JAIST Repository
https://dspace.jaist.ac.jp/

Title	声道の共振特性を考慮した歌声合成システムの構築に関する研究
Author(s)	長田，和也
Citation	
Issue Date	2015-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/12634
Rights	
Description	Supervisor:赤木正人，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

修 士 論 文

声道の共振特性を考慮した
歌声合成システムの構築に関する研究

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

長田 和也

2015 年 3 月

修 士 論 文

声道の共振特性を考慮した
歌声合成システムの構築に関する研究

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

1310049 長田 和也

提出年月: 2015 年 2 月

Copyright © 2015 by Nagata Kazuya

概 要

本稿では，声道の共振特性を考慮した歌声合成システムを構築するために，フォルマント制御モデルを構築する．そして，その有効性を評価実験を行い検討する．フォルマント制御モデル構築のためにフォルマント分析をした結果，F0 の変化に合わせて，4 つの区間で F1 と F2 が異なる特徴的な変化を行うことが確認された．1 つ目の区間は 300Hz 以下の区間である．この区間は modal で歌われている区間であり，F0 の変化に合わせて F1 はわずかに減少し，F2 はわずかに増加することが確認された．2 つ目の区間は 300Hz から 336.64Hz の区間である．この区間は modal から falsetto への変化が発生している区間であり，F0 の変化に合わせて F1 ，F2 ともに急激に減少することが確認された．3 つ目の区間は 336.64Hz から 454.64Hz の区間である．この区間は falsetto で歌われている区間のうち F0 が低い区間であり，F0 の変化に合わせて F1 はわずかに増加し F0 に近付く．また，F2 は減少して第二高調波に近付いていることが確認された．最後の区間である 4 つ目の区間は 454.64Hz 以上の区間である．この区間は falsetto で歌われている区間のうち F0 が高い区間であり，F0 の変化に合わせて F1 は F0 に沿って増加し，F2 は第二高調波に沿って増加することが確認された．この知見を元にフォルマント制御モデルを構築し，歌声合成音を合成し，評価実験を行った．その結果，提案手法による歌声合成音が相対的に不自然であると判断されたため，その原因を検討した．原因として，提案手法による歌声合成音における振幅の急激な変化や評価基準を挙げ，これらの改良を行い，再度評価実験を行った．その結果，振幅の制御前と比較して自然であると評価された割合が高くなった．つまり，振幅の影響により，提案手法の有効性が減少していたことが確認された．また，振幅の制御を行った音声は改良前の実験において最も自然であると評価された音声と同等に自然であると評価された．

目 次

第 1 章　序論 **1**
1.1　はじめに 1
1.2　研究の背景 1
1.3　本研究の目的・特色 2
1.4　本論文の構成 3

第 2 章　提案する歌声合成システムの概要 **5**
2.1　提案する歌声合成システムの前提条件 5
2.1.1　Klatt Formant Synthesizer 5
2.1.2　F0 制御モデル 7
2.1.3　LF モデル制御による声区表現 9
2.2　まとめ 11

第 3 章　声道の共振特性を考慮した歌声合成システムの構築 **13**
3.1　音声収録 13
3.1.1　収録機器 13
3.1.2　収録音声 13
3.2　フォルマント分析 15
3.2.1　分析条件 15
3.2.2　分析結果 15
3.3　フォルマント制御モデルの構築 16
3.3.1　F1 制御モデル 17
3.3.2　F2 制御モデル 18
3.4　まとめ 20

第 4 章　歌声合成音を使用した評価 **21**
4.1　評価実験 21
4.1.1　歌声合成音の作成 21
4.1.2　評価条件 27
4.1.3　評価結果 28
4.2　評価実験の改良 28
4.2.1　歌声合成音の分析 28

4.3 歌声合成音の制御 34
4.3.1 評価実験 35
4.3.2 評価結果 35
4.4 まとめ 36

第 5 章 結論 **37**
5.1 本研究のまとめ 37
5.2 今後の課題 38

第1章　序論

1.1　はじめに

本研究では，声道の共振特性を考慮して歌声合成を行うことにより，より自然な歌声を合成することを目的とする．歌声合成とは人工的に歌声を合成する技術である．代表的な歌声合成システムとして，楽器のように使用して演奏することが楽しまれているVOCALOID[1]や現代のオペラ歌手には歌うことの出来ない曲の再現を行ったFarinelli[2] が存在する．

歌声は話し声と比較して多様な発声法が存在する．人は声の生成系を利用することによってそれらの多様な表現を行うことが出来る．しかし，その仕組みは完全には解明されていない．声の生成系を考慮して音声を合成する研究は大きく分けて2つの手法に分けることが出来る．1つは，MRI等を使用して生成系の物理的な変位を計測し，物理モデルを計測結果に基づいて制御する手法[3][4]であり，もう一方は音声の音響的特徴量について分析し，ソースフィルタモデルに基づいて音響的特徴量を制御する手法[5]である．

本研究では，自然な歌声を合成するためにソースフィルタモデルに基づいた歌声合成システムを構築する．本研究によって自然な歌声の合成が可能になるだけでなく，声道の共振特性の解明につながる可能性がある．

1.2　研究の背景

歌声についての研究は古くから行われており，特にヨーロッパを中心としてオペラ歌唱についての研究が盛んに行われている．先に挙げたFarinelliはフランス国立音響音楽研究所（IRCAM）において作成された歌声合成システムであり，去勢された男性オペラ歌手（カストラート）の歌声を映画「カストラート」において再現している．また，スウェーデン王立工科大学（KTH）ではKlatt Formant Synthesizer [6]と呼ばれるフォルマント合成器が構築されており，Sundbergらにより歌声についての知見が多く報告されている[7]．歌声についての研究は，声の生成系をMRI等をもちいて物理的に計測し，計測結果を元に物理モデルを制御する手法が存在するが，物理量を計測するために大掛かりな装置が必要であったり，被験者の負担も大きいため，本研究は収録した歌声を分析し，その分析結果をもとに音響的特徴量を制御する手法を使用する．

歌声合成システムの手法として，VOCALOIDに代表される波形接続方式や隠れマルコフモデル（HMM）合成方式[8]といった手法が存在するが，これらの手法では声帯や声

道といった人が制御する器官に相当する部分を制御することができず，人がどの部分を制御することで歌声の発声を行っているかといった知見を得ることは出来ない．

また，歌声について先行研究において数々の知見が報告されている．まず，歌声のスペクトル包絡に含まれる特徴として歌唱フォルマント [9] と呼ばれるものがある．これは 3 kHz 付近に見られるフォルマントとは異なるスペクトルピーク成分のことである．また，ソプラノ歌手が基本周波数（F0）を変化させながら歌唱した場合，F0 が第 1 フォルマント（F1）より高くなるようなとき，F1 が第 1 高調波に近付くことが確認されている [10]．

F0 については，齋藤らの研究 [11] において歌声らしさの特徴量としてヴィブラートと呼ばれる，F0 と振幅包絡が 4〜6 Hz の変調周波数で振動する現象やプレパレーションやオーバーシュートといった歌声らしさに影響を与える F0 の変化と歌唱フォルマントを適用することによって歌声合成を行っている．また，元田らによる研究 [12] では ARX モデル [13][14] と Liljencrants-Fant （LF）モデル [15][16] を組み合わせた ARX-LF モデルを利用し，声帯モデルである LF モデルを制御することにより Vocal fry, Modal, Falsetto といった 3 つの声区を表現することを可能としている．

しかし，これらの歌声合成システムでは STRAIGHT[17] を使用しているため，声道の共振特性であるフォルマント成分について独立して制御することが難しく，STRAIGHT を使用している先行研究 [18][19] ではスペクトル包絡の制御によってフォルマント成分の制御を行っている．そこで，本研究ではフォルマント成分について独立して制御することが出来る Klatt Formant Synthesizer を使用して歌声合成システムを構築する．このシステムにフォルマント分析によって得られたフォルマント制御モデルを適用することでより自然な歌声の合成を行う．

1.3 本研究の目的・特色

本研究の目的は，声道の共振特性を取り入れた上で話声と音高から自然な歌声を合成することである．

本研究で提案する歌声合成システムの特色としてフォルマント制御モデルに基づいてそれぞれのフォルマントを独立して制御することにより歌声を合成する点である．既存の研究において，声帯音源特性の操作により人が幅広い音域をカバーするために声質を変化させる表現を歌声合成音として合成することは達成されているが，声道の共振特性に着目してスペクトル包絡全体の形状を変化させることによりフォルマントを制御して，人の声質変化を表現することは行われている．しかし，それぞれのフォルマントを独立して制御することにより人の声質変化を表現することは十分に達成されていない．それぞれのフォルマントを独立して制御することにより，フォルマントごとの変化を再現することでより声道の共振特性をより正確に反映した制御モデルを構築することが出来る．本研究で提案する歌声合成システムはこの制御モデルを導入することで，より自然な歌声の合成が可能になると考えている．

1.4　本論文の構成

本論文は 5 章で構成される．

第 1 章

本研究の背景と先行研究の課題を明らかにし，本研究の目的と特色を示す．

第 2 章

提案する歌声合成システムの概要を述べる．
本研究で提案する歌声合成システムでは合成器として，Klatt Formant Synthesizer を使用する．声道の共振特性を表現するために，それぞれのフォルマントについて独立して制御が出来る必要がある．Klatt Formant Synthesizer は複数の共鳴器の制御としてそれぞれのフォルマントのパラメータを独立して入力として与えられるので，声道の共振特性を表現するためには適している．また，Klatt Formant Synthesizer のもう 1 つの主要な入力として必要な音源を生成するために歌声らしさを表現するための F0 制御モデルと声区表現を可能にするための LF モデル制御を使用する．これらのモデルを使用することにより，声帯音源について歌声らしさと声区表現が含まれた音源を生成することが出来るようになる．
本章では Klatt Formant Synthesizer の概要を述べた後，音源生成に使用する F0 制御モデルと LF 制御モデルについて述べる．そして，提案する歌声合成システムの構成を述べる．

第 3 章

Klatt Formant Synthesizer の共鳴器を制御するための入力として声道の共振特性を反映したフォルマントパラメータを用意する必要がある．そこで，歌声を分析し，基本周波数ごとの声道の共振特性を反映したフォルマントパラメータを得る．そして，分析結果に従って基本周波数ごとにフォルマントパラメータを制御するモデルを構築する．
　本章では分析対象とする音声の収録について述べた後，収録音声のフォルマント分析とその結果について述べる．そして，フォルマント分析結果をもとに構築した提案するフォルマント制御モデルについて述べる．

第 4 章

構築されたフォルマント制御モデルの有効性を評価するために評価実験を行う．先行研究をもとに合成された歌声と比較し，提案手法の有効性を示す．
　本章では構築したフォルマント制御モデルを使用して歌声合成を行った評価実験について述べる．

評価実験に使用した音声の作成方法について述べた後，評価実験の評価条件を述べる．そして，評価実験の結果について述べる
h 結果として，提案手法より先行研究による歌声合成音の方が相対的に自然であると評価された原因について考察する．
提案手法より自然であると評価された歌声合成音と提案手法による歌声合成音を分析し，提案手法が不自然であると判断された原因を検討する．
それから，検討結果を踏まえて評価実験を再度行う．その結果得られた評価結果について考察を行う．

第 5 章

本研究で得られた結果のまとめと今後の課題について述べる．

第2章　提案する歌声合成システムの概要

提案する歌声合成システムの概要とシステムを構成する要素の詳細を述べる．提案する歌声合成システムの合成器は Klatt Formant Synthesizer を使用する．声帯音源は歌声らしさに関係する音響特徴量をもとに構築された F0 制御モデルと声区表現を可能とする LF モデル制御を使用する．

2.1　提案する歌声合成システムの前提条件

提案する歌声合成システムを構築する要素として使用する，合成器の Klatt Formant Synthesizer と声帯音源を生成に使用する F0 制御モデルと LF モデル制御について述べる．

2.1.1　Klatt Formant Synthesizer

Klatt Formant Synthesizer は図 2.1 のような処理を行い，歌声を合成する．このシンセサイザーはこれまでに提案されていた優れたシンセサイザーに共通する，直列フォルマントシンセサイザーと並列フォルマントシンセサイザーと呼ばれる 2 種類の基本構成を組み合わせたものである．

直列フォルマントシンセサイザーでは，直列につながれた共鳴器によって共鳴音が合成される．並列フォルマントシンセサイザーは，フォルマントの振幅に関係する振幅制御を各フォルマントの共鳴器の前に行う．直列フォルマントシンセサイザーの利点は母音において各フォルマントの振幅を独立して制御せずとも，振幅同士の関係が適切になる点と非鼻音の共鳴音を合成する場合において声道変換関数の精度が高くなるという点である，欠点は摩擦音と破裂音を生成するには並列に構成する必要があり，並列フォルマントシンセサイザーと比較して全体的に複雑になる点である．

