研究報告

音声波形画像のエッジ情報を用いた音声の生成
Generation of sound using edge information of sound waveform images

木村 凪
Nagi KIMURA
九州大学芸術工学部
School of Design, Kyushu University

城 一裕
Kazuhiro JO
九州大学芸術工学研究院
Faculty of Design, Kyushu University

概要

本研究では，音声波形画像から音声を生成する手法を提案する．Python 上で OpenCV を活用し，音声波形画像から得られるエッジ情報を抽出し，取得したエッジ座標の x 座標を時間，y 座標を振幅に変換することで音声を生成する．フォノトグラフによる音の視覚的な記録，トーキーのサウンドトラック，フォノトグラフに記録された歌声の復元，という歴史的な事例を踏まえ，図形と線で表される波形から音を生成するプロセスを検討する．本研究における波形から音声を生み出すプロセスを通して，音楽制作への活用，自らの手による音の形成，音声の波形としての保存，について考察する．

1. 序論

1.1. はじめに

本研究は，ジグソーパズルの組み立て過程から着想を得て，音声波形の断片をパズルのように組み合わせていくことは出来ないかという発想の元，音声波形の画像から，音声を生成することを目標としたものである．本項では，Python 上で OpenCV を活用して，音声波形画像から得られるエッジ情報を抽出し，取得したエッジ座標の x 座標を時間，y 座標を振幅に変換することで音声を生成する手法を提案する．以下本項では，これまでにも多くのアーティストやエンジニアにより試みられてきた，画像と音との変換に関わる事例の中から，フォノトグラフによる音の視覚的な記録，トーキーのサウンドトラックを用いた図形からの音の生成，フォノトグラフに記録された歌声の復元，について説明する。

1.2. フォノトグラフ

フォノトグラフは，1857 年にフランスの技師，エドワール＝レオン・スコット・ド・マルタンヴィルによって発明された，音声の振動を視覚的な波形として記録する装置である．この装置は空気振動としての音を振動膜によって捉え，その振動を豚の剛毛で作られた針に伝達し，油煙（スス）が塗布された紙の表面を削ることで波形として記録する仕組みを持つ．スコットはフォノトグラフを科学的な研究に役立つ「自然の速記者 (une sténographie naturelle, a natural stenography)」なるものとして構想し，空気振動をも視覚的に記録できたという点で偉大な功績を残した (Feaster，2008).

1.3. トーキーのサウンドトラック

1900 年代に発明されたトーキーでは，映像と音声とが同期して再生される．その中でも，リー・ド・フォレストらによって商業化がなされたサウンドトラックは，映像が記録されたフィルムの隣に音声情報を光学的に記録するという仕組みを持つ (谷口, 2015)．そこでの音声信号の光学的な記録には，音声波形の振幅を面積として表す可変面積型 (図 1 右) と，その一コマあたりの帯の面積量を濃度に変換した可変濃度型 (図 1 左) という 2 種類の変調方法が用いられている．

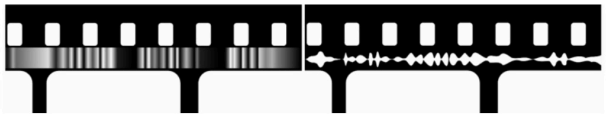


図 1: 2 種類の光学録音サウンドトラック　(左：可変濃度型，右：可変面積型)

このサウンドトラックの発明後，映像作家のオスカー・フィッシンガーによる，サウンドトラックの心

的イメージを視覚的に伝える手法 (涌井，2007) としての活用や，ノーマン・マクラレンによる，サウンドトラックへの直接的なペンの書き込みによる音声と映像の同期 (中川，2015) というように，可変面積型のサウンドトラック部分をキャンバスのように扱い，図形から音を生成しようとする幾つかの試みがなされている．

1.4. フォノトグラフからの復元

先述のように，フォノトグラフにとって最も重要なのは，音を視覚的に記録することであり，その音声を復元することではなかった (中川，2010)．しかし，2008 年に First Sounds Initiative がコンピュータ解析技術を用いて，フォノトグラフに記録された歌声を復元することに成功した．以下，その復元の過程をパトリック・フィースターによる資料 (Feaster, 2016, Playback Methods for Phonogram Images on Paper)，(Patrick Feaster, 2019, Enigmatic Proofs: The Archiving of Édouard-Léon Scott de Martinville's Phonautograms) に基づき順を追って説明する．

