研究報告

スピーチミュージックにおける「声の音色」再構築の試み -機械学習モデル RAVE を用いた音響変換による実践的研究-

English Title: Reconstructing “Vocal Timbre” in Speech Music Practical Study on Timbre Transformation Using RAVE

藤江 明香里
Akari FUJIE
名古屋市立大学芸術工学部
School of Design and Architecture,
Nagoya City University

松宮 圭太
Keita MATSUMIYA
名古屋市立大学大学院芸術工学研究科
Graduate School of Design and Architecture,
Nagoya City University

概要

本研究は、録音された発話を素材とする作曲実践を整理した Cathy Lane（2006）の枠組みを参照しつつ、同領域を音響的観点から捉え直し、「声の音色」の再構築を目指す制作研究である。本研究では、この種の実践を研究上の操作的定義として「スピーチミュージック」と呼ぶ。声を学習した機械学習モデル RAVE（Real-Time Audio Variational Auto-Encoder）を用い、意味をもたない器楽音にフォルマント様帯域や息成分、摩擦的ノイズといった声的特徴を付与することで、器楽音を「声らしき音声」へ変換する手続きを設計した。さらに、録音した話し声の韻律を MIDI 化して器楽で再合成し、その出力を RAVE で声へ還元する「声→楽器音→声」変換を導入し、声の旋律的時間構造を保持しつつ言語的意味を欠落させた声の音色を生成する方法を示した。以上の知見は、音響映像作品《奪われた声》に統合され、声が言語的意味を失い、音色へ解体され、再び声へ回帰する過程を段階的に構成する方法として具体化された。

This study reconsiders compositional practices that use recorded speech or the spoken word, drawing on Cathy Lane’s (2006) systematic overview of works and compositional techniques, and explores reconstructing vocal timbre from an acoustical viewpoint. In this paper, we use the term speech music as an operational label for such practices to clarify the analytical and practical scope.Using a voice-trained machine-learning model RAVE (Real-Time Audio Variational Auto-Encoder), we designed a procedure that transforms non-semantic instrumental sounds into “voice-like” audio by imprinting vocal attributes such as formant-like bands, breath components, and fricative noise. We further introduced a “voice → instrument → voice” process in which the prosody of recorded speech is converted into MIDI, resynthesized by instruments, and then mapped back to vocal timbre via RAVE, yielding timbres that retain melodic-temporal structure while losing linguistic meaning. These findings are integrated into an audiovisual work, *Ubauwareta Koe* (*The Stolen Voice*), which stages a gradual collapse of meaning, a decomposition into timbre, and a return to vocality.

1. はじめに

1.1. 研究背景

筆者は、MC として 8 年間、ナレーターとして 2 年間の活動経験を有している。言葉を「どのように音響化するか」を探求する過程で、話し声の韻律的特徴と音楽的パラメータとのあいだに、構造的な対応関係が存在するのではないかと考えるに至った。本研究は、この気づきを出発点としている。

1.2. 録音発話を用いる作曲実践の整理と本研究の位置づけ

本節では、録音された発話を素材とする作曲実践を、本研究の対象領域として定める。Cathy Lane（2006）はこの領域における作品類型と作曲操作を整理し、録音された発話を用いる作品に見られる操作を 19 の技法として列挙している。本研究ではこの整理を参照しつつ、録音発話を主要素材として編集・抽出・配置・変換

等の作曲操作により音楽的構造を形成する実践を、操作的定義として「スピーチミュージック」と呼ぶ。

本研究では、スピーチミュージックの音響的側面——すなわち「声の音色」——に注目し、機械学習による音色変換を通じて、この表現領域を拡張することを試みる。

1.3. 研究目的

本研究の目的は、機械学習による音色変換技術を用いて、意味をもたない音響素材を「声の音色」へと変換し、従来のスピーチミュージックが扱ってきた「意味／音響の緊張関係」を新たに拡張することである。

