修 士 論 文

声区表現を可能とする歌声合成を目的とした ARX-LF モデルの制御法に関する研究

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

元田 紘樹

2013 年 3 月

修 士 論 文

声区表現を可能とする歌声合成を目的とした ARX-LF モデルの制御法に関する研究

指導教員　赤木 正人 教授

審査委員主査　赤木 正人 教授
審査委員　党 建武 教授
審査委員　鵜木 祐史 准教授

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

1110061 元田 紘樹

提出年月: 2013 年 2 月

概 要

計算機上で人工的に歌声を生成・加工する歌声合成の分野は，音声科学における重要な分野の一つである．より高品質かつ多様な歌声合成システムを構築することは，音楽情報処理分野への貢献のみならず，音声の生成・知覚に関する新たな知見を与える上でも，重要な役割を担っている．これに対し，人のように自然で多様な歌声合成は，未だ実現に至っていない．その原因の一つとして，‘声区’の表現が挙げられる．

声区とは，人の声域を発声法と声質の相違によって区分したものである．人は，声区ごとの声帯振動様式の違いを歌唱訓練によって習得することで，広い音域を自然な声質で歌うことができる．一方で，歌声合成の分野では，そのような声区表現には十分に対応できていないため，高音域及び低音域で不自然な合成音を生じる．高音域及び低音域における合成音の自然性を向上させる方法として，声区ごとの声帯音源特性を付加することが考えられる．そのためには，声帯音源特性を記述できるモデルが必要となる．

本研究では，声区表現を可能とする歌声合成に向けた，声帯音源特性の制御法の検討を目的とする．目的を遂行するため，音声生成過程を模擬することで，声区表現のための声帯音源特性の制御が可能である ARX-LF モデルを適用する．ARX-LF モデルが持つ，声帯音源特性に対応する複数の ARX-LF パラメータを，音高の変化に伴い適切に変化させるように制御モデルを構築することで，声区ごとの声帯音源特性を付加できるようになる．声区表現を可能とするための歌声合成システムの枠組みを提案し，ARX-LF パラメータ制御モデルを構築した．そして，歌声合成音を作成し，客観評価と主観評価を実施することで ARX-LF パラメータ制御モデルの評価を行なった．これらの結果を報告する．

まず，ARX-LF モデルによる分析・制御・合成を行うことで，声区ごとの声帯音源特性を付加できる歌声合成システムを提案した．次に，声区表現に対する ARX-LF モデルの有効性を検証するために，声区ごとの ARX-LF パラメータを分析したところ，先行研究の声帯音源特性の知見に合致した結果が得られた．分析結果に基づいて，それぞれの ARX-LF パラメータ制御モデルを構築した．各声区内を線形で補間することで，作成する歌声合成音の音高ごとに，適切に各パラメータが制御されるようにした．

そして，提案システムによって歌声合成音を作成し，客観評価と主観評価を行なった．客観評価のために，低周波数域におけるスペクトル傾斜を分析し，声区ごとに比較を行なったところ，falsetto では急峻な傾き，vocal fry では緩やかな傾きが得られた．分析結果の妥当性を検証するために，人の歌声についても分析を行なったところ，歌声合成音と同様の傾向が得られ，歌声合成音におけるスペクトル傾斜の再現性が示された．主観評価のために，聴取実験を実施して高音域及び低音域における聴取印象の比較を行った結果，falsetto では気息性，vocal fry では粗慥性といった，声区ごとの典型的な聴取印象が得られた．さらに，音域が広いほど，声区ごとの声帯音源特性の付加が重要である可能性が示唆された．これらの結果から，声区表現が可能な歌声合成に対する，ARX-LF パラメータ制御モデルの有効性が示された．

本研究で提案した，音声生成機構からのアプローチに基づく歌声合成は，人のような歌声合成システムの実現だけでなく，音声生成機構・音響的特徴・知覚の相互関係性の解明にも繋がるものであると考える．

目 次

図 目 次

表 目 次

第1章　序論

1.1　はじめに

計算機上で人工的に歌声を生成・加工する歌声合成の分野は，音声科学における重要な分野の一つである．より高品質かつ多様な歌声合成システムを構築することは，音楽情報処理分野への貢献のみならず，音声の生成・知覚に関する新たな知見を与える上でも，重要な役割を担っている．これに対し，人のように自然で多様な歌声合成は，未だ実現に至っていない．その原因の一つとして，’声区’の表現が挙げられる．

声区とは，人の声域を発声法と声質の相違によって区分したものである．歌声は話声に比べて利用する音域が非常に広く，一つの声区でこの広範な音域をカバーするのではなく，複数の声区を使い分けて行っているといわれている．そのため，人は複数の声区を使い分けできるように歌唱訓練を受けることで，広い音域を歌えるようになると考えられる．一方で，歌声合成の分野は，このような声区表現に十分に対応できていないため，高音域や低音域で不自然な合成音を生じてしまう．

人の歌唱において，広い音高を自然に歌えることは最重要視される要素の一つである．人の歌声ような，自然かつ多様な歌声合成の実現にあたって，声区表現の問題は取り組むべき大きな課題であり，歌声合成分野の発展に不可欠であると言える．

1.2　本研究の背景

人の歌声のような歌声合成の実現を目指すにあたり，人の歌唱と歌声合成における声区表現の相違を明確にすることは重要である．人の歌唱における声区表現のメカニズムと，歌声合成における声区表現の問題点を，それぞれの先行研究を述べることで明らかにする．それらを踏まえた上で，問題解決のためのアプローチを提示する．

人の歌唱における声区表現のメカニズム

人の歌唱における声区表現に関して，音声生成機構・音響的特徴・知覚の様々な観点から，数多くの研究が行われている [1–8].

Fant の音源フィルタ理論によると，有声音の音声波は，声門での声帯振動により発生した声帯音源波が声道フィルタを通過し，開口部から放射される [9]．これらの音声生成機構のうち，声帯音源特性が声区に深く関連していることが知られており，声門開口率，

声帯の緊張，声門閉鎖時の乱流といった声帯振動様式が，声区ごとに大きく異なることが明らかとなっている．つまり人は，声区ごとの声帯音源特性の違いを歌唱訓練により習得して，使い分けていると言える．このような音声生成機構の使い分けにより，スペクトル傾斜をはじめとした音響的特徴が変化する．結果として，それぞれの声区特有の声質が得られ，高音域や低音域でも自然な歌声として知覚される [2, 4, 5, 10]．

歌声合成における声区表現の問題点

人の歌唱に対して，歌声合成の研究は声区表現にまだ完全には対応できていない．代表的な歌声合成システムである VOCALOID [11] は，素片接続型の合成方式であり，声質を大きく変化させることは難しい．声区に関連の深い声帯音源特性を制御することができないため，声区表現は集められた音素片データに依存してしまう．

これに対し，話声を歌声に自動変換する歌声合成システム，SingBySpeaking が提案されている [12, 13]．音声分析合成系・STRAIGHT [14, 15] を使用して，歌声らしさに関わる基本周波数とスペクトル包絡の制御規則を構築しているため，自然性の高い歌声合成を実現している．この手法に，声区表現の制御規則を設ければ声区表現が可能になることが考えられるが，STRAIGHT の枠組みでは声帯音源特性，特に声帯音源に関連する声質，を独立して制御することは困難であり，声区表現に十分に対応できない．声区表現に対応できていない歌声合成システムを用いて合成された歌声合成音では，高音域や低音域における歌声高音域や低音域のの自然性が損なわれてしまうことが問題点となる．

高音域や低音域において損なわれる歌声合成音の自然性を回復する方法として，声区ごとの声帯音源特性を付加することが考えられる．その方法を実現するためには，声帯音源特性を記述できるモデルについて考える必要がある．以下に，声帯音源特性を記述できるモデルの先行研究を示し，有効可能性が期待できるモデルを選定する．

声帯音源特性を記述可能なモデル

声帯音源特性を記述するためには，音源フィルタ理論に基づいて，有声音を声帯音源特性と声道フィルタに分離する必要がある．これまでに，Linear Predictive Coding (LPC) に基づいて，声帯音源信号を推定する逆フィルタリングの方法が提案されている [16, 17]．しかし，LPC では声帯音源信号をパルス列で表現しているため，声帯音源特性を十分に表現できていない．また，声帯音源特性に由来するスペクトルを，分離して独立制御することができない．STRAIGHT も同様の問題を抱えており，声区表現の問題を解決するためのアプローチには適していない．

これらに対し，分離が可能なモデルとして，Autoregressive with exogenous input (ARX) モデル [18,19] が提案されている．人の音声生成過程を模擬したモデルであり，有声音を声帯音源特性と声道フィルタに分離し，分析・変形・合成することが可能であるため，声帯音源特性の独立制御に適したモデルと考えられる．しかし，声帯音源特性の記述に用いている Rosenberg-Klatt モデル [20] はパラメータが少ないため，声区ごとの声帯音源特性を十分に表現できない．

一方で，声帯音源信号を近似するモデルとして，Lijencrants-Fant (LF) モデル [21,22] がある．RK モデルより多くのパラメータを持ち，声区ごとの声帯音源特性が各パラメータに対応しているため，声区表現に適していると考えられる．しかし，声帯音源特性だけでなく声道フィルタを推定するための別の方法が必要となる．

ARX モデルと LF モデルの問題点を解決するために，これらのモデルを組み合わせてそれぞれの短所を補った ARX-LF モデル [23–25] が提案されている．音声生成過程が模擬でき，かつ LF モデルによる声帯音源特性の詳細な記述が可能である．声区表現を可能とする歌声合成の実現に向けて，より適したモデルであると考えられる．

1.3 本研究の目的

本研究の目的は，声区表現を可能とする歌声合成に向けた，声帯音源特性の制御法の検討である．目的を遂行するため，上記の ARX-LF モデルを用いた，新たな枠組みの歌声合成システムを提案する．

本研究のアプローチの概念図を図 1.1 に示す．従来の規則ベースの歌声合成システム [12,26] では，歌声の知覚と音響的特徴の関連性を調査し，音響的特徴の制御規則を構築するといった知覚側からのアプローチがなされており，音声生成機構については考慮されていない．これに対し，本提案では，音声生成機構側からのアプローチを行うことで，より人に近い，直接的な歌声合成の枠組みを提供する．人の音声生成過程に基づいて，入力音声に対して分析・変形・合成を行うことにより，声帯音源特性の独立制御が可能となる．ARX-LF モデルは声帯音源特性に対応する複数のパラメータを持ち，これらのパラメータを音高の変化に伴い適切に変化させるように制御モデルを構築することで，声区ごとの声帯音源特性を付加できるようになる．本研究のアプローチによって，音声生成機構・音響的特徴・知覚における相互関係性について，より詳細な調査が可能となる．さらには，それらの過程で得られた調査結果を，様々な歌声合成システムに反映させ品質向上につなげる，といった応用可能性にも期待できる．

