創作ノート

音に変容する声—《目覚めよ、日本国民！》を例に
Voices Transformed into Sound: A case study of “Wake Up, Japanese Citizens!"

大久保 雅基
Motoki OHKUBO
名古屋芸術大学
Nagoya University of Arts

概要

筆者が作曲した《目覚めよ、日本国民！》は、音声合成技術によって出力された音声を用い、4 台のタブレット端末がそれぞれ異なる言葉を発する作品である。これらの音声は、ピッチや再生速度の操作、重なり合いを通じて合唱的な音響表現を生み出し、音声が音楽の素材へと変容する瞬間を探求している。

本作は、16 世紀から議論されている歌詞の理解可能性の問題や、佐近田による電子音が音声へ切り替わる瞬間の観察、さらに黒嵜が提起した「『PLAY』される声」の議論を背景に持つ。これらの議論を踏まえ、声という意味内容を持つ音がどのように音楽的素材としての純粋な音へと変容し得るのか、また音声合成による音声がいかにして新たな意味内容を付与されるのかを考察する。

本稿は、声の音楽的変容をテーマに、音声合成技術が音楽表現にもたらす新たな可能性とその意義を示唆することを目的とする。

The composition "Wake up, Japanese Citizens!" by the author is a work utilizing synthesized voices generated by speech synthesis technology, with four tablet devices each delivering distinct words. These synthesized voices, through manipulation of pitch and playback speed as well as overlapping textures, create a choral-like sonic expression, exploring the moment when speech transforms into a musical material.

This work draws on several theoretical and historical contexts: the long-standing debate on the intelligibility of song lyrics dating back to the 16th century, Sakonda’s observations of the transition from electronic sound to speech, and Kurosaki’s discourse on “the voice as PLAYed.” Building on these discussions, this work investigates how meaningful sounds—voices—can transform into pure sounds as musical materials, and how synthesized voices acquire new layers of meaning.

This paper aims to shed light on the musical transformation of voice, highlighting the novel possibilities and significance that speech synthesis technology offers to musical expression.

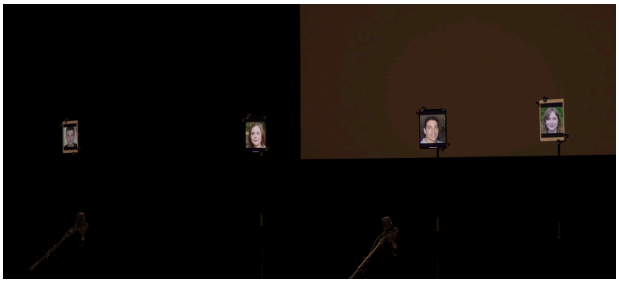

図 1: 《目覚めよ、日本国民！》の演奏の様子

1. 背景

本節では、音声合成の歴史と、歌声と言葉についての議論を振り返る。

1.1. 音声合成の歴史

本稿における音声合成とは「テキスト音声合成」を指す。これは、テキストを入力とし、音声データとして出力するものである。本節では音声合成の誕生から現在までの傾向を述べる。

最初の音声合成は、Voder(Voice Operation Demonstrator)(1939) である。これはオペレーターが複数の鍵盤を操作し、その周波数に応じた音声を電気回路で再現するものである。1980 年代には規則合成方式 (Formant Synthesis) が登場し、韻律やフォルマントの規則

を記述することで生成を行った。1990 年代にはコーパスベース音声合成 (Corpusbased Speech Synthesis) が登場する。音声や言語を集積したデータベースであるコーパスを用いて合成を行う。1990 年代〜2000 年代には、素片選択型音声合成方 (Unity Selection Synthesis) が登場し、音声を大量に録音し、音素などを接続する方法である。初代 VOCALOID もこの方法に従っている。2000 年代からは、統計的パラメトリック音声合成 (Statistical Parametric Speech Synthesis; SPSS) が登場し、機械学習に基づく手法が採用される。音声素片ではなく、音声パラメータを表現する統計モデルを保存している。2010 年代後半からは、一貫学習に基づく音声合成 (End-to-end Speech Synthesis) が登場する。統計的パラメトリック音声合成では、自然言語処理や音声信号処理の知識が必要だったが、機械学習による一貫学習によって不要になった。

