自分の声を知り, コントロールするための「自分声フィルタ」の提案

伏見 遼平[1,a)]　吉田 成朗[1]　鳴海 拓志[2]　谷川 智洋[2]　廣瀬 通孝[2]

概要：発声者が聞き返した時により「自分の声そのものだ」と感じやすい音声を出力する「自分声フィルタ」を提案する. 録音した自分の声を聞く不快感の軽減や, ボイストレーニング等への応用を目的とする. 実験を通してフィルタを構成し, パラメータの傾向や個人差について検討を加える. あわせて, このフィルタを用いた修正聴覚フィードバックによる発話への影響やその応用についても考察する.

Self Voice Image Filter for Modified Auditory Feedback

Ryohei Fushimi[1,a)]　Shigeo Yoshida[1]　Takuji Narumi[2]　Tomohiro Tanikawa[2]
Michitaka Hirose[2]

Abstract: We propose "Self voice filter" which outputs the sound similar self perception of the voice, in order to reduce unpleasantness during hearing recorded one's own voice and to control one's own voice more precisely. An experiment was conducted to constitute self voice filter and to examine appropriate parameters.

1. はじめに

自分の声を思ったとおりに操るのは難しい. つい早口で話してしまったり, ニュアンスが意図通りに伝わらなかったりと, 頭のなかでイメージしている声や話し方でそのまま発音することがうまくいかないといったことは成人でも多くある.

声のコントロールのためにボイストレーニングに通う人も多く, こういったトレーニングの市場規模は全世界で 7 億ドルにも達する. また日本人の 8 割が自分の声が嫌いという調査結果もあり, 実際に録音された自分の声を聞くとその割合は 9 割にもなる. その多くが自分の声のイメージと実際の声のズレを原因としてあげている. 録音された自分の声を聞くときの不快感は人種や文化を超えてよく知られている. このようなズレを小さく感じさせる, もしくはなくすことで, 声のコントロールはより容易になるのではないかと考える.

そこで本研究では, 自分の声のコントロールに役立てることや, 副次的に自分の声を聞く不快感を軽減することを目指し, マイクで録音した音声を後から聞き返した際に, もともとの音声よりもより自分の声らしいと思ってしまうような変換を行うフィルタを "自分声フィルタ" として提案する. さらにこのフィルタがどのような形をしているかを調べ, パラメータについて検討を加える.

発声者自身の知覚する音声と, 他の人に伝わる音声は異なる. このズレは主に気導音と骨伝導音の特性の違いによるものだとされており, 内耳の構造により説明がなされている [1]. 実際に骨伝導音を収集し, 発声者の知覚する音声の再現を行っても完全なものとならないという報告もある [2]. "自分声フィルタ"の構成には, このような単純な伝達特性だけではない非自明な成分が含める必要があることが示唆される.

本稿では最初に自分声フィルタに関する研究の背景や先行研究を記述した後, 自分声フィルタのパラメータ検討のために構築したソフトウェアの紹介を行い, これを用いて行った実験および考察を示す.

[1] 東京大学大学院学際情報学府
The University of Tokyo, 7-3-1 Bunkyo, Tokyo 103–0032
[2] 東京大学大学院情報理工学系研究科
The University of Tokyo, 7-3-1 Bunkyo, Tokyo 103–0032
[a)] fushimi@cyber.t.u-tokyo.ac.jp

2. 研究の背景

変調を行った自分の声の音声を聴かせる研究については印象を SD 法で評価させ, 不安等の印象が増大したことが調べられている [3] が, 自分の声のらしさについて評価した文献はない.

自己知覚については, 声よりも顔に関する研究が進んでおり [4], 記憶の中の自分の顔は, 実際よりもより魅力的な顔であることや [5], 目尻や口角を変形させた顔を鏡型のデバイスでフィードバックすることで情動体験や身に着けているものの選好に影響をあたえることができること [6] が明らかになっている. また, 自分の顔を見る行為 [7] や, 行動に関連する音を聞くこと [8] が “ミラーネットワーク”と呼ばれる自己行動の知覚に関わるニューロンネットワークを作動させるという知見もある. 提案システムを用いた応用や原理の解明には, 自己声知覚に関する研究だけではなく, 顔や動作を対象とした研究の成果も応用することができることが示唆される.

