

筑波大学 情報学群 情報メディア創成学類

卒業研究論文

変換聴覚フィードバックが話者の発話特性に
与える影響の調査

大沼 怜生

指導教員 川口 一画，志築 文太郎

2024年2月

概要

本研究では、発話音声に音響的変換処理を加えたものを話者にのみ聞かせる手法である変換聴覚フィードバックにおいて、話者の発話の音響特性を個別に変換した際の発話特性への影響について2回の調査を行った。

1回目の調査では、原稿を音読中の話者の発話のピッチ、抑揚、およびラウドネスをそれぞれ正負方向に変化させた変換聴覚フィードバックを行い、その際の話者の発話特性の変化を調査した。調査の結果、フィードバック音声の抑揚増加およびラウドネス減少において共に発話のシマの増加が確認された。また、フィードバック音声のラウドネス増加において発話のジッタおよびシマの減少、そしてラウドネスレベルの増加が確認された。調査結果および半構造化インタビューで得られた意見を基に変化の原因を考察し、調査に用いたシステムの課題を検証した。

2回目の調査は、1回目の調査で得られた課題に対する改善、およびピッチ、抑揚、ラウドネス変化に加えてハイパスフィルタおよびローパスフィルタを用いた変換聴覚フィードバックを行い、発話特性の変化を調査した。調査の結果、フィードバック音声のピッチ上昇において発話のジッタ減少、ピッチ下降において発話のZCR増加が確認された。また、フィードバック音声の抑揚増加において発話のHNR増加が確認された。フィードバック音声のラウドネスについては、増加および減少のそれぞれにおいてシマの増加が確認された。

2回の調査の結果から、フィードバック音声のピッチを上昇させることで話者の発話を明瞭にできると考えた。しかし、具体的な変換量およびその他の発話の変化を実現可能なパラメータについて、さらなる調査が必要である。

目次

第1章	はじめに	1
1.1	背景	1
1.2	目的とアプローチ	2
1.3	貢献	2
1.4	本論文の構成	2
第2章	関連研究	3
2.1	ボイスチェンジャによる発話特性操作の効果	3
2.2	遅延聴覚フィードバックによる発話特性操作の効果	3
2.3	AAFの効果および話者の心理状態に与える影響	4
2.4	AAFが話者の発話特性に与える影響	4
第3章	システムの設計	6
3.1	システムの構成	6
3.2	音響特性の変換方法	7
第4章	実験1	10
4.1	実験条件	10
4.2	仮説	10
4.3	実験参加者	11
4.4	実験タスク	11
4.5	実験環境	12
4.6	評価指標	12
4.7	実験結果	13
4.7.1	基本周波数	13
4.7.2	ジッタ	14
4.7.3	シマ	14
4.7.4	ラウドネスレベル	15
4.8	議論および考察	16
4.8.1	仮説の検証	16
4.8.2	ジッタ・シマおよびラウドネスレベルの関係	17
4.8.3	フィードバックの遅延	17

第 5 章 実験 2	18
5.1 改善点	18
5.1.1 システムの遅延短縮	18
5.1.2 評価指標の見直し	19
5.1.3 トレーニング条件の追加	20
5.2 実験条件	20
5.3 仮説	23
5.4 実験参加者	23
5.5 実験タスク	23
5.6 実験環境	24
5.7 評価指標	24
5.8 実験結果	25
5.8.1 基本周波数	25
5.8.2 ジッタ	26
5.8.3 シマ	26
5.8.4 ラウドネスレベル	27
5.8.5 ZCR	28
5.8.6 音響インテンシティ	29
5.8.7 HNR	30
5.9 議論および考察	31
5.9.1 仮説の検証	31
5.9.2 ピッチ変換の効果	32
5.9.3 考察のまとめ	32
第 6 章 本研究の制約および今後の課題	34
6.1 変換パラメータの選定	34
6.2 評価指標の選定	34
6.3 コミュニケーションへの応用	34
第 7 章 おわりに	35
謝辞	36
参考文献	37
著者論文リスト	41
付 録 A 実験 1 にて用いたアンケート	42
付 録 B 実験 2 にて用いたアンケート	44

目次

3.1	システム構成図	7
3.2	抑揚を増加させる設定	8
3.3	抑揚を減少させる設定	8
4.1	実験環境	12
4.2	パラメータ変換前および変換後の基本周波数	13
4.3	パラメータ変換前および変換後のジッタ、有意差が出た条件に*を示す	14
4.4	パラメータ変換前および変換後のシマ、有意差が出た条件に*を示す	15
4.5	パラメータ変換前および変換後のラウドネスレベル、有意差が出た条件に*を示す	16
5.1	改善後のシステム構成図	19
5.2	男性用ハイパスフィルタ	21
5.3	男性用ローパスフィルタ	21
5.4	女性用ハイパスフィルタ	22
5.5	女性用ローパスフィルタ	22
5.6	実験環境	24
5.7	パラメータ変換前および変換後の基本周波数	25
5.8	パラメータ変換前および変換後のジッタ、有意差が出た条件に*を示す	26
5.9	パラメータ変換前および変換後のシマ、有意差が出た条件に*を示す	27
5.10	パラメータ変換前および変換後のラウドネスレベル	28
5.11	パラメータ変換前および変換後の ZCR、有意差が出た条件に*を示す	29
5.12	パラメータ変換前および変換後の音響インテンシティ	30
5.13	パラメータ変換前および変換後の HNR、有意差が出た条件に*を示す	31
A.1	実験 1 にて用いたアンケートの内容	43
B.1	実験 2 にて用いたアンケートの内容	45

第1章 はじめに

本研究では、変換聴覚フィードバックを用いた際の話者の発話特性の変化を調査する。本章では、本研究の背景として、変換聴覚フィードバックの効果に関する既存の知見および課題を説明する。次に、本研究の目的とアプローチ、貢献、および本論文の構成を示す。

1.1 背景

近年、オンライン通話ツール [1, 2, 3] の普及に伴いオンラインでの通話機会が増加している。これらのツールはユーザのカメラ映像を有効にするか無効にするか選択可能である物が多い。カメラ映像が無効の場合、表情等の視覚情報による非言語情報が伝達されないため、参加者は相手の声の高さや抑揚といった発話特性から気分や感情等を判断する必要がある。しかし、十分なボイストレーニングを受けていない話者が自分の声の発話特性を意図通りに制御することは困難である。

このような問題を解決する手法の1つとして、ボイスチェンジャの利用が挙げられる。近年の機械学習の発展に伴い、モデル生成型ボイスチェンジャにおける音声の変換精度は向上している [4, 5]。しかし、ボイスチェンジャを通した音声は発話の音響特性を変換して生成された音声であるため、声質がボイスチェンジャを通す前の音声から大きく変化する可能性がある。

その一方、話者の発話特性を変化させる手法の1つとして、変換聴覚フィードバック (Altered auditory feedback, AAF) が存在する。AAFとは、マイクを通じて取得した話者の発話音声に音響的な変換処理を施し、変換した音声を話者のみに対してリアルタイムで聞かせる手法である。AAFを用いることで話者の感情や発話特性をある程度制御することが可能である [6, 7]。本研究では、このようなAAFの特性を活用することで、ボイスチェンジャを用いることなく、話者の発話特性を変化させるシステムを実現することができると考えた。発話特性の変化がインタラクションに与える影響として、発話音声のピッチ操作による話者の身体的および社会的優位性の向上 [8] や、発話音声のピッチおよびラウドネス操作による話者の発話の説得力の変化 [9] が示されている。このように、人間の発話において音響特性は非言語情報として意味を持つ。AAFによってボイスチェンジャを用いずにこれらの特性を変化させることができれば、話者の意図通りの気分や感情等を、話者に負担をかけることなく聞き手に伝えることが可能であると考えた。

しかし、AAFにおいて発話の音響特性に与える影響および変換する音響特性のパラメータの違いによる発話特性への影響については、基本周波数、フォルマント、摩擦音の周波数重

心, およびラウドネスに対する操作および分析が行われているのみである [10, 11, 12]. 変換する音響特性および発話特性への影響について調査することは, 対話支援システムを設計する上で発話特性の自在な変換を実現するために必要である.

1.2 目的とアプローチ

本研究の目的は, AAF を用いた対話支援システムを設計するにあたり, 発話の音響特性の変換が話者の発話特性に与える影響について調査することである. この目的に対するアプローチとして, 発話の音響特性を個別に変換可能な AAF システムを提案する. 変換する音響特性として, ピッチおよびラウドネスの他, これまでに調査が行われていない音響特性として抑揚にも着目する. 話者の発話のピッチ, 抑揚, およびラウドネスについて, AAF において発話の音響特性を個別に変換した際の話者の発話特性への影響を調査する.

ピッチ, 抑揚, およびラウドネスについての調査に際し, 特に抑揚に関しては 1.1 節にて挙げた発話特性である基本周波数, フォルマント, 摩擦音の周波数重心, およびラウドネスの分析では影響が判断できないと考えられる. そのため本研究では, 抑揚の分析項目として新たに音声波形の周期方向の乱れであるジッタおよび振幅方向の乱れであるシマを追加する.

1.3 貢献

本研究の貢献は以下の通りである.

- AAF における話者の音響特性を個別に変換した際の発話特性の変化について, 多岐にわたる調査を行った.
- AAF を用いた対話支援システムについて, 発話特性の変化に基づき実現可能な機能を提案した.

1.4 本論文の構成

本論文の構成は以下の通りである. 第 1 章においては, 本研究の背景, 目的とアプローチ, および貢献を示した. 第 2 章においては, 本研究に関連する研究を述べ, 本研究の位置づけを示す. 第 3 章においては, 本研究における提案システムを示す. 第 4 章においては, 本研究を評価するために行った実験 1 の内容, 実験 1 の結果, 実験 1 の議論および考察を述べる. 第 5 章においては, 実験 1 にて見られた課題を踏まえた改善点, 改善されたシステムおよび評価指標を用いて行った実験 2 の内容, 実験 2 の結果, 実験 2 の議論および考察を述べる. 第 6 章においては, 本研究における制約及び今後の課題を述べる. 第 7 章においては, 本研究の結論を述べる.

第2章 関連研究

本章では、これまで行われてきた発話特性操作に関する研究を提示し、本研究においてAAFを利用するに至った経緯を説明する。その後、AAFが話者の心理状態に与える影響を調査した研究、および話者の発話特性に与える影響を直接調査したAAF研究について述べる。

2.1 ボイスチェンジャによる発話特性操作の効果

ボイスチェンジャを用いたいくつかの研究において、話者の音声の音響特性を変化させることにより対話相手の話者に対する印象が変化することが報告されている。話者の音声のピッチを下降させることにより、話者の身体的および社会的優位性の向上が報告されており [8]、これらの優位性の向上は会議における参加者の説得力を向上させることが確認された [9]。また、若年成人男性の音声のピッチを上昇させフォルマントを下降させることにより、話者の魅力が減少することが示されている [13]。音声そのものの好みへの影響も調査されており、男性はピッチの高い女声を、女性はピッチの低い男声をそれぞれ好むことが報告されている [14]。しかし、ボイスチェンジャは話者の音声を変換して対話相手に向けて再生するため、変換する内容によっては再生される音声の話者の肉声から大きく変化する可能性が存在する。そのため、公的な会議のように話者本人の肉声が求められる場面において、ボイスチェンジャを用いることは難しいと考えられる。

この課題に対するアプローチとして、音声変換 (Voice Conversion, VC) の利用が挙げられる。VCとはボイスチェンジャの一種であり、発話の言語情報を維持しつつ非言語情報のみを変換する技術である [15, 16, 17]。統計処理および機械学習等の利用により、VCを用いることで人間の肉声に近い音声を生成することが可能である [18]。しかし、VCによる発話特性変換は、話者の音声を生成するターゲット音声に変換するためのパラメータを統計処理および機械学習等で決定する。そのため、VCの利用には事前学習に用いるターゲット音声および話者の音声のデータセット作成が必要である。

2.2 遅延聴覚フィードバックによる発話特性操作の効果

話者の発話特性操作の例として、吃音の治療を目的とした発話の明瞭化が挙げられる。いくつかの研究において、成人の吃音者に対し、本人の音声を 50 ms から 100 ms 程度発話から遅延させてフィードバックすることにより、吃音が改善したという結果が報告されている

[19, 20, 21]. しかし、この遅延聴覚フィードバックによる発話の明瞭化は吃音者にのみ有効であり、非吃音者に対しては発話の流暢性が低下すると報告されている [22, 23, 24].

以上の理由から、本研究では発話特性の操作手法として、非吃音者においても発話特性への影響が見られ、かつ事前準備を必要としない AAF を採用する. 1.1 節にある通り AAF は話者の発話をフィードバックする手法であるため音声のデータセット作成のような事前準備を必要とせず、対話相手が聞く声は話者自身の声であるため変換前後の声質の乖離は発生しにくい. また、フィードバック音声の遅延を短縮した AAF であれば上述した流暢性の低下は発生しないと考えられる.