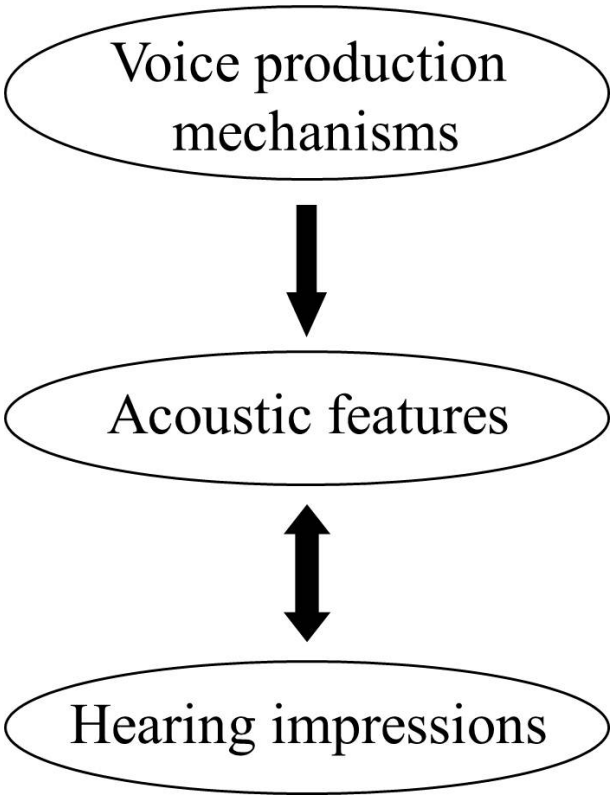

図 1.1: 提案アプローチの概念図

1.4 本論文の構成

本論文は 5 章で構成される．各章の概要を以下に示す.

第 2 章

本研究で提案する歌声合成システムの方略について述べる．構築する歌声合成システムの前提条件を提示し，システムに必要な ARX-LF モデルの概要について説明する．そして，歌声合成の手続きを示す.

第 3 章

第 3 章では，声区ごとの声帯音源特性を付加するための，ARX-LF パラメータ制御モデルについて説明する.，声区表現に向けた ARX-LF モデルの有効性を，声区ごとの ARX-LF パラメータの分析結果によって示す．分析結果に基づいて，各パラメータの制御モデルを構築する.

第 4 章

評価のために，提案システムによって作成された歌声合成音を用いた主観評価と客観評価を遂行する．まず，聴取実験による声質評価によって，提案システムによる声区の再現性を評価する．次に，音響的特徴の分析による客観評価を行う．最後に，音響的特徴の分析結果を従来法 [12] に反映させて主観評価を行うことで，本研究で得られた知見の有効性を示す.

第 5 章

本研究で得られた結果を要約し，今後の課題を述べる.

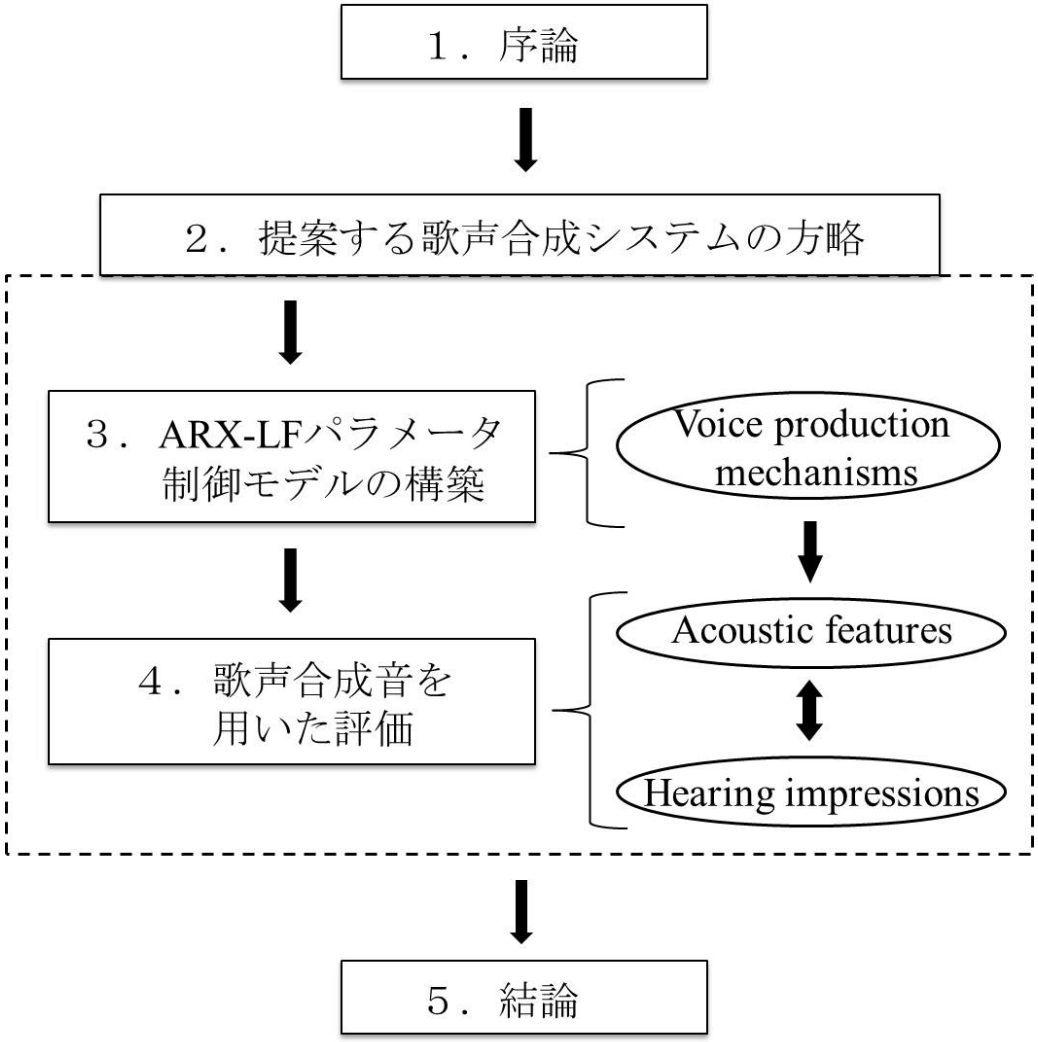
１．序論
２．提案する歌声合成システムの方略
３．ARX-LFパラメータ
制御モデルの構築
Voice production
mechanisms
４．歌声合成音を
用いた評価
Acoustic features
Hearing impressions
５．結論

図 1.2: 本論文の構成

第2章　提案する歌声合成システムの方略

2.1　はじめに

本章では，本研究で提案する歌声合成システムの方略について述べる．構築する歌声合成システムの前提条件を示し，前提条件を満たすために必要な ARX-LF モデルの概要について説明した上で，歌声合成の手続きを述べる．

2.2　提案する歌声合成システムの前提条件

実用的な歌声合成システムの構築に向けて，本研究で提案するシステムの前提条件を以下のように定義した．

前提条件 1: 典型的な 3 つの声区である vocal fry（低音域），modal（中音域），falsetto（高音域）を表現することで，幅広い音域を自然に歌唱できる．

前提条件 2: 入力データの個人性が保存され，出力である歌声に反映される．

前提条件 3: 人の歌声としての自然性が確保されている．

ARX-LF モデルを適用し，声区ごとの声帯音源特性を付加することで，前提条件 1 が満たされる．また，ARX-LF モデルによって入力データに分析・変形・合成を施し，出力することで前提条件 2 が満たされる．前提条件 3 については，斎藤らが提案した歌声らしさに関連する音響的特徴の制御モデル [12] を適用することによって対応する．

2.3 ARX-LF モデル

人の音声生成過程は，式 2.1 のように ARX-LF モデルによって模擬される．

$$s(n) + \sum_{i=1}^{p} a_i(n)s(n-i) = b_0(n)u(n) + \varepsilon(n) \tag{2.1}$$

ここで $s(n)$，$u(n)$ はそれぞれ音声信号，声帯音源信号である．ただし，$u(n)$ は LF モデルによって近似される．$a_i(n)$，$b_0(n)$ は声道フィルタに関する時変係数，$e(n)$ は残差である．式 2.1 を時不変と仮定して z 変換すると，式 2.2 となる．

$$S(z) = \frac{b_0}{A(z)} \cdot U(z) + \frac{1}{A(z)} \cdot E(z) \tag{2.2}$$

$U(z), E(z)$，$S(z)$ はそれぞれ声帯音源信号，残差，音声信号の z 変換である．$u(n)$ の形状は，図 2.1 に示すように，基本周期 T_0 と 4 つのパラメータ T_p, T_e, T_a，E_e によって表現される．E_e は，b_0 を用いて計算される．

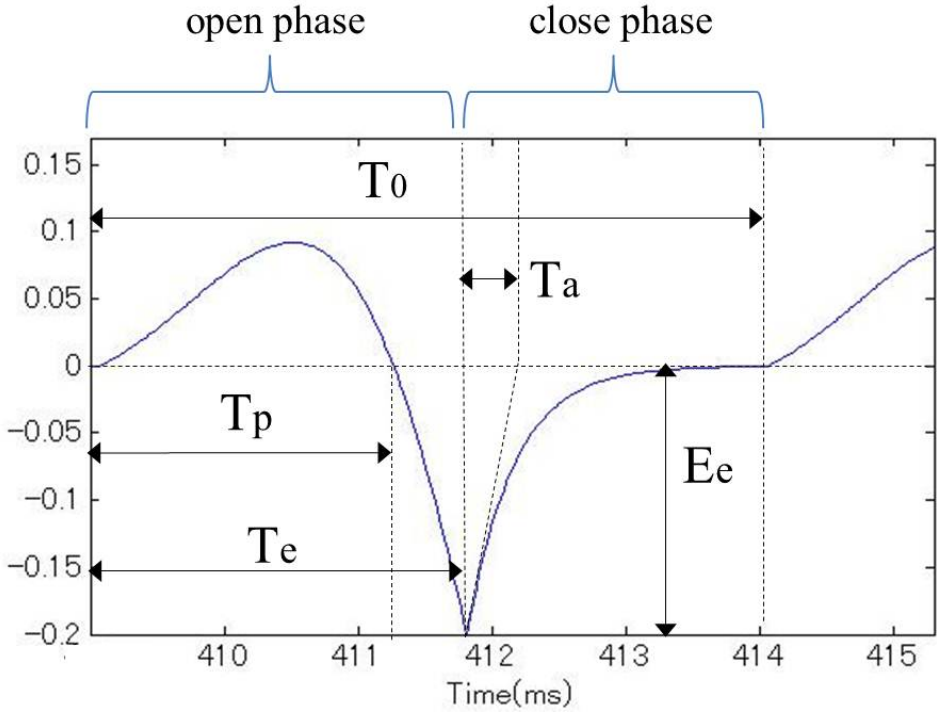

図 2.1: LF モデルによって得られる声帯音源信号

本研究では，制御を簡易化するため，3つのパラメータ O_q, α_m, Q_a を ARX-LF パラメータとして用いる．O_q は声門開口率，α_m は声帯音源信号の開口区間の左右対象性，Q_a は声門完全閉鎖までに要する戻り区間の時間率を表し，以下の式で算出される．