このフォルマントシンセサイザーは基本構造としてディジタル共鳴器から構成される．この共鳴器はフォルマント周波数 F とバンド幅 BW をパラメータとして，入力 $x(nT)$ から式 2.1 に従って出力 $y(nT)$ を生成する．

$$y(nT) = Ax(nT) + By(nT - T) + Cy(nT - 2T) \tag{2.1}$$

このとき，$y(nT - T)$ ，$y(nT - 2T)$ はそれぞれ出力 $y(nT)$ の 1 つ前，2 つ前のサンプルであり，A，B，C はフォルマント周波数 F とバンド幅 BW に依存する変数である．A，B，

C それぞれの変数とフォルマント周波数 F ，バンド幅 BW の関係は式 2.2 で表される．

$$
\begin{aligned}
C &= -\mathrm{e}^{-2\pi BW\,T} \\
B &= 2\mathrm{e}^{-\pi BW\,T}\cos(2\pi F\,T) \\
A &= 1 - B - C
\end{aligned}
\tag{2.2}
$$

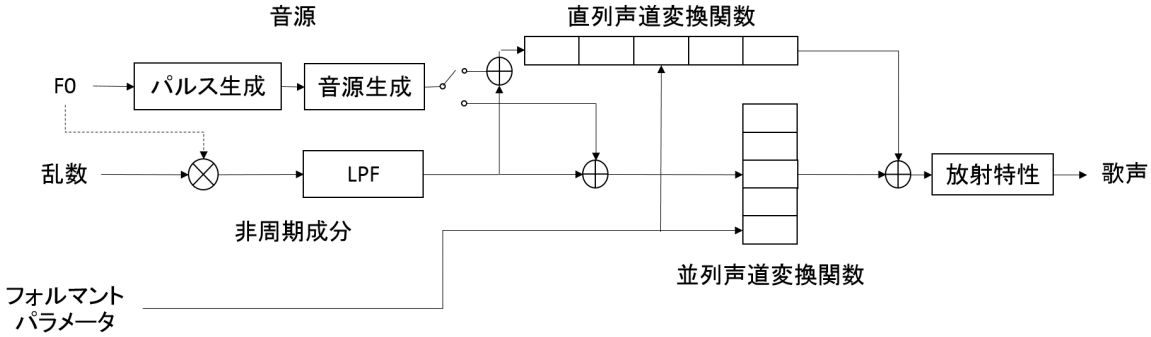

図 2.1: Klatt Formant Synthesizer の概要

2.1.2 F0 制御モデル

齋藤らによる研究 [11] において，歌声らしさの知覚モデルに基づいた歌声特有の音響特徴量が指摘されている．本研究では，歌声特有の音響特徴量の知見から齋藤らにより構築された F0 制御モデルを利用する．図 2.2 に本研究で使用する F0 制御モデルを示す．

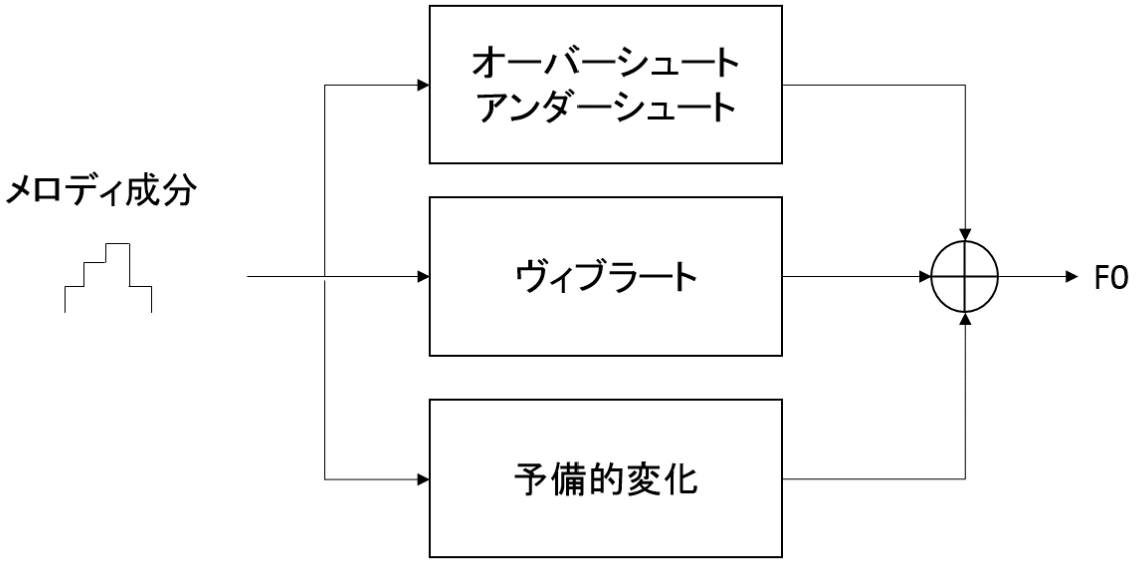

図 2.2: 本研究で使用する F0 制御モデルの概要

このモデルでは，矩形パルスで記述されるメロディ成分を入力とし，オーバーシュート・アンダーシュート，ヴィブラート，予備的変化を加えることにより，歌声における F0 変化にあわせて F0 を制御するモデルである．オーバーシュート・アンダーシュート，ヴィブラート，予備的変化といった変動を制動 2 次系のパルス応答で記述する．その制動 2 次伝達関数 $H(s)$ は式 2.3 で記述される．

$$H(s) = \frac{\Omega}{s^2 + 2\zeta\Omega s + \Omega^2} \tag{2.3}$$

このとき，ζ は減衰振動，Ω は固有振動数値であり，このシステムのインパルス応答 $h(t)$ は ζ によって式 2.4，2.5，2.6，2.7 のように場合わけされる．

$|\zeta| > 1$ のとき

$$h(t) = \frac{\Omega}{2\sqrt{\zeta^2 - 1}} \left(\mathrm{e}^{\lambda_1 \Omega t} - \mathrm{e}^{\lambda_2 \Omega t} \right), \lambda_{1,2} = -\zeta \pm \sqrt{\zeta^2 - 1} \tag{2.4}$$

$|\zeta| < 1$ のとき

$$h(t) = \frac{\Omega}{2\sqrt{1 - \zeta^2}} \mathrm{e}^{-\zeta \Omega t} \sin \sqrt{1 - \zeta^2 \Omega t} \tag{2.5}$$

$|\zeta| = 1$ のとき

$$h(t) = \Omega t \mathrm{e}^{-\Omega t} \tag{2.6}$$

$|\zeta| = 0$ のとき

$$h(t) = \sin \Omega t \tag{2.7}$$

このシステムは ζ,Ω という 2 つのパラメータを決定することで，異なる特性を持つインパルス応答を得ることが出来る．そこで，最適な F0 制御を行うためにはこの 2 つのパラメータを最適なものとする必要がある．そこで，歌声データと F0 制御モデルのフィッティングを行うことにより，最適なパラメータを求めている．F0 制御モデルのオーバーシュート・プレパレーション，ヴィブラート，予備的変動に最適なパラメータは表 2.1 の通りとなる．

表 2.1: パラメータ

歌声動的変動成分	Ω [rad/ms]	ζ
オーバーシュート・アンダーシュート	0.031	0.52
ヴィブラート	0.033	0
予備的変動	0.028	0.72

2.1.3 LF モデル制御による声区表現

元田による研究 [12] において，声区表現のための LF 制御モデルが提案されている．声区とは幅広い音域を発声するために，人が行う発声法や声質の変化によって区分したものであり，ボーカルフライ（vocal fry），地声（modal），裏声（falsetto）といった声区が存在する．本研究では，幅広い音域を歌い分けるときの声帯音源特性を取り入れるため，元田により構築された LF 制御モデルを使用する．図 2.3 に本研究で使用する LF 制御モデルを示す．

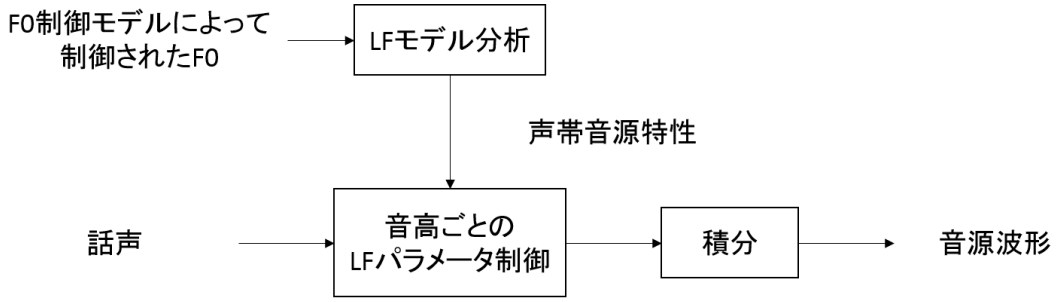

図 2.3: 本研究で使用する LF 制御モデルの概要

LF モデルは口唇の放射特性を含んだ微分声門波形を表現する．LF モデルは式 2.8 で表される．

$$E(t) = \begin{cases} E_0 e^{\alpha t} \sin(\omega t) & (0 \leq t \leq t_e) \\ \dfrac{-E_e}{\epsilon t_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}] & (t_e \leq t \leq t_c) \end{cases} \tag{2.8}$$

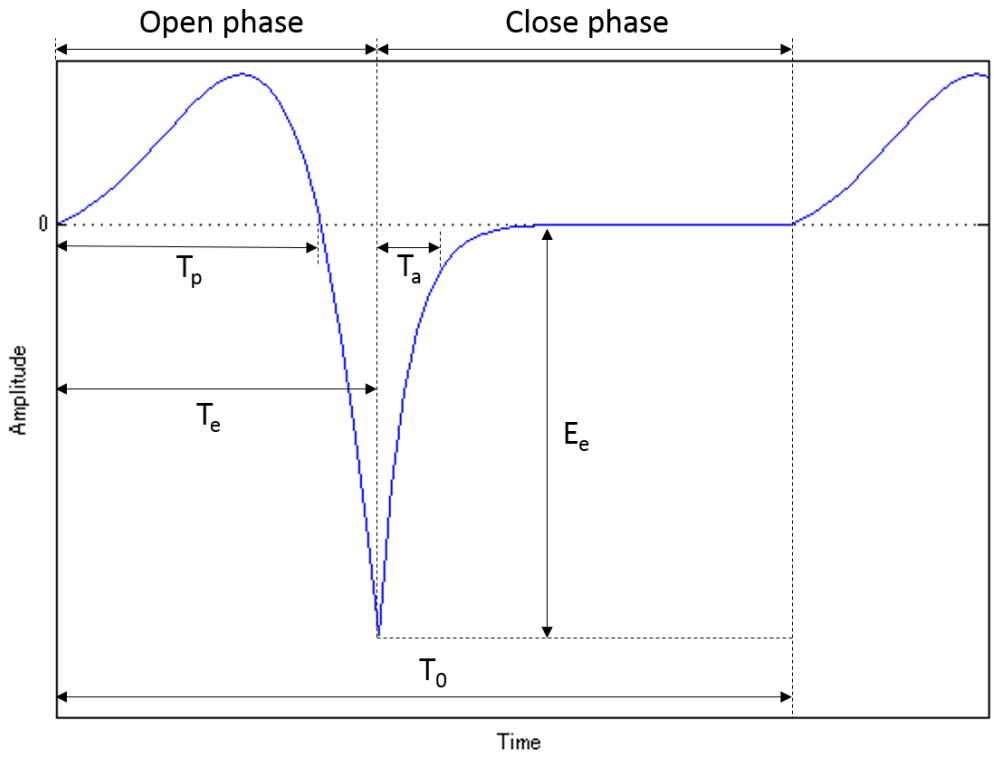

図 2.4: LF モデルにより生成される典型的な微分声門波形

図 2.4 に LF モデルで生成される典型的な微分声門波形を示す．

このとき，1 周期の終了時刻が T_0，振幅パラメータが E_e で表される．また，T_e は声門閉鎖地点，T_p は声門最大開口地点，T_a は声門完全閉鎖までに要する戻り区間である．

元田による LF モデル制御では，制御を簡易化するために O_q,α_m, Q_a という 3 つのパラメータを制御している．3 つのパラメータはそれぞれ O_q が声門開口率，α_m が声帯音源信号の開口区間の左右対称性，Q_a が声門完全閉鎖までに要する戻り区間の時間率を表しており，式 2.9 で表される．

$$
\begin{aligned}
O_q &= \frac{T_e}{T_0} \\
\alpha_m &= \frac{T_p}{T_e} \\
Q_a &= \frac{T_a}{(1-O_q)T_0}
\end{aligned}
\tag{2.9}
$$

