擦弦楽器の伝達特性を考慮したヴィオラ録音信号のピッチシフトによるチェロ音の合成

吉野 夏樹[1,a)]　長屋 龍司[2,b)]　田中 章[3,c)]

概要：本稿ではヴィオラの録音信号からチェロの音色を持つ音響信号を合成する手法を提案する．多くのヴァイオリン奏者はヴィオラの演奏技術も有するため，この手法によりヴァイオリン奏者 1 人による多重録音で弦楽四重奏曲の録音が可能になる．チェロはヴィオラ同じヴァイオリン属の擦弦楽器であり，かつ調弦がヴィオラの 1 オクターブ下であるため，ヴィオラの録音信号の音高を 1 オクターブ下げることでチェロに近い波形を得る．しかし，ピッチシフト操作のみで生成された楽器音の周波数スペクトルは現実のチェロのそれとはいくつかの点で異なる．この違いを吸収するフィルタ処理を追加で行うことでより本物に近い音色の合成を試みる．

1. はじめに

多重録音が身近な録音方法となったことを受け，近年では複数種類の楽器を演奏できる奏者が多重録音を行うことで少人数での楽曲制作が可能となった．例としてポピュラーミュージックの収録でヴァイオリンとヴィオラ，チェロを使用する場合には，各楽器ごとに奏者を 1 人ずつ集めることなく，1 人で 3 つ全ての楽器を弾ける奏者への参加依頼で十分となる．これには人件費を節約することができるなどのメリットが存在する．

すでにヴァイオリンを弾ける人物が新たにヴィオラの演奏技術を習得するのにかかる学習コストが比較的小さいため，両方の楽器を高水準で演奏できる奏者は少なくない．実際多くのプロ・ヴィオラ奏者はキャリアの途中までヴァイオリン奏者として活動し，音楽学校への入学や就職などのタイミングでヴィオラ奏者に転向している．一方で，ヴァイオリンとチェロの高い演奏技術を併せ持つ奏者は極めてすくない．これはヴァイオリンとヴィオラで楽器の構えかたがほとんど一致するのに対して，チェロはそれらとは異なる構え方をするため学習コストが高いためである．そのためヴァイオリンとヴィオラ，チェロすべてを高水準で演奏できる奏者は極めて少ない．

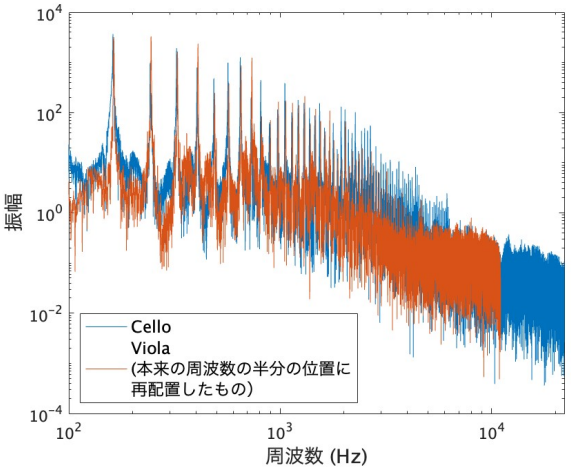

図 1　ヴィオラ録音信号（C 線の D3）とチェロの録音信号（C 線の D2）の振幅スペクトルの比較．比較のためヴィオラの録音信号の振幅スペクトルは元の周波数の半分の位置に配置した．

しかしながら，ヴィオラとチェロの間には多くの類似点がある．図 1 は両者の振幅スペクトルを比較したものであるが，基本周波数から 10^3Hz 以下の低次の倍音にかけてのピークの高さが大まかに一致していることがわかる．この特性を元にした信号処理をヴィオラの録音信号に対して適用することでチェロに類似した音色の合成が期待される．著者らは [1] でピッチシフトと人工的な倍音合成によるチェロに類似する音色合成手法を提案した．当該手法ではピッチシフト単体での音色合成ではナイキスト周波数の半分以上の高周波数領域に問題があることに対する対処を行っている．しかし周波数領域でのスペクトル形状の違いに起因