2.3 AAF の効果および話者の心理状態に与える影響

AAF を用いることによって、話者の気分や感情といった心理状態に対して影響を与えることが知られている. Arakawa ら [25] は、機械学習を用いた AAF を使用することで話者の自己表象に影響を与える可能性を示している. Jean ら [7] は、スピーチ中の話者に気づかれることなく、話者の感情を幸せ、悲しみ、および恐怖の 3 種類に変化させることに成功した. この研究では、ピッチ、フォルマント、声の震え、および帯域通過フィルタを利用している. Taguchi ら [26] は、フィードバック音声のピッチ上昇により、映像会議中に会議参加者の気分が向上し会話が活発になることを明らかにした. Jean ら [27] は、フィードバック音声のピッチ下降、フォルマント下降、およびローシェルフフィルタの利用により、対立中の話者の不安が軽減されることを発見した. また、フィードバック音声のピッチを下降させることにより、話者が自身をより力強く感じるようになると報告した. また、成瀬ら [28] は、フィードバック音声のピッチ下降およびハイシェルフフィルタの利用により、対面でのスピーチにおいて AAF の使用が話者の心理状態に影響を与える可能性を示唆している.

しかし、多くの AAF 研究は話者の心理状態を対象としており、音響特性のパラメータを対象とした AAF 研究は少ない. よって本研究では AAF における音声の音響特性に注目し、音響特性を示すそれぞれのパラメータが、話者の発話特性に与える影響について調査する.

2.4 AAF が話者の発話特性に与える影響

Burnett ら [11] は、AAF においてピッチを変更した際の話者の発話音声の変化を調査している. その結果、フィードバック音声のピッチが下がると発話音声のピッチが上がり、フィードバック音声のピッチが上がると発話音声のピッチが下がることを示した. Lane ら [12] は、AAF において音量を変更した際の話者の発話音声の変化を調査している. その結果、フィードバック音声の音量が上がると発話音声の音量が下がることを示した. Coughler ら [10] は、小児期の子供に対して AAF を用いた際の、基本周波数、フォルマント、および摩擦音の周波数重心への影響を調査した論文をまとめている. その結果、4 歳以上の子供はフィードバック音声のパラメータ変換と逆向きの発話特性の変化が見られることを発見した. また、いくつかの例において、大人よりも 4 歳以上の子供の方が AAF の効果が小さいことを報告した.

これらの研究により、ピッチ、ラウドネスに関する成人の発話特性への影響および子供の基本周波数、フォルマント、および摩擦音の周波数重心への効果は示されている。しかし、実際の会話への応用を考慮すると、成人に関してはピッチおよびラウドネスのみでは調査が不十分である。そのため本研究では、ピッチおよびラウドネスに加えて、これまでに調査が行われていない音響特性として、声の抑揚についての調査を行う。

また、抑揚は瞬間的なピッチの変動であることから、音声全体のピッチおよびラウドネスのみの分析では抑揚への影響が判断できないと思われる。そのため、分析項目として新たにジッタおよびシマを追加し、AAFがジッタおよびシマに与える影響の分析を行う。

第3章 システムの設計

本研究では, AAFにおいて変換する音響特性が発話特性へ与える影響を調査するため, ピッチ, 抑揚, およびラウドネスを個別に変換可能な実験用システムを実装した. 本章では, 本研究で用いたシステムの設計を述べる. 最初にシステムの構成を述べた後, 変換する音響特性について述べる.

3.1 システムの構成

本システムは, オープンソースソフトウェアプラットフォームである DAVID[29] を用いて実装を行った. DAVID を実行可能なコンピュータのサウンド入力を DAVID の入力とし, 仮想オーディオケーブルを介して変換後の音声を再生する. また, サウンド入力および仮想オーディオケーブルからそれぞれ発話音声およびフィードバック音声を取得し, Open Broadcaster Software (OBS Studio) [30] を用いて録音した. なお, 著者の環境において, システムに入力された音声に変換され再生されるまでの時間 (以降, 遅延と記す) は 120 ms 程度であった. システムの構成を図 3.1 に示す.

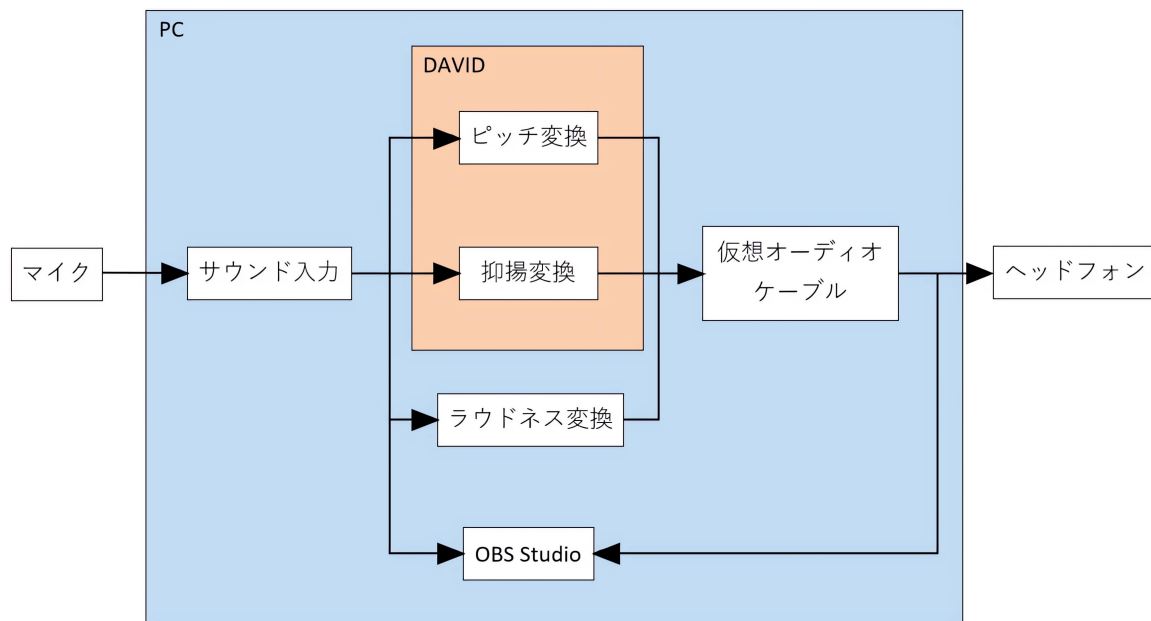


図 3.1: システム構成図

3.2 音響特性の変換方法

本研究で変換する音響特性のパラメータは、音声のピッチ、音声の抑揚、および音声のラウドネスの3種である。これらのパラメータは、音の三要素（音の高さ、音色、および音の大きさ）を基に、DAVIDで変換可能なものを考慮して選出した。

ピッチは、DAVIDのPITCHパラメータの操作により変換する。ピッチの影響を調査した先行研究[11]を参考に、100centのピッチ上昇および100centのピッチ下降を設定した。

抑揚は、DAVIDのINFLECTIONパラメータの操作により変換する。抑揚を増加させる設定を図3.2に、抑揚を減少させる設定を図3.3に示す。なお、これらの設定は、著者の主観に基づき変化が知覚可能な程度に設定された。

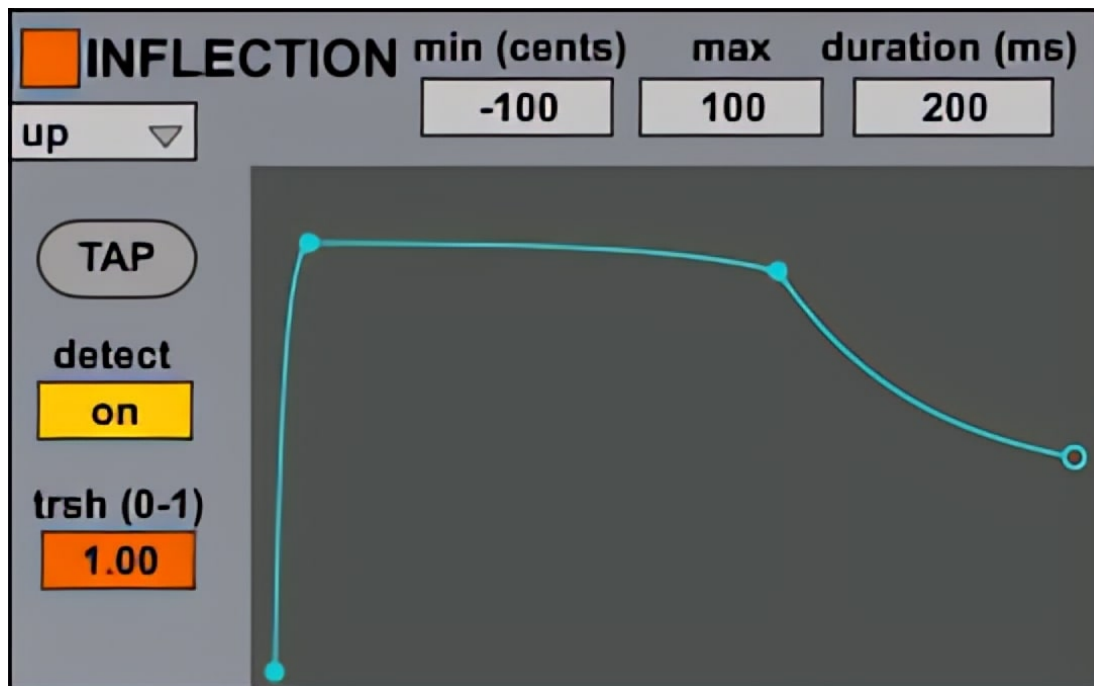


図 3.2: 抑揚を増加させる設定

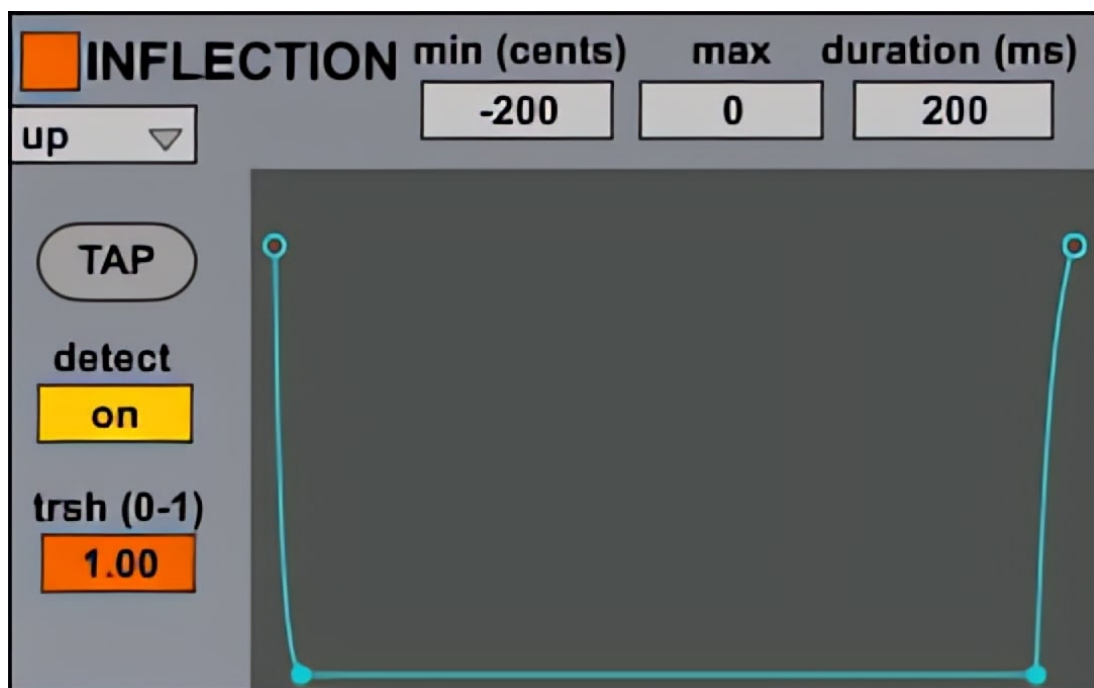


図 3.3: 抑揚を減少させる設定

ラウドネスは、コンピュータのサウンド出力ボリュームの操作により変換する。ラウドネスの増加においては出力ボリュームを5上昇させ、ラウドネスの減少においては出力ボリュームを5下降させる。なお、これらの数値設定も、著者の主観に基づき通常時から音量が50%程度増減するように設定された。

第4章 実験1

音響特性の変換が話者の発話特性に与える影響を調査するため、実装したシステムを用いて実験を行った。本章では、実験条件、仮説、実験参加者、実験タスク、実験環境、評価指標、実験結果、および考察を述べる。なお、本実験は筑波大学システム情報系の倫理審査委員会の承認を受けて実施した。

4.1 実験条件

実験条件は以下の6条件で、すべて参加者内配置で設計した。

- c1. ピッチ上昇
- c2. ピッチ下降
- c3. 抑揚増加
- c4. 抑揚減少
- c5. ラウドネス増加
- c6. ラウドネス減少

すべての条件において、実験者は参加者のタスク実施中に音響特性のパラメータ変換を行う。

c1では100 cent 上昇させ、c2では100 cent 下降させる。c3では抑揚増加のパラメータ変換を行い、c4では抑揚減少のパラメータ変換を行う。c5では、実験で使用したラップトップコンピュータのサウンド出力ボリュームを上昇させ、c6ではサウンド出力ボリュームを下降させる。具体的な変換方法に関しては3.2節にて述べた通りである。

4.2 仮説

本実験では、以下の3点の仮説を設定した。

- H1. フィードバック音声のピッチの変換と逆向きに発話のピッチが変化する。
- H2. フィードバック音声の抑揚を変換すると発話音声が震える。

- H3. フィードバック音声のラウドネスの変換と逆向きに発話のラウドネスが変化する.

