修 士 論 文

音声生成過程におけるフォルマント変換音声フィードバックの影響に関する研究

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

齋藤 和行

2004 年 3 月

修 士 論 文

音声生成過程におけるフォルマント変換音声フィードバックの影響に関する研究

指導教官　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 助教授
審査委員　宮原 誠 教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

210038 齋藤 和行

提出年月: 2004 年 2 月

概 要

雑踏のような騒音環境下での発話や電話の通話中に相手の声が聞き取りにくくなると通常より声が大きくなることがある．また，先天的難聴者は言語の獲得が困難であることなどから, 聴覚が発話動作の制御, 及び言語獲得において深く関与していると考えられる．しかしそのメカニズムに関しては未だ不明な点が多く残されている. 聴覚系と発話系の相互作用の存在を示すものとして, 聴覚フィードバックがある. 聴覚フィードバックは発話音声を聴覚系にフィードバックしながら発話動作の制御を行うものでこの機能により正常な発話が可能となっていると考えられている. 本研究では聴覚フィードバックが音声生成過程においてどのような役割を果たすかについて調べる.

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究の背景

音声の知覚・生成の相互作用として, 発話時における聴覚フィードバックの役割については, 古くから様々な研究がなされている.

雑音環境下では通常の発話より発話音声が大きくなり, 基本周波数も高くなる現象 (Lombard 効果) や [2] 発話音声が遅れて聴こえる条件 (遅延聴覚フィードバック：Delayed Auditory Feedback:DAF) では, 吃音や発話速度が遅くなる等の現象が生じることは聴覚フィードバックの効果を表す顕著な例である.[1]

これらの知見は聴覚フィードバックは発話にとって重要な役割を果たすことを示す証拠として考えられてきた. しかし, これらの報告は定性的な性質を述べるにとどまっており, これらの現象がどのようなメカニズムで生じるかについて説明するには不十分である.

また DAF のような発話動作を破壊するような実験では定性的な性質は観察できるが, 発話動作自体が破綻してしまうのでメカニズムを分析するうえで必要な定量的な分析を行うことが難しい.

そこで実時間により音響パラメータを変換するような, 定量的な分析を可能にする非破壊的な実験パラダイムの構築が必要となってくる.

また発話の全体像を理解するためには裏付けられた理論的枠組が必要である.

近年,MRI(核磁気共鳴画像) 等の観測技術発展に伴う神経回路や運動制御に関する多くの発見があったことや, 信号処理技術や計算機の処理能力の爆発的な向上よる実時間の音響パラメータの変換が可能になってきたことから, 音声の知覚・生成の相互作用について調べる上での要件を満たしつつあるといえる.

1.2　本研究の目的

発声・発話動作における聴覚フィードバックの影響に関して基本周波数に関しては定量的な分析が進められているが, 発話動作のフォルマント制御においては充分に調べられているとはいえない. またフォルマントの制御に聴覚フィードバックが用いられているかについても直接的な証拠に欠けているのが現状である.

本研究では発話動作における音声生成・知覚の相互作用に関する問題を運動制御モデルとして扱い, 入力を聴覚刺激, 出力を発話動作とした場合の発話動作のフォルマント制御モデルを構築するための知見を獲得することを目的とする.

1.3　本論文の構成

本論文は 6 章により構成される. 第 1 章では研究の背景等の本論文に関する導入部にあたる. 第 2 章は運動制御モデル等に基づいて聴覚フィードバックに関する挙動を予想し, それに応じて設定した実験パラダイムに関する記述である. 第 3 章は 2 章の実験パラダイムに基づいて行った実験の方法に関する記述である. 第 4 章は予備実験に関する記述である. 第 5 章は本実験に関する記述及び実験による結果の分析方法, 分析結果, そしてそれらについての考察について述べた. 第 6 章は結論として本論文のまとめと今回の実験の問題点や今後行われるべき内容について述べた

第2章　実験パラダイム

2.1　発話動作と運動制御モデル

2.1.1　運動制御モデル

運動制御モデルでは生体をシステムの一種として扱い, 自動制御理論を導入して生体の運動メカニズムを説明する方法が試みられている.

自動制御理論ではシステムの行動を入力と出力の視点からとらえ, フィードバックを利用するか, しないかにより 2 つの制御方式,「フィードバック制御」と「フィードフォワード制御」に分類する.

これらの制御方式はそれぞれ利点と欠点をもっている. 生体の運動メカニズムはどちらか一方のみで説明できるというものではなく, 一般的に 2 つの制御方式の組み合わせで説明される.

2.1.2　フィードフォワード制御モデル (中枢説)

フィードフォワード制御はフィードバックループを持たない制御方式で, システムは制御量の検出を待たずに外乱による制御量の変化を予測して補正反応を起こす.(図 2.1)

このためシステム特有の時間遅れはあまり生じず, フィードバック制御よりもずっと短期間に反応できる. しかし外乱が一定の規則に従わず予測が困難な状況下では正しい制御が困難である.

生体における運動調節モデルにおいてフィードフォワード制御は「中枢説」と呼ばれている. これは脳が運動パターンを決定する情報をあらかじめもっており, 末梢感覚の助けなしに運動命令 (中枢プログラム) を構成して動作を生成するというものである.

2.1.3　フィードバック制御モデル (末梢説)

フィードバック制御はフィードバックループを持つ制御方式で, 制御量に関する情報を検出部を通じて取得して調節部に戻すことで目標値とその誤差を比較し偏差を零にするように制御する. (図 2.2)

これによりシステムに外乱が加わった場合でも補正して最終的には目標値に辿り着く. しかし制御量の検出から補正反応までの間にはシステム特有の時間遅れが生じる. そのた

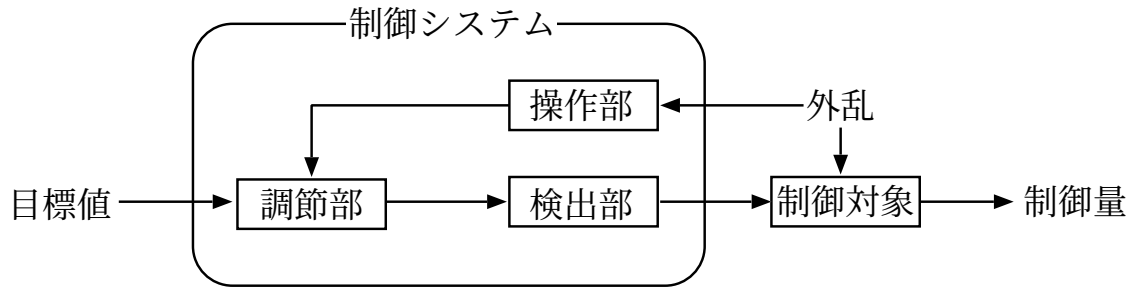


図 2.1: フィードフォワード制御 (文献 [8] より引用)

め外乱がこの時間遅れより短時間で急激に変化する場合, 補正反応が間に合わずかえってエラーを増大させることになる.

生体における運動調節モデルにおいてフィードフォワード制御は「末梢説」と呼ばれている. これは運動を生成するには運動器官や感覚器官のような末梢からの知覚情報が不可欠であるというものである.

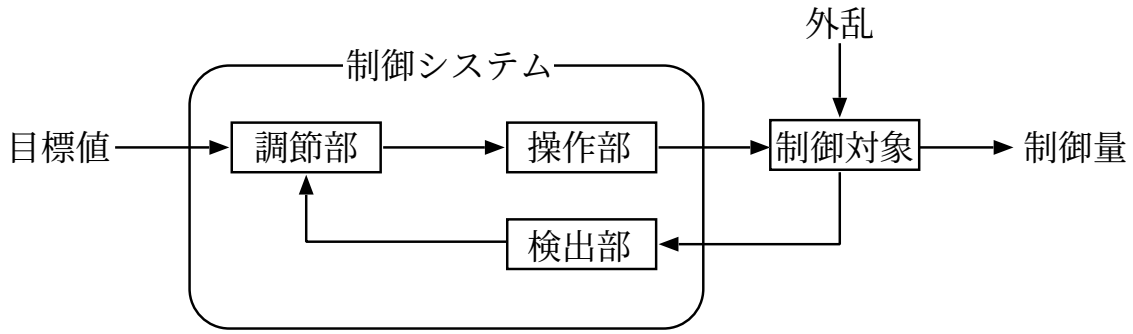


図 2.2: フィードバック制御 (文献 [8] より引用)

2.1.4 フィードバック誤差学習モデル

フィードバック誤差学習は, 生体の随意運動学習のモデルとして川人らにより提案されたモデルである.[11]

この方法は、制御に用いる基本的なフィードバック制御器の出力を学習のための誤差とみなし、その誤差を減らすように適応制御器を調節することにより制御性を改善する.