3 つのパラメータと声帯音源特性の関連性はそれぞれ次のように表される．

O_q
声帯振動にとって主要な情報である声門開口率（基本周波数 1 周期に対する声門が開いている時間の割合）に対応している．

α_m
声門開口・閉鎖の速さの比率を表し，声門抵抗や声帯の緊張の影響によって変化する．

Q_a
声門閉鎖が不完全である際に発生する乱流を表し，声門閉鎖の強さの影響によって変化する．

これらのパラメータの変化が 3 つの声区（vocal fry，modal，falsetto）ごとの歌声を分析することによって求められている．3 つの声区におけるそれぞれのパラメータの平均値は表 2.2 のようになる．

表 2.2: 声区ごとの LF パラメータ

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

2.2 まとめ

提案する歌声合成システムの構成を図 2.5 に示す．この歌声合成システムは入力としてメロディ成分である楽譜と話声を使用する．話声を LF モデルで分析することにより，話者の声帯音源特性を得る．その声帯音源特性を F0 制御モデルによって制御された F0 の音高ごとに LF モデル制御を行うことにより音源波を生成する．また，話声のフォルマント分析結果をもとにフォルマント制御モデルでフォルマント周波数を制御することにより，音高の変化による声道の共振特性を表現する．そして，Klatt Formant Synthesizer に音源波とフォルマント成分を入力することで歌声を生成する．

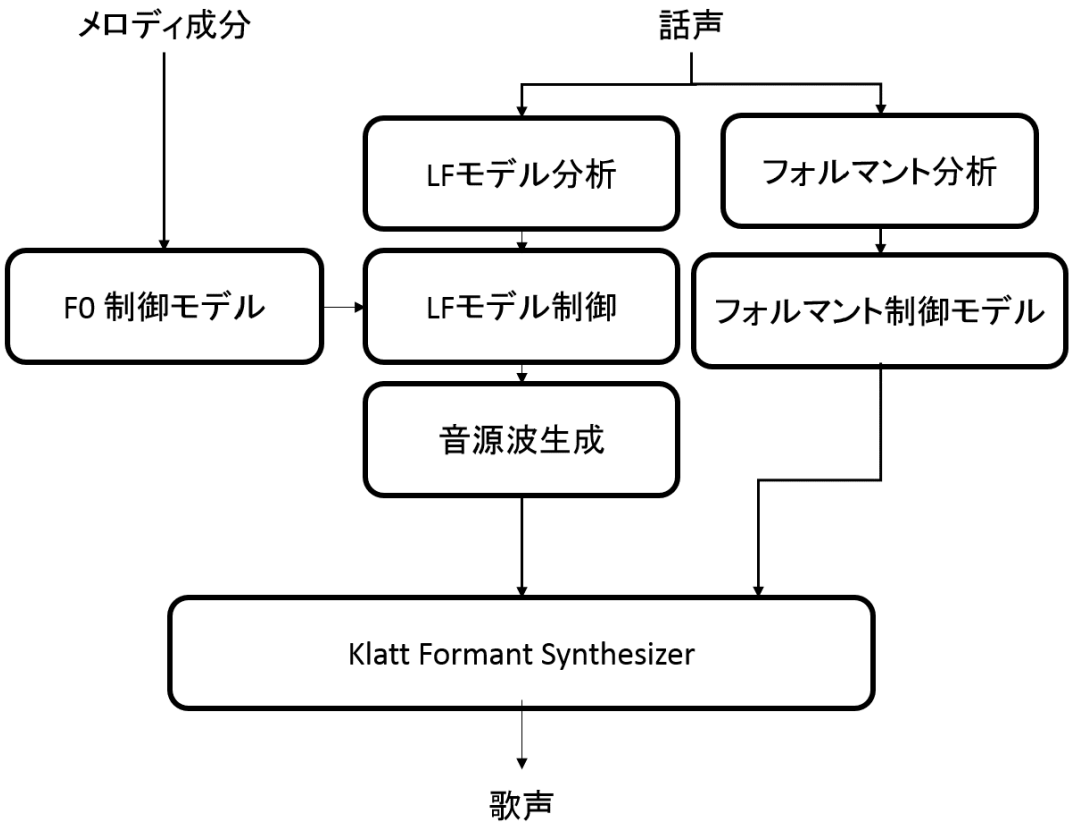

図 2.5: 提案するシステムの構成

第3章　声道の共振特性を考慮した歌声合成システムの構築

3.1　音声収録

フォルマント制御モデルを構築するために歌声のフォルマント分析を行う必要がある．そこで，声楽の経験がある男性1名の歌声と話声を収録した．

3.1.1　収録機器

収録に使用した機器を表3.1に示す．収録の際には被験者の首にEGGを装着して収録を行った．EGGは声門の変位を記録できるものであり，LFモデルにおける声門の最大開口部，最大閉口部の位置を推定するときに使用する．収録は防音室内で行い，マイクと口との距離が30cm　程度となるように指示した．これはEGGの波形と音声波形の時間差を1ms程度に固定するためである．

表3.1: 収録に使用した機器

使用機器	メーカー	型番
マイク	audio-technica	AT4050
EGG測定装置	Laryngograph	Speech Studio
アンプ	DENON	DRA-F109
iInterface	Roland	OCTA-CAPTURE

3.1.2　収録音声

収録音声は声楽経験のある男性1名の歌声と話声である．収録した音声の種類とデータ数を表3.2に示す．収録した音声は話声の/a/,/i/,/u/,/e/,/o/．歌声の/a/で図3.1のような，最低音と最高音の関係が五度となる上昇下降音形を歌ったものを収録した．このときの最低音はF3（174.61Hz）からB4（493.88Hz）まで半音ずつ変化させた19パターンのメロディーを各3回ずつ収録した．収録した歌声の音域に含まれる音名とF0の対応を表3.3に示す．

表 3.2: 収録音声

音声の種類	発話内容	データ数
歌声	音高変化のある/a/	19 種類，　各 3 テイク
話し声	/a/, /i/, /u/, /e/,/o/	3 テイク

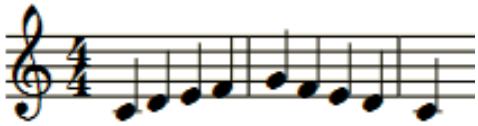

図 3.1: 収録した上昇下降音形の例（最低音： C4, 261.63Hz）

表 3.3: 音名と基本周波数の関係

音名	基本周波数 (Hz)	音名	基本周波数 (Hz)
F3	174.61	Fs4	369.99
Fs3	185	G4	392
G3	196	Gs4	415.3
Gs3	207.65	A4	440
A3	220	As4	466.16
As3	233.08	B4	493.88
B3	246.94	C5	523.25
C4	261.63	Cs5	554.37
Cs4	277.18	D5	587.33
D4	293.66	Ds5	622.25
Ds4	311.13	E5	659.25
E4	329.63	F5	698.46
F4	349.23	Fs5	739.99

収録した音声波形と EGG 波形の一例を図 3.2 に示す．EGG 波形に示した×印は声帯の最大閉口部，○印は声帯の最大開口部に相当する．上の音声波形と下の EGG 波形を比較すると，上の音声波形が変化を開始する時間は下の EGG 波形が変化を開始する時間に比べて 1ms 程度遅れていることが確認できる．

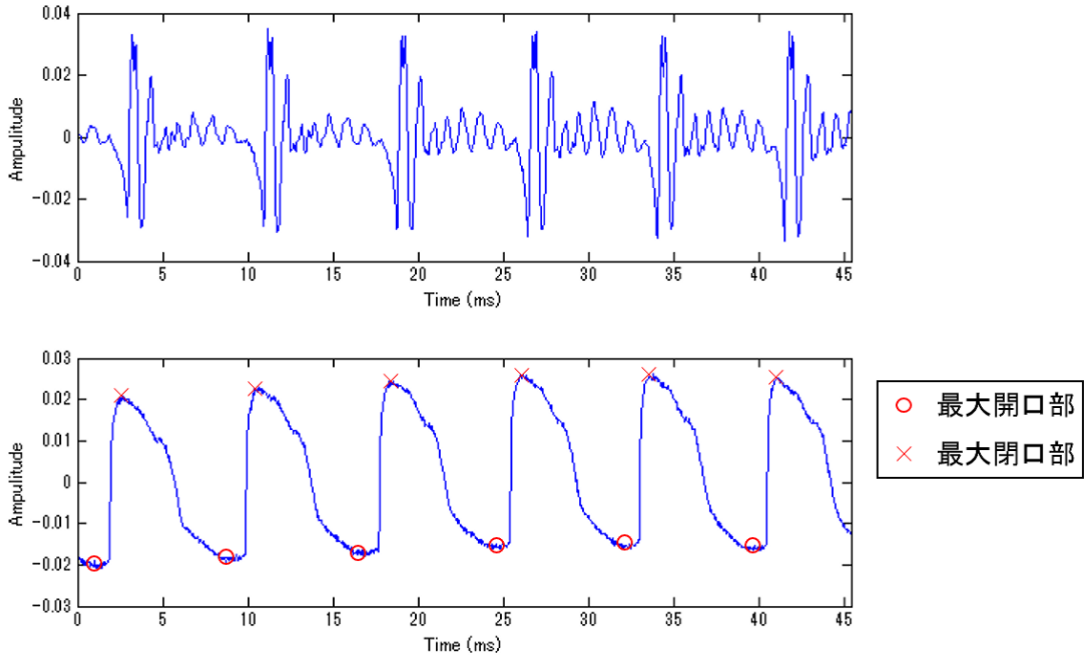

図 3.2: （上）　音声波形　（下）　EGG 波形

3.2　フォルマント分析

フォルマント制御モデルを構築するためにフォルマント分析を行う．

3.2.1　分析条件

分析条件を示す．収録した歌声を音高ごとに分割し，それぞれの音声信号にハミング窓を使用し，LPC 分析を行った．LPC 次数は 10 である．

3.2.2　分析結果

分析結果を図 3.3 に示す．点で示されたものが各音声のフォルマント分析結果であり，線で示されたものが 10 次のメディアンフィルタを使用して 10 近傍のサンプルに対する中央値を求めたものである．黒線は下から F0，F0 の 2 倍（第 2 高調波），F0 の 3 倍（第 3 高調波）を示している．

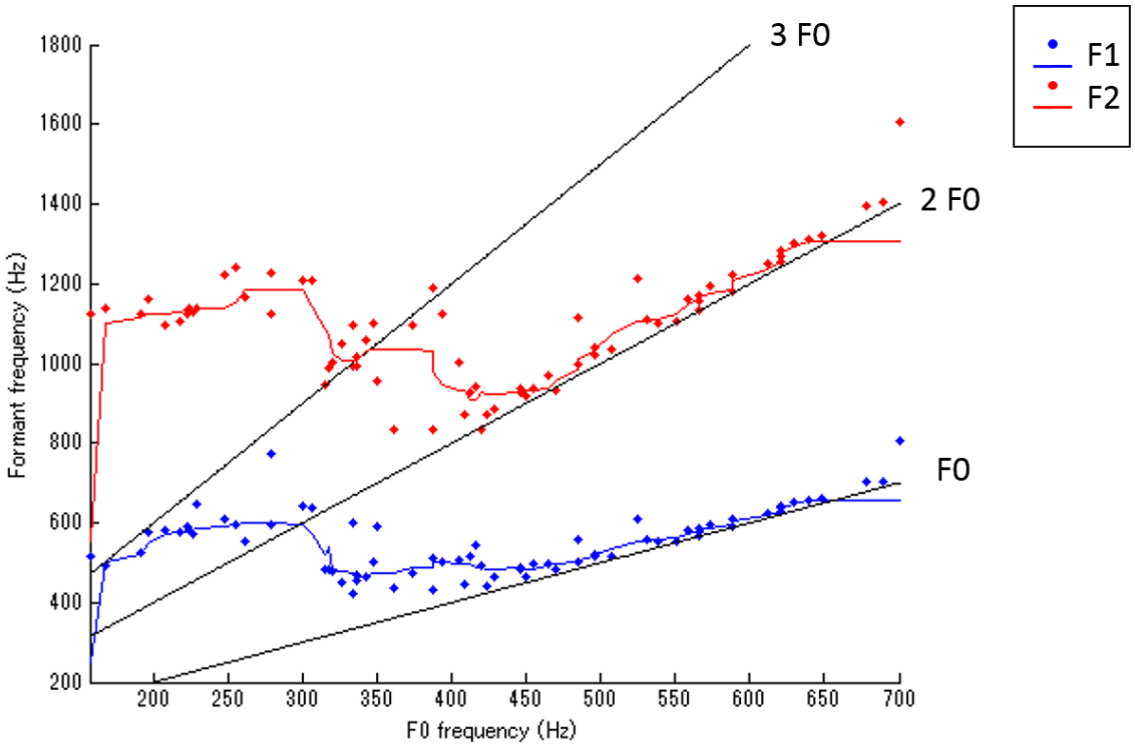

図 3.3: フォルマント分析結果

F0 が 300Hz 以下のすべて modal で歌われている区間については F1 は変化が少なく，F2 はわずかに増加している．F0 が 300Hz から 336.64Hz までの modal と falsetto が混在する区間については F1，F2 ともに急激に減少しており，F1 は第 2 高調波付近から F0 と第 2 高調波の中間付近，F2 は第 3 高調波付近へと移動することが確認される．F0 が 336.64Hz 以上のすべて falsetto で歌われている区間については大きく分けて 2 つの変化が見られる．まず，F0 が 336.64Hz から 454.65Hz の区間については，F1 が F0 と第 2 高調波の中間付近から F0 に近付いている．また，F2 については第 3 高調波から第 2 高調波に近付いている．そして，F0 が 454.64Hz 以上の区間については，F1 ，F2 それぞれが F0，第 2 高調波に合わせて移動していることが確認された．

