情報処理学会論文誌 Vol.61 No.4 807–816 (Apr. 2020)

声のピッチ感の錯覚と疑似歌声・疑似ささやき声による検討

内田 照久[1,a)] 森勢 将雅[2]

受付日 2019年6月26日, 採録日 2020年1月16日

概要：声道長の縮小や拡大を模して，スペクトル包絡の周波数軸を伸長圧縮した声質変換音声では，基本周波数（f_o）の高低関係と，声の高さの印象評価が逆転することがある．このピッチ感の錯覚が起こる条件を精査するため，標準 f_o 軌跡，平坦 f_o 軌跡，逆相 f_o 軌跡，疑似歌声，疑似ささやき声による評価実験を行った．その結果，標準 f_o 軌跡と逆相 f_o 軌跡ではピッチ感の錯覚が生起したが，平坦 f_o 軌跡では消失した．疑似歌声条件では錯覚量はやや低下したが，f_o の高低によるピッチ感への支配性は必ずしも検証されなかった．一方，f_o や調波構造が存在しない疑似ささやき声では，スペクトル周波数軸の伸縮に呼応した高低判断がなされていた．

キーワード：ピッチ，声質，声道長，スペクトル重心，WORLD

Investigation of Voice Pitch Illusion Using Quasi Singing Voice and Quasi Whisper

Teruhisa Uchida[1,a)] Masanori Morise[2]

Received: June 26, 2019, Accepted: January 16, 2020

Abstract: When the speech is converted by expansion and compression of the fundamental frequency axis of spectrum envelopes simulating shortening and extension of vocal tract length, the timbre of the voice changes systematically. In the converted voice, a reversal of the relationship between height of fundamental frequency (f_o) and impression of voice pitch is often observed. In order to examine the conditions that cause this voice pitch illusion, evaluation experiments were conducted using original f_o patterns (standard intonation), flat f_o patterns, reversed-phase f_o patterns, quasi singing-voice, and quasi whisper. As a result, a pitch illusion occurred in the original f_o patterns and the reversed-phase f_o patterns, then disappeared in the flat f_o patterns. The amount of illusion slightly decreased under the quasi singing-voice condition, but the dominance of pitch due to the height of f_o was not verified. On the other hand, in the case of quasi-whisper in which f_o and harmonic structure do not exist, the height of voice was evaluated in response to the expansion and compression of the spectral frequency axis.

Keywords: pitch, tonal quality, vocal tract length, spectral centroid, WORLD

1. はじめに

1.1 声の音色と声道長：スペクトル周波数軸の伸長圧縮

人の声の高さの知覚を検証することは，音楽，特に歌声の認知を考えるうえで不可欠である．本研究は，言葉としての音声と，音楽としての歌声を比較対照していくことで，音楽の知覚に迫ることを目的とする．

さて，人の発声器官を音響管ととらえた場合，声道の長さが変わると，声の音色は系統的に変化する．一般に，声道長が短いと「細い声」となり，長いと「太い声」になる．この声道長の違いに起因する声色の変化は，音声信号処理の観点では，音声の平滑化スペクトルの周波数軸を伸長，圧縮することで模すことができる [1]．このスペクトル周波数軸の伸縮によって，声道長（vocal tract length）の操作を模した声質の変換は，話者の秘匿などにも有効である [2]．

[1] 大学入試センター研究開発部
Research Division, the National Center for University Entrance Examinations, Meguro, Tokyo 153–8501, Japan
[2] 明治大学総合数理学部
School of Interdisciplinary Mathematical Sciences, Meiji University, Nakano, Tokyo 164–8525, Japan
[a)] uchida@rd.dnc.ac.jp

1.2 スペクトル重心とピッチ感への認知的バイアス

一般に，複合音のピッチは，基本周波数に従うとされる．しかし，スペクトル周波数軸を伸長して，声道長の縮小を模した音声では，元の音声よりもスペクトル重心（spectral centroid）が上昇する．この高スペクトル重心となった音声は，たとえ基本周波数（fundamental frequency，f_o）が元の音声と同一であっても「高い声」として知覚される．逆に，周波数軸を圧縮して，声道長の拡大を模した低スペクトル重心の音声では「低い声」の知覚印象が生起する [2]．

このスペクトル重心は，人が知覚する音の明るさに対応するとされる．ここで，音声のサンプリング周波数を f_s，スペクトル分析時の FFT 長を N，スペクトルの離散周波数番号 k における振幅を $A(k)$ とすると，スペクトル重心は，下記の式 (1) と式 (2) によって定義される [3]．

$$S_C = \sum_{k=0}^{N/2} f(k)A(k) \Big/ \sum_{k=0}^{N/2} A(k) \tag{1}$$

$$f(k) = k(f_s/N) \tag{2}$$

このスペクトル周波数軸の伸縮にともなうスペクトル重心の移動は，単に声の音色を系統的に変えるだけにとどまらず，声の高さの印象，声のピッチ感にも影響を及ぼす．著者らはこれまでに，このピッチ感への認知的バイアスは，基本周波数の高低の効果に重畳する形で，加算的に作用していることを報告している [4]．

1.3 声のピッチ感の錯覚

さらに，その認知的バイアスは，基本周波数の高低関係と，声の高さの印象評価を逆転させるほど，強い効果があることが分かってきた．実際に，著者の先行研究では，一定の条件の下では，「基本周波数が低い」音声であっても，「高い声」として誤って知覚されてしまうという「声のピッチ感の錯覚（voice pitch illusion）」が報告されている [4], [5]．

具体的には，短い声道長を模した高スペクトル重心の音色で，基本周波数が低い「細くて低い声」と，長い声道長を模して低スペクトル重心であるが，基本周波数は高い「太くて高い声」の，2 つの音声の対を聞いて比較をする．すると，基本周波数は低いにもかかわらず，高スペクトル重心の「細くて低い声」の方が「高い声」と評価される．すなわち，音色が異なると，基本周波数の低い方が高い声として知覚されてしまう現象が起こる．そして，それはかなり頑健に安定的に生起する [4]．

このような基本周波数が変動する声の高さの知覚については，声のビブラートのピッチ知覚の研究がある [6]．そこでは，ビブラートがかかっている音を聴いて，そのピッチと一致するように，ビブラートのない基本周波数を持つ音を調節するピッチ・マッチングの手法が用いられている．そして，音楽教育を受けた実験参加者がピッチ・マッチングを行った場合，ビブラート音のピッチ知覚は，基本周波数の変動のほぼ平均と対応することが示されている．

