I-018

口唇動作を用いた日本語発話トレーニング法の検討
Study of Japanese utterance training method using lip movement

菅沼 美由起† 山村 知生† 星野 祐子† 山田 光穂†
Miyuki Suganuma Tomoki Yamamura Yuko Hoshino Mitsuho Yamada

1. はじめに

近年，コンピュータ技術が進歩して多様な情報を用いる情報化社会となってきている．加えて，情報化社会に伴って個人情報等の多くの情報を紙媒体からインターネットを使用してのデジタル情報として扱われるようになった．デジタルデータに変換することでインターネットを通じて場所や時間の制限を受けることなく，閲覧・変更・改変を行うことが出来るようになり，大きな作業の効率化がはかられた．しかし，今まで紙媒体に記述されていた情報をデジタルデータとして扱えるようにするにはコンピュータへのキーボード操作による手入力が必要になる．こういった作業をより簡便で直感的に行えるようにする方法として音声認識の技術が注目されている．中でもスマートフォンの音声認識技術は目覚ましく性能が向上しており，エージェントとの雑談を楽しむことが出来るほどになった[1]．その技術を利用し，ヒューマノイドロボットなどに応用する例も数多くみられるようになった[2]．一方で，発話に伴う特徴に着目した生体認証システムの開発や，コンピュータ支援言語教育（以下 CALL）などにおいても音声認識技術が用いられている．音声認識を用いたデータ入力はキーボード入力と比較して，手が不自由な人や手を使うことのできない状況下であっても使用することが出来，使いやすいインタフェースとなっている．一方で音声認識技術には問題点もあり，雑音が多い場所等では目的の音声を認識することが難しく認識率の低下が懸念される．そのため，マイクを使用し指向性を持たせることで音声認識率の低下を防ぐ方法や音声認識に加え，音声を用いない認識を組み合わせてそれぞれの認識のメリットを活かしてデメリットを減らすマルチモーダル認識[3][4]といった方法などが考えられている．また岡田らは，人間は視覚情報を手がかりに，物事に関する概念を獲得することが可能であり，この概念は先天的に与えられるものではなく，人間自らが経験的に獲得していくものであり，もとの視覚情報と密接な対応関係を持つとしている[5]．本研究では，マルチモーダル認識の 1 つである口唇動作を用いた発話認識に注目した．しかし，研究を行う中で口唇動作による認識は個人差により認識率が低下することがわかった．以上のことから私たちは発話改善を目的としたトレーニング手法の提案と開発を行っている．また，口唇動作は個人差があることが知られている．先行研究においても個人で認識率の差が出ることが分かった．そのため，本研究では当研究室で行っている研究を基に開発された発話認識と口唇動作のトレーニングを行うことが出来る装置を用いたトレーニング法の評価を行った．本研究では，トレーニングを行うことで発話を流暢にすること，音声を用いてのトレーニングが困難な聾話者などの方々も含めて発話のトレーニングを簡単に自分自身で出来るような装置の作成および装置の評価を目的とする．発話学習者に自分の発音がどれだけ進歩したかの評価結果を示すことは大きなインセンティブとなり，学習への意欲を増進させる．また，このような自主トレーニング装置を用いる利点は個別学習が可能なことである．発話学習者それぞれのレベルや学習ペースに合わせて納得いくまで繰り返しトレーニングを行うことが出来る．加えて，先行研究からも課題とされていた口唇動作の個人差による発話認識率の低下を改善することも目的とする．

