身体構造推定に基づくキャラクター指向音声合成の設計枠組み

A Physically Grounded Design Framework for Character–Oriented Voice Synthesis

鷹宮　佳奈
独立研究者
kana.takamiya@kanade.tech

2026/04/17

概要

音声は声帯振動や声道形状、呼吸運動といった身体的制約のもとで生成される物理現象であり、話者の身体構造や身体状態と密接に関係している。近年の音声合成研究では、深層学習に基づく話者模倣型手法が主流となり、高品質な音声生成が可能となっている一方、声質を規定する身体的要因との対応関係は潜在表現として内部化され、明示的には扱われてこなかった。この傾向は、外見や身体設定が先行して設計されるキャラクター音声において、特に設計上の課題として現れる。

本稿では、心理学・認知科学における顔と声の統合知覚に関する知見を理論的動機として、キャラクターの外見から推定される身体的制約を音声生成の初期条件として位置づける。さらに、視覚情報から推定される体格、声道構造、声帯特性、呼吸容量などを物理・音響パラメータに対応付けることで、話者模倣や統計的類似性に依存しないキャラクター指向音声設計の設計基盤を提示する。

緒言

人の音声は、声帯振動、声道形状、呼吸運動といった身体的制約の下で生成される物理現象である。そのため、声質や話し方は話者の身体構造や身体状態と密接に関係していると考えられてきた。近年の音声合成研究は、物理モデルに基づく手法とデータ駆動型手法に大別され、高品質な音声生成を実現している。しかしその多くは、実在話者の再現または模倣を前提としており、実在人物の身体的あるいは統計的データを必要とする。

一方、心理・認知科学の分野では、顔と声が統合的に知覚されることが示されており、視覚情報と音声特性の間には一定の対応関係が存在することが報告されている。したがって、非実在のキャ

ラクターをデザインする場合も、視覚情報と音声の間における整合性がユーザ体験の上で重要な要素となる。しかしながら、存在しない仮想キャラクターのために、物理的一貫性を持ち、解剖学的に妥当な音声を構築する観点から整理された研究は少ない。

本研究では、この課題を「仮想話者デザイン問題（Virtual Speaker Design Problem: VSDP）」と定義する。VSDP とは、実在話者の音声データを前提とせず、キャラクターの外見および設定情報から推定される身体的制約を出発点として、物理的に一貫し、解剖学的に妥当な声を設計する問題である。

本稿が提示する枠組みは、キャラクターの外見から推定される身体的制約を音声生成の初期条件として位置づけ、音声を身体構造に根ざした物理現象として設計対象化する概念的枠組みである。この枠組みは、既存の音声合成と原理的に両立しつつ、実在話者の音声データを前提としない音声設計の視点を導入する点に特徴がある。視覚・文字情報から推定される体格、声道構造、声帯特性、呼吸容量、生理学的・文化的特徴などを、物理・音響パラメータに対応付けることで、外見と声の一貫性を保った音声設計を志向する。

本稿では、関連研究の整理に続き、身体構造推定と音声設計の枠組みを示し、その意義と今後の展開可能性について議論する。

1　先行研究

1.1　顔と声に基づく人物同一性の知覚

顔と声は、人間において独立に処理される情報ではなく、統合的に知覚されることが知られている。心理学および認知科学の分野では、顔画像と音声刺激を組み合わせた課題を通じて、顔と声の間に一定の知覚的一貫性が存在することが、繰り返し示されてきた（e.g. ,Kamachi et al. [1], Smith et al. [2], 伊東 et al.[3]）。これらの研究は、顔と声の対応関係が偶発的な連想ではなく、知覚処理に組み込まれた構造的傾向であることを示唆している。さらに工学分野においては、クロスモーダル人物同一性照合の文脈で、顔と声の対応関係を埋め込み空間上で学習する手法が検討されており、Nagrani et al. [4] や、Kim et al. [5] は、その学習可能性を示している。加えて、Mitchell et al. [6] は、顔と声の人間的リアリズムの不一致が不気味さの評価に影響しうることを示しており、外見と音声の不整合がユーザ体験を損なう可能性を示唆している。

これらの知見は、主として既存人物の照合や知覚・判断の段階を対象としており、視覚情報から音声を生成・設計する枠組みを直接的に提示するものではない。しかし、顔と声の一貫性が知覚的に重要であることを示す点で、キャラクターの外観情報を音声設計の入力として扱うことの妥当性を支持する基礎的知見と位置づけることができる。

1.2　従来型音声合成技術と話者性表現

従来の音声合成技術は、話者音声の収録と学習を前提とした話者模倣型のアプローチを主流として発展してきた。近年のニューラル TTS においては、話者性は埋め込み表現や潜在変数として内部化され、高品質な音声生成が可能となっている（Wang et al. [7, 8], Shen et al. [9], Jia et al. [10]）。

しかしながら、これらの話者埋め込みおよびスタイル制御型生成手法においては、話者特性が高次の潜在表現として統計的に学習される場合がほとんどであり、声質を規定する声帯長や声道形状、共鳴特性といった身体構造に由来する要因との対応関係は、明示的には扱われていない。

例えば Jia et al. [10] は、話者埋め込みが話者間の音声的類似性を効果的に捉えうることを示している一方で、その内部表現がどのような身体的・物理的要因に基づいて構成されているかについては明示的に論じていない。同様に、Global Style Token（GST）に基づくスタイル制御では、style token が複数の音響的・知覚的属性の混合として学習されることが報告されており、それらを個別の生理的・構造的要因へと対応付けて解釈することは困難である [8]。