$$O_q = T_e/T_0 \tag{2.3}$$

$$\alpha_m = T_p/T_e \tag{2.4}$$

$$Q_a = T_a/(1 - O_q)T_0 \tag{2.5}$$

それぞれの ARX-LF パラメータと声帯音源特性の関連性を，以下に示す．

O_q と声帯音源特性の関連性

O_q は，声帯振動にとって主要な情報である声門開口率 (ピッチ周期に対する声門が開いている時間の割合) に，直接対応している．

α_m と声帯音源特性の関連性

α_m は，声門の開き・閉じの速さの比率を表し，声門抵抗や声帯の緊張の影響を受ける．

Q_a と声帯音源特性の関連性

Q_a は，不完全な声門閉鎖の際に発生する乱流に対応しており，声門閉鎖の強さの影響を受ける．

2.4　歌声合成の手続き

本研究で提案する，ARX-LF モデルに基づく歌声合成システムのブロック図を図 2.2 に示す．歌声合成の手順を以下で説明する．

1. 楽譜情報を用いて，基本周期を計算する．F0 制御モデル [12] を用いて作成された F0 の逆数を取ることで，1 周期ごとの基本周期が得られる．

2. 　システムの入力となる朗読音声を ARX-LF を用いて分析し，1 周期ごとの声帯音源特性，声道フィルタ，残差の情報を保存する．

3. 手順 2 で得られた ARX-LF パラメータを，制御規則に基づいて音高ごとに適切に制御する．これにより，声区ごとの声帯音源特性を付加する．

4. 手順 2 で得られた残差を，基本周期に合わせて伸縮する．

5. 手順 2 で得られた声道フィルタ，手順 3 で得られた声帯音源特性，手順 4 で得られた残差を用いて，式 2.2 に基づいて再合成を行う．この際，スペクトル包絡制御 [12] を適用することで，歌声の自然性を向上させる．

手順 3 で必要な，声区ごとの ARX-LF パラメータの制御法については，次章で詳細を述べる．声道フィルタに関しては，声区表現との関連性が先行研究によって示されている [7,33] が，本研究では声帯音源特性の制御に着目し，声道フィルタは制御せずにそのまま用いる．

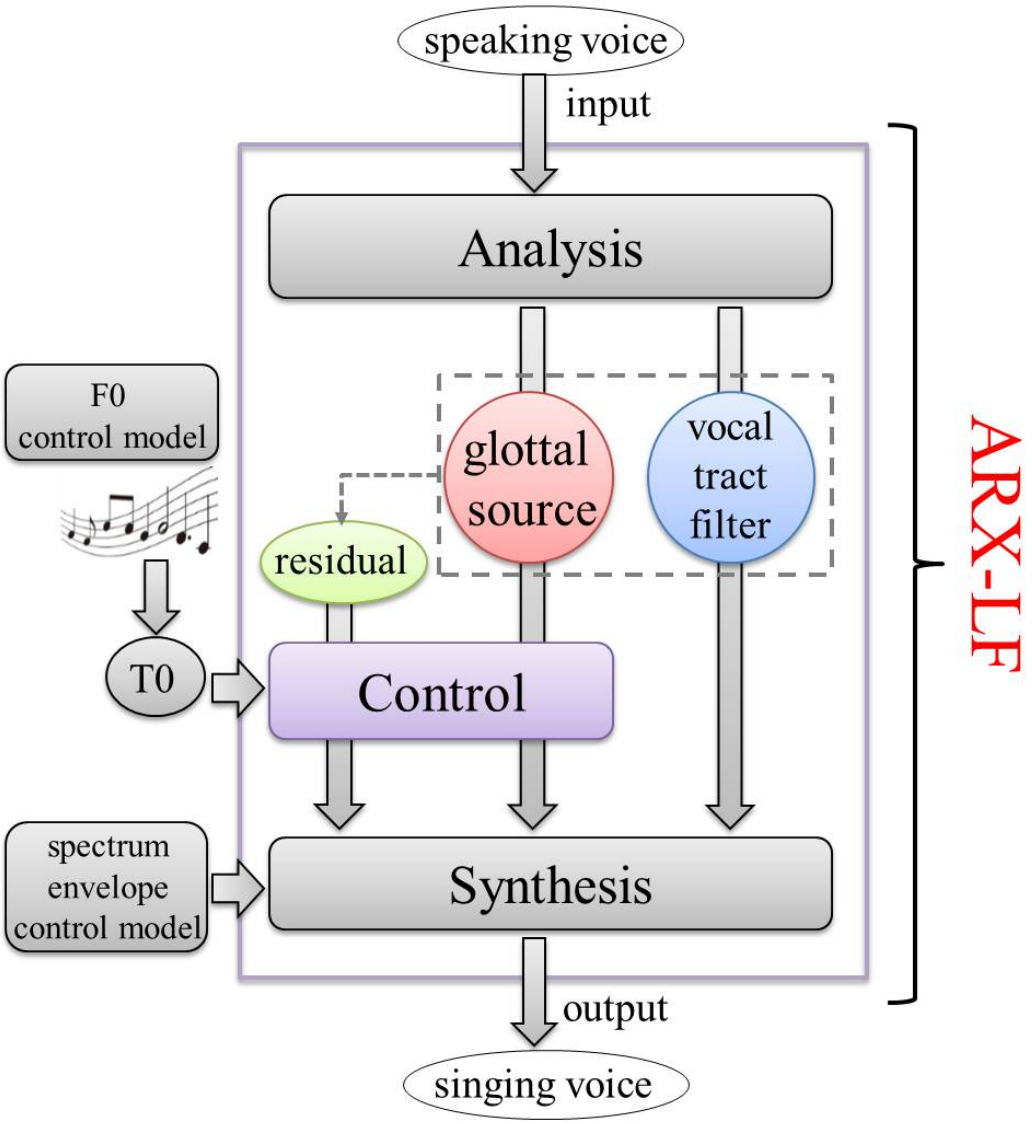

図 2.2: 提案システムのブロック図

2.5 まとめ

本章では，本研究で構築する歌声合成システムの方略について述べた．まず，歌声合成システムの前提条件を示し，前提条件を満たすために用いる ARX-LF モデルの概要について説明した上で，歌声合成の手続きを示した．これにより，ARX-LF パラメータの制御モデルを構築すれば，提案システムによる歌声合成音の作成が可能になることを明確にした．

第3章　ARX-LF パラメータの制御モデルの構築

3.1　はじめに

本章では，声区ごとの声帯音源特性を付加するための，ARX-LF パラメータ制御モデルを構築する．声区ごとの ARX-LF パラメータの分析結果を示し，ARX-LF パラメータで声区表現が可能であることを示した上で，各パラメータの制御モデルについて説明する．

3.2　ARX-LF パラメータの分析

ARX-LF パラメータの制御によって声区ごとの声帯音源特性を付加するという試みは，現在まで行われていない．まず，声区ごとの ARX-LF パラメータを分析し，分析結果の傾向を示すことにより，声区表現に向けた ARX-LF モデルの有効性を示す．

3.2.1　分析条件

歌声データベース「日本語を歌・唄・謡う」[28] を用いて分析を行った．有声音を取り扱うために，母音/a/を選出した．vocal fry と modal の境界，modal と falsetto の境界を，それぞれパラメータ V_b，F_b として任意に設定できるようにし，$V_b = 90$ Hz，$F_b = 400$ Hz とした．さらに，典型的な声区ごとの ARX-LF パラメータの傾向を分析するため，声区の重複部分を除いた三つの音域に分割し，分析対象とした．図 3.1 に示すように，F0 が 90 Hz 以下のデータを vocal fry データ，150 Hz 〜 300 Hz のデータを modal データ，400 Hz 以上のデータを falsetto データとして，分析を行なった．サンプリング周波数は 12 kHz とし，声道フィルタの次数は $p = 14$ とした．

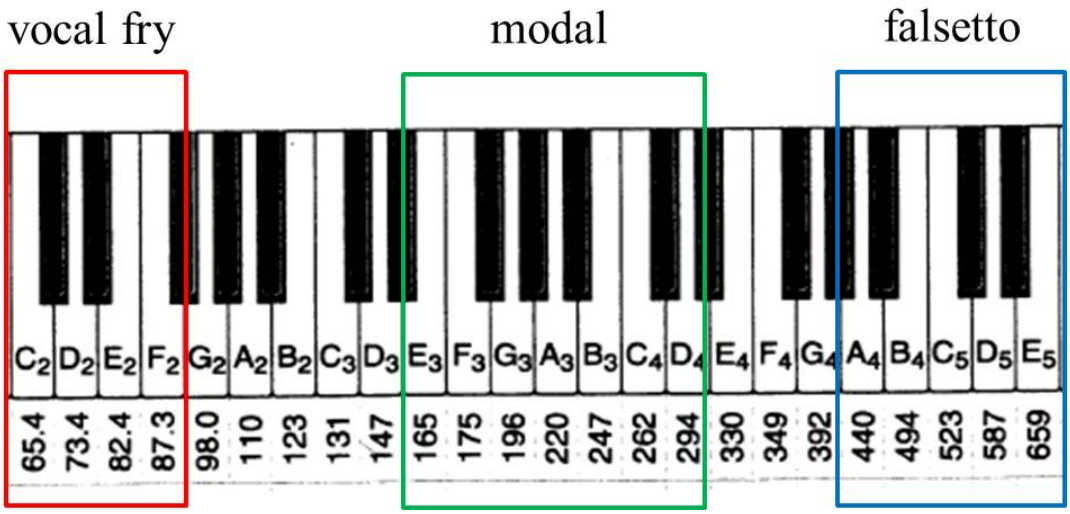

図 3.1: 分析対象とするデータの周波数範囲

表 3.1: ARX-LF パラメータを声区ごとに分析した平均値

声区	O_q	α_m	Q_a
vocal fry	0.226	0.826	0.015
modal	0.434	0.824	0.025
falsetto	0.824	0.773	0.116

3.2.2 分析結果

ARX-LF パラメータを声区ごとに分析した平均値を Table 1 に示す．ARX-LF パラメータと，声区ごとの声帯振動の知見に基づいて，抽出した ARX-LF パラメータ値を考察した．

声門開口率は，vocal fry では小さく，falsetto では大きいことが知られており，O_q の分析結果は，これらの知見に合致する結果となった．α_m は，falsetto で小さな値を取っている．falsetto では声帯が緊張し，部分振動となることが関連している．Q_a は，falssetto で非常に大きくなっている．falsetto は声門閉鎖が弱く，そのために発生する乱流が関連している．