† 東海大学，Tokai University

2. 発話トレーニングシステム

本研究で用いた発話トレーニングシステムについて述べる．実験に用いた口唇特徴点抽出装置の操作画面を図 1 に示す．図 2 は口唇トレーニング画面である．図 1 のように被験者が画面内に入ると自動的に顔の特徴点が認識される．これは SeeingMachines 社製の FaceAPI と Windows フォームアプリケーションの Visual Studio 2010 の C++言語を使用して当研究室で開発したものである[6]．faceAPI は高速かつ高精度な顔認識を行うシェアウェアのソフトである．取得できるデータは顔の向きや位置だけではなく目、鼻、口、眉などの情報も取得できる．口唇の点だけでも上唇と下唇，唇と皮膚の境界で計 16 点を取得できる．サンプリングレートは 30fps で，検出精度は外眼角間の距離が最小の 40 画素である場合 1cm 以下となる．また，faceAPI のバージョンは 3.2 となっている．操作手順は，使用するライブラリを図 1 右側の”Set Up”ボタンから設定し，被験者の名前を右下にファイル名として入力し，実験を開始する．”Start”ボタンをクリックすると画面左上の単語が口を閉じるように促す表示に切り替わる．その後，装置が口を閉じたことを認識すると再び単語の表示に切り替わる．そして被験者は発話を開始する．このとき，口を閉じたことが上手く認識されない場合は画面右側の”Re-recognition”ボタンをクリックすることで顔の特徴点を再認識することができる．発話終了後は”Stop”ボタンを被験者自身がクリックする．発話が終了すると，図 2 のトレーニング画面が表示される．ここで表示される赤い口の形の線はモデルデータである英語教員の口唇動作，黒い口の形の線は被験者の口唇動作を示している．図 2 画面右側の”Start”ボタンをクリックすると，発話した時の口唇動作がそのまま再現される．この口の動きをみながら被験者は発話トレーニングを行う．また，トレーニングでは被験者の発話音声の録音もヘッドセットを用いて同時に行っている．

3. 実験内容

先行研究として，口唇動作をトレーニングするために発音が正しいと考えられるアナウンサーの発話の口唇動作を取得した．その口唇動作のデータを基準として，本学の生徒に協力してもらい実験を行い，同様にして口唇動作のデータを取得した．口唇動作の取得には前項にも