このモデルでは始めはフィードバック制御器のみで制御が行われているため, 実際の軌道は目的軌道に追従できず, また動きもぎこちなくなってしまう. しかし, 学習の進行とともにフィードバック制御器の出力を小さくするようにフィードフォワード制御器に “逆モデル” を中枢プログラムとして, 徐々に獲得し, 運動を目標軌道に近づける.(図 2.3)

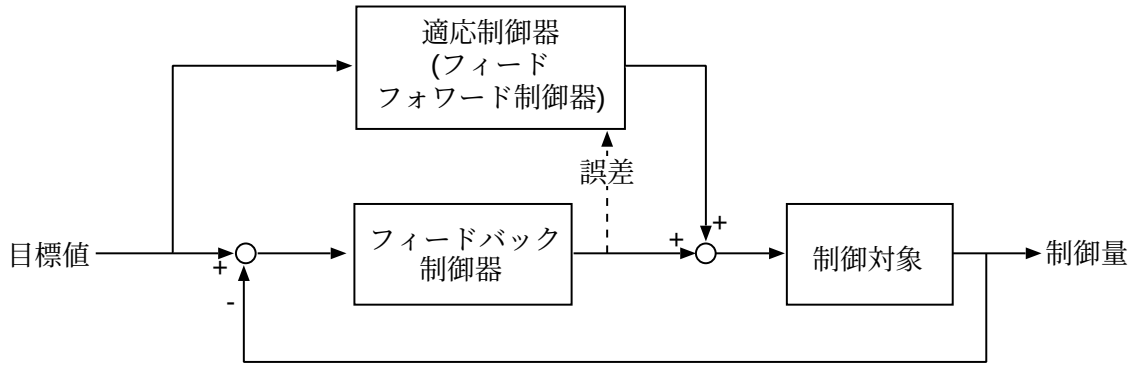


図 2.3: フィードバック誤差学習モデル (川人ら [11] による)

2.1.5 発声の基本周波数制御モデル

入力を聴覚刺激, 出力を発声・発話動作とした場合の制御モデルについての定量的に扱った研究として, 河原らの基本周波数制御と聴覚フィードバックに関する一連の研究がある.[14],[13],[12]

発声の基本周波数は、声門下圧や喉頭筋の緊張状態によって変化する不安定な系であるため, 基本周波数を一定に保つには何らかのフィードバックを用いて制御する必要がある.

一方, 発話時のアクセントやイントネーションのような急激な基本周波数の変化はフィードバックによる制御では間に合わないため, フィードフォワードによる制御も必要となってくる.

そこで, 河原らはこれらの動作を説明するため, フィードバックループを含む系と含まない系を用いて, それらが並列に動作する 2 次のむだ時間遅れをもつ線形システムで近似した.

また,TAF(Transformed Auditory Feedback:変換聴覚フィードバック) と呼ばれる実験手法に基づいた測定により聴覚フィードバックによる応答として速い応答と遅い応答の 2 種類の応答を確認している.

このうち速い応答はフィードバックループを含まないフィードフォワード制御, 遅い応答はフィードバックループを含むフィードバック制御であるとしている．図 2.4 は河原らによる運動の計算モデルに基づく発声の基本周波数制御の機能モデルである. 図中の A は聴覚系,P は発声系を表し,R は (小脳を介した) 反射系,C は (大脳を介した) 調整系を表している.

このうち A,C,P を含むループは遅い応答 (フィードバック) に対応し,A,R,P を含むループは速い応答 (フィードフォワード) に対応する.

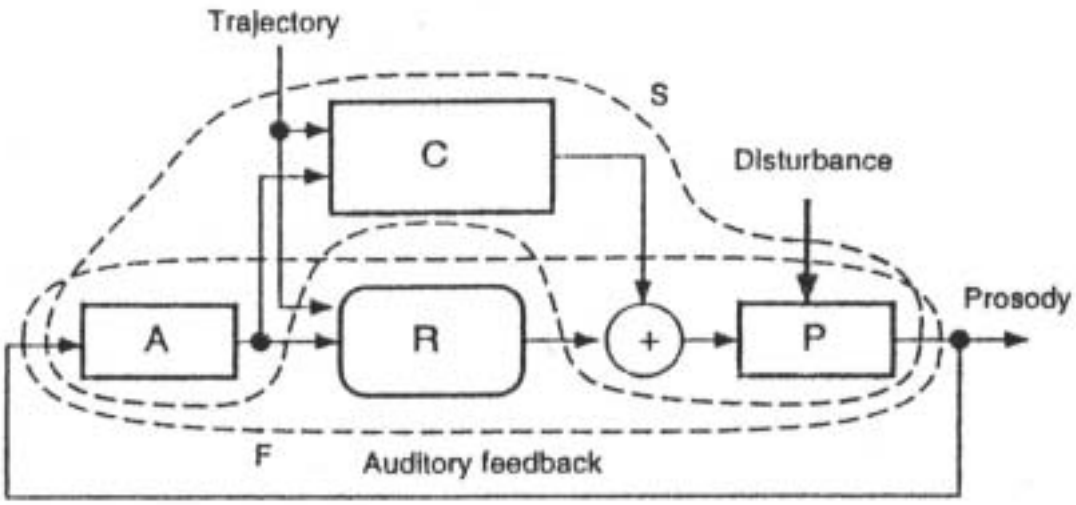


図 2.4: 運動の計算モデルに基づく発声の基本周波数の機能モデル (河原ら [14] による)

2.1.6 運動制御モデルと発話のフォルマント制御モデル仮説

発声の基本周波数制御がフィードバック制御とフィードフォワード制御の二つの制御モデルで説明が可能なことは既に述べた. 一方, 発話動作における音韻情報生成のようなスペクトル構造 (フォルマント) の制御においてモデル化された例は報告されていない.

ここでは過去に行われた研究をフィードフォワード制御による特徴を表しているもの, フィードバック制御による特徴を表しているもの, フィードバック誤差学習による特徴を表しているものの 3 つに分類し, 発話動作におけるフォルマントの制御について考察し, その推測を試みる.

まず発話動作においてフィードフォワード制御による特徴がよく表われているものとして,Lane らの Lombard 効果に関する研究がある.[2] これは雑音によって発話音声をマスクした環境下でも発話が可能であることを示している. また Lane らは言語習得後に聴力を失った後天性難聴者においても明瞭な発話が維持されることも報告されている.[3] これらの知見は共通して聴覚 (末梢) による情報がなくても発話 (運動の生成) が可能であることを示している.

次にフィードバック制御による特徴が表われているものとしてまず Lee らの DAF の実験が挙げられる.[1] これは発話音声が遅れて知覚されると発話に支障をきたすというもので, 発話時に自分の発した声を聞き取りながら話していることの証拠となっている. さらに音響情報を除去する DAF の実験では音響情報として第 1 フォルマント (以下 F1), 第 2 フォルマント (以下 F2) を除去した場合, DAF の影響が減少したことが報告されている.[16] これは F1,F2 が除去された音声はフィードバック音声としての効果が薄れた, 言い換えれば F1,F2 がフィードバック音声として重要で, 発話音声の F1,F2 を聞き取りながら発話して

いる可能性を示している. これらの知見は発話 (運動の生成) において聴覚 (末梢) からの情報が不可欠であることを示すものではないが, 少なくとも発話時に自分の発話音声を聞き取りながら発話し, なおかつフォルマントのようなスペクトル構造を参照していることを示している.

最後にフィードバック誤差学習による特徴が表われているものとして, Houde らの変換聴覚フィードバック環境下における発話動作の適応に関する実験がある. [4] 実験では発話音声のフォルマントを変換した音声がフィードバックされる, 変換聴覚フィードバック環境下で約 1 時間の発話が行われた. その結果として, 発話音声が変化し, 聴覚の入力が遮断されても変化が保持されたことや被験者によってはその効果が 1ヵ月保持されたということが報告されている. これは発話環境の変化に対して新たな “逆モデル” を獲得し, 聴覚 (末梢) からの情報が遮断された場合は “逆モデル” により発話しているものと思われる.

また Houde らはこの実験で, 聴覚入力が遮断されたときに比べて変換聴覚フィードバック環境下における発話の変化が大きくなることから, 学習のほかにフィードバックによる制御を行い, さらにその応答が発声の基本周波数制御と同様な補償動作である可能性を示唆している.

以上のことから聴覚と発話動作に関して次のようなことが予想される.

仮説 (a) 言語習得過程では発話における “逆モデル” を獲得するため, 発話時には主にフィードバック制御 (聴覚フィードバック) を用いられている.

仮説 (b) 言語習得後は滑らかな発話運動を実現するため, 発話時に主にフィードフォワード (“逆モデル” を用いた予測) を利用している.

仮説 (c) 発話者は発話時に聴覚フィードバックによる情報として, スペクトル構造に関する情報も利用している.

仮説 (d) 変換されたフィードバック音声に対して補償方向の応答が現れる.

2.2 観測対象と実験アプローチ

前述の仮説のうち, 特に (c),(d) は実験による直接的な証拠に乏しい. そこで本研究では発話時に聴覚フィードバックが発話動作のフォルマント制御に与える影響, すなわち発話動作の音声生成過程における聴覚フィードバックの役割について調べる.

フィードバック経路の存在を示すにはフィードバック経路を遮断するか外乱を与えることで発話動作に表われる変化を観察することにより可能である.

本研究ではフィードバック経路に外乱を与えることにより, 発話動作に現れる変化の観察を行う.

また過去の研究の多くでパラメータの変換による応答が音声により確認されたことや, 確立された分析手法が数多く存在し, 比較的扱いやすいという理由から音声を観測の対象とした.

2.3 実験要件

実験を行う上で以下の要件が満たされる必要がある.

要件 1 実時間による音声のパラメータ変換を行う.

要件 2 発話者の音響物理量をできるだけ多く残した自然性の高い変換を行う.

要件 3 変化に対して被験者が修正可能な摂動を与える.

要件 4 被験者が変化を知覚できる摂動をあたえる.

1つ目はフィードバック音声の変換は実時間処理により行われる必要があるというものである. その理由はフィードバック音声に基本周波数や振幅のような時間変化する音響物理量が保持されていることが重要であるためである. 例えば, フィードバック音声として予め被験者の音声を録音したものや合成したものを実験に用いたとすると, 発話時の時々刻々と変化する基本周波数や振幅包絡の時間情報等がフィードバック音声に反映されないという問題が生じる. 沢田らの報告 [16] によれば, 基本周波数や振幅包絡の時間情報もフィードバック音声として重要な音響物理量であることから, これらの情報の損失はフィードバック音声として不適切となる可能性がある. つまり基本周波数の時間変化や振幅包絡の時間情報等を保持したまま, フォルマントのみが変換されなければならない. さらにフィードバック音声の遅延は発話動作の破壊など実験に望ましくない影響を与える可能性があるため, 遅延を最小限に抑えることも重要である.

