2ZC-02

能の謡分析のためのU-Net音源分離を用いたメロディ抽出*

田本 篤喜†, 伊藤 克亘‡,

1 まえがき

音楽においてヴォーカルは主要な役割である．本研究では西洋の音楽ではなく，能を対象にする．能は，役に扮して舞台に立つ立方と，もっぱら音楽を受け持つ地謡方，囃子方とで成り立つ．立方のうち，主人公であるシテは，舞台の進行役を務める．囃子方は，笛方，小鼓方，大鼓方，太鼓方の四種の楽器で構成される．囃子は，声楽部である謡や動作部である所作とならぶ重要な表現要素である [1]．なお，謡のみの音源のことを素謡と言い，一人での素謡のことを独吟という．

能では，囃子の各楽器や立方の各役籍ごとに，流儀が複数存在する．実際の能の舞台は，異なる流儀の役籍が混じって構成されている．また謡の練習は，師匠の謡をまねるなどして練習する．そのため，謡の音高は，同じ流儀の師匠や個人の声質によって形作られていくことになる．これらの要因から，能では西洋音楽とは異なり音階がはっきりしないために，音階の客観的な研究が必要である．そのために，メロディを正確，詳細に求める必要がある．

絶対的な音高が決まっていない能においては，音階を分析するのにデータが多く必要である．多くの独吟の音源に対して謡を分析したいと思うが，入手できる音源は多くない．伴奏付きの謡の音源からも分析できることで，より多くの謡の分析が可能になる．本研究では，伴奏付きの謡の音源に対して謡のみのメロディをアカペラ並みに分析することを目的とする．

能の謡においては，演目が200程度しか存在しない．そのため，機械学習・深層学習等の手法を用いる場合は学習データの少なさも大きな問題である．

2 ニューラルネットワークによる時間周波数マスク推定に基づく音源分離

二つ以上の音源の音から構成される混合音を分離する手法を音源分離という．特に，音源に関する情報なしの音源分離をブラインド音源分離という．本研究では時間周波数マスクを用いた音源分離を実現する．

2.1 時間周波数マスキング

能における主人公の謡と楽器の音が混じった混合音から音声部分を強調・抽出する手法として，時間周波数マスキングを導入する [2]．時間周波数マスキングとは，混合音を時間周波数領域で分離する手法である．混合音がソース1とソース2の音源からの，2つの音が混じった音であると想定する場合，ソース1に関する理想的な時間周波数マスク $\mathbf{M}$ は以下の式 (1) のように定義できる．

$$\mathbf{M} = \frac{|\hat{\mathbf{y}}_{1_t}(f)|}{|\hat{\mathbf{y}}_{1_t}(f)| + |\hat{\mathbf{y}}_{2_t}(f)|} \tag{1}$$

ここで $\mathbf{y}_1$ はソース1の音源のスペクトログラム (時間周波数情報) を表している．また t は時間，f は周波数を表している．

計算した時間周波数マスク $\mathbf{M}$ を，混合音に適用することで，所望の音を抽出できる．ソース1に関する理想的なマスクを使用してソース1に対応する音を取り出したいときは，

$$\hat{\mathbf{s}}_{1_t}(f) = \mathbf{M}(f)\mathbf{X}_t(f) \tag{2}$$

の処理で取り出すことができる．なお，$\mathbf{X}_t$ は混合音のスペクトルを表している．

2.2 ニューラルネットワークによる時間周波数マスク推定

所望の音源からの音声だけを強調する理想的な時間周波数マスクを求めるためにニューラルネットワークを用いる．本研究では，U-Net[3] による時間周波数マスク推定を導入する．

U-Net は CNN ベースのネットワークであり，医療画像の高解像度化に関する研究に用いられた手法である．時間周波数マスク推定のための U-Net の構成を以下の図1に示す．

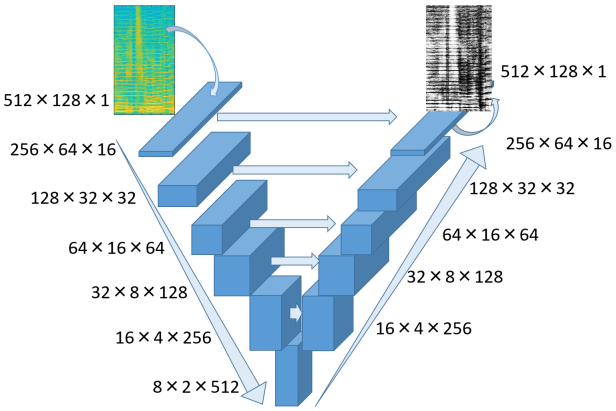

図 1. 時間周波数マスク推定 U-Net の構成

各畳み込み層での出力を同じレベルの逆畳み込み層の入力データに連結して演算する．この low level skip connection によって倍音成分まで高解像度でマスクが推定でき，効率的に学習できることを期待する．

3 メロディ推定器

3.1 概要

本研究では音源分離後にメロディ推定器を適用するが，音源分離後でも目的音のみを分離できているわけではなく目的音以外の音が混じっていることが考えられるため，音声時間波形の自己相関によってメロディを求めるべきでない．本研究では Melodia と CNN,LSTM によるメロディ推定器を比較する．

3.2 CNN,LSTM によるメロディ抽出

本研究では音源分離後の音声でメロディ推定 CNN,LSTM を学習することによる精度を確かめるため，音源分離後の謡が入力に対応する．

---

*: Melody extraction using U-Net source separation for analyzing Noh singing *Atsuki Tamoto (Hosei Univ.) et al.*