[1] 北海道大学情報科学院
N14W9, Sapporo-shi, Hokkaido, 060-0814, Japan
[2] 北海道大学工学部
N13W8, Sapporo-shi, Hokkaido, 060-0813, Japan
[3] 北海道大学情報科学研究院
N14W9, Sapporo-shi, Hokkaido, 060-0814, Japan
[a)] n_yoshino@ist.hokudai.ac.jp
[b)] r_nagaya@ist.hokudai.ac.jp
[c)] takira@ist.hokudai.ac.jp

する問題は高周波領域以外にも存在する．チェロの録音信号の振幅スペクトルには楽器の各部品の伝達特性に由来する山や谷が存在するが，ヴィオラの録音信号を単純に 1 オクターブピッチを低くするだけでは再現できないことである．本稿では [1] で提案した手法による合成に加えフィルタ処理を行うことでその問題に対する解決を試みる．

2. STN 分離とピッチシフトによるチェロらしい音色の合成 [1]

ヴィオラ弦のチューニングはチェロのそれの正確に 1 オクターブ上である．そのためヴィオラの録音からチェロに類似した音色を合成する際に第一に必要となる操作は 1 オクターブの音高下降である．音高を α 倍する手法としては，
(1) $1/\alpha$ 倍のサンプリング周波数でのリサンプリング
(2) 音高を保ったまま信号長を α 倍する
という手順を踏むものがある [2]．(1) の処理ではヴィオラの録音の振幅スペクトルを本来の周波数の 1/2 の位置に再配置しチェロの録音の低域成分を再現する．手順 (2) は Time-Scale Modification (TSM) と呼ばれる．

2.1 タイムスケール修正

多くの TSM アルゴリズムは入力信号 $\boldsymbol{x}$ を長さ N の短い分析フレームに分割する．分析フレームは分析ホップサイズ H_{a} でオーバーラップする位置から T 個作成される．

$$x_m[r] = \begin{cases} x[r + mH_{\mathrm{a}}] & \text{for } r \in [-\frac{N}{2}, \frac{N}{2} - 1] \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

m 番目の分析フレーム x_m から後述する方法により合成フレーム y_m を作成し，これらに窓関数 w を乗算したのちに合成ホップサイズ H_{s} でオーバーラップ加算を行うことで時間方向の伸縮を行う．x を伸縮率 $\alpha = H_{\mathrm{s}}/H_{\mathrm{a}}$ でタイムスケール修正した信号 y は，以下の式で定式化される:

$$y[r] = \frac{\sum_m w[r - mH_{\mathrm{s}}]\ y_m[r - mH_{\mathrm{s}}]}{\sum_n w[r - nH_s]}. \quad (2)$$

2.1.1 OLA

合成フレーム y_m を計算する最も素朴な方法は，前後のフレームとの整合性を無視して $y_m = x_m$ としてしまうものである．このアルゴリズムは Overlap Add (OLA) と呼ばれる．OLA は局所的な周期性を無視した合成を行うため調波音成分の時間伸縮には向かない．しかしながら，分析フレームサイズを小さく設定することで，打撃音やトランジェント成分のような短く周期性を持たない信号に対して良好に動作する．

2.1.2 WSOLA

Waveform Similarity-based Overlap Add (WSOLA) [3] は入力信号から分析フレームを切り出す位置をわずかなサンプル数 $\Delta_m \in [-\Delta_{\max}, \Delta_{\max}]$ ずらす手法であるり，分析フレームは

$$x'_m[r] = \begin{cases} x[r + mH_a + \Delta_m] & \text{for } r \in [-\frac{N}{2}, \frac{N}{2} - 1] \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

とあらわされる．Δ_m は x_m が x_{m-1} とオーバーラップする部分において x_{m-1} の周期性を受け継ぐように最適化される．WSOLA は分析フレーム内の最も目立つ周期性を保存するため単一音程の調波音に対して良好に動作するが，複数の音程が混ざっている楽器音には意図通りに動作しない．また，打撃音やトランジェント成分の存在が考慮されていない．