3. 自分声フィルタの構成

3.1 自分声フィルタとは

本研究では, まずマイクで録音した音声を後から聞き返した際に, もともとの音声よりもより自分の声らしいと感じるような変換を行うフィルタを”自分声フィルタ”として提案し, このフィルタがどのような形をしているかを調べ, パラメータについて検討を加える. さらにこのフィルタを自分の声のコントロールに役立てることを目指す. 自分声フィルタの模式図を 1 に示す.

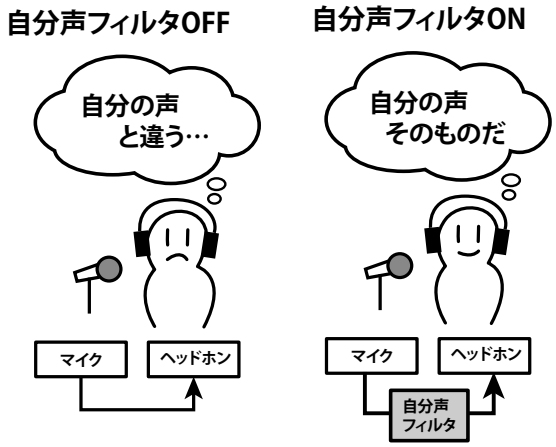

図 1 自分声フィルタの模式図

先に述べたとおり, 自分声フィルタの構成については, 伝達関数を使って骨伝導音を再現したフィルタのパラメータを聴取者に調整させた研究がある. しかし自分の声についてのイメージについて焦点を当て, 再生速度やピッチなどを変更した声を聴かせた研究はない.

自分声フィルタは様々な音声信号フィルタにより構成することができ, そのパラメータについての被験者間の傾向や予測モデルで作ることができると考えた. 本稿ではまず自分声フィルタを知られている信号フィルタにより構成し, そのパラメータについて検討を加える.

3.2 自分声フィルタに用いる処理

音声処理を行う上で一般的に用いられているフィルタとして次のようなものがある.

- FIR・IIR (イコライザ/EQ)
- タイムストレッチング・ピッチシフト
- フォルマントフィルタ

これらのフィルタのうち, 自分声フィルタに必要と考えられるものを抽出する.

このうち骨伝導音が付加されることによる影響は伝達関数の形で表すことができることが知られており [2], FIR を使ったイコライザを使って表現することができる. しかし, このフィルタは録音再生に用いる機具の特性や音響と独立ではないため, 環境からの影響を排除できず, これらを統制しないかぎりパラメータの検討が意味を成さない. 応用を考えるとこのような制約を満たすことは難しいため排除した.

タイムストレッチング・ピッチシフトは, 速度や持続時間を一定に保ったまま, ピッチだけを変更したり, ピッチを一定に保ったまま速度や持続時間を変更させるフィルタである. それぞれ周波数領域・時間領域で実現する実装としてヴォコーダや PSOLA が知られている. ピッチや再生速度は, 抑揚やイントネーション, 聞き取りやすさに影響する音声に重要な要素である. 実装方法もよく知られているため, 本研究ではこれらのフィルタを実装した.

フォルマントフィルタは音声に特有の周波数領域のピークを変位させるものである. フォルマント位置は母音の知覚に影響を与えることが知られている. 母音のはっきりとした発音は聞き取りやすさにつながるため, 各母音をそれぞれ分離させて発音するように誘導することができれば自分の声が意図通りにに伝わるようにコントロールすることができる. しかし, 母音の判別は話者性が強く影響するタスクであり, フィルタの構築に事前のトレーニングが必要になるため, 実装からはのぞいた.

今回は同時に検証できるパラメータ数の都合からタイムストレッチング・ピッチシフトのみに絞って, この 2 つのフィルタを用いて構成した自分声フィルタを用いて, ピッチ・再生速度の 2 つのパラメータについて検討を加える.

4. パラメータ検討のための実験

このシステムを用いて, 自分声フィルタのピッチ・再生速度パラメータについて実験協力者間で共通する傾向および

個人差の大きさについて調査することを目的として, 録音した音声を様々なパラメータで変調した音声を聞かせ「自分の声らしさ」を評価させる実験を行った. さらに社会的に好ましいと考える声, また自分が出そうと意識している声についてのアンケートを行い, 各個人内で計算したフィルタパラメータとの相関を調べた.

