JPHACKS における音声特徴量抽出技術の開発

Application Bug Creater 2021

2021 年 10 月 29 日

1　はじめに

時は昨年に遡る．筆者は，素晴らしい（クレイジーな）メンバーと共に JPHACKS で開発できて満足していた．そして，来年は研究が忙しくなり，ハッカソンに全力で打ち込むことはおろか，参加すら怪しくなると思っていた．

ところが実態はどうであろうか．今年の制作物は，似ている音声ファイルの検索となっ（てしまっ）た．もちろんライブラリなどなく，似たサービス（曲を除く）もなく，論文でようやく見るぐらいだ．

今年も楽しい研究ハッカソンの始まりだ．

2　導入

2.1　音の理論

そもそも音とは何であろうか．音とは，空気の疎密波である．よく音は波とよばれるが，空気の密度の疎な部分と密な部分が移動していくことで音というものが伝わっていく．それでは，人はどのようにその波を音として認識するのか．それは，耳の奥にある鼓膜が，空気の疎密波によって振動し，蝸牛でその振動を聴覚神経に電気信号に変換することで，人は音を認識する．蝸牛は，音の周波数によってそれぞれ共振する場所がある．その共振を聴覚神経が感じることで音というものを聴くことができる．

つまり，音は周波数によって聴いているのである．では，音色とは何か．また，どのように分析すればよいのだろうか．

2.2　先行研究

筆者は信号処理の知識はあるが，現在の音の特徴量抽出についての知識は無かったため，既存技術の調査を行った．以下はその代表的なものである．

2.2.1 パワースペクトル密度

先ほど，音は周波数によって聴いていると書いた．では，音を周波数ごとに分ければ音色の本質に近づくであろうことは自然だ．そのために，以下のような変換を用いる．

1. フーリエ変換 Fourier transform (FT)
2. 離散フーリエ変換 discrete Fourier transform (DFT)

コンピューター上のオーディオである以上，DFT を用いることになる．DFT を用い，音を周波数ごとに分けると以下の図 1 のようになる．（DFT の結果の絶対値の二乗をパワースペクトル密度という.）なお，DFT の詳細については後述する．

2.2.2 フォルマント

図 1 を見てもらえればわかるが，データとして多く，特徴検索に向かない．そこで，図 2 のような包絡線を描き，ピークとなるところ（青色で囲ったところ）で左側にある順から第一フォルマント・第二フォルマント…と呼ぶ．[1] こうすることで，音色というものを大きく圧縮することができる．

この特徴量は音声認識においてよく用いられ，フォルマントを調整することによって男性っぽい声・女性っぽい声と変えることができる．

2.2.3 離散全極型モデル

フォルマントは，非常に音を圧縮した表現となり，音声認識では問題ないが，音色としては逆に特徴量を捨てすぎているという問題がある．そこで，離散全極型モデル [2] というものがある．これは，パワースペクトル密度とフォルマントの中間的な表現であり，以下の図 3 で示すように，音の周波数ごとのピーク値と音量を特徴量としたものである．この方式であれば，充分音色を表現でき，かつ，よく圧縮できている．

