01

人工音声のアイデンティティ

モノローグ・オペラ『新しい時代』におけるフォルマント音声合成の挑戦

Identity of Artificial Voice

Note on a challenge for Formant Vocal Synthesis in Monologue Opera “The New Era”

映像メディア学科・教授
Department of Visual Media, • Professor
佐近田 展康 *Nobuyasu SAKONDA*

本稿は、2017年12月に上演された三輪眞弘＋前田真二郎作モノローグ・オペラ『新しい時代』[1]において、私が担当した「フォルマント音声合成」プログラミングに関する技術的覚書である。

本オペラは、主人公の《14歳の少年》が、ネットワーク上に存在する《新しい時代》と呼ばれる宗教に帰依し、自らの「声」を記号化することにより永遠の形而上的生命を得る —— 引き換えに、服毒して形而下の肉体を消滅させる —— 《聖なる儀式》の様子を描いている[2]。2000年に初演され、今回は17年ぶりの再演となる。再演にあたって脚本・音楽・演出・美術等の変更はなく、《14歳の少年》役も初演時と同じソプラノ歌手さかいれいしうが演じている。

作者により「一切のアップデートを封印する」方針が立てられた再演であるが、それは内容に関することであり、作品全体を背後で支える技術環境については17年間の進展を受け入れ、さらに一部には「当時の技術では不可能だったこと」への重要な「挑戦」も行われた。それが「フォルマント音声合成」である。初演の際には客席から本作品を体験した私であるが、今回は制作サイドとして、その「挑戦」に深くコミットすることができた。以下、その内容を紹介したい。

写真1：『新しい時代』公演の様子（撮影:羽鳥直志、提供:愛知県芸術劇場）

フォルマント、フォルマント音声合成とは

すべての説明の前に、「フォルマント」「フォルマント音声合成」の用語について簡単に触れておく必要があるだろう。

われわれが声を聞くとき、アイウエオなどの言語音素を識別できるのはなぜか？音声学では「フォルマント」の構成パターンの違いを聞き分けているからだと説明される。声帯の振動で生み出された声源音は、もともとブザーのような「音」であり、それが舌・顎・唇などの運動によって複雑に変形する経路（声道）を通過する過程で共鳴し、「声」になる。声に含まれる周波数成分を分解し、音響スペクトルとして視覚化すると、いくつもの特徴的な山の部分が見いだされる。それぞれの山の位置・大きさ・勾配は、音素の違い（アイウエオ）で大きく変化するが、声の高さや大きさを変えても形状はあまり変化しない。この山の部分を「フォルマント」と呼ぶ。われわれがアオウエオを識別できるのは、音素ごとにフォルマント

の構成パターンに違いがあり、その差異を聞き分けているからだ（図1上段の緑のスペクトル比較）。

年齢・性別・体格の異なる誰の声であっても、等しくアイウエオが聞き取れる以上、各音素のフォルマント構成パターンは、同じ言語を話す共同体内で共通性を持つ。しかし同時に、それは声の個性（アイデンティティ）を表す指標として個々人で異なるパターンも形成している。フォルマントは周波数の低い方から順に番号が付けられるが、一般に、第1、第2フォルマントは主として音素の弁別指標として機能し、第3フォルマント以降は声の個性の認知に強く影響を与えていると言われる。

本稿で扱う「フォルマント音声合成」とは、ソフトウェアでこのフォルマントを人工的にシミュレートする方法である。倍音成分を豊富に含んだ周期的波形（ノコギリ波など）の音源信号を、複数の並列バンドパス・フィルターを通すことで、特定の周波数成分を増幅したり、それ以外をカットすることにより目的の音色を得る（図1下段）。個々のフィルター特性は中心周波数・ゲイン・Q値という3つのパラメータで定義され、それぞれフォルマントの山の位置・高さ・勾配に相当する。これらのパラメータ値を動的に制御することにより人工的に音声を生成するのだ。