このように、従来の話者モデリングおよびニューラル TTS 研究は、音声品質や自然性の向上において顕著な成果を挙げてきた。しかし、話者性やスタイルは主として潜在空間上の統計的表現として扱われており、それらがどのような身体的・解剖学的制約に基づくのかという因果的対応は明示されていない。

1.3　視覚情報と統合した音声生成技術

音声生成に視覚情報を導入する試みは、主にマルチモーダル機械学習の文脈において研究されてきた。顔画像や映像から得られる顔特徴に基づいて話者埋め込みを推定し、それを TTS モデルに条件付けることで、未収録話者の音声生成を行う手法が提案されている（e.g., Um et al. [11], Goto et al. [12]）。また、逆方向のアプローチとして、音声認識の精度を向上するために、話者の口唇運動や顔映像と音声の動的対応を活用する手法も提案されている（e.g., Afouras et al. [13]）。

これらの研究は、顔と声の対応関係がデータ駆動的に学習可能であることを示している一方で、その多くは視覚情報を埋め込み空間上の整合性として扱っており、解剖学的・因果的な対応関係を明示しない。そのため、外見から身体構造を媒介として声質を設計するという本稿の目的に対しては、なお限定的である。

1.4　生体現象としての「発話」と、物理的モデル化

音声学および音響学の分野では、声道形状や声帯特性と音声特性との関係が、物理モデルに基づいて研究されてきた（Fitch [14, 15], Dang and Honda [16]）。

近年では、3D 形状や解剖学的情報を用いて、顔や頭部形状と音声特性の対応関係を検討する研究も報告されている（e.g., Dediu et al. [17], Yang et al. [18], Pawelec et al. [19], Fitch and Giedd [20]）。特に Yang et al. [18] は、3D 顔形状が骨伝導を含む音色形成に関与しうるという解剖学的観点に基づき、3D 顔形状から話者埋め込みを推定する枠組みを提案している。

さらに、Pisanski et al. [21] は、フォルマントに基づく声道長推定値が身長や体格の指標となりうることをメタ分析により示している。加えて、Pisanski et al. [22] は、複数文化圏に属する約 700 名の成人を対象とした分析により、フォルマント間隔に基づく声道長推定値と体格との間に有意な相関が存在することを示した。

Pawelec et al. [19] も、顔・頭頸部の寸法と音響特徴の間に一定の相関があることを指摘しており、これらの知見は、身体形態が音響パラメータを体系的に制約していることを示す経験的証拠と解釈できる。

法医学分野においても、音声の特徴を個人識別や身体的特徴と関連付けて扱う研究がみられ（e.g., Rohatgi et al. [23]）、声と身体的特徴との関係は応用的文脈においても関心の対象となっている。

1.5　構音空間の物理モデルと音声合成

音声学の知見を音声合成に反映する研究としては、古典的には Ishizaka and Flanagan [24] による two–mass glottal model が基盤となり、機械工学・流体力学における流体–構造相互作用の観点から様々な整理がなされている（e.g., Birkholz [25]）。これらの知見を基に、Weerathunge et al. [26] は、神経・運動制御を計算論的に取り入れ、物理声帯モデルと結び付ける試みを行っている。さらに、呼吸運動を導入してより自然な発話を試みる Bernardet et al. [27] や Székely et al. [28] のような研究もある。

これらの研究は、身体構造と音声特性の関係を説明可能な形で扱う点において非常に重要である。しかしながら、その主眼は、観測された発話現象を物理的・生理的に再現または説明することにあり、解剖学的制約を仮想エージェントの生成設計原理として明示的に定式化するものではない。Bernardet et al. [27] は仮想的な人体モデルを検討する点で本研究と類似するが、各呼吸パラメータが視覚情報や外見由来の身体制約と体系的に接続されていない点に課題が残る。

これらに限らず、現行の手法の多くは、音声生成自体を既存の多話者 TTS モデルに依存しており、身体構造を音声設計の制約条件として直接操作可能な形で扱う枠組みは提示していない。

1.6 キャラクターコンテンツへの応用

アニメーション、ゲーム、VTuber といったキャラクターコンテンツの分野では、外見と声の整合性が重要な要素となる。近年、Kikuchi et al. [29] などにより、アニメ調キャラクター画像から音声を生成する研究も報告されているが、多くは既存キャラクター表現の統計的再現や知覚的一貫性の最適化を目的としている。

これらの手法は実用的な成果を挙げている一方で、キャラクターの外見や身体構造を設計変数として解釈可能な形で音声生成に接続する枠組みは、十分に整理されていない。

1.7 本稿の位置づけ

以上の先行研究を踏まえると、顔や外見と音声特性の対応関係は、知覚的・統計的観点からは広く示されている一方、それらを身体構造に基づく設計指向の音声生成へと接続する枠組みは十分に整理されていない。

従来の話者模倣型 TTS や視覚条件付き音声生成手法は、実在話者の音声データや既存キャラクター表現を学習資源として用いることで、高品質な音声生成を実現してきた。しかし、この前提は、実在話者に依存しない架空キャラクター音声の設計や、外見設定そのものを出発点とする音声構築には、必ずしも適合しない。

