Vol.2020-MUS-127 No.16
Vol.2020-SLP-132 No.16
2020/6/6

声質変換手法を用いた自己音声生成の検討

野々山大樹 [1]　西野隆典 [2]

概要：発声時に自己が聞く音声と他者が聞く音声には違いがある．また，自己の音声を録音し，聴取すると大きな違和感を覚える．この要因の一つとして，音波の伝達経路の違いが考えられる．そこで本研究では，他者が聞いている自己の音声を，自らが聞いている声に変換することを検討する．実験では，録音した自らの音声を他者が聞いている音声とし，口から耳までの伝達関数を考慮するとともに，声質変換手法を用いて音声を変換した．作成した音声が録音音声よりも自己が聞く音声に近い自己音声であるかという観点で主観的に評価した．また，作成した自己音声には，どのような特徴があるかを併せて調査した．実験結果より，録音した音声から直接変換した音声が自己音声に近いと評価される割合が高いことが分かり，声質変換手法を用いて自己音声が生成可能であることが示唆された．

キーワード：**自己音声，声質変換，頭部伝達関数**

1. はじめに

自らの声を録音し聴取すると，大変な違和感を覚える．自らが聞いている自らの声は，口から発せられた音波が，頭部，耳介によって回折や反射して届く音波と，頭内を伝播してくる音波との合成であると考えられる[1]．一方，録音音声や他人が聞く音声は，空間の伝達特性の影響が主であると考えられる．これら音波の伝達経路の違いにより，録音した音声に違和感を覚えたり，自己と他者とで聞こえ方が変わると考えられる．この違和感の解消や，聞こえ方の差を認識することを通じて，個人の会話能力や音楽教育現場での発声技能の向上が可能になると考えられる．そこで本研究では，自己と他者の聞こえ方の差異の客観的理解が可能となるよう，自己音声の生成手法を検討する．

本研究は，録音した音声を自らが聞いている音声へ変換する手法の検討を行う．実験では，発音時に自身が聞いている音声を声質変換手法を用いて生成する．この生成した音声と録音した自身の声との音の違いについて調査を行い，両者の差異について調査した．音声の生成では，口から耳までの伝達特性の考慮の有無について検討するとともに，伝達特性の代替として声質変換手法の適用を試みた．また，生成した音声がどの程度自己音声に近いかをアンケート調査するとともに，処理前後でのスペクトルの差を求め，どのような特徴があるか検討を行った．

2. “自らが聞いている声”の生成手法

本研究では，録音音声をもとに“自らが聞いている声”(以降，自己音声)の生成を行う．被験者の音声を録音し，提案手法により複数の音声を作成する．その後，生成した音声を被験者に評価させ自己音声に最も近いと評価された音声を自己音声とした．

2.1 自己音声のモデル化

本研究で扱う自己音声のモデルを考える．口元での音声信号を $x(t)$ とし，自己音声信号を $y(t)$ とすると，

$$y(t) = h(t) * x(t) + g(t) * x(t)$$
$$= \{h(t) + g(t)\} * x(t) \quad (1)$$

と表すことができる．ただし，$h(t)$ は口元から耳までの頭部表面を伝う音波の伝達特性，$g(t)$ は口元から頭内を伝播する音波の伝達特性を表す．また，$*$ は畳み込み演算を表す．式(1)を周波数領域で表現すれば，

$$Y(f) = \{H(f) + G(f)\}X(f) = A(f)X(f) \quad (2)$$

と積で表現できる．ここで，$h(t)$ はダミーヘッドマイクロホンを用いるなどして計測可能であるが，$g(t)$ はいくつかの報告[2, 3]があるものの計測が困難な伝達特性であると考えられる．そこでこの $g(t)$ に関する伝達特性を声質変換手法により代替することを考える．声質変換手法を適用するため，式(2)を変形する．

$$Y(f) = \left\{1 + {G(f)}/{H(f)}\right\}H(f)X(f) = B(f)H(f)X(f) \quad (3)$$

本研究では，$x(t)$ を変換して得られた自己推定音声信号 $\tilde{y}(t)$ が $y(t)$ に近づくような処理を行うために，式(2)の $A(f)$，および式(3)の $B(f)$ を声質変換手法で代替する．この手順を図 1 に示す．

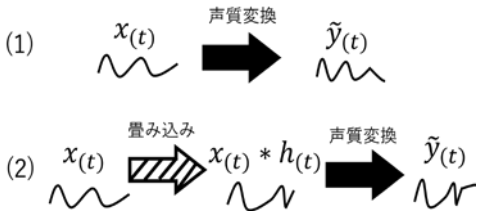

図 1　自己音声の生成手順

1 名城大学大学院都市情報学研究科
Graduate School of Urban Science, Meijo University
2 名城大学都市情報学部
Faculty of Urban Science, Meijo University