3.3 フォルマント制御モデルの構築

分析結果より，フォルマント制御モデルを構築する．分析結果より modal で歌われる部分，modal と falsetto の変化がある部分，falsetto のうち F0 が低い部分，falsetto のうち高調波に合わせてフォルマントの変化が見られる部分の 4 つの部分に分けてフォルマント制御モデルを構築した．modal から falsetto への変化を表現するために，話声の/a/

について分析を行い，F0，第 5 フォルマントまでの各フォルマント周波数（F1 から F5）を求め，その周波数を基準に変化させるモデルを構築した．話声の F0 は 157.50Hz，F1 は 640.21Hz，F2 は 1128.51Hz，F3 は 2586.74Hz，F4 は 2918.16Hz，F5 は 4797.41Hz である．

3.3.1 F1 制御モデル

F0 の変化による，声道の共振特性を表現するために，F1 制御モデルを構築した．F1 制御モデルは基本周波数 $F0$ によって式 3.1，3.2，3.3，3.4 のように場合わけされる．F1 が F0 に合わせて増加する部分は各サンプルごとに F1 と F0 の差を求め，中央値をとった結果として 15Hz を使用した．

$F0 \leq 300$ のとき

$$F_1 = -0.3201F0 + 690.6211 \tag{3.1}$$

$300 < F0 \leq 336.64$ のとき

$$F_1 = -3.3193F0 + 1590.4 \tag{3.2}$$

$336.64 < F0 \leq 454.64$ のとき

$$F_1 = 0.1517F0 + 421.9133 \tag{3.3}$$

$F0 > 454.64$ のとき

$$F_1 = F0 + 15 \tag{3.4}$$

F0 が 150Hz から 700Hz まで変化したとき，この F1 制御モデルによって制御された F1 の変化を図 3.4 に示す．図中の黒線は下から順に F0，第二高調波である．

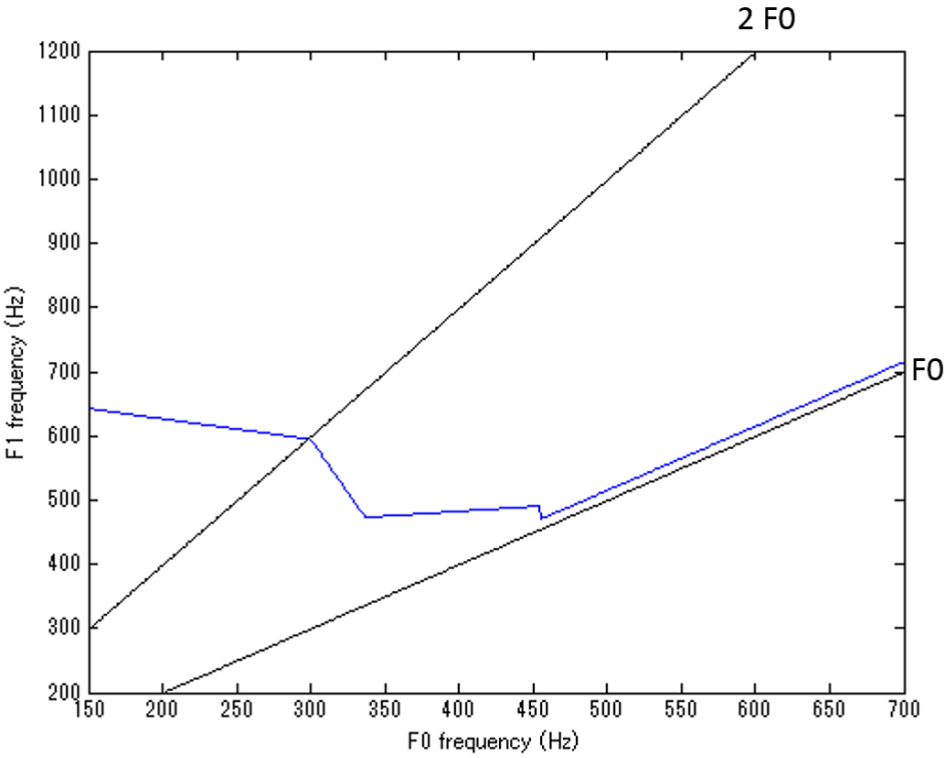

図 3.4: F1 制御モデル

3.3.2 F2 制御モデル

F1 制御モデルと同様に，F0 の変化による声道の共振特性を表現するために，F2 制御モデルを構築した．F2 制御モデルは基本周波数 $F0$ によって式 3.5，3.6，3.7，3.8 のように場合わけされる．F2 が第二高調波に合わせて増加する部分は各サンプルごとに F2 と第二高調波の差を求め，中央値をとった結果として 35Hz を使用した．

$F0 \leq 300$ のとき

$$F_2 = 0.4112F0 + 1063.8 \tag{3.5}$$

$300 < F0 \leq 336.64$ のとき

$$F_2 = -4.2303F0 + 2456.2 \tag{3.6}$$

$336.64 < F0 \leq 454.64$ のとき

$$F_2 = -0.8212F0 + 1308.5 \tag{3.7}$$

$F0 > 454.64$ のとき

$$F_2 = 2F0 + 35 \tag{3.8}$$

F0 が 150Hz から 700Hz まで変化したとき，この F2 制御モデルによって制御された F2 の変化を図 3.5 に示す．図中の黒線は下から順に第二高調波，第三高調波である．

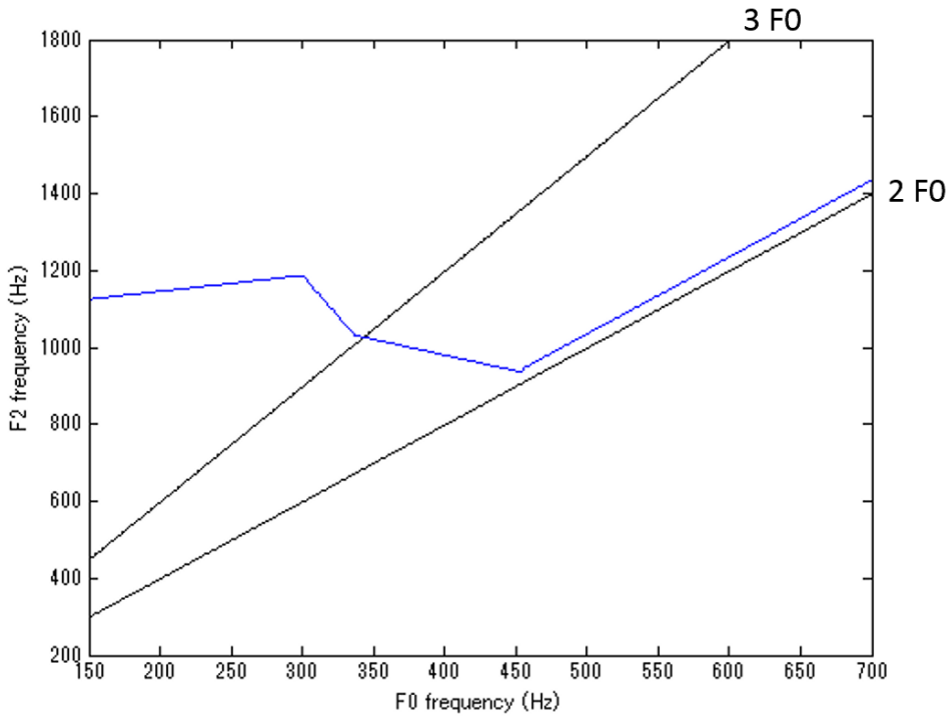

図 3.5: F2 制御モデル

3.4 まとめ

フォルマント分析結果から構築したフォルマント制御モデルを図 3.6 に示す．分析結果では，大きく分けて 4 つのフォルマント周波数の変化が見られた．4 つの区間のうち 3 つ目の区間については，先行研究 [20] で報告されている，裏声らしい歌声に見られるフォルマントの F0 への接近が見られている．このモデルによって，modal から falsetto へ移動するときのフォルマントの急激な変化やフォルマントが高調波に合わせて移動する変化を制御できるようになる．このモデルを歌声合成システムに使用することで，声道の共振特性に従った変化を合成される歌声にもたせる．

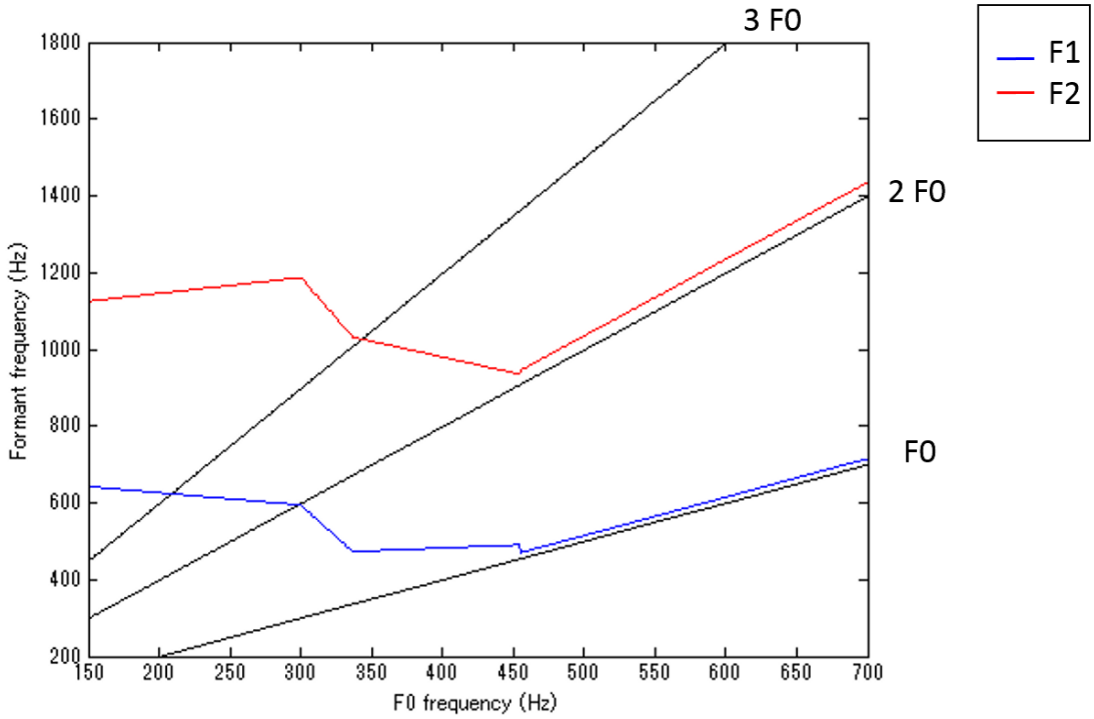

図 3.6: フォルマント制御モデル

第4章　歌声合成音を使用した評価

4.1　評価実験

フォルマント制御モデルの有効性を評価するために，歌声合成音を作成し，その評価を行う．

4.1.1　歌声合成音の作成

本研究において提案するフォルマント操作モデルの有効性を評価するために，4 種類の歌声合成音を生成した．そのうちの 1 種類として，声帯音源特性，声道の共振特性双方について制御を行わずに合成した音声として，声帯音源として Rosenberg 波を使用したものを用意した．Rosenber 波は式 4.1 で表現される．τ_1 は最大開口時間，τ_2 は最大開口時間から最大閉口時間の差，a は最大振幅値である．ここでは $\tau_1 = 0.40$，$\tau_2 = 0.16$ とした．

$$f(t) = \begin{cases} a\{3(\frac{t}{\tau_1})^2 - 2(\frac{t}{\tau_1})^3\} & (0 \leq t \leq \tau_1) \\ a\{1 - (\frac{\mathrm{t} - \tau_1}{\tau_2})^2\} & (\tau_1 \leq t \leq \tau_1 + \tau_2) \end{cases} \tag{4.1}$$

