1Q-05

リアルタイム発音検出のための動的しきい値自動最適化

松尾　章弘†　　土江田　織枝†　　山田　昌尚†

釧路工業高等専門学校†

1.はじめに

本研究はリアルタイム発音検出での動的しきい値のパラメータを自動的に最適化する方法を検討するものである．現在，我々はリズム感の向上を目的として，正確に自分の演奏したリズムを確認できるように発音検出を用いたリズム練習支援システムを開発中である[1]．発音検出に関しては一定の精度が得られたが，これまでは発音検出のための関数に用いるしきい値のパラメータを手動で設定してきた．そこで，システムの構成要素の 1 つとして，しきい値パラメータを自動で最適化する方法について検討する．

2.発音検出としきい値

音響信号から発音のタイミングを得ることを発音検出という．発音検出は一般的に，前処理，検出関数，ピーク抽出の 3 段階からなる[2]．第 1 段階の前処理としては正規化を施す．第 2 段階の検出関数として，周波数ごとの信号スペクトル強度を利用するスペクトルフラックスや HFC (High Frequency Content)，位相の変化を利用する方法など，各種が提案されている．本研究では，周波数ごとの信号スペクトル強度の変化が大きい場合に発音となるピークが現れるスペクトルフラックスを使用する．スペクトルフラックスは次式で表される．

$$SF(n) = \sum_{k=1}^{\frac{N}{2}-1} \max(0, |X(n,k)| - |X(n-1,k)|)$$

ここでNは FFT フレームのデータ数，nは時間，kは周波数であり，$X(n,k)$はスペクトログラム (時間周波数信号) を表す．0 とのmaxをとることでスペクトル強度が増加する場合のみを対象としている．第 3 段階のピーク抽出では，しきい値を超える局所最大値 (local maxima)を 検出し，その時刻を発音時刻とする．しきい値として次式による動的しきい値を用いる．

$$TH(n) = \delta + \lambda \cdot \mathrm{median}(SF(n-v_1 : n+v_2)) + \alpha \cdot \mathrm{mean}(SF(n-v_1 : n+v_2))$$

ここでδはしきい値の定数項，λおよびαはそれぞれ中央値，平均値に対する重みであり，v_1, v_2は動的しきい値の対象幅を表す．今回はこのδ，λ，αの最適値を求める．このしきい値を用いて，検出関数$DF(n)$および発音時刻$OD(n)$は次のように求められる．

$$DF(n) = SF(n) - TH(n)$$

$$OD(n) = \begin{cases} 1, DF(n) > 0 \ and \ \underset{w_1<m<w_2}{\mathrm{argmax}} \ DF(m) = n \\ 0, otherwise \end{cases}$$

3.システム構成

図 1 にシステム構成を示す．メトロノーム音は ChucK で生成し，スピーカから出力する．強拍と弱拍は周波数で区別する．強拍は 880Hz，弱拍は 440Hz とした．システムのユーザはこのメトロノーム音を聴きながら楽器を演奏する．Chuck から Processing へメトロノーム音と発音検出結果のタイミングを送るために OSC (Open Sound Control) を用いる．OSC は電子楽器およびコンピュータ間で音楽演奏データ等を送受信するための通信プロトコルである．OSC では URL 形式でデータを送るため，メトロノームのタイミングとオンセットを容易に区別して情報伝達をできる．

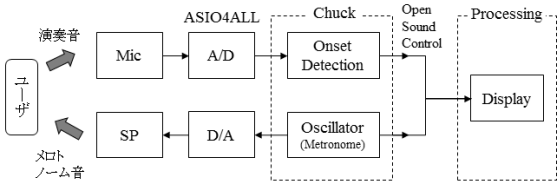

図 1 リズム練習支援システム構成図

4.検出精度の評価

発音検出の精度の評価として F 値 (F-measure) を用いる．その F 値を求めるのに必要な発音検出の結果の状態を分けるために，実際に発音されている場合で発音だと判断したものを True Positive(TP)，発音だと判断できなかったものを False Negative (FN)，実際に発音されていない場合で発音だと判断しなかったものを True Negative (TN)，発音だと判断してしまったものを False Positive (FP)とする．

次にこの 4 つの状態から，発音と判断したもののうち正しい時刻で発音と判断できた割合を示す再現率Pと正解の発音をどれだけ発音だと判断できたかの割合を示す適合率Rを求めることができる．

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \ , \ \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

そしてこの 2 つの調和平均である F 値をこのシステムの検出精度の指標とする．F 値は以下のように求められる．

Automatic Optimization of Dynamic Thresholding Parameters for Real-time Onset Detection

† National Institute of Technology, Kushiro College

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

本システムは FP，FN が多くなるとユーザへ正しい情報のフィードバックを行うことができなくなってしまう．よって，その 2 つを同時に評価することのできる F 値を精度の指標として扱う．

5.しきい値自動最適化

今回は最急降下法[3]によりしきい値の最適化を検討した．最急降下法によるパラメータの更新式は次式で表される．

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \alpha \frac{\partial F(\mathbf{X})}{\partial \mathbf{X}}$$

