発話者自身のモーラ音声に基づくテキスト音声合成における検討 —音声の基本周波数やスペクトル特性に関して—*

☆吉岡麻里子，荒井隆行（上智大・理工），安啓一（国リハ研究所），大月春花（NTT データ）

1　はじめに

コミュニケーションの手段として音声は非常に重要であるが，病気により声を失うことがある．例えば，筋萎縮性側索硬化症（ALS）は，筋肉が徐々に委縮し麻痺する病気である．そのため，呼吸筋の衰えにより人工呼吸器を装着する際に自身の声を失う．

現在では，TTS (Text-to-Speech)音声合成[1]や声質変換[2][3]による音声コミュニケーション支援が行われている．その一例として，発話者が必要な音声を録音・登録して利用する「マイボイス」[4]というソフトウェアを取り上げる．このソフトは，日本語の 50 音・濁音・拗音などを登録し，対応する音をつなげて再生することで文章の読み上げを可能にする．また，文の抑揚を出力音で再現するため，各モーラに対し基本周波数の異なる 6 種類の音声データを自動で作成し使用する．これは，従来の音声合成システムに比べ，短時間の録音で簡易にシステムを構築でき，自身の声での意思伝達を希望する患者の要望にも応えている[5]．しかし，マイボイスを利用する患者からは，使い始めてから様々な要望が出てくることが判明した．そのため，録音後の処理による選択肢を増やすことで，患者の要望に細やかに応えることを目指した．本研究では録音データの基本周波数とスペクトル特性に着目した前処理の効果を調査した．

2　実験

2.1 録音

刺激音作成のため，上智大学荒井研究室内の防音室で録音を行った．使用機器は，ディジタルレコーダ (Marantz PMD 660) および，単一指向性マイクロフォン (SONY ECM-23F5) である．録音はサンプリング周波数 48 kHz, 量子化レベル 16 bit の条件下で行った．録音の協力者は話者 1（20 代女性 1 名），話者 2（50 代の ALS 男性 1 名）の 2 名で，表 1 に示した条件で行った．録音した音声は，日本語の五十音・濁音・拗音を含む 140 音，および，日常使うフレーズ 35 語であった．

2.2 実験 1：基本周波数の処理

2.2.1. 刺激音の作成

まず，録音データのサンプリング周波数を Praat [6]で 44.1 kHz に変換し，以下の 3 種の方法でマイボイスに音声データを登録した．

(1) セット A：モーラごとに音声データをマイボイスに登録した．
(2) セット B：セット A に登録されている音声データの基本周波数の平均値をマイボイス用の Praat スクリプト[7] にある get mean pitch で求め，すべてのモーラの基本周波数が，各話者の平均基本周波数から±5 Hz 以内に収まるように手動で調整した．
(3) セット C:セット A の音声データに対し，マイボイス用の Praat のスクリプト[7] にある Adjust F0 で話者の平均基本周波数に調整したデータを作成した．

以上のように登録した音声からデータセットを作成し，15 文の刺激音を作成した．

2.2.2. 実験方法と実験環境

刺激音の評価実験には，Praat [6]を用いた．同じ文を読み上げた 2 種類の音声を流し，実験参加者にはより自然に聞こえた音声を選択させた．各刺激は無作為に提示され，最大 2 回まで聴取可能とした．聴取実験は，上智大学荒井研究室内の防音室で行われた．刺激音は，PC に接続した USB オーディオ・インタフェース (Roland, UA-25EX) を介し，ヘッドホン (SENNHEISER, HDA200) から騒音レベ

---

* Effect of Acoustic Adjustments in the Text-to-Speech Synthesis Based on Speaker's Moraic Sound: Focus on the Fundamental Frequency and the Spectral Characteristics of Speech, by M. Yoshioka, T. Arai (Sophia Univ.), K. Yasu, (Research institute national rehabilitation center for persons with disabilities), and H. Otsuki, (NTT data Corp.).

ルで 50 dB で提示した. 聴取実験の参加者は, 18 歳以上の男性 7 名，女性 13 名（平均年齢 21.4 歳）の日本語母語話者であった.