各パラメータについて，先行研究 [4–6, 8] で報告されている声帯音源特性の知見に合致した結果が得られ，ARX-LF パラメータによって声区ごとの声帯音源特性が表現できることが示された．

3.3 ARX-LF パラメータ制御モデルの構築

得られた分析結果に基づいて，3 つの ARX-LF パラメータの制御モデルを構築する．以下のようなコンセプトに基づき，モデルの構築を行なった．

- 各パラメータは，作成する歌声合成音の F0 ($F0_{\_syn}$) に基づいて制御される．
- 話者の個人性を保つため，入力音声の F0 ($F0_{\_ori}$) と，分析によって得られた各 ARX-LF パラメータ (O_{q_ori}，α_{m_ori}，Q_{a_ori}) を用いる．
- それぞれの声区内で線形補間を行うことにより，各パラメータを制御する．

各パラメータのデータごとの分布を図 3.2 〜 3.4 に示し，構築したそれぞれの制御規則について説明する．

3.3.1 O_q 制御モデル

図 3.2 に示すように，O_q の値は声区ごとに大きく異なる．さらに，それぞれの声区において，小さな傾きが見られる．傾きを表現するため，最小二乗法を用いて声区ごとに回帰直線を求めた結果を表 3.2 に示す．傾き a_{oq} と切片 b_{oq} を用いて，O_q 制御モデルを構築した．式 3.1，3.2 によって，制御した O_q の値 O_{q_syn} が得られる．

$$O_{q\_syn} = O_{q\_ori} + y_{oq}(F0_{\_syn}) - y_{oq}(F0_{\_ori}) \tag{3.1}$$

$$y_{oq}(x) = a_{oq} \cdot log_2 x - b_{oq} \tag{3.2}$$

a_{oq}，b_{oq} は x，V_b，F_b の値によって決定される．$x < V_b$ ならば表 3.2 の vocal fry の値，$V_b \leq x < F_b$ ならば modal の値，$F_b \leq x$ ならば falsetto の値が得られる．

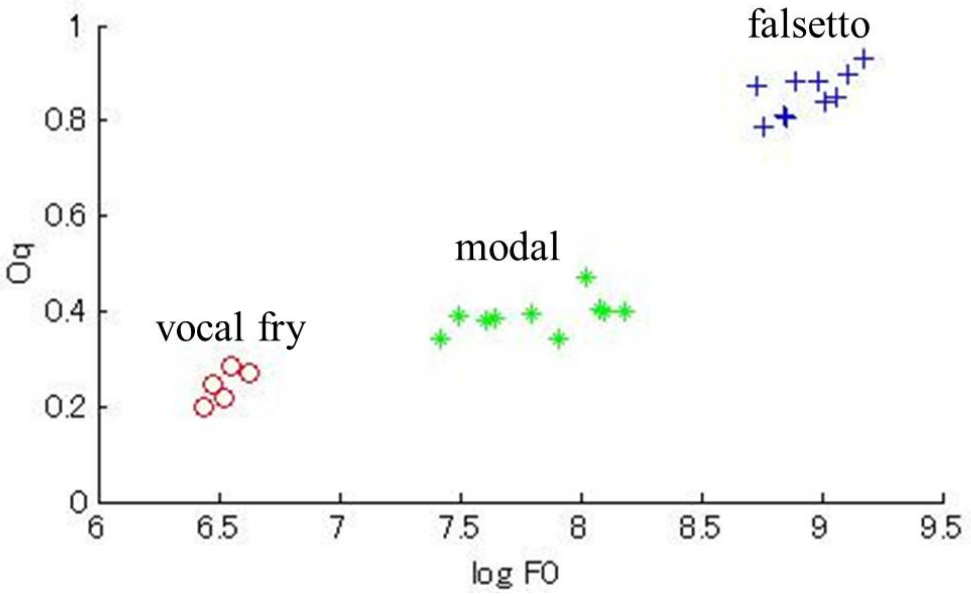

図 3.2: データごとの O_q の分布

表 3.2: 声区ごとの O_q の傾き a_{oq} と切片 b_{oq}

声区	傾き a_{oq}	切片 b_{oq}
vocal fry	0.119	-0.529
modal	0.050	0.047
falsetto	0.207	-1.001

3.3.2 α_m 制御モデル

falsetto における声帯の部分振動を表現するため，α_m 制御モデルを構築した．制御を簡易化するため，$F0_{\_syn}$ が高い場合のみ α_m の値を制御する．図 3.3 に示すように，α_m は falsetto 内で大きく異なる値をとっている．パラメータ α_r を設けることで，制御による値の変化率を任意に設定できるようにした．式 3.3 によって，制御した α_m の値 α_{m_ori} が得られる．

$$\alpha_{m\_syn} = \begin{cases} \alpha_{m\_ori} \cdot \alpha_r & (F_b \leq F0_{\_syn}) \\ \alpha_{m\_ori} & (F0_{\_syn} < F_b) \end{cases} \tag{3.3}$$

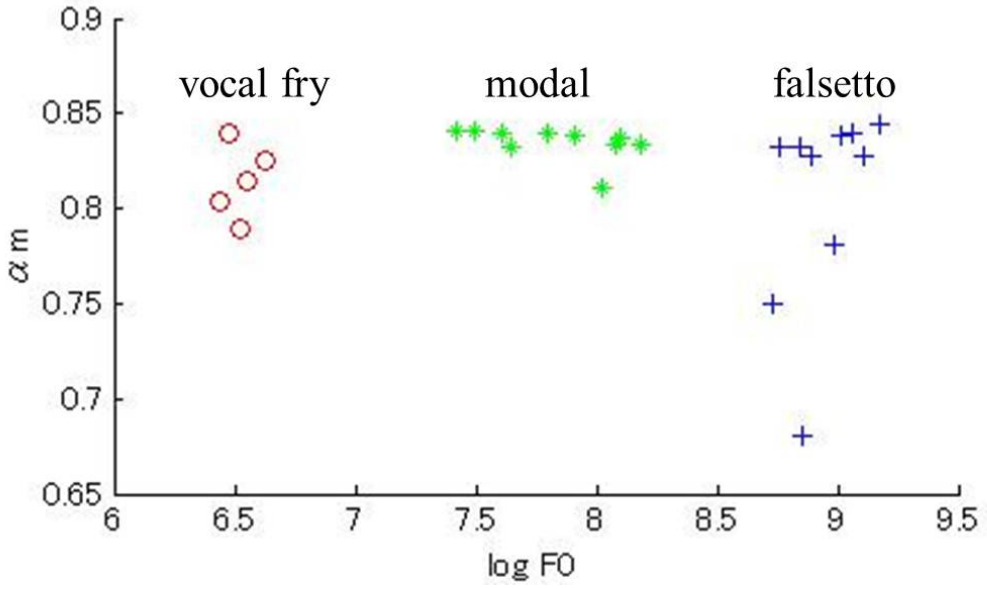

図 3.3: データごとの α_m の分布

3.3.3 Q_a 制御モデル

不完全な声門閉鎖によって生じる乱流の影響を表現するため，Q_a 制御モデルを構築した．図 3.4 に示すように，Q_a は falsetto で大きな値をとる．さらに，それぞれの声区において，小さな傾きが見られる．O_q 制御モデルと同様に，声区ごとに回帰直線を求めた結果を表 3.3 に示す．式 3.4，3.5 によって，制御した Q_a の値 Q_{a_syn} が得られる．

$$Q_{a\_syn} = Q_{a\_ori} + y_{qa}(F0_{\_syn}) - y_{qa}(F0_{\_ori}) \tag{3.4}$$

$$y_{qa}(x) = a_{qa} \cdot log_2 x - b_{qa} \tag{3.5}$$

a_{qa}，b_{qa} は x，V_b，F_b の値によって決定される．$x < V_b$ ならば表 3.3 の vocal fry の値，$V_b \leq x < F_b$ ならば modal の値，$F_b \leq x$ ならば falsetto の値が得られる．

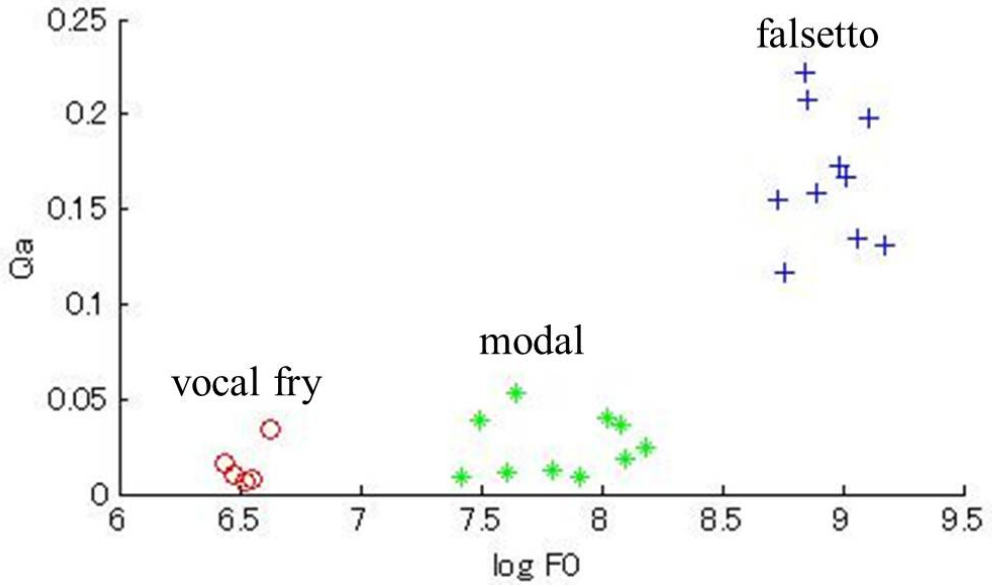

図 3.4: データごとの Q_a の分布

表 3.3: 声区ごとの Q_a の傾き a_{qa} と切片 b_{qa}

声区	傾き a_{qa}	切片 b_{qa}
vocal fry	0.038	-0.232
modal	0.009	-0.004
falsetto	0.015	0.035

3.4 まとめ

本章では，声区ごとの声帯音源特性を付加するための ARX-LF パラメータ制御モデルを構築した．声区ごとの ARX-LF パラメータの分析結果により，ARX-LF モデルが声区表現に有効であることを示した．そして，分析結果に基づいて ARX-LF パラメータ制御モデルを構築した．これにより，提案システムによる歌声合成音の作成が可能となった．

第4章　歌声合成音を用いた評価

4.1　はじめに

第 4 章では，ARX-LF パラメータ制御モデルの評価のために，音響的特徴の分析による客観評価と聴取実験による主観評価を行う．以下に，これら 2 つの評価項目を示す．

客観評価： 声区の再現性の客観評価を目的とする．歌声合成音の音響的特徴の分析結果を，声区ごとに比較する．

主観評価： 声区の再現性の主観評価を目的とする．聴取実験によって得られる歌声合成音の聴取印象を，声区ごとに比較する．

4.2　客観評価

声区の再現性の客観評価を行うため，提案システムによって作成された歌声合成音に対して，音響的特徴の分析を行う．falsetto と vocal fry の再現性を評価するために，以下に示す分析-F，分析-V を行う．

