Vol.2016-MUS-111 No.46
2016/5/22

楽器と音高の同時認識のためのRNN音響モデル

生田目 敬弘[1,a)]　亀岡 弘和[2,b)]　篠田 浩一[1,c)]

概要：複数の楽器を含む楽曲について楽器識別と音高認識を同時に行う RNN 音響モデルを提案する．楽器別に音高を推定するため，RNN の出力層を識別する楽器数に応じて増加させ，楽器を区別した教師信号を与えて学習させる．主流な音高推定手法では音高を得るために音源分離を試みているが，提案する RNN では難しい音源分離問題を回避して直接音高を推定できる．また，教師信号として音高に加えて楽器の種類を与えることで RNN がスペクトル形状を認識しやすくなることが期待できる他，楽器別に音高を得られることから，楽器別に学習した音楽言語モデルと統合することが可能になる利点がある．ヴァイオリンとクラリネットの 2 重奏曲について楽器推定と音高推定を行う実験により，クラリネットについて 0.9%とまったく検出できなかったが，ヴァイオリンについて 66.0%と，従来手法の 68.3%と同程度の推定性能を持つことが分かった．ヴァイオリンについて性能が良くクラリネットについて性能が良くない原因はデータセットの不足と質の悪さに起因すると考えられ，今後学習データの増加によって改善する見込みである．

1. はじめに

近年急速に発展を遂げている音楽情報処理の中でも，波形データの音楽音響信号から演奏内容を表す楽譜形式へと変換する自動採譜は，演奏を再現する目的だけでなく，カバー楽曲やアレンジ楽曲を含む音楽検索システムへの活用，また他者による楽曲の利用を検出する著作権管理技術への応用など，様々な分野で需要が高い．

自動採譜は，音高推定，リズム・拍推定，テンポ推定，調推定などの部分問題に分けられ，中でも音高推定はその核をなしており重要である．楽音のスペクトルは同一音高であっても楽器の種類によって大きく異なる場合がある．このため，単音を対象とした場合の音高推定であってもスペクトルからの単純なピーク検出だけでは不十分であり，スペクトル全体を手がかりにする必要がある．しかし，複数の楽音が混在した混合音スペクトルには，各周波数成分がどの楽音に由来したものであるかという情報が欠損しているため，複数音の音高推定は単音の場合よりさらに難しい問題となる．

音楽の音高推定は以上のような性質のため，従来は音源分離と音高推定をセットの問題として扱ったものが多い．例えば，音高を推定するステップと，その対応する調波成分を入力信号から減算するステップを反復して音高推定するアプローチ [12] や，パラメトリックモデルで調波構造を定式化し，音源分離に相当するステップと基本周波数パラメータを推定するステップを反復するアルゴリズムによる推定手法 [11] などが提案されている．

また，混合音のスペクトログラムを非負の行列と見なし，2 つの非負行列の積に分解する非負値行列因子分解（Non-negative Matrix Factorization; NMF）に基づくアプローチが近年注目されている．スペクトログラムを 2 つの非負行列の積で近似することは，各時刻の混合音スペクトルに非負のスペクトルテンプレートを非負の重みで混合したものをフィッティングさせていることに相当する．したがって，各スペクトルテンプレートを異なる音高の楽音スペクトルとし，各時刻の混合音スペクトルに最もフィットするように重み係数を推定することで，それぞれの音高がどれくらいの大きさで発音しているかを推定できる．これらの 2 つの非負行列をそれぞれ周波数と時間の条件付き分布と捉え，観測スペクトログラムを周波数と時間の同時分布から生成された周波数と時間のヒストグラムとみなせば，最尤基準により観測スペクトログラムから 2 つの条件付き分布を得ることができる．これはある特定のフィッティング基準を用いた場合の NMF と手続き的に等価であり，確率的潜在成分分析（Probabilistic Latent Component Analysis; PLCA）と呼ばれている．また，NMF や PLCA の拡張版として，楽音の調波構造において基本周波数と各倍音の間隔が対数周波数軸上でシフト不変となる性質に基づき，対数周波数スペクトルのテンプレートと音高の音量