2.3 実験 2：周波数フィルタの処理

2.3.1. 周波数フィルタ条件

本実験では, 14 種類の周波数フィルタを使用した．フィルタ設計に際し，境界周波数を 2500 Hz，3750 Hz，5000 Hz とし，4 つの周波数帯域 (Band1～Band4) に分割した. Band1 は F1 と F2, Band2 は F3, Band3 は F4, Band4 は F5 以上が主に含まれる帯域とし，表 3 に示す特定の帯域を 20 dB 強調するフィルタ処理を施した. 表中の○がついている Band は, 20dB 強調されている帯域を示している.

2.3.2. 刺激音の種類

本実験では，2.3.1.節に示した 14 種類の周波数フィルタと，処理なし（以下，Original）の刺激音を作成した．使用した文は，A:「爆音が銀世界の高原に広がる」及び，B:「栄養満点の料理をお母さんと一緒に作った」の 2 文である. 合計 30 種類の刺激音は全て，マイボイスで作成した.

表 1　録音条件

話者	録音方法
話者 1	基本周波数にばらつきが出るように録音
話者 2	基本周波数をほぼ一定にして録音

表 2　刺激音の処理条件

音声	処理
A	音声データを登録
B	話者の平均基本周波数から±5 Hz 以内に手動で調整
C	Praat スクリプトで話者の平均基本周波数に自動で調整

表 3　周波数フィルタの概要

処理条件	Band1 (2500 Hz以下)	Band2 (2500-3750 Hz)	Band3 (3750–5000 Hz)	Band4 (5000 Hz以上)
Proc1	○			
Proc2	○		○	
Proc3	○		○	○
Proc4	○			○
Proc5	○	○		
Proc6	○	○		○
Proc7	○	○	○	
Proc8		○	○	○
Proc9		○		
Proc10		○		○
Proc11		○	○	
Proc12			○	○
Proc13			○	
Proc14				○

2.3.3. 聴取実験

聴取実験は，上智大学荒井研究室内の防音室にて行った. 刺激音は, PC に接続した USB インタフェース(Roland, UA-25EX)を介し，ヘッドホン(SENNHEISER, HDA200)から提示した.

実験参加者は話者 2 と話者 2 の知人 4 名であった．話者 2 本人には，実験によって作成されたマイボイスの音声データを渡し，自分の声であるマイボイスとして使いたいか，という基準で評価をしてもらった．知人 4 名には，<Original>，および，<周波数フィルタ処理後（以下, Proc）>の 2 種類の音声を聞かせ，被験者のマイボイスとして使ってほしいかどうか, という基準で, ①1 番目の音声が良い，②2 番目の音声が良い，③どちらとも言えない，の三択で評価をさせた．音声は無作為に提示した．評価後に，音声を選んだ理由についてアンケート方式で答えさせた.

3　結果・考察

3.1 実験 1：基本周波数の処理

聴取実験の結果において, 話者ごとに A より B を選んだ回数，A より C を選んだ回数, B より C を選んだ回数の 3 種類を求めた．この回数から，話者ごとに Scheffe の一対比較法でセット A, B, C の総合的な評価を算出した．その結果，話者 1 では C > B > A の順に評価が高く, A-B 間, A-C 間の処理（主効果）に有意差が認められた [$F(2, 59) = 187.36$, $p < .01$] （図 1）．これにより，B, C の処理は，話者 1 のケースのように録音時点で基本周波数がばらついた音声データに対して，マイボイスの出力音の聞こえをより自然にする傾向が見られた．そして，基本周波数を手動で調整する処理と，Praat [9]で調整する処理に有意差は見られなかった．このことから，基本周波数のばらつきは，マイボイスの出力音を聞き取りにくくする要因の一つと考えられる.

また，文ごとに見た評価順序では，過半数の 10 文で C > B > A の順に自然に聞こえたと評価された．しかし，残り 5 文では，B > C > A と評価された．これにより，B, C の処理に関しては文ごとに評価が変わる可能性があることが分かった．そして，この 5 文の基本周波数曲線を比較した結果，曲線には大きな

差異がほとんど見られなかった．このことから，基本周波数以外の要因が，評価に影響を与えた可能性が考えられる．

一方，話者 2 では A > B > C の順に評価が高く，A-C 間，B-C 間の処理（主効果）に有意差が認められた [$F(2, 59) = 143.93$, $p < .01$]（図 2）．これにより，B, C の処理は，話者 2 のケースのように録音時点で基本周波数をほぼ一定にして録音した音声データに対して，必ずしもマイボイスの出力音の聞こえを自然にするわけではないと考えられる．特に，C の処理での評価が A よりも低く出た原因の一つとしては, Praat [6]が登録音声の基本周波数を誤って読み込んだ可能性がある. そのため, 基本周波数を調整した音声データが，実際の話者の基本周波数とは異なる値で出力されたと考えられる．