2つ目は自分の声とそれ以外の音声について弁別する能力が高い (自然側音により発話が妨害されないことや, 多くの情報が失われたフィードバック音声が発話に影響しないことと一致する) ため, 音響物理量あるいは, 自然性が損なわれることで実験に望ましくない影響が表われる可能性があるためである. そのためフィードバック音声としては被験者の音声にできるだけ近い音声を利用することが望ましい. また聴覚フィードバックの実験で用いるフィードバック音声の自然性の重要さについては Shimon ら [7] も指摘している.

3つ目の要件は摂動として与える変化が被験者により修正可能であることである. 発話のフォルマント制御モデルの仮説のうち (d) に直接対応するものである. もし摂動に対する応答が仮説の通り, 補償動作であった場合, 被験者の発話動作は発話の変化に対して元に戻そう働くことになる. このとき使用する摂動あるいは音韻, 音節によっては補償の方向が発話機構などの物理的に制限されてしまうことや, 言語習得時に獲得したモデル内に補償方向に該当する情報が存在しないことにより, 応答が充分に現れない可能性がある. また発話動作が破綻するような摂動も応答の測定にはふさわしくない. そのためには, これらの問題を避けるような摂動, および対象とする音韻, 音節の選択が重要である.

4つめは, 摂動が被験者によって知覚が可能なことである. 被験者に摂動が知覚されなければ当然, 発話動作での応答は確認できない. ここで注意すべきことは観察すべき応答に応じた摂動の選択である. つまり所望の応答が摂動に対する随意運動のような意識レベルによる応答であるか, 反射運動のような自動レベルによる応答に摂動を変える必要があ

る. 例えば, 自動レベルの応答に注意を払うとき, 摂動は被験者に意識レベル以上に知覚させることは望ましくない. またその逆も同様である. さらに検知感度などについても考察する必要がある. 例えばフォルマントの変化に対する検知感度は基本周波数のそれほど高くなく, 基本周波数に比べて系の制御が安定していることからも, 基本周波数に比べて与える摂動の大きさを上げる必要があるかもしれない. このように変換するパラメータに対応した摂動を与える必要がある.

第3章　方法

3.1　被験者

実験は正常な発話･聴覚能力を有する24〜30歳までの日本語話者の大学院生3名,AH(30),MO(25),TH(25)を被験者として行った. なお, 被験者は全て男性である.

3.2　装置

図3.1に実験で使用する実験装置の概要を示す. 実験は防音室内 (暗騒音 約 35dB(SPL)) で行う. 被験者により発話された音声はリアルタイム OS である RT-Linux 上のプログラムにより実時間で変換が行われ, 被験者にフィードバックされる.

また, 骨導音や自然側音をマスクするため, フィードバック音声には60dB程度のピンクノイズを付加する.

発話音声, 及び変換されたフィードバック音声はそれぞれ記録用の計算機に記録され, 分析に用いる.

被験者はヘッドホン (HDA-200) とマイクロホン (WM-C70) を身に着けた状態で発話を行う.

発話された音声はマイクロホン, マイクロホンアンプ (MA-8) を経て, 一方は計算機内の AD 変換ボード (PCI-3155) に入力され, もう一方は記録用の計算機に接続された AD 変換器 (DF-2021) を経て記録用の計算機に記録される. 計算機内の AD 変換ボードに入力された音声は実時間処理によりフォルマントの変換処理が行われ,DA 変換ボード (PCI-3336) を通じてへ外部に出力される.

DA 変換ボードから出力された音声は一方はミキサ (AT-MX50) によりピンクノイズを付加され, 防音室内のアンプ (AU- α 907MR) を経て被験者にフィードバックされる.

もう一方は AD 変換器を経て記録用の計算機に記録される.

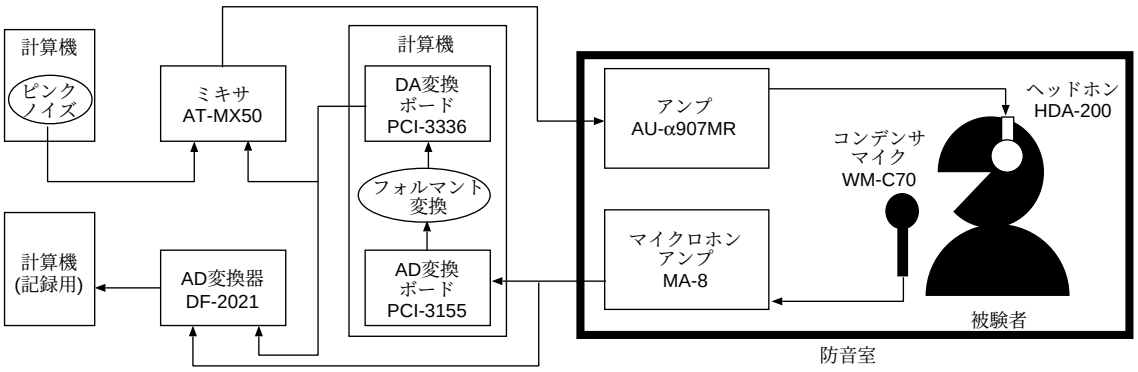


図 3.1: 実験装置

3.3 刺激

3.3.1 フォルマント変換処理

フォルマント変換には実時間による処理が必要であるという要請と高品質で自然性が高い音声が必要であるという要請を満たす必要がある. 一般にこの 2 つを同時に満たすことは困難である. ボコーダのような分析・合成系による方法では処理にかかる負荷が大きく実現が難しい, また群遅延等の問題で自然性が損なわれる危険性を含んでいる. そのためフォルマント変換処理は処理が簡易なフィルタ処理のみで実現する. フィルタ処理は FFT を用いた重複加算法による短時間合成を用いた畳み込みにより実現する. 図 3.2 はフォルマント変換処理の概要を表している.

入力音声 $x(n)$ は窓関数 $w(n)$ により切り出され, フーリエ変換により周波数表現 $X(k)$ を得る. 図中の右側のパスではフォルマント分析が行われ, 分析で得られた情報を基にフォルマントフィルタ $H(k)$ が生成される.

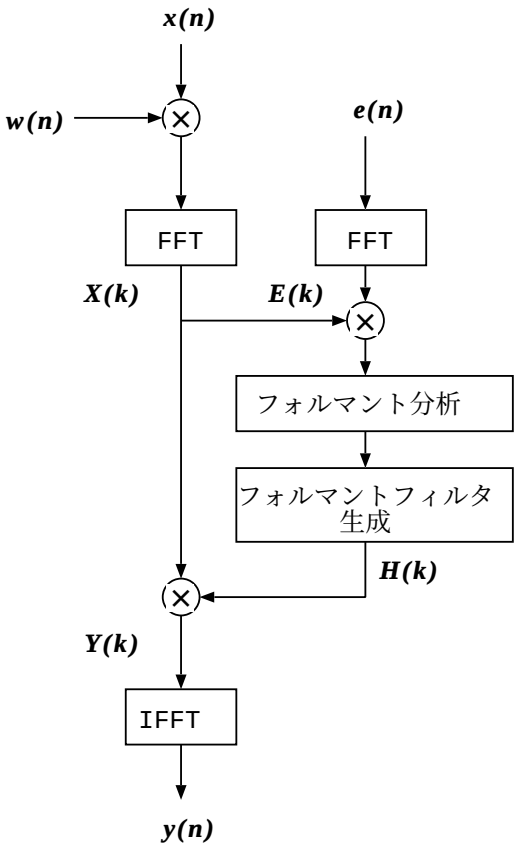
そして $X(k)$ に対してフィルタ処理がなされ, フィルタ出力 $Y(k)$ を逆フーリエ変換することでフィードバック音声 $y(n)$ が得られる.

なお, フォルマント変換による計算時間は $125\mu s$ 以下, アルゴリズムによる不可避の遅延として窓長の半分である 8ms,AD/DA 変換にかかる時間数十 μs 以下であるため. 全体としての遅延時間は 9ms 未満である. これは佐藤が実験により得た, 遅延が 20ms 以下であれば, 発話に影響しないという基準を満たしている [15]

3.3.2 重複加算法 (OLA 法) による短時間合成

重複加算法 (OverLap Addigtion method:OLA 法) は時間軸上で窓関数が重複するように一定の間隔シフトさせてながら分析, 変型を行い最後に足し合わせることで再合成を行う手法である.

入力信号 $x(n)$ の短時間スペクトル $X(e^{j\omega})$ を時間軸上で R サンプルの周期で標本化し



x(n)：発話音声
e(n)：高域強調フィルタのインパルス応答
E(n)：e(n)のフーリエ変換
w(n)：窓関数
X(k)：x(n)の短時間フーリエ変換
H(k)：フォルマントフィルタの伝達関数
Y(k)：y(n)の短時間フーリエ変換
y(n)：フィードバック音声

図 3.2: フォルマント変換処理の概要

たものを考える.

すなわち,$Y_r(e^{j\omega_k}) = X_{rR}(e^{j\omega_k})$, ここで r は整数,k は $0 \leq k \leq N-1$ である. 重複加算法は次のような式に基づいている.

$$y(n) = \sum_{r=-\infty}^{\infty} [\frac{1}{N} \sum_{k=0}^{N-1} Y(e^{j\omega_k}) e^{j\omega_k n}] \tag{3.1}$$

すなわち, 信号を再生するには,$Y_r(e^{j\omega_k})$ の逆変換を r の各値について計算して, 次の数列を求める.

$$y_r(n) = x(m)w(rR - m) \quad (-\infty < m < \infty) \tag{3.2}$$

そして時刻 n の点で重複している数列 $y_r(m)$ のすべての n における値を足し合わせれば時刻 n の信号が得られる. すなわち

$$y(n) = \sum_{r=-\infty}^{\infty} y_r = x(n) \sum_{r=-\infty}^{\infty} w(rR - n) \tag{3.3}$$

である. 右辺の和をとる項は,$w(n)$ が周波数帯域制限のフーリエ変換をもち, そして $X_n(e^{j\omega_k})$ の時間軸での標本化が適切ならば, すなわち R が十分小さくて ($R \leq L/4$) 時間の重複がなければ,n の値に無関係に

$$y(n) = \sum_{r=-\infty}^{\infty} w(rR - n) \approx W(e^{j0})/R \tag{3.4}$$

であることが示されるので式 (3.3) は

$$y(n) = x(n)W(e^{j0})/R \tag{3.5}$$

となる. すなわち, 式 (3.1) の合成規則によって波形の重複部分を重ね合わせたものは $x(n)$ の厳密な再生になっている.