[1] 東京工業大学
Tokyo Institute of Technology
[2] 東京大学
The University of Tokyo
[a)] namatame@ks.cs.titech.ac.jp
[b)] kameoka@hil.t.u-tokyo.ac.jp
[c)] shinoda@cs.titech.ac.jp

分布との畳み込みにより混合音スペクトルを表現できることに着目した手法も提案されており，shifted NMF[8]，シフト不変 PLCA[17] と呼ばれている．また，シフト不変の考え方とソースフィルタモデルを組み合わせたアプローチも近年提案されている [13]．Emmanouil らは，楽器の調波構造を事前に与えることで多楽器の楽曲について精度を改善させた多楽器シフト不変 PLCA[2] を提案した後，隠れマルコフモデルの導入によってスペクトル形状について時間方向に制約を与える時間的制約付き PLCA[3] を提案し，国際的コンテストである MIREX 2015 の多重基本周波数推定タスクにおいて最も高い性能を示している．

これらの従来法は，音源分離と音高推定を同時に解決しようという戦略に基づいていると言う点で共通しているが，音源分離は音高以外の情報（各楽音のスペクトルなど）も推定する必要があるという意味で本来解くべき音高推定の問題以上に難しい問題であるといえる．従って，音高さえ推定できれば十分であるような状況ではスペクトルから音高ラベルを直接推定する識別問題として捉えたアプローチをとる方が合理的である．このようなアプローチとしてこれまでサポートベクターマシーン（Support Vector Machine; SVM）やリカレントニューラルネットワーク（Recurrent Neural Network; RNN）を識別器として用いたアプローチが提案されている [14], [15]．RNN を用いた手法では，多重音の音高推定を各音高に対するマルチラベルの分類問題と捉えることで，ニューラルネットワークによって入力信号から直接音高を推定できる．

先に述べたように，音響信号に含まれるスペクトルの形状を推定すること，すなわち楽器の種類の推定は，音高を推定する上で極めて重要である．音響信号のスペクトルに含まれる楽器の調波構造が分かっていれば，スペクトルの各成分がどの楽器に由来するかを推定しやすくなり，音高を推定する手がかりになる．逆に音高が推定されれば，含まれる音のスペクトル形状を推定しやすくなり，楽器を推定する手がかりになる．このように音高推定と楽器推定は互いに補う関係にあると考えられる．また，楽器別に音高推定を行うことができれば，音楽のための言語モデルを導入するような応用 [15], [16] において，楽器別の言語モデルを生成することができる．これによって，楽器によって異なる旋律の傾向を捉えることができるようになると考えられる．

そこで本稿では，楽器推定と音高推定を同時に行うことにより楽器推定結果と音高推定結果が相補的に精度を向上させることを図り，かつ楽器別に音高推定を行うことが可能な RNN 音響モデルを提案する．

2. PLCA による音高推定

確率的潜在コンポーネント分析（PLCA）[18] は，Smaragdis らによって考案された，スペクトログラムを確率分布と捉えて分解し，音高，楽器，時間に関する確率分布を得ることで各楽器について全時間での音高を推定する手法である．現在最も性能が良い音高推定手法として，Emmanouil らが提案した，入力の音楽音響信号を定 Q 変換 [5] により対数周波数スペクトログラムに変換し，時間的制約付き PLCA を用いて音高を推定する方法 [3] について述べる．

PLCA は，スペクトログラム $V_{\omega,t}$ を対数周波数 ω と時間 t に関する確率分布 $P(\omega,t)$ と見なして複数の確率分布に分解する．確率分布 $P(\omega,t)$ は以下の式でモデル化される．

$$P(\omega,t) = P(t)\sum_{p,q,f,s} P(\omega|p,q,f,s)P(f|p,t)P(s|p,t)P(p|t)P(q|p,t) \tag{1}$$

p は音高，q は楽音の状態，f は周波数シフト因子，s は楽器を表す．$P(\omega|p,q,f,s)$ は楽器 s，音高 p，状態 q，周波数シフト f の音のスペクトルを表す確率分布であり，単音のデータを用いて事前に抽出できる．$P(t)$ はスペクトログラムの時間 t でのパワースペクトル密度 $\sum_\omega V_{\omega,t}$ と考えることができ，入力スペクトログラム $V_{\omega,t}$ より計算できる．