H1 については, Burnett ら [11] の研究にて, 話者はフィードバック音声のピッチ変換と逆向きに発話のピッチを調節したと報告されていたことから, 本研究においても同様の傾向がみられると考えた.

H2 については, 抑揚を増加させた場合フィードバック音声のイントネーションが変化するため, 話者が日常的に聞いているイントネーションとの乖離に話者の意識が向き, 部分的に声が震えることが予想される. 抑揚を減少させた場合においても, 発話のイントネーションが変化する点は抑揚増加と共通しているため, 抑揚増加と同様の変化が発生すると考えた.

H3 については, Lane ら [12] の研究にて, フィードバック音声の音量を増加させると発話の音量が減少したと報告されていたことから, 本研究においても同様の傾向がみられると考えた. また, フィードバック音声の音量を減少させた際にも, ピッチ操作と同様の効果として, 発話の音量が増加すると考えた.

4.3 実験参加者

実験参加者は計 18 名 (男性 15 名, 女性 3 名, 平均 22.2 歳, 標準偏差 1.18 歳) の大学生および大学院生である. 実験は参加者内配置で行われ, 各参加者が 6 条件すべてを実施した. 実験の所要時間は 1 時間程度であった. 実験参加者には筑波大学の規定に基づく謝金が支給された.

4.4 実験タスク

最初に, 参加者に対して実験で使用するシステムの音量調整を行った. 音量調整においては, ヘッドフォンから再生される音声の大きさと参加者の骨導音の大きさが, 参加者の主観で同じ大きさになるよう指示した.

その後, 参加者は実験者が用意した原稿を音読した. 原稿を読む速度や抑揚等, 読み方に関する指示は一切行わず, 参加者は原稿を自由に音読した. なお, 本研究は最終的に音声通話等における他者との会話中に使用することを想定しているが, 本実験ではその前段階として音響特性の変換が話者の発話特性に与える影響について詳細に調査することが目的である. そのため, 本実験で行うタスクは先行研究 [7] を参考に音読タスクとした. 原稿は条件毎に異なる内容で, すべて村上春樹の短編小説「ノルウェイの森」および「羊をめぐる冒険」から 900 字程度を抜粋したものである. 原稿および条件は対応しており, 条件の実施順序はランダム化した.

参加者が原稿を音読している間, 実験者はパラメータ変換の反映量を操作した. 原稿を文字数で 3 分割し, 最初の 300 字はパラメータ変換を行わず, 続く 300 字でパラメータ変換の効果を最大の状態まで線形に増加させ, 最後の 300 字ではパラメータ変換の効果が最大状態で固定した.

それぞれの条件に対して原稿の音読およびタスクに対するアンケートを行い，すべての条件が終了した後に実験全体に関する半構造化インタビューを行った．アンケートの具体的な内容を付録 A に記載する．

4.5 実験環境

ヘッドフォン（SONY WH-1000XM4）およびモニタ（BenQ GW2480T）を接続したラップトップコンピュータ（CPU: Intel Core i7-9750H, GPU: GeForce GTX 1650, メモリ: 16GB）を用いて実験を行った．ラップトップコンピュータはスクリーンを消灯した状態で実験参加者の正面に配置した．実験者は，参加者の視界に入らない位置に配置したモニタでシステムの監視および操作を行った．参加者はヘッドフォンを装着し，ラップトップコンピュータ内蔵のマイクに向かって声を出した．実験環境を図 4.1 に示す．



図 4.1: 実験環境

4.6 評価指標

評価指標として，実験参加者の発話音声に含まれる以下の 4 つのパラメータを用いる．

- 基本周波数 (f_0)
- ジッタ (PPQ5)
- シマ (APQ5)
- ラウドネスレベル

このうち基本周波数，ジッタ，およびシマは，Jean らの先行研究 [7] で用いられていた評価指標のうち，発話の音響特性に関連する項目として選出された．また，ラウドネスレベルは，特徴量抽出に用いた Python のライブラリである Surfboard[31] において抽出可能な特徴量から選出された．

4.7 実験結果

本節では、実験の各評価指標についての結果を示す。各評価指標について、パラメータ変換を行わない状態および変換効果が最大の状態の音声から特徴量を抽出し、比較および分析を行った。特徴量抽出後に等分散性および正規性を確認した。等分散性の検定には F 検定を、正規性の検定にはシャピロウィルク検定を行った。その後、正規性が認められた場合は対応のある t 検定を行った。正規性が認められなかった場合はウィルコクソンの符号順位検定を行った。また、検定において p 値が 0.05 未満であることを統計的に有意とみなした。なお、統計分析には R[32] を用いた。

4.7.1 基本周波数

パラメータの変換前および変換後における、抽出された基本周波数を条件ごとに箱ひげ図としてまとめた結果を図 4.2 に示す。ウィルコクソンの符号順位検定を行った結果、すべての条件において有意差が見られなかった ($c1: Z = 1.43, p = 0.15$, $c2: Z = 0.08, p = 0.93$, $c3: Z = 1.43, p = 0.15$, $c4: Z = 1.12, p = 0.26$, $c5: Z = 0.38, p = 0.70$, $c6: Z = 1.07, p = 0.28$)。

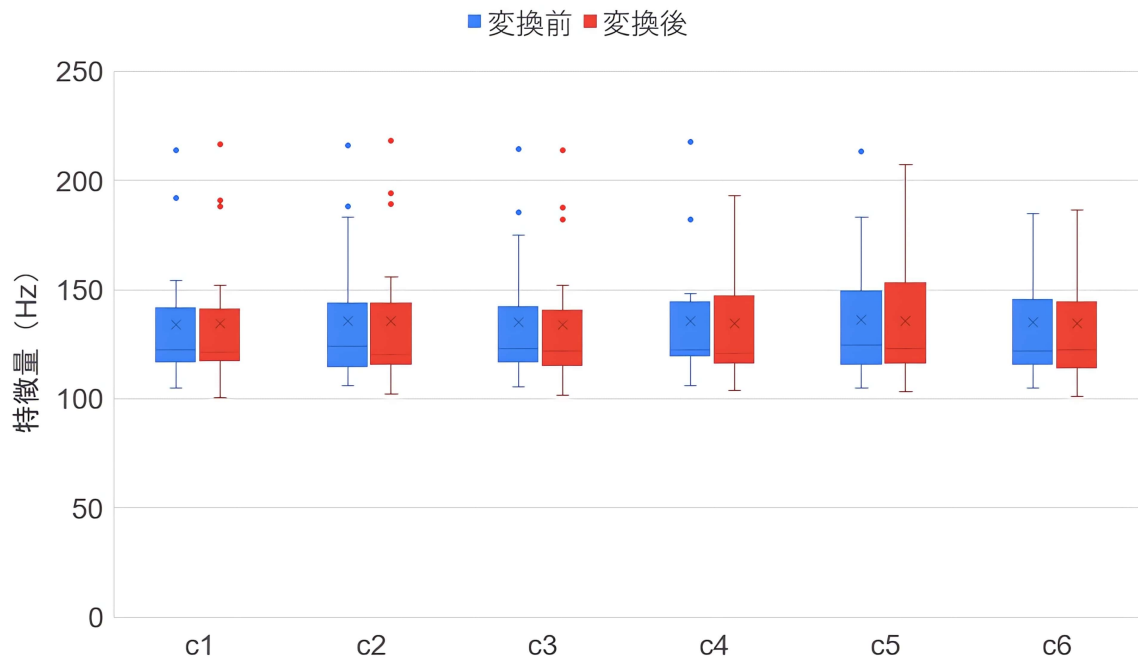


図 4.2: パラメータ変換前および変換後の基本周波数.

4.7.2 ジッタ

パラメータの変換前および変換後における，抽出されたジッタを条件ごとに箱ひげ図としてまとめた結果を図 4.3 に示す．対応のある t 検定を行った結果，c5 において有意差が見られた ($c1:t(17) = 1.61, p = 0.13$, $c2:t(17) = -0.41, p = 0.68$, $c3:t(17) = 1.58, p = 0.13$, $c4:t(17) = 0.58, p = 0.57$, $c5:t(17) = 2.83, p = 0.01$, $d = 0.67$, $c6:t(17) = 0.42, p = 0.68$)．

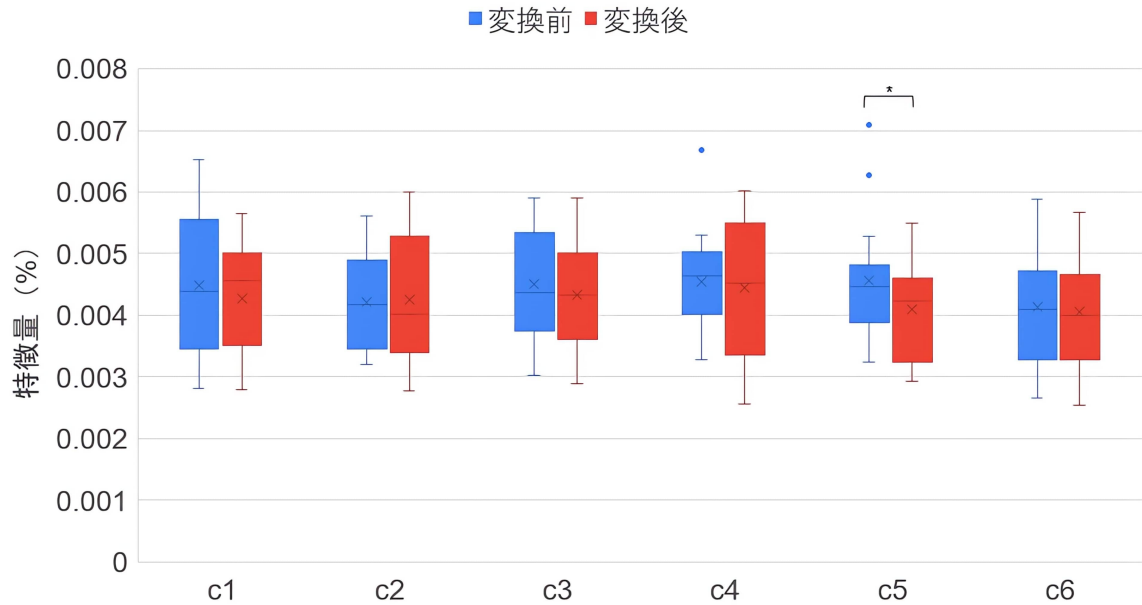


図 4.3: パラメータ変換前および変換後のジッタ．有意差が出た条件に*を示す．

4.7.3 シマ

パラメータの変換前および変換後における，抽出されたシマを条件ごとに箱ひげ図としてまとめた結果を図 4.4 に示す．対応のある t 検定を行った結果，c3, c5, および c6 において有意差が見られた ($c1:t(17) = -1.33, p = 0.20$, $c2:t(17) = -1.72, p = 0.10$, $c3:t(17) = -2.28, p = 0.03, d = 0.53$, $c4:t(17) = 0.43, p = 0.67$, $c5:t(17) = 2.49, p = 0.02, d = 0.59$, $c6:t(17) = -2.81, p = 0.01, d = 0.66$)．

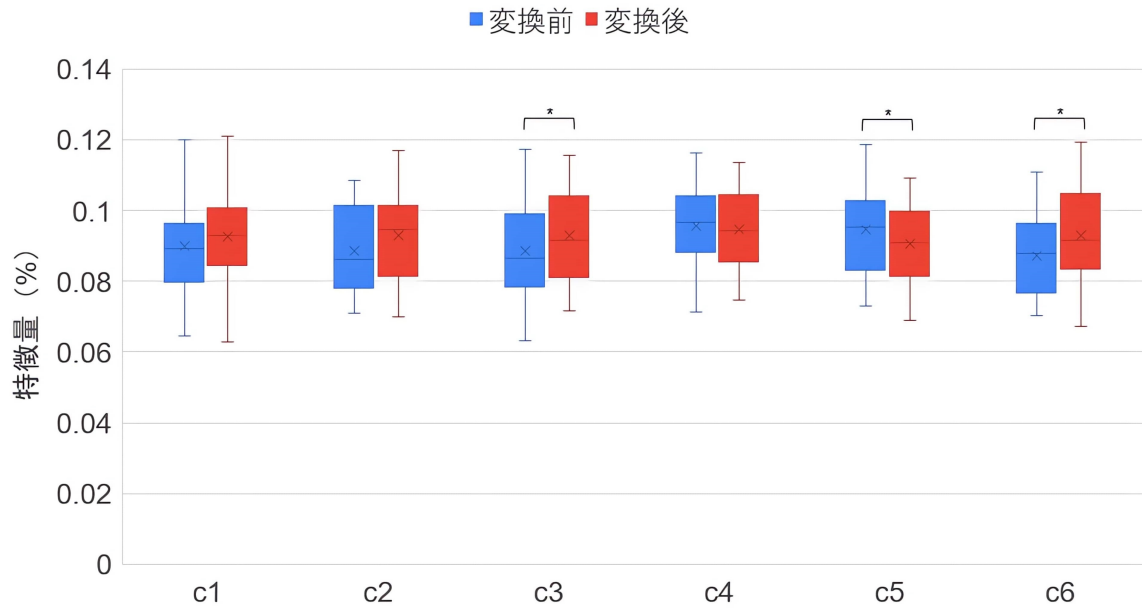


図 4.4: パラメータ変換前および変換後のシマ. 有意差が出た条件に*を示す.