2.1.3 PV-TSM

WSOLA の弱点を補ったものとして Phase Vocoder を基にした TSM（PV-TSM）[4] がある．Phase Vocoder は短時間フーリエ変換（Short-Time Fourier Transform: STFT）の周波数解像度の粗さを改善する．虚数単位を i で表し，式 (1) を用いて x の短時間フーリエ変換を

$$X[m,k] = \sum_{r=-N/2}^{N/2-1} x_m[r]w[r]\exp(-2\pi ikr/N) \quad (4)$$

と表すと，Phase Vocoder の出力は

$$X^{\mathrm{Mod}}[m,k] = |X[m,k]|\exp(2\pi i\phi^{\mathrm{Mod}}[m,k]) \quad (5)$$

で定められる．なお ϕ^{Mod} は合成位相 [4] を表す．PV-TSM では合成フレーム y_m を X^{Mod} の逆短時間フーリエ変換

$$y_m[r] = \frac{1}{N}\sum_{k=0}^{N-1} X^{\mathrm{Mod}}[m,k]\exp(2\pi ikr/N) \quad (6)$$

として計算する．最終的な出力は次式で表される:

$$y[r] = \frac{\sum_m w[r - mH_{\mathrm{s}}]y_m[r - mH_{\mathrm{s}}]}{\sum_n (w[r - nH_{\mathrm{s}}])^2}. \quad (7)$$

2.1.4 複数の TSM アルゴリズムの組み合わせ

PV-TSM は純粋な正弦波とその組み合わせに非常に効果的であり，入力信号にトランジェント成分が含まれている場合でも耳障りな副産物を生成しない．しかしトランジェントの存在感が薄れてしまうという問題点が存在する．これに対して [1] では前処理として STN（Sines-Transients-Noise）分離 [5] を行い，それぞれの成分に対して 2 倍のオーバーサンプリングを行ったのち

- 調波音成分を PV-TSM,
- トランジェント成分を OLA,
- 雑音成分を WSOLA

によりタイムスケール修正を行う手法を提案した．なお，STN 分離とは入力信号を調波音成分（Sines）と打撃音成

分（Transients），雑音成分（Noise）の計 3 つの成分への分離を行うタスクである．ヴァイオリン属の楽器音は調波音成分とトランジェント成分に加え，周期性をもつ雑音成分（非整数倍音を多く含む周期信号）から構成されることがわかっている [6] ため，これらを別々に処理することには意味がある．

2.2 STN 分離

STN 分離のアルゴリズムは複数提案されている [7] [8] が，本研究では二段階のバイナリマスキングによる方法 [5] を使用する．この手法では，まず所与の音響信号の短時間フーリエ変換 $X[m,k]$ を計算する．次に振幅スペクトログラム $|X[m,k]|$ に対して時間 (m) 方向，及び，周波数 (k) 方向にメディアンフィルタを適用したものを計算することで，調波音を強調した振幅スペクトログラム

$$X_h[m,k] = \mathcal{M}\big[|X[m-\frac{L_h}{2}+1,k]|,\ldots,|X[m+\frac{L_h}{2},k]|\big] \tag{8}$$

と打撃音を強調した振幅スペクトログラム

$$X_p[m,k] = \mathcal{M}\big[|X[m,k-\frac{L_p}{2}+1]|,\ldots,|X[m,k+\frac{L_p}{2}]|\big] \tag{9}$$

を得る．ここで $\mathcal{M}(\cdot)$ は引数の中央値を返す関数であり，L_h と L_p は時間方向および周波数方向のメディアンフィルタ長である．強調された振幅スペクトログラム $X_h[m,k], X_p[m,k]$ をもとに調波音らしさ

$$R_h[m,k] = \frac{X_h[m,k]}{X_h[m,k]+X_p[m,k]} \tag{10}$$

と打撃音らしさ

$$R_p[m,k] = 1 - R_h[m,k] = \frac{X_p[m,k]}{X_h[m,k]+X_p[m,k]} \tag{11}$$

を計算する．式 (10), (11) から調波音成分，打撃音成分および雑音成分の分離を行うためのマスク

$$S[m,k] = \begin{cases} 1 & \text{if } R_h[m,k]/R_p[m,k] > \beta \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$T[m,k] = \begin{cases} 1 & \text{if } R_p[m,k]/R_h[m,k] > \beta \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