2.2.4 相互相関

ふたつの信号がどれだけ似ているかというものに，以下の式で表されるような相互相関関数がよく用いられる．

$$(f * g)(m) = \sum_{n} f(n)g(m-n)$$

この関数の意味としては，時間 m だけシフトした際にどれだけ信号が似ているかというものである．そのため，この関数の最大値が大きいほど似た信号と判断できる．

しかし，相互相関は音の高さが異なると相関は低いと判定してしまうため，今回の音色を探すという目的には合わない．

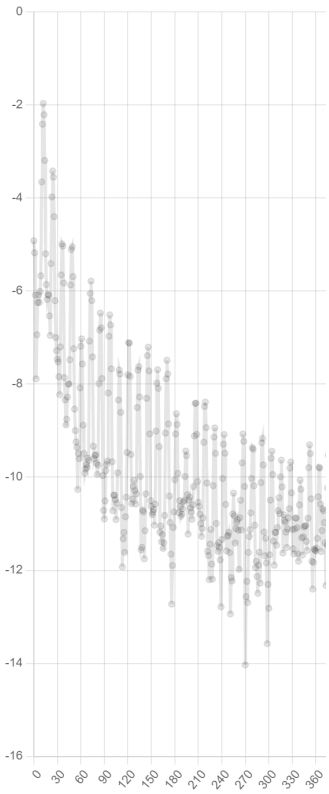

図 1　パワースペクトル密度

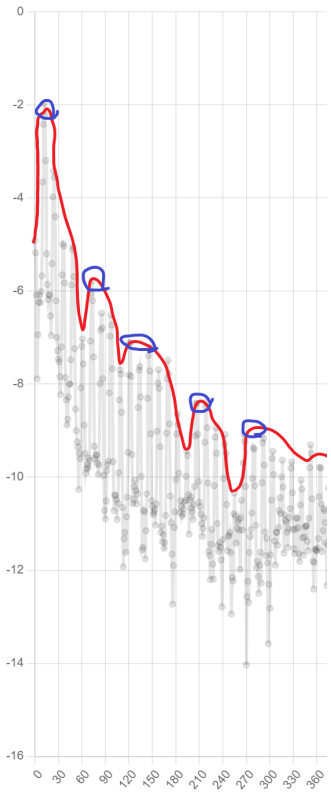

図 2　包絡線

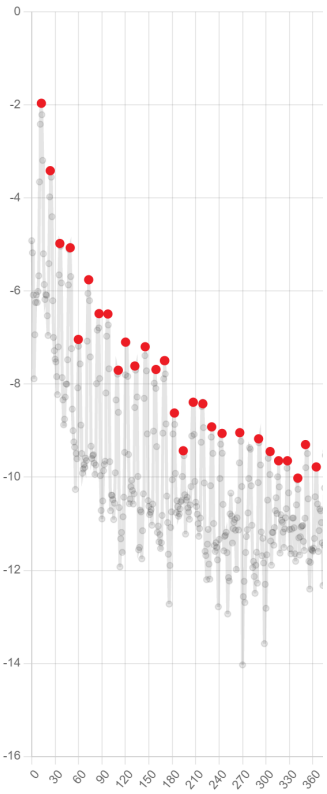

図 3　離散全極型モデル

3　特徴量抽出方針

先行研究の離散全極型モデルを元に，更に発展させる方針とする．

3.1　用いる公式

以下では，重要な式を紹介する．

3.1.1　DFT

離散データを周波数領域に変換する離散フーリエ変換．

$$F(t) = \sum_{x=0}^{N-1} f(x) exp(-j\frac{2\pi tx}{N})$$

3.1.2　窓関数

無限時間の DFT であれば目的のパワースペクトル密度を得られるが，現実には有限時間で区切る必要がある．そのため，目的以外の周波数であるサイドロープが付随してしまう．そこで，窓関

数を用いることで，サイドローブを減少させる．窓関数を使うことでサイドローブを減少するが，周波数分解能が下がるが，これは提案手法によって克服される．なお，今回用いたのはブラックマン - ナットール窓で，サイドローブの減少幅が大きい．

$$w(x) = 0.3635819 - 0.4891775\cos(2\pi x) + 0.1365995\cos(4\pi x) - 0.0106411\cos(6\pi x)$$

3.1.3 FFT

DFT について紹介したが，計算量が $O(n^2)$ と大きい．しかし，高速フーリエ変換によって計算量を $O(n\log n)$ に抑えることができる．

3.1.4 STFT

現実には，同じ音がなり続けるということはなく，変わっていくものである．そこで，短い期間で時間をずらしながら FFT することによって，音の変化も観測できる．これを短時間フーリエ変換と呼ぶ．

4 提案手法

我々が提案する手法は，短期間と長期間の二種類の STFT をして，離散全極型モデルの特徴量を抽出する．そして，短期間の物を中心にし，特徴量の補正をし，最も音色であると考えられる特徴量を選択する．この得られた特徴量を，基本周波数・電力で正規化したものを特徴量とする．これはベクトル表現でき，特徴量ベクトル間の距離を音色の近さの指標として提案する．

4.1 特徴量の補正

4.2 隣接する短期間 STFT

短期間の窓は 0.1 秒ほどであるため，隣接する STFT の要素は，ほぼ同じ音が鳴っていると考えられる．その性質を利用し，ノイズに埋もれたピークを取り出すことが可能である．具体的には，音量が小さい方のものに大きい方の特徴量をスケールし，ノイズと判定する閾値を下回れば，ノイズに埋もれたピークと判断し追加することで取り出すことができる．

4.2.1 短期間と長期間 STFT の組み合わせ

短期間 STFT は短い期間を分析でき，長期間 STFT は周波数分解能が高いといった特徴を持つ．まず，短期間 STFT で音色の特徴量の候補を抽出する．そして，その短期間 STFT が所属する長期間 STFT の特徴量で補正することができる．このように，ふたつを組み合わせることで双方の利点を利用できる．

4.2.2 楽器の性質

いわゆる物理的な楽器というものは，基本周波数というものがあり，それに加え，n 倍音成分を含むというものが多い．これは物理的な形状に起因するものである．（太鼓等の打楽器は球面調和関数といってまた別の成分も多いが…）そのため，特徴量を倍音ごとにグループ化し，そのグループの二乗誤差が最小となるよう基本周波数を変更する．この操作により，特に，相対的に荒い低周波数領域の補正が効果的に行われる．

4.3 特徴量の選択

ここまで上げた特徴量の候補を，同じ性質ものをグループとする．そして，そのグループの電力が最も大きいものをこの音の音色の特徴量として選択する．

4.3.1 考察

5 おわりに

結果として我々は, 音色を効果的に特徴量として表現し，近さを表現することができた.

6 謝辞

たった一週間で初めて参入する音の分野を学び利用する技術を作ろうという無謀にも見える挑戦を見事にやり遂げた, 筆者の無茶につきあってもらったメンバーに敬意を表したい.

参考文献

[1] 音声の音響分析の「いろは」 https://www.gavo.t.u-tokyo.ac.jp/ mine/japanese/nlp+slp/I-RO-HA.pdf

[2] 音楽のパーツ表現 http://sap.ist.i.kyoto-u.ac.jp/members/yoshii/slides/sigmus-2016-2-yoshii-slides.pdf