図 4.1 に典型的な Rosenberg 波を示す．また，図 4.2 に LF モデルにより生成される声門波形を示す．

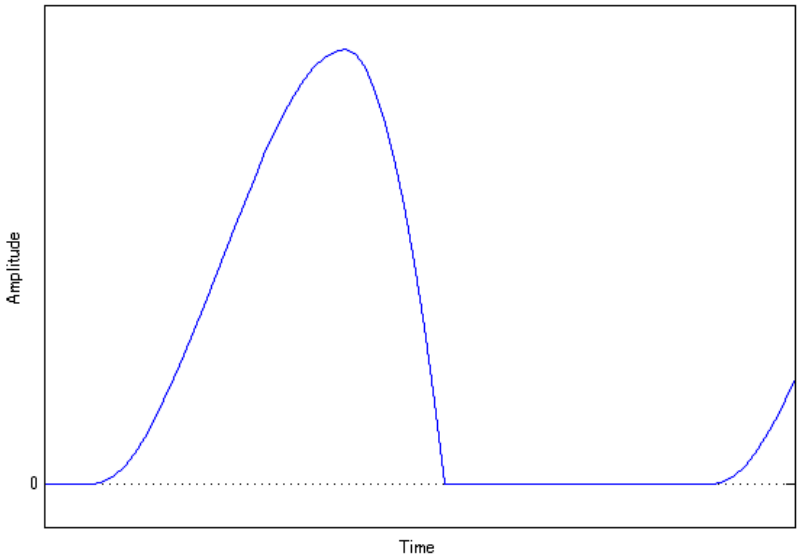

図 4.1: 典型的な Rosenberg 波

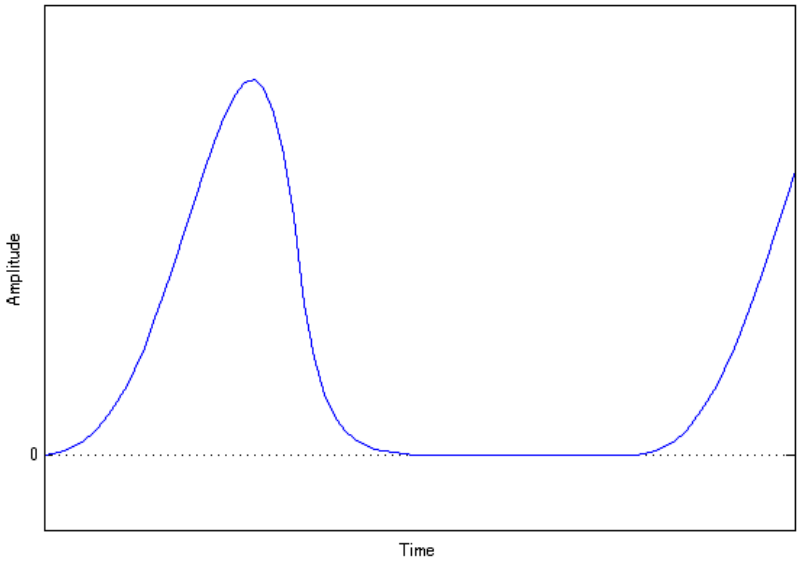

図 4.2: LF モデルにより生成される声門波形

表 4.1: 歌声合成音の種類

音声	音源	F0 制御モデル	LF モデル制御	フォルマント制御モデル
A	Rosenberg 波	未適用	未適用	未適用
B	Rosenberg 波	適用	未適用	未適用
C	LF モデル	適用	適用	未適用
D	LF モデル	適用	適用	適用

4 種類作成した歌声合成音の制御内容を表 4.1 に示す．作成した音声はそれぞれ，

- 音声 A

  声帯音源特性，声道の共振特性ともに操作なし．

- 音声 B

  声帯音源に対して歌声らしさに関係する音響特徴量の操作を行っている．声帯音源については歌声らしさを表現している．声道の共振特性については操作なし．

- 音声 C

  音声 B に加えて，LF モデル制御による声区表現を行っている．声帯音源については歌声らしさを表現しつつ声区表現も行っている．声道の共振特性については操作なし．

- 音声 D

  提案する歌声合成システムによって合成された歌声合成音．音声 C に加えてフォルマント制御モデルを適用することで声道の共振特性を表現している．歌声らしさを表現しつつ声区表現も行っている．

声帯音源特性

合成に使用するメロディ成分は C4（261.63Hz），E4（329.63Hz），G4（392Hz），C5（523.25Hz）である．これらの音高はそれぞれ modal で歌われる部分，modal と falsetto の変化がある部分，falsetto のうち F0 が低い部分，falsetto のうち高調波に合わせてフォルマントの移動が見られる部分に含まれる音高である．図 4.3 に歌声合成に使用するメロディ成分の周波数を示す．

この F0 に対して F0 制御モデルを適用し，歌声らしさに関係する歌声動的成分としてオーバーシュート・アンダーシュート，ヴィブラート，予備的変動を付与した．このとき

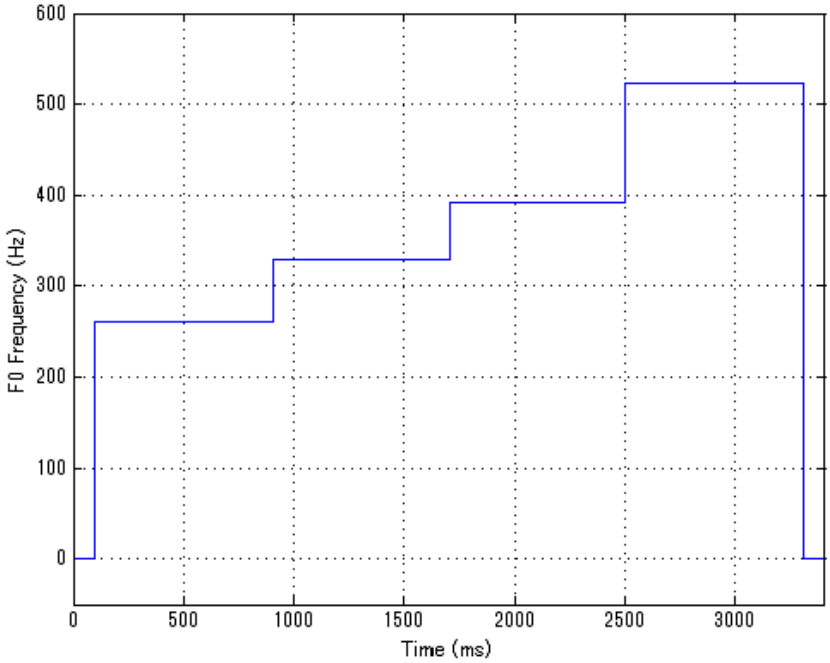

図 4.3: メロディ成分（F0 制御モデル未適用）

オーバーシュート・アンダーシュートの Ω は 0.0312，ζ は 0.512，予備的変動の Ω は 0.0312ζ は 0.512，ヴィブラートのヴィブラートの変動周波数は 5.495Hz である．図 4.4 に F0 制御モデルを適用した F0 の変化を示す．

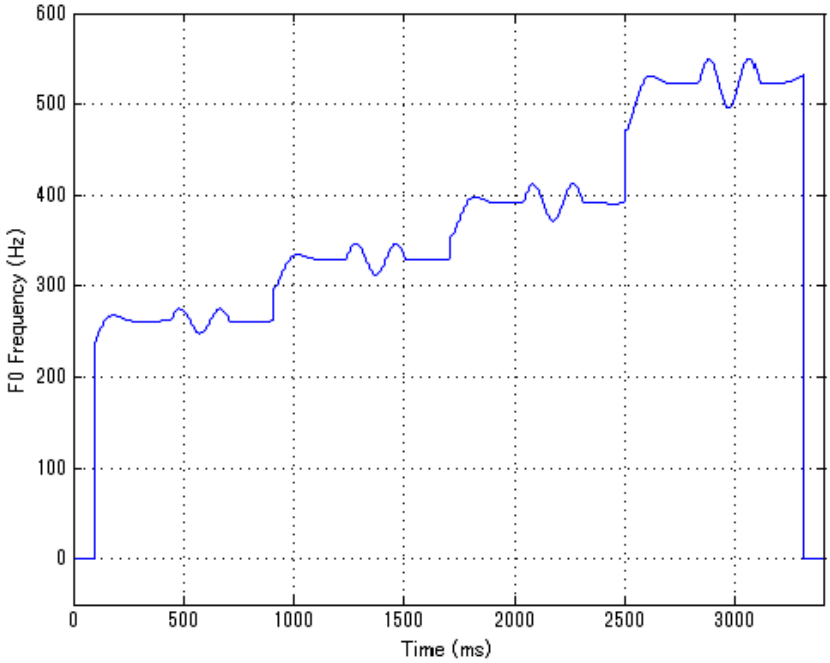

図 4.4: F0 制御モデルを適用した F0

声道の共振特性

合成に使用するフォルマント周波数は話声をフォルマント分析した結果得られた F1 から F5 までを使用する．図 4.5 にフォルマント制御モデルを使用せずに合成に使用するときのフォルマントのうち F1，F2 について示す．

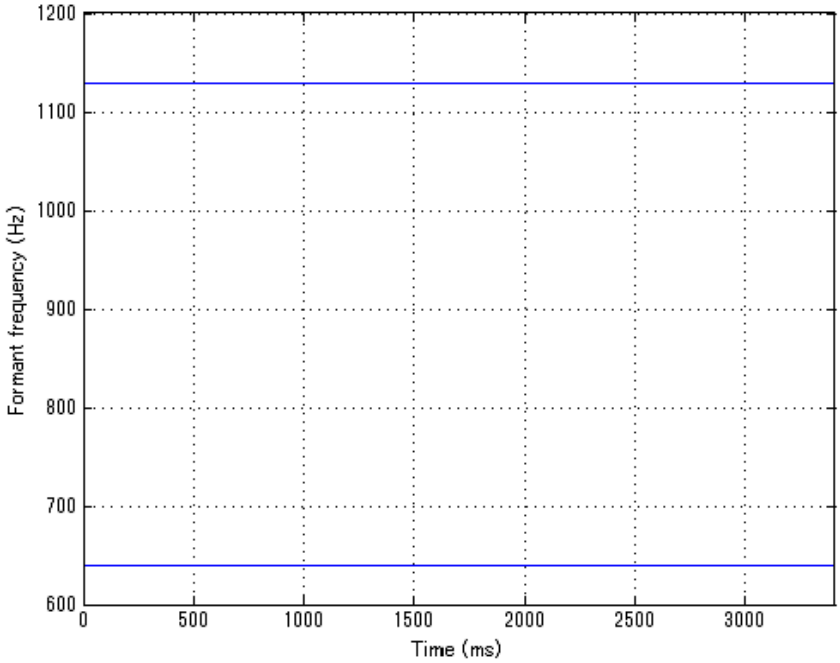

図 4.5: フォルマント制御モデル未適用のフォルマント

合成に使用するフォルマントのうち，F1，F2 についてフォルマント制御モデルを使用して制御を行った．これにより，声道の共振特性を歌声に取り入れる．歌声を分析した結果から得られた 4 つの制御モデルそれぞれについて制御することで，フォルマント周波数の特徴的な変化が再現することができる．図 4.6 にフォルマント制御モデルを適用してフォルマント制御を行った F1，F2 を示す．

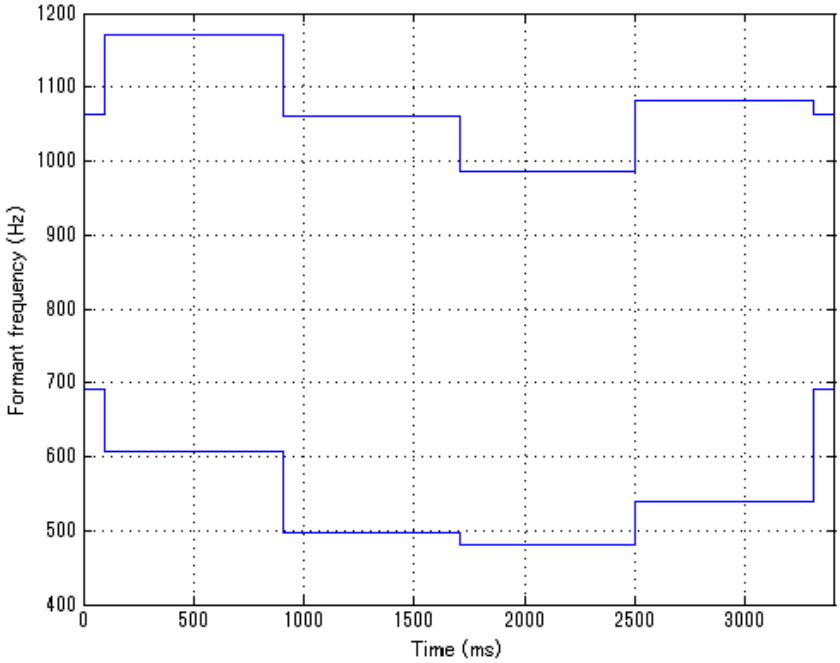

図 4.6: フォルマント制御モデルを適用したフォルマント

4.1.2 評価条件

評価実験を行うときの使用機器や評価の条件を示す．評価実験に使用した機器を表 4.2 に示す．

表 4.2: 評価実験に使用した機器

使用機器	メーカー	型番
ヘッドフォン	SENNHEISER	HDA200
アンプ	audio-technica	AT-HA21
インタフェース	Synthax	FIREFACE UCX