Vol.2020-MUS-127 No.16
Vol.2020-SLP-132 No.16
2020/6/6

2.2 声質変換手法による伝達特性の代替

実験では，伝達特性を代替するための声質変換手法として WORLD [4-6] を用いた．WORLD は高品質で柔軟な音声変換技術であり，録音した音声から自己音声に変換する際に細かな変化を加えることが可能であると考えた．WORLD は入力音声に対し，音声を構成する 3 つの要素である f_o (基本周波数), スペクトル包絡, 非周期性指標を抽出し再合成が可能な声質変換ツールである．声質変換は，f_o のシフト，スペクトル伸縮，時間伸縮の 3 つのパラメータで行われる．本実験では，この 3 つのパラメータの内，f_o のシフトとスペクトル伸縮の 2 つのパラメータを変えて自己音声の候補 $\tilde{y}(t)$ を生成する．

実験では, 2 種類の音声信号に対し WORLD を適用して，自己音声の生成を行った．

(1) 録音音声の変換

この条件は，式(2)の伝達特性 $A(f)$ を代替することに相当する．すなわち，録音した音声信号 $x(t)$ を用いて WORLD で変換し，自己音声 $\tilde{y}(t)$ を生成する．

(2) 頭部伝達関数を考慮した音声の変換

この条件は，式(3)の伝達特性 $B(f)$ を代替することに相当する. ここで, 口元から耳までの伝達特性 $h(t)$ はダミーヘッドマイクロホン(B&K 4128)で収録した信号を用いた．伝達特性 $h(t)$ と録音した音声信号 $x(t)$ とが畳み込まれた信号を WORLD で変換し，自己音声 $\tilde{y}(t)$ を生成する．

3. 実験条件

3.1 音声の録音，および候補音の生成

録音環境は雑音の少ないセミナー室で行った．録音は PCM レコーダ(OLYMPUS, LS-P4)を用い，サンプリング周波数 44.1 [Hz]とした．被験者と機材の配置は 図 2 に示すとおりである．なお，PCM レコーダの高さを被験者の口の高さと同じになるように設定した．

被験者は聴力に問題のない 20 代学生 6 名(男性 5 名，女性 1 名)である。それぞれの被験者には，定型のあいさつ文を普段と同じように話すように指示をした. あいさつ文は，「こんにちは　わたしのなまえは　○○(被験者氏名)です」とした．

候補音の生成条件は，f_o のシフトを 0.6 から 0.1 きざみで 1.2 までの 7 通り，スペクトル伸縮を 0.8 から 0.1 きざみで 1.2 までの 5 通りとした．これより，候補音は合計 35 となる．

3.2 候補音の評価

生成した候補音 $\tilde{y}(t)$ がどの程度，自己音声に類似しているかを聴取実験により評価した．評価方法，音声録音を行った被験者自身に，録音音声 $x(t)$ と候補音 $\tilde{y}(t)$ を対にして聞かせ，3 段階で評価させた．評価の段階は次の通りである．

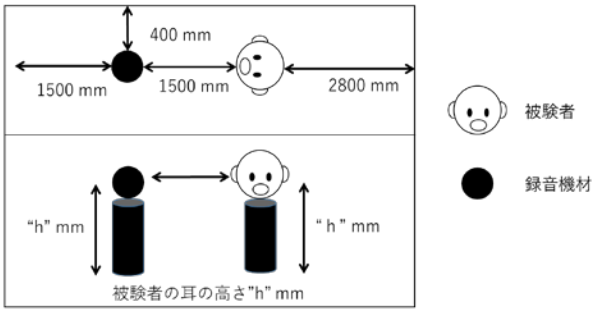

図 2　録音環境

1. 候補音は録音音声より劣っている
2. 候補音は録音音声と同等程度
3. 候補音は録音音声よりも似ている

併せて，候補音の中で最も自己音声に近い音声を一つ挙げさせた．

聴取実験では，雑音の少ないセミナー室にて，PC(HP, ProBook 400)で音声ファイルを再生し，ヘッドホン(Audio-Technica, ATH-A900Z)により聴取させた．また，被験者は評価対象の音を繰返し聴取して良いこととした．

4. 実験結果

4.1 被験者自身の評価結果

被験者自身による候補音の評価結果を表 1 と表 2 に示す. 表 1 は録音音声をそのまま変換した条件，表 2 は頭部伝達関数を考慮した音声の変換の条件の結果をそれぞれ示している. また, 自己音声に最も近いと回答した候補音をゴシック体で表記した．