加えて、話者埋め込みやスタイルトークンに代表される近年の手法では、声質や話者性は高次の潜在表現として学習されることが多く、生成された声がどのような身体的・物理的制約を反映した結果であるかを設計者が解釈・制御することは容易ではない。この点は、キャラクター音声を「演技の結果」ではなく、「身体構造に基づいて生成される声」として捉えたい場合において、設計上の制約となり得る。

これに対し、本稿は、顔や外見と音声特性の対応関係を知覚的一貫性や統計的相関として捉えるだけでなく、身体構造に基づく設計制約として再編成する立場を取る。すなわち、外見情報から推定される身体的制約を音声生成の初期条件として位置づけることで、話者模倣や統計的類似性とは異なる設計指向の音声生成フレームワークを提示する。

この点で、本稿は従来の調音合成と立場を異にする。従来の調音合成が実在個人の解剖学的計測データに基づく再構築を主眼としてきたのに対し、本稿は、解剖学的モデリングを架空存在の設計空間を規定する制約体系として再解釈する。

したがって、本稿の新規性は、新たな音響モデルの導入そのものにあるのではなく、物理モデリ

ングの目的を「観測された話者の再現」から「原理に基づく仮想話者の構築」へと転換し、そのための概念的基盤を提示する点にある。この立場は、データ駆動型音声合成を否定するものではなく、それと両立しながら、話者模倣や統計的類似性に依存しない音声設計の視点を与えるものである。

2　全体パイプライン

本章では、キャラクターの外見から推定される身体的構造を出発点とし、音声生成における物理・音響パラメータを構成・調律するための概念的枠組みを示す。本稿の立場は、音声を話者模倣の対象としてではなく、身体構造に制約された物理現象として捉える点にある。したがって、本枠組みは概念的整理を主目的とし、具体的な実装や性能評価そのものを扱うものではない。

本枠組みは三つの層から構成される。第一に、視覚情報から身体構造を推定する層である。第二に、推定された身体構造を音声生成パラメータへと対応付ける層である。第三に、それらを創作者が介入可能な設計変数として整理・調律する層である。

本稿で提案するキャラクター指向音声合成の全体的な処理理念を Figure 1 に示す。

本枠組みは、キャラクターの立ち絵や 3D モデル、設定資料といった視覚情報を入力とし、外見から解釈される身体的特徴を経て、身体構造の範囲推定、物理・音響パラメータの生成、既存 TTS モジュールによる音声生成へと至る一連の流れを明示的に整理したものである。

特に、身体構造に由来する制約、身体・精神状態に依存する可変要素、および創作者が個別に調整可能な要素を区別して扱う点に、本枠組みの特徴がある。この整理により、視覚情報から直接声を推定するのではなく、身体構造とその状態を媒介とした因果的な設計プロセスとして、音声生成を位置づけることが可能となる。

2.1　視覚情報に基づく身体構造推定

キャラクターの外見には、頭部形状、顎や頸部の太さ、体格、年齢感、性差といった身体的特徴が視覚的に表現されている。本研究では、これらの視覚情報および設定資料に記述された文字情報を、声帯長や声道長、共鳴腔の相対的容量といった音声生成に関与する身体構造推定の入力情報として扱う。

これらの推定は単一の決定値を与えるものではなく、「一定の幅を持つ制約条件」として表現されることを想定する。すなわち、外見から直接一意に声を決定するのではなく、「この外見であれば取り得る身体構造の範囲」を定め、その範囲内で音声設計を行うという立場を取る。このような推定の考え方は、身体構造と音声特性の相関を示した音声学・法医学・認知科学の先行研究と整合的であり、外見と声の一貫性を説明可能な形で扱うための基盤となる。

加えて、キャラクター設計においては、視覚と並行して、実年齢、心理特性、生活習慣、発話特徴（方言や構音癖）といった、外見から一律に推定することが困難な特徴も文字情報によって与えられる。これらは、上記の推定範囲をさらに具体化する上で有効である。たとえば、生理的状態や発達段階に由来する解剖学的・機能的制約、あるいは特徴的な構音特性に関わる変化範囲を絞り込