2. RAVE による音色変換

近年、機械学習技術の発展によって、声と楽器音との関係はさらに拡張されている。Google Magenta チームによる DDSP（Differentiable Digital Signal Processing）は、歌声を楽器音にリアルタイムで変換できることを示した。また、IRCAM が開発した機械学習モデル RAVE（Real-Time Audio Variational Auto-Encoder）は、高品質な音色変換をリアルタイムで実現するフレームワークとして注目されている。

作曲家・大久保雅基による音楽劇《声のゆくえ》(2022) はこの技術を用いた作品の一例である。この作品では、俳優が発する言葉をその場で録音し、DDSP によって構築された機械学習モデルが、声をヴァイオリンの音色に変換している。

本研究では、DDSP ではなく RAVE を用いた。RAVE は、入力音の音色を別の音色へ変換することを可能にする。たとえば、高次倍音が少ないピアノ音を入力音とし、高次倍音が多い発話音声を学習したモデルで変換した場合のフォルマント帯域を図 1 に示す。変換前のピアノ音では、楽器固有の倍音構造が時間的に安定しているが、RAVE を通した後のピアノ音では、この倍音構造が声のフォルマント帯域に対応するように再配置され、「声らしい共鳴」を帯びるようになったことが確認できる。

このように、意味をもたない楽器音を「声」として提示することで、聴覚的意味の再構築を促すことができる。ここで「声の音色の再構築」とは、RAVE による声の音響的特徴（フォルマント帯域や倍音構造など）の再現として定義する。

3. 制作内容

本研究の知見は、音響映像作品《奪われた声》として統合された。本作は童話『人魚姫』における魔女と人魚姫の場面を舞台に、「声が奪われる」出来事を、言

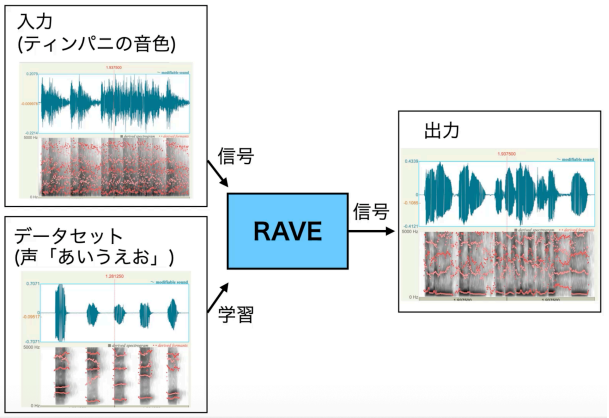

図 1: RAVE による音色変化のイメージ（音声分析ソフト praat を用いて作成）

語意味と音響質感の両面から段階的に崩壊させる構成を採用した。奪われた声が意味（言語内容）を失い、音色としての声が解体され、最終的に旋律のみが残存した楽器音として提示されるよう設計し、その後楽器音から再び声の音色へ回帰する過程を導入した。

3.1. MIDI による「声の音色→楽器音」変換

筆者の制作において、録音したセリフ「わたしのこえ」の音程やリズムを、Cubase（Steinberg 社の DAW。録音・編集・作曲・ミックスを一括して行える音楽制作ソフト）のピッチ検出機能によって抽出し、同じ高さとリズムをもつ MIDI データへと変換した（図 2）。これをもとに、楽器音を再生することで、スティーブ・ライヒ《Different Trains》（図 3）と同様に、話者の声の旋律を楽器がなぞる構造をもたせた。さらに、楽器音と録音された声を同時に再生することで、ヘテロフォニーを形成した。