4.7.4 ラウドネスレベル

パラメータの変換前および変換後における, 抽出されたラウドネスレベルを条件ごとに箱ひげ図としてまとめた結果を図 4.5 に示す. 正規性が認められた c1, c4, c5, および c6 にて対応のある t 検定を行った結果, c5 において有意差が見られた ($c1:t(17) = -0.88, p = 0.39$, $c4:t(17) = -0.07, p = 0.94$, $c5:t(17) = -4.26, p = 0.0005, d = 1.00$, $c6:t(17) = 1.47, p = 0.16$). また, 正規性が認められなかった c2 および c3 にてウィルコクソンの符号順位検定を行った結果, それぞれの条件において有意差が見られなかった ($c2:Z = 1.29, p = 0.20$, $c3:Z = 0.90, p = 0.37$).

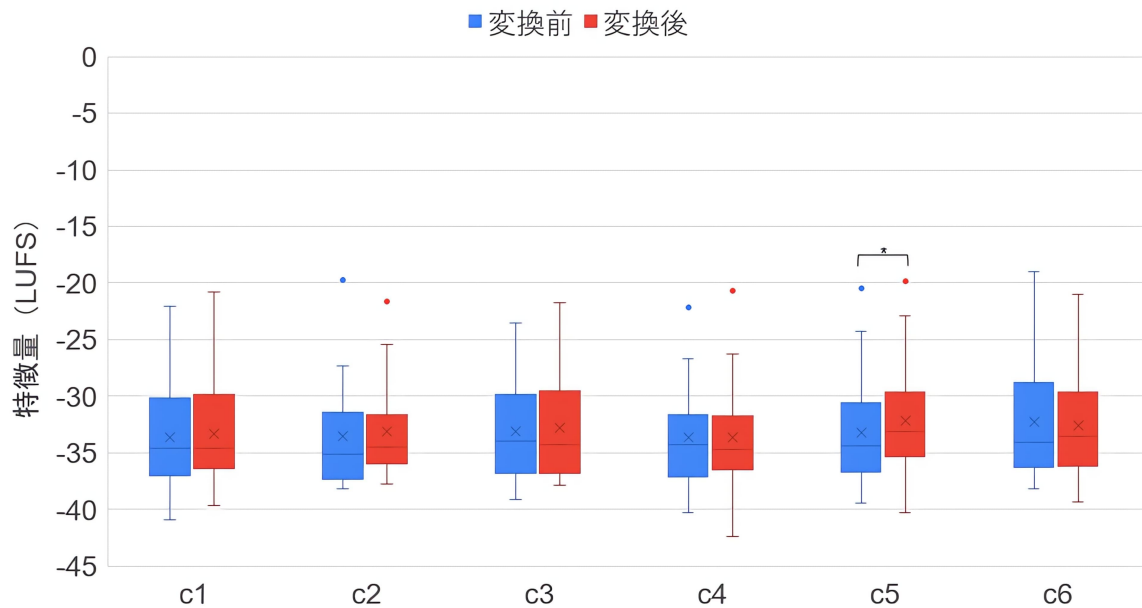


図 4.5: パラメータ変換前および変換後のラウドネスレベル。有意差が出た条件に*を示す。

4.8 議論および考察

本節では、4.2 節で設定した仮説の検証を行う。その後、4.7 節で得られた一部の結果に対して議論を行う。

4.8.1 仮説の検証

「H1. フィードバック音声のピッチの変換と逆向きに発話のピッチが変化する。」について、c1 および c2 において基本周波数に有意差は見られなかった。したがって H1 は支持されなかった。半構造化インタビューにて「自分の声を聞くことに慣れておらず、原稿を音読するだけで精一杯だった」(P1, P13, P14) という意見を得た。そのため、実験にトレーニング条件を追加し自分の声を聞きながら発話することに慣れた状態でタスクを実施することで、結果に変化が見られる可能性がある。

「H2. フィードバック音声の抑揚を変換すると発話音声が震える。」について、c3 におけるシマのみ有意に増加した。したがって H2 は部分的に支持された。しかし、シマの増加が抑揚の変化ではなくラウドネスレベルの変化に起因する可能性がある。したがって、H2 の検証のために今後さらなる調査が必要である。

「H3. フィードバック音声のラウドネスの変換と逆向きに発話のラウドネスが変化する。」について、c5 においてラウドネスレベルが有意に増加した。また、c5 以外においてラウドネスレベルに有意差は見られなかった。したがって H3 は支持されなかった。半構造化インタ

ビューにて、「遅延があるため骨導音を集中して聞いていたが、スピーカ側の音が大きくなったため話しにくかった」(P7)という意見を得た。この意見から、参加者が話しにくさを解消するために骨導音を大きくすることを試み、発話のラウドネスレベルが増加したと考えられる。そのため、遅延を短縮することで結果に変化が見られる可能性があり、システム設計を見直したうえでの再調査が必要である。

4.8.2 ジッタ・シマおよびラウドネスレベルの関係

実験の結果、c5においてジッタおよびシマの減少とともにラウドネスレベルの増加が確認された。ジッタおよびラウドネスレベル、シマおよびラウドネスレベルの間にそれぞれ負の相関があることが報告されており [33]、本実験においても報告を再現したものとして考えられる。なお、報告で分析の対象とされていたのはラウドネスレベル (LKFS) ではなく音圧レベル (dB) であるが、LKFS および dB は同等のものとして扱うことが可能である [34]。

一方、c6においてシマのみが有意に増加している。これはシマおよびラウドネスレベルの相関に反する結果である。この原因を考察するためには情報が不足しており、シマおよびラウドネスレベルの関係や両者に影響を与える要因についての詳細な調査が必要である。

4.8.3 フィードバックの遅延

半構造化インタビューの中で一部の参加者から、「システムで再生される音声の遅れにより発話が阻害された」(P2, P6, P9, P12, P18)という意見を得た。聴覚フィードバックにおいてフィードバック音声に 30 ms から 300 ms 程度の遅延が存在する場合、発話の流暢性が低下することが報告されている [22, 23, 24]。本システムにおける遅延は、3.1 節にある通り著者の環境で 120 ms 程度であり、遅延が原因となって参加者の発話が阻害された可能性がある。今後はシステムの遅延を、発話の阻害が発生しない程度に短縮する必要がある。

第5章 実験2

実験1の結果から、システムおよび実験設計に課題が見られた。そのため、システムおよび実験設計にいくつかの改善を行い、実験1と同様の実験を行った。本章では新たに行った実験における改善点、実験条件、仮説、実験参加者、実験タスク、実験環境、評価指標、実験結果、および考察を述べる。なお、本実験は筑波大学システム情報系の倫理審査委員会の承認を受けて実施した。

5.1 改善点

4.8節の考察を踏まえて、システムおよび実験設計に3つの改善を行った。本節では具体的な改善内容について述べる。

5.1.1 システムの遅延短縮

実験1ではサウンドの入出力およびオーディオドライバにラップトップコンピュータ内蔵のものを用いており、これが主な遅延の原因であった。そのため、新たなシステムではサウンドの入力にコンデンサマイクを、オーディオドライバ含むサウンド入出力の管理にオーディオインタフェースをそれぞれ用いた。

また、実験1ではフィードバック音声録音のために仮想オーディオケーブルをシステムに組み込んでおり、これも遅延の原因であった。そのため新たなシステムでは仮想オーディオケーブルを外し、DAVIDから直接フィードバック音声を録音する設計に変更した。新たなシステムの構成を図5.1に示す。

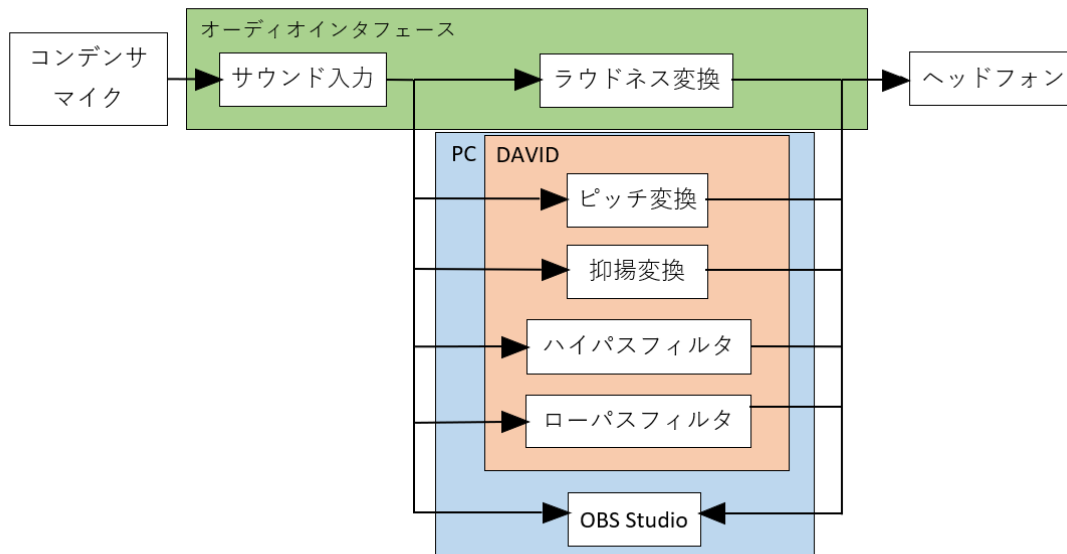


図 5.1: 改善後のシステム構成図

これらの改善を行った結果、著者の環境における遅延が 120 ms 程度から 23 ms 程度に短縮された。この値は 4.8.3 項にて示した発話の流暢性が低下する遅延の範囲よりも小さいため、新たなシステムにおいて遅延による話者の発話阻害が発生する可能性は低いと考えられる。

5.1.2 評価指標の見直し

実験 1 では 4.6 節に示すように、先行研究および特徴量抽出に用いた Python ライブラリから評価指標を選出した。本実験では実験 1 の方針に従いつつ、音の三要素を考慮して評価指標として用いるパラメータを再検討した。

本実験では、実験 1 にて用いた基本周波数 (f_0)、ジッタ (PPQ5)、シマ (APQ5)、およびラウドネスレベルに加えて、新たに音声信号の正負が切り替わる割合を表す ZCR (Zero Crossing Rate)、調波成分および雑音成分のエネルギー比を表す HNR (Harmonic to Noise Ratio)、および単位面積あたりに音波がなす仕事率を表す音響インテンシティを追加した。このうち ZCR および HNR については、発話の囁声度の指標として採用した [35, 36]。また、音響インテンシティについては、ラウドネスレベルが心理量である [37] ことから、ラウドネスレベルに対応する物理量として採用した。

なお、上述した音の三要素について、音の高さには基本周波数 (f_0) が対応する。また、音色にはジッタ (PPQ5)、シマ (APQ5)、ZCR、および HNR が対応する。音の大きさにはラウドネスレベルおよび音響インテンシティが対応する。

5.1.3 トレーニング条件の追加

実験 1 では、自分の声を聞きながら音読を行うというタスク内容により、参加者の経験が実験結果に影響した可能性があった。そのため本実験ではすべての条件に先駆けて音響特性の変換およびデータの計測を行わないトレーニング条件を追加した。これにより実験結果が参加者の音読経験に影響されにくくなると考えられる。

5.2 実験条件

実験条件は以下の 8 条件で、すべて参加者内配置で設計した。

- c1. ピッチ上昇
- c2. ピッチ下降
- c3. 抑揚増加
- c4. 抑揚減少
- c5. ラウドネス増加
- c6. ラウドネス減少
- c7. ハイパスフィルタ
- c8. ローパスフィルタ

すべての条件において、実験者は参加者のタスク実施中に音響特性のパラメータ変換を行う。

c1, c2, c3, および c4 については実験 1 と同様の操作を行う。c5 および c6 についてはラップトップコンピュータのサウンド出力ボリュームの代わりにオーディオインタフェースのボリュームを操作する。なお、変換量については実験 1 と同様に、著者の主観に基づき通常時から音量が 50% 程度増減するように設定された。

c7 および c8 については、ハイパスフィルタおよびローパスフィルタ単体の AAF における影響調査が行われていないため、新たに追加した。フィードバック音声の周波数帯を一部削減することで、フィードバック音声の変化による違和感を軽減できると考えられる。変換量については、著者の主観に基づき変化が知覚できる程度として、声の第 1 フォルマントおよび第 2 フォルマントの値域の上方 30% および下方 30% を通過させるように設定された。なお、フォルマントの値域の関係から、フィルタは男性および女性で別に実装された。男性用のハイパスフィルタを図 5.2 に、男性用のローパスフィルタを図 5.3 に、女性用のハイパスフィルタを図 5.4 に、女性用のローパスフィルタを図 5.5 に示す。