足早な素描だが、要するにフォルマント音声合成は、人間の肉声由来の素材をいっさい使わず、単純な周期的波形の音源信号をもとに倍音成分を彫刻のように削り、声に聞こえる音響へと仕立てる、純粋に機械的な音響合成法だということだ。詳細については別に発表した拙稿を参照してほしい[3]。

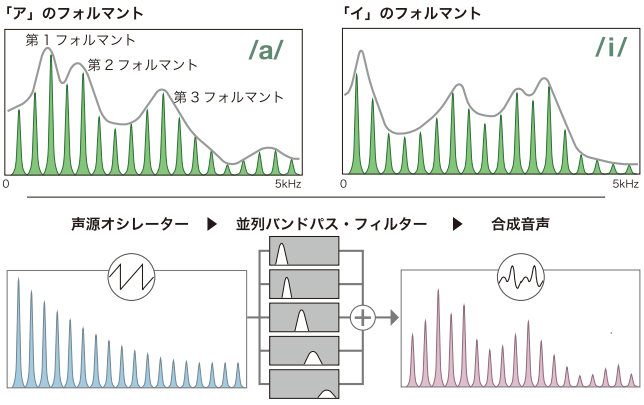

図1：フォルマントとフォルマント音声合成の概念図

前代未聞の挑戦

さて、本オペラにおける「挑戦」とは、《14歳の少年》役のさかいれいしうの歌声を、この「フォルマント音声合成」により生成させ、舞台上で共演するというものだ。私は20年来人工音声の研究を重ねており、この技術をベースに、三輪とは「フォルマント兄弟」としてユニット活動も行っている。2009年に発表した『フレディの墓／インターナショナル』[4]では、亡きロック歌手フレディ・マーキュリーに似せた人工音声を制作しており、筆者にとって「特定の個人の声を作る」ことは初めてではない。ただし、この時の人工音声は、フレディ・マーキュリーの声質および歌い方を分析したうえで、その特徴を誇張し、いわば「戯画化」したものであり、もし聴衆が「フレディに似ている」と感じたとすれば、それは各々の記憶に刻まれた彼の歌声のイメージとの対比においてであった。

しかし、今回の制作では状況が決定的に異なっている。というのも、舞台上のさかいれいしう本人が肉声で歌い、8秒遅れてまったく同じフレーズがエコーのように追いかけるシーンを、人工音声により実現することが目論まれたからだ（初演時はデジタル・ディレイにより生の歌声をそのまま遅延させていた）。審美的な評価以前に、連続する2つの声が「似ている／似ていない」の成否は、すべての観客の耳が自動的かつ正確に判定するだろう。もし似ていなければ、即座に「違う」と認知され、作品解釈を混乱させるか、あるいは最悪の場合にはシリアスなシーンが滑稽なものとなり、観客を興醒めさせるだろう。そして何より本オペラのコンセプトの根幹にも抵触する。なぜならこのシーンは《14歳の少年》が自らの声を記号化する（データ化する）決定的に重要な場面であり、追従するエコーは「記号化された当人の声」として観客の耳に「聞こえる」必要があるからだ。つまり、今回目指すべきゴールは「人工音声だと気付かれないくらい似ていること」に他ならない。これまで私、いや恐らく世界中のコンピュータ音楽家の誰も、経験したことのない高いハードルである。

最初に構想を聞かされたとき、困難な挑戦であることはすぐに理解したが、密かに勝算はあった。というのも当該シーンの歌はほとんどアイウエオの母音だけで歌われており、この5音素だけに集中して制作すればよい。また人工音声は生の歌声を反復する「ディレイ」あるいは「エコー」なので、メロディや強弱については歌手がその場で歌うピッチ（声の高さ）と音量を解析して自動制御する方法が使える。つまり「声質」は人工でも「歌い方」は生の歌唱から「借用（サンプリング）」できるということだ。それが「似せる」うえで大きなアドバンテージになることは経験上分かっていたので、かなりの程度まで成功するだろうと直感した。ただし「人工音声だと気付かれない」クオリティにまで到達できるかどうかは全く未知数である。それゆえ、満足な成果が得られなかった場合の安全策として、初演時に使用したサンプリング、ディレイ技術も留保しておいた。

