Vol.2021-MUS-130 No.30
Vol.2021-EC-59 No.30
2021/3/17

歌唱テクニックの識別における hand-cratfted 特徴量と深層学習抽出特徴量の比較

山本 雄也[1,a)]　Juhan Nam[2,b)]　寺澤 洋子[1,c)]　平賀 譲[1,d)]

概要：本研究は歌唱テクニックの識別において，専門知識に基づき設計された特徴量（hand-crafted 特徴量）と深層学習によって自律的に獲得した特徴量の識別性能を比較するものである．歌唱テクニックは歌手が歌唱中に音高・音色・音量を変動させることにより表現する技法である．歌唱テクニックの様相は様々であり，中には特性が未解明なものや非自明なものも含まれているため，その特徴をとらえるのは難しい．本研究では深層学習による自律的な特徴獲得によって，歌唱テクニックの明示的モデリングを回避する方法について検討する．検証事項として特徴抽出の良し悪しのみを考えるため，分類器の条件を一定にする．従来の音声分類問題に用いられた hand-crafted 特徴量と深層学習により抽出した特徴量を用いて同分類器を学習させ比較する．10 種類の歌唱テクニック分類実験の結果，深層学習による特徴抽出では 73.6%の正解率が得られた．この数値は hand-crafted 特徴量での結果を 2.6%上回っており，明示的モデリングなしでも hand-crafted 特徴量を用いた場合と同等の性能が得られることを確認した．特に極端な歌唱テクニックにおいて hand-crafted 特徴量を用いた場合より正解率が高く，深層学習による特徴の自動獲得が有用である可能性を示した．

A Comparison of Hand-crafted Feature and Deep-extracted Feature on Singing Technique Classification

1. はじめに

楽曲を歌唱する歌手は，しばしば歌声の音高・音量・音色・タイミングを細かく変動させて表情豊かな歌唱を行う．研究ではそれらを「歌唱テクニック」と定義し，歌声に含まれる歌唱テクニックを識別することを目的とする．歌唱テクニックにはピッチを変動させるビブラートや，呼気を混ぜるブレシー，ざらついたフライの声を出すボーカルフライ等がある．歌唱テクニックにはこのように様々な性質を有しているものがあるが，現在定量的な分析を通して性質が明らかになっている歌唱テクニックの音響的特性は一部である．そもそも音響特徴を解明されていない歌唱テクニックもある．また，研究事例のある歌唱テクニックであっても未だに完全なモデル化には至っていない．歌声の構造は複雑であり，歌唱テクニックの定義との繋がりも非自明であるため明示的なモデリングが難しい．作成できたとしても，データに沿ったものとなっている保証はなく，識別精度が頭打ちになる場合がある．加えて，特徴量の設計自体，歌唱テクニックに関して深く観察を行う必要があり，容易ではない．

一方深層学習モデルにより特徴抽出を行う方法は，音響的特性を明示的に強く仮定することなく，データから特徴を自動獲得することが可能という利点がある．歌唱テクニックのような，非自明な音響的特性を持ち，研究者によって特徴に関する仮定をおくのが難しい対象においては，特に性能を発揮することが期待できる．

このような背景から，本研究では歌声に関する hand-crafted 特徴量と深層学習モデルによって自動的に抽出された特徴量を同一の分類器に入力し学習させ，その識別性能を比較し考察する．

[1] 筑波大学
1-2 Kasuga, Tsukuba-shi, Ibaraki 305-8550 Japan
[2] KAIST
291 Daehak-ro, Eoeun-dong, Yuseong-gu, Daejeon, 34141 South-Korea
[a)] s1921652@s.tsukuba.ac.jp
[b)] juhan.nam@kaist.ac.kr
[c)] terasawa@slis.tsukuba.ac.jp
[d)] hiraga@slis.tsukuba.ac.jp

Vol.2021-MUS-130 No.30
Vol.2021-EC-59 No.30
2021/3/17

2. 関連研究

2.1 歌唱テクニックの研究

従来の歌唱テクニック識別の研究にはビブラート [10], [13], [14]，ポルタメント [15]，声区の識別 [16]，声質の識別 [17] などの事例が存在する．各々，識別対象の歌唱テクニックに対してその音響特性を仮定するアプローチが中心であった．総合的に歌唱テクニックを扱う研究には，池宮らの歌唱スタイル転写 [18] がある．これはビブラートと 3 種のポルタメントの基本周波数（f0）の軌跡をいくつかのパラメータで表現し，歌唱スタイルの要素として抽出し他の歌手の歌声に転写する手法である．