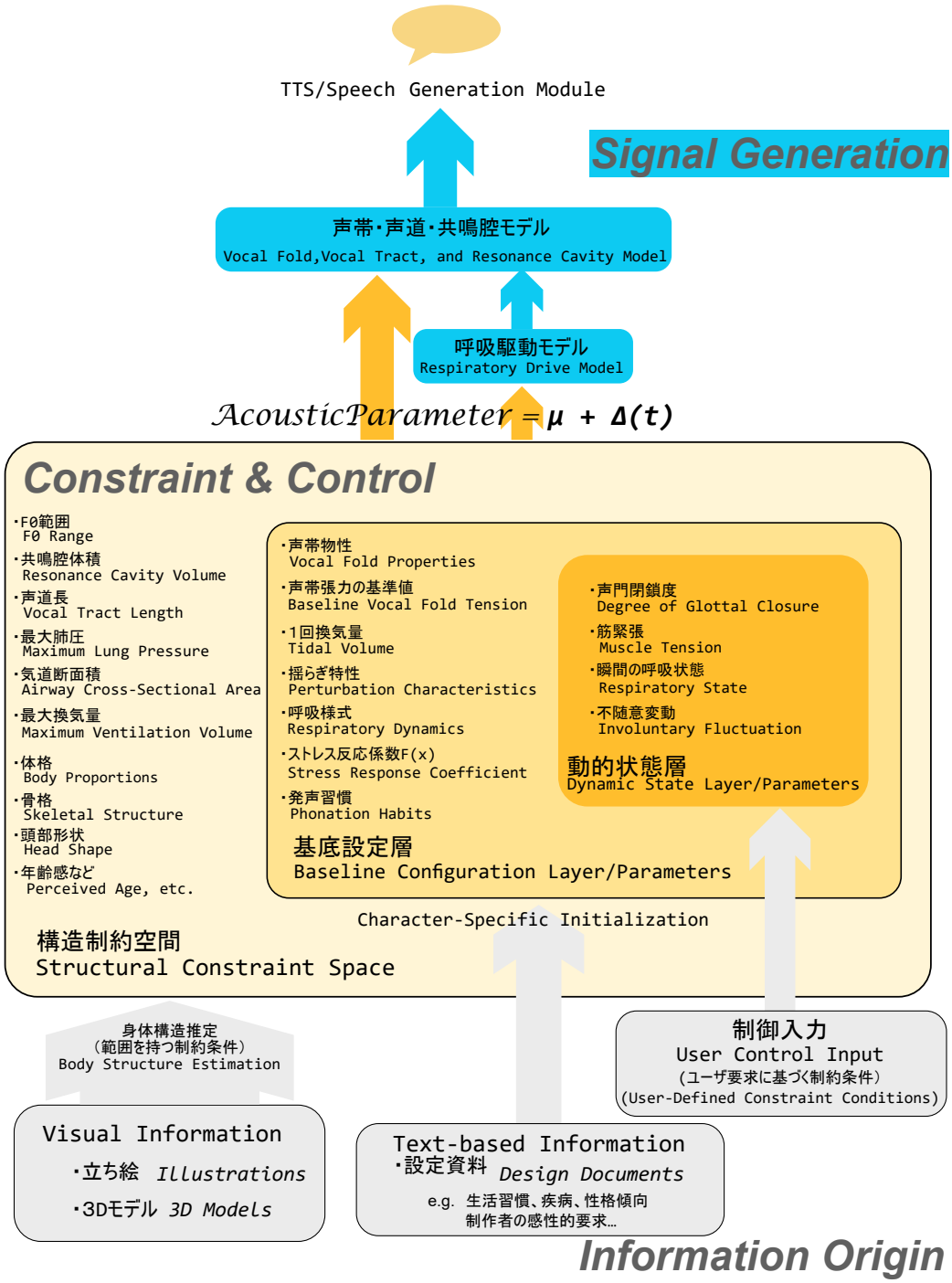

図 1　Overview of the proposed character–oriented voice synthesis framework. Visual and text–based character information is mapped to a structural constraint space, from which baseline and dynamic parameters are organized and connected to speech generation modules.

む手がかりとなる。

このような複合的視点からの推定により、一次的に生成パラメータを設定する上での調整可能範囲が与えられる。この推定された調整可能範囲は、身体的特徴に基づくという特性上、比較的長期にわたり安定的であると想定される。したがって、本稿ではこれらを音声設計における「基底変数」と定義する。

2.2 動的に変動しうるパラメータ

発声は、横隔膜および胸郭の運動によって生じた呼気流が声帯を振動させ、その振動が声道内で共鳴することにより音響が生成される現象である。この過程は、生体において随意・不随意の区別はあるものの、いずれも神経刺激とそれに伴う筋活動によって制御されている。

筋活動と音響特性の関係については、声帯振動の物理的基礎および基本周波数との関係（Titze [30]）、甲状披裂筋活動と基本周波数制御の関係（Titze et al. [31]）、ならびに声道形状の変化と共鳴特性の関係（Sundberg [32], Story [33]）など、多くの研究により検討されてきた。これらは、音声特性が神経–筋制御に依存する動的過程であることを示す知見である。したがって、発話に関与する呼気圧、声帯緊張度、構音器官の位置、共鳴腔形状といった諸量は、身体的特徴によって与えられる構造範囲内において、時間とともに連続的に変動する動的量として定式化することができる。

本稿では、これらの変化量を「動的変数」と呼び、基底変数によって規定される許容状態空間の内部を時間的に遷移する制御関数として概念化する。このとき、基底変数は状態空間の幾何学的境界を与え、動的変数はその内部における軌道を形成する。

2.3 意図・状態入力と制御層

発話は、身体構造と神経–筋制御のみによって決定されるものではない。実際の音声生成においては、発話意図、情動状態、文脈的要求、さらには演技指示や社会的役割といった要因が、動的変数の時間発展に影響を与える。本稿の枠組みにおいて、これらの要因は「意図・状態入力」として位置付けられる。すなわち、基底変数が状態空間の幾何学的境界を与え、動的変数がその内部を時間的に遷移する軌道を形成するのに対し、意図・状態入力はその軌道の方向、速度、振幅を決定する制御項として機能する。

ここでいう意図・状態入力には、少なくとも以下のような要素が含まれる。

- 発話意図（強調、疑問、命令などの発話機能）
- 情動状態（緊張、安心、怒り、悲嘆など）
- 身体状態（疲労、覚醒度、ストレス反応）
- 創作者による演技・表現指示

これらは、基底変数によって規定される許容範囲を逸脱することなく、動的変数の時間的変化様式を変調させる入力として働く。たとえば、情動状態の変化は呼気圧や声帯緊張度の変動幅に影響を与え、発話意図は構音運動の速度や共鳴特性の時間変化を調整する。