ここで$\mathbf{X}_t = (\delta, \lambda, \alpha)$とし，αは学習率，Fは最大化する目的関数である F 値である．終了条件は $F(\mathbf{X}_{t+1}) - F(\mathbf{X}_t) > 0.001$ とした．発音検出に用いたデータは，サンプリング周波数 44.1kHz で保存された 48 個の発音があるホルンの演奏の音響データである．正解の発音時刻は 3 人でアノテーションし，その平均時刻を各発音の発音時刻とした．$v_1 = 50ms$，$v_2 = 0$ として，このデータに対して最急降下法によるパラメータの最適化を行った．今回はしきい値のパラメータの自動最適化の手法を検討するため，計算時間を短縮する目的でフリーの数値解析ソフトである Scilab を用いて実験を行った．

λ, αをそれぞれ 0～1.5 の範囲で 0.15 刻みで 11 個，δを 0～0.2 の範囲で 0.02 刻みで 11 個用意し初期値を変えて 1331 個の組み合わせで最急降下法によるパラメータの最適化を行った．F 値が 0.9 を越えるものが 299 個となった．0.8 を下回るものはそれぞれ 100 個以下であり，最急降下法で，F値の高くなるパラメータの組み合わせを得ることができた．

次に最急降下法で初期値を変えて最適化を行った組み合わせの例を例 1～4 として表 1 に示す．その結果を縦軸を F 値，横軸を繰り返し回数として図 3 に示す．初期値によって，収束する値が異なることが図 3 からわかる．これらは F 値の極大値が複数あるため起こると考えられる．例として，δを固定したときのλ, αの変化に伴う F 値の変化を 3 次元グラフとして図 4 に示す．図 4 は極大となる山がいくつかあることがわかり，このような山があると最適値を求めることが難しくなってしまう．また，F 値が最も高くなる値が複数のパラメータの組み合わせで存在することが確認でき，最適値は 1 つではないことを示している．

6.まとめ

現在開発中のリズム練習支援システムの構成要素の 1 つとして今まで手動で行ってきたしきい値の設定を自動化する方法について検討した．結果は，F値の高いパラメータの組み合わせを求めることができた．最適値は複数あることと，最急降下法により最適値を求める場合，初期値の設定が結果に影響することがわかった．今後の課題としては，今回はホルンのみで発音検出を行ったので，他の種類の楽器についても同様の結果が得られるかの確認も必要である．また，システムに実装する場合，最適値以外のところで収束したものを選択しないように複数の初期値の組み合わせを同時に試し，F 値が最も高くなるように組み合わせを選ぶなどの改善が必要である．

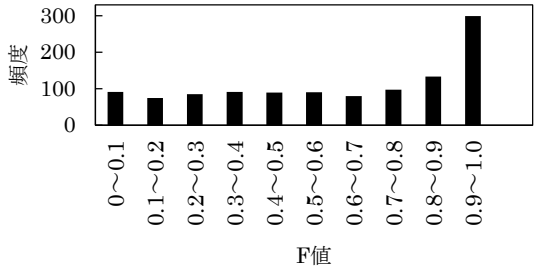

図 2　最急降下法による F 値のヒストグラム

表 1　最急降下法での初期値の組み合わせ

	α	λ	δ
例 1	0.2306	0.5452	0.4151
例 2	0.2306	0.2370	0.0607
例 3	0.2306	0.1104	0.4151
例 4	0.0519	0.1104	0.0607

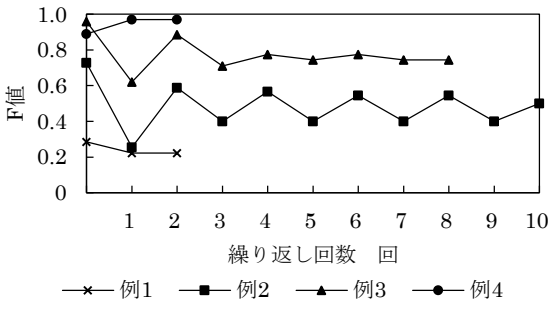

図 3　最急降下法での F 値の変化

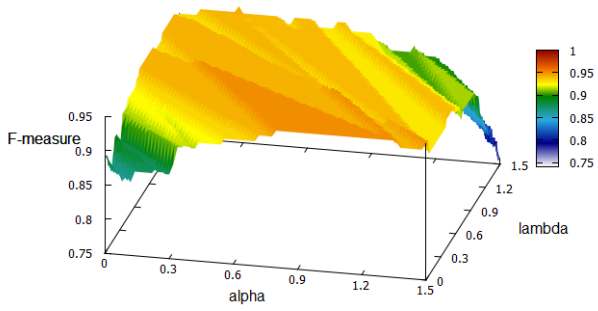

図 4　F 値の変化（$\delta = 0.050$）

参考文献

[1] 松尾章弘，原佑輔，土江田織枝，山田昌尚，“リズム練習支援のためのリアルタイム発音検出手法の検討”，第 14 回情報科学技術フォーラム(FIT)論文集，第 2 分冊，pp. 205-206(2015)

[2] Bello, Juan Pablo, et al., “A tutorial on onset detection in music signals”, Speech and Audio Processing, IEEE Transactions, Vol.13, No.5, pp.1035-1047 (2005)

[3] 北村充，“数理計画法による最適化　実際の問題に生かすための考え方と手法”，森北出版株式会社，p. 72(2015)