ニコニコ動画や YouTube などの動画配信サイトで公開されている動画にしばしば採用されている「ゆっくりボイス」や「VOICEVOX」もこれらの音声合成である。「ゆっくりボイス」は AquesTalk 1(2006)、AquesTalk 2(2011)、AquesTalk 10(2017) [1] に搭載されている。AquesTalk は規則合成方式である。また「VOICEVOX」(2021) [2] は End-to-End モデルを使用せず、独自の深層学習によるモデルによって開発している (藤本 2021)。

ここで着目しておきたいのは、コーパスベース音声合成の登場前後で合成方式が変化することである。それ以前は電気的に、もしくは電子音によって声を再現していた。しかしその後は、実際の人の声を録音し、それを素材に音声合成を行っている。その生成される声は、実際に存在する人の声が元になっている。

1.2. 歌声と言葉の議論

歌を伴う音楽には、音そのものの表象だけではなく、歌詞による文学的・言語的な表現が含まれる。そのため、音楽そのものの本質を議論する際には、歌詞が意味を持つ「歌」は、純粋な音楽表現から外されることが多かった。声という音には、ピッチや音色といった物理的なパラメータに加え、言葉という意味内容が付随する。そのため、声を聴く際には、音響的要素とともに言語的意味を聞き取ろうとする二重の知覚が発生する。本稿で声が音楽として知覚されるのはどのような瞬間であるかを議論するために、本節では音声や歌声にまつわる論考を確認していく。

1.2.1. 言葉と音の間

1554 年から 1563 年にかけて行われたトリエント公会議では、カトリックの儀式、哲学、神学などが確認、刷新された。ここでは、音楽における複雑さが問題となった。音楽が過度に複雑で、それを表現するはずの宗教的な歌詞を会衆が聴くことができず、理解できなくなったというものだ。教会関係者は音楽はテキストの主人ではなくその従者であるべきだと主張した (Kevy 2002)。16 世紀の時点でも、音楽の複雑性により歌詞が聴き取れなくなったことが問題となっている。ここから、歌もある程度の複雑性を持つことによって、言語的意味を理解することが難しくなることが読み取れる。言語として聴き取ることが難しくなった瞬間に、それは音楽のテクスチャに変わり、歌声が言語的意味を持たない音楽の素材に変わるのかもしれない。

佐近田は自作のフォルマント合成による音声合成プログラムを開発している。音声をスペクトログラムで確認すると、山や谷の分布が見え、その山の頂上の周波数の低い方から順に第一フォルマント、第二フォルマント、第三フォルマント...と呼ばれる。母音の「アイウエオ」ではそれぞれに異なったフォルマントの分布パターンが見られる。佐近田はブザー音を鳴らしその音のフォルマントの値をバンドパスフィルターによって変更することで、音声に聞こえる音響を作っている。彼は授業などで、ブザーにしか聞こえない電子音から、徐々にフィルターを効かせて、ピッチを揺らせて「アイウエオ」をランダムに歌う少女の声へと「連続的」に聞かせる実験を行っている。そこでは、ブザー音が徐々に声に変化するのではなく、途中でパチっとスイッチが切り替わるように声だと聞こえる体験をすると語っている (佐近田 2024) [3]。電子音でしかない音がフォルマントを獲得していくことで、言語として聞き取れる瞬間が存在することを示唆している。

これら２つのケースから、言葉が音になる瞬間、音が言葉になる瞬間というものがあり、両者はそれぞれ推移が可能であると考えられる。

1.2.2. 「*PLAY*」される声

近年の深層学習を用いた音声合成手法は、事前に大量の音声の断片であるコーパスを必要とする。そのため、事前に録音された音声が深層学習の学習に用いられ、出力される際にはそれらが合成され、再生される。そのために「再生される音声」について取り上げた議論に触れる必要がある。黒嵜はそのような音声について、バルトと川田の言説を用いて議論を展開している。

バルトは２人の歌手、シャルル・パンゼラとフィッシャー＝ディスカウの歌声を聴き比べ、その差異につ

[1] AquesTalk シリーズ: https://www.a-quest.com/products/aquestalk.html

[2] VOICEVOX: https://voicevox.hiroshiba.jp/

[3] 佐近田展康. 2024.「声という迷宮　人工音声をとおして考える〈声〉の哲学序説」『ユリイカ』第 56 巻, 12 号, pp.222-223.