一方，本研究で扱う言語音声は，不規則に変動する基本周波数の軌跡を持ち，それが，数秒から 10 数秒にまで及ぶ．そのため，ここでいうところの声の高さは，単一の音高に対応させることのできるピッチ・マッチング可能な音の高さとは性質が異なり，むしろ，声の高さの印象に近い．そこで，本稿で扱う声の高さについてはピッチとは呼ばず，ピッチ感と区別して呼ぶ．

1.4 f_o 軌跡の形状と認知的バイアスの変動

この高さの印象の逆転を引き起こす原因となるピッチ感への認知的バイアスの強度は，f_o 軌跡の形状によって，大きく変動することが分かってきた [4]．

まず，標準的な普通の発話のイントネーションのときは，認知的なバイアスが非常に強く効き，基本周波数の高低と声の高さの印象が逆転する．すなわち，ピッチ感の錯覚が生起する．

しかし，ロボット声のようなフラットな f_o 軌跡のときには，認知的バイアスの効果は非常に弱くなり，高さの印象は，基本周波数の高低関係に従うようになる．その場合，錯覚は消失することになる．

これらの知見からは，スペクトル重心の違いから生じる認知的バイアスはつねに安定している訳ではなく，f_o 軌跡の動的特性に応じて変動していることが分かる．なかでも f_o の変動幅の大きさは，認知的なバイアスの強度の変移を引き起こしている．すなわち，変動幅の拡大は認知的なバイアスの効果を促進し，縮小はバイアスを抑制する．

この振舞いについて，ピッチ感の評価に影響を与える音響属性の支配性の観点からとらえ直してみる．f_o 軌跡がフラットなときは，「基本周波数の高低」がピッチ感への支配的な役割を担っている．しかし，f_o の変動幅が大きくなるにつれて，今度は「スペクトル重心の高低」が支配的役割を担うように変わる．これは，f_o 変動幅の大きさに応じて，ピッチ感を支配する音響属性が，緩やかに移行しているととらえることができる．

1.5 疑似歌声と疑似ささやき声による検討

しかし，この認知的バイアスやピッチ感の錯覚に関する詳細な生起条件や消失条件などの全体像は，まだほとんど分かっていない．f_o 変動幅の要因についても，フラットな f_o 軌跡区間がどのくらいの時間長になったら「基本周波数の高低」の支配性が優位になるのか，といったことも明らかでない．

一方で，歌声では音楽を成立させるうえで，基本周波数が音高として支配的な役割を担っているのは明らかである．f_o 軌跡は，音楽的旋律として遷移する．声色の異なる別の歌い手が発する歌声であっても，基本周波数が同じであれ

ば，楽音としての音の高さは等価であると見なされる．すると，歌声では，「基本周波数の高低」がピッチ感の評価でも支配的な役割を担っていると想定される．

また，f_o 軌跡の形状だけに原因をもとめるのではなく，平滑化スペクトルそのものの影響についても検討する必要がある．ピッチ感への認知的バイアスは，スペクトル周波数軸を伸縮させた結果として生じている．すると，スペクトル包絡の違いによって生ずる効果についても検討しておかなければならない．そして，その測定のためには，f_o や調波構造が存在しない，ささやき声のような音声を用いた検証が有効であると考えられる．

そこで本研究では，従来からの f_o 変動幅の要因に加えて，疑似歌声と疑似ささやき声を用いて，声のピッチ感にかかる認知的バイアスについて検討する．

歌声については，f_o 軌跡を操作して疑似的に模擬する．比較的短時間のフラットな f_o が，音楽的音階のステップで遷移する疑似歌声（quasi singing-voice）を生成して用いる．

また，従来の研究では，スペクトル包絡の違いによる単独の効果量を検討するために，雑音駆動音声（noise-vocoded voice）を用いてきた [4], [7]．しかし，当然のことながら，雑音駆動音声は聴感上，どうしても不自然な印象が生ずる．そこで，実験条件の基幹部分では雑音駆動方式を用いることで，f_o や調波構造を取り除くことができるメリットを生かしつつも，より自然な声に近い，疑似ささやき声（quasi whisper）を生成して用いる [8]．

2. 変換音声の対比較による評価実験

2.1 聴覚実験の条件設定

本実験では，まず従来の研究 [4] での音声（16 kHz サンプリング）と比べて，より高品質な広帯域の音声（44.1 kHz サンプリング）を用いて f_o 変動幅要因の追試を行う．

言語音声のデータ通信が目的ならば，狭帯域音声の検証だけで十分である．しかし，音楽をターゲットとした場合には，より高品質な広帯域の音を前提とした検証が不可欠である．そこで本研究では，新たに広帯域の音源を用いて，ピッチ感の錯覚の生起，認知的バイアスの発現についての検証実験を行う．

そこでは，実験条件 1：標準的な f_o 軌跡，実験条件 2：フラットな f_o 軌跡，実験条件 3：その間となる中庸な f_o 変動幅であるが f_o 軌跡の高低が逆相となる条件，を設定する．なお，ここで f_o 軌跡が逆相となる条件を設定するのは，言語音声に特有のアクセントやフレーズ成分などの特徴を取り除くためである．逆相 f_o 軌跡にしてアクセントやフレーズ成分を破壊することにより，純粋に f_o 変動幅の大きさの関数として，ピッチ感の認知的バイアス量を測定することを意図している．

また，歌声でのピッチ感の評価への f_o の高低の支配性の検証のために，実験条件 4：疑似歌声の条件を設定する．さらに，ピッチ感へのスペクトル包絡単独の効果量の測定のために，実験条件 5：疑似ささやき声を用いる実験条件を設ける．

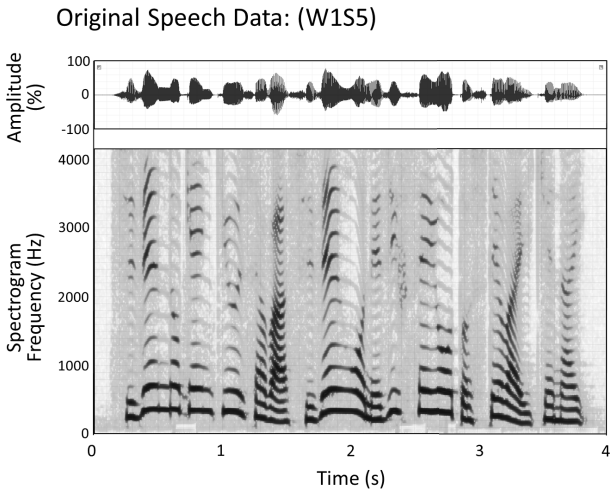

図 1 原音声の音声波形とスペクトログラム（W1S5：自然の研究者は，自然をねじ伏せようとしてはいけない）

Fig. 1 Example waveform and spectrogram of original speech (W1S5).