残りの未知の 4 パラメータ $P(f|p,t), P(s|p,t), P(p|t), P(q|p,t)$ を期待値最大化法（EM アルゴリズム）により推定することで，音高の情報を持つ $P(p|t)$ 及び楽器の分布を表す $P(s|p,t)$ が得られる．

$P(s|p,t)$ 及び $P(p|t)$ より，

$$P(p,s,t) = P(p|t)P(s|p,t)\sum_\omega V_{\omega,t} \tag{2}$$

を得，生起を平滑化するため 5 サンプルのメディアンフィルタ処理を加えた上でしきい値 θ によって 2 値化し，各楽器，各時間の音高が推定される．

PLCA はスペクトログラムを分解する手法であり，音源分離と音高推定を同時に解決しようとしている．しかし，音源分離は音高以外にも各楽音のスペクトルや時間による変化など，様々な情報を推定する必要があり，単に音高を推定するより難しい問題に取り組んでいると考えられる．そのため，多重音の音高推定を各音高に対するマルチラベルの分類問題と捉え，次章で述べるようなニューラルネットワークによって信号から音高を直接推定する方法が提案された．

3. RNN による音高認識

リカレントニューラルネットワーク（RNN）は時系列データの学習に適した統計モデルである．RNN は図 1 のような入力層 $\boldsymbol{x}$，隠れ層 $\boldsymbol{h}$，出力層 $\boldsymbol{y}$ から成る 3 層構造のネットワークで，入力データ系列 $\{\boldsymbol{x}_t\}$ に対して

$$\boldsymbol{h}_t = \sigma(W_{hx}\boldsymbol{x}_t + W_{hh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_h) \tag{3}$$

$$\boldsymbol{y}_t = \sigma(W_{yh}\boldsymbol{h}_t + \boldsymbol{b}_y) \tag{4}$$

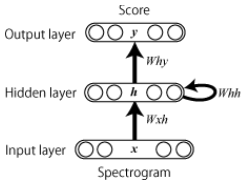

図 1　通常の RNN

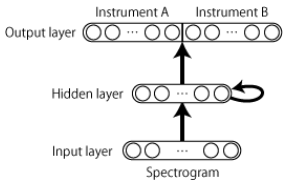

図 2　提案する RNN

によって予測系列 $\{\boldsymbol{y}_t\}$ を得る．W_{hx}, W_{hh}, W_{yh} は各層間の重みパラメータ，$\boldsymbol{b}_h, \boldsymbol{b}_y$ はそれぞれ隠れ層および出力層のバイアスパラメータを表す．$\sigma(\boldsymbol{x})$ は要素ごとに計算される活性化関数を表し，通常シグモイド関数 $\sigma(x) = (1 + e^{-ax})^{-1}$ が用いられる．RNN のパラメータの学習は通時的誤差逆伝播法（BPTT）によって行う．

Siddharth らが提案した RNN を用いた音高認識 [15] では，楽曲の音響信号から定 Q 変換で得たスペクトログラムの一部を入力データ系列として与え，出力の教師信号として音高を，2 値 {0,1} の 88 次元ベクトルの形で与える．

RNN は隠れ層同士をつなぐ回帰結合を持つことにより，隠れ層が内部状態を保持する役割を持ち，時間方向の相関を考慮して音高を推定できる．学習された RNN は入力スペクトログラムに対して各フレームの音高を推定するモデルとなる．出力は $(0, 1)$ の値をとり，これは各音高の生起確率に対応し，しきい値によって 2 値化できる．

4. 楽器と音高を同時認識する RNN

前章で述べた標準の RNN は，楽器を区別しない音高を教師データとして与えられ学習しているため，入力音響信号に含まれる楽器を推定できず，楽器別に音高が得られない問題がある．楽器を区別しない音高推定結果は，楽譜の獲得や楽曲検索への活用という目的に沿わない．また，楽器ごとに大きく異なるスペクトル形状をすべて同一の音高と認識するのは難しく，特に複数の音の周波数成分が混ざったとき音高認識は困難になると考えられる．そこで，図 2 のように出力層を識別させたい楽器数だけ増やし，楽器別に分離した音高ラベルを並列に並べたものを教師信号として学習させる手法を提案する．