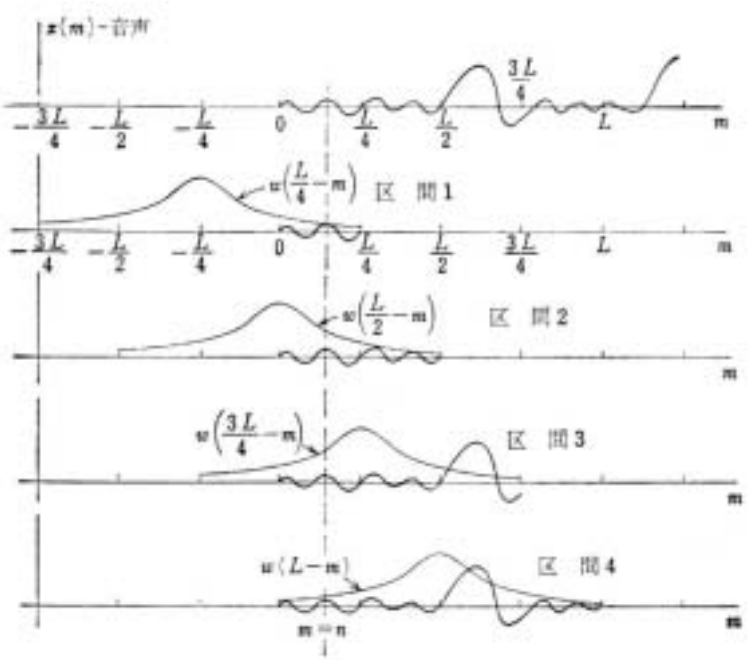


図 3.3: OLA 法 (文献 [6] より引用)

3.3.3 フォルマント分析

フォルマント分析では短時間フーリエスペクトルを用いたピーク検出により, フォルマント周波数を推定する.

表 3.1: 分析条件

標本化周波数	8kHz
窓長	16ms(128 点)
窓関数	hamming 窓
F1 の範囲	400 〜 900Hz
F2 の範囲	1000 〜 2500Hz
高域強調フィルタ係数 $(1-\alpha z^{-1})$	$\alpha = 0.9$

フォルマント分析に LPC 分析による方法や準同型分析による方法 [5] 等, 様々な方法が提案されているが, 実験には実時間での信号処理が必要なため, 計算負荷ができるだけ小さいことが望まれる.

そこで本研究では計算が簡易な短時間フーリエスペクトルによる分析を行った. 一般にフーリエスペクトルを有声音の分析に適用する場合, 声道形状の特徴の他に声門パルス波による調波構造が含まれる.

調波構造はときにフォルマント推定に望ましくない影響を与えることがある. これに対しては窓長を短くし, 周波数分解能を下げることで調波構造の発生を低減した. また窓長を短くすることは計算時間や窓長による遅延の短縮にも繋がるため実時間処理には有利であるといえる.

ピーク検出には標準的なフォルマントの範囲を設定し, その範囲内でパワーの最大値を与える周波数インデックス f_1,f_2 を求め, これをフォルマント周波数とした.

本研究では表 3.1 を分析条件とした.

入力信号 $x(n)$ の振幅スペクトル $|X(k)|$ は短時間フーリエ変換

$$X(k) = \sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2\pi k}{N}} \tag{3.6}$$

を用いて

$$|X(k)| = \sqrt{Re[X(k)]^2 + Im[X(k)]^2} \tag{3.7}$$

で得られる. ここで分析条件の F1 の範囲 (表 3.1) で

$$f_1 = max[|X(k)|] \tag{3.8}$$

となる周波数インデックス k を F1(f_1) とする. 同様に F1 の範囲で

$$f_2 = max[|X(k)|] \tag{3.9}$$

となる周波数インデックス k を F2(f_2) とする.

3.3.4 フォルマントフィルタ

フォルマント分析により得られたパラメータに基づき, フォルマント形状を Gauss 関数により近似を行うものとする.

$$G(\omega) = A\exp(-\frac{(\omega - f)^2}{2B^2}) \tag{3.10}$$

f,A,B はそれぞれフィルタの中心周波数 (Hz), 利得 (dB), 帯域幅 (Hz) を表す. 伝達関数の振幅特性は次のようになる.

$$|H(\omega)| = 10^{G(\omega)/20} \tag{3.11}$$

このフォルマントフィルタは 1 つのフォルマントに対して 2 つのフィルタが対応する. すなわち発話音声のフォルマント除去を行うフォルマント逆フィルタ $H(\omega)^{-1}$ とフォルマントの追加を行うフォルマントフィルタ $H(\omega)^{-1}$ である. これらは互いに

$$H(\omega)H(\omega)^{-1} = 1 \tag{3.12}$$

となる性質を持つ.

3.3.5 音声データを用いたシミュレーション

計算機上でのシミュレーションを行った. 音声データには本研究で用いる実験環境で録音された音声を使用した.

サンプル音声は男性話者の単母音/e/と連続母音/aeae . . . /とした. シミュレーションでは発話音声の母音/e/の部分を/a/に変換した.

図 3.4,3.5 はそれぞれのシミュレーションによって得られた音声のサウンドスペクトログラムである.

変換音声の音韻の変化, 音色について主観評価を行ったところ, 特に問題がないことを確認した.

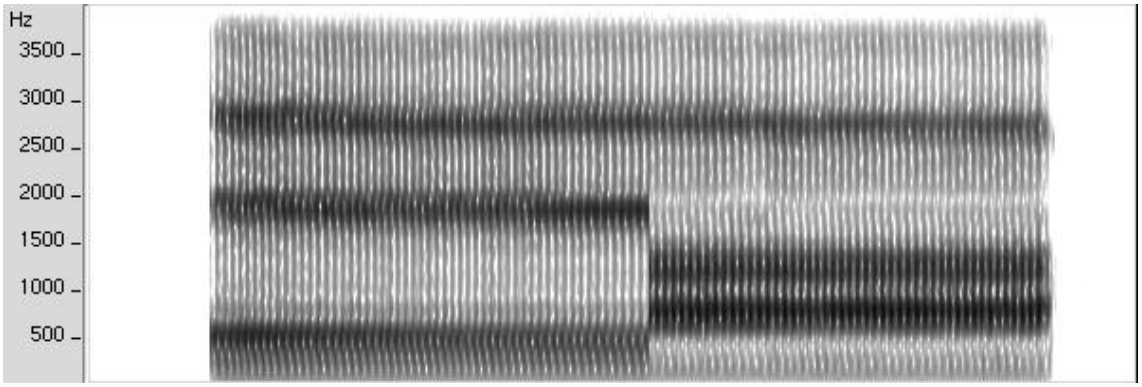


図 3.4: 単母音/e/のシミュレーション結果

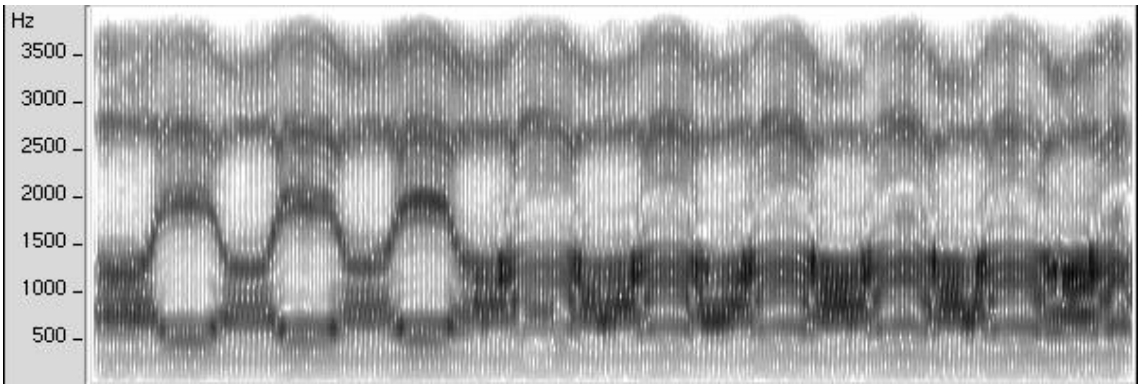


図 3.5: 連続母音/aeae . . . /のシミュレーション結果

第4章　予備実験

4.1　目的

- 被験者のフォルマントに関する情報を取得する
- 被験者を実験環境に慣れさせる
- 実験環境における違和感の有無を調査する

4.2　実験手順

実験では発話音声を連続母音/eaea . . . /とした.

実験環境に対する違和感や被験者を実験環境に慣れさせることも目的として含んでいるため, 摂動を与えることを除いて本実験に近い条件で行った.

発話の開始, 及び終了は装着されたヘッドフォンを通じて発話開始信号, 発話終了信号により被験者へ通知される.

発話開始信号と発話終了信号の間隔は 5 秒間で, その間被験者は持続して発話を行う. これら一連の動作を 1 セットとして各被験者毎に 5 セットずつ行った.

各セットの間は休憩として 3 秒間の間隔を設けた.