そして，各条件の音声ごとに，スペクトルの周波数軸を伸縮する．さらに，実験条件 1〜3 ではそれぞれの f_o 軌跡の形状を mel 軸上で維持したまま，また，実験条件 4 の疑似歌声では形状を log 軸上で維持したまま，f_o 軌跡全体を昇降させる．そのうえで，「低スペクトル重心・高 f_o 軌跡」と「高スペクトル重心・低 f_o 軌跡」となる音声の比較対を組む．そして，声の高さの印象評価を含む音声の印象評定を行い，対比較による聴覚実験を実施する．

2.2 原音声

音声データベース SRV-DB [9] の ATR25 文の中から，男女各 2 名の計 4 名が発話した，異なる 8 種類のモノローグの音声を選んで，原音声として使用した（表 1）．なお，話者の性別の要因は実験条件ではないが，f_o の平均やスペクトル重心に関係する声道長が系統的に異なるので，後段の分析では分けて扱う．

ここで，原音声（W1S5）の音声波形とスペクトログラムを**図 1** に示す．スペクトログラムをみると，基音とその倍音からなる縞模様が等間隔に並んでいる．この縞の一番下にある曲線が，声帯音源の基本周波数の f_o 軌跡に相当する．

さらに，このスペクトログラムを垂直方向に見てみると，縞模様のなかには濃淡があり，特定の周波数帯域で倍音成分が強かったり，弱かったりする様子が読み取れる．音声を発するとき，口や顎，舌や唇などの動きによって，声道の中には，いくつか異なる大きさの空間が作られる．それぞれの空間は固有の共鳴周波数を持つため，それに近い倍

音は強調されることになる．それが，縞模様の濃淡を作り出している．なお，高調波が強調される周波数を，低いものから順に，第 1，第 2，第 3 フォルマント（formant）と呼ぶ．

なお，この縞模様の細かい構造から，声道の共鳴特性を抽出して，より滑らかに表現したものは平滑化スペクトルと呼ばれる．

2.3 平滑化スペクトルと f_o 軌跡の変換

声道長の操作を模したスペクトル周波数軸の伸縮，および f_o の操作には，WORLD v0.2.0_7 [10], [11] を用いた．

WORLD は，vocoder ベースの分析合成方式として，音声を極めて高い品質で様々に変換することができる．まず，原音声を分析することで，音声を形作る音響的要素に分解する．はじめに，声帯振動に相当する f_o を抽出し，その後，声道伝達関数に相当する平滑化スペクトルをもとめる．さらに，スペクトルの帯域ごとに，調波構造によらない非周期性の成分が，どのくらい含まれているかを表す非周期性指標を算出する．

これらの要素は，それぞれ独立に操作することができる．そして，操作した要素を組み合わせて音声を再構成することで，品質を維持したまま，任意の変換音声を合成することができる．そこで，本実験では，後述する各種の変換処理に，この WORLD を使用した．

(1) スペクトル周波数の伸縮操作

声の音色の操作のために，原音声の平滑化スペクトルの周波数軸を 0.9 倍，1.111（1/0.9）倍の 2 段階で伸縮を行った．これは，声道長の拡大と縮小に相当する．この方法は，文献 [3] で紹介しているスペクトル包絡の伸縮法によるものである．α 倍の音響管の長さの伸縮は，共鳴周波数を $1/\alpha$ 倍にすることに相当する．すなわち，パラメータ α は，スペクトル包絡を α 倍に伸縮する効果があり，これは声道全体を $1/\alpha$ 倍することと等価であると解釈される．なお，文献 [3] での提案法に準じて，変換時には品質維持のために，平滑化されたパワー・スペクトルを，いったん，対数パワー・スペクトルに変換する．そのうえで，周波数軸を α 倍したものを線形補間でもとめ，再びべき乗してパワー・スペクトルに戻して変換に使用した（**図 2**）．

この方法を適用した場合，平滑化スペクトル周波数軸を伸長するとスペクトル重心が上昇し，圧縮すると下降する．しかし，同一のスペクトルであっても，駆動する f_o 軌跡が異なると高調波の位置も変わるため，最終的なスペクトル重心は変ってくる可能性がある．したがって，本実験でのスペクトル重心の操作は，スペクトル包絡の周波数軸の伸縮による近似的な操作である．

なお，原音声の伸長時，圧縮時のいずれでも，変換後の音声のスペクトル成分が存在する帯域の範囲を揃えておくため，帯域の上限を 16 kHz に制限した．この高域制限は，

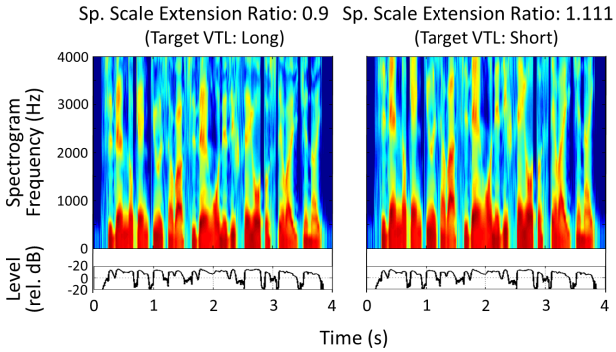

図 2 変換音声の平滑化スペクトルと音圧レベル（W1S5）
Fig. 2 Example smoothed-spectra and level of converted speech (W1S5).

変換後の音声の信号処理でなく，スペクトルに直接重みづけをして，合成時に高域制限を実現するものである．ここでは遮断周波数の ±2,000 Hz の範囲に遷移域を設けて，パワー・スペクトルを減衰させたものを用いた．具体的には，14 kHz〜18 kHz のスペクトル周波数（f）の範囲で式 (3) の w 値でスペクトルに直接重みづけをして高域の制限を行った．

$$w = -0.00025f + 4.5 \tag{3}$$

(2) f_o 軌跡の昇降操作と形状変換

f_o 軌跡のすべての操作は，音高の心理的な知覚的尺度として提案されている mel 尺度上で行った．これは，mel 軸上での値の差が同じであれば，人が感じる音高の差が同じになるように意図して提案されている尺度である．

具体的な手続きとしては，周波数 f（Hz）を式 (4) によってメル値 m（mel）に変換したうえで，f_o の操作を行った [12]．

$$m = 2595 \log_{10}(1 + f/700) \tag{4}$$

そのうえで，実験条件 1-4 の音声に対して，以下の変換操作を行った．

f_o 軌跡の昇降　比較対を作るための基本周波数の昇降のため，f_o 軌跡の高低位置を ±10 mel で昇降させた．これは，後述する f_o 軌跡の形状にかかわる実験条件 1-3 で生成するすべての音声に適用した．