提案する RNN では，出力は各楽器，各音高の生起確率を表していると捉えることができ，楽器別の音高推定結果が得られる．これにより応用として [15], [16] のように旋律をモデル化する音楽言語モデルと組み合わせるとき，全ての楽器について区別せずに旋律をモデル化するのではなく，楽器ごとにそれぞれ言語モデルを生成して利用することができる．楽器によって旋律の傾向はまったく異なるため，楽器を区別しない言語モデルでは記述するのは無理があると考えられる．また，単旋律と仮定できる楽器の言語モデルによる音高予測はシングルラベルの分類問題となり，予測性能が向上することが期待できる．

この手法では音高だけでなく楽器の種類も教師信号として与えることになるため，楽器ごとのスペクトル形状を識別できるようになり，複数の楽器の音が混ざっているときでも個々のスペクトル形状を認識できるようになる狙いもある．

表 1　RNN の設定

Constant-Q Transform	Sampling frequency[kHz]	44.1
	Length of a frame [ms]	40
	Q	1.0
RNN	Hidden layer[dim]	250*2
	Activation function	Sigmoid
	Input length[frames]	60
	Stepsize[frames]	20
	Batchsize	128
	Loss function	Cross entropy
	Optimization algorithm	AdaDelta

5. 評価実験

提案する楽器認識と音高認識を同時に行う RNN 音響モデルの性能を評価するため，ヴァイオリンとクラリネットの 2 重奏曲を用いて比較実験を行った．

5.1 実験条件

定 Q 変換 [5] および PLCA[3] には，文献 [4] に記載のものを用いた．PLCA に与えるパラメータはプログラムコード内のデフォルト値を使用した．また，RNN と条件を近づけるため，入力される楽曲はヴァイオリンとクラリネットのみで構成されていると仮定を置いている．しきい値 θ はいくつかの値を選んで実験し，最も結果が良いものを選んだ．

RNN には，単楽器楽曲の音高推定に RNN を採用した手法 [15] を参考に，表 1 のようにパラメータを定めた．

PLCA に使用する楽器別のスペクトルテンプレートは，RWC 研究用音楽データベース（楽器音）[10] からヴァイオリン，クラリネットのノーマル奏法の演奏データ（RWC-MDB-I-2001 No.15，No.31）を使用した．

提案する RNN 音響モデルの学習には，RWC 研究用音楽データベース（クラシック音楽）[9] のうち対象となる 2 楽器のいずれかが含まれる 24 曲，RWC 研究用音楽データベース（楽器音）よりクラリネットのノーマル奏法の 3 曲（RWC-MDB-I-2001 No.31），様々なクラシック音楽が録音された Su Dataset[19] のうち対象の楽器を含む 6 曲，木管 5 重奏の楽曲が楽器ごとに録音された MIREX 2007 Development Dataset[1] のうち対象の楽器を含む 2 曲，合計 34 曲を使用した．RWC 研究用音楽データベースに付属する楽譜データは実際に演奏された楽曲と同期していないため，Müller らによる Music Synchronization for RWC Music Database[7] を使用した．

評価には，クラリネット，ヴァイオリン，バスーン，

表 2　音高推定結果

	Violin	Clarinet
[3]	68.30%	78.05%
Proposed	66.00%	0.86%

サックスによる 4 重奏の楽曲が 10 曲収録された Bach10 Dataset[6] を用いた．Bach10 Dataset の楽曲データは楽器ごとに収録されており，この実験ではクラリネットとヴァイオリンを重ねたデータを使用した．

時間単位を 1 フレーム=10 ミリ秒と定め，フレームごとに推定した音高を報告する．フレーム t で正解と一致した音高の数を $TP[t]$，出力のうち不正解の音高の数を $FP[t]$，正解のうち出力されなかった音高の数を $FN[t]$ とし，

$$\mathcal{P} = \frac{\sum_t TP[t]}{\sum_t (TP[t] + FP[t])} \tag{5}$$

$$\mathcal{R} = \frac{\sum_t TP[t]}{\sum_t (TP[t] + FN[t])} \tag{6}$$

$$\mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P} + \mathcal{R}} \tag{7}$$