2.2 深層学習による特徴表現について

本研究に類似する研究として従来，基本周波数の軌跡を分類する問題において，1 次元畳み込みニューラルネットワーク（1DCNN）を用いた特徴抽出器が hand-crafted 特徴量と同等の性能が得られることを明らかにしたものがあった [2]・また，環境音識別で用いられた深層学習モデル（VGGish）による特徴抽出や，スペクトログラムを画像とみなし 2 次元フーリエ変換を処理した特徴を歌唱テクニック識別に利用した研究も存在する [19]．しかし，ピッチだけでなく様々な要素が変動する歌唱テクニックにおいて深層学習を用いてデータから学習した特徴と，従来数多く音楽の識別に用いられている hand-crafted 特徴量の比較はこれまで行われておらず，歌唱テクニックの識別に有用かどうかは未だに明らかになっていない．

3. 問題設定と用いるデータセット

本研究の問題設定は，時間サンプル数 T_{sample} の 1 次元歌声波形 $X \in R^{T_{sample}}$ から次元数 M の中間特徴量 $Z \in R^M$ を得た上で，K 種類中の歌唱テクニックから最も出力値の高い $Y = \arg max_k f(Z)$ を識別結果とする．$f(Z)$ は中間特徴量 Z を入力すると各歌唱テクニックの K 次元確率ベクトル $Y_{act} \in [0,1]^K$ を出力する分類器である．

今回，データセットとして VocalSet[3] を用いた．VocalSet は異なる 10 種類の歌唱テクニックをアルペジオ・スケール等様々なコンテキストで，5 つの母音それぞれで歌われた歌声を収録した大規模なデータセットである．VocalSet に含まれる歌唱テクニックの詳細を表 1 に示す．なお，VocalSet は 1 ファイルにつき 1 つの歌唱テクニックのみが含まれているものとしてラベリングがされている．

表 1 VocalSet のラベル

ラベル名	歌唱テクニック名	概要	総時間 [min]
straight	通常発声	通常に発声する	71.65
belt	ベルティング歌唱	力強く発声する	26.24
breathy	呼気を伴う歌唱	歌声に呼気を混ぜる	28.00
vocalfry	ボーカルフライ	歌声にフライを混ぜる	34.10
vibrato	ビブラート	音高を揺らす	57.79
trill	トリル	2 音高間を行き来する	18.45
trillo	トリッロ	音量を揺らす	14.54
inhaled	吸気発声による歌唱	息を吸いながら発声する	9.95
liptrill	リップトリル	唇を震わせながら発声する	24.40
spoken	話声	話すように歌う	4.06

4. 音響特徴量の抽出について

4.1 Hand-crafted 特徴量

本研究では音色と音高変動に関する 2 種類の hand-crafted 特徴量を用いる．音色の特徴量にはメル周波数ケプストラム係数（MFCC）を用いた．MFCC は音声認識，音楽分類などにおいて従来多くの分類問題において，音響特徴量として用いられている．歌声の識別においても用いられた事例はいくつか存在する [17]．音高変動の特徴量には Vibrato extend，Vibrato rate を用いた．それぞれビブラートのピッチの変動の振幅，変動の周波数に該当する．これら 2 種類の特徴量は歌手識別の先行研究 [1] でも用いられた．

4.2 深層学習モデルによる特徴抽出

次に深層学習モデルを用い，データから特徴量を自動獲得する方法を述べる．図 1 に示すような 2 次元畳み込みニューラルネットワーク（以降，CNN）を用いて特徴量の抽出を行う．このモデルは入力層でメルスペクトログラムに変形する．その後に畳み込み層，バッチ正則化層，活性化層，プーリング層，ドロップアウト層で構成される畳み込みブロックが 4 つ続く. 全ての活性化層の活性化関数には ReLU，全てのプーリング層には Max Pooling を用いた.4 つの畳み込 みブロックの後には全結合層（Dense）が続く．全結合層によって出力される特徴量の次元数は M となるようにする．このモデルは学習時のみ，さらにもう一つの全結合層に入力することによって識別クラス数に等しい次元数を持つベクトルを出力で得ることができる．このベクトルに Softmax 関数を適用することによって種類方向の正規化処理を行うと歌唱テクニックの出力確率ベクトルを得ることができる. この確率ベクトルと正解ラベル間とのクロスエントロピー損失を最小化する．後述するが，評価実験においてデータセットは教師データとテストデータに分割する. その分割した教師データのみを用いて学習を行う．