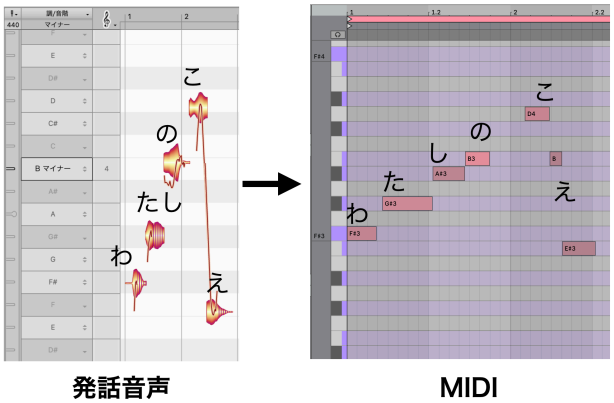

図 2: 声の MIDI ノート化

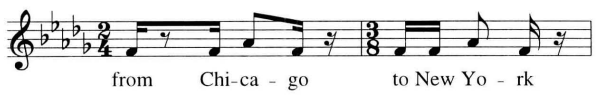

図 3: スティーブ・ライヒ《Different Trains》(1988) より抜粋 © Boosey & Hawkes Music Publishers Ltd.

本作では、声の旋律を模倣させる対象楽器としてマリンバを採用した。マリンバは明確なピッチを持ち、アタックが速く、減衰が比較的澄んだ打楽器であるため、声の連続的な音色変化とは異なる時間特性を示す。したがって、録音された声とマリンバを同時に提示した際に、両者の差異が聴取上明確になり、声と非声的な楽器が同一旋律を共有しているという緊張関係を作りやすい。結果として、声と楽器音が同一の旋律輪郭を保ったまま並走するヘテロフォニーが形成され、声の崩壊と、旋律の存続を同時に提示することができた。[1]

3.2. RAVE モデルを用いた「楽器音→声」変換

3.2.1. RAVE モデルの作成

本研究では、Real-time Audio Variational Auto-Encoder (RAVE：IRCAM が開発したリアルタイム音色変換用の深層学習モデル) を用い、楽器音を学習させることで楽器音を声の音色に変換させて制作を行なった。RAVE の学習には、標準設定ファイルである「v1.gin」を使用した。サンプリングレートは 48 kHz、バッチサイズ 8、最大訓練ステップ数 400,000、1 サンプル長は約 2.7 秒 (131,072 samples) である。特別なデータ分割は行わず、すべての音声データを訓練用に用いた。学習に用いたデータセットは、『星の王子さま』の朗読音声 (総計 2 時間) である。このデータセットには、破裂音・摩擦音など多様な子音を含む日本語音声が幅広く収録されており、モデルは日本語の音響的および音韻的特徴を包括的に学習したと考えられる。

3.2.2. RAVE モデルによる音色変換と段階的切替 (擬似モーフィング)

作成した RAVE モデルを用いて、複数の器楽パートの一部を RAVE で声の音色へ変換したうえで、DAW 上で段階的に原音へ戻すことで、聴感上のモーフィング (連続的な変化) を擬似的に実現した (図 4). 制作手順は以下の通りである。

[1] スティーブ・ライヒの《Different Trains》(1988) は、弦楽四重奏とテープのための作品である。録音されたインタビューの言葉や汽笛などの旋律やリズムが、弦楽器によって模倣されている。

1) MIDI データにより、flute / trombone / trumpet / clarinet / viola / violin / doublebass の各パートの旋律を作成した。

2) このうち trombone / viola / violin の 3 パートについて、当該楽器音を RAVE モデルに入力し、声の音色へ変換した。

3) DAW 上で、上記 3 パートが「RAVE 変換された声的音色」から「原楽器音」へ入れ替わるように配置した。このとき、全パートが同時に切り替わるのではなく、時間差をもって順次切り替わるよう設計することで、変換点を分散させ、より自然なモーフィングとして知覚されることを狙った。

このようにして、複数の楽器音の旋律と、複数の声へと変換された旋律とが重なり合うポリフォニーを構成した。モノフォニーではなくポリフォニーの構成を採用したことで、単純な音量変化によるものであってもより自然なモーフィングを実現できた。

また、声と弦楽器は、ともに連続的なピッチ変化に基づく音高変化を共有し、聴取上近似した印象を生みやすい。そのため、切替対象として violin / viola を含めることで、RAVE 変換音と原楽器音のあいだに過度な断絶が生じにくい条件を作り、違和感の少ないモーフィングを得ることを試みた。