実験条件 1-3：f_o 変動幅と形状の操作　f_o 軌跡の mel 平均を中心軸とし，標準偏差（SD）を f_o の変動幅として操作した．そして，実験条件 1：オリジナル 1.0 倍，実験条件 2：ほぼフラットな 0.02 倍，実験条件 3：f_o を中心軸で高低反転した逆相 f_o 軌跡（$r.p.$）での 0.4 倍，を設定した．

この実験条件 1-3 の操作の後，f_o 軌跡をメル値 m（mel）で計算してきた変換値は，式 (5) で逆変換して周波数 f（Hz）に戻す．

$$f = 700(10^{(m/2595)} - 1) \tag{5}$$

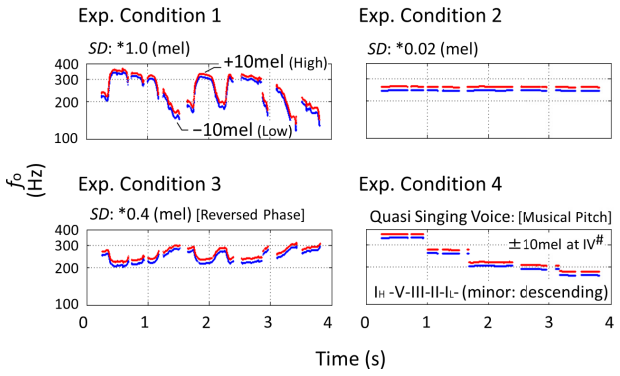

図 3 変換音声の f_o 軌跡の例（W1S5）

Fig. 3 Example f_o contours of converted speech (W1S5).

表 1 原音声の属性・提示順序と割り当てた疑似歌声条件

Table 1 Attributions of original speeches and assigned quasi singing-voice condition.

Speech wave data (ID) *	Quasi Singing Voice Condition **
M1S1 †: ATR_AM01_0800_000	Major: Ascending
M1S2 ‡: ATR_AM01_0800_004	minor: Descending
M2S3 †: ATR_AM03_0800_011	Major: Descending
M2S4 ‡: ATR_AM03_0800_018	minor: Ascending
W1S5 ‡: ATR_PF00_080_005	Major: Ascending
W1S6 †: ATR_PF00_080_010	minor: Descending
W2S7 ‡: ATR_PF01_080_003	Major: Descending
W2S8 †: ATR_PF01_080_008	minor: Ascending

Sampling frequency: 44.1 kHz, Resolution: 16 bit linear
* M: Men, W: Women, S: Script
** Ascending: I_L-II-III-V-I_H / Descending: I_H-V-III-II-I_L
† [Low Sp. - High f_o] –> [High Sp. - Low f_o]
‡ [High Sp. - Low f_o] –> [Low Sp. - High f_o]

そして，得られた物理量である周波数 f（Hz）に基づいて，WORLD で変換音声を合成する（**図 3**）．

実験条件 4：疑似歌声の生成 f_o 軌跡を音楽的な音階に変換した．長調（Major）と短調（minor），それぞれの上行（Ascending：I_L-II-III-V-I_H），下行（Descending：I_H-V-III-II-I_L）の各条件を，原音声ごとに割り当てた（**表 1**）．音高を安定させる区間の f_o の変動幅は，実験条件 2 と同じ標準偏差（*SD*）の 0.02 倍とした．また，比較対での f_o 軌跡の昇降の位置は，原音声の mel 平均を ±10 mel した高さが，各音階での IV# となるように変換した（図 3）．

具体的な操作は次のとおりである．時間面では，それぞれ時間長が異なる原音声ごとに音声区間を等分に 5 分割する．分割された各部分の f_o に，後述する音階からの 5 つの音高に相当する周波数を割り当てる．したがって，音韻の途中で音高が変化することがある．

次に，各実験刺激での音階のキーにかかる周波数を定める．はじめに原音声の f_o の mel 平均をもとめる．それを ±10 mel した値 [mel] から，2 つの周波数 [Hz] をもとめる．そして，それぞれを 1 オクターブのちょうど真ん中にあたる IV# として扱う．Major の場合は，移動音階でのファ # に相当する．そして，その IV# の周波数を固定した形で平均律の音階を形成する周波数を算出する．はじめに Major：Ascending は I_L-II-III-V-I_H なので，ド（低）・レ・ミ・ソ・ド（高）となる．+10 mel を規準としたものと，−10 mel を規準としたものの，高低 2 系統の音程のメロディを持つ f_o 軌跡を生成する．

次に minor の場合，IV# は，移動音階のレ # に相当する．そこを規準に音階の周波数をもとめる．すると，minor：Ascending の I_L-II-III-V-I_H は，ラ（低）・シ・ド・ミ・ラ（高）となる．なお，Descending は，いずれも提示順を反転させたもので，今度は逆に音が下がってくるメロディになる．

なお，この生成の手続きから分かるように，各実験刺激での音階のキーに相当する周波数は，必ずしも A = 440 Hz で調律された音階上にない場合がある．

なお，実験条件 1-4 の最終的な変換音声は，「低スペクトル重心・高 f_o 軌跡」と「高スペクトル重心・低 f_o 軌跡」となる比較対を作る組合せの条件で合成した（**図 4**）．

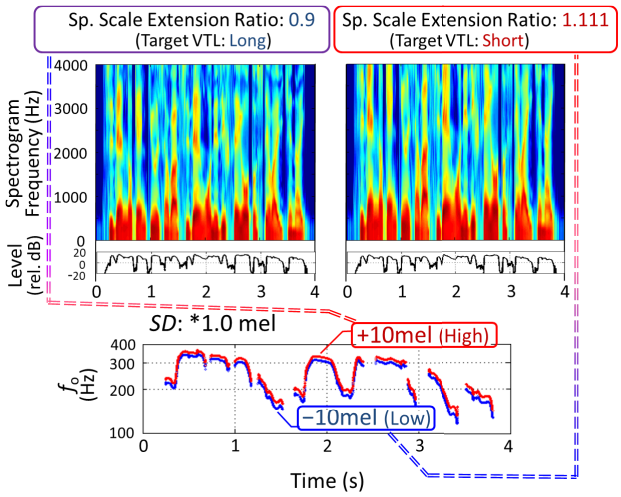

図 4 平滑化スペクトルと f_o 軌跡の組合せの例

Fig. 4 Combination of smoothed-spectra and f_o contours.