この評価実験では，ヘッドフォンで 2 つの音を順番に呈示し，どちらの音声がより地声から裏声への変化が自然な歌声であるかを選択させる．音声は 4 種類のうちの二 2 つの音声の組み合わせ 12 種類を各 10 回分，計 120 試行分用意した．これをランダムに並べ替え，評価実験は 1 セット 60 試行，1 人当たり 2 セットを行った．評価実験開始前に歌声合成音のメロディ成分にあわせた人の歌声を被験者に聞いてもらい，人の歌声の変化を参考に評価をしてもらった．評価基準は，2 つの音のうちどちらがより modal から falsetto への変化が自然な歌声であるかを選択してもらい，評価結果を集計して他の音声より自然な

変化であると選択された割合を評価する．ヘッドフォンから呈示される音は 70dB に調節した．

4.1.3 評価結果

評価結果を示す．評価基準は 2 つの音声を比較したとき，より modal から falsetto への変化が自然な歌声であるかという基準であり，他の音より自然であると評価された割合を比較する．図 4.3 に各被験者ごとの評価結果と全員の合計を示す．

表 4.3: 評価結果

	他の音声より自然であると評価された割合（%）			
被験者	A	B	C	D
A	6.67	56.67	55	81.67
B	1.67	85	75	38.33
C	1.67	58.33	85	55
D	0	71.67	80	48.33
E	1.67	51.67	86.67	60
F	0	75	73.33	51.67
合計	1.94	66.39	75.83	55.83

6 名の被験者に対して評価実験を行ったが，提案手法による歌声合成音が一番地声から裏声への変化が自然であるという評価になった被験者は 1 名，他の手法による歌声合成音の方が地声から裏声への変化が自然であるという評価になった被験者は 5 名となり，全体で見ると一番地声から裏声への変化が自然な歌声合成音は元田によるモデルを使用して合成された歌声合成音，二番目は齋藤らによるモデルを使用して合成された歌声合成音，提案手法は三番目という結果となった．

4.2 評価実験の改良

4.2.1 歌声合成音の分析

評価実験の結果，提案手法による歌声合成音は先行研究による歌声合成音と比較して相対的に不自然であると判断された．評価実験の後，被験者にどういった基準で地声から裏声への変化の自然さを評価したのかといったアンケートをとった結果，呈示音の 4 音目である，F1，F2 がそれぞれ高調波に合わせて移動するという音域の合成音について不自然であるという意見が半数以上から寄せられた．そこで，4 種類の歌声合成音の違いを比較する．図 4.7，4.7，4.7，4.7 に評価実験に使用した音声 A から D までの 4 種類の歌声合成

音の音声波形とスペクトログラムを示した．分析した音声のサンプリング周波数は 16kHz であり，スペクトログラムは 64 次のハミング窓とフーリエ変換長 4096 使用した．

評価実験において一番自然であると評価された歌声合成音である音声 C や二番目に自然であると評価された音声 B と提案手法による歌声合成音を比較すると．提案手法の 2 音目，3 音目の振幅が極端に小さく，逆に 4 音目の振幅が特に大きくなってしまっている．またスペクトルグラム上でも 4 音目とそれ以前ではエネルギーに大きな違いが見られる．これらの極端な変化により，提案手法による自然さの向上より不自然さの増加が上回ってしまい，先行研究での音声と比較して相対的に不自然であるという評価が行われたのではないかと推測した．そこで，自然な歌声の振幅包絡として最適であるとは言いきれないものの，評価実験で最も自然であると評価された音声の振幅包絡線を使用して，振幅包絡線にあわせて音声の振幅制御を行い，再度評価実験を行うこととした．