$$N[m,k] = 1 - S[m,k] - T[m,k] \tag{14}$$

を構成し，これらを入力信号のスペクトログラムに乗算して分離を行う．なお β は分離のための閾値であり，アルゴリズムのユーザが指定する．

マスキングによる分離は短時間フーリエ変換の分析窓長に大きく依存することが知られている [5]．具体的には，長い窓長は周波数分解能が高いため調波音の分離に優れ，一方で短い窓長はトランジェント成分の抽出に適している．このトレードオフ関係を解消するため，分離は（1 段目）長い分析窓長による STN 分離 と（2 段目）短い窓長による STN 分離 の 2 段階で行われる．1 段目の分離では調波音成分とそれ以外に分離される:

$$x_s[r] = \mathrm{ISTFT}\left[S_1[m,k]X[m,k]\right] \tag{15}$$

$$x_{\mathrm{res}}[r] = \mathrm{ISTFT}\left[(T_1[m,k]+N_1[m,k])X[m,k]\right]. \tag{16}$$

ここで S_1, T_1, N_1 は 1 段目の分離のために計算されたマスクであり，ISTFT は逆短時間フーリエ変換を表す．続いて，2 段目の分離として (16) に対して短い分析窓長を用いた分離を行う:

$$x_t[r] = \mathrm{ISTFT}\left[T_2[m,k]X_{\mathrm{res}}[m,k]\right] \tag{17}$$

$$x_n[r] = \mathrm{ISTFT}\left[(S_2[m,k]+N_2[m,k])X_{\mathrm{res}}[m,k]\right]. \tag{18}$$

なお，S_2, T_2, N_2 は 2 段目の分離のために計算されたマスクであり，X_{res} は x_{res} の短時間フーリエ変換である．図 2 に 2 段階の分離の流れ図を示す．

2.3 高周波成分の補完

先に説明した 1 オクターブのピッチ下降は 2 倍のサンプリング周波数でのリサンプリングを伴うため，元信号のナイキスト周波数の半分以上からナイキスト周波数にかけては原理的にまったくパワーを持たない．しかし現実のチェロはより高い周波数（少なくとも 48kHz 近辺）まで倍音成分を持つため，この点で本来のチェロの音色との違いがある．楽器音の音色はその周波数スペクトルの形状に強く影響を受けるため，それらの補完は楽器音の再現を行う上で重要となる．[1] では人工的な倍音付加と原信号の高域成分を用いて高周波数領域での問題解決を試みた．

3. 楽器の周波数特性を考慮したフィルタ処理

楽器音の音色を決定する要素は様々である．先に述べたようにチェロの楽器本体の大きさはヴィオラのおよそ倍である．特に振動源となる弦の長さが，チェロはヴィオラの倍程度であるため，ヴィオラの録音信号に対して 2 倍のリサンプリングを行うことでチェロに類似する音色の合成が期待される．しかし駒に代表されるようにヴィオラとチェロの間で単純な相似関係にない構成部品が複数存在する．それらの部品の伝達特性の違いがあるため，ヴィオラの録音信号に対し 1 オクターブのピッチ下降を行うだけでは振幅スペクトルの再現は困難である．

本稿では，それらの伝達特性の違いを吸収するフィルタ処理手法を提案する．このフィルタ処理は理想的には複数のチェロとヴィオラの伝達特性を計測し，それらの平均値などから設計するべきである．チェロについては，伝達特性の詳細な分析を行った先行研究 [9] が存在するが，ヴィオラについては著者らの知る限り存在しない．そのため本