(3) 雑音駆動による疑似ささやき声の生成（実験条件 5）

WORLD での音声合成時に，普通なら f_o 軌跡を描く声帯音源を模した音源信号で駆動させるところを，白色雑音で差し換えて合成する．すると，調波構造が存在しない雑音駆動音声が得られる．それを元に，自然なささやき声を模して低域を取り除いた音声（truncated noise-vocoded voice）を生成して，疑似ささやき声として用いた [8]．

まず，原音声のスペクトル周波数軸を圧縮，伸長して得られるスペクトル包絡の低域をそれぞれ減衰させた．次に，それらを雑音で駆動して疑似ささやき声を生成して，比較対を作成した（**図 5**）．ここでの低域抑圧は，音声の変換後に行う信号処理ではなく，スペクトルに直接重みづけをして，合成時に抑圧を実現する．雑音駆動音声で音韻

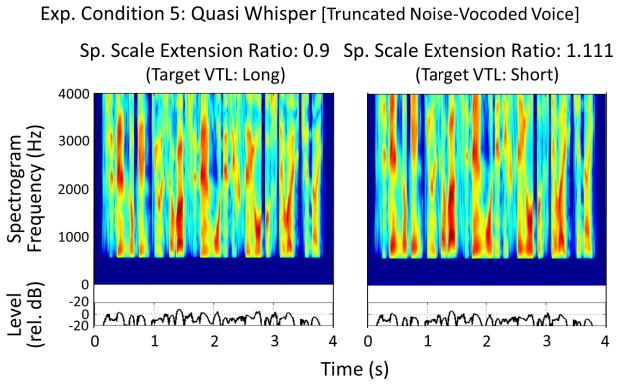

図 5 雑音駆動音声の平滑化スペクトルの低域を抑圧した疑似ささやき声のスペクトルと音圧レベル（W1S5）

Fig. 5 Example truncated-spectra and level of converted quasi whisper (W1S5).

の了解性に重要とされる 570 Hz～1,370 Hz の帯域の範囲を遮断の遷移域に設定して，低域の抑圧を行う．具体的には式 (6) の w 値でスペクトルに直接重みづけをして低域抑圧を行った [8]．

$$w = [(f - 550)/(1350 - 550)]^e \qquad (6)$$

2.4 実験音声の比較対の配置

実験条件 1-4 の比較対は「低スペクトル重心・高 f_o 軌跡」と「高スペクトル重心・低 f_o 軌跡」となる音声の組合せである．これは，スペクトル重心の高低によるピッチ感への影響の方向と，基本周波数の高低による影響の方向とが，互いに逆向きの音声の組合せである．ピッチ感の錯覚は，この組合せの条件のときにみられる．この比較対の相互関係を，スペクトル軸の伸縮率と f_o 昇降のシフト量に基づいて図 6 に示した．

実験条件 5 の疑似ささやき声では f_o が存在しないので，「低スペクトル重心」と「高スペクトル重心」の組合せとなる（図 6）．したがって，この音声比較対でのピッチ感へのバイアス量は，スペクトル重心の違いによって生ずる効果量であると見なすことができる．

実験全体では，80 の変換音声からなる，40 対を用いた．提示順序は，同条件での同性話者による対間で反転させた．

2.5 実験の手続き

参加者 東京都内の 6 つの国立大学の 1 年生 140 名（男性：106 名・女性：34 名，18～23 歳）が参加した．参加者を 10 群に分割して，音声比較対も 10 系列に分けて配置した．

評定項目 音声の比較評価には，声の高さの印象評価を含む，声質表現語 [13] の項目など [2], [4] を用いた（表 2）．

手続き 携帯型の CD プレーヤ（Sony：D-EJ002）から，付属のステレオ・イヤホンで音声比較対を提示した．参加者には，2 人の声の印象について，項目ごとに，どちらの

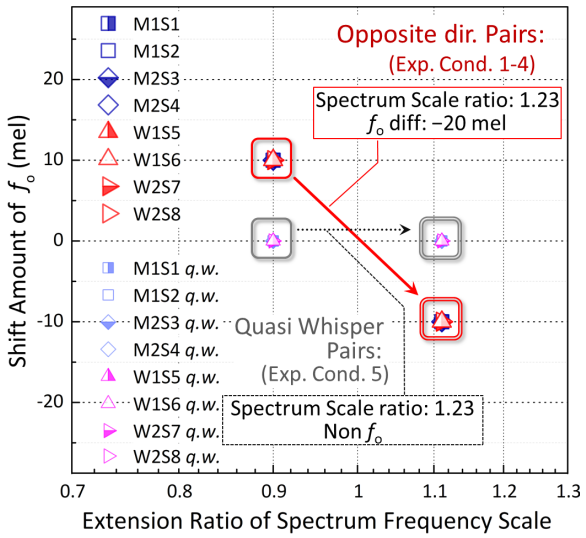

図 6 音声比較対におけるスペクトル周波数軸の伸縮量と f_o シフト量の布置．音声対をつなぐ矢印は，起点のものを評価音声，終点のものを比較音声として分析時に扱うことを示す．提示順序は順序効果の相殺のために同一話者による別スクリプトの音声間で反転している．原音声と提示順序の割当は表 1 に示す（$q.w.$：疑似ささやき声条件）

Fig. 6 Example assignments of comparison pairs ($q.w.$: Quasi Whisper).

表 2 主観評価のための形容詞を用いた評定項目

Table 2 Adjectives to express subjective impression.

Trait	Adjectives
Voice Quality: 声質表現語	高い声，男性的な声，かすれた声，落ち着いた声，迫力のある声，若い感じの声，太い声，張りのある声，鼻声，ささやき声
Naturalness: 自然性	自然な，わかりやすい はっきりした，聞き取りやすい
Speaker's Body Size: 話者の体格	体の大きい，背の高い 太った，がっしりした

声がその項目にあてはまるかを，10 段階で評定するように指示した．実際には，「この課題では，2 人の話し手が発音する文章を続けて一度だけ聞きます。そして，2 人の話し方や話し手の印象について，評価の観点ごとに，どちらの音声がその項目によりあてはまるかを評定します。制限時間に注意すると共に，記入漏れや記入欄の間違いなどがないようにして下さい。」と教示した．なお，対ごとの評定の制限時間は 60 秒であった．

3. 比較評定実験の結果と考察

3.1 評価観点ごとの評価指標と尺度得点

スペクトル重心の影響の方向を固定してとらえるために，音声比較対の中で「高スペクトル重心・低 f_o 軌跡」の声を評価音声とし，もう一方の「低スペクトル重心・高 f_o 軌跡」を対比音声として扱った．