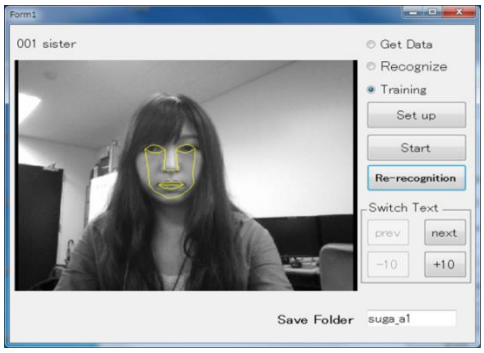

図 1. 口唇特徴点取得画面

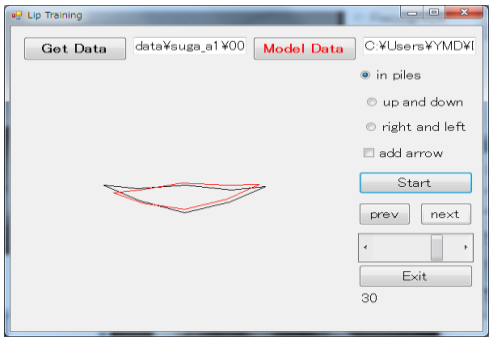

図 2. 口唇トレーニング画面

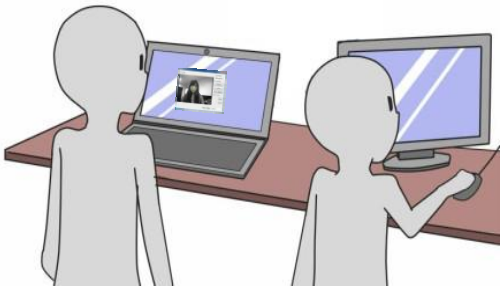

図 3. 発話データの取得の様子

挙げた当研究室で開発した自動特徴点抽出装置を用いている．使用したカメラは Windows パソコン内蔵の web カメラを用い，カメラの解像度は 640×480 に設定し，被験者とカメラの距離を 40cm にして撮影を行った．取得したデータは日本語の発話訓練用書籍「声が良くなる簡単トレーニング」[7]のうち「あ」から「お」を多く含む文章をそれぞれ 2 文ずつ，計 10 文を抜粋し使用した．表 1 に使用したデータを示す．先行研究において，NHK 放送研修センターのベテランアナウンサーの方々の男女の発話データを取得した．図 3 に発話データの取得の様子を示す．左がアナウンサーで右が実験者である．被験者の画面には文章を表示し，操作は実験者が行っている．実験者の操作画面によってアナウンサーの方の気が散らないよう斜めに設置し，画面が見えない状態にした．データの取得は NHK 放送研修センター内の視聴覚室で行った．なお，本実験でのデータ取得は本学内のスタジオで行った．本実験では被験者に図 2 に示される口唇トレーニング画面を確認しアナウンサーの口唇動作と自身の口唇動作を比較してもらい，トレーニング装置を使用することにより，発話の改善がされたのか評価を行った．1 回の実験で 1 文につき 5 回のデータを取得する．1 回目から 5 回目まで，それぞれデータを確認しトレーニングを行った状態でのデータ取得を行うこととし，一連の流れを全 10 文行うことで 1 セットとした．トレーニングは継続的に行うことが予測されることから日を空けず，かつ長期的に行うことを目的とし，1 日につき 2 セットまでトレーニング可能としてデータを取得した．本実験では被験者のタイミングでデータを取得，口唇動作をトレーニングできるように操作は被験者が行い，実験者は取得されたデータに誤りがないかの確認のみ行っている．発話時に顔の動きが大きい場合は，口唇特徴点の差異が多くなる懸念があるため顔はできるだけ動かさないよう注意も行ったが，出来る限り自由な発話の取得をしたいと考え，頭部固定などは行わなかった．被験者は本学学生の男性 3 名女性 1 名の計 4 名のデータを取得した．この際，口唇動作データのみではなく音声のデータも取得し，3 つの評価項目を設けた音声比較を行い，トレーニングによる発話の改善がみられているかを 1・3・5・7・10 セット目の音声データをそれぞれランダムに選び図 4 のように 1・3, 3・5, 5・7, 7・10 セット目の音声を一対比較することで客観的評価を行った．3 つの評価項目には，アーティキュレーション(歯切れの良い発音)，声の速さ，声の大きさを取り上げて比較した．以上の 3 つの項目についてはモデルであるアナウンサーの声の特徴[8]である語音の明確さやテンポ，強さ，高さなどの要素を参考としてトレーニングによって効果的に得られるだろうと考えられる成果を独自に設定および定義した．

表 1. 使用した日本語データ

会ったら愛想よく挨拶しなさい
憧れの相手に会う
生きがいを求めていこう
今以上の思いを入れる
歌を歌って憂さ晴らし
迂闊に上手いウソ
栄誉よ、栄光よ、永遠なれ
えらい絵描きさんが選んだ絵
オオカミの大きな遠吠え
おいしいお菓子をお裾分け

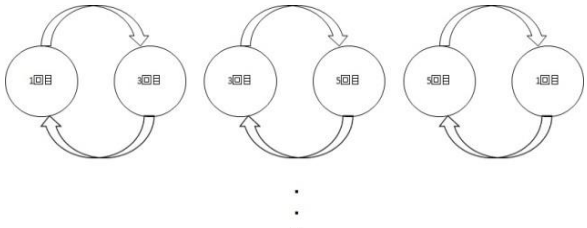

図 4. 一対比較の例

4. 結果・考察

4.1 口唇動作履歴について

先行研究で取得したアナウンサーの口唇動作をモデルとしたトレーニング法を用いた実験を行った．また，本学学生を一般発話者とし，アナウンサーの発話と一般の発話間に違いがあるのか動作履歴を確認し行った．ここでは「オオカミの大きな遠吠え」についての結果を例として示す．図 5 はアナウンサーの口唇動作履歴を示したものであり，図 6,7 はそれぞれの被験者が発話した口唇動作履歴を 1, 5, 10 セット目まで示したものである．横軸は時間で，縦軸は移動量を示している．青い実線が左唇特徴点の移動を示し，赤い実線が下唇特徴点の移動を示している．これらの図においては，1 セット目の発話ではどの被験者も左唇，下唇特徴点ともにアナウンサーの動作履歴に比べ振幅が小さく不安定な動作を見せているものがほとんどであった．しかし，5 セット目 10 セット目とトレーニング回数を重ねたのちの動作履歴では不安定な動作は少なくなり，振幅も大きくなったことからよりはっきりとした大きな声で発話していたことがわかった．特に，図 6 の被験者 T が発話した「オオカミの大きな遠吠え」においては 10 セット目で動作履歴の波形は規則的であり，振幅の大きさの変化も顕著に出ていた．

4.2 音声比較について

本実験では前項で示した口唇の動作履歴の他にも同時にトレーニング時の音声を取得しており，ここではその音声を比較した結果を示す．今回はトレーニングに協力してもらった学生も含め，全部で 20 代の学生 10 名に協力してもらった．この際，自分のトレーニング音声を聴いてもらうことはなく自身とは別の被験者の音声を比較してもらった．