当初，Feaster らは，レコード上の針の動きのように，波形の線を光学的に追従することを検討したものの，広い滲みや，時間に逆行した波形 (図 2) により，この試みはうまくいかなかった．

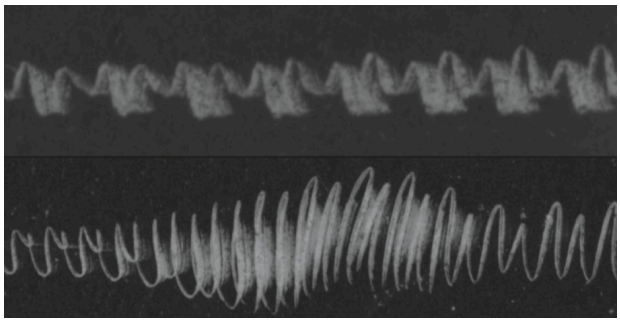


図 2: 時間に逆行した波形

続いて，画像を音声へと変換するソフトウェア ImageToSound(`https://www.softpedia.com/get/Others/Miscellaneous/ImageToSound.shtml`) を用いて，先述のトーキーのサウンドトラックと同様に各列のピクセルの輝度の平均の値を音声に変換することを試みた．しかし，この方法では波形の上または下の領域を白で塗りつぶす作業を行わないといけないため (図 3)，次の方法に切り替えた．

この方法では，数値計算ソフトの GNU Octave を用い，グレースケール画像のピクセル強度 (pixel intensities) を行列として，各列のピクセル強度の平均の座標を波形の位置へと対応させている．以下，縦軸が時間軸の 5 列の行を用いた例を示す (図 4)．

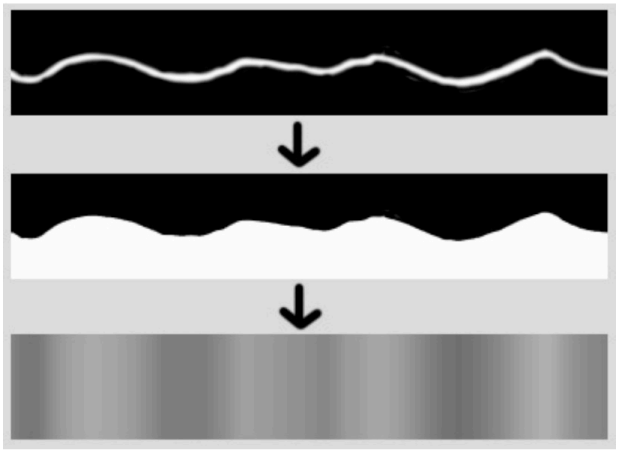
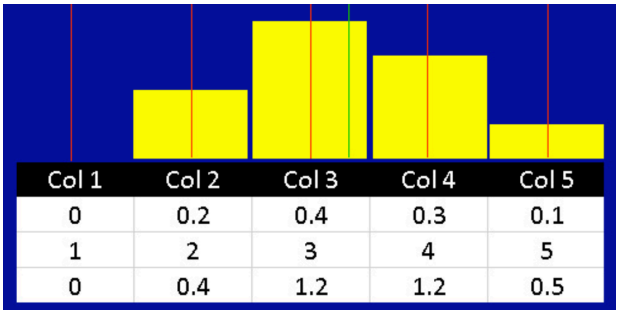


図 3: ImageToSound 使用前の処理



Col 1	Col 2	Col 3	Col 4	Col 5
0	0.2	0.4	0.3	0.1
1	2	3	4	5
0	0.4	1.2	1.2	0.5

図 4: GNU Octave での動き (表の上段：全画素強度，中央：列番号，下段：全画素強度と列番号の積)