5. 評価実験

本章では hand-crafted 特徴量と深層学習モデルによる特徴量を 10 クラス識別問題により性能を比較した実験ついて報告する．本章では簡単のために hand-crafted 特徴量を用いる手法と深層学習モデルによって特徴量を自動抽出した方法をそれぞれ hand-crafted 手法，CNN-extracted 手法と記す．

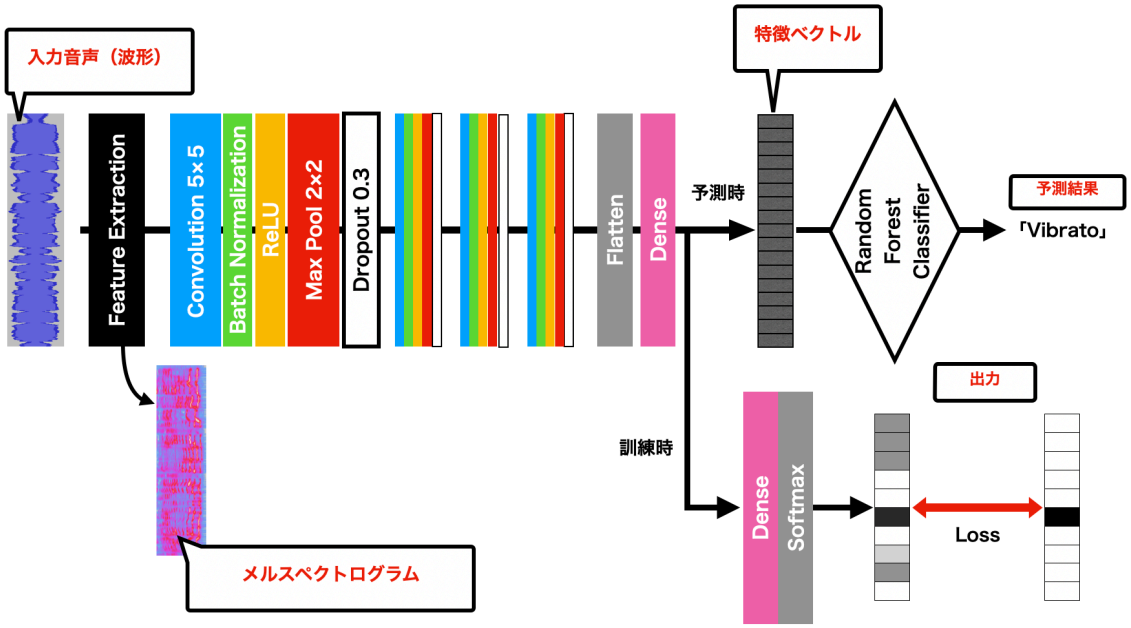

図 1 深層学習による歌唱テクニック特徴量抽出

5.1 実験条件

データセットは訓練データとテストデータの割合を 8:2 になるよう分割した．全クリップ数は訓練データ：テストデータ＝ 3921：980 である．音声はサンプリング周波数 44100 で読み込んだ後，3 秒のクリップにオーバーラップなしで分割し，無音部分のみが含まれるクリップは除いた．

5.1.1 hand-crafted 手法の設定

各々の特徴は全て時間平均をとる．MFCC は次元数 20，ビブラート特徴量は vibrato extend と vibrato rate で次元数 2 とし，結合して 22 次元の特徴ベクトルとした．MFCC の計算は音響信号処理のライブラリである librosa[5]，ビブラート特徴量は CREPE[6] によって抽出された基本周波数（f0）を入力に，音楽音響信号処理のライブラリである Essentia[7] のビブラート検出モジュール [10] を用いた．

5.1.2 CNN-extracted 手法の設定

入力の特徴量に用いるメルスペクトログラムは，ハン窓を窓関数に窓幅 25 ミリ秒，シフト幅 10 ミリ秒の短時間フーリエ変換を用いて振幅スペクトログラムを作成したのち，バンド数 128，最高周波数 8000Hz，最低周波数 20Hz のメルフィルタバンクを適用して作成した．この処理で得られるメルスペクトログラムはサイズ 301×128 の行列である．中間特徴量については，hand-crafted 手法と公平な比較のため次元数を揃え全結合層の出力サイズは同じく 22 とした．この CNN モデルの学習は 40 エポック行い，バッチサイズは 32 とした．最適化手法には Adam[9] を用いた．