4.1 実験用ソフトウェアの制作

自分の声を録音し，これを変調した音声を聞き返すことのできる実験用のソフトウェアを作成した．ピッチ・再生速度の変調は AVFoundation フレームワークに組み込まれているフィルタを用いて実装した．一般には再生速度やスピードの一方を変化させるともう一方も変化してしまうが, 2 つのパラメータを独立して変調できるような実装を行い, ピッチを変えても再生速度は不変，再生速度を変えてもピッチは不変となる．再生速度では± 5%、ピッチでは± 50 セント (=3%) の範囲で違和感が少なく変調できることを確認したため，この範囲で実験を行った.

ソフトウェアは Swift を用いて，Cocoa フレームワーク, AVFoundation フレームワーク上に実装した．Mac OSX を搭載した MacBook Pro 上で動作させた．実験用のインターフェイスとして，ログ記録ボタン等実験の記録に必要なインターフェイスのほか，録音・停止ボタン，変調音声再生ボタンを用意し，これらは必要に応じて実験協力者自身が操作できるようにした．図 2 に作成したシステムのスクリーンショットを示した.

図 2 実験用に作成したソフトウェアのスクリーンショット

4.2 実験手順

実験手順を図解したものを図 3 に示す.

本実験では実験協力者が読み上げる文章として 6 つの文章を用意した．予備実験では読み上げに 10 秒近くかかる文章を使っていたが, すべてを聞き返すときに退屈・苦痛に感じるという声もあった．また, 実験協力者への負担を考慮し全体の実験時間を 1 時間に収めるために 5 秒程度で読み上げられるものとした.

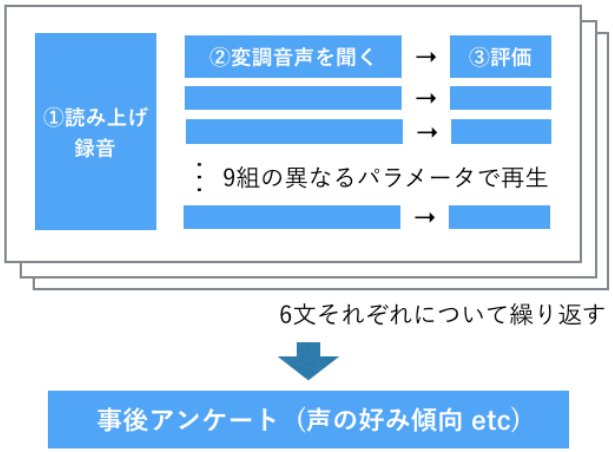

図 3 実験手順の模式図

文章は下記の 6 つを用いた.

(1) 自分の実力は, 自分が一番良く知っているはずだ.
(2) 日本人は決して, ユーモアと無縁な人種ではなかった.
(3) この企画書の提出を, 急いでもらえませんか?
(4) 小松菜がなかったら, ほうれん草でも大丈夫
(5) 繰り返し練習したら, 簡単になるかもしれないよ.
(6) 大学までの電車の中で, 何をしているの?

音声領域で頻繁に使われている, 様々な音素をバランスよく含む音声データベース用文『ATR 音素バランス文 503』に掲載されいてる文章を, 短くアレンジしたもの (1,2,5) を含めた．また, これらの文章は日常会話には不自然なものも多かったため, 疑問文 (3,6) や日常会話に近い文章 (4) を創作したものを追加した.

それぞれの文章について, 文章を読み上げて録音させその後異なるパラメータで変調した音声を聞き, 自分の声らしさを 7 段階で評価することを 9 組のパラメータについてそれぞれ繰り返した.

パラメータはピッチ・再生速度についてそれぞれ 3 通りずつ用いた．ピッチは 50 セント高い音声・オリジナルと同じ音声・50 セント低い音声をそれぞれ HIGH,MID,LOW 条件, 再生速度は 5%速い音声・録音と同じ・5%遅い音声をそれぞれ FAST,MID,SLOW 条件と呼ぶ．これらのピッチ 3 通り×再生速度 3 通りの計 9 通りのパラメータで変調した音声をランダムに並べ替えて呈示をおこなった.