声作りのプロセス

声作りの作業に取りかかったのは2017年8月中旬だった。さかいに当該シーンのスコアを歌って録音してもらい、まず「ア」のシラブルで歌っている部分だけを切り取って編集する。そのオーディオファイルをループ再生し、歌声のピッチと音量のリアルタイム解析によって人工音声エンジンを駆動しながら、人工音声のスペクトルの形状が肉声のそれに近似するように、7つのバンドパ

ス・フィルターのパラメータ（3×7＝35個の数値）を手探りで調整して行く。いわば「粗彫り」の段階だ（図2）。

次に、もっぱら耳で調整する第二段階に入る。視覚的なスペクトルの形状は十分に似ているのに、聴感上は似ていないというケースは少なくない。スペクトル表示では見抜けなかったフォルマントが隠れていたり、ひとつの山として見えていたものが実は複数のフォルマントの相互作用の結果だったりするからだ。パラメータ値のひとつを変えれば他の複数の値にも影響するので、まるで何枚もの皿を同時に回す曲芸師のように作業を進める。しばらく調整を続けても見込みがない場合は第一段階に戻ってやり直しをしなければならない。さらに辛いのは「耳がバカになる」という自身の生理的問題との格闘である。長時間作業していると聴覚が馴化し客観的な判断ができなくなるのだ。モニターに使うスピーカーやヘッドフォンの音響特性のクセも判断を狂わせる。そんな時は耳を休ませ、スマートフォンの貧弱な内蔵スピーカーであえて鳴らしてみたり、三輪の客観的な感想で補正するなど、つねに自身の耳を疑う努力を重ねた。

図2：肉声と人工音声のスペクトルを比較しながらフォルマント・パラメータ値を調整する様子

2ヶ月以上こうした地道な作業を続けるなかで、それまで余所余所しかった別人の「ア」が、ある日一挙に、声主その人となって立ち上がる瞬間が訪れる。まるで手書きのイラストが高精細な写真に転じるような瞬間だ。しかし、どうやって突破したのかを遡って説明するのは難しい。音声学やデジタル音響合成の知見に基づく論理的なアプローチだけではなく、これまでの経験による直観と、さらに言えば「運」にも助けられているからだ。

いったん満足のいく「ア」ができると、個人の声紋の決定的特徴を数値化できたことになり、それを基準に「イウエオ」を作って行くので、この後の作業は比較的早い。こうして5種類すべての母音パラメータが完成したのが10月末だった。

人工音声だと気付かれないクオリティの実現

結果は予想以上の出来映えだった。誰の耳にも「人工音声だと気付かれない」クオリティが夢想ではなくなった。もちろん、この段階では録音されたテスト用歌唱との類似であり、実演時にそのクオリティになるかどうかは分からない。何より、生の歌声のピッチと音量検出によって人工音声をリアルタイムに駆動するこの制御方法は、舞台の現実空間で行おうとすると、音のフィードバック干渉により検出エラーを起こすリスクを伴う。そのためにスコア進行に基づくマイク入力のゲート処理、歌唱ピッチの範囲に特化したフィルター処理、エラー値のカットなど、想像できるトラブルに対して何重にも安全回路をプログラミングした。

これらすべての成果が試されたのは、12月1日に情報科学芸術大学院大学（IAMAS）で行われた立ち会い稽古の時である。さかいれいしうの生の歌唱から8秒遅れて人工音声のエコーが聞こえると、スタッフさえ（冒頭では私自身さえ）気付かないほどに人工音声は「歌手その人」の声であった。アイウエオすべての母音が、高低すべての音域で、強弱のダイナミクスを含めて、申し分なく似ていた。心配していたフィードバックによるトラブルも安全回路が機能して大きな問題にはならなかった。