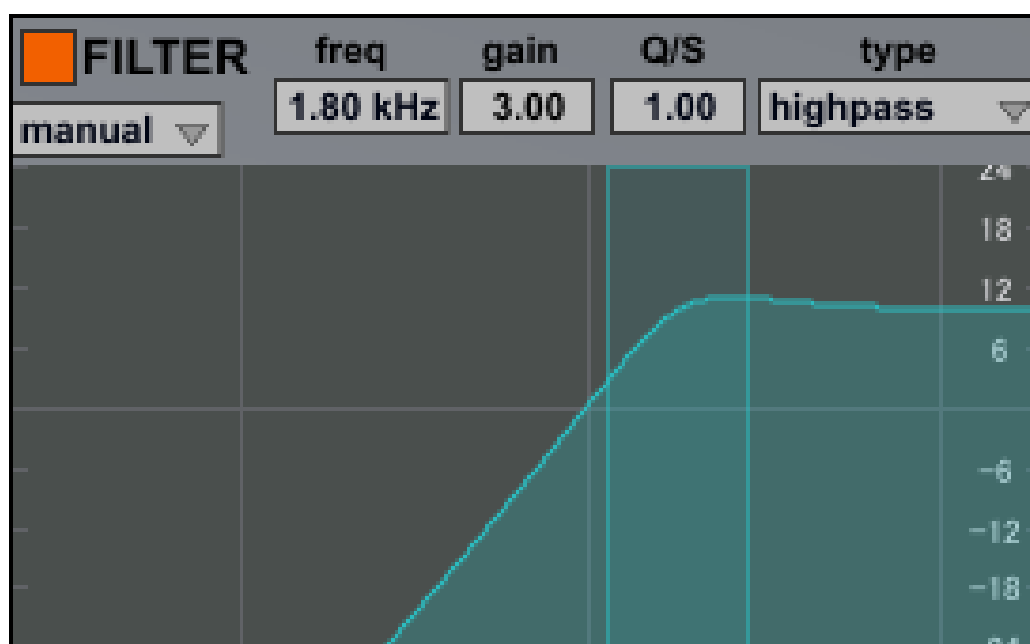


図 5.2: 男性用ハイパスフィルタ

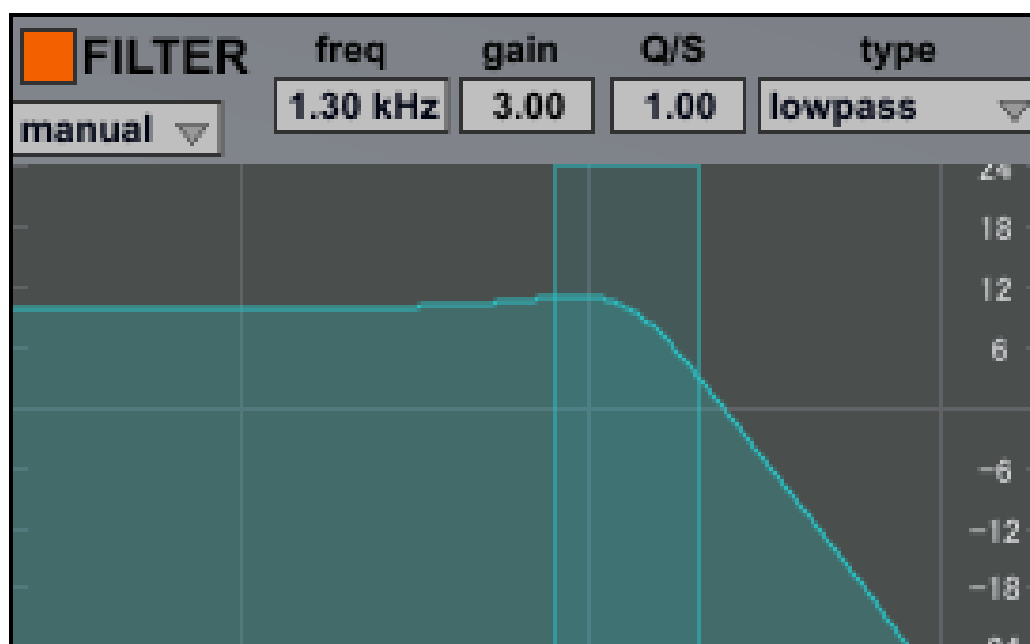


図 5.3: 男性用ローパスフィルタ

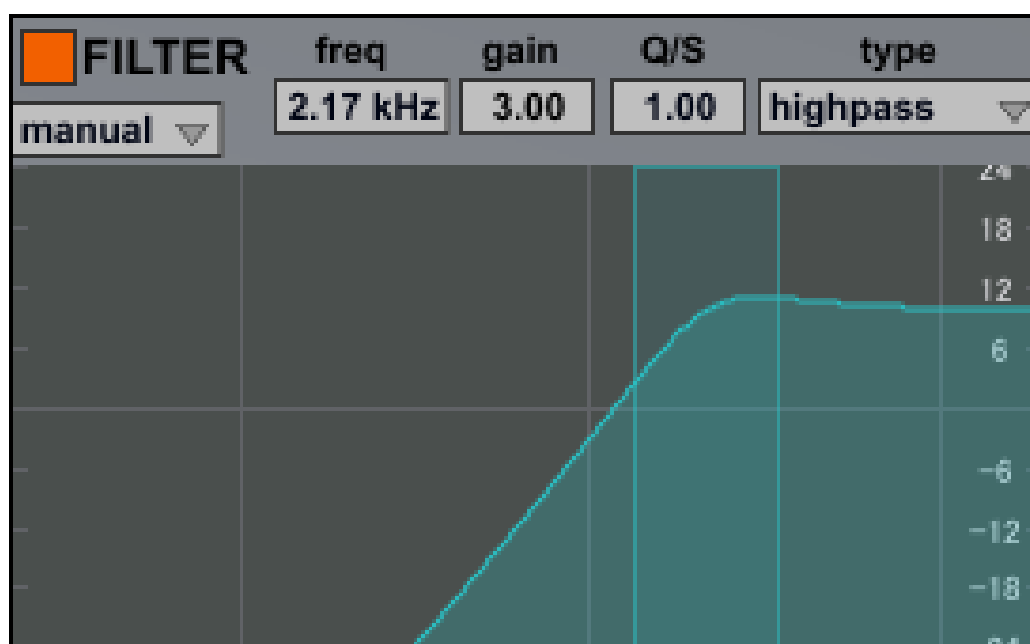


図 5.4: 女性用ハイパスフィルタ

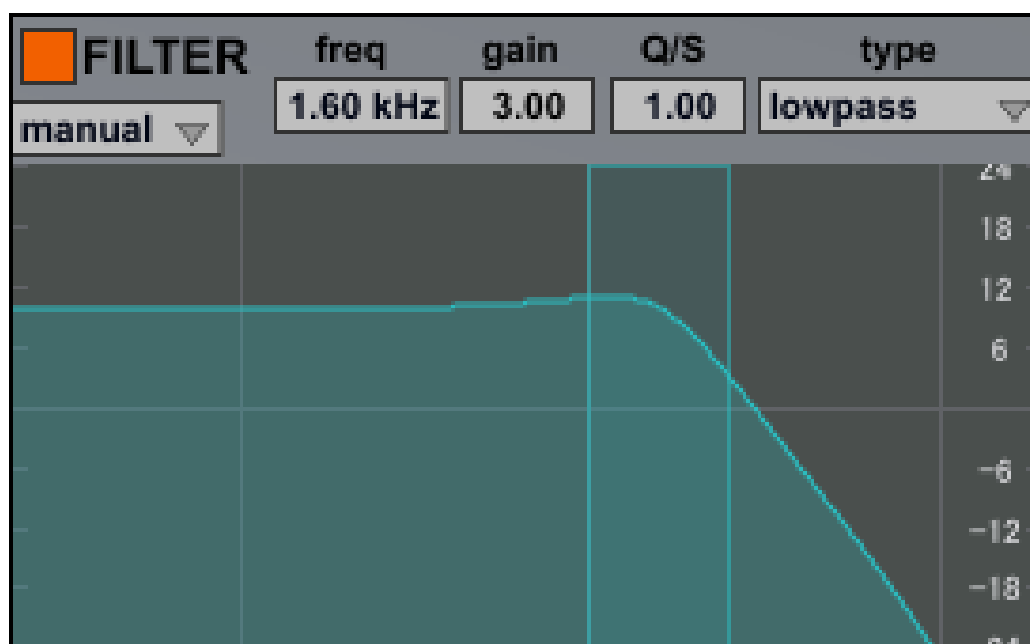


図 5.5: 女性用ローパスフィルタ

5.3 仮説

本実験では、以下の5点の仮説を設定した。このうちH1、H2、およびH3については実験1と同様である。

- H1. フィードバック音声のピッチの変換と逆向きに発話のピッチが変化する。
- H2. フィードバック音声の抑揚を変換すると発話音声が震える。
- H3. フィードバック音声のラウドネスの変換と逆向きに発話のラウドネスが変化する。
- H4. ハイパスフィルタを通して音声をフィードバックすると発話のラウドネスが増加し、発話が明瞭になる。
- H5. ローパスフィルタを通して音声をフィードバックすると発話が明瞭になる。

H4については、ハイパスフィルタを通した音声および通さない音声を比較した際の差分と逆の変化が発生すると考えた。著者1人を対象とした予備調査の結果、ハイパスフィルタを通したフィードバック音声は元の音声と比較してラウドネスが減少し、発話の明瞭度が減少した。したがってハイパスフィルタを通したフィードバック音声を聞いた話者の発話では、ラウドネスの増加および発話の明瞭化が確認できると考えた。

H5についてもH4と同様に仮説を設定した。著者1人を対象とした予備調査の結果、ローパスフィルタを通したフィードバック音声は元の音声と比較して発話の明瞭度が減少した。したがってローパスフィルタを通したフィードバック音声を聞いた話者の発話では、発話の明瞭化が確認できると考えた。

5.4 実験参加者

実験参加者は計18名（男性17名、女性1名、平均22.7歳、標準偏差0.93歳）の大学生および大学院生である。このうち13名は実験1と同じ参加者であった。実験は参加者内配置で行われ、各参加者が8条件すべてを実施した。実験の所要時間は1時間程度であった。実験参加者には筑波大学の規定に基づく謝金が支給された。

5.5 実験タスク

最初に参加者はシステムの音量調整を行い、トレーニング条件を実施した。その後実験1と同様に条件毎に原稿の音読を行いアンケートに回答、すべての条件が終了した後に半構造化インタビューに回答した。アンケートの具体的な内容を付録Bに記載する。実験者も実験1と同様に、参加者の音読中にパラメータ変換の反映量を操作した。

実験1との違いとして、原稿と条件を一对一で対応させずに実験を行った。原稿は条件毎に異なる内容で、すべて村上春樹の短編小説「1Q84」、「街とその不確かな壁」、「海辺のカフ

カ」,「世界の終りとハードボイルド・ワンダーランド」, および「スプートニクの恋人」から900字程度を抜粋したものである。原稿および条件の実施順序はそれぞれランダム化した。

5.6 実験環境

オーディオインタフェース (NATIVE INSTRUMENTS KOMplete Audio 6) を接続したラップトップコンピュータ (CPU: Intel Core i7-9750H, GPU: GeForce GTX 1650, メモリ: 16GB) を用いて実験を行った。オーディオインタフェースにはコンデンサマイク (audio-technica AT2050) およびヘッドホン (SONY WH-1000XM4) を接続した。実験者は、参加者の視界に入らない位置に配置したオーディオインタフェースおよびラップトップコンピュータでシステムの監視および操作を行った。参加者はヘッドホンを装着し、コンデンサマイクに向かって声を出した。実験環境を図 5.6 に示す。



図 5.6: 実験環境

5.7 評価指標

評価指標として、実験参加者の発話音声に含まれる以下の7つのパラメータを用いる。

- 基本周波数 (f0)
- ジッタ (PPQ5)
- シマ (APQ5)
- ラウドネスレベル
- ZCR (Zero Crossing Rate)
- 音響インテンシティ

- HNR (Harmonic to Noise Ratio)

なお、各パラメータの選出理由については、5.1.2 項の通りである。

5.8 実験結果

本節では、実験の各評価指標についての結果を示す。実験 1 と同様に、音声から特徴量を抽出した後 F 検定およびシャピロウィルク検定を行い、等分散性および正規性を確認した。正規性が認められた場合は対応のある t 検定を、正規性が認められなかった場合はウィルコクソンの符号順位検定をそれぞれ行った。なお、検定において p 値が 0.05 未満であることを統計的に有意とみなした。

5.8.1 基本周波数

パラメータの変換前および変換後における、抽出された基本周波数を条件毎に箱ひげ図としてまとめた結果を図 5.7 に示す。ウィルコクソンの符号順位検定を行った結果、すべての条件において有意差が見られなかった ($c1:Z = 1.12, p = 0.26$, $c2:Z = 0.04, p = 0.97$, $c3:Z = 0.51, p = 0.61$, $c4:Z = 1.25, p = 0.21$, $c5:Z = 0.38, p = 0.70$, $c6:Z = 0.51, p = 0.61$, $c7:Z = 0, p = 1$, $c8:Z = 0.17, p = 0.87$)。