評定結果については，各項目の内容が，評価音声にあてはまるものが 0.5～4.5 点，対比音声にあてはまるものが −0.5～−4.5 点になるよう変換した．

参加者の評定は，1～10 の番号で回答している．その値を，項目の内容に評価音声があてはまると答えた場合に正数，比較音声があてはまるとした場合に負数となるように変換した．そこでは，元となった 10 段階の評定で，どちらがあてはまるかがちょうど拮抗する 5.5 点を，原点となる 0 点とした．すると，評価音声側が 0.5～4.5 点，対比音声側が −0.5～−4.5 点となる．いずれも，その絶対値が大きくなる程，それぞれが項目によりあてはまることを示す．

なお，音声対では，提示順序の影響を相殺するために，評価音声を先に提示する課題と後に提示する課題があった．そこでは，元の回答の 1～10 の値が持つ意味が逆転する．そこで，提示順序を考慮して評価音声にあてはまるものが正数となるように揃えて集計した．

そして，各音声対に割り当てられた参加者による項目別の評定値の平均を，各比較対の項目ごとの評価指標とした．なお，自然性と話者の体格については，各 4 項目からなる合計尺度得点から算出した（表 2）．

3.2 f_o 軌跡の形状の違いによるピッチ感の変動

スペクトル重心の高低と，f_o 軌跡の高低による，ピッチ感への影響が逆向きに相反する条件の下で，f_o の変動幅や形状，時間的安定性や音楽的音階での遷移が，声の高さの評価にどのような影響を与えたか分析した．

また，疑似ささやき声条件での結果から，f_o や調波構造が存在しないスペクトル包絡単独の影響による効果量と，f_o や調波構造が存在する普通の音声での結果を比較した．

ここでは，まず，声質表現語の中の「高い声」の項目の回答から得られた評価指標の分析を行う．声の高さの評価指標の平均について，話者の性別の要因（2 水準：男声，女声）と，f_o の軌跡の要因（5 水準：変動幅（SD）：1.0，0.02，$r.p.$0.4，疑似歌声，疑似ささやき声）の 2 要因分散分析を行った．

まず，性別要因の主効果が有意だった（$F_{(1,30)} = 4.56$，$p > .05$）．男声と女声では，スペクトル周波数軸の伸縮率が 1.23 倍で同一であっても，女声で評価音声がより高いと評価された．すなわち，スペクトル重心の高低による影響，ピッチ感への認知的なバイアスは，女声でより強く見られた．

次に，f_o 軌跡の主効果が高度に有意であった（$F_{(4,30)} = 11.54$，$p > .0001$）．f_o 軌跡全体の高低差は同じでも，f_o 軌跡の形状が異なると．ピッチ感に差異が見られた．なお，性別と f_o 軌跡の交互作用に有意差は見られなかった（$F_{(4,30)} = 0.23$，$n.s.$）．そこで，これ以降は，f_o 軌跡の条件間での多重比較の結果に基づいて検討する．そこで，声の高さの評価指標について，実験条件ごとでの結果を図 7 に示す．

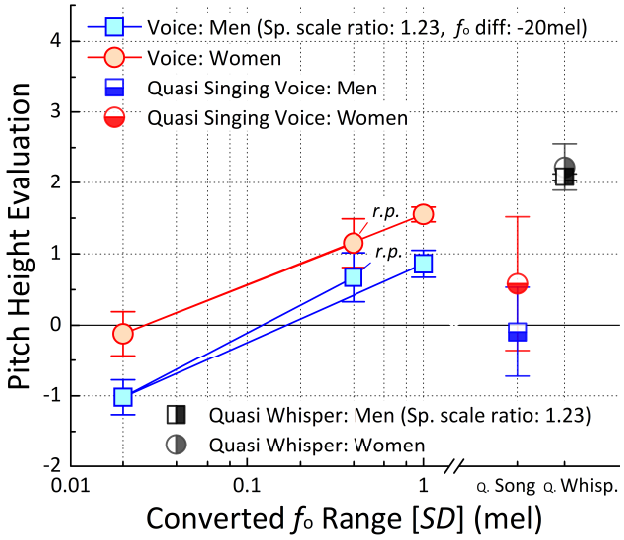

図 7 f_o 軌跡の変動幅別，および，疑似歌声，疑似ささやき声の比較対での声の高さを比較した評定値の平均と標準誤差．横軸は，連続量の範囲は f_o 軌跡の変動幅の大きさで，分割記号より右側は実験条件別のカテゴリ区分．縦軸については，指標が正数方向にあると評価音声（高スペクトル重心・低 f_o 軌跡）の方が，対の比較音声（低スペクトル重心・高 f_o 軌跡）よりも高い声だと評価されたこと示す．逆に，指標が負数方向の場合は，評価音声（高スペクトル重心・低 f_o 軌跡）の方が低いと評価されたことを示す（Sp. Scale ratio：平滑化スペクトル周波数軸の伸縮率，$r.p.$：逆相 f_o 軌跡，Q. Song：疑似歌声条件，Q. Whisp.：疑似ささやき声条件）

Fig. 7 Means and standard errors of pitch scores for converted f_o range, quasi singing-voice and whisper (Sp. Scale ratio: Spectrum Scaling ratio, $r.p.$: reversed-phase f_o contour voice, Q. Song: Quasi Song, Q. Whisp.: Quasi Whisper).

(1) 実験条件 1-3：f_o 変動幅とピッチ感の錯覚

原音声と同じオリジナルの f_o 軌跡の変動幅（SD：1.0 倍）と，それよりやや小さい逆相 f_o 軌跡（$r.p.$）での変動幅（SD：0.4 倍）では，実際には f_o 軌跡の平均が低い評価音声の方が「高い」と評価されており，スペクトル重心が高いものの方が「高い声」だとしていた．すなわち，f_o の高低関係と高さの印象評価が逆転していた．

しかし，f_o 軌跡がフラット（SD：0.02 倍）になると逆転し，f_o 軌跡の高低関係に従う評価に変わって，f_o の高低関係と高さの印象評価の逆転現象が消失した．

なお，ピッチ感の評定指標は，変動幅の対数値に対して線形に上昇していた．そこでは f_o 軌跡の高低反転による輪郭形状（contour）の違いは，必ずしもピッチ感には影響を与えていないと考えられる．すると，f_o の変動幅が，ピッチ感の認知的なバイアス量の変動を引き起こす主たる原因である，と考えるのが順当であろう．