- 分析-F：提案システムにより作成した modal と falsetto の歌声合成音について，音響的特徴を分析し，歌声データの分析結果と比較
- 分析-V：提案システムにより作成した modal と vocal fry の歌声合成音について，音響的特徴を分析し，歌声データの分析結果と比較

表 4.1: 分析-F における歌声合成音の F0 と音名

歌声合成音	Mf1	Mf2	Mf3	F1	F2	F3
F0 (Hz)	262	294	311	349	392	465
音名	C4	D4	E4♭	F4	G4	B4♭

表 4.2: 分析-V における歌声合成音の F0 と音名

歌声合成音	Mv1	Mv2	Mv3	V1	V2	V3
F0 (Hz)	130	123	110	87	82	73
音名	C3	B2	A2	F2	E2	D2

4.2.1 歌声合成音の作成

分析-F，分析-V それぞれにおいて，音高の異なる歌声合成音を 6 つずつ作成した．分析-F では，modal の歌声合成音 Mf1，Mf2，Mf3 と falsetto の歌声合成音 F1，F2，F3 を作成した．それぞれの F0 と音名を表 4.1 に，F0 の時間変化を図 4.1 に示す．分析-V では，modal の歌声合成音 Mv1，Mv2，Mv3 と vocal fry の歌声合成音 V1，V2，V3 を作成した．それぞれの F0 と音名を表 4.2 に，F0 の時間変化を図 4.2 に示す．$\alpha_r = 0.9$，$V_b = 100$, $F_b = 310$ とした．

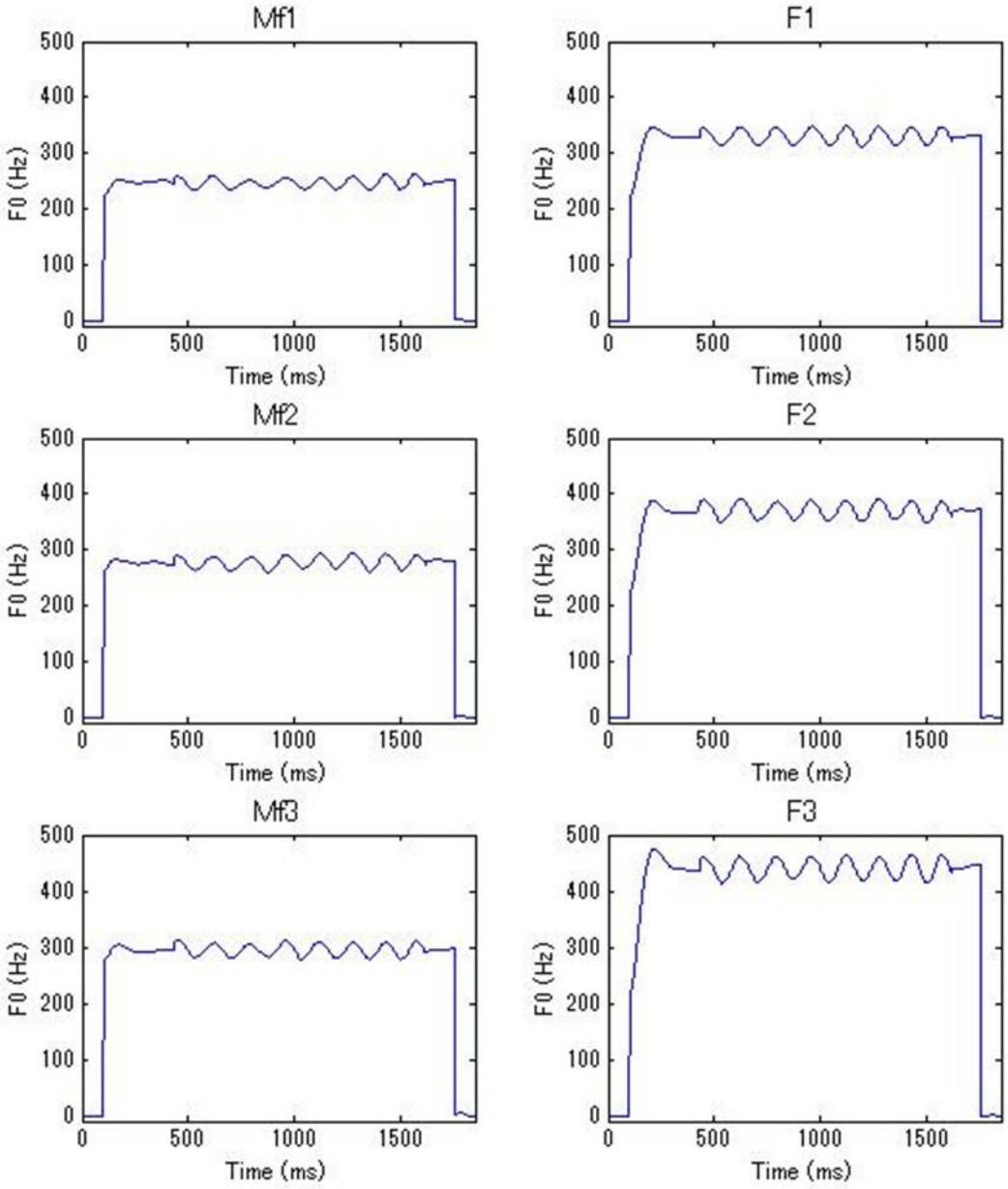

図 4.1: 分析-F における歌声合成音の F0：Mf1（左上），Mf2（左中），Mf3（左下），F1（右上），F2（右中），F3（右下）

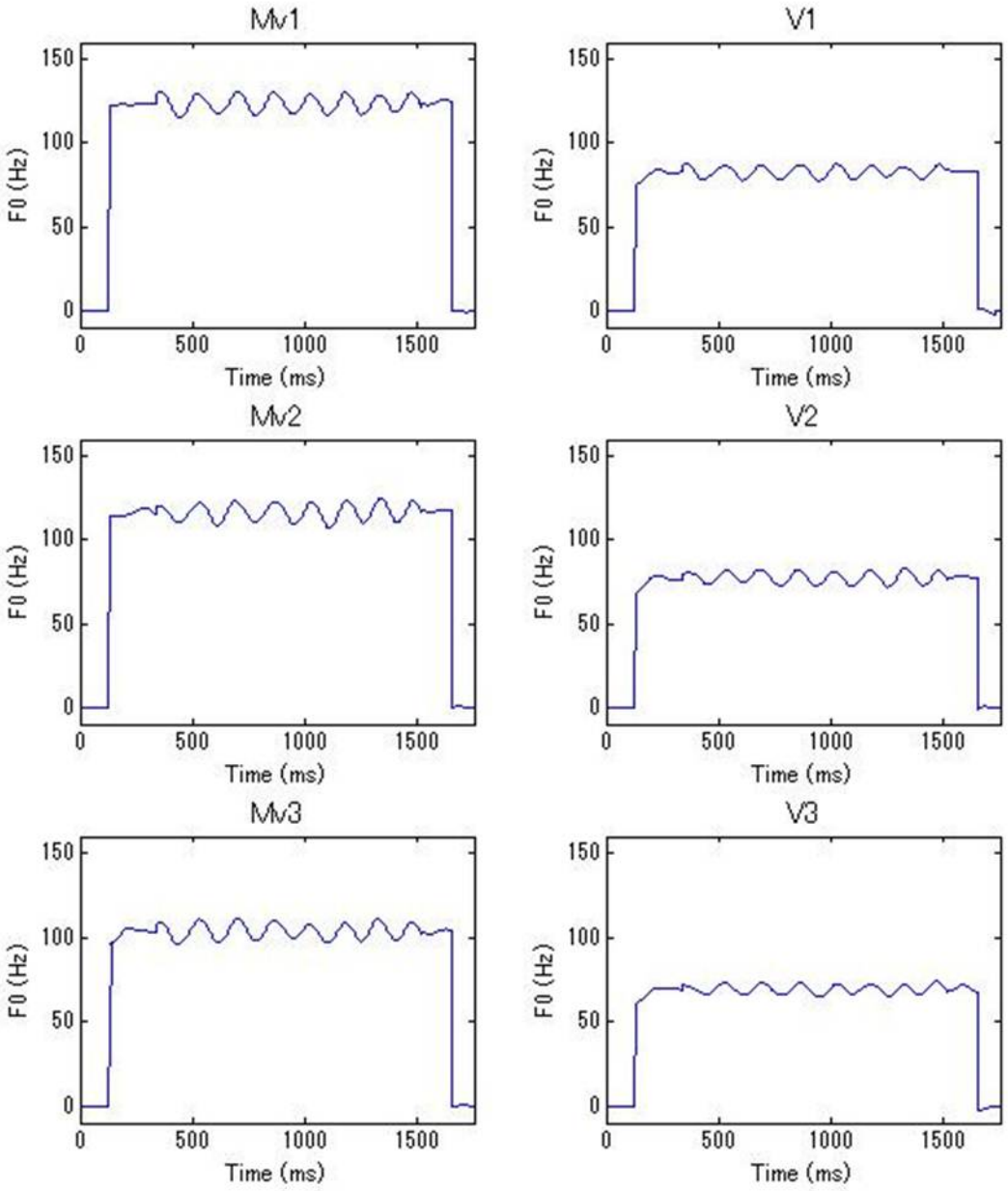

図 4.2: 分析-V における歌声合成音の F0：Mv1（左上），Mv2（左中），Mv3（左下），V1（右上），V2（右中），V3（右下）

4.2.2 分析条件

分析対象として，声質に関連の深い典型的な音響的特徴である，スペクトル傾斜を用いた．スペクトル傾斜は声帯音源特性によって異なり，falsettoでは急峻な傾斜，vocal fryでは緩やかな傾斜が得られることが知られている [5,29]．例として，3.2節におけるARX-LF分析で得られた，歌声データの声帯音源波のスペクトル包絡を図4.3に示す．先行研究で述べられているような，声区ごとの傾斜の違いが読み取れる．提案システムによって，これらの典型的なスペクトル傾斜の違いが得られるかを評価するため，歌声合成音の声帯音源波のスペクトル包絡の傾斜を求め，声区間で比較した．得られた結果の妥当性を評価するため，歌声データに対しても同様にスペクトル傾斜を求め，歌声合成音の結果と比較した．スペクトル傾斜は，回帰直線の傾きによって求めた．