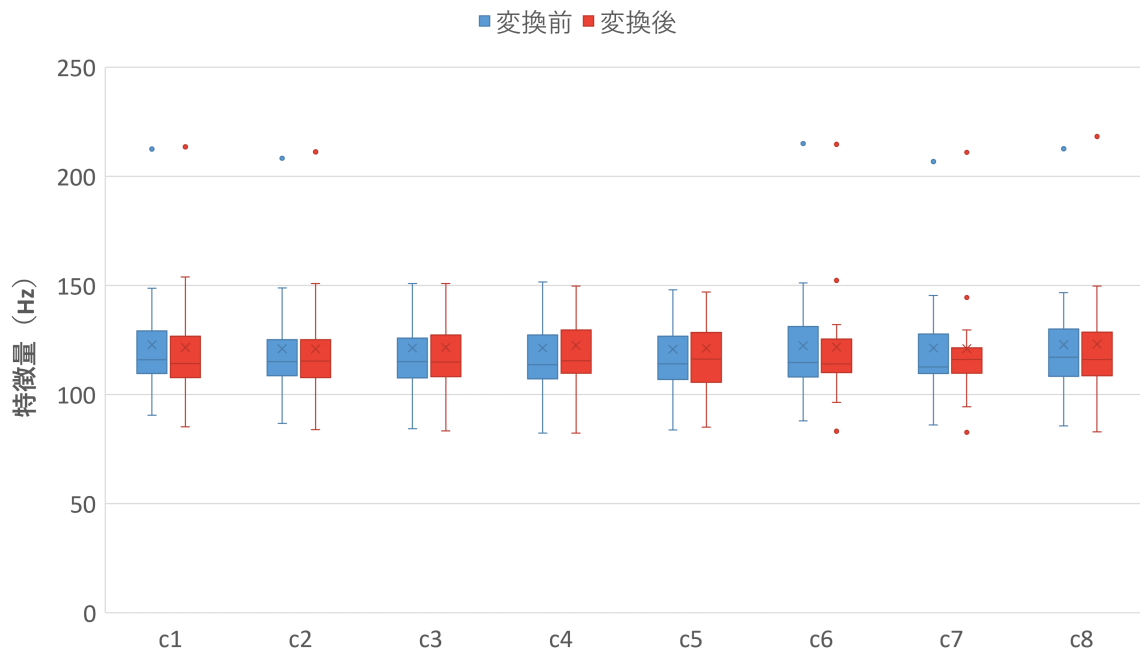


図 5.7: パラメータ変換前および変換後の基本周波数.

5.8.2 ジッタ

パラメータの変換前および変換後における，抽出されたジッタを条件毎に箱ひげ図としてまとめた結果を図 5.8 に示す．対応のある t 検定を行った結果，c1 において有意差が見られた ($c1:t(17) = 2.59, p = 0.02, d = 0.61$, $c2:t(17) = -0.30, p = 0.76$, $c3:t(17) = 0.41, p = 0.68$, $c4:t(17) = 0.91, p = 0.37$, $c5:t(17) = 0.12, p = 0.91$, $c6:t(17) = 0.92, p = 0.37$, $c7:t(17) = -0.91, p = 0.37$, $c8:t(17) = 0.18, p = 0.86$)．

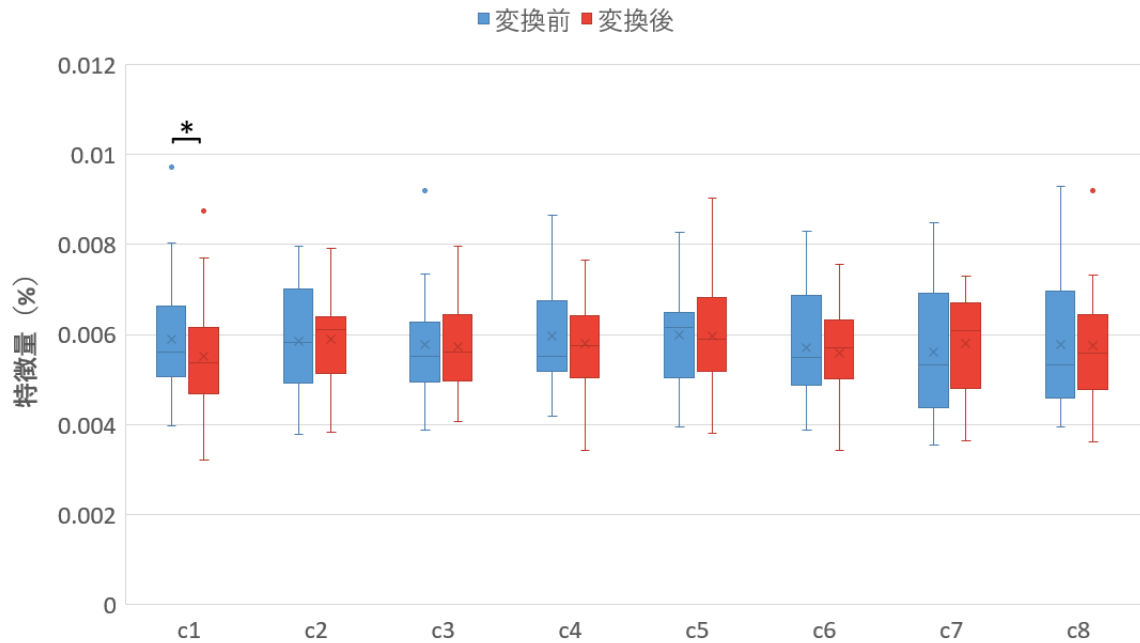


図 5.8: パラメータ変換前および変換後のジッタ．有意差が出た条件に*を示す．

5.8.3 シマ

パラメータの変換前および変換後における，抽出されたシマを条件毎に箱ひげ図としてまとめた結果を図 5.9 に示す．対応のある t 検定を行った結果，c5 および c6 において有意差が見られた ($c1:t(17) = 1.83, p = 0.08$, $c2:t(17) = -1.13, p = 0.27$, $c3:t(17) = 0.83, p = 0.42$, $c4:t(17) = -0.39, p = 0.70$, $c5:t(17) = -2.43, p = 0.03, d = 0.57$, $c6:t(17) = -3.39, p = 0.004, d = 0.80$, $c7:t(17) = 0.59, p = 0.56$, $c8:t(17) = -0.20, p = 0.84$)．

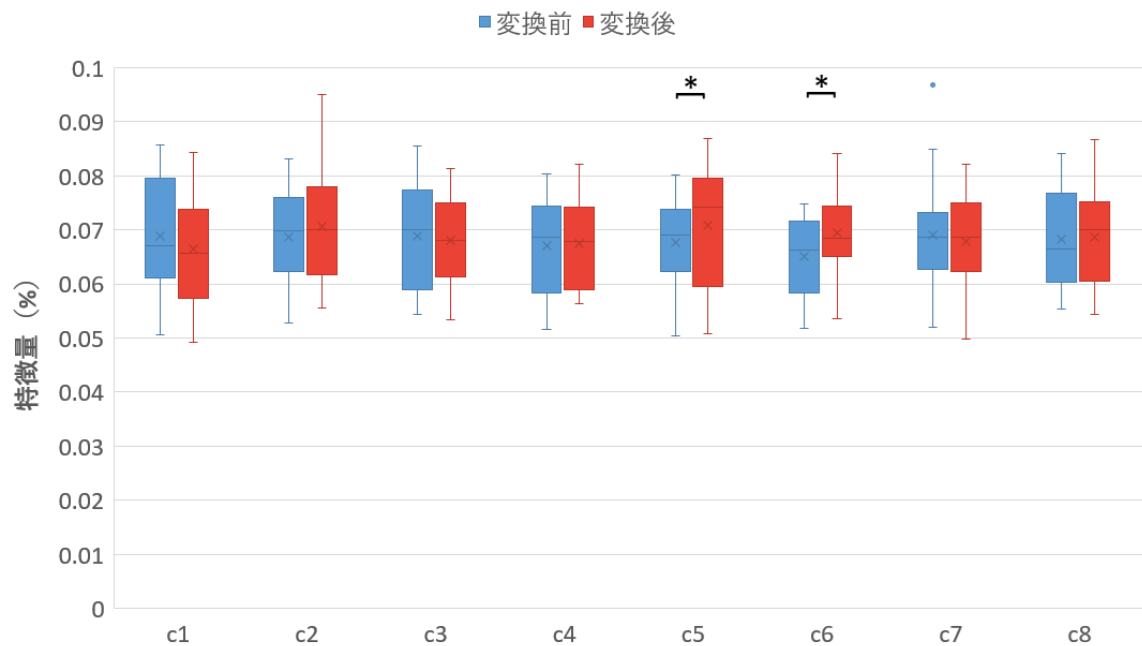


図 5.9: パラメータ変換前および変換後のシマ. 有意差が出た条件に*を示す.

5.8.4 ラウドネスレベル

パラメータの変換前および変換後における, 抽出されたラウドネスレベルを条件毎に箱ひげ図としてまとめた結果を図 5.10 に示す. 対応のある t 検定を行った結果, すべての条件において有意差が見られなかった ($c1:t(17) = 0.25, p = 0.80$, $c2:t(17) = -0.32, p = 0.76$, $c3:t(17) = 0.49, p = 0.63$, $c4:t(17) = -0.60, p = 0.56$, $c5:t(17) = -0.64, p = 0.53$, $c6:t(17) = -0.04, p = 0.97$, $c7:t(17) = -0.008, p = 0.99$, $c8:t(17) = -1.83, p = 0.09$).

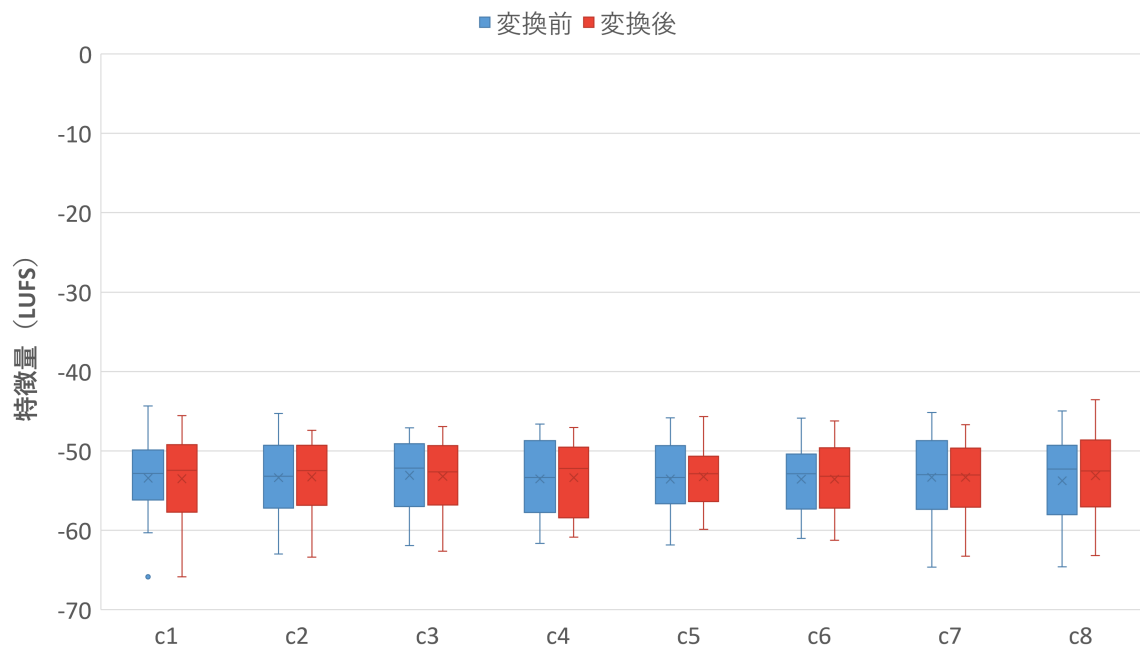


図 5.10: パラメータ変換前および変換後のラウドネスレベル.

5.8.5 ZCR

パラメータの変換前および変換後における、抽出された ZCR を条件毎に箱ひげ図としてまとめた結果を図 5.11 に示す. 正規性が認められた c1, c3, c4, c6, c7, および c8 にてにて対応のある t 検定を行った結果, それぞれの条件において有意差が見られなかった ($c1:t(17) = 0.84, p = 0.41$, $c3:t(17) = 0.29, p = 0.77$, $c4:t(17) = 0.72, p = 0.48$, $c6:t(17) = -1.59, p = 0.13$, $c7:t(17) = -0.81, p = 0.43$, $c8:t(17) = -0.55, p = 0.59$). また, 正規性が認められなかった c2 および c5 にてウィルコクソンの符号順位検定を行った結果, c2 において有意差が見られた ($c2:Z = 2.72, p = 0.007, r = 0.64$, $c5:Z = 0.60, p = 0.55$).