3.2 実験 2：周波数フィルタの処理

聴取実験の結果より, Original よりも<Proc>が良いと答えた人数の割合を求めた（図 3）．その結果, Proc7 がもっとも高い評価 (58.3%) を得た．Proc7 の次に Proc3 の評価が高く，53.9%であった．<Proc>が良いと評価した人のフィルタ処理による分散分析を行った結果, 周波数フィルタ処理の違いによる有意差は見られなかった．また，各周波数帯域の有無による評価の違いを t 検定により求めた(図 4)．その結果，Band1 のみ周波数帯域の有無による有意差が見られた($p < .05$). Band2～4 では有意差は見られなかった．アンケートの結果から，<Proc>が良いと評価した理由については，「a:音声が聞き取りやすい」が半数近くを占めていた．

<Original>よりも良いと評価された Proc7 と Proc3 の 2 つに着目する．この 2 種類は，Band1, Band3 が強調され，母音・子音の識別が十分に可能である点が共通していた．また図 4 より，Band1，および Band3 のある処理が高い評価を得る傾向があったことから，母音や子音の識別に重要なフォルマントが強調されていた方が良いと考えられる．さらに，t 検定の結果から，Band1 の有無による評価の違いには有意差が見られたが，Band3 では有意差が見られなかったので，子音よりも母音の識別の方が個人性判断に重要であることが示唆される．

また，文により，評価の高いフィルタが異なるという結果が得られた．これは，各音のフォルマントが異なるためであると考えられる．これにより，音ごとにフォルマントの値を考慮したフィルタをかけることによって，より有効なフィルタの効果が得られることが示唆される．また，参加者ごとに周波数フィルタの嗜好が見られた．

他にも、聴取実験で,「③どちらとも言えない」と選択した処理にも着目した．アンケートからは，Original と Proc の音声に違いが感じられないという理由であった．特に，半数の人が③と選択した Proc6 は F1，F2，F3 を

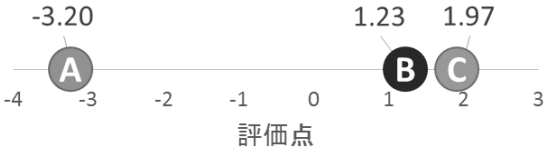

図 1　話者 1 における心理尺度の評価点

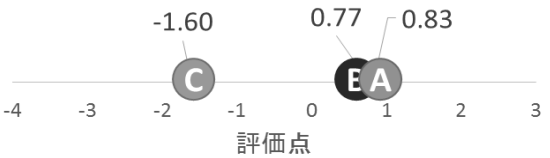

図 2　話者 2 における心理尺度の評価点

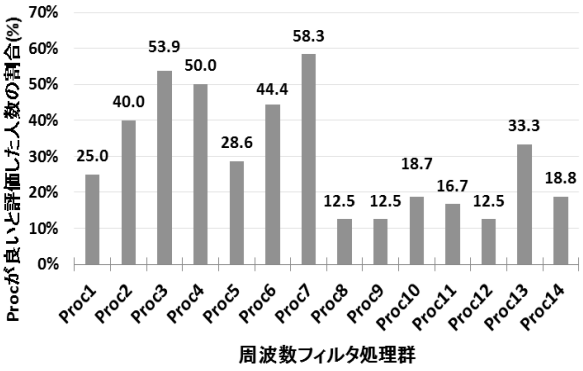

図 3　< Proc >が良いと評価した人数の割合

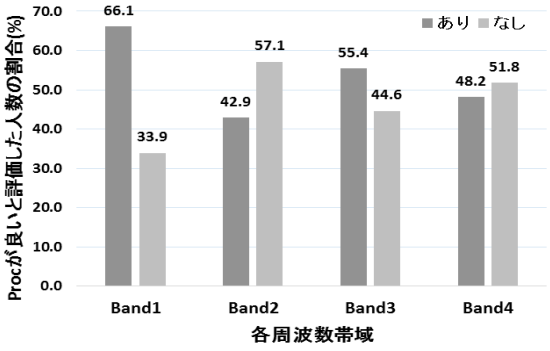

図 4　Band1～4 の有無による評価の違い

強調していた．このことから，F1，F2 により母音の識別が十分可能な場合，F3 が強調されていても音声に大きな変化はないと考えられる．ただし，話者の声が普段から低めであるため，参加者からは「低くて響きのある声が，より話者らしい」という意見があった．そのため，人間の耳に敏感で音声の響きに貢献すると言われている F3 が強調されていると，音声に響きが出て，話者らしさをマイボイスでも表現することができたと考えられる．

