N-009

# 能の謡分析のためのブラインド音源分離を用いた F0 抽出 *

田本 篤喜 †,　伊藤 克亘 ‡,

## 1　まえがき

日本の伝統芸能である能楽は, 重要無形文化財に指定され, ユネスコ無形文化遺産にも登録されている. 能楽は, 狂言と能の総称である. 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]. 謡のみによって構成される場面, 謡と囃子がともに演奏される場面, 囃子のみが演奏される場面, が能を構成する音として挙げられる

入手可能な能の音源として, 独吟, 複数人による素謡, 囃子入り謡などがある. 謡の基本周波数 (F0) を分析したい場合, これらのような音源を用いて分析することになるが, 独吟の音源は少ない. 比較的入手可能な複数人による素謡の音源や囃子入り謡に含まれる謡を分析する場合, 囃子や他の謡などが混じっているため, 素謡と同じ条件での分析が不可能である.

謡のみの F0 を推定する前処理として音源分離を施すことで, 主人公の謡以外の音を抑圧した音に対して F0 の分析が可能になることを示す. 能の演目の少なさを考慮した, 音源分離器の学習データ増量, F0 推定器の検討を行う. 複数人による素謡が含まれる音源にも対応できることを目指すが, 本研究では, 囃子入りを対象とする.

## 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. 二つの音源分離器を実装し, 比較する. 時間周波数マスク推定のための DNN の構成を以下の図 1 に示す.

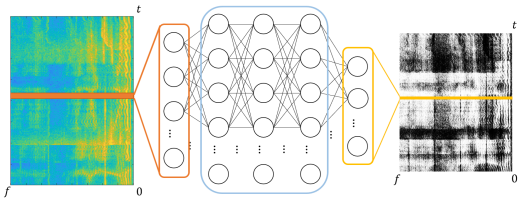


**図 1.** 時間周波数マスク推定 DNN の構成

音声のつながり具合を特徴としてニューラルネットワークに学習させる. つまり, 学習時には混合音のスペクトル情報を複数フレーム連結させたデータを入力, 事前に計算した理想的な時間周波数マスクを正解として与えて学習させる. 実際の推論時には, 学習時の入力と同じフレーム数分のスペクトル情報を学習済みのニューラルネットワークに入力することで, 時間周波数マスクが出力される.

次に, 時間周波数マスク推定のための U-Net の構成を図 2 に示す. 畳み込み層での出力を同じレベルの逆

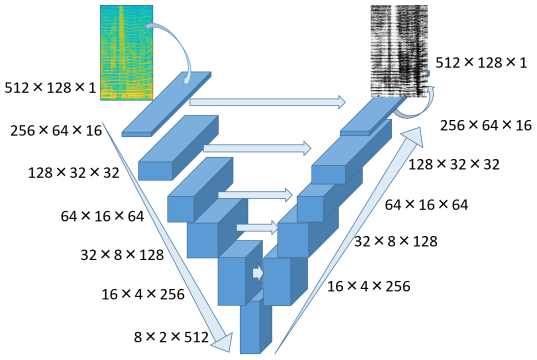


**図 2.** 時間周波数マスク推定 DNN の構成

畳み込み層に連結して演算を行う. このスキップコネクションによって倍音成分まで高解像度でマスクが推定できることを期待する.

---

*: F0 estimation using blind 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. All results(%)

| | 1. Melodia | 2. DNN+Melodia | 3. U-Net+Melodia | 4. DNN+F0 推定 CNN | 5. U-Net+F0 推定 CNN |
|---|---|---|---|---|---|
| RPA | 72.3 | 48.6 | 92.7 | 58.2 | 93.0 |
| OA | 76.5 | 59.2 | 91.9 | 69.2 | 92.8 |