ここでは，緑の線が全画素強度の中心 (この場合の 3.3) を表し，元の画像に近づけた場合は図 5 のようになる．

この方法においては，時間軸上の各点における定量化された中心の位置が返される．また画素強度の数値について，元ピクセルの強度値を最適なべき乗まで上げてから計算を行うことで，ノイズが完全に減衰するポイントが現れる．

2. 画像情報処理

以上の先行事例を踏まえた上で，本研究では音声波形画像から音声を生成する上で，先述のような画素強度を用いずに，2 値化 (白と黒) された波形画像を使い，その波形のエッジ情報のみを数値化することで，簡易に音を生成することを目指す．以下本項では，波形画像から音声を生成する上で必要な画像情報の特性について説明する．

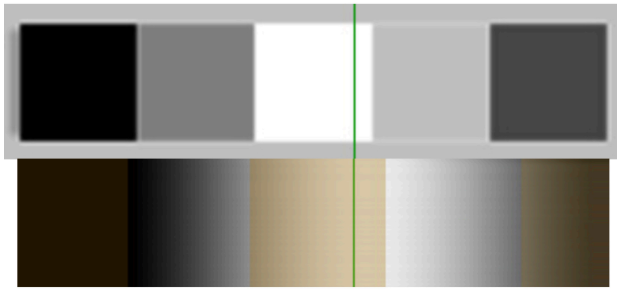


図 5: 上：実際のピクセル強度値との相対的に表した場合
下：ピクセルに分ける前のアナログなデータ

2.1. 画像の表現方法

画像を表現する方法には大別すると，ラスタ画像とベクター画像の二つがある．ここでラスタ画像とは，色と輝度の情報が含まれる正方形のピクセルを最小単位として，それらのピクセルを平面上に配置することにより画像を構成する．一方で，ベクター画像では，数学的に表される点，線，曲線，多角形によるパスを組み合わせて画像を表現する．この内，本研究ではラスタ画像で音声波形画像を取り扱う．

2.2. 画像の座標系

先述のようにラスタ画像では左上を原点 (0, 0) としてピクセルが 2 次元上に配置されている (三浦，2013) ため，本研究の手法では，画像のエッジ情報の座標を取得する際に，左上から右下に向かって行列の値が増していくことになる．

3. 予備実験

3.1. プロトタイプ

本研究では予備実験として以下の手順により，波形画像のエッジを検出し，取得したエッジ座標の x 座標を時間，y 座標を振幅に変換することで音声を生成することを試みた．

(1). Python 上で，指定したパスの画像を読み込み，OpenCV の Canny エッジ検出アルゴリズムを使用し，画像からエッジを抽出する．
(2). 各々の x 座標において，エッジが波の上下に現れる (図 2) ため，その中心の座標を y 座標とする．
(3). 画像の中心から各エッジまでの垂直方向の距離をピクセル値で計算し，振幅データを生成する．
(4). (3) を適切な値で割って振幅の範囲が-1 から 1 になるように調整する．
(5). 生成する音声のサンプリングレートと再生時間を決める．
(6). np.arange を使用して，時間軸に沿った配列を作成する．
(7). scipy.io.wavfile モジュールを使用して，作成した配列を音声を生成可能な WAV ファイルとして保存する．

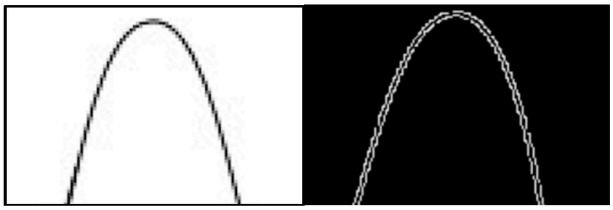


図 6: 左: 読み込んだ波形の一部 右: 検出されたエッジ
波の輪郭に沿って 2 本の線のようにエッジが検出される．

4. 結果

上のコードで正弦波 (のような図形) を読み込んだ結果を図 7 に示す．

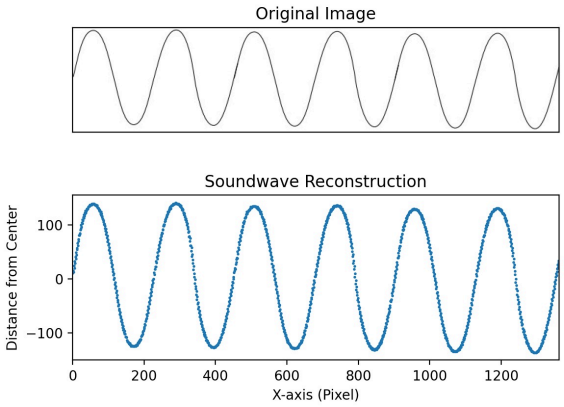


図 7: 上：読み込んだ元の画像　下：中心からの距離 (ピクセル値)