これらの結果は，著者の先行研究の結果とも合致する [4]．しかるに，広帯域の高品質音声においても，f_o の高低関係

と高さの印象評価の逆転現象の生起と消失が確認されたと言ってよい．

(2) 実験条件 4：疑似歌声でのピッチ感の評価の不安定性

疑似歌声では，1 オクターブにおよぶ f_o の変動範囲がある．それにもかかわらず，疑似歌声条件で f_o が低い評価音声は，オリジナルの f_o 軌跡変動幅（SD：1.0 倍）条件の評価音声より，男声では 0.86 点から −0.09 点へと，0.95 点低下した．また，女声でも 1.55 点から 0.58 点へと，0.97 点低下している．したがって，本来 f_o が低い声であるにもかかわらず，高い声だと評価してしまう認知的なバイアスは，疑似歌声では，オリジナルの f_o 軌跡の変動幅の実験条件 1 のものよりも，低減していると考えることができる．

しかし，この疑似歌声条件では，f_o 変動幅の実験条件 1-3 に対して有意な差は見られなかった．標準誤差を見ると，他の条件よりも大きく，条件内でのバラツキがかなり大きかったことがうかがえる．この結果からは，「歌声では，f_o の高低が，ピッチ感への支配的な役割を担っているはずである」，とする当初の想定は，必ずしも完全には支持されなかった．

疑似歌声の下位の生成条件別に集計してみると，上行（Ascending）よりも，下行（Descending）の方が，f_o の高低関係に対応した高低判断がなされる傾向が見られた．しかし，原話者の違いによる影響も見られるため，統一的な解釈には至らなかった．

ここで，ピッチ感の評定指標のバラツキが大きくなった原因を検討してみる．まず，今回の疑似歌声の生成法では，変換音声が真に音楽的な歌声に至っていない可能性がある．そのため，参加者ごとに，「言葉としての音声」として聞く者と，「音楽としての歌声」として聞く者が，分かれてしまっている可能性がある．

また，音楽の歌声として聞いた者であっても，その者が相対音感保有者か，絶対音感保有者かによって，絶対音高での高低判断の精度が異なっている可能性もある．

したがって，歌声を念頭においたアプローチに関しては，上記の点もふまえたうえで，さらなる検証が必要である．

(3) 実験条件 5：疑似ささやき声の高低判断

疑似ささやき条件では，スペクトル重心の高い評価音声でのピッチ感の評定指標が，他のどの条件よりも高かった．疑似ささやき声には f_o が存在しないので，スペクトル包絡の違いが，直接的にピッチ感の評価に反映しているものと考えられる．また，この結果からは，たとえ f_o や調波構造が存在しなくても，音声の高低判断は安定してなされていることが分かる．

この結果は，著者の先行研究で雑音駆動音声を用いた実験での結果と，ほぼ同じ傾向を示す結果であった [4]．この疑似ささやき声は，雑音駆動音声の低域除去によって，本来ならば備えていた情報の一部分が欠落したスペクトル包絡から生成されたものである．しかし，自然なささやき声を模擬する範囲内での情報の欠損であれば，雑音駆動音声と同程度の，かなり強い認知的バイアスが生じることが見出された．

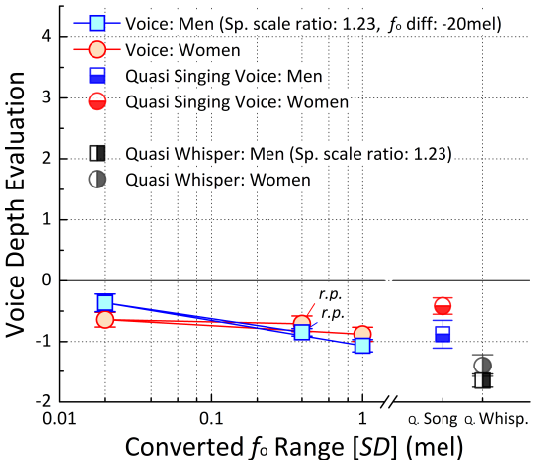

図 8 声の太さに関する評定値の平均と標準誤差

Fig. 8 Means and standard errors of deep voice scores for converted f_o range, quasi singing-voice and whisper.

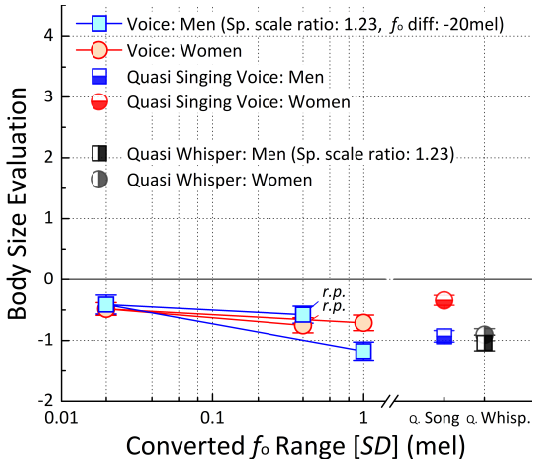

図 9 話者の体格に関する評定値の平均と標準誤差

Fig. 9 Means and standard errors of body-size scores for converted f_o range, quasi singing-voice and whisper.

3.3 f_o 軌跡のピッチ感への影響の選択的特異性

スペクトル周波数軸の伸縮による音色の系統的な変化は，声の太さや，話者の体格の印象評価に影響を与える [1], [2], [4]．そこで，ピッチ感に強い影響を与えていた f_o 軌跡の変動幅や形状が，声の太さや話者の体格にはどんな影響を与えているか検討した（**図 8**，**図 9**）．

ピッチ感と同様の 2 要因の分散分析の結果，スペクトル重心が低いと，声が太く，体格が大きいと評価されていた．しかし，f_o 軌跡の形状の違いによる有意な差は見られなかった．また，話者の性別との交互作用も見られなかった．したがって，f_o 軌跡の違いは，ピッチ感に特徴的な影響を及ぼしていると見られる．

4. まとめと課題

(1) 声のピッチ感の錯覚の生起と消失

この実験から，広帯域の高音質な音声でも，声の音色が異なると，f_o 軌跡が低い音声でも高いと知覚される錯覚が生起することが確認された．しかし，音色が高さの知覚に影響を与えること自体は，無限音階，シェパード・トーンの例からも分かるように，特段に新しい発見ではない [14]．

シェパード・トーンではスペクトル包絡を固定している．その場合，1 オクターブの音が循環して最初の音に戻ったときに，本来の 1 オクターブ下であるべき音に戻ったという聴覚印象が起こらない．これは，通常の場合ならば生じるはずのスペクトル包絡上の不連続性が，最小限に留められているからだとも考えられる．すると，通常のピッチ判断においても，そのようなスペクトル重心の移動が，高さの印象評価の手がかりになっている可能性もある．