いて前者は「声のきめ」があり、後者にはないと述べた。「声のきめ」は言葉の意味、その形式、メリスマ、演奏スタイルの向こう側にある、胸腔の奥、筋肉、粘膜、軟骨からあたかも演奏者の内部の肉と彼の歌う歌と同じ皮膚が覆っているかのように、聴く者の耳にもたらされる、歌手の身体そのものであるような何かである。そしてそこには意味作用がもたらされるとされる (バルト 1984)[4]。

川田は言述には「演戯性」があり、口承文芸のなかで重要な意味を持っていると指摘する。演戯性とは、例えば古典落語や昔話のように、情報としては聞き手が全く分かっており、新しく知る情報はないが、何度聞いても面白いような言述としている (川田 1988a)[5]。そして、マイケル・ジャクソンの歌に痺れる現代の若者や、羽左衛門の口跡に酔った明治・大正の若者の、アイドル化された声の中毒現象について、フランシス・ベーコンの語を用いて「イドラ（熱愛の対象）」と読んだ (川田 1988b)[6]。

黒嵜は音声が録音再生装置によって「PLAY」された場合の、「きめのない／ある」声について、「再生」されたものだとするか、「演戯」されたものだとするか、との峻別だと述べている (黒嵜 2024)。前者についてはバルトの「声のきめ」に含まれる意味作用を取り上げ、それを持たない、つまり文字を読んでいないかのような声をさす。後者については声から詞や楽譜のような意味作用を想像させ、何度も聞きたくなるような「イドラ」を持つものだとしている[7]。「PLAY」される音声が意味作用を持つかどうかの問いは、本稿にとって有用である。もしその音声が意味作用を持たず「再生」されるだけの場合、それは言語的認識をせずに音として聞くことに変化しているからである。

2. 作品の概要

《目覚めよ、日本国民！》は４台のタブレットから、音声合成によって生成した音声を再生し、ピッチをもたせ、重ね合わせることで合唱のような音響表現を生み出す作品である。各タブレットの画面には GAN によって生成された、この世に存在しない人の顔が映し出されている。音声に合わせて口が動くように処理されている。哲学的な問いを ChatGPT に投げかけ、その回答を歌詞としている。個人が物事を考えるよりも、集合知としての人工知能の答えのほうが優れているという思想をテーマにした作品である。

合成音声には日本語音声コーパスである JSUT(Japanese speech corpus of Saruwatari-lab., University of Tokyo)[8] を学習した、Tacotoron 2[9] による一貫学習に基づく音声合成システム[10] を使用している。

本作は 2022 年 12 月 8 日に、日本電子音楽教会主催の「日本電子音楽協会 30 周年記念演奏会」にて初演された (大久保 2022)。この記録映像は YouTube にアップロードされている[11]。

3. 作品の構成

本節では作品の構成について述べる。なおセクションごとの再生時刻は、YouTube にて公開している記録映像のものを参照している。

セクション１　00:10-02:25

冒頭は４つのタブレットそれぞれが１つの台詞を読み上げる。その後２台ずつのペアになり読み上げていく。2:47 からは音声生成の結果、音素が繰り返し生成されるエラーが起きたものをそのまま採用している。今回使用しているモデルでは、テキストをいくつかの文字数を超えた長文にすると、このような音素が繰り替えされるエラーが出力されることを発見し、それを音響的な表象に捉え直している。

セクション２　02:26-03:48

１台ずつが異なる歌詞を読み上げ、重なることで音声の内容を聞き取ることを困難にさせ、音響テクスチャに変化させる効果を持たせる。そしてだんだんと再生速度が早くなりピッチも上昇し、内容が聴き取れないくらい高速になることで、さらに高音の音響テクスチャを生成させる。

セクション３　03:49-05:08

それぞれのタブレットが歌詞の一部を切り取った音声を繰り返すことで、台詞の抑揚を、リズムを持ったテクスチャに変化させる。

セクション４　05:09-5:56

冒頭は切り取った台詞をそれぞれのタブレットが順番に再生する。途中から短時間のループやピッチの変更などの処理が加えられることで、言語の意味内容よりもリズムやピッチの変化に意識を向けさせるように促す。

[4] ロラン・バルト著. 沢崎浩平訳. 1984「声のきめ」『第三の意味 映像と演劇と音楽と』. みすず書房. pp.189-190.

[5]「はなしの演戯性」『口承文芸研究』. 第 11 号, pp. 88-103. https://ko-sho.org/download/K_011/SFNRJ_K_011-09.pdf. 2024 年 11 月 25 日参照. p.88

[6] 川田順造. 1988b.『聲』. 筑摩書房. pp.249-250.