によって計算された F 値 $\mathcal{F}$ を評価に使用する．

5.2　実験結果

PLCA および RNN による音高推定結果を各楽器について正解の楽譜と比較した際の F 値は表 2 のようであった．

また，提案する RNN 音響モデルによる出力の一例を，正解の楽譜とともにそれぞれ図 3，図 4 に示す．

全 10 曲の 2 重奏楽曲について，ヴァイオリンの音高推定は従来手法と変わらない精度であったが，クラリネットについて検出できなかった．図 3，4 からも，ヴァイオリンについては音高を認識している一方で，クラリネットについてはまったく反応を示していないことが分かる．

原因としては，データセットの不足と，この問題に適したデータセットでなかったことが考えられる．データセットにはクラリネットを含む楽曲が 9 曲しかなく，全体でも 33 曲しか用意できなかった．またデータ長の観点から，学習データのほとんどを RWC 研究用音楽データベースが占めており，同じ環境で録音されたデータに偏っていることが，汎化性能に悪影響を及ぼしていると考えられる．

また，学習時にヴァイオリンとクラリネットの音高情報を教師信号として与えていたが，学習データには識別の対象となっていないビオラ，チェロ等の楽器音が含まれていた．そのため，入力には未知の楽器音が含まれており，それらについては無視するよう学習されている．学習データのクラリネットとテストデータのクラリネットが同一の楽器であると識別されず，未知の楽器の音として処理されたためにクラリネット部分の出力が反応しなかったと考えられる．

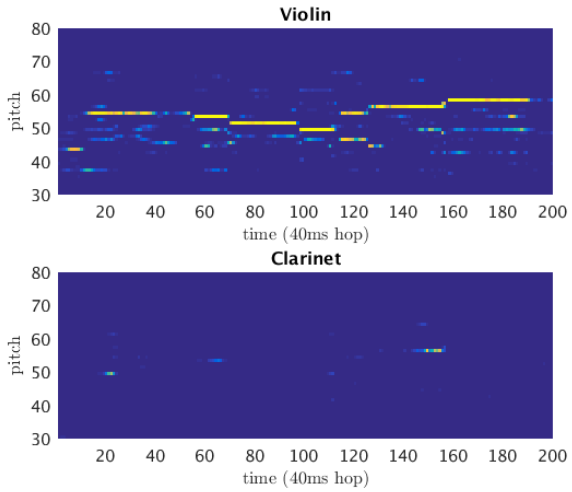

図 3　RNN による出力例．上がヴァイオリン，下はクラリネットについての音高推定結果を表している．

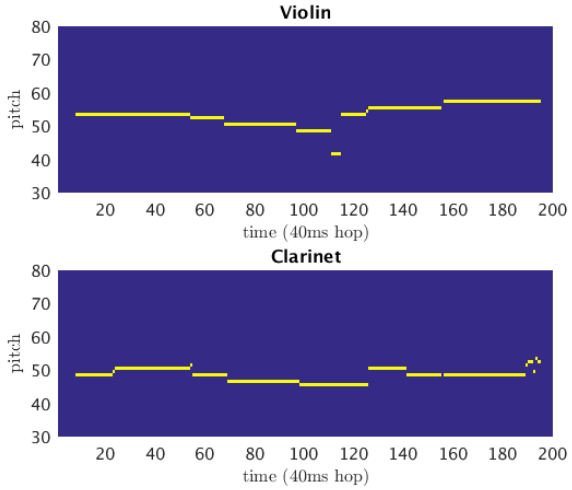

図 4　対応する正解の楽譜．同じく上がヴァイオリン，下がクラリネットの楽譜である．

6.　おわりに

本報告では，リカレントニューラルネットワークを用いて楽器認識と音高認識を同時に行う手法を提案し，楽器別音高推定の精度向上を試みた．

比較実験の結果，提案する RNN ではヴァイオリンについて従来手法である PLCA と同程度の音高推定ができたが，クラリネットについてはまったく検出できなかった．

現状，対象とする楽器や学習するデータセットに偏りがあり検出できる楽器が限定されているため，今後はデータセットの幅を広げて汎用性を高めることや，少ないデータセットでも学習が可能なように同じデータからより多くの情報を学習できるような方法の考案に取り組んでいきたい．

参考文献

[1] Benetos, E. and Dixon, S.: Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription, *Selected Topics in Signal Processing, IEEE Journal of*, Vol. 5, No. 6, pp. 1111–1123 (2011).