それよりもむしろ，f_o 軌跡が単にフラットになるだけで，声の高さの印象が f_o の高低関係に従うようになり，錯覚が消失している．この点については，まだ明確な解釈がなされていない．今後の詳細な検証が望まれる．

(2) ピッチ感の錯覚が生ずる本質的な原因の解明

声のピッチ感の錯覚の生起と消失について，広帯域音声を使った実験によって，その頑健性を示すことができた．しかし，この錯覚のより詳細な生起条件や，消失条件などはいまだに分かっていない．さらに，聴覚処理のどの段階で起こっているのか，そもそも，この現象はどうして起こるのか，この現象は聴覚や音声認知のうえでどのような意味を持つのか，といった，より本質的な部分はまったく分かっていない．

さらに，本実験では，音の高さや音色の明るさなど，類似した印象を与え得る音の物理的な特性が，測定上，「高さ」という 1 つの評定語に集約されている．そのため，どちらかの物理量が判断基準として有用かが逐次的に切り替えられているために生じた結果である可能性もある．これは，ピッチ知覚にかかわる初期の聴覚理論での時間説と場所説の論争 [15] にまで立ち戻って，改めて考え直す必要があるかもしれない．また，この現象が，人の音声知覚に特化した現象である可能性も排除できない．その場合，声のピッチ知覚と，調音器官としての声道長の違いによる音色の知覚との関係についても，再検討が必要になる [16]．課題は山積している．

(3) 聴覚での一般化を視野に入れたピッチ感の錯覚の検討

検討の 1 つの方向として，声を “動的特性を持つ複合音” としてとらえ直してみるのも一考の価値があるかもしれない．「ピッチ感の錯覚の生起と消失」は，音声の場合に限ってだけ起こる現象なのか，それとも，より一般の複合音全般で見られる現象なのか，検討する余地があると思われる．

今回の実験の疑似歌声では，ピッチ感の安定した評価が得られなかった．これからの検証にあっては，実験参加者の音楽経験に関する個別の情報収集は不可欠である．今後さらに，歌声や楽器による楽音などを用いた検証とともに，より厳格な統制が可能な “動的な複合音” を用いた研究が進めば，聴覚の認知的特徴としての一般化を視野に入れた，検証の糸口になるのではないか，と考えるところである．

付記 本研究は，JST さきがけ（JPMJPR18J8），および JSPS 科学研究費補助金（JP15K04103）の援助を受けました．また，本研究の一部は，日本音響学会 2018 年秋季研究発表会，および，音学シンポジウム 2019（第 123 回音楽情報科学研究会・第 127 回音声言語情報処理研究会共催研究会）[17] で発表しました．

また，本稿には，声のピッチ感の錯覚に関連して，一般向けに分かりやすく構成したマルチメディア・データ作品「声色の罠：—高いのに低い声？—」の動画データが添付されています．

参考文献

[1] Smith, D.R.R. and Patterson, R.D.: The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age, *Journal of Acoustical Society of America*, Vol.118, No.5, pp.3177–3186 (2005).
[2] 内田照久：話者の匿名性の確保を目的とした声道長の制御を模した声質変換の評価，日本音響学会誌，Vol.73, No.3, pp.151–162 (2017).
[3] 森勢将雅：音声分析合成，日本音響学会（編）音響テクノロジーシリーズ 22，コロナ社 (2018).
[4] Uchida, T.: Reversal of relationship between impression of voice pitch and height of fundamental frequency: Its appearance and disappearance, *Acoustical Science & Technology*, Vol.40, No.3, pp.198–208 (2019).
[5] 日本基礎心理学会：錯視・錯聴コンテスト 2016（第 8 回），入手先 ⟨http://www.psy.ritsumei.ac.jp/~akitaoka/sakkon/sakkon2016.html⟩（参照 2019-05-16）.
[6] Sundberg, J.: Acoustic and psychoacoustic aspects of vocal vibrato, STL-QPSR, Vol.35, No.2-3, pp.45–68 (1994).
[7] Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J. and Ekelid, M.: Speech recognition with primarily temporal cues, *Science*, Vol.270, No.4234, pp.303–304 (1995).
[8] 内田照久，森勢将雅：実用的なささやき声の生成法：Phantom Silhouette 方式とその評価，日本音響学会 2019 年秋季研究発表会講演論文集，pp.973–976 (2019).
[9] 高橋弘太：高橋弘太研究室音声データベース，入手先 ⟨http://www.it.cei.uec.ac.jp/SRV-DB/⟩ 入手先 ⟨2019-05-16⟩.
[10] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: A vocoder-based high-quality speech synthesis system for real-time applications, *IEICE Trans. Information and Systems*, Vol.E99-D, No.7, pp.1877–1884 (2016).
[11] Morise, M.: D4C, a band-aperiodicity estimator for high-quality speech synthesis, *Speech Communication*, Vol.84, pp.57–65 (2016).
[12] O'Shaughnessy, D.: *Speech Communication: Human and Machine, 2nd edition*, Wiley-IEEE Press, New York, pp.109–139 (1999).
[13] 木戸 博，粕谷英樹：通常発話の声質に関連した日常表現語—聴取評価による抽出，日本音響学会誌，Vol.57, No.5,

pp.337–344 (2001).
[14] Shepard, R.N.: Circularity in judgements of relative pitch, *Journal of Acoustical Society of America*, Vol.36, No.12, pp.2346–2353 (1964).
[15] 大串健吾：音のピッチ知覚，日本音響学会（編）音のサイエンスシリーズ 15，コロナ社 (2016).
[16] Gaudrain, E.: Can spectral centroid explain voice pitch and vocal-tract length perception in normal-hearing and cochlear implant listeners?, *Journal of the Acoustical Society of America*, Vol.140, No.4, p.3439 (2016).
[17] 内田照久：声色の罠〈声のピッチ感の錯覚〉：疑似歌声・疑似ささやき声による検討，情報処理学会研究報告，Vol.2019-SLP-127, No.32, pp.1–6 (2019).

内田 照久

1964 年生．1988 年名古屋大学卒業．1993 年同大学大学院博士課程修了．1996 年博士（教育心理学）取得．1994 年大学入試センター研究開発部助手．2017 年より同教授．リスニングテストの開発，および，聴覚心理の研究に従事．

森勢 将雅 （正会員）

1981 年生．2004 年和歌山大学卒業．2008 年同大学大学院博士課程修了，博士（工学）取得．関西学院大学，立命館大学，山梨大学を経て，2019 年より明治大学総合数理学部専任准教授．人間の知覚情報を活用した音声分析・合成・デザインの研究に従事．