重要なのは、本枠組みがこれらの入力を「感性的属性」としてではなく、状態空間に対する制御入力として概念化する点にある。この整理により、キャラクター音声は単なる統計的類似性の結果ではなく、構造的制約・動的運動・意図的入力の三層構造によって生成される因果的過程として理解される。

このように神経制御と物理声道モデルの統合を志向する研究としては、DIVA モデルおよびその拡張である LaDIVA [26] が挙げられる。これらのモデルは、発話運動を感覚フィードバックを含む制御理論的枠組みとして記述しうることを示している。本枠組みは、これらの知見と整合しつつ、その制御構造を創作キャラクターの設計空間へ接続するための抽象的整理を与える点に特徴がある。

2.4 三層構造に基づく設計指向フレームワーク

以上の三層構造により、本稿は、視覚情報から身体構造を推定し、その構造的制約のもとで動的制御と意図入力を統合する設計指向の音声生成フレームワークを提示する。本枠組みでは、基底変数が構造的制約を与え、動的変数がその内部を時間的に遷移し、意図・状態入力がその軌道を制御する。この整理により、音声生成は統計的模倣ではなく、構造・運動・意図の相互作用として記述される設計問題として理解される。

3 身体構造推定と基底変数

本章では、前章で提示した三層構造のうち、身体構造推定および基底変数に関わる層を具体化する。具体的には、外観由来の制約空間を Constraint Ranges として整理し、その内部から選択される長期安定な基準値（基底変数）との対応関係を明示する。特に本章では、声道モデルおよび声帯モデルとの接続面に着目し、基底変数がどの物理量に対応するかを整理する。

本稿では、身体構造と音声生成を接続するための基底変数として、声道幾何（声道長および断面積関数）と声帯物性（質量、剛性、減衰）を最小構成として採用する。この選択は、声道を一次元伝送線路として近似するモデル（e.g., Story [33]）および two-mass glottal model（Ishizaka and Flanagan [24]）に基づく。これらの変数は、物理的音声生成において一次的な支配因子として広く採用されている。加えて、外観や身体構造と解剖学的に接続可能な量であるという点で、設計指向の枠組みに適している。

また、これらに対する入力条件の一つである呼気圧については、呼吸機能の器質的側面に着目し、最大換気量と最大腹圧を基底変数として設定する。

なお、本稿は網羅的な生理変数体系を提示するものではなく、設計問題として最小限必要な変数群を明示することを目的とする。

3.1 外観から Constraint Ranges への写像

Constraint Ranges は、声道長、断面積分布、声帯質量および剛性などの物理量に対し、外観と既存知見に整合する範囲として設定される。この範囲は、統計的推定に基づく場合と、創作上の設

計意図に基づき再調整される場合とを含むが、いずれも物理的一貫性を維持する条件のもとで定義される。以降、本稿では Constraint Ranges を物理パラメータの許容領域として扱う。

外観から身体構造を推定する研究としては、顔画像から体格や年齢を推定する画像認識手法、ならびに保存された解剖学的構造から発声可能性を再構成する研究などが報告されている（e.g., Angulu et al. [34], Wen and Guo [35], Baudouin et al. [36]）。このような研究では、推定値に加えて、その精度や適用限界が議論されることが多い。

本研究では、このような推定値を Constraint Ranges の中心値（中央値または代表値）として扱い、推定精度に応じた誤差幅を含む範囲を制約領域として定義することが可能であると考える。すなわち、外観から得られる推定値 $\hat{\theta}$ に対し、その推定信頼性に応じて一定の幅を持つ領域を設定することで、物理的に許容される構造空間を与える。この範囲は、統計的根拠に基づいて設定される場合と、創作上の要請に応じて再調整される場合とがあるが、いずれも物理的一貫性を維持することを前提とする。

3.2 声道における基底変数

声道は、声門から口唇に至る音響伝達路として機能し、その幾何学的構造は生成音声のスペクトル特性を一次的に規定する。一次元伝送線路モデルにおいて、声道は位置 $x \in [0, L]$ に沿った断面積関数 $A(x)$ により記述される。

本稿では、声道に関する基底変数として、

1. 声道長 L
2. 断面積関数 $A(x)$ の基準分布

を最小構成として採用する。

声道長 L は、フォルマント間隔に対して一次的な支配因子として働く。理想化された閉管共鳴近似においては、共鳴周波数 F_n は概ね

$$F_n \approx \frac{(2n-1)c}{4L}$$

と表される。この関係は、身体的スケールが音響特性に直接的に影響を及ぼすことを示す代表的な例である。

一方、断面積関数 $A(x)$ は、母音間のスペクトル差異や共鳴ピークの鋭さ、帯域幅などを規定する。顎の形状、口腔容積、顔面輪郭、頸部の太さといった外観情報は、これらの解剖学的形状に間接的に対応するため、外観由来の Constraint Ranges を $A(x)$ の許容分布として解釈することが可能である。

したがって、本稿における声道の Constraint Ranges は、

$$L \in [L_{\min}, L_{\max}]$$

および

$$A(x) \in \mathcal{A}_{\text{allowable}}$$

として定式化される。ここで $\mathcal{A}_{\text{allowable}}$ は、物理的連続性および解剖学的妥当性を満たす断面積分布の集合を表す。

これらの許容領域の内部から選択される代表値が、声道における基底変数として機能する。すなわち、基底変数は声道構造の長期安定な基準形状を与え、動的変数はその形状に対する時間的変形として扱われる。