## 3 F0 推定器

### 3.1 直接推定

音源分離によって主人公の謡以外の音を抑圧した音に対して F0 を推定する．自己相関によって求める代表的な F0 推定手法では, 生の音声時間波形の自己相関を用いるため, F0 を求めたい目的音以外の音が対象のファイルに混じっている場合, それらの音に強く影響されてしまう．本研究では音源分離後に適用するが, 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 F0 を求めるべきでない．そこで本研究ではニューラルネットワークを用いた F0 抽出器を導入する．CNN を用いた F0 推定器がロバストであるとの報告がある [3]．そこで本研究では CNN を用いた F0 推定器を導入する．

### 3.2 CNN による F0 抽出

混合音の時間周波数情報から直接 F0 を推定する CNN を構成する．CNN の構成を以下の図 3 に示す．入力データとして混合音のスペクトログラム, ラベルデータには混合音に含まれる謡部分のみの F0 に対応したセントの値を設定し学習する．なお, 正解ラベルに対応する F0 の決定には, 既存の F0 推定器を用いる [4].

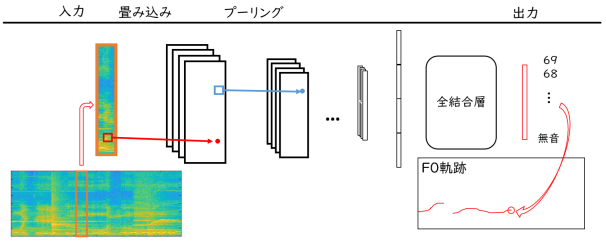


図 3. CNN を用いた F0 推定器の構成

## 4 評価

音源分離を施した後の音に対して F0 推定器を適用し, 評価する．

### 4.1 使用するデータ

音源分離の DNN,U-Net と F0 を推定する CNN の学習データとして, 実際に収集した素謡と, 伴奏 (囃子) の音源を使用する．推論時には, それらを混合させたデータを入力とする．能のデータの少なさを考慮し, 素謡, 伴奏を全組み合わせで学習データとする．

## 5 実験条件

音源分離部では, 音声のつながり具合を学習させるために, 512 点でフーリエ変換した 257 点のスペクトル情報を, 対象のフレームとその前後 7 フレーム分, 1 方向に連結させ, DNN の入力とする．中間層は 4 層あり, 入力側から 4096-2048-2048-1024 のノードを持つ．出力層では, 対象の 1 フレーム分のノードを持つ．損失関数には, 正解マスクと推定されたマスクの二乗誤差を最小化する式を導入している．U-Net の構造は [5] と同様である．CNN の構造は, [3] と同様であるが能の演目の少なさを考慮し, 600 ノードを持つ層を追加している．

## 6 結果と考察

混合音に直接 Melodia を適用した結果 (1), DNN,U-Net 音源分離後に Melodia を適用した結果 (2,3), DNN,U-Net 音源分離後に F0 推定 CNN を適用した結果 (4,5) を表 1 に示す．

2. と 4. を比較すると, F0 推定 CNN の方が良い結果となっているが, 1. よりも悪い結果となった．RPA と OA の差に注目すると DNN 音源分離によって無音区間の性能は改善されたが, 音声区間の音源分離性能が良くないことがわかる．5. の結果が最も良い結果となった．音源分離後の音で学習した F0 推定 CNN が良い結果であるとともに, U-Net による音源分離自体の性能が良いことが, RPA と OA の差からわかる．

## 7 あとがき

音源分離後の音で F0 推定 CNN を学習する手法を実装し評価を行った．能の演目の少なさを考慮した学習データの検討や CNN のネットワークの改良を行うことで, F0 推定性能が改善された．現在は, 囃子方によって演奏される楽器音と掛け声のみを対象としているが, 実際には地謡方というコーラス舞台による謡が加わる．コーラスによるシテの謡の F0 への影響もあるため, 考慮する必要がある．また, 音源分離と F0 推定のネットワークを連結することで, F0 推定性能が向上するように学習する時間周波数マスク推定器を検討する．

## 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] 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.

[4] 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.

[5] A.Jansson, E.J.Humphrey, N.Montecchio, R.M.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.