4.3　実験結果

表は各被験者の/a/及び/e/の第 1, 第 2 フォルマントである.

実験において遅延等による発話への影響はなく, 特に問題がないことを確認した.

表 4.1: 被験者の母音フォルマント

	/a/		/e/	
被験者	F1(Hz)	F2(Hz)	F1(Hz)	F2(Hz)
AH	694	1536	530	1856
TH	777	1266	622	1713
MO	700	1204	550	1820

第5章　本実験

5.1　目的

- フォルマント変換フィードバック音声を呈示したときの発話音声を測定する.

5.2　実験手順

実験では発話音声を連続母音/eaea ... /とし, 摂動として発話音声の/e/の部分を/a/に変換し, 被験者にフィードバックした.

図は実験手順を表している. 発話の開始, 及び終了は装着されたヘッドフォンを通じて発話開始信号, 発話終了信号により被験者へ通知される.

発話開始信号と発話終了信号の間隔は 30 秒間でその間被験者は発話を行う.

最初の 10 秒間は" 摂動なし" の音声がフィードバックされ, 次の 10 秒間は変換した "摂動あり" の音声がフィードバックされ, 最後の 10 秒間は再び" 摂動なし" が音声をフィードバックされる.

息継ぎは各自のタイミングで自由に行ってよいものとした.

これら一連の動作を 1 セットとして各被験者毎に 3 セットずつ行った.

各セットの間は休憩として 10 秒間の間隔を設けた.(図 5.1)

また被験者の意識を聴くことに向け, フィードフォワードによる要因を減少するため, 発話音声に変化が現れている間, ボタンを押すというタスクを与えた.

5.3　結果

5.3.1　分析データ

分析データは摂動を与える前の摂動なしの区間, 摂動ありの区間の前半, およびその後半, そして摂動を与えた後の摂動なしの区間の 4 つの区間を対象とし, 各区間から標本点として母音/e/の部分 3 つずつを取りだしたものとした. (図 5.2)

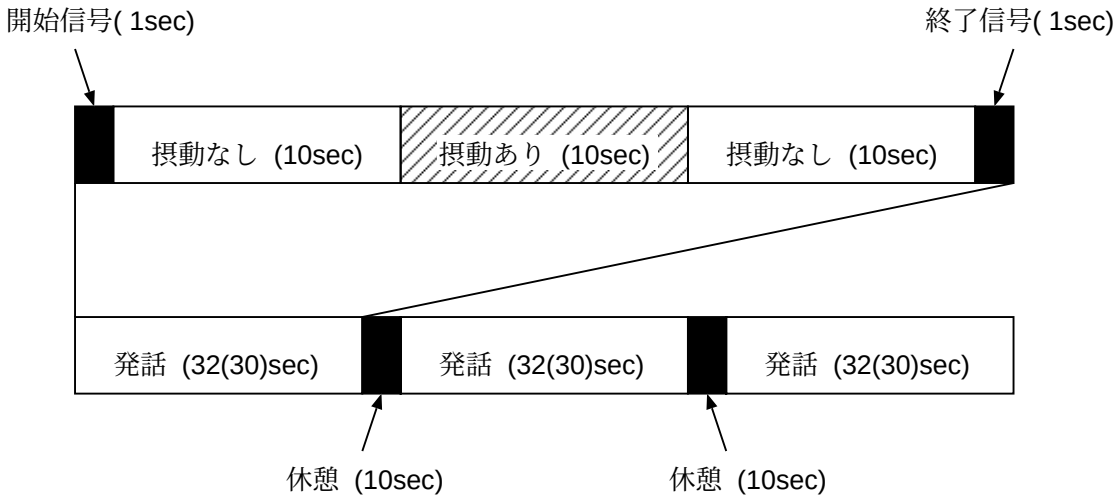


図 5.1: 実験手順

表 5.1: 分析条件

標本化周波数	8kHz
窓長	64ms(512 点)
窓関数	hamming 窓
フレームシフト	128 点 (16ms)
ケプストラム次数	30 次

5.3.2 スペクトルの変動分析

分析には発話音声の短時間スペクトルを用いた変動分析を行った.

ここでは短時間スペクトルのスペクトル推定において, 不偏推定量を与える, 不偏推定法 [9] による分析を行った. 図 5.3,5.4, 5.5 は被験者 MO,TH,AH の発話音声の分析結果である.

分析条件は表 5.1 の通りである.

図 5.3,5.4,5.5 は不変推定法によって得られた, 被験者ごとの短時間対数スペクトルのスペクトル包絡を重ねて表示したものである.

縦軸は対数パワー, 横軸は周波数に相当する.

図の上の 2 つは “摂動なし” の区間での分析結果, 下の 2 つは “摂動あり” の分析結果に対応する.

また図 5.6, 図 5.7, 図 5.8 はそれぞれ図 5.3,5.4,5.5 の周波数ごとの分散を表したもので, 縦軸は分散, 横軸は周波数に相当する.

表 1〜3 は図の分散の平均をとったもので. また図 5.9 は表 5.2 をグラフにプロットした

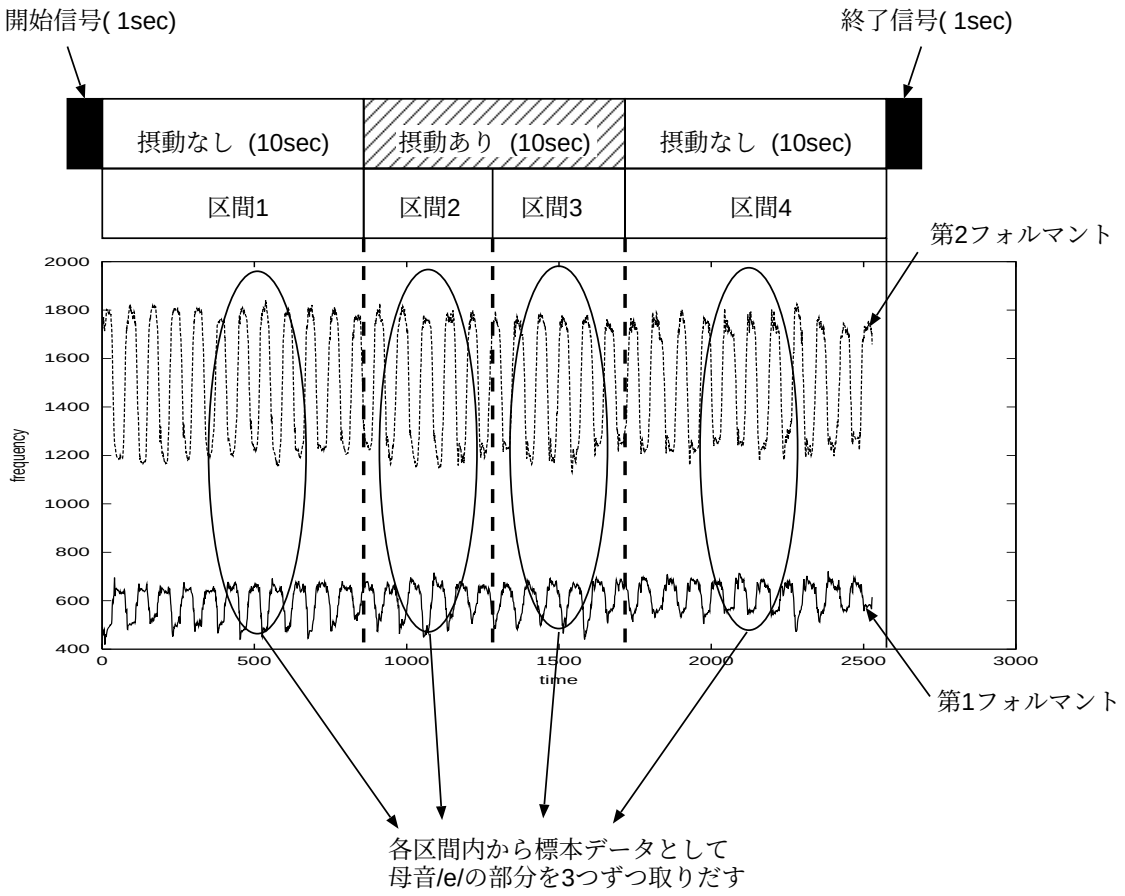


図 5.2: 分析対象データ

ものである.

グラフの縦軸は分散平均, 横軸は区間 (1〜4) を表している.

全体として (AH を除く) のスペクトルの変動分析において摂動がない状態の区間 1 から摂動が加わっている区間 2, 区間 3 のに向けて分散の値が大きくなる傾向があり, さらにそこから摂動がない区間 4 に向けて若干分散の値が小さくなる傾向がみられる. この結果は “摂動あり” の区間で “摂動なし” の区間と比較してスペクトルの分散が大きくなっていると解釈することができる.

また実験後の被験者による報告では “摂動なし” のときと比較して “摂動あり” の場合に発話に関する違和感を訴えている.