こうして当初の目論みはほぼ理想的な形で実現できたのだが、そうなると贅沢にも新たな欲望が掻き立てられる。ここまで出来たのだからディレイでは不可能なこと、人工音声だからこそ可能になる表現を加えてもいいのではないか…。さまざま議論した結果、このシーンを機械学習の過程だとみなし、ブザー的な機械音が歌声へと徐々に変化したり、歌手が「イ」や「エ」で歌っているのにエコーは「ア」で返す…といった「演出」を控え目ながら追加することにした。ただし、この「演出」の当否には公演前から現在まで迷いが残っている。じっさい観客の感想のなかには「わざとらしく感じた」「トラブルかと思い集中力が途切れた」などの否定的意見も聞かれた。ただ、三輪も私も、もうひとつ別の難題の解決に意識を集中させていて、この迷いに対して別案を試すまでには至らなかった。

超えられない壁

「もうひとつ別の難題」とは、この人工音声エンジンをそのまま鍵盤キーボードで演奏しても、さかいれいしうの声には聞こえないという問題であった。本オペラで人工音声が登場するのは、上述のシーンに加え、その背後で延々と歌い続ける7音の多声コーラス、そして続くシーンにおいて4人のキーボード奏者が主人公の声で《天使の合唱》を演奏する場面の計3カ所だ。前二者は申し分のない成果を得たが、最後のキーボード演奏だけがどうしてもうまく行かない。

このことは、さかいれいしう「らしさ」、つまり特定個人の「声の

アイデンティティ」が、フォルマントのパラメータのみに還元できるものではないことを如実に示している。長年の経験から確信を持って言えるのだが、われわれが「声を聞く体験」の本質は「時間的変化の知覚」にある。これは音素の弁別から歌声のリアリティ認知まで一貫して妥当する話だ。その時間的変化のなかに、「らしさ」を知覚するうえでの重要な属性が含まれているのだ。

声は休むことなく絶えず変化しており、たとえ一定の音高と声量で「アー」と延ばされた声であっても、ピッチ、音量、フォルマントは目まぐるしく変動している。フォルマント音声合成のエコノミーは、ある瞬間の声成分のスナップショットからパラメータを抽出し、それを他の瞬間に適用（代用）しても、概ねうまく行く点にある。しかし、今回のように「声のアイデンティティ」を追求する極めて繊細なレベルになると、これでは大雑把にすぎる。これまで紹介して来た「人工音声のエコー」が満足できる結果を生んだ理由は、最適なフォルマント・パラメータを見つけられたことに加えて、ピッチと音量の特徴的な時間的変化を、生の歌声からそのまま「借用」できたことが大きい。

ピッチと音量の特徴的な時間的変化とは、マクロに見ればメロディの「歌い方」に現れる歌手の個性であり、ミクロに見れば微妙な「揺らぎ」に含まれる個性である。それらがキーボード演奏ではうまく表現されないのだ。公演直前まで、いや公演中もギリギリまで調整の努力を続けた(5)。

その努力のなかでも生の歌唱からミクロな揺らぎを「借用」したことの効果については少し触れておきたい。通常、人工音声に「揺らぎ」を与えてヒューマナイズする（人間らしくする）には、低周波のサイン波オシレータやランダム・ジェネレータ等を用いてピッチや音量をゆっくりと変化させ、人工的なビブラートやトレモロ効果を加える。しかし、いまは揺らぎに内包される「個性」の知覚が問題になっている。そこで、さかいが一定ピッチで素直に歌った録音をプログラムの中でループ再生し（もちろん無音のままで）、そのピッチ＋音量解析から揺らぎ成分を抽出し、キーボート演奏で入力される音高の整数値に加算する方法を試してみた。すると劇的な変化が生じた。それはビブラートやトレモロといった明示的に聞き分けられる運動（波のうねり）というより、声の「テクスチャー」あるいは「肌理」とでも表現すべき識別閾以下の変化（水面のさざ波）である。にもかかわらず確かに耳は捉えており、そこに歌手の声の個性を聞き取るのだ。