4.2.1 一対比較について

音声比較は一対比較法で行い，1 回の比較につき 3 つの評価項目を設け主観的に比較してもらった．比較項目は，アーティキュレーション(歯切れの良い発音)，声の速さ，声の大きさを取り上げた．これらの項目はモデルであるアナウンサーの発話から得られるだろう効果の要素を独自で選び定義した．アーティキュレーションは「オオカミの」「大きな」「遠吠え」というように，1 語 1 語がしっかりと発音されているかを基準として比較してもらった．また，声の速さは音声が聞き取りやすい適切な速度であるかどうかを，声の大きさは純粋に比較している音声のどちらの音量が大きいかを基準として比較してもらった．これらの比較項目については実験前に説明を行った．特に声の大きさについては比較開始時に被験者に音声のボリュームを調整してもらい，それ以降は変更しないように留意してもらった．それぞれの評価者は被験者 1 人につき 4 回の音声比較

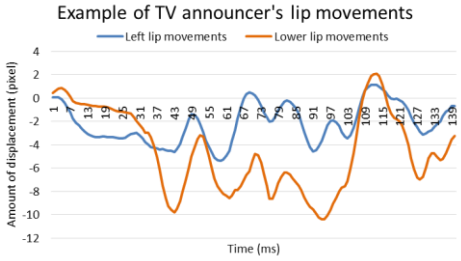

図 5. アナウンサーの口唇動作履歴の例

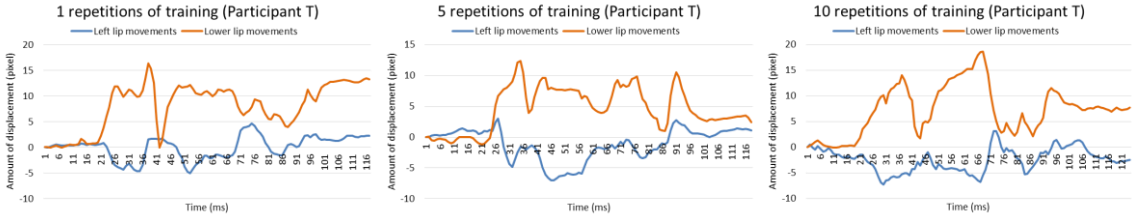

図 6. 被験者 T 「オオカミの大きな遠吠え」について

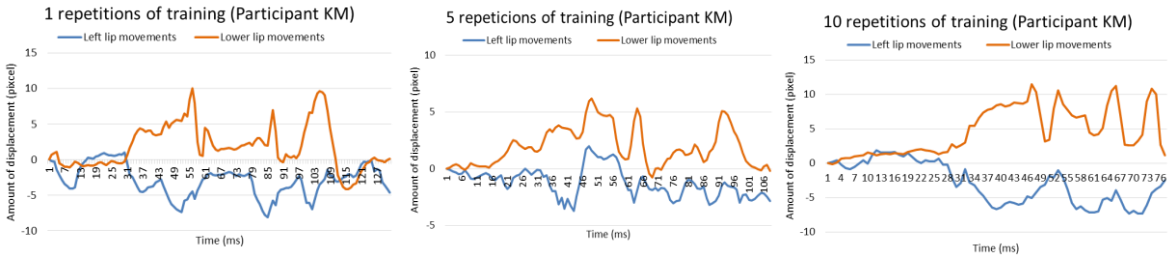

図 7. 被験者 KM 「オオカミの大きな遠吠えについて」

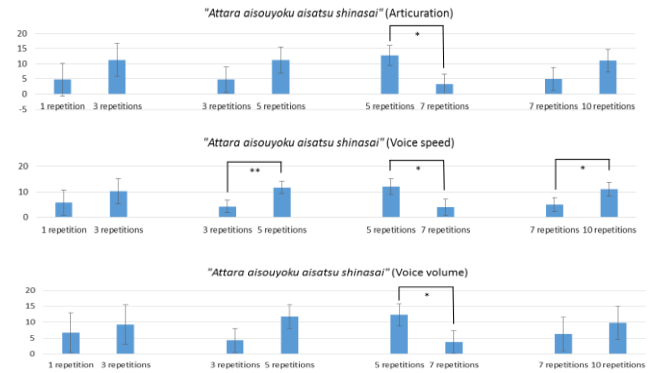

図 8.　「会ったら愛想よく挨拶しなさい」について

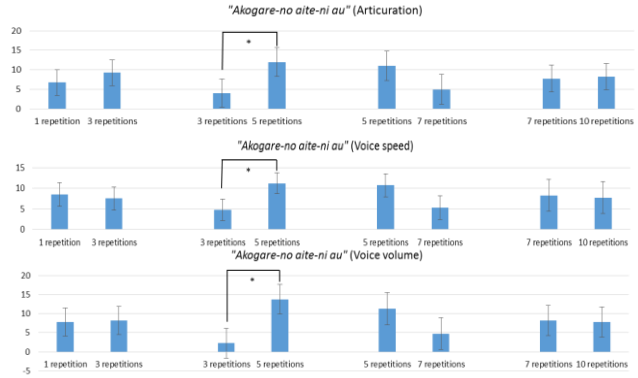

図 9.　「憧れの相手に会う」について