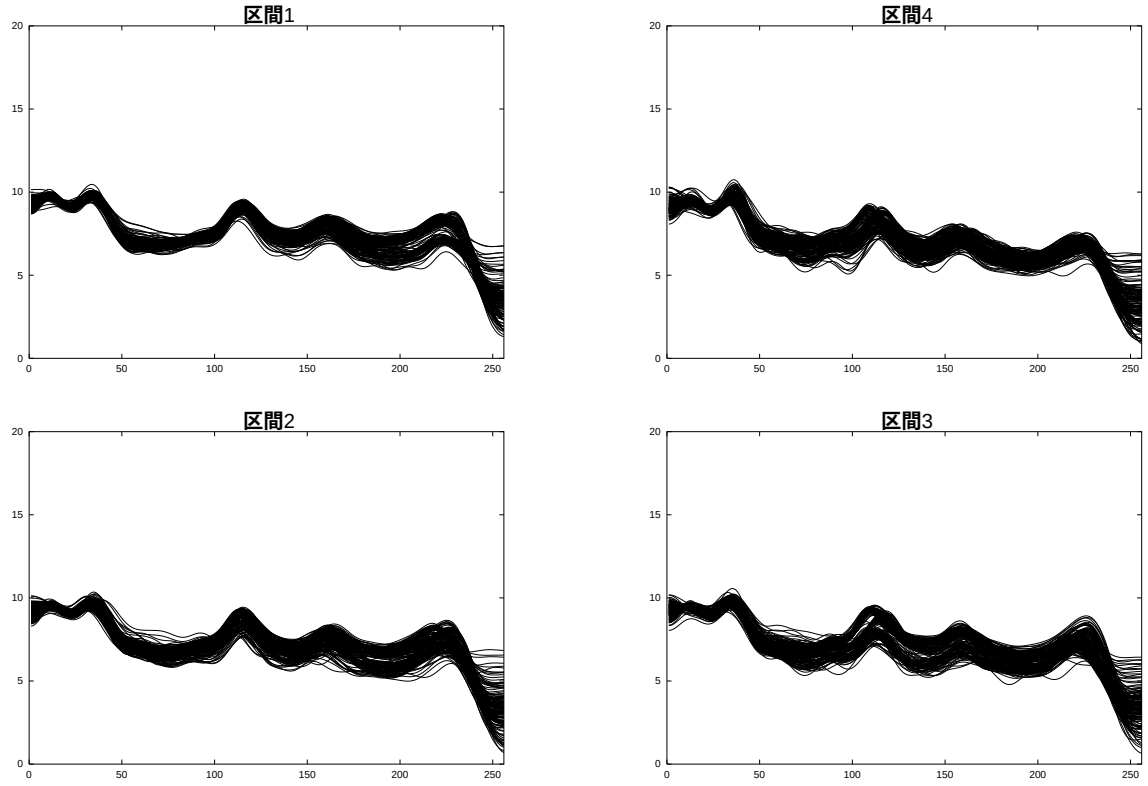


図 5.3: 被験者 MO の短時間スペクトル

5.4 考察

実験の結果から摂動を与えることにより発話が変化したものと思われる. その要因の一つとしてフィードバック音声の変更により, 発話における制御特性の悪化により発話動作が不安定になっていることが可能性の一つとして考えられる.

これは発話時にフォルマントの情報を聞き取りながら発話していることの証拠として考えることができる.

また今回の実験では被験者にフィードフォワードの影響を低減するためボタンを押すというというタスクを与えたが, それ以前に行ったボタンを押すというというタスクがない,

表 5.2: 被験者毎のスペクトル変動

被験者	区間 1	区間 2	区間 3	区間 4
AH	0.17545	0.18362	0.12333	0.12729
TH	0.29044	0.34167	0.33778	0.31435
MO	0.18518	0.24488	0.30932	0.23419

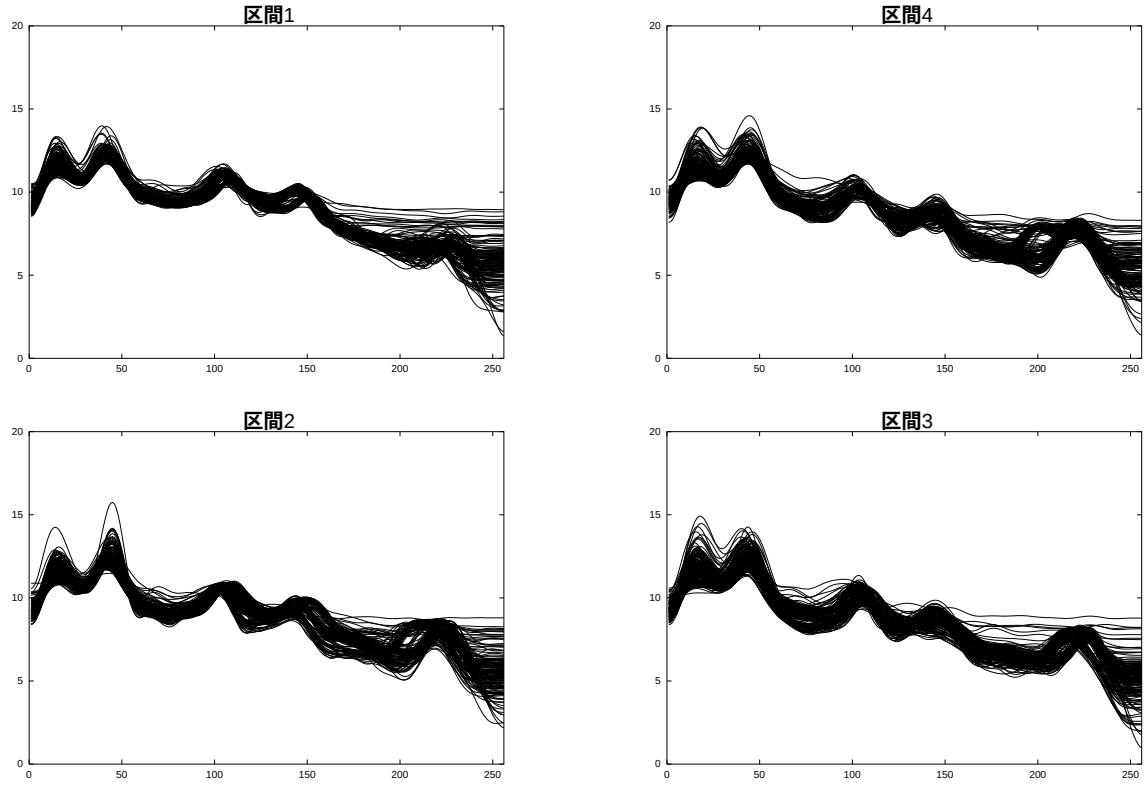


図 5.4: 被験者 TH の短時間スペクトル

同様の実験と比較しても発話が困難になったと報告している.

これは通常の発話においては主にフィードフォワード制御によって発話の制御がなされている可能性を示している.

また被験者 AH に目立った応答が確認できないことや, 被験者ごとに分散のが大きくなる周波数やその度合いが異なることから, 変換聴覚フィードバックの影響は言語習得過程等の違いによる個人差が存在するものことが推測される.

しかし, 今回の実験では音声の変換に対して補正するような応答を確認することはできなかった. その原因としていくつかの理由が考えれられる.

まず始めに言語習得時に獲得した発話運動の制御に関する情報, すなわち” 逆モデル” の影響があまりに強いため, フィードバック制御による応答がフィードフォワード制御の応答に埋もれた可能性がある.

これは言語習得後に聴力を失った被験者が発話可能なことや Houde らの実験により変換フィードバックによる学習効果が 1ヵ月保たれる被験者がいたことからも” 逆モデル” の影響の大きさを窺うことができる.

最後に被験者がフィードバック音声の変換の補償方向について無知であったため補正できなかった可能性が考えられる. Houde らの実験で補償方向の応答が表われたのは,1 時間の学習のなかで, 補償方向に関する情報を獲得したことが可能性として考えられる.

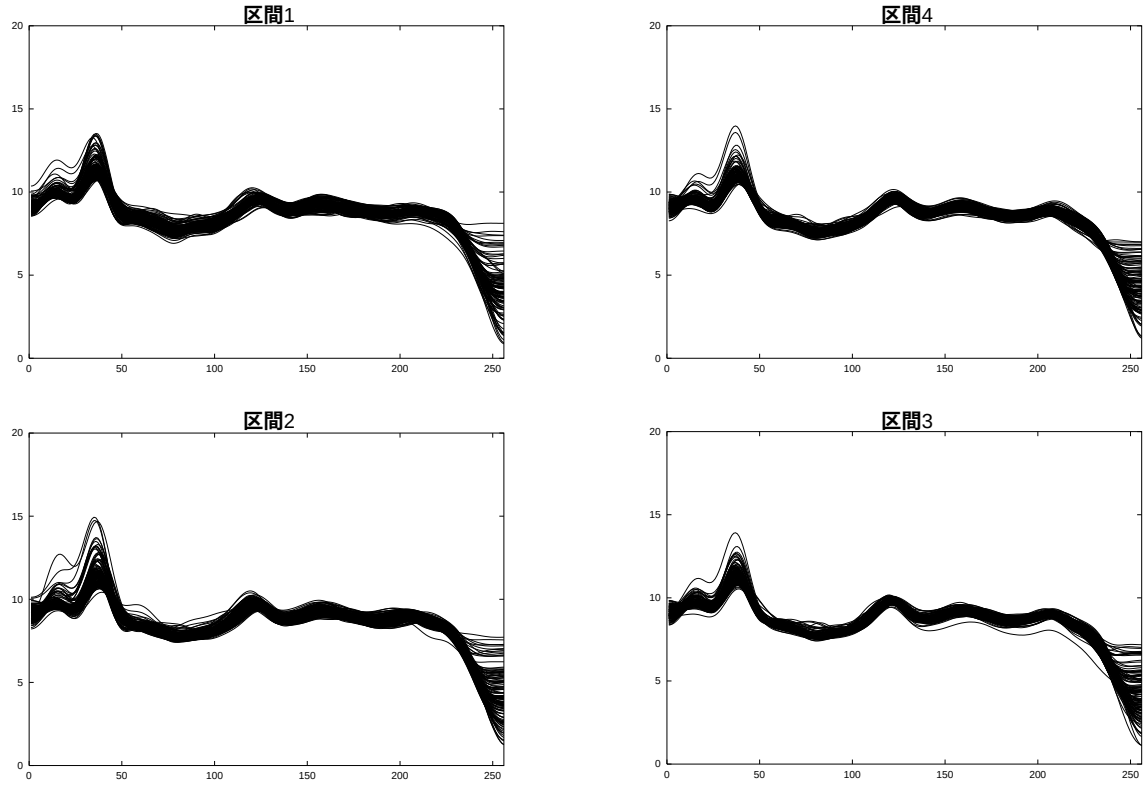


図 5.5: 被験者 AH の短時間スペクトル

以上のことから言語習得後の音声生成過程には主にフィードフォワード制御により行われている可能性が高いものと思われる.

