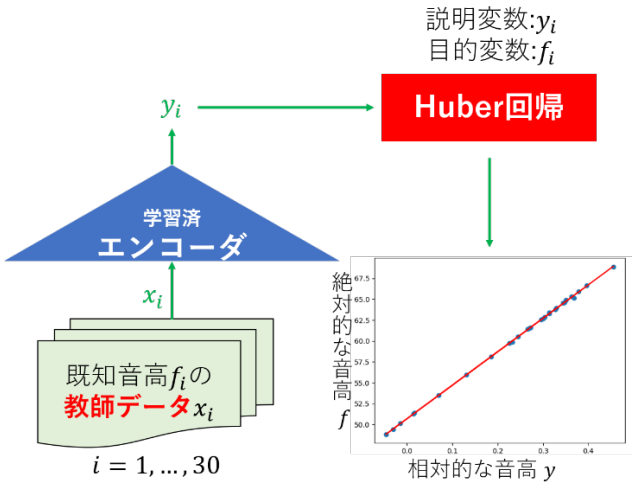

図 3：提案手法の概要

本研究では，少量のラベル付きデータを用いてキャリブレーションを最適化する手法を提案する．提案手法では合成音は使用せずに，あらかじめ人手で音高を付与した少量の教師データのみを用いてキャリブレーションをおこなう．

しかし，人手で正確な音高を推定することは困難であるため，誤った正解ラベルが付与されることがある．ラベル付きデータは少量であることが望ましいが，外れ値の影響が大きくなることを考慮する必要がある．そこで提案手法では，Huber 回帰を用いて外れ値の影響を軽減する．

4. 評価実験

実験データセットとして，MIR-1K の 1000 曲を用いた．MIR-1K とは歌唱に対する音高があらかじめ付与されているデータセットである．900 曲を訓練データ，100 曲をテストデータとした．提案手法では，訓練データの 30 サンプルをキャリブレーションに用い，残りのサンプルでエンコーダの訓練を行った．式(2)により得た推定音高fと真の音高の差が，半々音以内であった場合，推定が正解したとみなした．表 1 に，SPICE を用いて 10 回音高推定した際の平均精度を示す．

	平均（%）	標準偏差
従来手法	88.32	0.96
提案手法	89.70	0.18

表 1：実験結果

提案手法（少量のラベル付きデータ＋Huber 回帰）の精度は，従来手法（合成音＋線形回帰）よりも 1.38 ポイント向上した．加えて，提案手法の精度の最高値は 89.9%，最低値は 89.2%であり，ばらつきを抑制できることがわかった（差分 0.67）．

5. おわりに

SPICE は，キャリブレーションにおいて合成音を用いるため，音声や楽器音に対して最適なキャリブレーションを実現できないという問題があった．本研究では合成音に代わって，少量の教師データを用いた SPICE のキャリブレーションを提案した．実験の結果，推定精度が向上し，ばらつきが改善することが確認された．

提案手法では，少量ではあるが作成コストが高い正解ラベル付きデータを用いるため，より少量のデータ数でより正確なキャリブレーションができることが望ましい．今後は，ラベル付きデータ数とキャリブレーション精度の相関性を調査し，ラベル付与のコストを低下させることが考えられる．

参考文献

[1] J.W. Kim, J. Salamon, P. Li, and J.P. Bello, “Crepe: A convolutional representation for pitch estimation,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IEEE, pp.161–165 2018.

[2] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirovi´c, ”Spice: Self-supervised pitch estimation,” IEEE/ACM Transactions on Audio,Speech, and Language Processing, vol.28, pp.1118–1128, 2020.