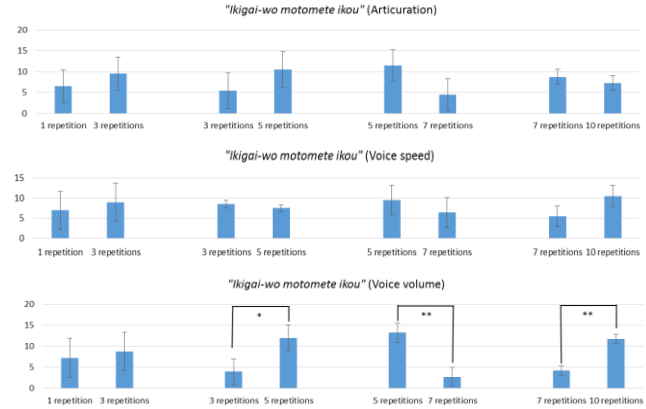

図 10.　「生きがいを求めていこう」について

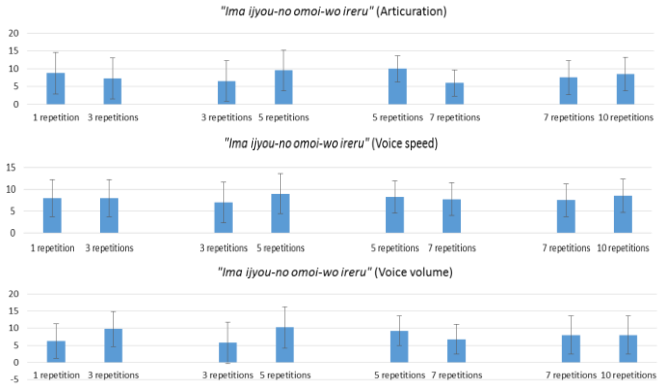

図 11.　「今以上の思いを入れる」について

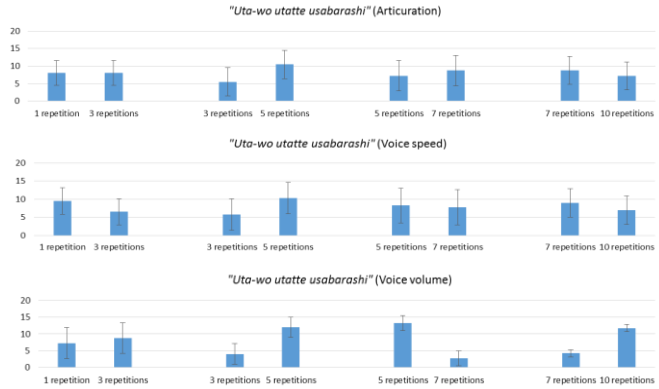

図 12.　「歌を歌って憂さ晴らし」について

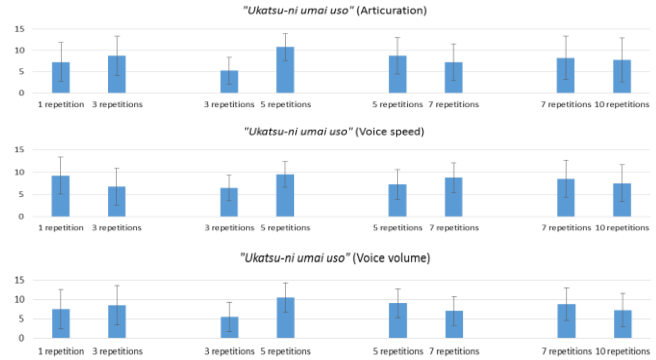

図 13.　「迂闊に上手いウソ」について

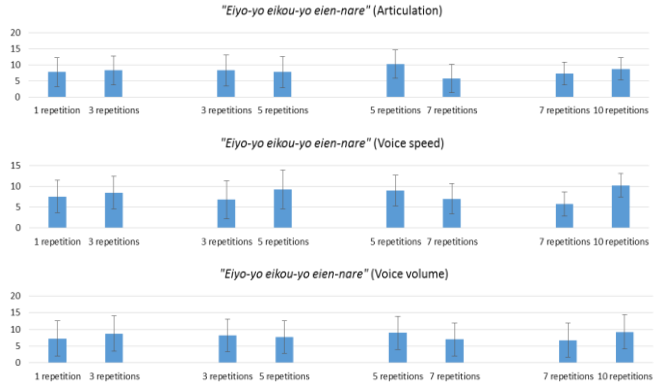

図 14.　「栄誉よ，栄光よ，永遠なれ」について

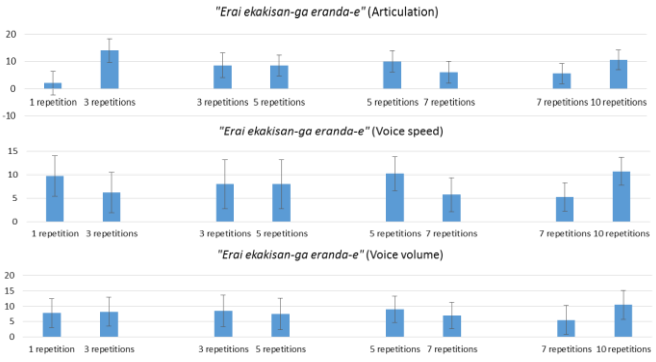

図 15.　「偉い絵描きさんが選んだ絵」について

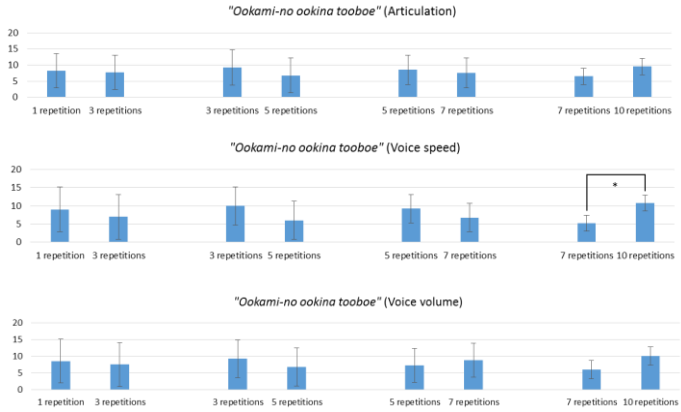

図 16.　「オオカミの大きな遠吠え」について

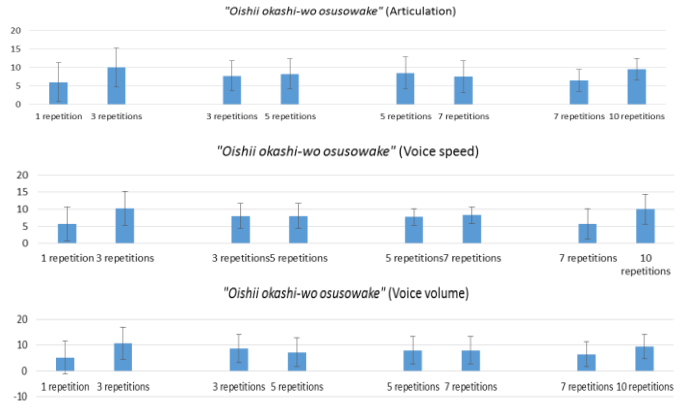

図 17.　「おいしいお菓子をお裾分け」について

を行い，全部で 16 回行ってもらった．図 8 は「会ったら愛想よく挨拶しなさい」の結果について示している．このグラフは一対比較の全データの平均と標準偏差を示したものである．1 列目はアーティキュレーションの結果について示したものである．2 列目は声の速さの結果について，3 列目は声の大きさの結果について示したものである．図 9 から図 17 までについても同様に示している．