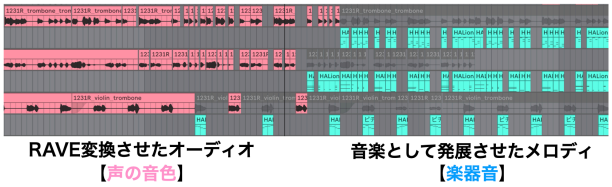

図 4: RAVE を用いた「声→楽器音」のなめらかな移行

3.3. RAVE モデルを用いた「声→楽器音→声」変換

3.3.1. 声の再合成

本研究では、同一の RAVE モデルを用いて、元の声を直接加工するのではなく、いったん楽器音を経由させたうえで声の音色へ還元する手続きを設計した (図 5)。手順は以下の通りである。

1) 録音したセリフ「わたしのこえ」の音程とリズムをピッチ検出機能によって抽出し、同一の高さとリズムを持つ MIDI データへ変換して、楽器音源へ割り当てた。

2) 生成した楽器音を、発話音声を学習させた RAVE モデルに入力し、声の音色へ変換した。

このように、声をいったん「旋律情報としての MIDI」へ抽出し、楽器音として再合成した後に、再び声の音色へ還元することで、元のセリフ「わたしのこえ」と似た時間構造を持ちながらも、言語的な意味は失われ、響き（声質）も異なる「声らしき音声」を得ることができた。したがって本手続きは、声の複製ではなく、「声の再合成」として位置づけられる。

図 5: RAVE を用いた「声→楽器音→声」変換

3.3.2. 入力楽器による *RAVE* 変換後の声質の差異

本シーンでは、同一旋律を 6 種の楽器（violin / trombone / flute / doublebass / trumpet / Breath Noise）でそれぞれ再生した音源を RAVE により声の音色へ変換し、6 種類の異なる響きを持つ音声を制作した（図 7）。これらを「声が崩壊するシーン」から「声が完全に意味を失い、音色となるシーン」への切替点に配置し、単一の変換音ではなく、複数の声質が交代・重畳するテクスチャとして提示した。

Praat で声質指標を比較した結果を図 6（表 1）に示す。器楽由来の 5 条件では周期性指標 HNR が約 10.45–21.57 dB の範囲で条件間差を示し、trombone・flute・doublebass 条件は高い HNR（約 18.8–21.6 dB）を示した一方、violin・trumpet 条件は約 10–11 dB に留まり、相対的にノイズ混入が大きい可能性が示唆された。CPPS はいずれも正の値を示し、voice-like な周期構造は保持されつつも、調波優勢さの度合いに差があることが確認された。さらに高域寄与について、LTAS から 500 Hz・3000 Hz・4000 Hz のスペクトルレベルを取得し、500 Hz を基準とした差分（高域比）として評価した。その結果、条件間で高域寄与の程度に差がみられ、入力楽器の倍音構造の違いが変換後のスペクトル傾きに影響し、「母音優勢に聴こえる出力」と「摩擦・破裂に相当する成分が強調される出力」が混在する傾向が観察された。

一方、Breath Noise 条件は周期性について HNR が −0.137 dB、CPPS が 2.453 dB と低い値域に留まり、調波優勢な周期構造が明瞭な音というより、ノイズ成分が相対的に支配的で周期性が弱い音であることが示唆される。この Breath Noise 変換音は、楽器音主体の場面から“もとの声”へ回帰する際のトリガーとして配置した。単に「声っぽい音」を再提示するのではなく、周期性を強く固定しない中間的素材として Breath Noise 変換音を用いた。

また、上記の HNR/CPPS/LTAS による比較に加え、制作上の判断へ接続するため、Praat で得た指標に基づき各条件を 2 次元平面上に配置し、条件間関係を可視化した声質マップを作成した（図 8）。本図では、横軸に HNR（全区間平均）をとり、値が高いほど調波成分が相対的に優勢で、ノイズ混入が相対的に少ない方向を示す。縦軸には LTAS 由来の高域比（LTAS(3000)−LTAS(500)）をとり、差分が 0 に近いほど高域寄与が相対的に大きく、スペクトル傾きが浅い方向を示す。なお、本図は各条件の特徴を指標空間で整理するための補助的表現であり、聴感上の印象を直接測定したものではない。