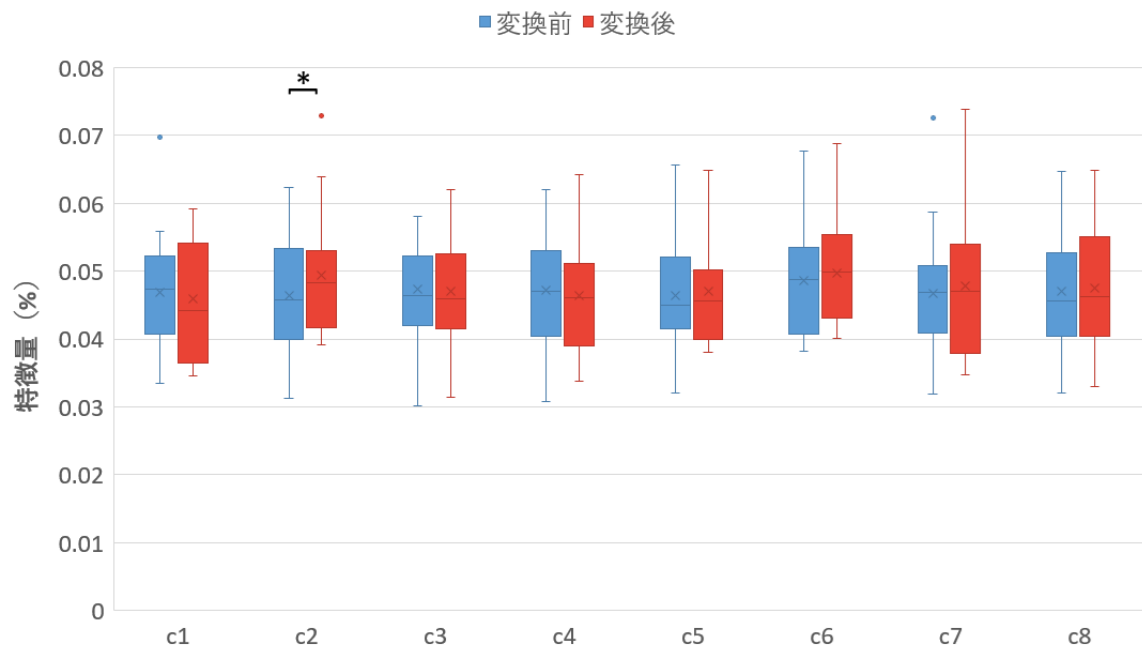


図 5.11: パラメータ変換前および変換後の ZCR. 有意差が出た条件に*を示す.

5.8.6 音響インテンシティ

パラメータの変換前および変換後における、抽出された音響インテンシティを条件毎に箱ひげ図としてまとめた結果を図 5.12 に示す. ウィルコクソンの符号順位検定を行った結果, すべての条件において有意差が見られなかった (c1: $Z = 0.17, p = 0.87$, c2: $Z = 0.13, p = 0.90$, c3: $Z = 1.95, p = 0.05$, c4: $Z = 0.38, p = 0.70$, c5: $Z = 0.55, p = 0.58$, c6: $Z = 0.26, p = 0.80$, c7: $Z = 0.43, p = 0.67$, c8: $Z = 1.56, p = 0.12$).

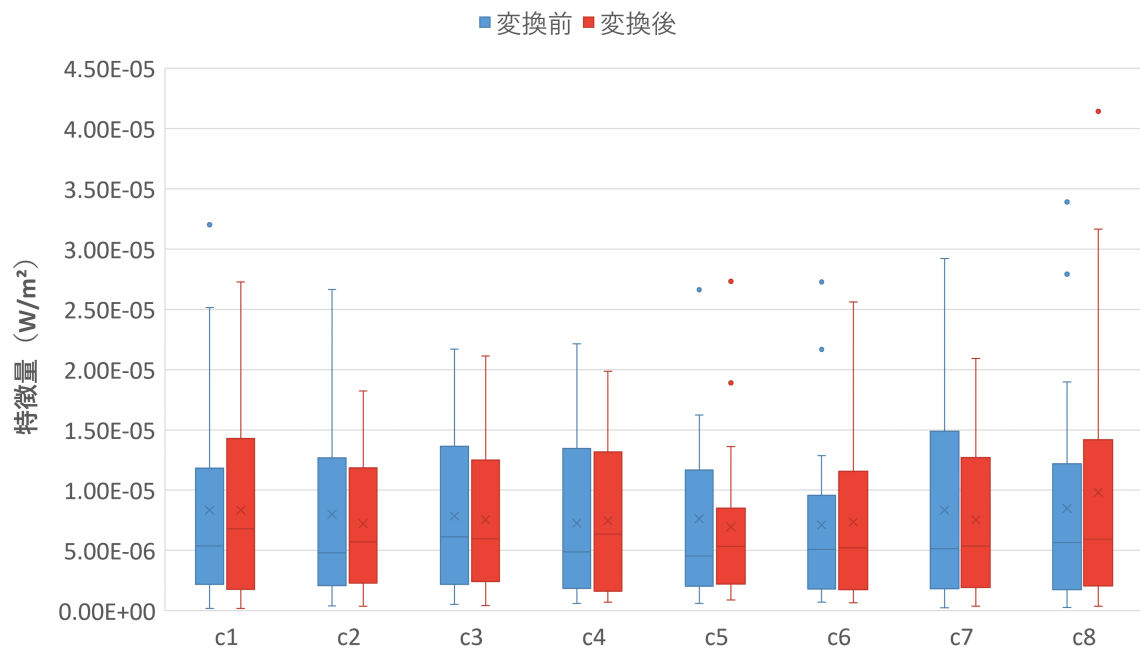


図 5.12: パラメータ変換前および変換後の音響インテンシティ.

5.8.7 HNR

パラメータの変換前および変換後における、抽出された HNR を条件毎に箱ひげ図としてまとめた結果を図 5.13 に示す. 正規性が認められた c1, c2, c5, c6, c7, および c8 にて対応のある t 検定を行った結果, それぞれの条件において有意差が見られなかった ($c1:t(17) = -2.10, p = 0.05$, $c2:t(17) = -1.14, p = 0.27$, $c5:t(17) = -1.72, p = 0.10$, $c6:t(17) = -0.19, p = 0.85$, $c7:t(17) = -0.26, p = 0.80$, $c8:t(17) = -1.70, p = 0.11$). また, 正規性が認められなかった c3 および c4 にてウィルコクソンの符号順位検定を行った結果, c3 において有意差が見られた ($c3:Z = 2.99, p = 0.003, r = 0.70$, $c4:Z = 1.47, p = 0.14$).

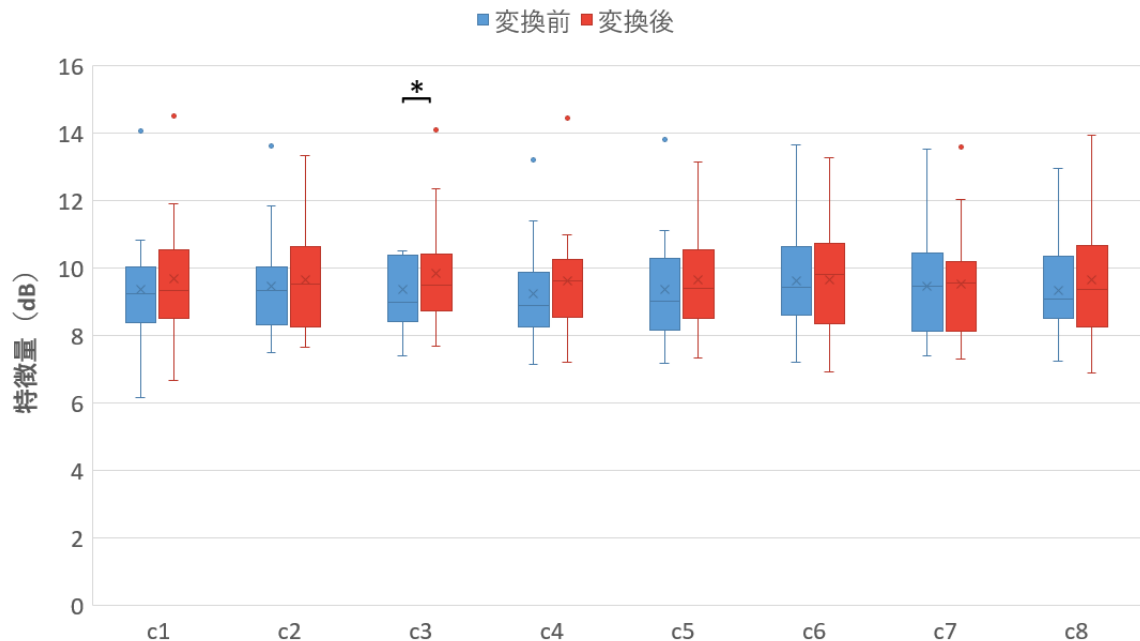


図 5.13: パラメータ変換前および変換後の HNR. 有意差が出た条件に*を示す.

5.9 議論および考察

本節では、5.3 節で設定した仮説の検証を行う。その後、5.8 節で得られた一部の結果に対して議論を行う。

5.9.1 仮説の検証

「H1. フィードバック音声のピッチの変換と逆向きに発話のピッチが変化する。」について、c1 および c2 において基本周波数に有意差は見られなかった。したがって H1 は支持されなかった。AAF により発生した話しにくさが結果に影響したと考えられる。5.9.2 項にて詳しい議論を行う。

「H2. フィードバック音声の抑揚を変換すると発話音声が震える。」について、c3 および c4 においてジッタおよびシマに有意差は見られなかった。したがって H2 は支持されなかった。本実験では抑揚の変換量を著者の主観に基づき設定したため、変換量を変えることで結果に変化が見られると考えられる。

「H3. フィードバック音声のラウドネスの変換と逆向きに発話のラウドネスが変化する。」について、c5 および c6 においてラウドネスレベルおよび音響インテンシティに有意差は見られなかった。したがって H3 は支持されなかった。実験 1 において、発話のラウドネスレベル増加は、フィードバック音声の遅延に起因する発話障害への対応の結果であると考察した。

また、本実験で実施した半構造化インタビューにおいて、実験1にも参加していた参加者より「遅延を感じなくなり、話しやすかった」(P3, P7, P9, P16)という意見を得た。遅延短縮の結果、発話阻害が改善されたことにより、実験1にて見られた発話のラウドネスレベルの変化が生じなかったと考えられる。

「H4. ハイパスフィルタを通して音声をフィードバックすると発話のラウドネスが増加し、発話が明瞭になる。」について、c7においてラウドネスレベルおよび音響インテンシティに有意差は見られなかった。また、ジッタ、シマ、ZCR および HNR についても同様に有意差は見られなかった。したがって H4 は支持されなかった。実験で用いるハイパスフィルタの周波数通過帯域を変更することで、話者の発話特性に変化が発生する可能性がある。そのため、AAFにて発話特性を変化させるハイパスフィルタの設定について、さらなる調査が必要であると考えられる。

「H5. ローパスフィルタを通して音声をフィードバックすると発話が明瞭になる。」について、c8においてジッタ、シマ、ZCR および HNR に有意差は見られなかった。したがって H5 は支持されなかった。半構造化インタビューにて「声が小さく、こもった感じに聞こえて話しにくかった」(P1, P11, P12, P13, P16, P17)という意見を得た。そのため、変換量を適切に変更した上での再調査が望まれる。また、具体的な変換量についても追加の調査が必要であると考えられる。

5.9.2 ピッチ変換の効果

c1 および c2 に該当するフィードバック音声について、半構造化インタビューの中でほとんどの参加者から「声が二重に聞こえた」(P1, P2, P3, P4, P5, P9, P10, P11, P17)「ロボットの話し声のように聞こえた」(P6, P7, P8, P12, P13, P14, P16)という意見を得た。このうち12名の参加者が「フィードバック音声の原因で話しにくく感じた」と報告している(P1, P2, P3, P4, P5, P7, P9, P10, P11, P12, P13, P17)。システムの遅延を短縮した結果、参加者がフィードバック音声と骨導音をほぼ同時に認識し、周波数の違う2つの音声を同時に聞いたと考えられる。これが参加者の感じた話しにくさの原因である可能性が高い。

また、本実験の結果より、c1においてジッタが有意に減少したほか、シマおよびHNRがそれぞれ減少および増加の有意傾向にあった。これらの変化は発話の明瞭化を意味している。したがって、参加者は周波数の違う2つの音声を同時に聞いたことにより感じた話しにくさを解消するため、普段よりも明瞭に話すことを試みた可能性がある。

5.9.3 考察のまとめ

本研究の最終目標は、AAFの特性を活用し、ボイスチェンジャを用いることなく話者の発話特性を変化させるシステムを実現することである。システムの実装にあたり、今回の実験で得られた知見より、以下のような機能の実現が考えられる。

- 話者の発話が不明瞭な場合，AAFで発話のピッチを上昇させることにより，発話のジッタおよびシマを減少させ，HNRを増加させて発話を明瞭にする．

ただし，実際にコミュニケーションで用いる上では，上記の機能のみでは効果が限定的であると考えられる．そのため，AAFを用いた場合の発話特性の変化について，今後より多くの知見を明らかにしていく必要がある．

今回の実験で得られた知見が限定的であった理由として，実験に使用したシステムや評価項目に課題が存在したことが考えられる．6章にて，本研究の制約および今後の課題について述べる．

第6章 本研究の制約および今後の課題

本章では、システムに対する議論を行い、本研究の制約および今後の課題を述べる。

6.1 変換パラメータの選定

3.2節で説明した通り、本研究では変換する音響特性としてピッチ、ラウドネス、抑揚、ハイパスフィルタ、およびローパスフィルタを設定した。しかし、これらのパラメータは DAVID の仕様に大きく依存している。今後はこれら以外の音響特性の調査を想定し、DAVID に依存しないシステムの設計が必要とされる。

また、パラメータの変換量に関しても再考を要する。本研究では抑揚、ラウドネス、ハイパスフィルタ、およびローパスフィルタの変換量を著者の主観で設定した。そのため AAF において発話特性に変化を与える変換量の閾値については考慮されていない。今後は変換パラメータの選定とともに、パラメータの変換量についても精査が求められる。