録音音声をそのまま変換した条件では，被験者 6 名全員が 3 点を付けた候補音が 1 つ以上存在した．これより，録音音声から自己音声に近い音声が生成可能であると考えられる．また，自己音声に近いとされた候補音を生成するパラメータは，被験者ごとに異なることが確認された．

これに対し，頭部伝達関数を考慮した音声の変換の条件では，3 点を付けた候補音が 1 つ以上存在した被験者 6 名中 3 名であった．この原因として，口から耳までの伝達特性には頭部や耳介の形状の個人差があり，今回はダミーヘッドマイクロホンの伝達特性を用いたため，被験者と特性が合わなかったことが考えられる．

4.2 他者による評価結果

この実験では，録音音声 $x(t)$ と，録音音声から変換した候補音の中で被験者自身が最も自己音声に近いと回答した音声を，被験者以外に聴取させ，どちらの音声が普段聞いている被験者の音声に近いかを回答させた．録音した 6 名の音声に対し，被験者と異なる 10 名が聴取実験を行った．なお，評価者の条件として，録音の被験者と週に 1 回以上会話を行っている者とした．

評価結果は，10 名全員がそれぞれの被験者の録音した音声が普段の音声であると回答した．

Vol.2020-MUS-127 No.16
Vol.2020-SLP-132 No.16
2020/6/6

表 1 被験者自身の評価結果
(録音音声をそのまま変換した条件)

(被験者: 男性 A)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	1	2	1	1	1	1
1.0	3	**3**	3	3	3	2	1
1.1	2	3	3	2	2	2	1
1.2	1	1	1	1	1	1	1

(被験者: 男性 B)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	2	1	1
0.9	1	1	3	2	2	1	1
1.0	1	1	**3**	2	2	2	1
1.1	1	1	1	2	1	1	1
1.2	1	1	1	1	1	1	1

(被験者: 男性 C)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	1	1	1	2	**3**	1
1.0	1	1	1	2	2	2	2
1.1	1	1	1	1	2	1	1
1.2	1	1	1	1	1	1	1

(被験者: 男性 D)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	1	1	1	1	3	1
1.0	1	1	1	2	2	**3**	2
1.1	1	1	1	1	1	1	1
1.2	1	1	1	1	1	1	1

(被験者: 男性 E)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	1	1	1	1	1	1
1.0	1	1	2	**3**	2	2	2
1.1	1	1	1	1	1	1	1
1.2	1	1	1	1	1	1	1

(被験者: 女性 A)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	1	1	1	1	1	1
1.0	1	1	2	**3**	2	2	1
1.1	1	1	1	1	1	1	1
1.2	1	1	1	1	1	1	1

表 2 被験者自身の評価結果
(頭部伝達関数を考慮した音声の変換の条件)

(被験者: 男性 A)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	2	1	2	2	1	1
1.0	2	2	**3**	3	2	2	1
1.1	1	2	2	2	2	2	1
1.2	1	1	1	1	1	1	1

(被験者: 男性 B)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	1	1	1	1	2	1
1.0	1	1	2	**2**	2	2	1
1.1	1	1	1	2	1	1	1
1.2	1	1	1	1	1	1	1

(被験者: 男性 C)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	1	2	2	1	1	2
1.0	1	1	**3**	2	2	2	1
1.1	1	1	2	1	1	1	1
1.2	1	1	1	1	1	1	1

(被験者: 男性 D)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	1	1	1	1	1	1
1.0	1	1	1	1	3	**3**	1
1.1	1	1	1	1	1	2	1
1.2	1	1	1	1	1	1	1

(被験者: 男性 E)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	1	1	1	1	1	1
1.0	1	1	2	2	**2**	2	1
1.1	1	1	1	1	1	1	1
1.2	1	1	1	1	1	1	1

(被験者: 女性 A)

スペクトル伸縮 \ f_0 のシフト	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.8	1	1	1	1	1	1	1
0.9	1	1	1	1	1	1	1
1.0	1	1	1	2	**2**	2	1
1.1	1	1	1	1	1	1	1
1.2	1	1	1	1	1	1	1

Vol.2020-MUS-127 No.16
Vol.2020-SLP-132 No.16
2020/6/6

5. 考察

評価結果から，自己音声に近い音声が生成可能であることが示された．また，他者からの評価により，録音音声よりも生成した音声の方が被験者に近いことが確認できた．ここでは，録音音声と生成処理により得られた音声とにどのような差があるか調査した．録音音声 $x(t)$ と生成した自己音声 $\tilde{y}(t)$ に対し，それぞれフーリエ変換を行い，周波数領域で差分 $|X(f)| - |\tilde{Y}(f)|$ を求めた．