しかしフィードバックによる制御も同時に行われており, その際フィードバック音声として F1,F2 のような音韻情報も利用している可能性はますます高まった.

このことから発話におけるフォルマント制御モデルは本質的には発声の基本周波数の制御モデルと同様にフィードフォワード制御とフィードバック制御の並列動作により説明が可能であると推測される.

ただ制御対象の違いによりフィードバックとフィードフォワードの重要さや獲得する逆モデルが異なると考えられる.

系の制御が難しい基本周波数の制御ではフィードバックによる情報が非常に重要なものとなるが, 系の制御が比較的容易な発話の制御ではフィードバックによる情報があまり必要ではないと考えられる.

生体には負のフィードバック経路が存在することが知られているが, それらのゲインは小さく, 比較的応答が速い体性感覚についても信号が感覚受容器への到達から実際の運動に移るまでの時間は 50ms〜100ms と長く, 腕の運動にかかわる視覚のフィードバック経路において 100 数十 ms〜200ms に達する.

これに対して一般に発話運動は運動としては速い分類に入り, かつ調音結合の特徴から

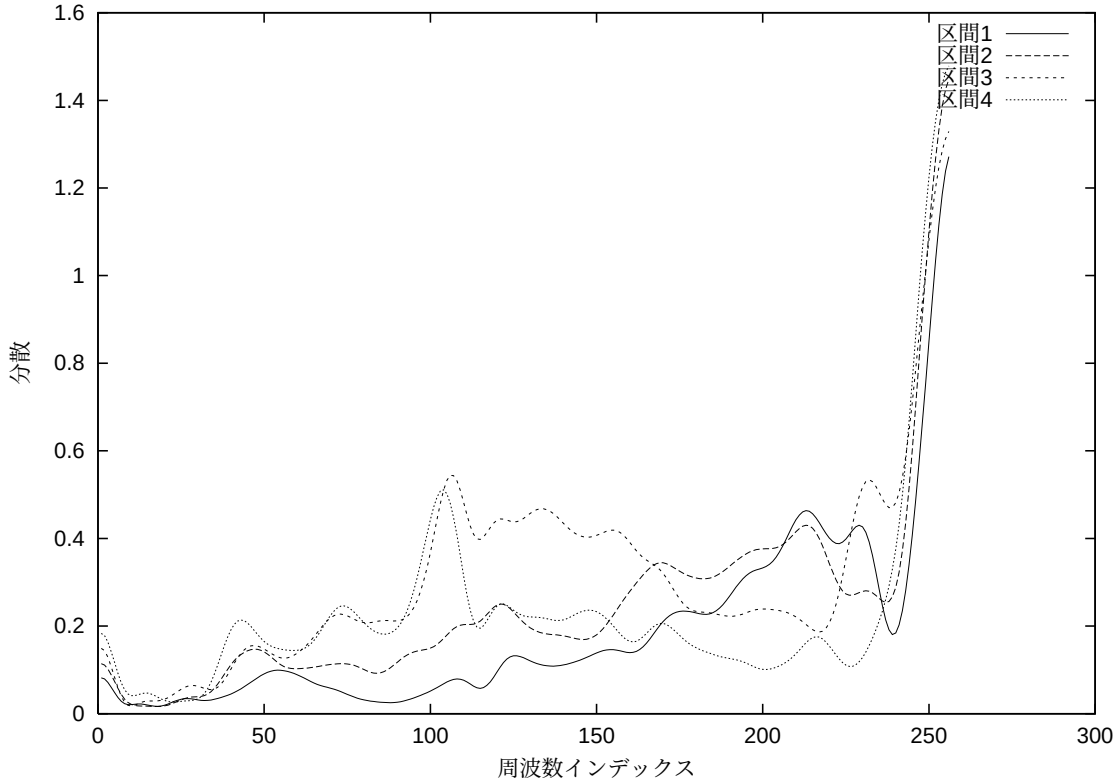


図 5.6: 被験者 MO の周波数毎のスペクトル変動

も分かるようにその動きは非常に滑らかである. このような運動を行うには, 予測による運動が必要とされる. もし仮にここでフィードバックによる干渉が入ったとするとかえって発話動作を破壊してしまいかねない.

さらにそのフィードバックゲインが非常に大きいものだとすると発話動作がスティフネスの大きないわゆる “硬い” 運動になるため, 調音結合のような滑らかな運動が実現できなくなる.

最後に, 多分に主観的ではあるが, 発話運動の生成における聴覚フィードバックの役割についての考えを述べる.

実験の結果からフィードバックによる影響は発話音声に微弱に現れた程度で発話動作を大きく変化させるものではなかった. これは発話運動を行う上での戦略上, フィードバックゲインを敢えて小さくすることで運動を妨害しないようにして, なおかつフィードバックゲインを完全にゼロにしてしまうのではなく, 僅かに残すことにより, 環境の変化による外界と外界の内部モデル (“逆モデル”) の誤差を監視し, 変化があれば調整するような目的に聴覚フィードバックが用いられているのではないかと考えられる.