6.2 評価指標の選定

5.1.2 節で説明した通り、本研究では実験の評価指標として、最終的に発話音声の基本周波数、ジッタ (PPQ5)、シマ (APQ5)、ラウドネスレベル、ZCR (Zero Crossing Rate)、音響インテンシティ、および HNR (Harmonic to Noise Ratio) を用いた。しかし、これらのパラメータが評価指標として必要十分であるか否かは不明である。そのため評価指標に用いる音響特性について精査する必要がある。

6.3 コミュニケーションへの応用

5.9.3 項で説明した通り、本研究の最終目標は AAF を用いて話者の発話特性を変化させるシステムを実装し、会話によるコミュニケーションを支援することである。しかし、今回行った2つの実験はどちらも話者の発話特性に与える影響の調査にとどまっている。今後、変換パラメータおよび評価指標の最適化を行った上で、実際の会話に与える影響についての調査が必要である。

第7章 おわりに

本研究では、AAFにおいて発話の音響特性を個別に変換した際の、話者の発話特性への影響について2回の調査を行った。1回目の調査では、変換する音響特性としてピッチ、ラウドネス、および抑揚を設定し、発話の基本周波数、ジッタ（PPQ5）、シマ（APQ5）、およびラウドネスレベルを分析した。

調査の結果、フィードバック音声の抑揚を増加させると発話のシマが増加した。フィードバック音声のラウドネスを増加させると発話のジッタおよびシマが減少し、ラウドネスレベルが増加した。また、フィードバック音声のラウドネスを減少させるとシマが増加した。また、どの音響特性を変換しても発話の基本周波数に有意差は見られなかった。なお、考察の中で、フィードバック音声が再生されるまでの時間（遅延）の長さおよび評価指標に課題が発見された。

2回目の調査は、システムの遅延を短縮した上で行った。2回目の調査では変換する音響特性として、1回目の調査に加えてハイパスフィルタおよびローパスフィルタを設定した。また、評価指標には1回目の調査に加えてZCR（Zero Crossing Rate）、音響インテンシティ、およびHNR（Harmonic to Noise Ratio）を設定し、分析を行った。

調査の結果、フィードバック音声のピッチを上昇させると発話のジッタが減少し、ピッチを下降させると発話のZCRが増加した。フィードバック音声の抑揚を増加させると発話のHNRが増加した。フィードバック音声のラウドネスについては、増加および減少のそれぞれにおいてシマが増加した。なお、どの音響特性を変換しても発話の基本周波数、ラウドネスレベル、および音響インテンシティに有意差は見られなかった。

本研究の最終的な目標は、ボイスチェンジャの代替として利用可能な発話特性変化システムをAAFを用いて実装することである。2回の調査の結果を踏まえ、システムの機能としてフィードバック音声のピッチ上昇による発話の明瞭化が実現可能であると考えた。

今後はシステム設計の見直しおよび評価指標の精査を行い、AAFにおいて発話特性に変化を与える変換量の閾値について調査する。また、発話特性を変化させるパラメータおよびその変換量を特定した上で、会話中にAAFを使用した際の影響を調査する。

謝辞

本研究を進めるにあたり，川口一画先生，志築文太郎先生，高橋伸先生には多くのご指導を頂きました。心から感謝申し上げます。特に，川口先生には，研究の進め方や論文執筆などの研究の基礎をご指導頂きました。また，研究内容や研究生活における不安などの相談に対してもご快諾頂き，多くのご助言を頂きました。ここに深く感謝の意を表します。

インタラクティブプログラミング研究室の同輩，先輩方には，研究活動において大変お世話になりました。COMMUNICATION チームの皆様には，チームゼミにおいて研究に関するご意見や論文の添削だけでなく，研究生活において多くのご支援を頂きました。特に，市川あゆみ氏には論文執筆や研究発表をはじめ，日頃から研究に関する様々な相談に快くご対応頂き，多くのご助言を頂きました。深く感謝いたします。また，多くの皆様に実験にご協力頂きましたことを大変感謝しております。

最後に，学生生活においてお世話になった皆様，そして，私の学生生活を支えて頂いた家族に深く感謝申し上げます。

参考文献

- [1] Zoom. Zoom Meetings. <https://explore.zoom.us/ja/products/meetings/> (Accessed on 01/30/2024).
- [2] Discord. Discord. <https://discord.com/> (Accessed on 01/30/2024).
- [3] Microsoft. Microsoft Teams. <https://www.microsoft.com/microsoft-teams/group-chat-software> (Accessed on 01/30/2024).
- [4] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, 2021.
- [5] RVC-Project. RVC-Project/Retrieval-based-Voice-Conversion-WebUI.
- [6] Michelle Lincoln, Ann Packman, and Mark Onslow. Altered auditory feedback and the treatment of stuttering: a review. *J Fluency Disord*, Vol. 31, No. 2, pp. 71–89, 2006.
- [7] Jean-Julien Aucouturier, Petter Johansson, Lars Hall, and Katsumi Watanabe. Covert digital manipulation of vocal emotion alter speakers’ emotional states in a congruent direction. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 113, No. 4, pp. 948–953, 2016.
- [8] David Andrew Puts, Steven Gaulin, and Katherine Verdolini. Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, Vol. 27, No. 4, pp. 283–296, 2006.
- [9] Tzu-Yang Wang, Ikkaku Kawaguchi, Hideaki Kuzuoka, and Mai Otsuki. Effect of Manipulated Amplitude and Frequency of Human Voice on Dominance and Persuasiveness in Audio Conferences. *Proc. ACM Hum.-Comput. Interact.*, Vol. 2, No. CSCW, 11 2018.
- [10] Caitlin Coughler, Keelia Quinn de Launay, David W. Purcell, Janis Oram Cardy, and Deryk S. Beal. Pediatric Responses to Fundamental and Formant Frequency Altered Auditory Feedback: A Scoping Review. *Frontiers in Human Neuroscience*, Vol. 16, , 2022.
- [11] Theresa A. Burnett, Jill E. Senner, and Charles R. Larson. Voice F0 responses to pitch-shifted auditory feedback: a preliminary study. *Journal of Voice*, Vol. 11, No. 2, pp. 202–211, 1997.

- [12] Harlan Lawson Lane, A Charles Catania, and Stanley Smith Stevens. Voice Level: Autophonic Scale, Perceived Loudness, and Effects of Sidetone. *The Journal of the Acoustical Society of America*, Vol. 33, No. 2, pp. 160–167, 1961.
- [13] David R. Feinberg, Benedict C. Jones, Anthony C. Little, Dr Michael Burt, and David Ian Perrett. Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour*, Vol. 69, No. 3, pp. 561–568, 2005.
- [14] Benedict C. Jones, David R. Feinberg, Lisa M. DeBruine, Anthony C. Little, and Jovana Vukovic. A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour*, Vol. 79, No. 1, pp. 57–62, 2010.
- [15] Tomoki Toda, Alan W. Black, and Keiichi Tokuda. Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.
- [16] Yannis Stylianou, Olivier Cappé, and Eric Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, 1998.
- [17] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp. 132–157, 2021.
- [18] Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi. Multidimensional scaling of systems in the Voice Conversion Challenge 2016. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pp. 38–43, 2016.
- [19] Joseph Kalinowski, Andrew Stuart, Sarah Sark, and Joy Armson. Stuttering amelioration at various auditory feedback delays and speech rates. *International Journal of Language & Communication Disorders*, Vol. 31, No. 3, pp. 259–269, 1996.
- [20] Garen Sparks, Dorothy E. Grant, Kathleen Millay, Delaina Walker-Batson, and Linda S. Hyman. The effect of fast speech rate on stuttering frequency during delayed auditory feedback. *Journal of Fluency Disorders*, Vol. 27, No. 3, pp. 187–201, 2002.
- [21] Ayoub Daliri and Ludo Max. Stuttering adults’ lack of pre-speech auditory modulation normalizes when speaking with delayed auditory feedback. *Cortex*, Vol. 99, pp. 55–68, 2018.
- [22] Bernard S. Lee. Artificial Stutter. *Journal of Speech and Hearing Disorders*, Vol. 16, No. 1, pp. 53–55, 1951.
- [23] John W. Black. The Effect Of Delayed Side-Tone Upon Vocal Rate And Intensity. *Journal of Speech and Hearing Disorders*, Vol. 16, No. 1, pp. 56–60, 1951.

- [24] Andrew Stuart, Joseph Kalinowski, and Michael P Rastatter abd Kerry Lynch. Effect of delayed auditory feedback on normal speakers at two speech rates. *The Journal of the Acoustical Society of America*, Vol. 111, No. 5, pp. 2237–2241, 2002.
- [25] Riku Arakawa, Zendai Kashino, Shinnosuke Takamichi, Verhulst Adrien, and Inami Masahiko. Digital Speech Makeup: Voice Conversion Based Altered Auditory Feedback for Transforming Self-Representation. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, No. 9 in ICMI '21, pp. 159–167, New York, NY, USA, 2021. Association for Computing Machinery.
- [26] Wakana Taguchi, Fumio Nihei, Yutaka Takase, Yukiko I. Nakano, Shinichi Fukasawa, and Hiroko Akatsu. Effects of Face and Voice Deformation on Participant Emotion in Video-Mediated Communication. In *Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct*, ICMI '18, New York, NY, USA, 2018. Association for Computing Machinery.
- [27] Jean Costa, Malte F. Jung, Mary Czerwinski, François Guimbretière, Trinh Le, and Tanzeem Choudhury. Regulating Feelings During Interpersonal Conflicts by Changing Voice Self-Perception. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 1–13, New York, NY, USA, 2018. Association for Computing Machinery.
- [28] 成瀬加菜, 吉田成朗, 世田圭佑, 鳴海拓志, 谷川智洋, 廣瀬通孝. リアルタイムな変換聴覚フィードバックによる緊張緩和効果の基礎的検討. ヒューマンインタフェース学会研究報告集, Vol. 20, pp. 105–112, 2018.
- [29] Laura Rachman, Marco Liuni, Pablo Arias, Andreas Lind, Petter Johansson, Lars Hall, Daniel Richardson, Katsumi Watanabe, Stéphanie Dubal, and Jean-Julien Aucouturier. DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. *Behavior Research Methods*, Vol. 50, pp. 323–343, 2018.
- [30] Hugh "Jim" Bailey. Open Broadcaster Software. <https://obsproject.com/> (Accessed on 01/30/2024).
- [31] Raphael Lenain, Jack Weston, Abhishek Shivkumar, and Emil Fristed. Surfboard: Audio Feature Extraction for Modern Machine Learning, 2020.
- [32] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.
- [33] Meike Brockmann, Claudio Storck, Paul Nicholas Carding, and Michael J. Drinnan. Voice loudness and gender effects on jitter and shimmer in healthy adults. *J Speech Lang Hear Res*, Vol. 51, No. 5, pp. 1152–1160, 2008.

- [34] 難波精一郎. 知っているようで知らないラウドネス. 日本音響学会誌, Vol. 73, No. 12, pp. 765–773, 2017.
- [35] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. A Cough-Based Algorithm for Automatic Diagnosis of Pertussis. *PLOS ONE*, Vol. 11, No. 9, pp. 1–20, 09 2016.
- [36] 平野滋. 音声障害評価法. 専門医通信, Vol. 120, No. 5, pp. 752–753, 2017.
- [37] 青木直史. ゼロからはじめる音響学. 講談社, 2014.

著者論文リスト

参考論文

本論文の主な内容は，下記にて公表済みである．

- 査読あり国内会議論文
 - 大沼怜生，市川あゆみ，川口一画，変換聴覚フィードバックが話者の発話特性に与える影響の調査，情報処理学会研究報告，コラボレーションとネットワークサービスワークショップ 2023，情報処理学会，2023 年 11 月，8pages.

付 録A 実験1にて用いたアンケート

実験1において、各条件に対して参加者に行ったアンケートの内容を図 A.1 に示す.

タスク後アンケート

onuma@iplab.cs.tsukuba.ac.jp
[アカウントを切り替える](#)

共有なし

参加者ID

回答を入力

年齢

回答を入力

実験条件

☐ ピッチ上昇
☐ ピッチ下降
☐ フォルマント上昇
☐ フォルマント下降
☐ 音量上昇
☐ 音量下降

流れてくる音声の変化に気づきましたか？

1

2

3

4

5

気づかなかった

☐

☐

☐

☐

☐

気づいた

音声の変化を不快に感じましたか？

1

2

3

4

5

感じなかった

☐

☐

☐

☐

☐

感じた

送信

[フォームをクリア](#)

図 A.1: 実験 1 にて用いたアンケートの内容

付 録 B 実験 2 にて用いたアンケート

実験 2 において，各条件に対して参加者に行ったアンケートの内容を図 B.1 に示す．

タスク後アンケート

onuma@iplab.cs.tsukuba.ac.jp

[アカウントを切り替える](#)

共有なし

参加者ID

回答を入力

年齢

回答を入力

セッション番号

選択

流れてくる音声の変化に気づきましたか？

12345

気づかなかった

☐

☐

☐

☐

☐

気づいた

音声の変化を不快に感じましたか？

12345

感じなかった

☐

☐

☐

☐

☐

感じた

送信

フォームをクリア

図 B.1: 実験 2 にて用いたアンケートの内容