本図には、5 条件に加えて比較参照として「わたしの声？」という実際の発話音声も配置した。ここで、元音声は「条件間比較（5 条件の平均値比較）」の対象ではなく、制作上の素材選定を検討するための参照点として用いた。その結果、violin および doublebass は、他条件に比べて元音声の近傍に位置し、とくに高域寄与の程度において乖離が小さい傾向を示した。そこで本研究では、録音音声と同時提示した際にスペクトル傾きの差異が過度に強調される可能性を相対的に抑えられる素材として、この 2 条件を録音音声が崩壊していくシーンに混合した。ただし、ここでいう“混合の適性”は本研究の制作上の判断であり、心理音響実験等によって検証したものではない。

以上の分析から、RAVE 変換音は一様に「声らしく」なるのではなく、入力素材の倍音構造やノイズ性が、変換後の周期性とスペクトル傾きに反映され、結果として複数の声質が生成されることが確認された。そこで本研究では、この差異を単なる誤差として扱うのではなく、声が崩壊して音色へ移行する局面でのテクスチャ形成に積極的に利用し、Breath Noise を中間的素材として配置しつつ、violin・doublebass を元音声との同時提示に適した混合素材として用いた。以上より、本節で得られた指標空間の整理は、変換音の選別と配置の判断を支える設計図として機能したといえる。

4. おわりに

本研究で使用した RAVE モデルは「楽器音を学習させたモデル」ではなく「声を学習させたモデル」である。ゆえに本研究の要点は、楽器音の旋律・時間構造を保持したまま、出力側に声に固有な音響特徴（フォルマント様帯域、息成分、摩擦的ノイズ等）を付与し、

	HNR	CPPS	LTAS500	LTAS3000	LTAS4000
violin	11.257	6.725	51.130	26.218	29.603
tronbone	18.779	8.042	54.185	27.816	27.199
flute	19.349	8.162	54.042	17.227	22.605
doublebass	21.566	8.671	47.185	23.326	18.911
trumpet	10.454	8.610	46.022	32.750	29.157
breathnoise	-0.137	2.453	15.464	26.595	31.824
元音声	10.058	5.592	52.448	30.347	29.753

図 6: 表 1 声質指標の比較（HNR(dB)/CPPS(dB)/LTAS）

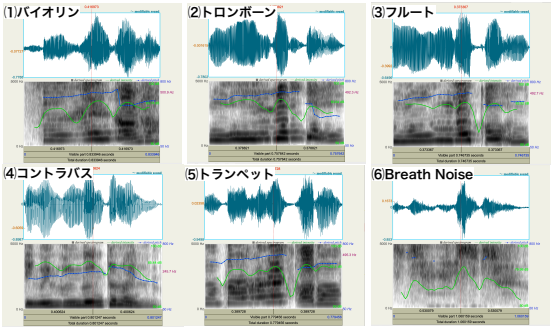

図 7: RAVE を用いた 6 種類の変換後音声

楽器音を「声の音色」へ変換する点にある。従来のスピーチミュージックでは「声→楽器音」方向の技法が中心であったが、本研究はその逆方向「楽器音→声」を機械学習により実装し、十分に体系化されてこなかった手続きを作曲・音響表現の方法として明確化した。また、声の旋律抽出→器楽化→声学習 RAVE による還元、という“楽器音色を媒介にした声の再変換”は、意味だけを欠落させた非人称的な声を構成する手段として位置づけられる。

5. 今後の展望

5.1. モデル設計の高度化

破裂音・摩擦音・破擦音・弾音など子音グループごとに独立した RAVE モデルを訓練し、複数モデルを使い分ける戦略が有効と考えられる。子音は短時間・非定常で知覚上の鍵となる一方、大量学習では埋没しやすいため、ターゲットを絞ったデータ設計が重要である（筆者は/p/・/t/・/k/のみ抽出した約 12 時間データで学習を進行中である）。