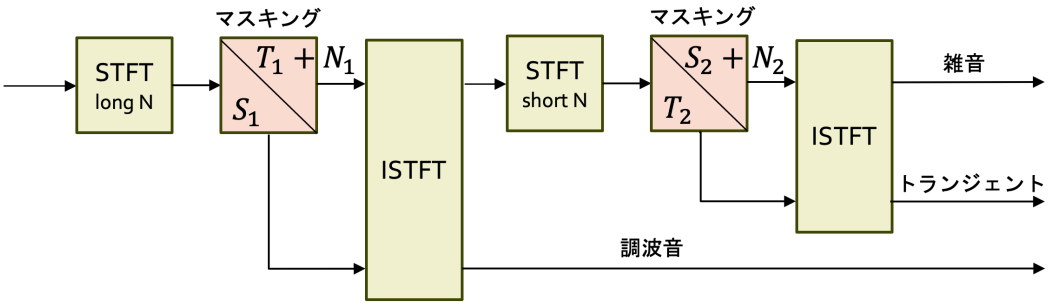

図 2 2 つの分析窓長を組み合わせた 2 段階での STN 分離のブロックダイアグラム [1]

研究ではチェロとヴィオラの録音信号を元にフィルタの設計を行った．

チェロとヴィオラの録音信号は University of Iowa Electronic Music Studios によって提供されている楽器音のデータベース[*1]のものを使用した．なお，チェロとヴィオラでは同一音程を複数の弦で演奏能である一方，弦楽器を演奏する際には駒から遠い場所を押さえる機会が多く，またチェロとヴィオラの音色は駒に近い場所を押さえて演奏した場合に類似した音色となることが知られている．これらの理由からチェロ，ヴィオラともに各弦の最低音から 8 半音の録音信号を使用した．

これらの録音から所望のフィルタを設計する手順を次に示す．

(1) [1] によりヴィオラの録音信号からチェロに類似する音色を合成
(2) 手順 (1) で合成した信号の Yule-Walker 法により音程ごとに振幅スペクトルの包絡線を計算
(3) 手順 (2) のスペクトル包絡の平均を計算
(4) 手順 (2),(3) と同様にチェロの録音信号のスペクトル包絡及びその平均を計算
(5) 手順 (4) で計算したスペクトル包絡を手順 (3) で計算したスペクトル包絡で周波数成分ごとに除算
(6) 手順 (5) で計算したスペクトル包絡の比を周波数領域での利得とする線形位相フィルタを計算

なお，手順 (2) と (4) でスペクトル包絡を計算する際の AR モデルの次数は 18 とし，高速フーリエ変換には 1024 点を使用した．

[1] により合成した信号に対して以上の手順により計算されたフィルタを適用したものが，フィルタ処理前と比べて実物に近づいているかを 2 つの例を用いて確認する．図 3 は [1] を用いてヴィオラ音 (D2) から合成した信号に対する提案フィルタ処理の前後の振幅スペクトルである．なお，

[*1] ヴィオラの録音信号は https://theremin.music.uiowa.edu/MIS-Pitches-2012/MISViola2012.html，チェロの録音信号は https://theremin.music.uiowa.edu/MIS-Pitches-2012/MISCello2012.html から入手可能である．

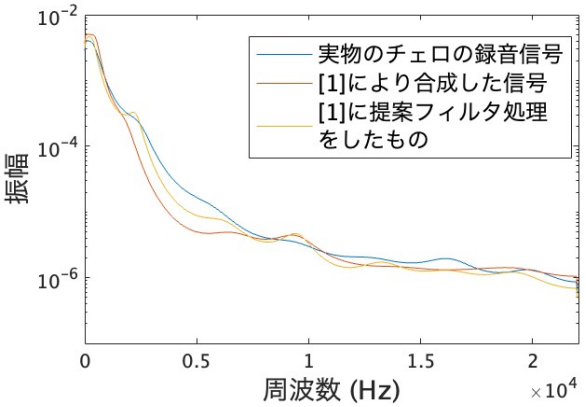

図 3 [1] を用いてヴィオラ音 (D3) から合成した信号に対する提案フィルタ処理の適用前後での振幅スペクトルの比較

一連の処理の目標値であるチェロ音 (D2) の振幅スペクトルの包絡線についても併せて表示している．この例では提案フィルタ処理によって，およそ 2kHz から 8kHz にかけてのスペクトルの違いを補うことに成功していることが確認できる．また提案フィルタ処理によってより本物に近い音色を得たことを聴感上でも確認した．次に同様の比較をヴィオラ音 (E5) から合成した信号に対して行った（図 4）．こちらは先ほどの例とは異なり，スペクトル包絡の形状を目標であるチェロのものに近づけることができていない．聴感での確認も行ったが，期待通りの結果が得られたとは言えない．