[2] Benetos, E. and Dixon, S.: A shift-invariant latent variable model for automatic music transcription, *Computer Music Journal*, Vol. 36, No. 4, pp. 81–94 (2012).
[3] Benetos, E. and Dixon, S.: Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model, *Journal of the Acoustical Society of America*, Vol. 133, No. 3, pp. 1727–1741 (2013).
[4] Benetos, E. and Weyde, T.: An efficient temporally-constrained probabilistic model for multiple-instrument music transcription, *Proc. of the 16th International Conference on Music Information Retrieval (ISMIR 2014)*, pp. 701–707 (2015).
[5] Brown, J. C.: Calculation of a constant Q spectral transform, *Journal of the Acoustical Society of America*, Vol. 89, No. 1, pp. 425–434 (1991).
[6] Duan, Z., Pardo, B. and Zhang, C.: Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions, *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 18, No. 8, pp. 2121–2133 (2010).
[7] Ewert, S., Müller, M. and Grosche, P.: High resolution audio synchronization using chroma onset features, *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, pp. 1869–1872 (2009).
[8] Fitzgerald, D., Cranitch, M. and Coyle, E.: Shifted non-negative matrix factorisation for sound source separation, *Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*, IEEE, pp. 1132–1137 (2005).
[9] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC music database: Popular, classical and jazz music databases, *Proc. of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pp. 287–288 (2002).
[10] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC music database: Music genre database and musical instrument sound database, *Proc. of the 4th International Conference on Music Information Retrieval (ISMIR 2003)*, pp. 229–230 (2003).
[11] Kameoka, H., Nishimoto, T. and Sagayama, S.: A multipitch analyzer based on harmonic temporal structured clustering, *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 15, No. 3, pp. 982–994 (2007).
[12] Klapuri, A. P.: Multiple fundamental frequency estimation based on harmonicity and spectral smoothness, *Speech and Audio Processing, IEEE Transactions on*, Vol. 11, No. 6, pp. 804–816 (2003).
[13] Nakamura, T. and Kameoka, H.: Shifted and convolutive source-filter non-negative matrix factorization for monoral audio source separation, *Proc. of 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 489–493 (2016).
[14] Poliner, G. E. and Ellis, D. P.: A discriminative model for polyphonic piano transcription, *EURASIP Journal on Advances in Signal Processing*, Vol. 2007 (2007).
[15] Sigtia, S., Benetos, E., Boulanger-Lewandowski, N., Weyde, T., d'Avila Garcez, A. S. and Dixon, S.: A hybrid recurrent neural network for music transcription, *Proc. of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 2061–2065 (2015).
[16] Sigtia, S., Benetos, E., Cherla, S., Weyde, T., Garcez, A. and Dixon, S.: An RNN-based music language model for improving automatic music transcription, *Proc. of the 15th International Society for music Information Retrieval*, International Society for Music Information Retrieval, pp. 53–58 (2014).
[17] Smaragdis, P. and Raj, B.: Shift-Invariant Probabilistic Latent Component Analysis, Technical Report TR2007-009, MERL - Mitsubishi Electric Research Laboratories, Cambridge, MA 02139 (2007).
[18] Smaragdis, P., Raj, B. and Shashanka, M.: A probabilistic latent variable model for acoustic modeling, *Advances in Models for Acoustic Processing Workshop, NIPS* (2006).
[19] Su, L. and Yang, Y.-H.: Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription, *Proc. of the 11th International Symposium on Computer Music Multidisciplinary Research*, pp. 221–233 (2015).