図 9 は「憧れの相手に会う」，図 10 は「生きがいを求めていこう」，図 11 は「今以上の思いを入れる」，図 12 は「歌を歌って憂さ晴らし」，図 13 は「迂闊に上手いウソ」，図 14 は「栄誉よ，栄光よ，永遠なれ」，図 15 は「えらい絵描きさんが選んだ絵」，図 16 は「オオカミの大きな遠吠え」，図 17 は「おいしいお菓子をお裾分け」の結果について表している．10 回のトレーニングを通して，全母音に共通して 1 から 3 回目のトレーニングにかけては改善の変化が少なく，3 から 5 回目で改善傾向があり，5 から 7 回目では逆効果になることがあり，7 から 10 回目ではまた改善傾向になる．トレーニングの回数を終えるごとに着実に改善されるわけではなく，スランプと言える 5 から 7 回目のトレーニングを経過して再び改善傾向が見られた．これは 1 つの学習特性を表していると考えられる．さらに，一対比較の評価結果の平均について t 検定を行った．有意差が得られたかどうかの結果は図 8 から図 17 のなかに* ($p<0.05$)と** ($p<0.01$)で示している．図 8, 9, 10, 17 から，/a/, /i/, /o/の母音では有意差が見られたことがわかる．図 8 の「会ったら愛想よく挨拶しなさい」の声の速さについての 3 回目と 5 回目のトレーニングでは有意差が見られた ($p=0.008<0.01$)．また，図 10 の「生きがいを求めていこう」の声の大きさについての 5 回目と 7 回目，7 回目と 10 回目のトレーニングでも有意差が得られた ($p=0.008<0.01$)．一部例外が見られるものの，すべての母音について 3 つの評価項目において，3 回目から 5 回目，7 回目から 10 回目で改善傾向が見れらた．