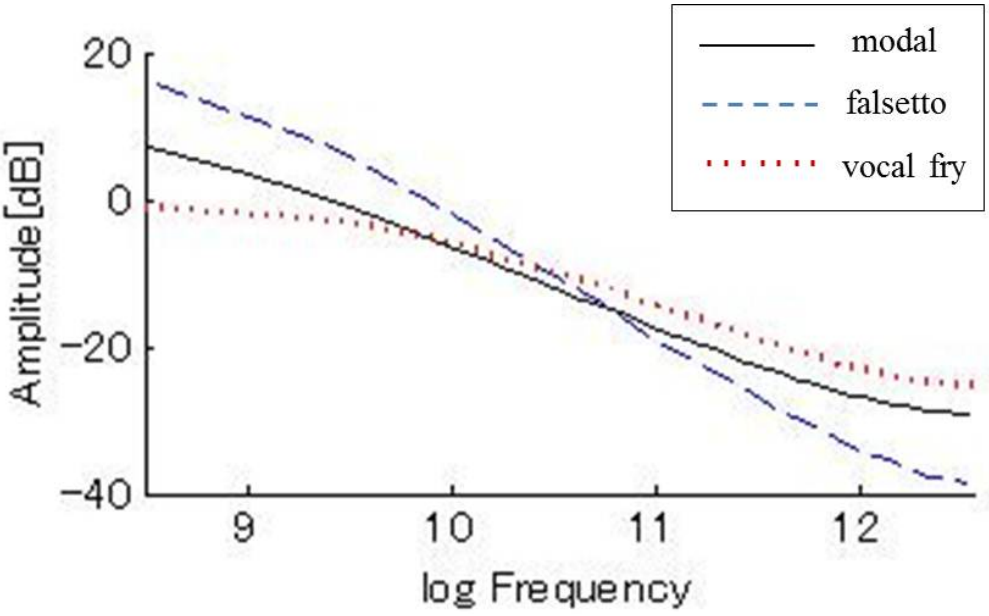

図 4.3: ARX-LF 分析によって得られた各声区の声帯音源波のスペクトル包絡

表 4.3: 分析-F における歌声合成音の声帯音源スペクトルの傾斜

歌声合成音	Mf1	Mf2	Mf3	F1	F2	F3
傾斜 (dB/oct.)	-10.74	-10.40	-10.12	-14.21	-14.92	-14.63

表 4.4: 分析-F における歌声データの声帯音源スペクトルの傾斜

	modal の歌声データ			falsetto の歌声データ		
F0 (Hz)	277	329	349	370	415	465
傾斜 (dB/oct.)	-10.08	-10.96	-10.54	-14.88	-13.99	-14.59

表 4.5: 分析-V における歌声合成音の声帯音源スペクトルの傾斜

歌声合成音	Mv1	Mv2	Mv3	V1	V2	V3
傾斜 (dB/oct.)	-10.57	-10.81	-11.06	-8.14	-7.52	-6.97

表 4.6: 分析-V における歌声データの声帯音源スペクトルの傾斜

	modal の歌声データ			vocal fry の歌声データ		
F0 (Hz)	146	138	130	98	92	82
傾斜 (dB/oct.)	-10.47	-10.21	-10.48	-7.55	-8.09	-7.39

4.2.3 分析結果

分析-F で得られた歌声合成音と歌声データのスペクトル傾斜を，それぞれ表 4.3，表 4.4 に示す．歌声合成音において，modal と falsetto の間で明確な違いが見られ，平均 4.17 (dB/oct.) の傾斜の差が得られた．歌声データについても，平均 3.96 (dB/oct.) の差が得られ，歌声合成音の妥当性を示す結果となった．これより，falsetto 特有の急峻な声帯音源スペクトルの傾斜が，歌声合成音によって表現できていると言える．

分析-V についても，分析-F と同様の傾向が得られた．分析-V で得られた歌声合成音と歌声データのスペクトル傾斜を，それぞれ表 4.5，表 4.6 に示す．歌声合成音では，modal と vocal fry の間で 3.27 (dB/oct.) の傾斜の差が得られ，歌声データでは平均 2.71 (dB/oct.) の差が得られた．vocal fry 特有の緩やかな声帯音源スペクトルの傾斜が表現できていると言える．これらの結果より，人の歌声における声区ごとの声帯音源スペクトルの傾斜を，歌声合成音で表現できていることが示された．

4.3 主観評価

提案システムによって作成された歌声合成音を用いて，声区の再現性の主観評価を聴取実験にて実施する．falsetto と vocal fry の再現性を評価するために，以下に示す実験-F，実験-V を行う．

- 実験-F : 提案システムにより作成した modal と falsetto の歌声合成音の聴取印象を，falsetto の典型的な声質 ‘気息性’ に基づいて比較評価
- 実験-V : 提案システムにより作成した modal と vocal fry の歌声合成音の聴取印象を，vocal fry の典型的な声質 ‘粗慥性’ に基づいて比較評価

4.3.1 歌声合成音の作成

実験-F で作成する歌声合成音の音高は，前節の分析-F と同様である．ただし，聴取印象の判断を簡易化するため，1 音目に F0=233 (Hz) の歌声を提示した上で，2 音目に目的の音高に移行するようにした．それぞれの F0 の時間変化を図 4.4 に示す．実験-V の音高については，分析-V と同様である．1 音目に F0 = 146(Hz) の歌声を提示し，2 音目に目的の音高に移行するようにした．それぞれの F0 の時間変化を図 4.5 に示す．$\alpha_r = 0.9$，$V_b = 100$, $F_b = 310$ とした．