保存した音声の波形を音声情報処理ソフトウェアの Audacity 上で確認する．

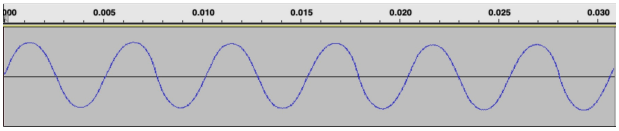


図 8: Audacity 上での波形

4.1. 考察

以上の手法により波形画像から音声を生成することができたものの，手順 4 に示した「適切な値」が元の画像サイズや，画像に対しての波形の大きさによって異なるため，毎回値を調整しなければならない． また，この手法では振幅 0 の値が読み込んだ画像の高さの中心となっているため，画像によっては正しく中心が取れない可能性がある．

4.2. 改善

以上の点を踏まえて実施した修正を以下に記す．

4.3. 修正 1

読み込む画像の波形の位置によらず音を生成するため，取得したエッジの最大値と最小値の中心の値を振幅 0 とした．

図 3 の正弦波を上にずらした画像を読み込ませた結果を図 9 に示す．

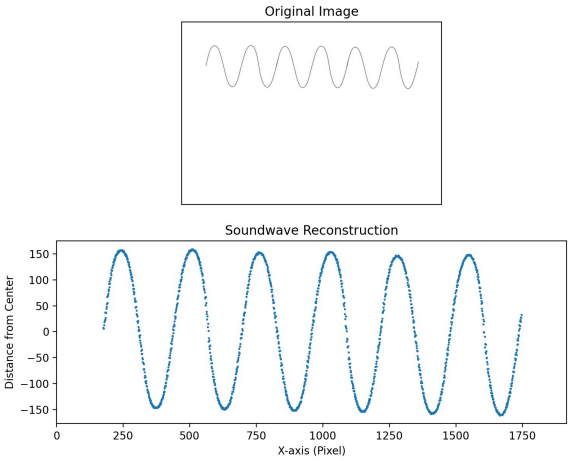


図 9: 上：読み込んだ元の画像　下：中心からの距離 (ピクセル値)