図 3 に被験者ごとの 300 ~ 1500 [Hz]のスペクトル差分を示す．各図において実線は差分が 0 を表す．実線より上であれば，録音音声のスペクトルが大きいことを表し，実線より下であれば，生成音声のスペクトルが大きいことを表す. 大まかな傾向として, 男性 A, B では 300～500 [Hz]で，女性 A では 600~800 [Hz]で，それぞれ生成音声のスペクトルが大きいという結果が現れた．これらの被験者では，自己音声は録音音声に比べて声の高さが低いと認識していると考えられる．一方，男性 C は 500~700 [Hz]，男性 D は 300~400 [Hz]で，それぞれ録音音声のスペクトルが大きいという結果が現れた．これより，両被験者は自己音声が録音した音声に比べて声の高さが高いと認識していると考えられる．また，男性 E は差分の変動が大きいが，声の高さに関しては，他の被験者との共通点を発見できなかった．

6. まとめと今後の課題

本研究では，“自ら聞いている声”すなわち自己音声の生成方法として声質変換手法を用いる方を提案した．提案手法を用いて， 2 種類の作成方法で自己音声の生成を行った結果，録音音声を変換した場合の方が，より自己音声であると評価される割合が高いことが確認された．これより，声質変換手法を用いて自己音声が生成可能であることが示唆された．

ただし，被験者が男性 5 名，女性 1 名と非常に少ないことから，さらに被験者数を増やすことが必要であると考える．併せて，頭部伝達関数の個人差の影響についても検討が必要であると考える．また，得られた変換パラメータと身体特徴量などを，統計的アプローチなどで関連付けることが可能であれば，自己音声の生成が容易となると考えられる．

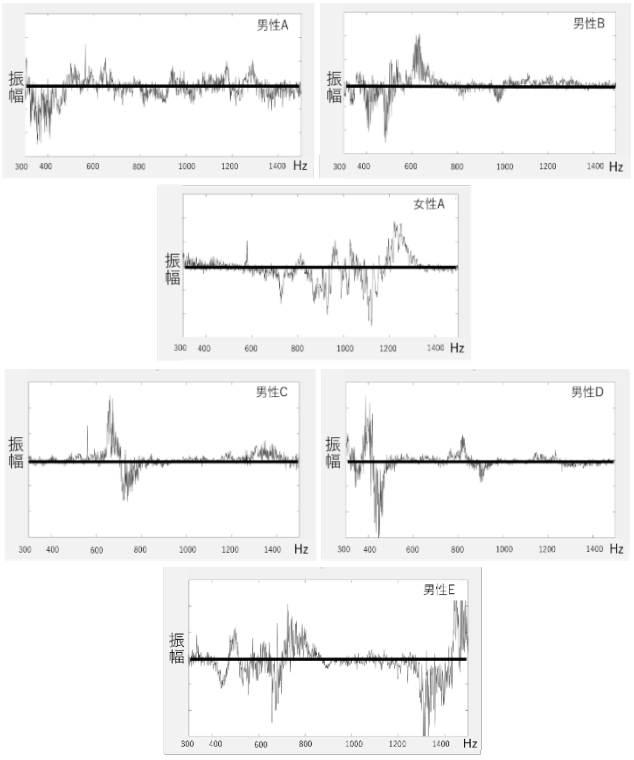

図 3　スペクトル差分

謝辞

本研究の一部は JSPS 科研費 19K12193 の助成を受けたものである．

参考文献

[1] 森幹男，吉田千夏，荻原慎洋，谷口秀次，高橋謙三．自己聴取音に占める気導音と骨伝導の割合の推定．電気学会論文誌，2007，vol.127，no.8，pp.1268-1269.

[2] 鳥谷輝樹，鵜木祐史．話者が知覚する骨導音声の音響的特徴の検討．日本音響学会秋季講演論文集，2017，3-P-10.

[3] 鳥谷輝樹，Peter Birkholz，鵜木祐史．長時間平均スペクトルに着目した骨導音声の伝達特性の分析．日本音響学会春季講演論文集，2018，3-P-5，pp.1297-1300.

[4] M. Morise, F. Yokomori, and K. Ozawa. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE transactions on information and systems, 2016, vol. E99-D, no. 7, pp. 1877-1884.

[5] M. Morise. D4C, a band-aperiodicity estimator for high-quality speech synthesis. Speech Communication, Nov. 2016, vol. 84, pp. 57-65.

[6] 森勢将雅．オープンソース音声合成システム WORLD の現状と課題．情報処理学会研究報告，2014，vol. 2014-MUS-103，no. 68，pp. 1-6.