以上の整理により、外観から推定された身体的特徴は、声道長および断面積分布の制約条件として物理モデルに接続される。

3.3 声帯における基底変数

声帯は、呼気流をエネルギー源として自励振動を生じる流体–構造連成系であり、その振動特性は生成音声の基本周波数および波形形状を一次的に規定する。two-mass glottal model に代表される物理モデルでは、声帯は質量–ばね–ダンパ系として近似され、その動的挙動は幾何学的スケールおよび組織物性によって記述される。

本稿では、声帯に関する基底変数を、外観および解剖学的知見と接続可能な最小構成として以下の三群に整理する。

(1) 幾何学的基底変数

- 声帯有効長 l_g
- 下質点および上質点の有効厚み d_1, d_2
- 必要に応じて後部間隙（posterior chink）の基準値

声帯長は、身体スケールおよび性差と解剖学的に関連しうる量であり、振動特性および張力分布に影響を与える。厚みは、粘膜波動の伝播様式および有効質量の決定に寄与する幾何学的要因である。

(2) 組織力学的基底変数

- 有効質量 m_{eff}
- 有効剛性 k_{eff}
- 有効減衰 r_{eff}
- 上下質点間結合剛性 k_c
- 必要に応じて衝突剛性

two-mass モデルの単純化近似において、声帯振動の基本周波数は概ね

$$F_0 \approx \frac{1}{2\pi}\sqrt{\frac{k_{\text{eff}}}{m_{\text{eff}}}}$$

と表される。この関係は、組織の弾性特性および有効質量が基本周波数を規定する主要因であることを示す代表的な近似式である。

ここで重要なのは、これらの量が発話ごとに大きく変動するものではなく、個体の組織特性や身体構造に依存する長期安定な基準値として扱える点である。本稿では、これらを声帯側の基底変数

と位置づける。

(3) Constraint Ranges との対応　外観および設定情報に基づく Constraint Ranges は、声帯側では以下のように形式化される。

$$l_g \in [l_{\min}, l_{\max}]$$

$$m_{\text{eff}} \in [m_{\min}, m_{\max}]$$

$$k_{\text{eff}} \in [k_{\min}, k_{\max}]$$

これらの範囲は、解剖学的統計、年齢層、性差、体格情報、あるいは創作上の設計意図に基づき設定される。ただし、その定義は常に物理的一貫性（自励振動が成立する条件）を満たす範囲内でなければならない。

この許容領域の内部から選択される代表値が、声帯の基底変数として機能する。

(4) 動的変数との分離　本節で定義した基底変数は、声帯構造および組織物性の長期安定な基準値である。一方、発話中に時間的に変動する要素としては、

- 声門下圧
- 内転度および前発声形状
- 筋活動（輪状甲状筋、甲状披裂筋など）

などが存在する。

これらは基底変数を直接変更するのではなく、その上で有効剛性や張力を時間的に変調する制御入力として機能する。本研究ではこれらを動的変数および制御項として次章で扱う。

(5) 接続面の整理　声帯モデルの出力は、声門面積または体積速度として声道モデルへ接続される。この接続面の定義により、声帯の基底変数は声道側の音響応答と因果的に結び付けられる。次章では、このモデル間の接続について概説する。

4　動的変数と微細変動の取り扱い

前章では、身体構造および組織物性に対応する長期安定な基準値を基底変数として整理した。本章では、それらの基底変数の上で時間的に変化する要素を動的変数として整理する。

音声生成は、声帯振動、呼吸運動、調音器官の運動などが時間的に変化することによって実現される動的過程である。したがって、音声波形の実際の時間構造は、前章で定義した基底変数だけでは決定されず、それらに対する時間的な変調によって形成される。

本稿では、この時間的変調を与える要素を動的変数として整理し、基底変数とは概念的に分離して扱う。

4.1　呼吸駆動と声門下圧

音声生成のエネルギー源は呼気流であり、声門下圧（subglottal pressure）は声帯振動を駆動する主要な入力変数である。声門下圧は呼吸運動および呼気制御によって時間的に変化し、発話強度や

声量の変化に直接的に対応する。その範囲は、体幹部の外見および設定資料から推定される最大換気量や呼吸様式によって制約される。

声帯振動モデルにおいては、声門下圧は自励振動を成立させる主要な駆動条件として扱われる。したがって、本稿では、声門下圧を声帯側の動的入力として扱い、基底変数として定義された声帯質量や有効剛性などを時間的に変調する駆動項として位置づける。

この整理により、身体構造に依存する長期的特性と、発話時の呼吸制御に依存する短期的変動を明確に区別することが可能となる。

4.2 喉頭筋活動と声帯張力制御

声帯振動特性は、輪状甲状筋（CT）、甲状披裂筋（TA）などの喉頭筋活動によって調節される。これらの筋活動は、声帯張力および有効剛性を変化させることで、基本周波数や声質の変化に寄与する。

本稿では、これらの筋活動を、基底変数を直接変更する要素としてではなく、基底変数に対する時間依存の制御入力として扱う。

すなわち、

- 声帯質量
- 声帯幾何
- 組織弾性

などの長期安定な基準値は基底変数として固定され、その上で筋活動が有効剛性や張力を変調する形で振動状態が決定される。

この整理は、声帯構造そのものが発話ごとに変化するわけではなく、筋活動によって振動条件が調整されるという生理学的理解とも整合する。

4.3 調音運動と声道形状の時間変化

声道形状は、顎、舌、口唇などの調音器官の運動によって時間的に変化する。前章で定義した断面積関数は、声道の基準形状としての役割を持つが、実際の発話ではこれが時間関数として変形する。