5.1.3 分類器

分類器には Random Forest[8] を用いた．Random Forest は決定木を用いたアンサンブル学習の手法であり，教師データと予測に用いる特徴量をランダムサンプリングして複数の決定木を学習し，データへの過適合を防ぐ手法である．先行研究 [2] でも用いられた．今回，決定木の数は 50 個とした．また，Vocalset は表 1 に示す通りクラスによってデータ数に偏りがあるため，クラスごとのサンプルに重み付けをした．Random Forest の実装には Scikit-learn[11] を用いた．前節で分割された教師データを用いて学習を行い，テストデータで評価する．評価は正解率（Accuracy）によって行う．

6. 結果と考察

6.1 識別結果

両手法の正解率を表 2 に示す．結果として，CNN-extracted 手法が全体の正解率では 2.6%上回った．クラスごとの正解率に着目すると，Hand-crafted 手法は straight や vibrato，vocal fry において正解率が高かった．一方，CNN-extracted 手法は belt，breathy，lip trill，inhaled，spoken 等通常の歌唱法と程遠い歌唱テクニックの正解率が高かった．これらの歌唱テクニックはビブラート等と比べ先行事例が少なく，新たに hand-crafted 特徴量を個別に設計するのは容易でなかった．しかし本実験から，CNN を用いることによって，明示的なモデリングなしにこれらの歌唱テクニックも識別可能であることが示唆された．

また，各クラスのベクトルが実際にどのクラスに識別したかを表す混同行列を図 2,3 に示す．どちらの手法においても straight と breathy,trill と trillo の誤分類がみられた．この結果はこれら誤分類が互いに多いテクニックは形態が類似しているテクニックであることを示唆している．

表 2　比較実験の結果

手法	正解率 (全クラス)	belt	breathy	inhaled	lip trill	straight	trill	trillo	spoken	vibrato	vocalfry
Hand-crafted	0.710	0.645	0.505	0.364	0.789	**0.900**	0.269	0.500	0.182	**0.890**	**0.598**
CNN-extracted	**0.736**	**0.724**	**0.61**	**0.577**	**0.948**	0.806	**0.563**	**0.574**	**0.444**	0.875	0.588

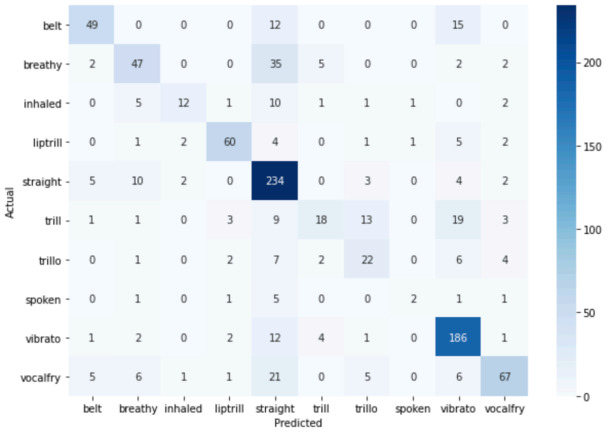

図 2　hand-crafted 手法での識別結果

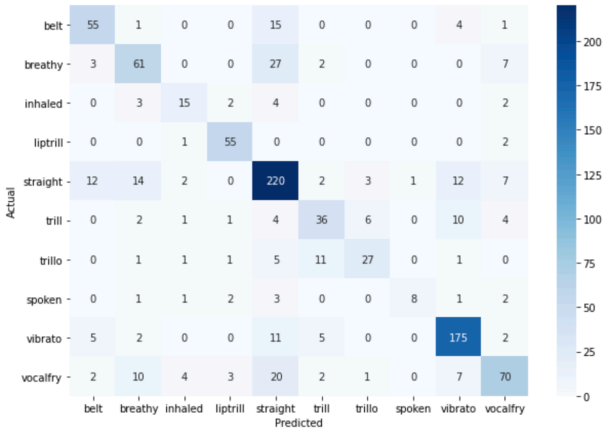

図 3　CNN-extracted 手法での識別結果

6.2　特徴量の可視化

両手法においての特徴ベクトルを t-Distributed Stochastic Neighbor Embedding(t-SNE) [20] を用いて 2 次元に圧縮し可視化した．それぞれの手法の可視化の結果を図 4,5 に示す．特に着目すべき点として，CNN-extracted 手法（図 5）において各クラスがよりまとまっている．特に，他の歌唱テクニックと様相が異なる lip trill が密になっている（図中赤丸）．