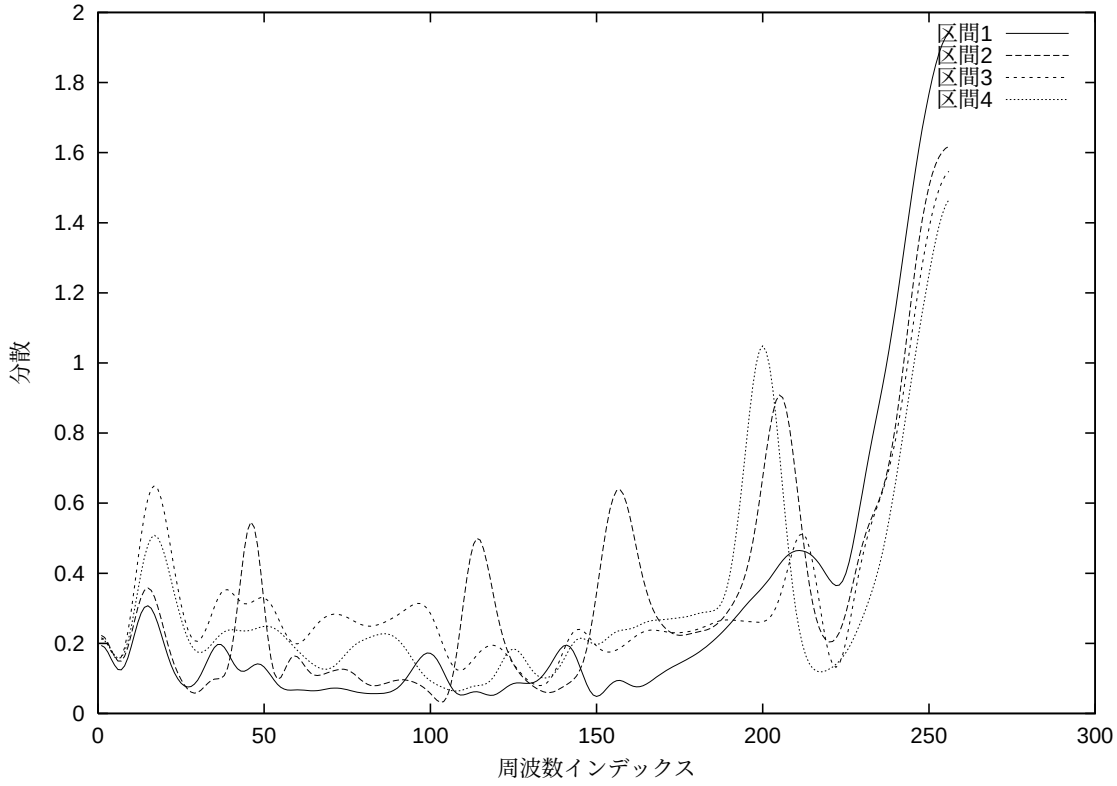


図 5.7: 被験者 TH の周波数毎のスペクトル変動

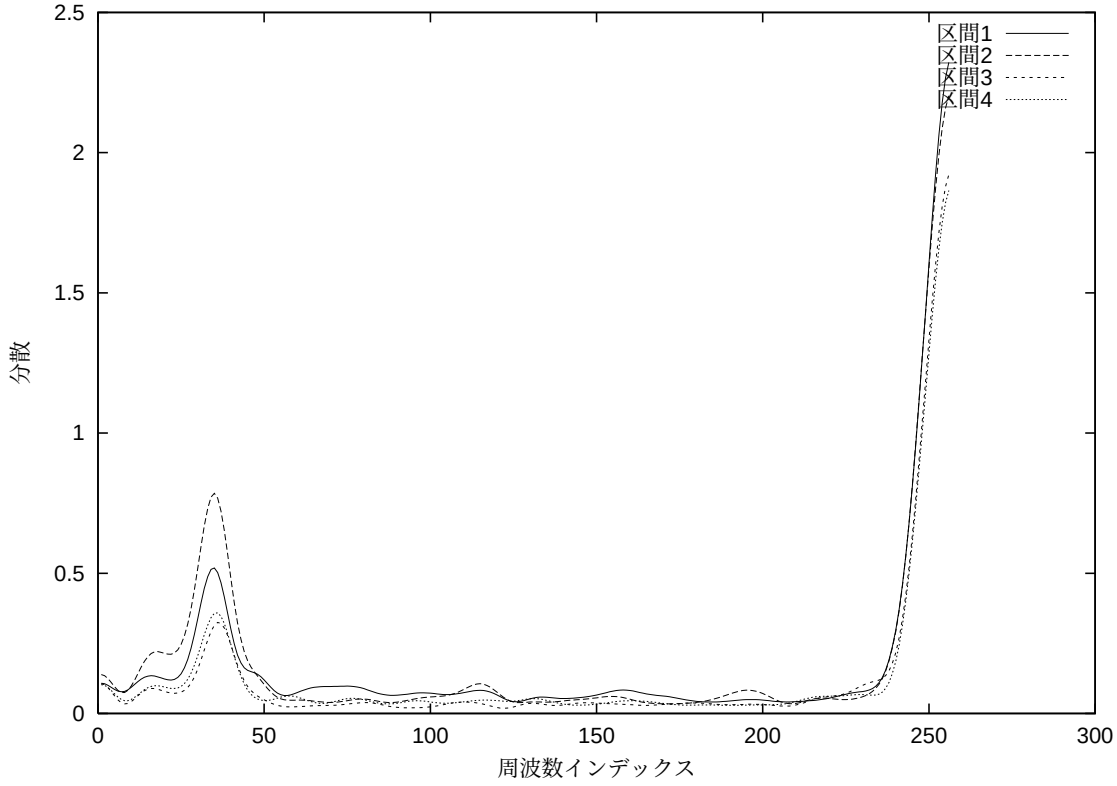


図 5.8: 被験者 AH の周波数毎のスペクトル変動

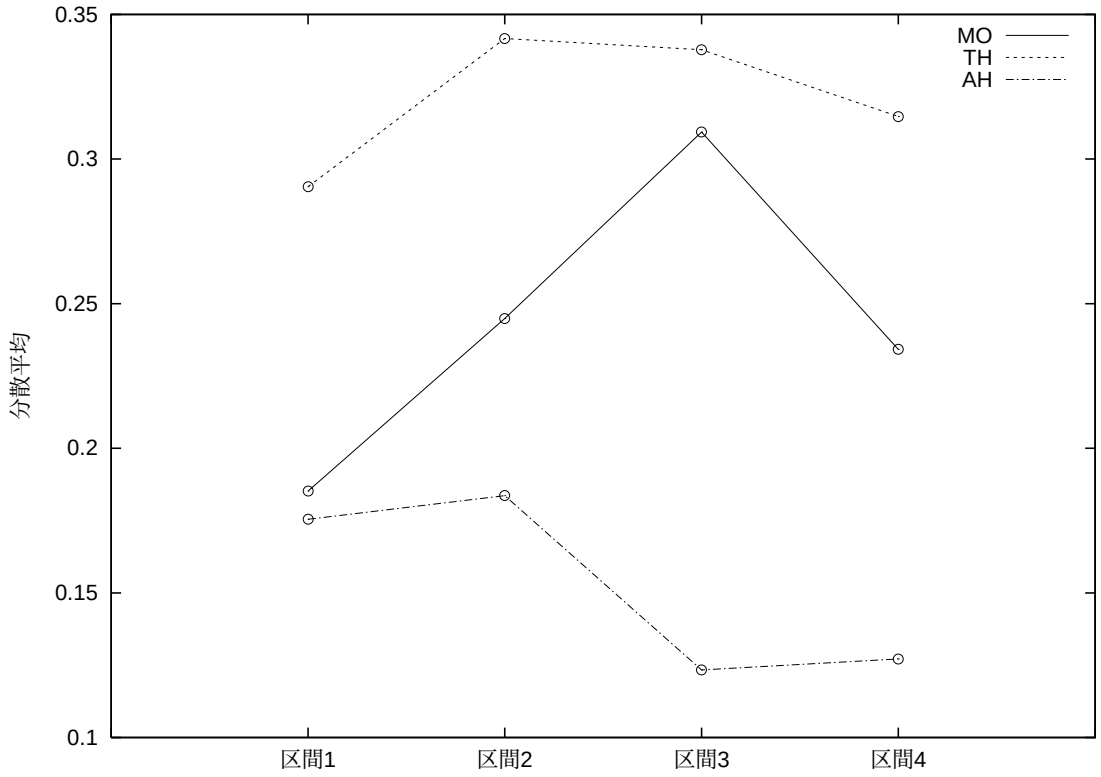


図 5.9: 被験者毎のスペクトル変動

第6章　結論

6.1　本論文のまとめ

本研究では実時間でフォルマントを変換し被験者にフィードバックすることが可能な実験系を構築し, 実験を行った. その結果, フォルマントを変換したフィードバック音声の呈示区間においてスペクトルの分散が大きくなる傾向がみられた. このことからフィードバック音声の変更が, 発話動作に何らかの影響を与えているものと解釈することができる. この結果は発話時にフォルマントのようなスペクトル構造に関する情報を利用していることを支持する結果である. またそれらの応答は被験者により異なる. その原因として音声習得過程等による個人差が存在するものと思われる. 今回行った実験等による結論として, 通常の発話においては主にフィードフォワードによる制御が行われていると思われる, しかしフォルマントに関する情報を利用している可能性が高いことからフィードバック制御も同時に行われているものと思われる. つまり本質的には発声の基本周波数制御モデルと同様にフィードフォワード制御とフィードバック制御の並列動作により説明が可能であると推測される.

6.2　今後の課題

今回の実験では変換された音声を, 補償するような動作を確認するには至らなかったが, 被験者は摂動を与えられた状態では発話しにくい等の, 発話に対する違和感を訴えていることから, 筋電計やカメラによる口の周りの筋肉の動きの測定など音声以外の測定方法を用いることでより明確な応答が確認できる可能性がある.

また今回の実験では/eaea . . . /という単純な繰り返しの音節であったため, フィードフォワードによる応答がフィードバックによる応答の発生を抑圧したことが考えられる. さらに付け加えると日本語という言語による問題も挙げられる. そもそも日本語は母音の数が 5 つと他の言語に比べて圧倒的に少ない. これはそれだけ母音に関する弁別が他の言語に比べて不利である. これらはフィードバック制御の発生を妨げている要因ではないかと思われる. このため, フィードフォワード制御の要素を除去, もしくは弱めるなどフィードバックの要素を引き出すような実験パラダイムが必要があると考えられる.

その方法として普段使わないような, 音韻や音節を利用することや, 何らかの方法でフォルマントに関する情報を視覚的にもフィードバックしてやることなどが考えられる. また音韻, 音節の違い, 与える摂動の種類や大きさ, 摂動を与えるタイミングなどによる影響に

ついての, 定量的な評価は今後の課題である.

参考文献

[1] B.S.Lee “Effect of Delayed speech feedback.”, J.Accoustic.Soc.Am, 22,824-826(1950).

[2] H.Lane, B.Tranel. “The Lombard sign and the role of hearing in speech.”, Jornal of Speech and Hearing Research, 14,677-709(1971).

[3] H.Lane, J.Webster. “Speech deterioration in postlingually deafened adults.”, Jornal of the Acoustical Society of America, 89,859-866(1990).

[4] J.Houde, M.Jordan. “Sensorimotor Adaptation of Speech I: Conpensation and Adaptation”, Jornal of Speech Language, and Hearing Reserach, 45, pp.295-310(2002).

[5] R.W.Schafer and L.R.Rabiner. “System for Automatic Formant Analysis of Voiced Speech”, J.Accoust.Soc.Am, Vol.47, No.2, pp.24-33 (1977).

[6] R.W.Schafer and L.R.Rabiner (著) 鈴木 (訳) “音声のディジタル信号処理 (下)”, コロナ社, (1983).

[7] Shimon Sapir, Elizabeth DeRosier, Andrea M.Simonson, and Amy Wohlert. “Effects of frequency modulated tones and vowel formants on perioral muscle activity during isometric lip rounding”, Jornal of Voice,and Hearing, Vol.4, No.2, pp.152-158(1990).

[8] 甘利, 外山 “脳科学大事典” 朝倉書店, (2000).

[9] 今井, 古市 “対数スペクトルの不変推定”, 電子通信学会論文誌, ’87 /3 Vol.J70-A No.3. pp.471-480 (1987).

[10] 金井 “音・振動のスペクトル解析”, コロナ社, (1999).

[11] 川人 “脳の計算理論”, 産業図書, 1996.

[12] 河原 “音声知覚･生成相互作用の伝達特性について”, 音響学会聴覚研究会資料, H-95-35, pp.223-226,(1995).

[13] 河原 “変換聴覚フィードバックによる音声生成・知覚相互作用の検討 - 非定常ピッチ変換による影響の解析-”, 音響学会聴覚研究会資料, H-93-24, (1993).

[14] 河原, 加藤,J.C.Wiliams “聴覚による発声の制御モデルとシミュレーション” 秀英出版,(1964).

[15] 佐藤 “スペクトル変型聴覚フィードバックによる音声生成・知覚の相互作用に関する研究” 北陸先端科学技術大学院大学, (2003).

[16] 沢田, 筧 “聴覚フィードバックに利用される音声情報の物理的特徴” 日本音響学会聴覚研究会資料,Vol.33, No.2, H-2003-21 pp.117-122 (2003).

[17] 電子通信学会 “聴覚と音声”, コロナ社, (1980).

[18] 森友, 薬師, 馬場 “RTLinux リアルタイム処理ハンドブック”, 秀和システム, (2000).

謝辞

本研究を進めるにあたり, 日頃から多くの貴重な御助言, 御指導をいただきました, 北陸先端科学技術大学院大学 情報科学研究科 赤木 正人教授, 党 建武助教授, 鵜木 祐史助手, 並びに本学教官の皆様に深く感謝致します. また御多忙の中, 御助言, 御討論を頂き, また, 実験にご協力いただいた研究室の皆様に感謝致します. 最後に, 研究を進めるにあたり日頃からあたたかく見守って下さった家族や友人に深く感謝致します