本稿では、この声道形状を

$$A(x,t) = A_{\mathrm{base}}(x) + \Delta A(x,t)$$

のように表現し、

$$A_{\mathrm{base}}(x)\colon \text{基底形状}$$

$$\Delta A(x,t)\colon \text{時間変形}$$

として整理する。

この分離により、身体構造に由来する声道形状と、発話内容に応じた調音運動を概念的に区別することが可能となる。

4.4　心理状態および認知的要因

音声特性は生理的要因だけでなく、感情状態や心理状態によっても変化することが知られている。例えば、緊張状態では声門下圧の増加や声帯内転の変化が観察される場合があり、声質の変化として知覚される。精神的緊張状態の亢進は、一般的に不随意的な筋緊張の上昇を介して、呼吸制御、喉頭筋活動、調音運動の各系に広く影響しうる。このような中間的制御層を導入することで、感情状態を直接音響特徴へ写像するのではなく、神経–筋制御の変調として記述する、より簡潔な演技的変調の枠組みが得られる。

したがって、これらの場合も、主として以下の動的要素の変調として整理できる。

- 呼吸制御
- 声門下圧
- 喉頭筋活動
- 調音運動

以上より、心理状態や感情状態は、神経–筋運動関係の観点から動的変数の制御項として扱うことが可能であると考える。この整理は、キャラクターの感情表現を音声生成に導入する際にも有用であり、身体構造に依存する基底特性と、演技的表現としての声質変化を区別して扱うことを可能にする。

5　将来の展望

本稿では、キャラクターの外観情報から身体構造を推定し、それを物理的音声生成モデルへ接続する概念的枠組みを提示した。本章では、この枠組みの今後の発展方向について整理する。

5.1　外観から身体構造推定への拡張

近年のコンピュータビジョン研究では、顔画像や全身画像から体格、年齢、姿勢などを推定する手法が多数提案されている。これらの手法は主として人物認識や人体再構成の文脈で発展してきたが、身体寸法推定という観点では音声生成モデルとの接続可能性を持つ。

今後は、頭頸部形状推定、顎形状および顔面骨格推定、身体スケール推定などの技術を用いることで、外観情報から声道長や声帯長などの生理パラメータをより体系的に推定できる可能性がある。

5.2　物理モデルとデータ駆動モデルの統合

本研究では、声道モデルおよび声帯モデルを中心とした物理モデルを前提としている。一方、近年の音声合成研究では、ニューラルネットワークを用いたデータ駆動的な音声生成が主流となっている。今後は、物理モデルによる制約空間の生成とニューラルボコーダによる波形生成を組み合わせた、ハイブリッド型音声合成が有望な方向性と考えられる。

このような構成では、物理モデルは説明可能な設計空間を提供し、ニューラルモデルは音響的自然性を補完する役割を担う。

5.3 キャラクターデザインへの応用

本稿で提示した枠組みは、音声合成技術だけでなく、キャラクターデザインの観点からも応用可能である。従来、キャラクターの声は声優の演技や既存音声の選択に依存する場合が多かったが、本研究の枠組みでは、外観、身体設定、声質設計を統合的に扱うことが可能となる。

これにより、キャラクターの身体設定と音声特性の整合性を設計段階で検討することが可能となり、より一貫したキャラクター表現を実現できる可能性がある。

6 結語

本研究では、キャラクターの外観情報から身体構造を推定し、物理的音声生成パラメータへ接続する概念的枠組みとして、仮想話者デザイン問題（Virtual Speaker Design Problem: VSDP）を提案した。

提案枠組みでは、

1. 外観情報から身体構造を推定する層
2. 身体構造を物理パラメータの制約空間（Constraint Ranges）として整理する層
3. 基底変数および動的パラメータを用いて音声生成モデルへ接続する層

という三層構造を明示した。

さらに、声道幾何および声帯物性を基底変数として整理し、それらと呼吸駆動や筋活動などの動的要素を概念的に分離することで、身体構造に基づく音声生成の設計空間を示した。

本稿は、具体的な音声合成システムの実装を目的とするものではなく、外観情報、身体構造推定、Constraint Ranges、基底変数、動的変数という各層の関係を設計問題として記述するための概念的基盤を提示するものである。

このような整理は、音声学、認知科学、音声合成研究の知見を横断し、創作キャラクターの音声設計に対して説明可能な設計視点を与えるものと考えられる。

参考文献

[1] M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson, “Putting the face to the voice’: Matching identity across modality,” *Current Biology*, vol. 13, no. 19, pp. 1709–1714, 2003.

[2] H. M. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey, “Matching novel face and voice identity using static and dynamic facial images,” *Attention, Perception, & Psychophysics*, vol. 78, no. 3, pp. 868–879, 2016.

[3] 伊東 裕司，高山 宏，日比谷 潤，渡辺 茂，“顔と声の関連性の判断: 人物の同一性について”,

哲學，vol. 98，pp. 123–139，1995.

[4] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8427–8436.

[5] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 276–292.

[6] W. J. Mitchell, K. A. Szerszen Sr, A. S. Lu, P. W. Schermerhorn, M. Scheutz, and K. F. MacDorman, "A mismatch in the human realism of face and voice produces an uncanny valley," *i-Perception*, vol. 2, no. 1, pp. 10–12, 2011.