以上のような例を複数確認したところ，問題点として 2 つのことがらが確認できた．1 つ目は高い音程のヴィオラ音に対して [1] の手法を適用すると，高周波成分が過剰に生成されるということ，2 つ目は演奏する弦によっても周波数スペクトルの特性が大きく変化するため，提案フィルタ処理のような単純な処理では対応が困難という点である．

4. おわりに

本研究ではヴィオラの録音信号からチェロに類似する音色を持つ信号を合成する際に，より実際のチェロの録音信号に近い音色を得るためのフィルタ処理を提案した．いく

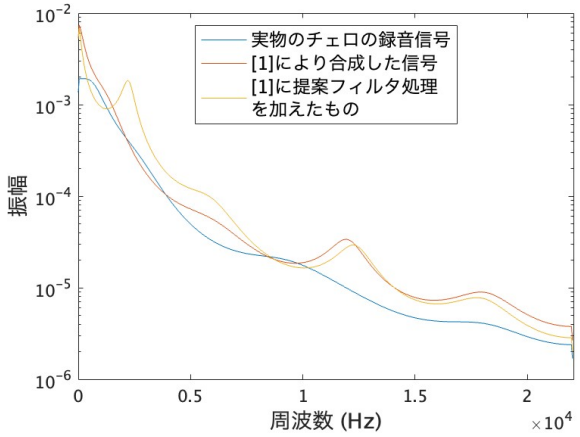

図 4 [1] を用いてヴィオラ音 (E5) から合成した信号に対する提案フィルタ処理の適用前後での振幅スペクトルの比較

つかの例ではこの処理は期待通りの結果を示したが，提案フィルタ処理が有効となるヴィオラ録音信号の音域が限定的であることを確認した．任意の音程の録音信号を取り扱うことができる音色合成手法の開発が今後の課題である．

謝辞 本研究の一部は JST 次世代研究者挑戦的研究プログラム JPMJSP2119 の支援を受けたものである．

参考文献

[1] N. Yoshino, and T Akira “Cello-like Sound Synthesis from Viola Recordings Using Pitch Shifting and Harmonic Generation”, IEICE Technical Report, EA2023-114 (2024-03)
[2] Driedger, Jonathan, and Meinard Müller “A Review of Time-Scale Modification of Music Signals”. Applied Sciences 6, no. 2:57. 2016
[3] W. Verhelst, M. Roelands, “An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech”, In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Minneapolis, MN, USA, 27–30 April 1993.
[4] J. Laroche and M. Dolson, ”Improved phase vocoder time-scale modification of audio”, in IEEE Transactions on Speech and Audio Processing, vol. 7, no. 3, pp. 323-332, May 1999.
[5] J. Driedger and S. Disch, “Extending harmonic-percussive separation of audio signals”, 15th International Society for Music Information Retrieval Conference (ISMIR 2014) pp.611–616, Oct. 2014
[6] Matthias Demoucron, “On the control of virtual violins - Physical modelling and control of bowed string instruments”, Acoustics [physics.class-ph], Université Pierre et Marie Curie - Paris VI; Royal Institute of Technology, Stockholm, 2008.
[7] L. Fierro and V. Välimäki, “Towards Objective Evaluation of Audio Time-Scale Modification Methods”, In Proceedings of the 17th Sound and Music Computing Conference. Axea sas/SMC Network, pp.457-462, 2020.
[8] R. Füg, A. Niedermeier, J. Driedger, S. Disch and M. Müller, “Harmonic-percussive-residual sound separation using the structure tensor on spectrograms”, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 445-449, 2016.
[9] A. Zhang, J. Woodhouse and G. Stoppani “Motion of the cello bridge”, Journal of the Acoustic Society of America, vol. 140, issue 4, pp.2636–2645 (2016)