5.2. 音響分析の精緻化

同一旋律・同一レベル条件で入力楽器を変えた際の出力の遷移を、スペクトル包絡の帯域集中、無声区間やノイズ帯域の相対量、時間変動などで指標化し、「母

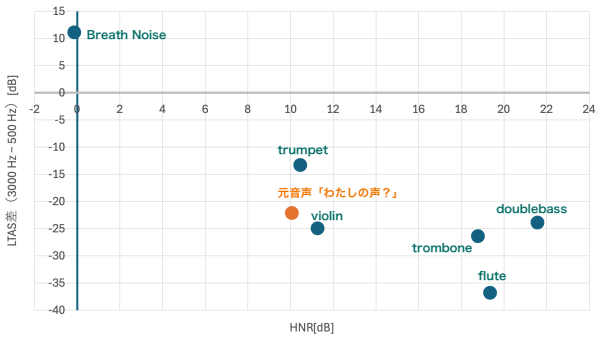

図 8: HNR× 高域比（LTAS(3000)-LTAS(500)）による声質マップ

音優勢／子音優勢」の印象をより再現性高く比較する余地がある。

5.3. 限界への対処

高音域で雑音化して変換が破綻する例がみられ、doublebass は高音域破綻を避けるため 1 オクターブ下げて変換した。学習データの F0 レンジが狭い場合に入力高音域が想定外となる可能性があるため、ピッチシフト等による分布拡張を含め、声らしさを損なわず音域適応性を得る学習設計が論点となる。

5.4. 制作応用

「意味が欠落しても声らしさが立ち上がる」変換を、作品制作のみならず音響演出・サウンドデザインへ展開する可能性がある。たとえば、意味をもたない環境音・楽器音へ声の身体感覚を付与して「誰かがいる」感覚を立ち上げる、あるいは台詞・ナレーションを意図的に“非人称化”して空間に配置するなど、主体の在／不在を音色操作として扱える。Max/MSP と nn~を介した実装は、固定作品だけでなくライブエレクトロニクスやインスタレーションへも拡張しやすく、変換プロセス自体を作曲操作として組み込む設計論（どの段階で周期性を落とし、どの段階で高域ノイズを増やすか等）を、制作研究として継続できると考える。

6. 参考文献

Lane, Cathy. 2006. “Voices from the Past: Compositional Approaches to Using Recorded Speech.” *Organised Sound* 11(1), 3–11. Cambridge University Press.

Engel, J., Hantrakul, L., Gu, C., & Roberts, A. 2020. “DDSP: Differentiable Digital Signal Processing.” *International Conference on Learning Representations (ICLR)*.

Caillon, A., & Esling, P. 2021. “RAVE: A variational autoencoder for fast and high-quality neural audio synthesis.” *arXiv preprint* arXiv:2111.05011.

Reich, Steve. 1988. *Different Trains: For String Quartet and Tape*. Boosey & Hawkes.

音楽劇 声のゆくえ. (n.d.). motokiohkubo.net. 最終閲覧日 2025 年 10 月 14 日.

7. 著者プロフィール

藤江 明香里 (Akari FUJIE)

名古屋市立大学芸術工学部在学中。神戸薬科大学を中途退学後、人間の声を音楽的素材として捉える表現に関心を持ち、制作活動を行っている。中部地方を拠点に MC・ナレーターとしても活動し、CM、VP、映像出演、トークショー MC など多様な現場で話し声を扱ってきた経験を持つ。話し声の韻律的特徴と音楽的パラメータとの構造的対応関係、ならびに声と楽器音の音響的関係に着目し、意味を担う声が音へと移行・変質する過程を探るアプローチで作品を制作している。

松宮 圭太 (Keita MATSUMIYA)

名古屋市立大学芸術工学研究科准教授。サウンドデザイン／作曲。環境音・電子音響と器楽の統合、振動体を用いたハイブリッド楽器の研究制作に従事。主なテーマは「自然音の記録と再解釈」「演奏と機械処理の関係」「発音源認知の揺らぎ」。