一方 Hand-crafted 手法では CNN-extracted 手法に比べばらつきが大きく，特に小節 6.1 で述べたような通常の歌唱法と程遠い歌唱テクニックにおいて，ばらつきが見られた．今回用いた hand-crafted 特徴量のみでは十分に特徴を捉えられてないことが示唆される．

また，興味深い点として，音高や音量を周期的に変動させる vibrato（図中黄丸），trill（図中茶丸），trillo（図中桃丸）が近い位置にクラスタを形成していることが観察できる．このことから，CNN は自律的に歌唱テクニック間の類似関係をも獲得している可能性が示唆される．

7.　終わりに

本研究は歌唱テクニックの識別問題において，hand-crafted 特徴量と深層学習により自動抽出された特徴量の識別性能を比較した．hand-crafted 特徴量には音色と音高変動に関する 2 種類の特徴量，深層学習による特徴量の自動抽出には CNN を用いた．分類器の条件を同一にして識別実験を行った結果，深層学習による手法が全体の正解率を上回った．特に特殊な歌唱テクニックにおいても自律的に特徴を学習し，明示的な知識を必要とせずに識別が可能であることが示唆された．

今後はデータセットのサイズや入力特徴表現を変更し，どのような条件下で深層学習が機能するかを調査する予定である．深層学習手法の利点には，特徴の自動獲得に加えてスケーラビリティの高さがある．すなわちラベル付きデータをより多く獲得することができれば，さらなる性能向上も期待できる．しかし逆に言えばデータ・ラベルが不足している状況下では深層学習による手法は期待できないということでもある．どの程度のラベルデータ量で hand-crafted 特徴量を下回るか，また深層学習と hand-crafted 特徴量を併用した場合での識別性能を調査したい．

また，検出における状況等問題設定が異なるが，本研究の延長として既存曲など現実に存在する歌声データを用いて歌唱テクニックを自動検出することに取り組みたい．既存曲に対し検出を行う場合は，学習すべきラベルデータの不足が懸念材料であることは山本ら [12] によって指摘されている．こうしたラベルデータの不足の状況下で，識別に有用である汎用的な特徴表現を抽出し，十分な識別性能を得ることは未だ挑戦的な課題となっている．

参考文献

[1] N. Kroher and E.Gomez. ”Automatic singer identification for improvisational styles based on vibrato, timbre and statistical performance descriptors.” in ICMC-SMC 2014, 2014.
[2] J. Abesser and M. Muller, ”Fundamental Frequency Contour Classification: A Comparison between Hand-crafted and CNN-based Features,” in ICASSP 2019, 2019.
[3] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, “Vocalset: A singing voice dataset,” in ISMIR 2018, 2018.
[4] O. Sergio, et al. ”Multi-label music genre classification from audio, text and images using deep features.” in ISMIR 2017 , 2017.
[5] B. McFee, C. Raffel, D. Liang, D.PW Ellis, M. McVicar,