さらに事後アンケートでは, 社会的に好ましいと考える声, また自分が出そうと意識している声について調べることを目的とし, 下記の 4 つの状況・対象についてそれぞれ, ピッチ・再生速度に対応する「高い声のほうが、低い声よりも好ましいと思う」「早い話し方のほうが、遅い話し方よりも好ましいと思う」という 2 つの文章を 7 段階のリッカート尺度で評価させた.

(1) 講義やビジネスやスピーチなどの状況で, 自分と同性の他人についてどんな声や話し方が好ましいと考えるか
(2) 友人や家族, 恋人として話す場面において, 自分と同性の他人についてどんな声や話し方が好ましいと考えるか
(3) 講義やビジネスやスピーチなどの状況で, あなた自身が出したい声/したい話し方
(4) 友人や家族, 恋人として話す場面において, あなた自身が出したい声/したい話し方

4.3 実験協力者

実験には 14 名が参加したが, 事後アンケートで自分の声を日常的に録音再生していると申告があった協力者や, 音声信号処理に詳しく、変換されていない声を 100%当てることができた参加者は除いて分析を行った. 結果として分析の対象となったのは 11 名 (男性 5 女性 6, age: 21-26 mean=23.5) であった.

4.4 結果

4.4.1 各条件の平均値

変調したピッチ・再生速度を独立変数とし, 各試行・条件において実験協力者が回答した自分の声らしさの 7 段階リッカート尺度の値 (1〜7 の整数値を取る) を従属変数とした分析を行う. まず図 4 に全 9 条件における平均値の結果を示した. 結果として変調していないオリジナルの音声 (MID-MID) が最も高い結果となった.

ただし, 事後アンケートの自由記述欄について, 「音質の差を手がかりに自分の本当の声を当てることができた」という協力者がいたため, 音質の差が MID-MID のスコアを押し上げている可能性が考えられる.

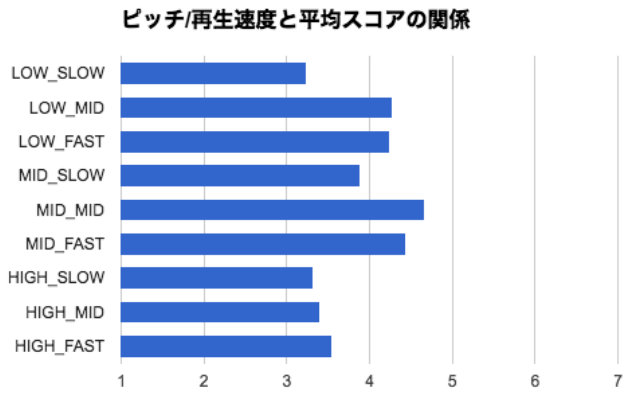

図 4 ピッチ及び再生速度条件と自分の声らしさの関係

図 5,6 に, ピッチ・再生速度別に平均値を示した. ピッチパラメータに関しては、オリジナルと同じピッチが最も自分の声らしいと判断されたが, これを除くと自分の声よりも低いほうが自分の声だと感じやすい傾向が見られる. また再生速度パラメータは, 自分の声よりも速いほど自分の声だと感じやすいという傾向が見られる.

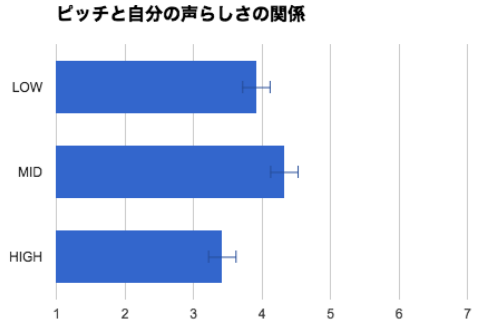

図 5 ピッチ条件と自分の声らしさの関係

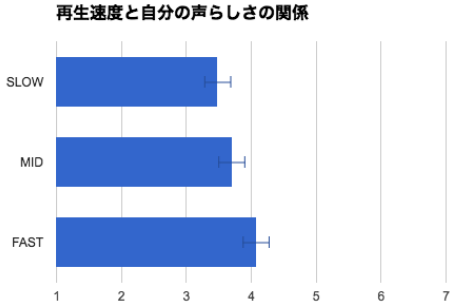

図 6 再生速度と自分の声らしさの関係

4.4.2 分散分析

ピッチ・再生速度についての 2 要因分散分析を行った. 分析は 11 名の実験協力者から得られた 65 試行のデータが用いられた (1 試行は正しく記録ができていなかったため削除した).