[7] バルトは「歌声のきめ」について論述しているが、黒嵜はそれを「音声のきめ」に置き換え、さらに川田の「演戯性」は音声による音声の想起を指しているが、黒嵜はそれをテキストを想起させる「声のきめ」の意味作用と重ねているため、さらなる精査は必要であると思われる。

[8] JSUT: https://sites.google.com/site/shinnosuketakamichi/publication/jsut

[9] Tacotron 2: https://github.com/NVIDIA/tacotron2

[10] 使用したモデル: https://github.com/r9y9/ttslearn

[11]《目覚めよ、日本国民！》記録映像: https://www.youtube.com/watch?v=IliBBhug2JM

セクション５　**05:57-6:10**

タイムストレッチによって音声を高速度で再生することで、意味内容を認識できないが、音声のテクスチャである様相を残す。そしてピッチシフトをかけていくことで、声から電子音としての音に変化させる。

セクション６　**6:11-8:53**

再度４つのタブレットは、それぞれの台詞を読み上げる。ただし下手から順番にピッチを固定してハーモニーを形成していいく。ここではそれぞれがピッチシフトにより１つのピッチに留まるように処理している。最後の方には、セクション１で採用した、長文を読み上げると音素が繰り返されて出力されるエラーを再び使用している。これによって声のトレモロがうねりのあるハーモニーを作る。

セクション７　**8:54-9:55**

再び各タブレットが台詞を読み上げるが、１つの台詞を４種の声それぞれに分割して読み上げていく。

4. 考察

4.1. 言葉と音の間

本作では全体を通して、言葉として認知できる箇所と、声のテクスチャを残しながらも言葉として認知が難しく、音として聞こえる箇所を組み合わせたり、推移する表現を採用した。

セクション１のように、タブレット端末がそれぞれ音声を再生する場合、意味内容を理解することは容易である。しかし、セクション２のように、４つの異なる音声が同時に再生されるとき、１つの音声のみに集中することも難しく、４つの音声が重なった１つのテクスチャとして聞かれるだろう。これは言葉が音に変わる瞬間と言えるかもしれない。

またセクション５において、タイムストレッチ処理によって高速で再生される音声は、意味内容は理解が難しいものの、声としてのテクスチャは残っている。しかしピッチシフトが加わることで、声というテクスチャを離れ、電子音として聞こえるようになる。これは、声にピッチというパラメーターが加わったこと、そして音色的に声のフォルマントを逸脱したことが考えられる。

本作では音を言葉に推移する表現は扱わなかったが、言葉が「声のテクスチャ」という音、そして電子音に推移することを試みている。また、16 世紀の例では歌声が複雑になり音として聞こえる例を挙げたが、本作の事例は音声においても起こり得ることを示す。

4.2. 再生される声

黒嵜による議論においては、録音された音声が扱われていていたが、本作のような音声合成でもその議論は引き継げるかもしれない。

近年の深層学習を用いた音声合成では、コーパスという様々な言葉を実際の人間の音声を録音した音声データが使用されている。録音時には、文章として成立するテキストを読み上げている。そのため、素材となる音声は原稿となるテキストを読み上げた声であり、自然な会話の声ではない。この音声からは、「声のきめ」を持つ、つまりテキストという意味作用を持つ声が採用されていると言えるだろう。それが深層学習処理によって分解され合成された場合、本来の音声に乗った意味内容は失われるかもしれないが、合成による音声の規則性により、また新たなテキスト性が生じるだろう。これは音声合成の音声を聴いた時に感じられる不自然さでもある。この自然な音声とは異なる違和感が「テキストから変換された音声」を感じさせ、この音声の奥にテキストを感じさせる。

これを前提とした上で、「PLAY」される音が本作でどのように扱われるかをみていく。セクション１においては、それぞれの音声の意味内容を理解することは可能であり、声のきめがあると言えるだろう。セクション３では、意味内容を持つ言葉が切り取られループ再生される。何度も繰り返されるうちに、言葉の内容ではなく、その言葉がもつアクセントやイントネーションによるリズムとして聴くようになる。こうなった聞き手は、音声を「声のきめ」のある、意味作用を持つ声ではなく、単に「再生」された音として聴いている状態だと言えるだろう。

4.3. 音声として聞く

深層学習を用いた音声合成では、学習過程においてコーパスの音声データが分解され、出力時に合成される。つまり厳密には、言葉という意味内容を持った音声を録音した音源ではなく、意味内容を持った言葉として聞こえる音声として合成されたものである。