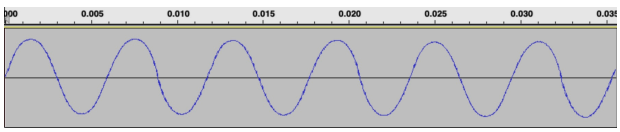


図 10: 図 5 の上の画像を読み込み Audacity 上で表示したもの

4.4. 修正 2

前後のエッジとの隙間を埋め，波形を滑らかにするため，得られたデータに対して補完と平滑化フィルタを適用した．

図 3 の上の画像を読み込ませた結果を図 11 に示す．

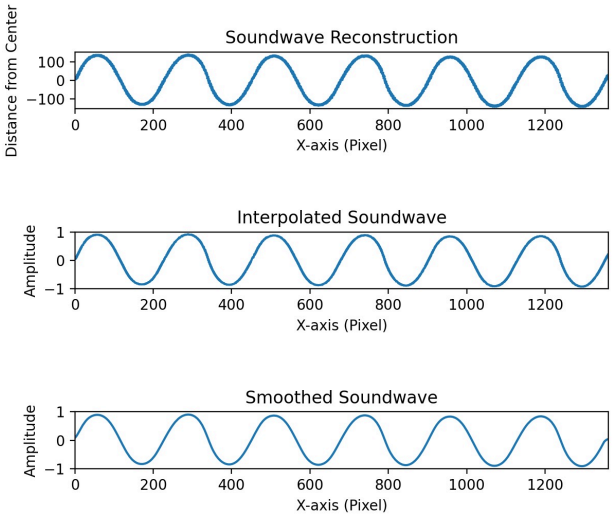


図 11: 上段：中央と各エッジの差 中央：各エッジの間を補完
下段：滑らかになるように平滑化フィルタを導入

Audacity 上での波形を図 12 に示す．

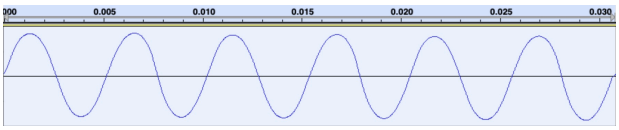


図 12: (2) の作業を加えた後生成された音声の Audacity 上での波形

以上の修正により，図 8 と図 12 を比較した場合，滑らかな波形になっていることがわかる．

4.5. 修正 3

長時間の音声の生成を可能とするために，横方向のピクセル数を拡大した．本手法では，音の長さが横方向のピクセル数により規定されているため，画像処理により複数の画像を横方向に足し合わせることで，秒数の増加，音高の変更を可能とした．なお，実験においては，Audacity 上でドラムならびに人の声の音声データを，それぞれはっきりと波形が見えるまで拡大し，スクリーンショットを複数枚とって貼り合わせた画像を

作成し，それらの色を Adobe Photoshop を用いて調整した画像データを使用した．

読み込ませた画像と生成された音声波形を比較する．

ドラムの波形

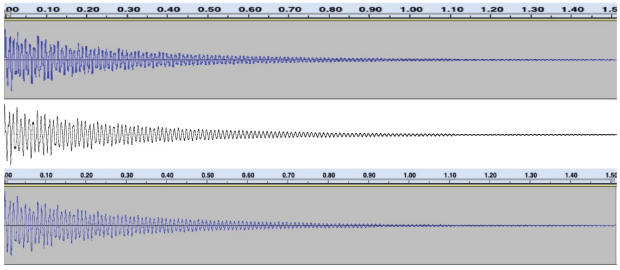


図 13: 上：音源の波形
中央：画像処理後の読み込ませた画像
下：生成された音声の波形 (画素数 1．5 倍の処理)

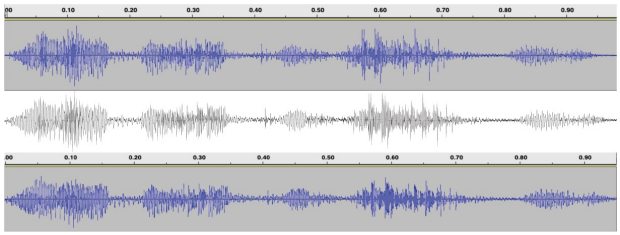


図 14: 上：音源の波形　中央：画像処理後の読み込ませた画像　下：生成された音声の波形 (画素数 2．63 倍の処理)

この修正により，画素数を適切な値に拡大することで，元の音声データに近い音声が再現可能となった．

5. 本実験

以上の修正を加えた上で，手書きの波形からの音声の生成ならびに逆再生音声の生成を試みた．実験では手書きの波形および印刷した波形をブックスキャナーで読み取った．

実験では 4 種類の手書きの波形をスキャンした上で，各画像の拡大倍率を調整し繋ぎ合わせることで，一繋ぎの音声の生成を試みた．図 15 は元の手書きの波形と生成された音声の波形の比較である．

この実験により，自分の手で描いた 4 つの波形から，8 秒ほどの音が生成された．

ここで，ノイズが多数発生しているが，その原因の一つとして，波形の逆行がある．本研究の手法では，先述のフォノトグラフからの再現における光学的な追従と同様に時間方向での波形処理を行っているため，波形の逆行に対応することが難しい．