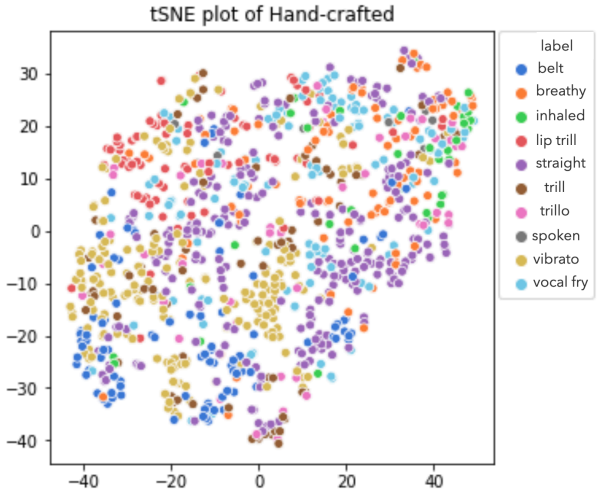

図 4 t-SNE による 2 次元可視化：hand-crafted 手法の特徴ベクトル

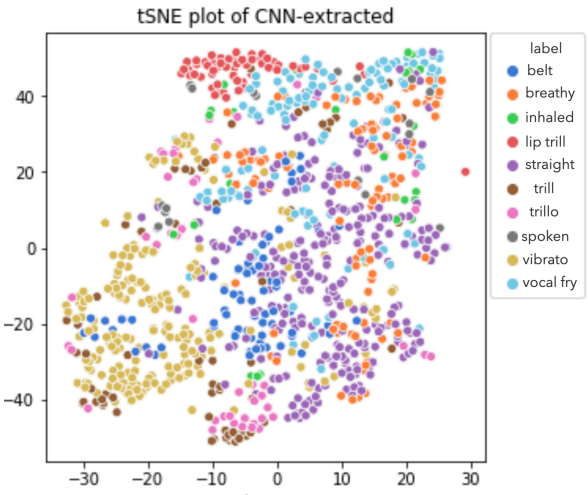

図 5 t-SNE による 2 次元可視化：CNN-extracted 手法の特徴ベクトル

E.Battenberg, and O. Nieto. "librosa: Audio and music signal analysis in python," in Proceedings of the 14th python in science conference, Vol. 8, pp. 18-25, 2015.

[6] J. Kim, J. Salamon, P. Li, and J. Bello. "Crepe: A convolutional representation for pitch estimation" in ICASSP 2018, 2018.

[7] D. Bogdanov, N. Wack, E. Gomez, S. Gulati, H. Boyer, O. Mayor, G.R.Trepat, J. Salamon, J. R. Z. Gonzalez, X. Serra, et al. "Essentia: An audio analysis library for music information retrieval," in ISMIR 2013, 2013.

[8] L. Breiman. "Random forests," Machine learning, Vol. 45, No. 1, pp. 5-32, 2001.

[9] D. P. Kingma, J. Ba. "Adam: A method for stochastic optimization". arXiv preprint arXiv:1412.6980, 2014.

[10] J.salamon, B. Rocha, E. Gomez. "Musical genre classification using melody features extracted from polyphonic music signals.", in ICASSP 2012, 2012.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michael, B. Thirion, O. Grisel, M. Blomdel, P. Prettemhofer, R. Weiss, V. Dubourg. "Scikit-learn: Machine learning in python." the Journal of machine Learning research, Vol. 12, pp. 2825-2830, 2011.

[12] 深山覚，藤田良祐，河原英紀，大矢隼士，下道雄太，山本雄也，須山夢菜，丸山新世，関晋之介，澤田恭平，類家怜央，渡邉一樹，城田晃希，塚本康太，中島凱斗，櫻沢繁，武田郁弥，名畑皓正．"第 2 回音楽情報科学萌芽・デモ・議論セッション"，情報処理学会研究報告，Vol.2020 - MUS - 126，No.4 (2020).

[13] T. Nakano, M. Goto, and Y. Hiraga. "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features." in Ninth International Conference on Spoken Language Processing, 2006.

[14] J. Driedger, S. Balke, S. Ewert, and M. Muller. "Template-based vibrato analysis in music signals." in ISMIR 2016, 2016.

[15] L. Yang, S. Rajab, E. Chew. "Ava: an interactive system for visual and quantitative analyses of vibrato and portamento performance styles." in ISMIR 2016, 2016

[16] 平山健太郎, 伊藤克亘. "ポピュラー歌唱における高音域の声区と発声状態の判別手法." 研究報告音声言語情報処理 (SLP), Vol. 2012, No. 16, pp. 1-6, 1 2012.

[17] D. Stoller, S. Dixon. "Analysis and classification of phonation modes in singing." in ISMIR 2016, 2016.

[18] Y. Ikemiya, K. Itoyama, H. G. Okuno. "Transferring vocal expression of f0 contour using singing voice synthesizer.", in IEA/AIE 2014, 2014.

[19] F. Pishdadian, B. Kim, P. Seetharaman, and B. Pardo. "Classifying non-speech vocals: Deep vs signal processing representations." in Acoustic Scenes and Events 2019 Workshop (DCASE2019), 2019.

[20] L. Maaten, and G. Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11, 2008.