我々が/a/, /i/, /o/の音を発声する際，通常は/a/, /o/の場合は縦に，/i/の場合は横に口を大きく開く．そのため，有意差が得られたのではないかと考えている．これらの結果から，トレーニング回数を重ねるほど被験者の発話の評価結果が高くなることが分かった．

4.2.2 フォルマント解析について

次に一対比較による評価に加え，フォルマント解析による評価を行った．森は，第 1 フォルマントと第 2 フォルマントで性別や人種によらず母音が区別できるとしている[9]．また大塚は，音声のスペクトログラムで第 1 フォルマントと第 2 フォルマントを検出し，周波数値をもとに耳で聞いた評価と比較して，より具体的で細かな母音発音指導ができると報告している[10]．音声学では，音声の第 1 フォルマントと第 2 フォルマントをプロットすることで，疑似的に発話時の口の開け方と舌の位置を表すことができる IPA（International Phonetic Association）チャートというものがある．これは，国際音声記号として英語教育において世界共通で用いられている．図 18 は日本人男性の平均フォルマント値を示している．ここでは，図 19 に「栄誉よ，栄光よ，永遠なれ」についての 1 回目から 10 回目のトレーニング結果についての散布図を示す．この文章は/e/の音が多く含まれているため，例として示すのにふさわしいと考えた．丸でプロットしているのが 1 回目のトレーニング，四角でプロットしているのが 5 回目のトレーニング，ひし形でプロットしているのが 10 回目のトレーニングについてのフォルマント値である．図からわかるように，それぞれのフォルマント値が 10 回のトレーニングを通して集ま

