Vol.2016-MUS-111 No.55
2016/5/22

リズム練習システムの発音検出精度向上に関する検討

原佑輔† 松尾章弘† 山田昌尚†

概要：近年，楽器の演奏データから音高推定や発音検出を行う研究が多くある．特に発音検出はテンポ推定や自動採譜に用いられるだけでなく，検出結果をグラフィカルに表示することでリズム練習の支援にも用いることができる．本発表では，我々の研究室で開発を行っているリズム練習支援システムの発音検出精度向上を図るため，検出処理におけるしきい値のパラメータ設定についての検討と，発音検出の前処理としての楽器音とノイズの識別を行う．

キーワード：発音検出，ケプストラム，MFCC，SVM，

1. はじめに

近年，楽器の演奏データから音高推定や発音検出を行う研究が多くある．特に発音検出については，テンポ推定や自動採譜などに応用されるほか，検出結果をグラフィカルに表示することでリズム練習の支援にも用いることができる．そこで我々の研究室ではリズム練習支援のためのシステムの開発を行っている[1].しかし,楽器の演奏を録音する際に様々な雑音が混入し，誤検出されることがある．この誤検出を防ぎ，発音検出精度の向上を図るために，検出処理におけるしきい値のパラメータ設定について検討と，発音検出の前処理として楽器音とノイズの識別を行う．

2. 発音検出について

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．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．今回は，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．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．

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

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

我々の研究室ではこのしきい値の重みであるδ，λ，αの最適値を，最急降下法を用いて自動で値を求めたが，初期値によって収束する値は大きく異なったため，最適な値を得るために引き続き検討が必要である．

3. 楽器音とノイズの識別

発音検出の前処理として楽器音とノイズの識別を行う．識別を行う際に，音源をそのまま識別に用いるのは冗長であるため，識別に有用な特徴を抽出した特徴量に変換する必要がある．特徴量についてはケプストラム，メル周波数ケプストラム係数を比較する.識別には SVM(support vector machine)を用いる．

3.1 ケプストラム

楽器音は楽器の特徴である音色情報と，音の高さを表す音高情報の畳み込みで表される．入力を音色情報h，音高情報gの畳み込みである$h * g$とする．そのスペクトルは$F(h) * F(g)$で表される．このスペクトルの対数をとることで(1)式のように入力を和で表すことができる．

$$\log(h * g) = \log(F(h)) + \log(F(g)) \quad (1)$$

ケプストラムは(1)式を時間信号と見て，さらにフーリエ変換をした結果である．

MFCC はメル尺度を考慮したケプストラムの係数である．メル尺度とは人間の聴覚の特性を考慮した音高の尺度である．この尺度は人間がなんらかの周波数の変化を聞いたと

† 釧路工業高等専門学校
National Institute of Technology, Kushiro College

き，その周波数が低いと変化が大きく感じ，周波数が高いと変化は小さく感じることに対応している．
MFCC には 20 次元のメルフィルタバンクを用いて分類のための特徴量とした．また MFCC と同様に実験を行うため，ケプストラムについても周波数軸に等間隔なフィルタバンクを用いて 20 次元に圧縮し，特徴量とした．以降はこの 20 次元に圧縮したケプストラムをケプストラムと表記する．

3.2 SVM(support vector machine)

SVM とは教師あり学習を用いるパターン認識モデルの一つで，主に二値分類に用いられる．特徴として，高い汎化性能があげられる．2 クラスの学習データが与えられたとき，データ点との距離が最大となるような分離面を求める．図(1)の場合，分離面は中心の点線で，矢印で示されるマージンが最大になるようにパラメータを学習し，分離面を求めていく．

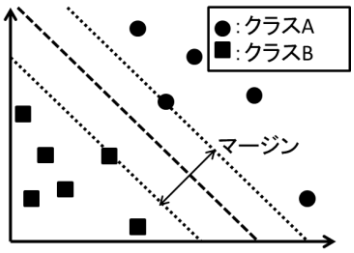

図 1　SVM による 2 クラス 2 次元分類の例

3.3 実験

ケプストラム，MFCC それぞれを用いて，正しく分類できるかどうか実験を行った．サンプルとして，トランペットを用い，楽器音 80 個とノイズとして息を吹き込む音，スライドを動かした音，バルブを押した音を合計 80 個用意した．録音環境は無響室で測定用のマイクを用いた．まず始めに発音のタイミングの入力信号をケプストラムおよび MFCC に変換し，データセットを作成した．分類にはフリーウェアの Weka を用いた．データを標準化し，交差検証を行った．結果はケプストラム，MFCC ともにすべてのデータが正しく楽器音とノイズに分類された．

次に，データセットのケプストラム，MFCC それぞれの高次元のデータを減らしていき，正しく分類するにはどこまでの次元が必要かを調べた．結果，ケプストラムでは 2 次元あれば正しく分類できるのに対し，MFCC では 3 次元まであれば正しく分類できることがわかった．

3.4 考察

ケプストラム，MFCC どちらを用いてもすべてのデータが正しく分類することができた．また分類には低次元のデータが有効であることがわかった．図 2，図 3 に楽器音のケプストラム，MFCC の例を示す．ケプストラムでは周波数軸に対し，均等にデータが得られているのに対し，MFCC では低周波側で密にデータを取っている．そのため MFCC の低次元側はひとつの次元が表す周波数範囲が狭いので，2 つ目の実験で次元を減らしたときにケプストラムに比べ，分類に必要な次元の数が多くなったと考えられる．また MFCC の 3 次元までのデータがあれば分類できることから，1[kHz]程度の周波数範囲があれば分類ができることがわかった．

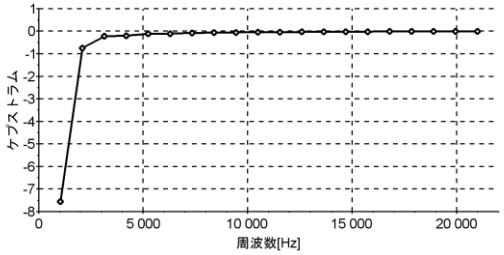

図 2　楽器音のケプストラムの例

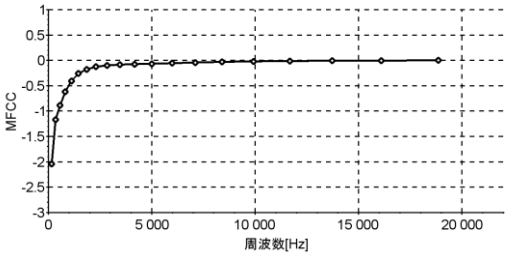

図 3　楽器音の MFCC の例

4. まとめ

リズム練習支援システムの発音検出精度の向上を図るために，楽器音とノイズの識別を行った．ケプストラム，MFCC ともに分類することができた．ただしサンプル数が少なく，また録音を無響室で行ったため，実際にシステムを使うことを想定した録音，実験が必要である．

検出処理のパラメータの調整については現在検討中である．

参考文献

[1] 松尾章弘，土江田織枝，山田昌尚，“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,”情報処理学会第 78 回全国大会，第 2 分冊，pp. 437-438, 2016.