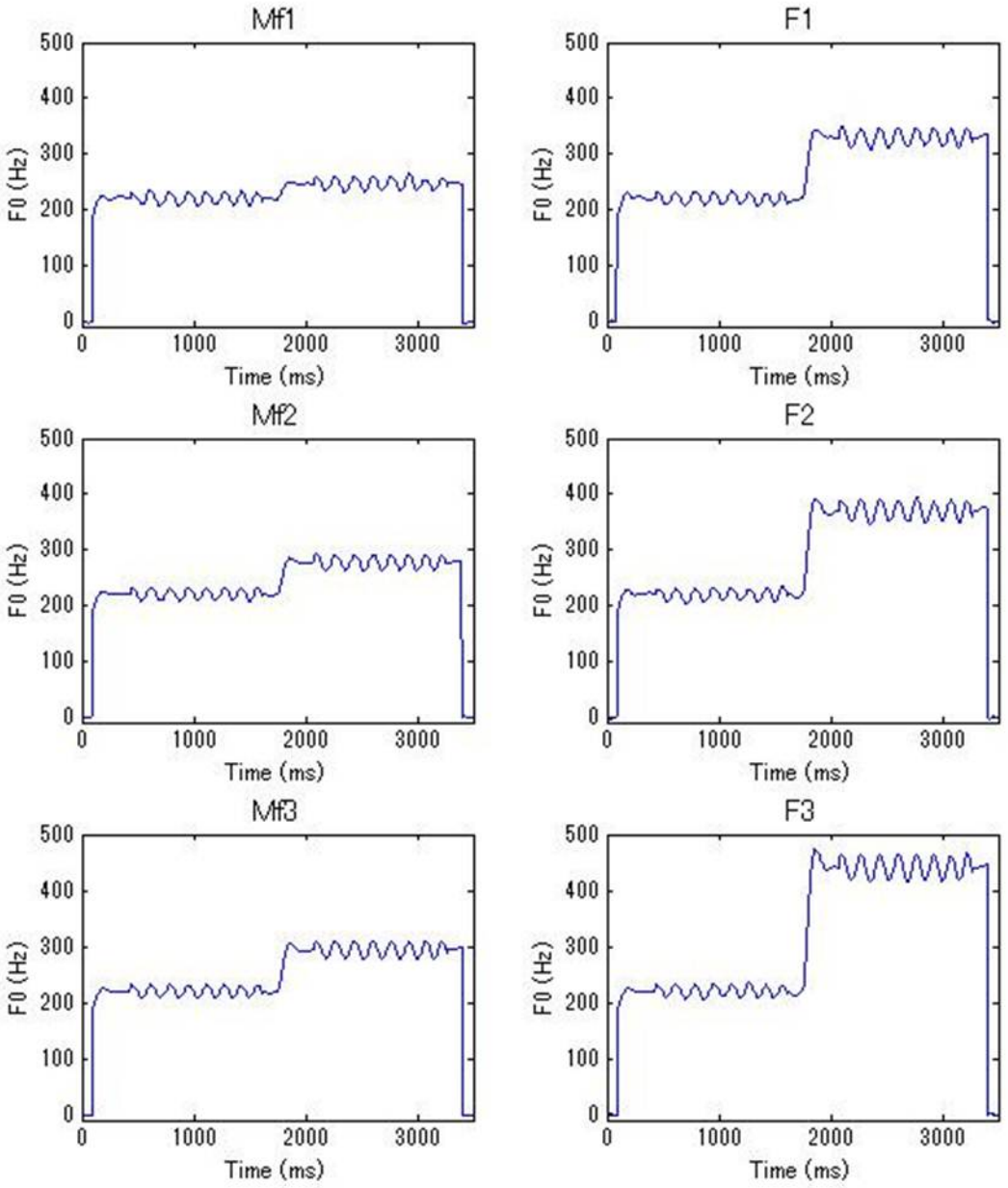

図 4.4: 実験-F における歌声合成音の F0：Mf1（左上），Mf2（左中），Mf3（左下），F1（右上），F2（右中），F3（右下）

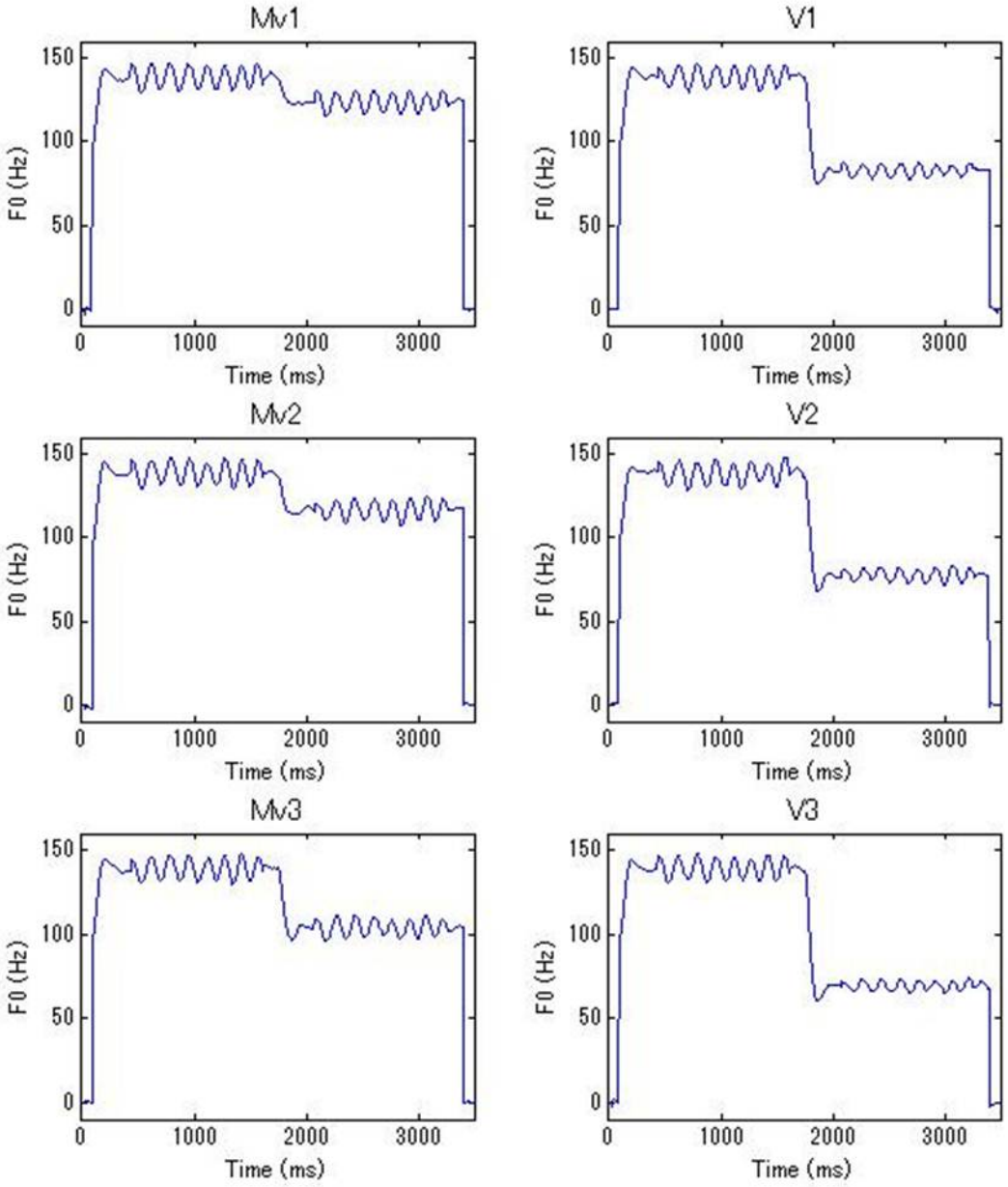

図 4.5: 実験-V における歌声合成音の F0：Mv1（左上），Mv2（左中），Mv3（左下），V1（右上），V2（右中），V3（右下）

4.3.2 実験条件

シェッフェの一対比較法 (浦の変法) [30] によって聴取実験を行った．被験者は，大学院生 8 名である．刺激順序の違いも考慮した 6 × 5=30 対の歌声合成音を，それぞれの実験において被験者に提示した．図 4.6 に，歌声合成音の提示順序を示す．

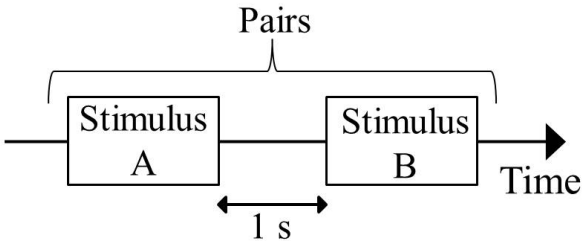

図 4.6: 歌声合成音の提示順序

4.3.3 実験手続き

被験者には，実験-F では ‘気息性 ’，実験-V では ‘粗慥性 ’といった，それぞれの声区の典型的な聴取印象を評価させた．実験-F の際に，聴取者には以下のような教示を与えた．

ヘッドホンから 2 つの歌を対にして聴いてもらいます．前の歌と後の歌の，それぞれ 2 音目同士を聴き比べて，どちらがより ‘気息性’ のある歌声かを，7 段階の評価尺度（図 4.7）に従って判断してください．前の歌声がより ‘気息性’ があると判断したら正の値（3 〜 1）に，後の歌声がより ‘気息性’ があると判断したら負の値（-3 〜 -1）を選択してください．どちらも同程度だと判断した場合は 0 を選択してください．

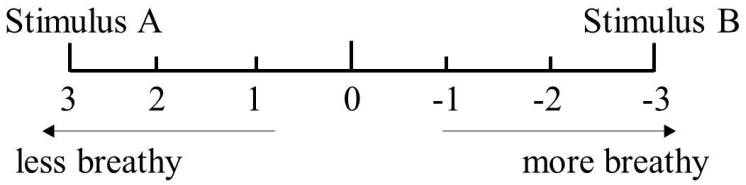

図 4.7: 実験-F で用いた聴取印象 ‘気息性’ に関する七段階評価尺度

実験-V に関しても，‘粗慥性’ を評価すること以外は，同様の教示を与えた．聴取印象の判断を容易にするため，被験者には予め modal，falsetto，vocal fry の歌声データを複数提示し，気息性のある歌声と粗慥性のある歌声について学習させた．

表 4.7: 実験-F における母数 σ の推定値

歌声合成音	Mf1	Mf2	Mf3	F1	F2	F3
母数 σ	-1.46	-1.43	-1.14	1.04	1.23	1.77

表 4.8: 実験-V における母数 σ の推定値

歌声合成音	Mv1	Mv2	Mv3	V1	V2	V3
母数 σ	-1.75	-1.64	-1.14	0.95	1.54	2.05

4.3.4 実験結果と考察

実験-F，実験-V について推定した母数 σ を，それぞれ表 4.7，4.8 に示す．また，母数の値に従って，歌声合成音の距離関係を直線で示したものを，それぞれ図 4.8，4.9 に示す．

実験-F について，母数が正の大きな値であるほど，‘気息性’ のある歌声だと判断されたことを表す．modal と falsetto で，明確な差が得られており，falsetto 特有の ‘気息性’ を表現できていると言える．音高が高くなるに従って，より ‘気息性’ のある歌声だと判断されており，広い音域であるほど声区ごとの声帯音源特性の付加が重要であることが示唆された．実験-V についても，実験-F と同様の傾向が得られており，vocal fry 特有の ‘粗慥性’ が表現されていると言える．これらの結果により，提案システムによって作成された歌声合成音において，声区特有の声質が得られることが示された．

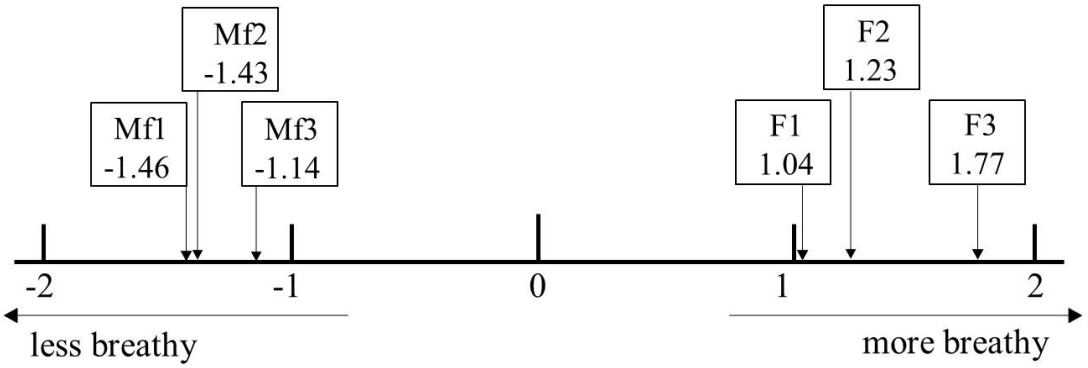

図 4.8: 実験-F における歌声の気息性の関係

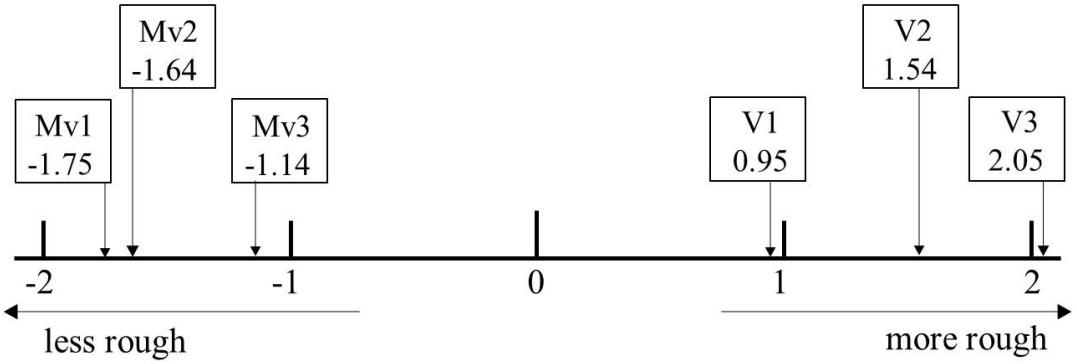

図 4.9: 実験-V における歌声の粗慥性の関係

4.3.5 まとめ

本章では，ARX-LF パラメータ制御モデルの評価のために，音響的特徴の分析による客観評価と聴取実験による主観評価を行なった．客観評価のために，歌声合成音の声帯音源スペクトルの傾斜を声区ごとに比較したところ，人の歌声における声区特有の傾斜を表現できていることが示された．主観評価のために，声区ごとの聴取印象を比較したところ，声区ごとの典型的な声質が得られた．さらに，広い音域であるほど，声区ごとの声帯音源特性の付加が重要である可能性が示唆された．

第5章　結論

5.1　本研究のまとめ

本研究では，声区表現が可能な歌声合成の実現に向けて，ARX-LF モデルの制御法を提案した．声区ごとの声帯音源特性を付加するため，ARX-LF パラメータ制御モデルを構築した．提案システムによって歌声合成音を作成し，客観評価と主観評価を行なった．得られた結果を，以下に要約する．

- 声区ごとの ARX-LF パラメータを分析した結果，先行研究の声帯音源特性の知見に合致した結果が得られ，ARX-LF モデルによって声区ごとの声帯音源特性を表現可能であることが示された．
- 声区ごとの ARX-LF パラメータの分析結果によって，Oq と Qa は声区ごとに異なるだけでなく，同一声区内でも音高変化に伴った傾きを持つことが分かった．
- 声区の再現性の客観評価のために，歌声合成音のスペクトル傾斜の分析結果を声区ごとに比較したところ，falsetto では急峻な傾斜，vocal fry で緩やかな傾斜が得られた．歌声データの分析結果においても同様の傾向が得られ，歌声合成音のスペクトル傾斜の妥当性が示された．
- 声区の再現性の主観評価のために，歌声合成音の声区ごとの聴取印象を比較したところ，falsetto では ‘気息性’ ，vocal fry では ‘粗慥性’ といった，声区ごとの典型的な声質が得られた．
- 聴取実験において，歌声合成音の音高が高いほど ’気息性’ がある歌声，低いほど ’粗慥性’ がある歌声であると判断された．これにより，広い音域であるほど，声区ごとの声帯音源特性の付加が重要である可能性が示唆された．

5.2 今後の課題

■ ARX-LF モデルに関する課題

より高品質で多様な歌声合成を実現するための ARX-LF モデルに関する課題を，以下に列挙する．

声帯音源モデルの改良

今回用いた LF モデルでは，実際の声帯音源信号に含まれる雑音成分 [31] を表現できていない．人の音声生成機構をより適切に表現するため，声帯音源モデルの改良が必要である．