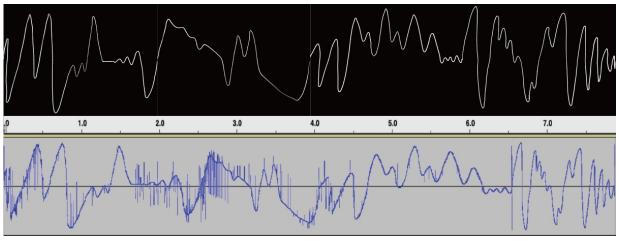


図 15: 上：手書きの波形　下：生成された音声の波形

5.1. 逆再生音声の生成

筆者により発声された母音 (i) と子音 (ka) の音声データを Audacity 上で ika となるように配置し，その音声を確認した上で，各々の波形を画像として印刷し 180 度回転させさらに順番を反対にしてスキャンした画像から aki と発音させることを試みた．

結果として，生成された子音の立ち上がりの潰れ，並びに多数のノイズの混入により，正確に聞きとることは難しかったものの，母音を意識すると逆再生音声として聞き取ることができた．

子音の立ち上がりの潰れの原因としては，使用したスキャナーの画素数では印刷した波形の細かい部分を読み取れなかった，ということが考えられる．

また，2 つの実験に共通するノイズの混入については，スキャナーのライトが紙に反射し，うまく 2 階調化できなかったということが原因として考えられる．

6. 考察

6.1. 先行事例との比較

研究の手法は，First Sounds Initiative の手法と比較して，エッジ情報のみを用いているため，簡便なものとなっている．一方で，First Sounds Initiative が，グレイスケール画像の 1 列の平均化，光学処理，ピクセル値のべき乗からの強度平均，によりノイズの除去を実現した逆行した波形については，その音声としての生成ができていないが，波の頂点と次の頂点との位置関係から順方向に収めることで，逆行部分を時間方向に戻すことを今後検討していきたい．

6.2. 新たな音の生成方法として

普段私たちが耳にする音は振幅の正負がほとんど対称に交互に現れる．手書きでの波形はそのような概念は取り除いた新たな音が生成できる．また，波形の拡大の倍率や組み合わせる順番をランダムにすることで「音楽のサイコロ遊び」(モーツァルト, 1797) や「易の音楽」(ケージ，1951) といった，これまでの音楽史上の試みに連なる，不確定性のある曲の創造に

貢献できる可能性がある．これまでにも，波形をモチーフにしたウォールアート (Bespoken Art, https://www.bespokenart.com/index.html) やリング (Encode Ring, https://encodering.com/) のように音声波形を視覚的に表した事例はあるものの，本研究の手法は，波形の視覚的な造形の面白さだけでなく，実際の音声としてその構造を視覚的に示すことができる，という点が特徴的である．

6.3. 逆再生音声の生成

逆再生とは, 元の音声波形を逆から辿ることである．過去にも，母音は全く性質を変えないこと，子音も後に母音がくると元の子音として聞こえること，従って母音も子音も前後転倒自由の性質を持っているということを，トーキーフィルムを逆に回す実験により確認した事例がある (田口，1943)．本研究は，およそ 80 年前のこの事例と同様の手法を，画像処理という現代のテクノロジーで検証したものと見なすこともでき，母音と子音の任意の逆再生を可能とする本手法は，音声学における音声構造の理解を促す可能性もある．

6.4. 音声のパズル化

本手法を応用することで，音声波形の画像を縦（振幅方向）と横（時間方向）に分割し，自由に組み替えたり，一部を抜いたりすることで任意の音声を生成する，ということが考えられる．当初の着想のように，分割した音声波形をピースとしてジグソーパズルのように音を組み立てる，ということの他，この手法の応用例としては, 例えば, 音声データの横軸を時間, 縦軸を振幅とした上で一定の区画ごとに平均パワーを求め，それらの区画を雑音で埋めたモザイク音声 (上田，2022) を対象することが考えられる．音の高さを手がかりとした音声の了解度を検証するために用いられるこの手法において，処理前後の各々の区画を個別の画像として印刷し，それらをパズル上に組み合わせることにより，あたかもパズルの「ピースを取り除く」ような，新たな実験手法を生み出すことが期待できる．