交互作用が認められた ($F(4,64) = 2.41, p = 0.049, \eta^2 = 0.0087$) ので, 単純主効果の検定を行ったところ, ピッチ・再生速度それぞれについて有意差が認められた (ピッチ: $F(2,64) = 6.18, p = 0.003, \eta^2 = 0.0406$, 再生速度: $F(2,64) = 24.9, p = 0.000, \eta^2 = 0.027$).

Shaffer の補正のもと多重比較 (有意水準 $p < .05$) を行ったところ, ピッチパラメータについては MID<HIGH, LOW<HIGH の 2 水準間にそれぞれ有意な差があり, また再生速度パラメータについては SLOW<FAST, SLOW<MID の 2 水準間に有意な差があった.

ピッチ・再生速度ともに自分の声らしさに関与しており, 特に再生速度に関しては, ピッチのパラメータ条件にかかわらず変調していないオリジナル音声よりも早く再生した音声のほうが自分の声らしく判断されるという結果となった.

4.4.3 傾向値の標準偏差および男女差

次に最も自分らしい声として選択した音声の各パラメータの分布を調べることで, 個人差について評価を行う.

最も自分らしい声として選んだパラメータの値を

各実験協力者で平均を取り、この値を選択傾向値とした. 例えば, 選んだ音声のピッチのパラメータが LOW,LOW,LOW,LOW,MID,HIGH だった場合は -1,-1,-1,-1,0,1 の平均を取ることで 選択傾向値 -0.5 (ピッチが低めを選ぶ傾向がややある) を得る.

この値の標準偏差は個人間のパラメータのばらつきを示すと考えられるが, 標準偏差は再生速度が 0.14 に対してピッチが 0.4 という大きな値をとった. 再生速度に比べると, ピッチは個人差が大きいことが示唆された.

男女によって差があるかどうか調べた結果, ピッチに関しては有意差がある (t 検定, $p < .05$) が, 再生速度に関しては有意差がないという結果となった.

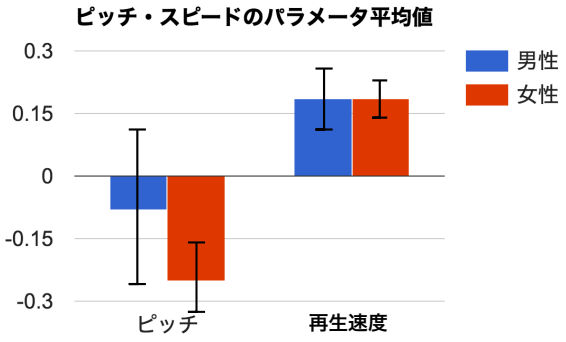

図 7 男女間のピッチ・再生速度の選択傾向値の差 (エラーバーは標準誤差)

4.4.4 傾向値と事後アンケートの相関

この個人差について, 男女要因以外に予測できる因子がないか調べた. 事後アンケートで 7 段階リッカート尺度を用いて評定させた項目のそれぞれについて, 選択傾向値との相関がないか調べたところ, ピッチパラメータについて, 事後アンケートで評定させた社会的に好ましいピッチ傾向との逆相関 ($r = -0.50$) が見られた. 他の事後アンケートの評定項目とは相関は見られなかった.

ピッチ・再生速度それぞれとの相関を図 8,9 に示す.

この結果は, 個人差が比較的大きいピッチパラメータについて, 発声者に関する情報からこのパラメータの値を予測することができることを示すものである.

5. 議論

5.1 適切なパラメータの値域

今回は± 5%の範囲で再生速度を変調した. ピッチパラメータについては「変調なし」を中心にスコアは山形の結果となったが, 再生速度パラメータについては最も自分の声らしいと判断されるパラメータは今回検討した± 5%の範囲外にある可能性が高い. 今後この検討を追加で行いたいと考えている.