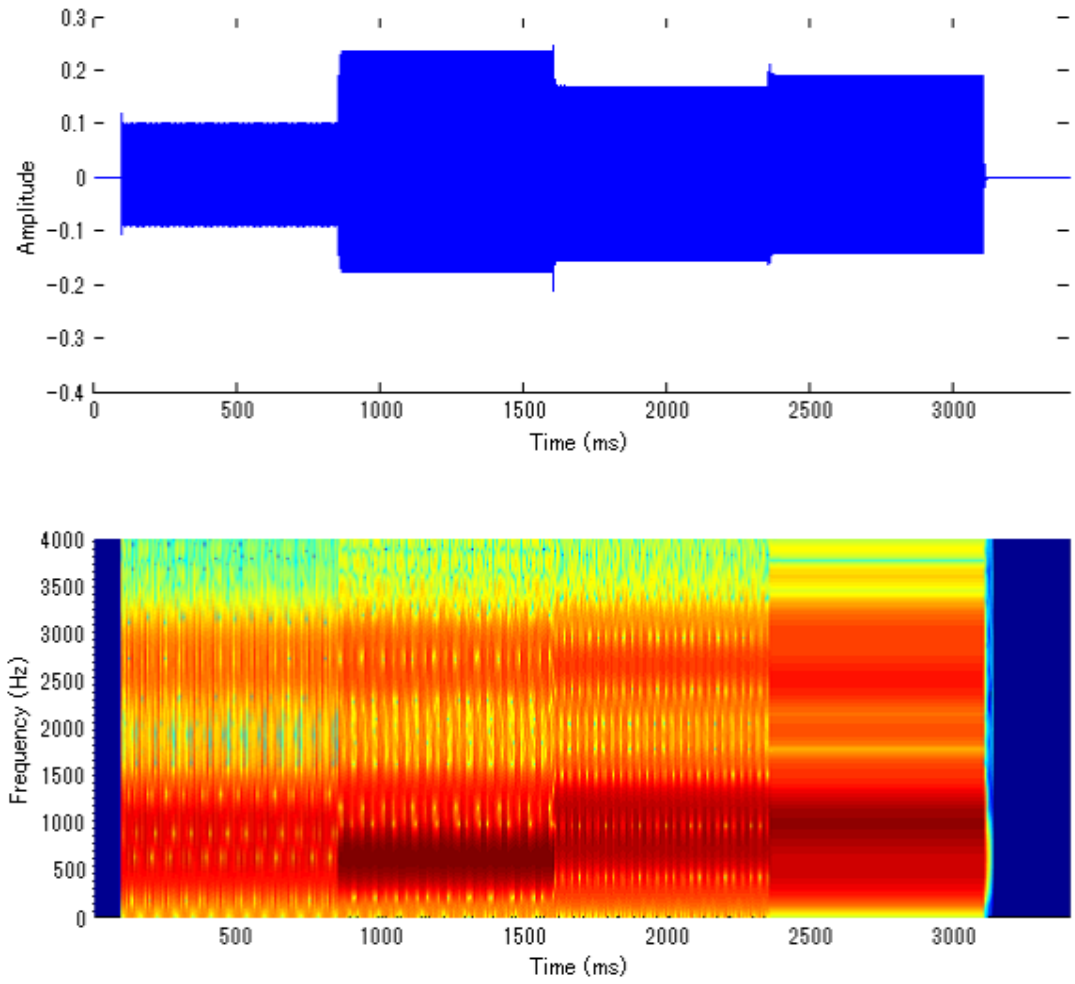

図 4.7: 音声 A の分析結果：（上）音声波形　（下）スペクトログラム

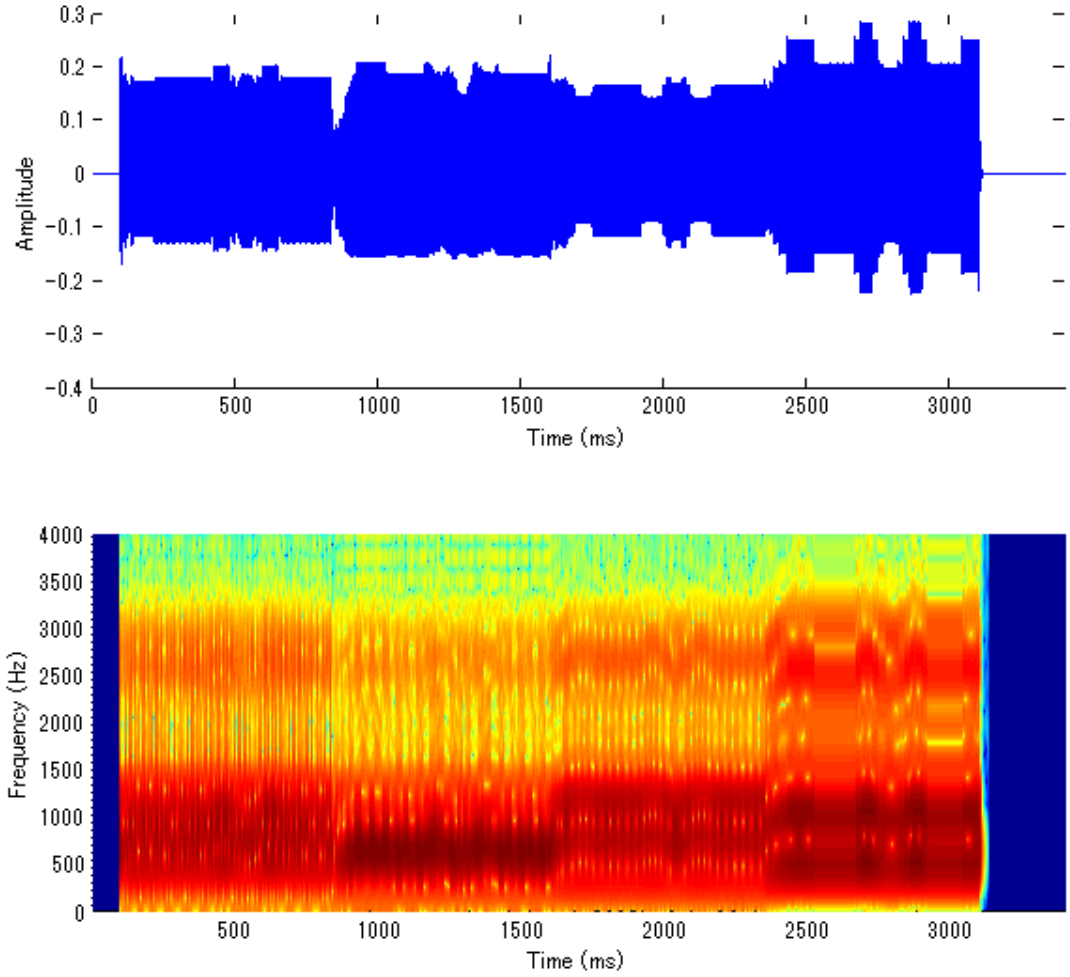

図 4.8: 音声 B の分析結果：（上）音声波形　（下）スペクトログラム

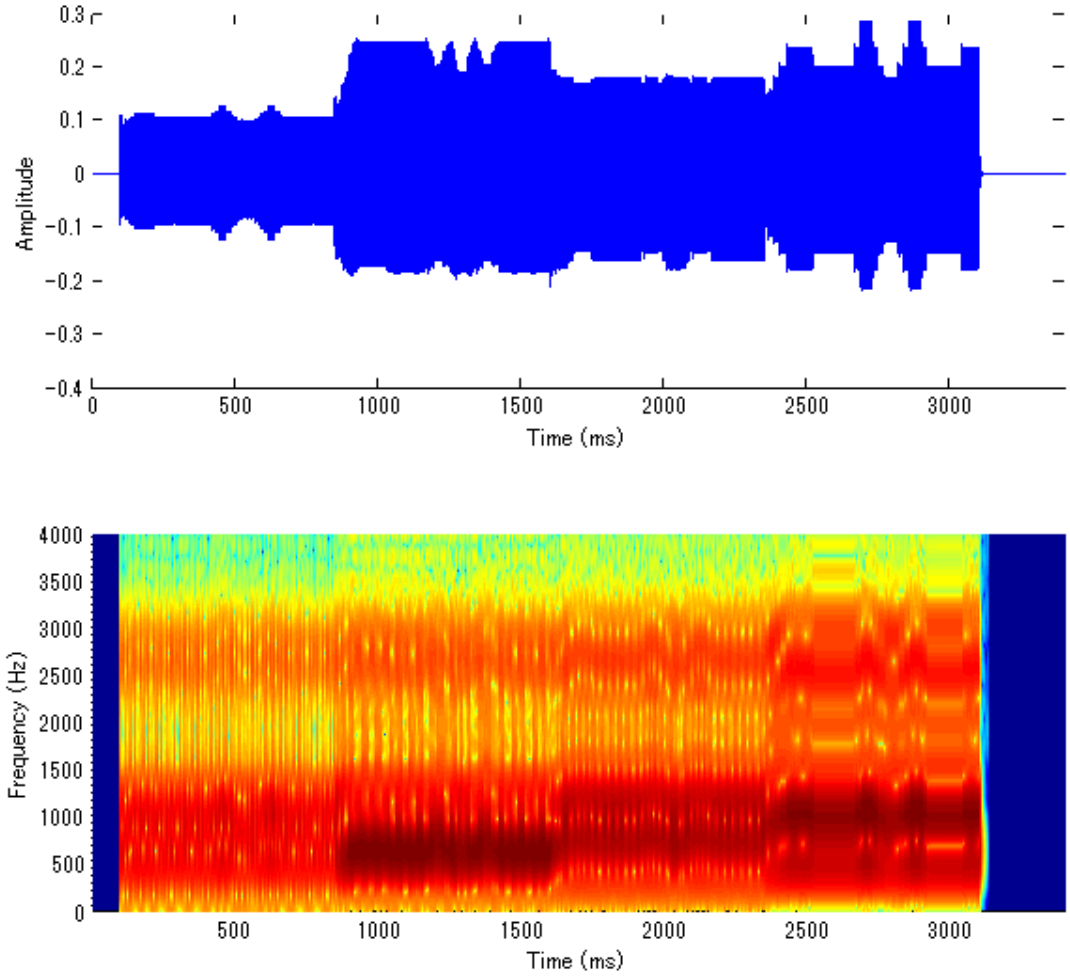

図 4.9: 音声 C の分析結果：（上）音声波形　（下）スペクトログラム

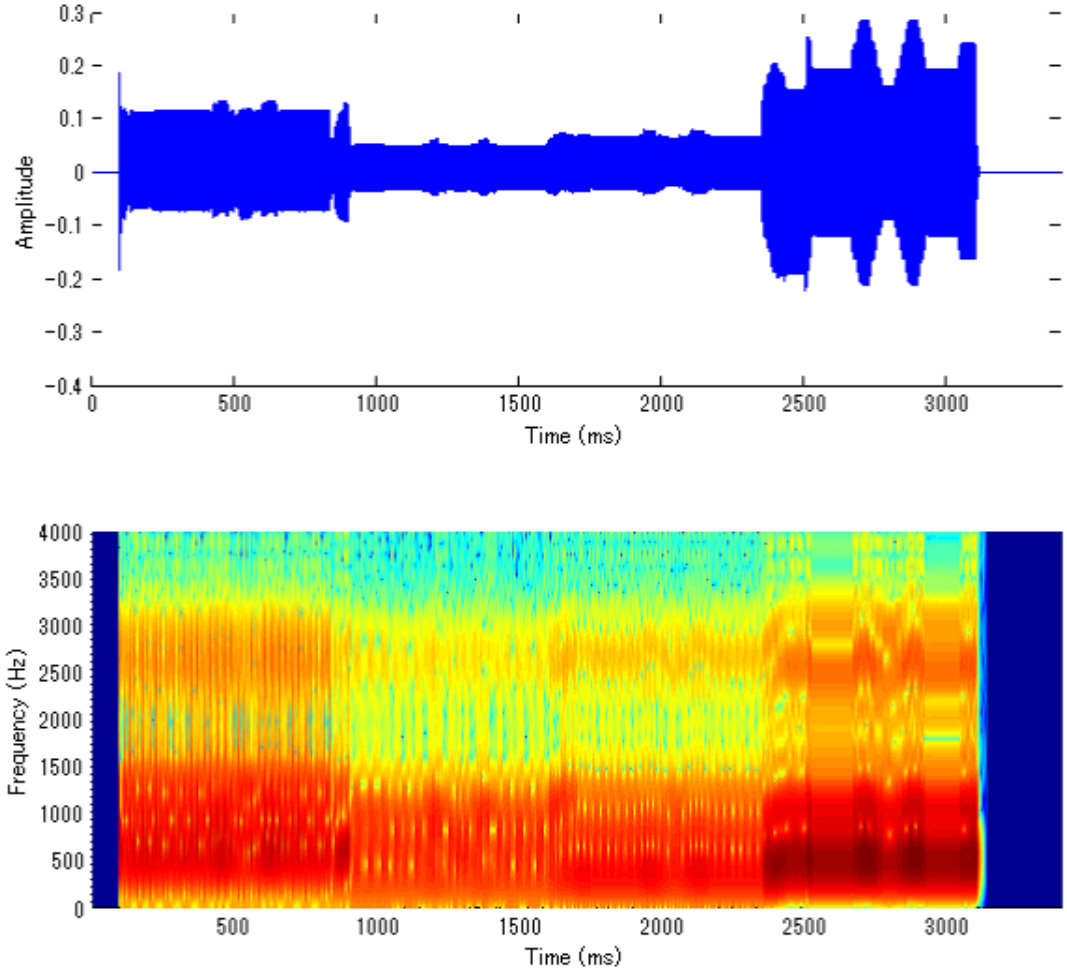

図 4.10: 音声 D の分析結果：（上）音声波形　（下）スペクトログラム

4.3 歌声合成音の制御

提案手法での歌声合成音と音声 C の振幅包絡を揃えるために，それぞれの振幅包絡線を求めた．提案手法の振幅包絡線を図 4.11 に，音声 C の振幅包絡線を図 4.12 に示す．

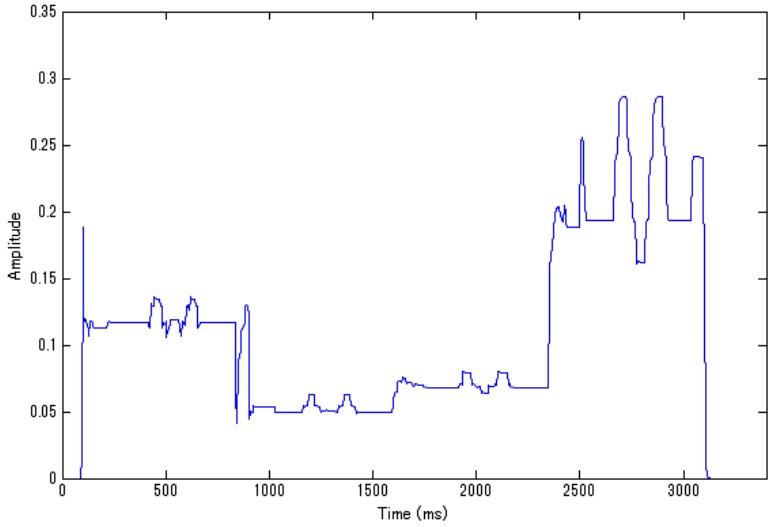

図 4.11: 提案手法による歌声合成音の振幅包絡線

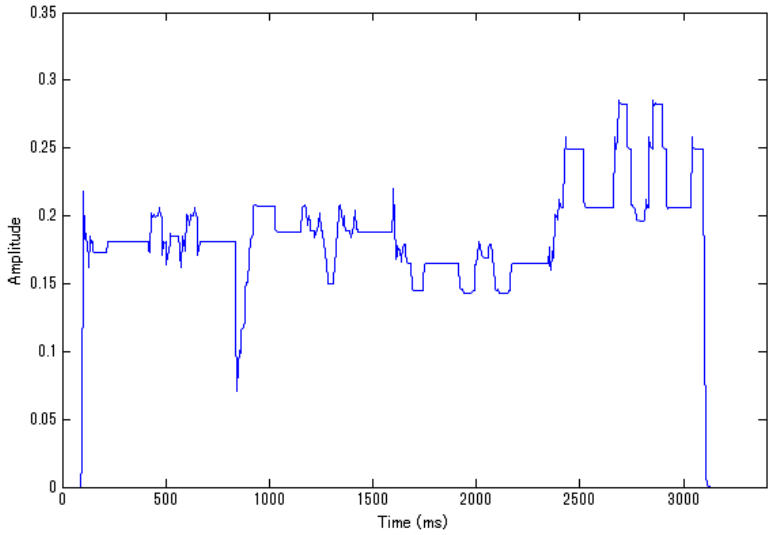

図 4.12: 音声 C の振幅包絡線

このとき，音声 C の振幅包絡線に合わせて提案手法の歌声合成音の振幅を制御した音声の振幅包絡線を図 4.13 に示す．提案手法に加えて振幅制御を行った音声を音声 E として，再度評価実験を行う．

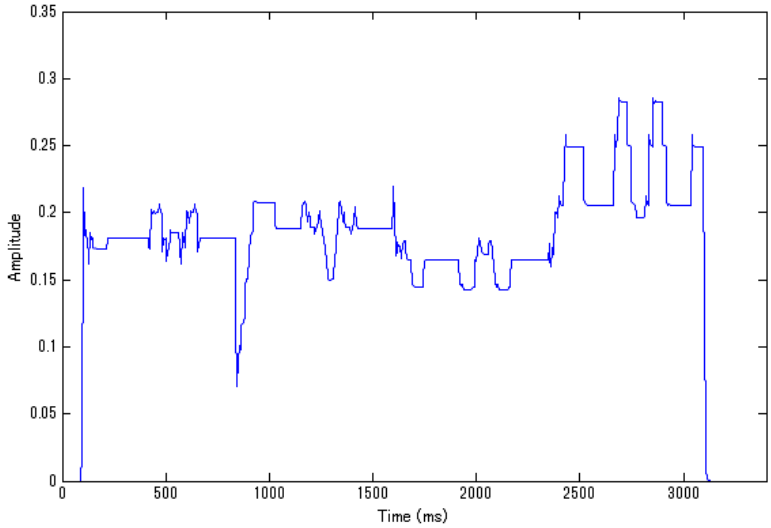

図 4.13: 音声 E の振幅包絡線

4.3.1 評価実験

評価実験として，4 章で使用した音声，音声 A，B，C，D に加えて，前節で音声 D に対して振幅操作を行った音声 E を加えた，また，収録した人の歌声のうち評価実験で合成する音高の音声を作成し，評価実験開始前に 3 回繰り返し呈示し，人の歌声を参考，人の歌声により近いものを選択するように評価基準を変更した．

4.3.2 評価結果

被験者 3 名に対して評価実験を行った．評価結果を表 4.14 に示す．評価結果より，振幅制御を行うことにより，人の歌声に近い自然な歌声に他の音声より近いと判断される割合は大きく上昇した．つまり，急激な振幅変化による相対的な自然さの低下が発生していたと考えられる．また，振幅制御を行った後の音声の相対的な自然さは先行研究による歌声合成音より自然であると評価された．

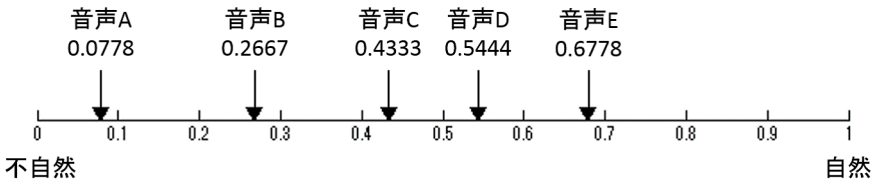

図 4.14: 評価結果

4.4 まとめ

4 章 1 節で行った実験において相対的に不自然であると評価された原因として，振幅の急激な変化をその原因の 1 つとして挙げた．そこで，提案手法による歌声合成音の振幅を最も自然であると評価された音声 C の振幅包絡線に合わせる制御を行った．そして，振幅制御を行った音声を音声 E として 4 章 1 節での評価実験の音声に加え再度評価実験を行った．その結果，振幅制御を行った音声は振幅制御を行う前の音声と比較して，他の音声より相対的に自然であると評価される割合が高くなった．また，振幅制御を行った音声は 4 章 1 節で最も自然であると評価された音声より自然であると評価された．

第5章　結論

本研究全体を通しての成果のまとめと本研究によって考えられる今後の課題を述べる．

5.1　本研究のまとめ

本研究では歌声を生成するときの1要素である声道の共振特性を制御することで自然な歌声を合成することを目標とした.．歌声合成システムに声道の共振特性を取り入れるためフォルマント制御モデルを構築した．フォルマント制御モデル構築のために，歌声のフォルマント分析を行った．その結果，F0 に合わせて大きく分けて4つの区間においてフォルマントの特徴的な移動が見られた．1つ目の区間は 300Hz 以下の区間である．この区間は modal で歌われている区間であり，F0 の変化に合わせて F1 はわずかに減少し，F2 はわずかに増加している．2つ目の区間は 300Hz から 336.64Hz の区間である．この区間は modal から falsetto への変化が発生している区間であり，F0 の変化に合わせて F1，F2 ともに急激に減少している．3つ目の区間は 336.64Hz から 454.64Hz の区間である．この区間は falsetto で歌われている区間のうち F0 が低い区間であり，F0 の変化に合わせて F1 はわずかに増加し F0 に近付く．F2 は減少して第二高調波に近付いている．最後の区間である4つ目の区間は 454.64Hz 以上の区間である．この区間は falsetto で歌われている区間のうち F0 が高い区間であり，F0 の変化に合わせて F1 は F0 に沿って増加し，F2 は第二高調波に沿って増加している．