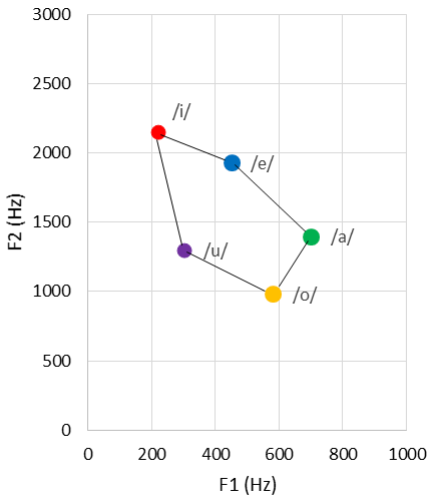

図 18.　日本人男性平均母音フォルマント値

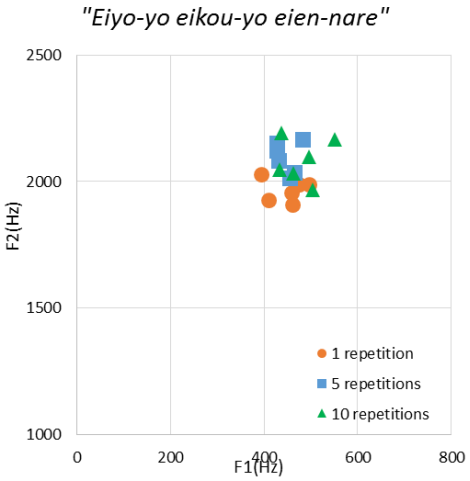

図 19. 被験者の「栄誉よ，栄光よ，永遠なれ」についての各トレーニングにおけるフォルマント散布図

ってきているのが分かる．さらに，図 18 に示した平均フォルマント値と比較すると，「栄誉よ，栄光よ，永遠なれ」のフォルマント値と平均の/e/のフォルマント値が非常に近いことが分かった．我々が/e/から/i/の発話をするとき，縦方向から横方向の口の動きをするため，被験者ははっきりと発話することを心掛けやすかったと考えられる．

5. おわりに

本研究では，個人差による認識率の低下および個人の発話能力の改善を目的として先行研究で開発された口唇特徴点抽出装置を用いたトレーニング法の評価を行った．トレーニングには前回の実験で取得したアナウンサーの口唇動作をモデルとして使用した．トレーニングによって取得した被験者の口唇動作とアナウンサーの口唇動作をまとめ，音声も取得し複数の評価項目を設け一対比較を行った．口唇動作履歴については 1 セット目から 10 セット目にかけて安定した動作になり，移動量も大きくなったことから，トレーニング前の発話よりもはっきりと大きな声で発話していることが分かった．音声比較については，特に/a/，/i/，/o/の発話において，10 回のトレーニングで十分な評価結果が得られ，発話の改善が示された．また，回数ごとにトレーニング効果が得られるわけではなく，5 回目から 7 回目など逆にトレーニング効果が見られなくなるところが有り，スランプもしくは中だるみと言った学習プロセスが示された．5 回目から 7 回目の要因を探り，飽きさせない工夫をすることにより，トレーニング回数を重ねるほど良い評価結果が得られると考えられる．フォルマント解析においても，「栄誉よ，栄光よ，永遠なれ」のように顕著な改善が見られたものもあった．以上のことから，本研究で検証したトレーニング法の有用性が示唆された．今後は，一対比較におけるトレーニング比較の間隔をもう少し狭めて，発話改善の傾向を詳しく見ていきたい．

謝辞

データベースの作成にあたりご協力頂いた NHK 放送研修センターの方々に感謝致します．また本研究は科研費（25330418）（16K01566）の助成を受けたものである．ここに深く謝意を表する．

参考文献

[1] 林勇吾，クーパーエリック，クリサノフビクター，浦尾彰，小川均，“対話エージェントのコミュニケーションにおける心理特性―スキーマと擬人化に関する検討―，”日本感性工学会論文誌，vol.11, no.3, pp.459-467 (2012).
[2] 岡田将吾，賀小淵，小島量，長谷川修，“自己増殖型ニューラルネットワークを用いたヒューマノイドロボットの発達的言語獲得，”人工知能学会論文誌，vol.22, no.5, pp.493-507 (2007).
[3] 吉本 正祥, 篠崎 隆宏, 岩野 広司, 古井 貞熙，“軽量な画像特徴量を用いたマルチモーダル音声認識，”信学論(D), J95-D(3), pp.618-627 (2012).
[4] 速水 悟, 竹沢 寿幸，“マルチモーダル情報統合システムの研究動向，”人工知能学会誌, vol.13, no.2, pp.206-211 (1998).
[5] 相原 政徳，小田嶋 和幸，畑岡 信夫，“カーナビにおける音声インタフェースの実車評価，”情報科学技術フォーラム講演論文集，8(2), pp.351-352 (2009).
[6] 若松英輝，寺尾太志，山崎夢子，山田光穂，"口唇動作による発話トレーニング法の提案，”信学技報，Vol.113(468), pp.277-280 (2014).
[7] 福島英，“声がよくなる簡単トレーニング，”成美堂出版 (2006).
[8] 鈴池 静，“コミュニケーションにおける声のノンバーバル要素の重要性に関する研究，”早稲田大学，pp40-42 (2010).
[9] 森崇人，“音響・調音音声学でのフォルマントによる多言語の母音比較分析，“名古屋大学学生論文コンテスト(2015).
[10] 大塚貞子，“フォルマント周波数値を利用した母音発音指導の可能性についての一考察，”東京女子大学紀要論，vol.64, no.2, pp.311-333 (2014).