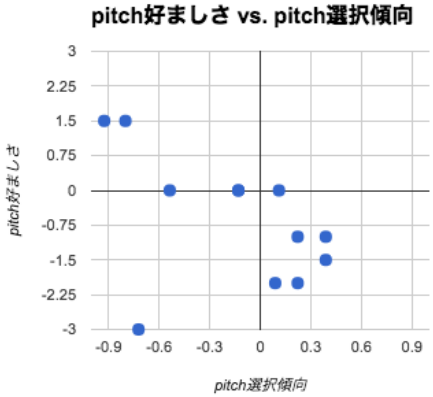

図 8 自分の声として選んだピッチパラメータの傾向と, 事後アンケートで評定した社会的に好ましいピッチ傾向の関係

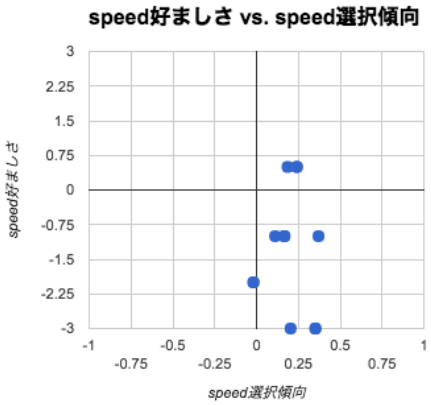

図 9 自分の声として選んだ再生速度パラメータの傾向と, 事後アンケートで評定した社会的に好ましい再生速度傾向の関係

5.2 読み上げ時に想定する状況による違い

今回の実験では, 読み上げさせた文章はパブリック (講義や面接など他人に対する)・プライベート (友人や家族等に対する) 両方に対応するシチュエーションが混在したものだった. 事後アンケートの項目も, パブリック・プライベートな状況について実験協力者数の不足から, 意味のある分析を行うことはできなかった.

しかし現実的な状況を考えると, 理想とする声や実際に観察される話し方は, 対する相手や会話の流れなどのシチュエーションにより大きく変わる. 対応してシチュエーションにより自分声フィルタのパラメータが変化することも十分考えられる. 後述するワークショップでは様々なシチュエーションを想定した実験を行う予定であり, この実験でより多くのデータを収集し分析を行う.

5.3 声のコントロールに向けて

本研究では “自分声フィルタ” を最終的に自分の声をうまくコントロールすることに役立てることを目指しているが, このためのシステム構築に援用しようとしている修正聴覚フィードバックに関する理論を紹介する,

修正聴覚フィードバックは, 発声した音声や動作音にマイクで録音し, 修正を加えてからヘッドホン等でフィードバックするという手法である. 修正聴覚フィードバックでは様々なフィルタを用いて多様な実験がなされており, 調音声のフォルマント [9] や, 筆記中の体験やタスク成績 [10], 情動体験 [11] が変化するという現象が指摘されている.

中でも赤木らは調音系に焦点を当て, 修正聴覚フィードバック下で母音フォルマントのピークについて補償動作 (フィードバックの変調と逆方向に, 無意識下で音声を調整する現象) が起こることを報告している [2].

修正聴覚フィードバックに自分声フィルタを用いることで, この効果を応用し自分の声の特徴にあわせて自分の声のコントロールをより容易にするシステムを構築することができるのではないかと仮説を立て, 現在検証を進めている.

6. おわりに

本研究では, 自分の声を録音しこれを変調した音声を聞き返すことのできるシステムを用いて, “自分声フィルタ”の構成およびピッチ・再生速度の 2 つのパラメータについて検討を加えることを試みた.

結果として, ピッチパラメータについては実際の声よりも低いほうが自分の声だと感じやすいという結果が得られたが, 個人差が大きかった. 事後アンケートの傾向との相関を調べたところ, 社会的に好ましいと考えるピッチパラメータの傾向と逆相関が見られた. また再生速度のパラメータについては実際の声よりも早く再生したほうが自分の声だと感じやすいという結果となった. この結果については個人差は小さく, また事後アンケートの項目との相関はなかった.