この知見を元に4つの区間において異なる変化を見せるフォルマント制御モデルを構築した．構築したフォルマント制御モデルの有効性を評価するための評価実験の結果，1名にとっては提案手法による歌声合成音が最も自然であるという結果であったが，評価実験全体としては提案手法による歌声合成音は先行研究による歌声合成音と比較したとき相対的に不自然な歌声であると評価された．評価実験後に行ったアンケートの結果4音目が特に不自然であるという意見があったため，特に4音目に注目して評価実験に使用した歌声合成音を分析した．その結果，提案手法の歌声合成音のみ2音目，3音目の振幅が極端に小さく，4音目に変化するとき振幅が急激に大きくなっていた．

た，被験者の結果の半数である3名において地声から裏声への変化についての操作を行っていない歌声合成音に対して，地声から裏声に対して自然な変化が感じられているという結果があるため，被験者によって modal から falsetto へ変化するというものがどういった音の変化であるかという基準が異なっていることが考えられた．そこで，再度評価実験を行うに際して2つの変更を行った．1つ目は，振幅の制御である．提案手法での

歌声合成音と一番自然であると評価された音声 C に対して振幅包絡線を求めた．そして，音声 C の振幅包絡線に合わせて提案手法による歌声合成音の振幅を制御し，提案手法による歌声合成音の振幅を制御したものを音声 E として再度評価実験を行った．2 つ目は，評価基準の変更である．裏声と地声という主観的な評価ではなく，収録した音声により近いものが自然であるという基準に変更した．

その結果，提案手法による歌声合成音と音声 E がそれぞれ他の音声より自然であると評価される割合を比較したとき，音声 E の方が自然であると評価される割合が高くなった．また，振幅を制御した歌声は 4 章の実験で最も自然であると評価された音声より自然であると評価された．

5.2 今後の課題

本研究の結果，F0 の変化によって 4 つ区間での異なるフォルマントの変化が発生していることが判明した．また，評価実験の結果，声帯音源に対する操作に加えてフォルマント周波数の操作を行うことで，一部の人にとってはより自然であると感じられるという結果が得られたが，誰が聞いてもより自然であるという歌声合成音は得られなかったが，その原因として，音声の振幅包絡の影響があったということが分かった．今回の結果から，フォルマント周波数の操作による歌声の自然さへの良い影響はあるものの，フォルマント周波数の操作だけでは，高調波に近付くことでエネルギーが大きくなりすぎてしまい，相対的に不自然に聞こえてしまうということが判明した．

そこで，声道の共振特性を忠実に再現するためには，フォルマント周波数以外のパラメータについても操作を行っていく必要があると考える．また，今回のフォルマント制御モデルのみの制御ではフォルマントは階段状に変化しており，これは人の声道形状が瞬時に大きく変動する状態に相当してしまっているため，音高が変化する部分についての制御についても考える必要がある．そして，今回は男性 1 名の歌声を分析することによって得られた知見をもとにフォルマント制御モデルを構築したが，汎用性のある歌声合成システムを構築するためには，より多くの人の歌声に対して分析を行い，modal から falsetto への変化が行われる部分に共通点などが発見されれば，汎用性のある歌声合成システムを構築することが出来るのではないかと考える．また，STRAIGHT による分析を行うことにより得られた STRAIGHT spectrum を図 5.1，5.2，5.3 に示す．modal で歌われている歌声では LPC による分析と大きな違いは無いが，falsetto で歌われている歌声については，F1 と F2 が非常に近くなりスペクトル上に 1 つの大きなピークとして現れている．また音高が高くなっていくに従ってピークの中心周波数は増加し，帯域幅も広くなっていることが確認された．LPC による分析結果と STRAIGHT による分析結果の双方に共通して現れていることは，modal の音域において音高が高くなっていくと F1，F2 ともに増加する傾向があるということであるが，F1，F2 の変化については異なる挙動が見られた．STRAIGHT による分析結果に現れた変化についてもフォルマント制御モデルを構築し，操作の有効性を検討することでどの分析手法による分析結果がフォルマント変化を正確に

表しているかを検討する必要があると考える.

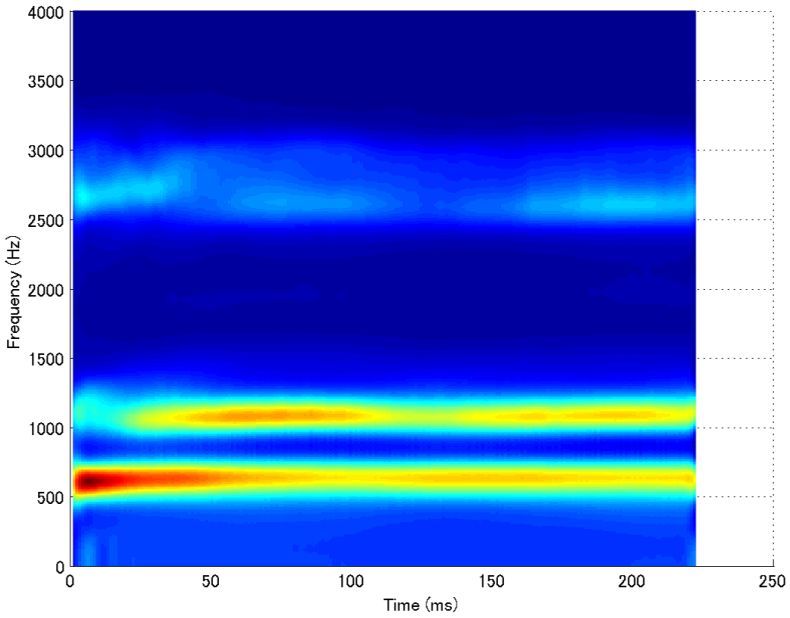

図 5.1: STRAIGHT spectrum (modal の歌声：F0 = 207.04 Hz)

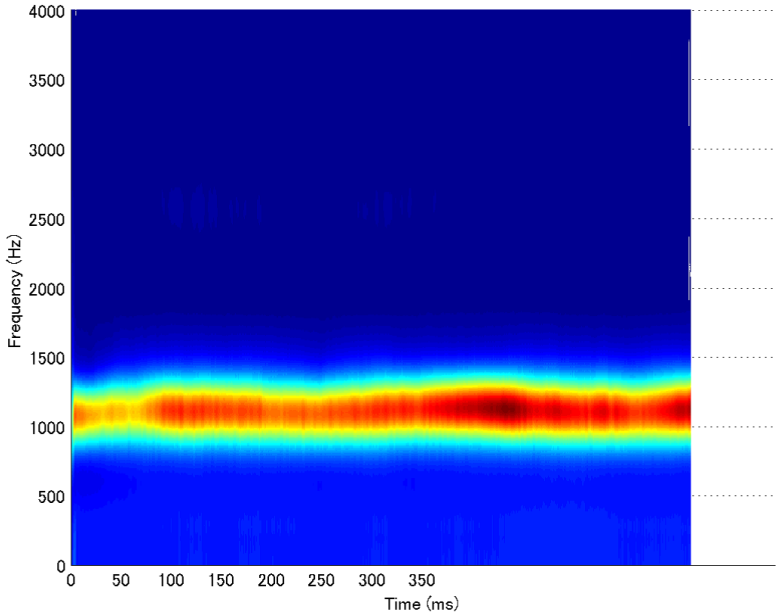

図 5.2: STRAIGHT spectrum (falsetto の歌声：F0 = 347.24 Hz)

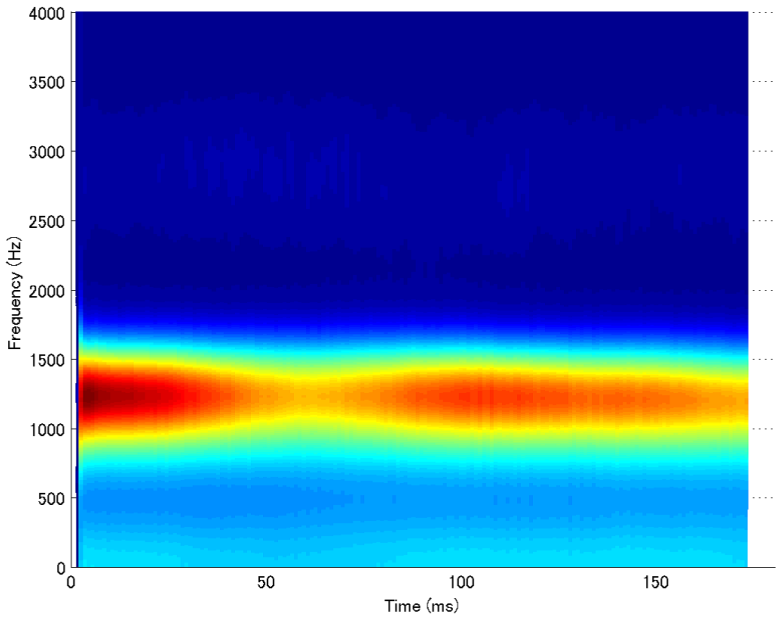

図 5.3: STRAIGHT spectrum (falsetto の歌声：F0 = 621.13 Hz)

謝辞

本研究を進めるにあたり，多大なるご指導ならびにご鞭撻を賜りました赤木正人教授に深く感謝の意を表します．

また，御有益な助言を賜りました党建武教授，ならびに鵜木祐史准教授に心より感謝致します．

ご多忙の中，評価実験にご参加頂いた方々に深く感謝致します．

そして，日ごろから多大なる討論と激励を頂きました赤木研究室，鵜木研究室の皆様に厚く御礼申し上げます．

最後に，大学院での貴重な研究生活を支えて頂いた，家族や友人に心から感謝致します．

参考文献

[1] 剣持秀紀, ”歌声合成システム VOCALOID の開発”, Institute of Systems, Control and Information Engineers, Vol.56, No.5, pp. 244-248, 2012.

[2] P. Depalle, G. Garcia and X. Rodet, “A virtual castrato”, Proc. ICMC ’94, pp. 357-360, 1994.

[3] 北村達也, 正木信夫, ”MRI 観測を基礎にした音声生成系研究の進展”, 日本音響学会誌 62 巻 5 号, pp.385-390, 2006.

[4] 北村達也, ”音声生成機構に基づく合成技術の動向”, 日本音響学会誌 67 巻 1 号, pp.28-33, 2011.

[5] 河合孝時, 鏑木時彦, ”音源・声道モデルを用いた音声分析法に関する検討”, TECHNICAL REPORT OF IEICE, SP2002-99, WIT2002-39, pp.31-36, 2002.

[6] Dennis H. Klatt, ”Software for a cascade / parallel formant syntesizer”, J. Acoust. Soc. Am. 67 (3), 1980.

[7] Johan Sundberg, ”The KTH Synthesis of Singing”, Advance in Cognitive Psychology, vol. 2, No. 2-3, pp.131-143, 2006.

[8] 酒向慎司, 宮島千代美, 徳田恵一, 北村正, ”隠れマルコフモデルに基づいた歌声合成システム”, IPSJ SIG Technical Report, 2003-MUS-51 (13), pp77-82, 2003.

[9] Johan Sundberg, ”Articulatory interpretation of the ”singing formant””, J. Acoust. Soc. Am. , Vol. 55, No. 4, pp.838-844,1974.

[10] Johan Sundberg, ”Formant technique in a professional female singer”, Acustica, Vol. 32, pp.89-96, 1975.

[11] 齋藤毅, 辻直也, 鵜木祐史, 赤木正人 “歌声らしさの知覚モデルに基づいた歌声特有の音響特徴量の分析”, 日本音響学会誌 64 巻 5 号, pp.267-277, 2008.

[12] 元田紘樹, “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”, 日本音響学会春季研究発表会, 2-7-13, 2013.

[13] Ding W., KasuyaH, Adachi S,“Simultaneous Estimation Basede on an ARX model”, IEICE TRANSACTIONS, vol.E78-D, vol.6, pp.738-743, 1995.

[14] 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 巻 7 号, pp.386-397, 2002.

[15] Fant G., Liljencrants J., Lin Q., “ A four-parameter model of glottal flow ”, STL-QPSR, vol.85, no.2, pp.1-13, 1985.

[16] Li Yougwei, Akagi Masato,“Glottal source analysis of emotional speech”, NCSP’14 , pp. 513-516, 2014.

[17] Hideki Kawahara, ”STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds”, Acoust. Sci. & Tech, 27, 6, pp.349-353, 2006.

[18] 赤木正人, 清水一郎, ”STRAIGHT を用いた話声からの歌声合成”, TECHNICAL REPORT OF IEICE, AP2003-37, pp.13-18, 2003.

[19] 辻直也, ”歌声らしさの要因とそれに関連する音響特徴量の検討”, 聴覚研究会資料, 34 (1), pp.41-46, 2004.

[20] 森下亮祐, 齋藤毅, 三好正人, ”歌声の地声-裏声変換のためのフォルマント操作の有効性”, IEICE Technical Report, EA2014-57, pp.83-87, 2014.