考えてみれば、キーボードとはひとつの「抽象」であり、声の時間的変化がつねに連続的であるのに対し、キーボードは音高の配列も、各鍵盤のON/OFF仕様も、つねに離散的である。この声の連続性と楽器（広く言えば「西洋音楽」というシステム）の離散性との橋渡しすることこそ、まさに「フォルマント兄弟」が挑戦し続けているテーマのひとつである。そのために兄弟は、和音平均化アルゴリズムによる微分音の指定や、アコーディオンの蛇腹による声帯緊張度の制御など、次々とユニークな工夫を考案して来たのだった。しかし、これらで培った経験を持ってしてもキーボードで演奏する人工音声の「アイデンティティ」にはあと一歩届かなかった。

以上、モノローグ・オペラ『新しい時代』におけるフォルマント音声合成の挑戦について書いて来た。われわれが、かくも執拗に「似せる」ことに粉骨砕身したのは、ひとえに本オペラの物語が、「声」を主人公の「実存」そのものとみなし、テクノロジーによる声の記号化を実存の昇華とみなす《教義》を掲げていたからに他ならない。《教義》はもちろんフィクションだが、フィクションと現実の混交、あるいはテクノロジーによってフィクションの一部が現実化される様を、演出効果ではなく、知覚や認知のレベルで「本当に」実現したいという欲望があったことは否めない（「本当」が何を意味するのかは極めて複雑な問題ではあるが）。

そうして出来上がった声は、声の生成メカニズムが完全に人工的・機械的であるのに対し、その制御は生きた身体からの「借用」に基づくハイブリッドなものであった。人間ー機械の「キメラ的結合の声」と言ってもいい。ただし、《教義》とは裏腹に、この声は生きた身体から離れて自律的に歌うことはない。主人公が現世で幻聴した《天使の合唱》を歌うことができない(6)。図らずもキーボードによる人工音声演奏の困難さが象徴的にそれを物語っている。

これを技術的な未熟さ、途上段階と評することはたやすい。また、最近のAIの深層学習的手法を持ってすれば、劇的な解決を期待できるのかも知れない。ただ、このキメラの声の周囲には、主体、実存、身体、精神、現実、虚構、現象、表象、などをめぐる問いの深淵が口をあけている。しばらくはここに立ち止まり、思索を深めたい。

【註】

(1) 三輪眞弘(作曲・脚本・音楽監督)、前田真二郎(演出・映像)、主演：さかいれいしう(14 歳の少年)
愛知公演：2017 年 12 月 8 日・9 日　愛知県芸術劇場小ホール（名古屋）
大阪公演；2017 年 12 月 16 日　ザ・フェニックスホール（大阪）
http://www.operanewera.com/

(2) 作品の詳細については次を参照されたい。 三輪眞弘、『三輪眞弘音楽藝術——全思考 一九九八 ー 二〇一〇 』、アルテスパブリッシング、2010 年、18-42 頁。

(3) 佐近田展康、「“兄弟式リアルタイム音声合成演奏システム”の概要と背景」、名古屋学芸大学メディア造形学部研究紀要、vol.6、pp.21-33、2013.

(4)「フレディの墓／インターナショナル Le tombeau de Freddie/L'Internationale」(2009)」亡きロックスター、フレディ・マーキュリーが日本語で革命歌「インターナショナル」を歌う「フォルマント兄弟」の録楽作品。Prix Ars Electronica 2009/Digital Music 部門 /Honorary Mention 賞受賞(オーストリア)。
同作品のビデオおよび「フォルマント兄弟」の諸活動については次を参照されたい。
http://formantbros.jp/

(5) ちなみに三輪からのリクエストは「キーボードから聞こえる声は、すでに記号化された段階なので機械的でよく、人間らしく聞こえる必要はない。ただし、さかいれいしうの声には聞こえてほしい」というものだった。

(6) キーボードによる人工音声の演奏は、 主人公の次のエピソードをリアライズしようとしている。
「告白します／偉大なる「新しい時代」の神よ／きのうの夜明け前に不思議なことが起こりました／ 4 人の天使が窓の外に現れ、 ボクの声でうたっていたのです／まるでボクが 4 人いるみたいでした／その歌は、 やがて高い声に変わり今まで聴いたこともないような、 美しい旋律になっていきました」
——『新しい時代』《告白》のシーンより。