それ以前の規則合成方式等は、音声を録音しているのではなく、電子音から音素として聞こえる音に合成している。自然な人間の声と比べると、近年の深層学習方式よりも再現度が劣り、言葉として認識するには少しの慣れが必要でもある。しかし、現在の 2000 年頃から、様々な動画にて「ゆっくりボイス」として定着し、現在でも使われていることから、規則合成方式のような古い音声合成方式でも、意味内容をもった言葉として聞こえる効果を持っている。

私達は港町が描かれた絵画を鑑賞する際、これはキャンバスと絵の具であると認識するのではなく、港町の風景を見ている、として観る。そのように、電子音や

サンプリングによって合成された音も、音声として再現された場合は、そこに意味内容をもつ音声として聴くことができるだろう。

5. 参考文献

1988a.「はなしの演戯性」『口承文芸研究』. 第 11 号, pp. 88-103. https://ko-sho.org/download/K_011/SFNRJ_K_011-09.pdf. 2024 年 11 月 25 日参照.

川田順造. 1988b.『聲』. 筑摩書房.

黒嵜想. 2024.「(6)きめのない、きめのある｜声をかく」『ちえうみ PLUS』. https://chieumiplus.com/article/koewokaku06. 2024 年 11 月 25 日参照.

佐近田展康. 2024.「声という迷宮　人工音声をとおして考える〈声〉の哲学序説」『ユリイカ』第 56 巻, 12 号, pp.221-230.

藤本健. 2021.「商用でも利用可能な AI 音声合成ソフトウェア『VOICEVOX』がオープンソースとして無料でリリース」『DTM STATION』. https://www.dtmstation.com/archives/41014.html. 2024 年 11 月 25 日参照.

難波優輝.2017.「ピーター・キヴィ『音楽哲学入門』読書ノート」. https://lichtung.hatenablog.com/entry/2017/07/14/202452. 2024 年 11 月 25 日参照.

ロラン・バルト著. 沢崎浩平訳. 1984「声のきめ」『第三の意味　映像と演劇と音楽と』. みすず書房. pp.185-199.

山本龍一, 高道慎之介. 2021.『Python で学ぶ音声合成　機械学習実践シリーズ』. インプレス.

Kevy, Peter. 2002. "Introduction to a Philosophy of Music". Oxford University Press.

6. 参考作品

大久保 雅基. 2022.《目覚めよ、日本国民！》 第 24 回日本電子音楽教会定期演奏会 -仮想と伝承-. https://www.youtube.com/watch?v=IliBBhug2JM.

7. 著者プロフィール

大久保 雅基 (Motoki OHKUBO)

宮城県仙台市出身。コンピュータ音楽の作曲家。テクノロジーと音楽の関係を創作、文化の面から見直すことで、オルタナティブな創作を行う。名古屋芸術大学、愛知淑徳大学、相愛大学非常勤講師。洗足学園音楽大学 音楽・音響デザインコースを卒業。情報科学芸術大学院大学 [IAMAS] メディア表現研究科 修士課程修了。Contemporary Computer Music Concert にて ACSM116 賞 (2010/東京)、Wired Creative Hack Award 2019 にて Sony 特別賞を受賞。ARTE PUBBLICA E METAVERSO(2023) にて現代音楽部門 3 等賞を受賞、The 2024 Living Music Summit: Sound Image Gesture に入賞。MUSICACOUSTICA-HANGZHOU 2024 Electroacoustic Music Composition Competition ファイナリスト。塩竈市杉村惇美術館若手アーティスト支援プログラム Voyage 2021 に採択された。Musica Viva Festival 2010 “Sound Walk”、同 2013 ”Close, Closer”、千代田芸術祭 2014 音部門 LIFE LIKE LIVE、Muestra Internacional de Música Electroacústica 2015、横浜スマートイルミネーションアワード 2014、ACOUSTIC FOR THE PEOPLE III “RAW”、ICMC2024、ISEA 2015、TENOR2023、Festival Atemporánea 2020、2023、2023 SONIC MATTER、やまなしメディア芸術アワード 2022、2023-24 に入選。Contemporary Computer Music Concert、Festival Futura、関西・アクースマティック・アート・フェスティバル、富士電子音響芸術祭 、サラマンカホール電子音響音楽祭等で作品が上演されている。先端芸術音楽創作学会、日本電子音楽協会会員、日本 AI 音楽学会。ACSM116 運営委員。