声道フィルタの制御モデルの構築

本研究では，声道フィルタの制御は行なっていないため，falsetto の歌声合成音において，声帯音源特性と声道フィルタのミスマッチが原因と考えられる音韻性の欠如が目立った．声区ごとの声道フィルタの性質について調査を行い，声道フィルタ制御モデルの構築が必要である．Nguyen ら [32] が提案しているスペクトル変形法を適用すれば，声道フィルタの適切な制御が期待できる．

残差の制御法の改良

今回，残差の性質については時間方向への伸縮のみを行っており，振幅の制御は行なっていない．声区ごとの残差の性質をより詳細に調査し，制御法を検討する必要がある．

声区の境界部分における ARX-LF パラメータの調査

本研究では，声区の境界部分は分析対象から除外している．声区の境界部分における ARX-LF パラメータの遷移について，先行研究の知見 [6,33] を参考にしつつ調査を行い，制御モデルを改良すれば，声区の境界部分において滑らかに声区変換が可能な，高品質な歌声合成が期待できる．

ARX-LF モデルの分析精度の向上

上記で述べた課題において，正確な調査結果を得るために，ARX − LF モデルの分析精度の向上は重要である．周波数ドメインに着目した手法 [34] といった，分析精度向上のための手法を検討する必要がある．

無声音の表現

ARX-LF モデルでは，無声音の音源を適切に表現することができない．無声音の音源モデルを確立することができれば，無声音を含んだ様々な歌詞を歌唱できる歌声合成が実現可能となる．

データベース

今回，声区ごとの典型的な歌声を選定して分析対象としているが，複数の歌唱者データを用いているため，個人性の影響が含まれていると考えられる．音高変化に伴うARX-LF パラメータの変化をより正確に調査するには，同一歌唱者が幅広い音域を歌った歌声データを使用するべきである．声区表現に関するデータベースの構築が必要となる．

■ 客観評価，主観評価に関する課題

より詳細な評価を行うための客観評価，主観評価に関する課題を，以下に列挙する．

客観評価で分析する音響的特徴

今回，客観評価の分析対象としてスペクトル傾斜のみを扱っている．falsetto における雑音成分や，vocal fry 特有のサブハーモニック [5,35] といった，声区特有の音響的特徴を調査できていない．上記の ARX-LF モデルの改良を施した上で，声区に関連する音響的特徴について，詳細に調査する必要がある．

主観評価で用いる聴取印象

今回，主観評価で用いる聴取印象として，典型的なもののみを選定しているが，声区に関連する様々な聴取印象が先行研究によって挙げられている．複数の聴取印象を選定し，調査する必要がある．

一連の課題を遂行し，体系化することで，より高品質で多様な歌声合成システムの実現だけでなく，音声生成機構・音響的特徴・知覚の相互関係性の解明にも繋がるものである．本研究で用いた手法や，本研究で得られた知見が，今後の歌声合成分野の発展，ひいては音声科学の発展のために活かされれば，幸いである．

謝辞

本研究を進めるにあたり，多大なる御指導ならびに御鞭撻を賜りました赤木 正人 教授に深く感謝致します.
本研究を進めるにあたり，日頃から熱心な御指導ならびに御鞭撻を賜りました鵜木 祐史 准教授に心より感謝致します.
本研究を進めるにあたり，日頃から熱心に御討論頂き，また御助言を賜りました宮内 良太 助教に心より感謝致します.
本研究を進めるにあたり，熱心に御討論頂き，また御助言を賜りました党 建武 教授，末光 厚夫 助教，川本 真一 助教に心より感謝致します.
本研究を進めるにあたり，数々の御指導と御助言を賜りました金沢大学 自然科学研究科 齋藤 毅 助教に深く感謝致します.
また，本研究を進めるにあたり，日頃から熱心な議論と激励をいただきました，音情報処理分野の諸先輩方，及び諸氏に熱く御礼申し上げます.
本研究における聴取実験のために，貴重な時間を割いて頂きました実験協力者の方々に感謝の意を表します.
最後に，本学での研究生活を支え，温かく見守ってくれた両親に心から感謝致します.

参考文献

[1] Garcia, M., “Observations on the human voice,” Proc. Royal Soc., 3, 399-408, 1855.

[2] Childers, DF., Lee, CK., “Vocal quality factors: analysis, synthesis, and perception.,” J. Acoust. Soc Am. 90, 2394-2410, 1991.

[3] 今泉 敏，斉田 晴仁，H.Abdoerrachman，廣瀬 肇，新美 成二，志村 洋子，“音響分析による声の可制御性の評価：声区とヴィブラートについて,” 電子情報通信学会技術研究報告, 93(266), 25-29, 1993.

[4] Titze, I.R., “Principles of Voice Production,” Allyn & Bacon, 1994. References.

[5] Sakakibara, K., “Production Mechanism of Voice Quality in Singing,” J. Phonetic Society of Japan, 7(3), 27-39，2003.

[6] Roubeau, B., Henrich, N., Castellengo, M,. “Laryngeal vibratory mechanisms: The notion of vocal register revisited,” Journal of Voice, 23(4), 425-438, 2009.

[7] Tokuda, I., Zemke, M. kob, M., Herzel, H., “Biomechanical Modeling of Register Transitions and the Role of Vocal Tract Resonators,” Journal of Acoustic Society of America 127(3), 1528-1536, 2010.

[8] 今川 博，榊原 健一, 徳田 功，大塚 満美子，田山 二郎，“立体内視鏡とハイスピードカメラによる声門面積関数の計測,” 音声研究 14(2), 37-44, 2010.

[9] Fant,G., “Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations,” Mouton, 1970.

[10] 粕谷 英樹, 楊 長盛, “音源から見た声質,” 日本音響学会誌, 51(11), 869-875, 1995.

[11] Kenmochi, H., Ohshita, H., “VOCALOID — Commercial singing synthesizer based on sampleconcatenation,” INTERSPEECH, 4011-4010, 2007.

[12] 齋藤 毅，“歌声知覚・生成機構の解明に向けた歌声合成システム構築に関する研究,” JAIST 情報科学研究科博士論文，2006.

[13] Saitou, T., Goto, M., Unoki, M., Akagi, M., “Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices,” WASPAA, 215-218, 2007.

[14] 河原 英紀, “聴覚の情景分析が生み出した高品質 VOCODER: STRAIGHT,” 日本音響学会誌, 54(7), 521-526, 1998.

[15] Kawahara, H., “STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds,” Acoustic Science and Technology, 27(6), 349-353, 2006.

[16] Alku, P., “Glottal wave analysis with Pitch Synchronous iterative Adaptive inverse Filtering,” Speech Communication, 11, 109-118, 1992.

[17] Akande, O., Murphy J., “Estimation of the vocal tract transfer function with application to glottal wave analysis,” Speech Communication, 46, 15-36, 2005.

[18] Ding, W., Kasuya, H., Adachi, S., “Simultaneous Estimation of Vocal Tract and Voice Source Parameters Based on an ARX Model,” IEICE TRANSACTIONS, E78-D, 6, 738-743, 1995.

[19] 大塚 貴弘, 粕谷 英樹, “音源パルス列を考慮した頑健な ARX 音声分析法,” 日本音響学会誌 58(7), 386-397, 2002.

[20] Klatt, D., Klatt, L., “Analysis synthesis, and perception of voice quality variations among female and male talkers,” J. Acoust. Soc. Am., 87, 820—857, 1990.

[21] Fant, G., Liljencrants, J., Lin, Q., “A four-parameter model of glottal flow,” STL-QPSR, 85(2), 1-13, 1985.

[22] Fant, G., “The LF-model revisited.Transformations and frequency domain analysis,” STL-QPSR, 36(2-3), 119-156, 1995.

[23] Vincent, D., Rosec, O., Chonavel, T., “Estimation of LF glottal source parameters based on arx model,” INTERSPEECH, 333-336, 2005.

[24] Vincent, D., Rosec, O., “A new method for speech synthesis and transformation based on a ARX-LF source-filter decomposition and HNM modeling,” ICASSP, 525-528, 2007.

[25] Agiomyrgiannakis, Y., Rosec, O., “ARX-LF-based source-filter methods for voice modification and transformation,” ICASSP, 3589-3592, 2009.

[26] Minematsu, N., Matsuoka, B., Hirose, K., “Prosodic Modeling of Nagauta Singing and Its Evaluation,” ISCA, 487-490, 2004.

[27] Garnier, M., Hhnrich, H., Wolfe, J., Smith, J., “Vocal tract adjustments in the high soprano range,” Journal of the Acoustical Society of America, 127(6), 3771-3780, 2010.

[28] 中山 一郎, “日本語を歌・唄・謡う,” 日本音響学会誌, 59, 688-693, 2003.

[29] Gordon, M., Ladefoged, P., “Phonation types a cross-linguistic overview,” J. of Phonetics, 29, 383-406, 2001.

[30] 天坂 格郎, 長沢 伸也, “官能評価の基礎と応用,” 日本規格協会, 2003.

[31] Iijima, H., Miki, N., Nagai, N., “Glottal impedance based on a finite element analysis of two-dimensional unsteady viscous flow in a static glottis,” IEEE trans, sp, 40(9), 2125-2135, 1992.

[32] Nguyen, B., Akagi, M., “A flexible spectral modification metod based on temporal decomposition and Gaussian mixture model,” Acoustical Science and Technology, 30(3), 170-179, 2009.

[33] Garnier, M., Henrich, N., Smith, J., Wolfe, J., “Vocal tract adjustments in the high soprano range,” Acoust Soc Am., 127(6), 3771-3780, 2010.

[34] O Cinneide, A., Dorran, D., Gainza, M., Coyle, E., “A Frequency Domain Approach to ARX-LF Voiced Speech Parameterization and Synthesis,” INTERSPEECH, 57-60, 2011.

[35] Gerratt, B. R., Kreiman, J., “Toward a taxonomy of nonmodal phonation,” J. of Phonetics, 29(5), 365-381, 2001.

本研究に関する研究業績

国際会議

- Motoda, H., Akagi, M., "A singing voices synthesis system to characterize vocal registers using ARX-LF model," *Proc. 2013 RISP International Workshop on Nonliner Circuits, Communications and Signal Processing*, (to appear).

研究会

- 元田 紘樹, 赤木 正人, "声区の違いによる声質の変化と声帯音源特性の関連性," 日本音響学会聴覚研究会資料, 42(7), 585-590, 2012.
- 元田 紘樹, 赤木 正人, "声区表現可能な歌声合成を目的とした ARX-LF パラメータの制御法の検討," 日本音響学会聴覚研究会資料, 43(1), 37-42, 2013.
- 元田 紘樹, 赤木 正人, "ARX-LF に基づく声区表現を組み込んだ歌声合成システムの構築," 日本音響学会 2013 年春季研究発表会, (to appear).