7. まとめ

本研究では 2 値化された音声波形画像からエッジ情報を抽出し，取得したエッジ座標の x 座標を時間，y 座標を振幅に変換することで音声を生成した．先行事例であるフォノトグラフからの復元と比較すると，一部同等の処理をしているものの，より簡易な手法で音声を生成することができる．ただし，実験で確認されたように，ノイズの混入が見られたことは今後の課題であり，逆行する波形の処理ならびにスキャニングの手法を改善することで，改良をはかっていきたい．また，当初の着想であるジグゾーパズルのように音声を組み立てていくことを実現する上では，スキャンする際の解像度（画素数）が大きな課題となっていくであろう．筆者にとって，本研究において画像から音声を生成し，その波形を再び視覚的に確認するという体験は興味深いものになった．これまで，音楽をはじめとした音声情報というものはあくまでも聴く対象，という理解であり，波形そのものをあまり意識したことはなかったが，今回の研究において，波形さえあれば音声を生成できる，という知見が得られたことは良い経験となった．

8. 謝辞

本研究の一部は, 日本学術振興会科研費 [JP23H00591] の支援を受け実施された．

9. 参考文献

[1] 川克志, (2010), 音響記録複製テクノロジーの起源—帰結としてのフォノトグラフ, 起源としてのフォノトグラフ, 京都精華大学紀要, 第 36 号, pp.1-20.

[2] easter, Patrick. (2008-1), “Working Paper 1, Edouard-Leon Scott de Martinville’s “Principes de Phonautographie” (1857) A Critical Edition with English Translation and Facsimile.”, Facsimile by David Giovannoni, FirstSounds.org.

[3] easter, Patrick, (2008-2), “Working Paper 2, Edouard-Leon Scott de Martinville’s 1857 Phonautograph Patent (31470) and 1859 Certificate of Addition A Critical Edition with English Translation and Facsimile.”, Facsimile by David Giovannoni, FirstSounds.org.

[4] easter, Patrick, (2008-3), “Working Paper 3, Edouard-Leon Scott de Martinville’s “Fixation Graphique de la Voix” (1857) A Critical Edition with English Translation and Facsimile.”, Appendix: Facsimile of the Photograph News, 15 April 1859, pp.62-64, FirstSounds.org.

[5] 井隆, (2007), オスカー・フィッシンガーと寺田寅彦, 言語文化論集, 25(2), pp.211-223.

[6] 口文和, 中川克志, 福田裕大, (2015), 音響メディア史, ナカニシヤ出版.

[7] easter, Patrick, (2016), Playback Methods for Phonogram Images on Paper, In Sustainable Audiovisual Collections Through Collaboration:

Proceedings of the 2016 Joint Technical Symposium (p. 90), Indiana University Press.

[8] atrick Feaster, (2019), Enigmatic Proofs: The Archiving of Édouard-Léon Scott de Martinville's Phonautograms, Technology and culture, Johns Hopkins University Press, Volume 60, Number 2 Supplement, April 2019, pp.S14-S38.

[9] 浦靖, (2013), デジタル画像解析, Nippon Shokuhin Kagaku Kogaku Kaishi, Vol.60, No.5, pp.242-256.

[10] 口泖三郎 (1943) 音と音楽 人文書院

[11] 田和夫, (2022), 劣化音声の知覚的修復: 音声の断片をつなぎ合わせて意味のあるまとまりにする聴覚の働きとその限界, 九州大学, 研究・産学官民連携, 研究の取組紹介, 芸術工学研究院研究紹介. https://www.kyushu-u.ac.jp/ja/research/information/artdesign/design/design51

10. 著者プロフィール

木村 凪 (Nagi KIMURA)

2001 年愛媛県出身. 2020 年に九州大学芸術工学部音響設計コースに入学. 大学院進学に失敗し, 大学卒業後はソフトウェアの SE として就労予定.

城 一裕 (Kazuhiro JO)

1977 年生まれ. 博士（芸術工学）. 英国ニューカッスル大学 Culture Lab, 東京藝術大学芸術情報センター [AMC], 情報科学芸術大学院大学 [IAMAS] を経て, 2016 年 3 月より九州大学 大学院芸術工学研究院 音響設計部門 准教授. 専門はメディア・アート. 現在の主なプロジェクトには「Life in the Groove」,「The SINE WAVE ORCHESTRA」,「phono/graph」などがある.