現在日本基礎心理学会「心の実験パッケージ」開発委員会の協力を得て, 中高生を対象としてワークショップ『自分の声と仲良くなろう』のプログラムを制作している. このワークショップでは, 本稿で記述した実験にあたる「自分の声をあててみよう」を含んでおり, グループの中で読み上げ役の参加者よりも他の参加者のほうが変調していない声を当てやすいことや, 読み上げ役が間違えてしまった音声の変調パラメータには一定の傾向があることなどを実際の実験を通して確かめることができる. さらに,『叱る』『謝る』など具体的なシチュエーションを想定して発声させた音声について, グループ内のほかの参加者に意図が伝わっているかを評価させている.

このワークショップは, 全国の科学教室や科学館施設で開催することを予定し, さらに細かい・広いパラメータ幅での実験実施・データ収集を行い検討に生かす. すでに 2016 年 8 月に, お茶の水女子大学の協力を得てワークショップの内容をプレ実施し, その結果をもとにプログラムの改善にあたっている.

本研究では個人差について男女差や, 自己アンケートで評定させた社会的に好ましいピッチ傾向との関係を調べることで, 大きな個人差について要因を調べようとしたが, 年齢やその人自身の声の特徴なども大きな要因として働くことが考えられる. 今回はそうした点について詳細な分析を行えなかった. ワークショップを通して幅広い年齢層からデータを収集し, この要因分析をさらに進めていく.

7. 謝辞

本研究およびワークショップの実施は, 日本基礎心理学会「心の実験パッケージ」開発委員会の協力のもと行いました. ここに感謝を申し上げます.

参考文献

[1] Békésy, G. V.: The structure of the middle ear and the hearing of one's own voice by bone conduction, *The Journal of the Acoustical Society of America*, Vol. 21, No. 3, pp. 217–232 (1949).

[2] 中井孝芳，高尾論司：発声者自身の音声の知覚経路，電子情報通信学会技術研究報告. EA, 応用音響， Vol. 101, No. 73, pp. 15–22 (2001).

[3] Rousey, C. and Holzman, P. S.: Some effects of listening to one's own voice systematically distorted, *Perceptual and Motor Skills*, Vol. 27, No. 3 suppl, pp. 1303–1313 (1968).

[4] Koole, S. L., DeHart, T., Nimrod, D., Fryer, J., Salvator, J. and Knihnicki, J.: Self-affection without self-reflection: Origins, models, and consequences of implicit self-esteem, *The self*, pp. 21–49 (2007).

[5] Epley, N. and Whitchurch, E.: Mirror, mirror on the wall: Enhancement in self-recognition, *Personality and Social Psychology Bulletin*, Vol. 34, No. 9, pp. 1159–1170 (2008).

[6] Yoshida, S., Tanikawa, T., Sakurai, S., Hirose, M. and Narumi, T.: Manipulation of an emotional experience by real-time deformed facial feedback, *Proceedings of the 4th Augmented Human International Conference*, ACM, pp. 35–42 (2013).

[7] Uddin, L. Q., Kaplan, J. T., Molnar-Szakacs, I., Zaidel, E. and Iacoboni, M.: Self-face recognition activates a frontoparietal “mirror” network in the right hemisphere: an event-related fMRI study, *Neuroimage*, Vol. 25, No. 3, pp. 926–935 (2005).

[8] Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V. and Rizzolatti, G.: Hearing sounds, understanding actions: action representation in mirror neurons, *Science*, Vol. 297, No. 5582, pp. 846–848 (2002).

[9] 赤木正人：音声コミュニケーションにおける知覚と生成の相互作用に関する研究，平成 16 年度 平成 18 年度科学研究費補助金 (基盤研究 (B)) 研究成果報告書 (2007).

[10] 金ジョンヒョン，橋田朋子，大谷智子， 苗村健：筆記音のフィードバックが単純な筆記作業に及ぼす影響の検討，日本バーチャルリアリティ学会論文誌，Vol. 17, No. 3, pp. 289–292 (2012).

[11] Aucouturier, J.-J., Johansson, P., Hall, L., Segnini, R., Mercadié, L. and Watanabe, K.: Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction, *Proceedings of the National Academy of Sciences*, Vol. 113, No. 4, pp. 948–953 (2016).