4　周波数フィルタ処理の応用

周波数フィルタ処理の応用として，マイボイスで作成したボイスメールに処理をかける実験も行った．話者 3（日本人女性）は，人工呼吸器を装着しており，自身の声による発話が困難である．そのため，話者 3 は自身の声を失う前に録音した音声データセットによる「マイボイス」を使用している．話者 3 がそのデータセットを録音したときには症状が進行しており，その結果「マイボイス」の音声はささやき声となっている．

刺激音に用いる音声は，著者が実際に話者 3 とメールを交換した際に添付されていた音声とした．この音声は，話者 3 が普段使用している話者自身の声による「マイボイス」により作成されたものである．この実験では，28 種類の周波数フィルタを準備した．表 1 に示す Proc1～Proc14 の他に，概要は表 1 と同じで，境界周波数のみを変更した Proc15～Proc28 を作成した．女声の平均的なフォルマントの値を参考に，Proc15～Proc28 において，境界周波数は 3000 Hz, 4500 Hz, 6000 Hz とした．ただし，ささやき声は通常発声のときとは周波数構造が異なるため，フォルマントは関係なしに複数のフィルタを用いて処理した後，その中から良いものを選ぶこととした．

音声評価は，話者 3 にのみ 10 種類の刺激音を音声ファイルとして電子メールで送り，原音声と比較して評価を行った．その結果，実験 2 とは異なり，Proc6 と Proc13 が原音声と比較して高い評価を得た．

5　おわりに

本研究では，マイボイスに登録する音声に，基本周波数とスペクトル特性に着目した処理を行った．基本周波数に対する処理に対しては，聴取実験で出力音の評価を行った．その結果，録音時点で基本周波数をほぼ一定で録音できれば，出力音の聞こえは自然になり，基本周波数を揃える処理が必ずしも必要がないことが分かった．また，録音時点で基本周波数にばらつきがある音声データに対しては，基本周波数を揃える処理がマイボイスの出力音の聞こえをより自然にする可能性が示唆された．一方，周波数フィルタ処理では，処理をしていない音声よりも処理後の音声を使いたいという回答を得た処理が存在した．このことから，フィルタ処理によって多様な音質の声を提供し，患者の選択の幅を広げる可能性がある．

今後は，音響的特徴を考慮したフィルタの設計や，当事者自身で音質を簡単に調整できるインタフェースの開発などにより，更なる発展が期待される．また，モーラや母音ごとのフォルマントに合わせたフィルタをかけることで，より話者の嗜好に合ったフィルタの効果が期待される．

謝辞

本研究を行うにあたり，マイボイス開発者の吉村隆樹さん，東京都立神経病院の本間武蔵先生，慶応義塾大学の川原繁人先生，音声データを提供してくださった藤元健二さん，酒井惠子さん，論文執筆にあたりご助言くださった上智大学特別研究員の井下田貴子さん，そして実験に協力してくださった皆様に感謝申し上げます．

参考文献

[1] A. Iida *et al.*, Int. J. of Speech Tech., 6, pp.379-392, 2003.
[2] 加島他，音講論（秋）, 251-252, 2006.
[3] Yamagishi *et al.*, 音響誌，67(12), 587-592, 2011.
[4] HeartyLadder, Retrieved from http://takaki.la.coocan.jp/hearty/, 2014.
[5] 篠原他，第 28 回人工知能学会全国大会, Retrieved from https://kaigi.org/jsai/webprogram/2014/pdf/17.pdf, 2014.
[6] Praat: doing phonetics by computer, Retrieved from http:.//www.praat.org/
[7] マイボイスの紹介およびマイボイス用の Praat scripts, Retrieved from http://user.keio.ac.jp/~kawahara/myvoice.html, 2014.