[7] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[8] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 5180–5189.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[10] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[11] S. Um, J. Kim, J. Lee, and H.-G. Kang, "Facetron: A multi-speaker face-to-speech model based on cross-modal latent representations," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 281–285.

[12] S. Goto, K. Onishi, Y. Saito, K. Tachibana, and K. Mori, "Face2speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image." in *Interspeech*, 2020, pp. 1321–1325.

[13] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[14] W. T. Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," *The Journal of the Acoustical Society of America*, vol. 102, no. 2, pp. 1213–1222, 1997.

[15] ——, "The evolution of speech: a comparative review," *Trends in cognitive sciences*, vol. 4,

no. 7, pp. 258–267, 2000.

[16] J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3374–3383, 1996.

[17] Dediu Dan, Jennings Emily M, van' t Ent Dennis, Moisik Scott R, Di Pisa Grazia, Schulze Janna, de Geus Eco JC, den Braber Anouk, Dolan Conor V, Boomsma Dorret I, "The heritability of vocal tract structures estimated from structural mri in a large cohort of dutch twins", Human genetics, vol. 141, no. 12, pp. 1905–1923, 2022.

[18] Z. Yang, Z. Wu, Y. Shan, and J. Jia, "What does your face sound like? 3d face shape towards voice," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 905–13 913.

[19] Ł. P. Pawelec, K. Słowik, and A. Lipowicz, "Dimensions of the face, head, and neck affect acoustic parameters in polish males and females," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7659–7663.

[20] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[21] K. Pisanski, P. J. Fraccaro, C. C. Tigue, J. J. O'Connor, S. Röder, P. W. Andrews, B. Fink, L. M. DeBruine, B. C. Jones, and D. R. Feinberg, "Vocal indicators of body size in men and women: a meta-analysis," *Animal Behaviour*, vol. 95, pp. 89–99, 2014.

[22] K. Pisanski, B. C. Jones, B. Fink, J. J. O'Connor, L. M. DeBruine, S. Röder, and D. R. Feinberg, "Voice parameters predict sex-specific body morphology in men and women," *Animal behaviour*, vol. 112, pp. 13–22, 2016.

[23] S. Rohatgi, V. Gupta, B. Yadav, and B. Yadav, "Forensic phonetics: A linguistic approach," *Journal of Punjab Academy of Forensic Medicine and Toxicology*, vol. 18, no. 2, pp. 36–41, 2018.

[24] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell system technical journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[25] P. Birkholz, "A survey of self-oscillating lumped-element models of the vocal folds," in *Konferenz Elektronische Sprachsignalverarbeitung*. TUDpress, Dresden, 2011, pp. 47–58.

[26] H. R. Weerathunge, G. A. Alzamendi, G. J. Cler, F. H. Guenther, C. E. Stepp, and M. Zañartu, "Ladiva: A neurocomputational model providing laryngeal motor control for speech acquisition and production," *PLoS computational biology*, vol. 18, no. 6, p. e1010159, 2022.

[27] U. Bernardet, S.-H. Kang, A. Feng, S. DiPaola, and A. Shapiro, "Speech breathing in virtual humans: An interactive model and empirical study," in *2019 IEEE Virtual Humans and Crowds for Immersive Environments (VHCIE)*. IEEE, 2019, pp. 1–9.

[28] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Breathing and speech planning in spontaneous speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics,*

Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7649–7653.

[29] 菊池遥斗，能勢隆，伊藤彰則，“キャラクター画像からの音声合成に向けた動向分析”，情報処理学会研究報告，vol. 2024-MUS-140，no. 11，pp. 1–10，2024.

[30] I. R. Titze, “The physics of small-amplitude oscillation of the vocal folds,” *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1536–1552, 1988.

[31] I. R. Titze, E. S. Luschei, and M. Hirano, “Role of the thyroarytenoid muscle in regulation of fundamental frequency,” *Journal of Voice*, vol. 3, no. 3, pp. 213–224, 1989.

[32] J. Sundberg, *The Science of the Singing Voice*. DeKalb, IL: Northern Illinois University Press, 1987.

[33] B. H. Story, “A parametric model of the vocal tract area function for vowel and consonant simulation,” *The Journal of the Acoustical Society of America*, vol. 117, no. 5, pp. 3231–3254, 2005.

[34] R. Angulu, J. R. Tapamo, and A. O. Adewumi, “Age estimation via face images: a survey,” *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, p. 42, 2018.

[35] L. Wen and G. Guo, “A computational approach to body mass index prediction from face images,” *Image and Vision Computing*, vol. 31, no. 5, pp. 392–400, 2013.

[36] R. Baudouin, A. Amelot, S. Nicolleau, I. Huynh-Charlier, L. Crevier-Buchman, S. Hans, and P. Charlier, “Voice of mummified king henri iv recreated via 3d functional vocal tract model,” *Journal of Voice*, 2026.

利益相反

著者は、本稿に関して開示すべき利益相反はない。

AI 利用に関する注記

本稿の執筆にあたり、著者は汎用人工知能（AI）プログラムを、文章表現の推敲、構成整理、英語表現案の作成、参考文献表記の整形補助のために利用した。研究の着想、論旨の構成、引用文献の選定、本文内容の確認および最終的な記述の決定はすべて著者が行った。