†法政大学大学院 情報科学研究科
‡法政大学 情報科学部

表 1. 実験結果 (%)

	1. Melodia	2. CNN	3. U-Net+Melodia	4. U-Net+CNN	5. U-Net+Clean CNN	6. U-Net+LSTM
RPA	72.3	79.8	92.7	88.4	82.2	89.7
OA	76.5	78.8	91.9	88.3	84.5	87.5

CNN の構造は, [4] で実装しているネットワークを基に構成する. 能の謡のバリエーションの少なさを考慮し, [4] の構成に加えて, 全結合層を一層追加する. LSTM の構造は, [5] に基づき, 2000unit の LSTM 一層としている. CNN の入力データのコンテキストは前 7 フレーム, 後ろ 8 フレーム, LSTM の入力データのコンテキストは前 12 フレーム, 後ろ 13 フレームとした. ラベルデータは, 正解周波数をセントに変換し, 50 ノードごとにまとめて 1 ノードとしている. なお, 正解ラベルに対応するメロディの決定には, 既存のメロディ推定器を用いる [6].

4 DATA AUGMENTATION

入手可能な能の謡や囃子の音源は少ない. そこで, データを増量する処理を施す. ひとつの謡に対していくつかの組み合わせで囃子を足し合わせる. 能の謡の演目は絶対的に少なく, 歌い方に特徴もあるため, このようにすることで, 多くのバリエーションの混合音に対するマスクを学習できることになる.

5 評価

Raw Pitch Accuracy(RPA) と Overall Accuracy(OA) の尺度で評価を行う [7].

5.1 使用するデータ

実際に収集した独吟と囃子の音源を用いる. メロディ推定 CNN,LSTM の学習データには, 音源分離部で使用しなかったデータで混合音を作成し, 音源分離を適用した後の音を用いる. U-Net の学習データ量は約 104 時間, メロディ推定 CNN, LSTM の学習データ量は約 4 時間である. 評価には, 30 秒ごとにランダムに抽出した 60 ファイル分のデータとした.

5.2 結果と考察

混合音に対して直接 Melodia を適用した結果 (1), メロディ推定 CNN を混合音に対して直接適用した結果 (2), U-Net 音源分離後の音に対して Melodia を適用した結果 (3), U-Net 音源分離後の音に対してメロディ推定 CNN,LSTM を施した結果 (4,6), U-Net 音源分離によって分離した音に対してクリーンな謡で学習したメロディ推定 CNN を適用した結果 (5) を表 1 に示す. なお, メロディ推定 CNN,LSTM では, 入力にコンテキストフレームを含むため, 評価の際, コンテキストのフレームは含めていない.

4. と 5. を比較すると, 4. の方が良い結果であることがわかる. この結果から, U-Net 音源分離をメロディ推定の前処理として施す場合でも, 音源分離後の謡でメロディ推定 CNN を学習することで, ロバストなメロディ推定が可能であることがわかる. 3. が全ての結果の中で最も良い結果であった. スペクトログラム上の倍音構造が重要な Melodia を適用する前処理として, 高解像度でマスクを推定する U-Net による音源分離を施したことで高性能なメロディ推定が実現できた. CNN と LSTM を比較すると, 音声区間においてはコンテキストをより柔軟に考慮できる LSTM の方が良い結果が得られた. 深層学習手法に関して, 学習データとして能の謡と囃子を大量に集め, 能のための音源分離器を構成することができたことで, 良い性能が得られたと考えられる.

6 結論と今後の展望

本研究では, メロディ推定器の前処理として音源分離を施す手法を実装した. U-Net と Melodia を用いた手法で最も良い性能が得られた. 能の演目が少なく, 今回大量に能のデータを収集できたことで, 良い精度でのメロディ推定ができた. 加えて多くの組み合わせを考慮する data augmentation によってこの良い結果が得られたと考えられる.

本研究では, 独吟に対して囃子が混じっている音を対象とした. 実際の能では, 地謡方の謡も含まれるため, より汎用性を高めるためには, 地謡方を考慮する必要がある. 複数人が同じ謡を謡っていることもあるため, 独吟だけでなく, 複数人による謡も考慮する必要がある. 音源分離とメロディ推定のネットワークを連結することで, さらなる精度向上を狙う.

参考文献

[1] “囃子”, 新版 能・狂言事典, JapanKnowledge Lib, https://japanknowledge.com, (参照 2019-07-24)

[2] P.-S.Huang, M.Kim, M.Hasegawa-Johnson, and P.Smaragdis, “Deep learning for monaural speech separation”, in Proc.IEEE Int. Conf.Acoust.,Speech,Signal Process. (ICASSP), 2014,pp.1562-1566.

[3] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, “Singing voice separation with deep u-net convolutional networks,” in Proc. 18th Int. Soc. Music Inf. Retrieval Conf., 2017, pp. 23-27.

[4] Hong Su, Hui Zhang, Xueliang Zhang, and Guanglai Gao, “Convolutional neural network for robust pitch determination,” in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016, pp. 579-583.

[5] H. Park and C. Yoo. “Melody extraction and detection through LSTM-RNN with harmonic sum loss,” In IEEE ICASSP, pages 2766-2770, 2017.

[6] J. Salamon and E. G´ omez, “Melody extraction from polyphonic music signals using pitch contour characteristics,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 6, pp. 1759-1770, 2012.

[7] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, “mir eval: a transparent implementation of common MIR metrics,” in Proc. of the 15th ISMIR, Taipei, Taiwan, 2014.